ISSN 1847-3946

mipro 2020

organizer mipro



43rd

international convention

September 28 - October 2, 2020, Opatija, Croatia

Lampadem tradere



mipro - innovative promotional partnership

mipro proceedings

My profession.
My organization.
My IEEE.



Discover the benefits of IEEE membership.

Join a community of more than 365,000 innovators in over 150 countries. IEEE is the world's largest technical society, providing members with access to the latest technical information and research, global networking and career opportunities, and exclusive discounts on education and insurance products.

Join today
www.ieee.org/join

IEEE

MIPRO 2020

43rd International Convention

September 28, 2020 – October 2, 2020
Opatija, Croatia

Proceedings

Conferences:

Microelectronics, Electronics and Electronic Technology /MEET

Data Science and Biomedical Engineering /DS-BE

Telecommunications & Information /CTI

Computers in Education /CE

Computers in Technical Systems /CTS

Intelligent Systems /CIS

Robotics Technologies and Applications/RTA

Information Systems Security /ISS

Business Intelligence Systems /miproBIS

Digital Economy and Digital Society /DE-DS

Information and Communication Technology Law /ICTLAW

Engineering Education /EE

Software and Systems Engineering /SSE

MIPRO Junior - Student Papers /SP

Smart, Sustainable And Resilient Cities And Infrastructure /SSRCI

Optoelectronics and Photonics /OPHO

Dew Computing /DEWCOM

Edited by:
Karolj Skala

International Program Committee

Karolj Skala, General Chair (Croatia)
Lejla Abazi-Bexheti (North Macedonia)
Enis Afgan (United States),
Miimu Airaksinen (Finland),
Saša Aksentijević (Croatia),
Slaviša Aleksić (Germany),
Slavko Amon (Slovenia),
Krešo Antonović (Croatia),
Michael E. Auer (Austria),
Viktor Avbelj (Slovenia),
Dubravko Babić (Croatia),
Snježana Babić (Croatia),
Tadej Bajd (Slovenia),
Ante Bakić (Croatia),
Marko Banek (Croatia),
Mirta Baranović (Croatia),
Bartosz Bebel (Poland),
Nina Begičević Ređep (Croatia),
Ladjel Bellatreche (France),
Adrian Boukalov (Belgium),
Ricardo Branco (Portugal),
Ljiljana Brkić (Croatia),
Marian Bubak (Poland),
Andrea Budin (Croatia),
Željko Butković (Croatia),
Patrizio Campisi (Italy),
Željka Car (Croatia),
Jesús Carretero Pérez (Spain),
Bojan Cukic (United States),
Alfredo Cuzzocrea (Italy),
Duško Čakara (Croatia),
Stipo Čelar (Croatia),
Marina Čičin-Šain (Croatia),
Dragan Čišić (Croatia),
Davor Davidović (Croatia),
Vlado Delić (Serbia),
Radoslav Delina (Slovakia),
Matjaž Depolli (Slovenia),
Saša Dešić (Croatia),
Dražen Dragičević (Croatia),
Todd Eavis (Canada),
Maurizio Ferrari (Italy),
Tiziana Ferrari (Netherlands),
Nikola Filip Fijan (Croatia),
Renato Filjar (Croatia),
Tihana Galinac Grbac (Croatia),
Enrico Gallinucci (Italy),
Dragan Gamberger (Croatia),
Paolo Garza (Italy),
Tom Gavazzi (Croatia),
Gordan Gledec (Croatia),
Matteo Golfarelli (Italy),
Stjepan Golubić (Croatia),
Vera Gradišnik (Croatia),
Simeon Grazio (Croatia),
Andrej Grgurić (Croatia),
Stjepan Groš (Croatia),
Nina Gumzej (Croatia),
Marjan Gusev (North Macedonia),
Jaak Henno (Estonia),
Bojan Hlača (Croatia),
Željko Hocenski (Croatia),
Tatjana Holjevac (Croatia),
Vlasta Hudek (Croatia),
Darko Huljenić (Croatia),
Robert Inkret (Croatia),
Ivo Ipšić (Croatia),
Mile Ivanda (Croatia),
Marina Ivašić-Kos (Croatia),
Hannu Jaakkola (Finland),
Tomislav Jagušt (Croatia),
Darko Jardas (Croatia),
Vojko Jazbinšek (Slovenia),
Leonardo Jelenković (Croatia),
Bojan Jerbić (Croatia),
Dragan Jevtić (Croatia),
Alen Jugović (Croatia),
Admela Jukan (Germany),
Oliver Jukić (Croatia),
Irena Jurdana (Croatia),
Ozren Jureković (Croatia),
Marko Jurić (Croatia),
Đani Juričić (Slovenia),
Nikola Kadoić (Croatia),
Jurij Matija Kališnik (Slovenia),
Ivan Kaštelan (Serbia),
Zlatko Katalenić (Slovenia),
Ana Katalinić Mucalo (Croatia),
Tihomir Katulić (Croatia),
Pekka Kess (Finland),
Tonimir Kišasondi (Croatia),
Zalika Klemenc-Ketiš (Slovenia),
Mario Konecki (Croatia),
Marko Koričić (Croatia),

Gregor Kosec (Slovenia),
Igor Kotenko (Russia),
Božidar Kovačić (Croatia),
Miklos Kozlovszky (Hungary),
Danica Kragić Jensfelt (Sweden),
Goran Krajačić (Croatia),
Dieter Kranzlmüller (Germany),
Marjan Krašna (Slovenia),
Srećko Krile (Croatia),
Lene Krøl Andersen (Denmark),
Benjamin Kušen (Croatia),
Marko Lacković (Croatia),
Erich Leitgeb (Austria),
Jadran Lenarčič (Slovenia),
Tomislav Lipić (Croatia),
Hrvoje Lisičar (Croatia),
Dražen Lučić (Croatia),
Duško Lukač (Germany),
Igor Ljubi (Croatia),
Zongmin Ma (China),
Goran Marković (Croatia),
Leslie Martinich (United States),
Ludek Matyska (Czech Republic),
Mladen Mauher (Croatia),
Igor Mekterović (Croatia),
Željka Mihajlović (Croatia),
Branko Mikac (Croatia),
Anđelko Milardović (Croatia),
Hrvoje Mlinarić (Croatia),
Thor Moen (Norway),
Lueny Morell (Puerto Rico),
Gorana Mudronja (Croatia),
Neeta Nain (India),
Jadranko F. Novak (Croatia),
Dario Ogrizović (Croatia),
Predrag Pale (Croatia),
Dana Paľová (Slovakia),
Nikola Pavešić (Slovenia),
Mile Pavlić (Croatia),
Branimir Pejčinović (United States),
Ana Perić Hadžić (Croatia),
Dana Petcu (Romania),
Juraj Petrović (Croatia),
Duc Truong Pham (UK),
Damir Pintar (Croatia),
Vincenzo Piuri (Italy),
Tonka Poplas Susič (Slovenia),
Andreja Pucihar (Slovenia),
Aleksandra Rashkovska (Slovenia),
Robert Repnik (Slovenia),
Libuša Révészová (Slovakia),
Slobodan Ribarić (Croatia),
Vittorio Rosato (Italy),
Dubravko Sabolić (Croatia),
Ioan Sacala (Romania),
Davor Salamon (Croatia),
Jörg Schulze (Germany),
Bruno Siciliano (Italy),
Sandro Skansi (Croatia),
Zoran Skočir (Croatia),
Ivanka Sluganović (Croatia),
Mladen Sokele (Croatia),
Ana Sović Kržić (Croatia),
Mario Spremić (Croatia),
Vlado Sruk (Croatia),
Uroš Janez Stanič (Slovenia),
Vjeran Strahonja (Croatia),
Tomislav Suligoj (Croatia),
Aleksandar Szabo (Croatia),
Dina Šimunić (Croatia),
Frano Škopljanac-Mačina (Croatia),
Dejan Škvorc (Croatia),
Zorislav Šojat (Croatia),
Andreja Špernjak (Slovenia),
Vitomir Štruc (Slovenia),
Velimir Švedek (Croatia),
Darko Švelec (Croatia),
Viktor Švigelj (Slovenia),
Zheng-Hua Tan (Denmark),
Nikola Tanković (Croatia),
Antonio Teixeira (Portugal),
Edvard Tijan (Croatia),
Paul Timmers (UK),
A Min Tjoa (Austria),
Ivan Tomašić (Sweden),
Roman Trobec (Slovenia),
Tibor Vámos (Hungary),
Mladen Varga (Croatia),
Lucija Vejmelka (Croatia),
Matjaž Veselko (Slovenia),
Linda Vicković (Croatia),
Marijana Vidas-Bubanja (Serbia),
Davor Vinko (Croatia),
Goran Vojković (Croatia),
Mihaela Vranić (Croatia),
Miroslav Vrankić (Croatia),
Boris Vrdoljak (Croatia),
Slavomir Vukmirović (Croatia),
Yingwei Wang (Canada),
Mario Weber (Croatia),
Roman Wyrzykowski (Poland)

organized by

MIPRO Croatian Society

technical cosponsorship

IEEE Region 8
IEEE Croatia Section
IEEE Croatia Section Computer Chapter
IEEE Croatia Section Electron Devices/Solid-State Circuits Joint Chapter
IEEE Croatia Section Education Chapter
IEEE Croatia Section Communications Chapter
EAI European Alliance of Innovation

under the auspices of

Ministry of Science and Education of the Republic of Croatia
Ministry of the Sea, Transport and Infrastructure of the Republic of Croatia
Ministry of Economy, Entrepreneurship and Crafts of the Republic of Croatia
Ministry of Public Administration of the Republic of Croatia
Ministry of Regional Development and EU Funds of the Republic of Croatia
Ministry of Environment and Energy of the Republic of Croatia
Ministry of Demography, Family, Youth and Social Policy of the Republic of Croatia
Ministry of Agriculture of the Republic of Croatia
Primorje-Gorski kotar County
City of Rijeka
City of Opatija
Croatian Regulatory Authority for Network Industries - HAKOM
Croatian Power Exchange - CROPEX

patrons

University of Zagreb
University of Rijeka
Juraj Dobrila University of Pula
Ruđer Bošković Institute, Zagreb
University of Zagreb, Faculty of Electrical Engineering and Computing
University of Zagreb, Faculty of Organization and Informatics, Varaždin
University of Rijeka, Faculty of Maritime Studies
University of Rijeka, Faculty of Engineering
University of Rijeka, Faculty of Economics and Business
Zagreb University of Applied Sciences
Croatian Academy of Engineering - HATZ
Croatian Regulatory Authority for Network Industries - HAKOM
Ericsson Nikola Tesla, Zagreb
T-Croatian Telecom, Zagreb
Končar - Electrical Industries, Zagreb
HEP - Croatian Electricity Company, Zagreb
A1 Hrvatska, Zagreb

sponsors

HEP - Croatian Electricity Company Zagreb
Končar-Electrical Industries Zagreb
Storm Computers Zagreb
InfoDom Zagreb
A1 Hrvatska Zagreb
Mjerne tehnologije Zagreb
Selmet Zagreb
Institute SDT Ljubljana
Nomen Rijeka

All papers are published in their original form

For Publisher:

Karolj Skala

Publisher:

Croatian Society for Information, Communication and
Electronic Technology – MIPRO
Office: Kružna 8/II, P. O. Box 303, HR-51001 Rijeka, Croatia
Phone/Fax: (+385) 51 423 984

Printed by:

GRAFIK, Rijeka

ISSN 1847-3946

Copyright © 2020 by MIPRO

All rights reserved. No part of this book may be reproduced in any form, nor may be stored in a retrieval system or transmitted in any form, without written permission from the publisher.

CONTENTS

LIST OF PAPER REVIEWERS

LIST OF AUTHORS

FOREWORD

MICROELECTRONICS, ELECTRONICS AND ELECTRONIC TECHNOLOGY

INVITED PAPER

Infrared Quantum Detectors 5
D. Grubišić, D. Babić

PAPERS

Ge-on-Si Avalanche Photodiodes for LIDAR Applications 15
M. Wanitzek, M. Oehme, D. Schwarz, K. Guguieva, J. Schulze

Processing Sequence for a PureB Bipolar Junction Transistor 20
A. Čaušević, H.S. Funk, D. Schwarz, K. Guguieva, J. Schulze

Carrier Mobilities in Heavily Doped Pseudomorphic Ge1-xSnx-epilayers 24
M.M. Dettling, D. Weißhaupt, H.S. Funk, M. Kern, F. Berkmann, C. Clausen, M. Oehme, D. Schwarz, J. v. Slageren, J. Schulze

On the Modelling of Interface Roughness Scattering in AlGaN/GaN Heterostructures 29
I. Berdalović, M. Poljak, T. Suligoj

Modeling of Electrical Properties of Al-on-Ge-on-Si Schottky Barrier Diode 35
L. Marković, T. Knežević, T. Suligoj

Material and Device Properties of Bismuthene Nanoribbons from Multi-orbital Quantum Transport Simulations 41
M. Poljak

Single-band Quantum Transport Study of Resonant Tunneling Diodes Based on Silicene Nanoribbons 46
M. Mihaljević, M. Širić, M. Poljak

Predicting the Transport Properties of Silicene Nanoribbons Using a Neural Network 51
T. Župančić, I. Stresec, M. Poljak

Introduction to the Design and Simulation of Reflectionless Filters 56
G. Lemire, B. Pejčinović

Microwave Interferometry Measurements of Yeast Cell Suspension and Sediment Process 61
M. Zhang, T. Marković, B. Nauwelaers

Energy Harvesting on Power Amplifiers Based on Application of Thermoelectric Generators 65
I. Škalic, I. Marinović

Doherty Power Amplifier in Horizontal Current Bipolar Transistor (HCBT) Technology 69
Ž. Osrečki, J. Žilak, M. Koričić, T. Suligoj

Design Considerations for Mid-Power Receiver in Resonant Wireless Power Transfer System 74
D. Vinko, D. Bilandžija I. Biondić

Semi-Analytical Estimation of On-Chip Intertwined Rectangular Transformer Parameters in 180 nm CMOS Technology 78
I. Brezovec, J. Mikulić, G. Schatzberger, A. Barić

Multi-Topology DC-DC Converter for Low-Voltage Energy Harvesting Systems 84
M. Kováč, D. Arbet, L. Nagy, M. Šovčík, V. Stopjaková

Design and Measurements of Low Power 32-kHz Oscillators and a Test Interface in 180-nm CMOS Technology 90
I. Kuljak, I. Tomić, R. Bertolan, J. Mikulić, G. Schatzberger, J. Fellner, A. Barić

Programmable Low-Frequency Divider in 180-nm CMOS Technology 96
D. Tomić, J. Mikulić, G. Schatzberger, J. Fellner, A. Barić

Design of a Tunable Temperature Coefficient Voltage Reference with Low-Dropout Voltage Regulator in 180-nm CMOS Technology 100
D. Osmanović, I. Skeledžija, K. Špoljarić, D. Tomić, J. Mikulić, G. Schatzberger, J. Fellner, A. Barić

Mismatch Reduction Techniques for Current-Mirror Based Potentiostats 106
M. Haberler, I. Siegl, C. Steffan, M. Auer

Design of CMOS Temperature Sensors Based on Ring Oscillators in 180-nm and 110-nm Technology 111
I. P. Tolić, J. Mikulić, G. Schatzberger, A. Barić

A Circuit for Identifying Oldest Ready Instructions in Reservation Stations 116
D. Spasov

An Improvement in the Convergence of Superscalar Processors 121
D. Spasov

Sequential Register Renaming 125
D. Spasov

Improved Circuits for a Random Pulse Computer 130
M. Batelić, M. Stipčević

Two Approaches for Solving Rubik's Cube with Hardware-Software Co-design 135
E. Baručija, A. Akagić, S. Ribić, Ž. Jurić

A Methodology for Frequency-Measurement Characterization Based on Increment of Input Signal Phase 141
A.N. Serov

Comparative Analysis of the Active Power Measurement Methods in Time Domain 147
A.N. Serov, A.A. Shatokhin, N.A. Serov

Comparison of Pearson Correlation Coefficient and Distance Correlation in Correlation Power Analysis on Digital Multiplier 153
J. Kundrata, D. Fujimoto, Y. Hayashi, A. Barić

A Realization of Adaptive Compressive Sensing System 159
K. Sever, T. Vlašić, D. Seršić

Delta/Alpha Ratio for Diagnosis of Autism Spectrum Disorder 165
Z. Šverko, S. Vlahinić, M. Vrankić, I. Markovinović

Group Delay of Fractional n+alpha-Order Bessel Filters 170
A. Ćoza, V. Županović, D. Vlah, D. Jurišić

DATA SCIENCE AND BIOMEDICAL ENGINEERING

INVITED PAPER

Information Communication Society - a Possible Ecology of the Future 181
Z. Šojat

PAPERS

DATA SCIENCE

Comparison of Machine Learning Methods in Classification of Affective Disorders 193
I. Kinder, K. Friganović, J. Vukojević, D. Mulc, T. Slukan, D. Vidović, P. Brečić, M. Cifrek

Evaluating Microsoft Face API in the Context of Student Classroom Attendance 198
M. Marjanović, T. Kramberger, R. Kramberger, I. Cesar

A Scalable K-Nearest Neighbor Algorithm for Recommendation System Problems 202
A. Sagdić, C. Tekinbas, E. Arslan, T. Kucukyilmaz

Retinal OCT Image Segmentation: How Well do Algorithms Generalize or How Transferable are the Data? 208
M. Melinščak, S. Lončarić

Smartphone Based Range of Motion Measurement in Physiotherapy 213
M. Njirjak, E. Otović, M. Budimir, H. Vlahović, M. Tomić, V. Marijančić

Comparative Analysis of Network Embeddings for Functional Annotation in Protein Interaction Networks 219
I. Ivanoska, E. Petreska, B.R. Stojkoska, S. Kalajdziski, K. Trivodaliev

Modernization of the PIC Codes for Exascale Plasma Simulation 225
I. Vasileska, P. Tomšič, L. Kos

GPU Accelerated RBF-FD Solution of Poisson's Equation 230
M. Jančič, J. Slak, G. Kosec

Parallel Point Sampling for 3D Bodies 235
U. Duh, M. Depolli, J. Slak, G. Kosec

Sarajevo War Tunnel - Mobile Virtual Cultural Heritage Application 240
I. Prazina, I. Ivković-Kihić, T.A. Chahin, A. Jajčanin, S. Rizvić, V. Okanović

Modeling Agricultural Production Activities Using Weather and Soil Parameters 244
T. Kovačević, L. Mrčela, A. Merćep, Z. Kostanjčar

Improving 3D Printing of Garments by Using HPC Cloud 249
D. Tomić, D. Davidović, T. Šubić, J. Mesarić, K. Skala

Global Repeat Map Algorithm (GRM) Reveals Differences in Alpha Satellite Number of Tandem and Higher Order Repeats (HORs) in Human, Neanderthal and Chimpanzee Genomes – Novel Tandem Repeat Database 253
I. Vlahović, M. Glunčić, K. Dekanić, L. Mršić, H. Jerković, I. Martinjak, V. Paar

Air Traffic Control Competencies Measurements Based on Functional Near-Infrared Spectroscopy 259
I. Kesedžić, J. Božek, M. Radoš, S. Popović, K. Ćosić

Public Transport Analyzing Using Bee Colony 265
M. Babič, B. Soldo, J. Povh

Measure for Traffic Anomaly Detection on the Urban Roads Using Speed Transition Matrices 268
L. Tišljarić, Ž. Majstorović, T. Erdelić, T. Carić

Calculating the Fire Weather Index for US Wildfires Using a GPU 276
J. Kuzmanova, M. Gusev, V. Zdraveski

Iris Identification Using Wavelet Wecomposition and Gabor Filter 2816
H. Nitz Petterson, J. Rehnholm, S. Vikström, M. Åslund, E. Åstrand, I. Tomašić

An Analytics Toolbox for Cyber-Physical Systems Data Analysis: Requirements and Challenges 287
M. Zanin, E. Menasalvas, A. Rodriguez González, P. Smrz

Assisting Motorists Using Parking Prediction through a Car App 293
C. Attard, A. Naudi, S. Mallia, D.Gauci, R. Farrugia

BIOMEDICAL ENGINEERING

Automatic Brain Extraction in Nissl-Stained Histological Sections 301
T. Smolčić, T. Petković, S. Škokić

Laser Speckle Stereo System in Biomedical Imaging 306
F. Salgueiro, P.A. Assunção, R. Fonseca-Pinto

Lung Cancer Typology Classification Based on Biochemical Markers Using Machine Learning Techniques 310
J. Nuhić, J. Kevrić

Self-Help for Elderly Individuals with Morning Gymnastics "1000 Movements" and "Method 23+5" 316
B. Širaiy, U.J. Stanič, Z. Katkič, N. Grishin

Design of a Non-invasive ECG-based Glucose Measurement System 321
M. Gusev, E. Guseva, L. Poposka

Analysis of Diabetes Dataset 327
L. Beqiri, A. Velinov, B. Fetaji, L. Loku, A. Buçuku, Z. Zdravev

Trends from Minimally Invasive to Non-invasive Glucose Measurements 333
M. Gusev, L. Poposka, E. Guseva, M. Kostoska, B. Koteska, M. Simjanoska, N. Ackovska, A. Stojmenski

Respiration Extraction from Single-Channel ECG using Signal-Processing Methods and Deep Learning 339
E. Merdjanovska, A. Rashkovska

Simultaneous Phonocardiography and Electrocardiography Using Smartphone in Dogs, Cats and Horses 345
V. Kadunc Kos, M. Brložnik, A. Domanjko Petrič, V. Avbelj

Phonocardiography and Electrocardiography with a Smartphone 350
V. Avbelj, M. Brložnik

Detection of Uninterpretable ECG Signal Segments 355
E. Ajdaraga Krluku, M. Gusev

A Programmatic Approach for Development of the ViewHRV Service Platform with Accurate and Reliable Results 361
E. Shaqiri, M. Gusev

Impact of Subthreshold Transcutaneous Auricular Vagus Nerve Stimulation on the Heart Rate Variability and Atrial Arrhythmias 367
R. Trobec, M. Lindén, M. Šinkovec

Clinical Decision Support Systems in Practice: Current Status and Challenges 373
A. Jović, I. Stančin, K. Friganović, M. Cifrek

MOVIDA.eros : an eHealth Solution for Cardiac Rehabilitation Programs 379
R. Fonseca-Pinto, E. Silva, R. Martinho, R. Rijo, F. Januário, A. Antunes

Smart Healthcare Ecosystem for Elderly Patient Care 383
A. Abugabah, N. Nizamuddin

High-Throughput Virtual Screening Web Service Development and its Applications to SARS-CoV-2 Target Protein Inhibitors Drug Design 389
V. Miletić, M. Ašenbrener Katić, Ž. Svedružić

Drug Screening with the Autodock Vina on a Set of Kinases without Experimentally Established Structures 395
D. Tomić, D. Davidović, V. Janđel, J. Mesarić, K. Skala, T. Lipić

TELECOMMUNICATIONS & INFORMATION

INVITED PAPERS

Cybersecurity Risk is a Board-Level Issue 407
B. Cerin

Govorna komunikacija čovjek-stroj X & Y vs Z & α generacije 409
V. Delić

PAPERS

TOWARDS 5G NETWORKS

LTE and 5G NR – Coexistence and Collaboration 415
T.B. Iliev, G.Y. Mihaylov, I.S. Stoyanov, E.P. Ivanova

Decoding of LTE Turbo Codes Initialized with the Two Recursive Convolutional Codes 419
D. Spasov

Evaluation of Commercial Off-The-Shelf LTE Antennas for Use on Unmanned Aerial Vehicles 423
K. Kainrath, H. Flühr, W. Stocksreiter, R. Findenig, E. Leitgeb, M. Gruber

Techno-Economic Analysis of Deployment Options for Converged 5G Wireless-Optical Access Networks 427
M. Bongard, K. Grunert, S. Aleksić

Resource Utilization – QoS Isolation Trade-Off in 5G Networks Considering Network Slicing Reconfiguration Interval 433
S. Demirović, P. Kiš, J. Janković, Ž. Ilić

Bounding Reliability in Service Function Chaining 439
A. Engelmann, W. Bziuk, A. Jukan

INTERNET DEVELOPMENT & ENERGY EFFICIENCY

IoT Cloud-Based Services in Network Management Solutions 447
O. Jukić, I. Heđi, E. Ciriković

Innovation Platfirm – A Novel Energy Service Utility 453
A. Višković, D. Šimunić, V. Franki

Modeling of Energy Consumption of Sensor Nodes 459
F. Tsvetanov, I. Georgieva

Čimbenici adresiranja i povezivanja za IoT usluge i aplikacije 465
D. Švelec, M. Blažeković

SMART CITIES & CHALLENGES

Establishing the Utility Charges Spatial Database Using Digital Twin Technology 473
V. Mihoković, L. Zalović, V. Zalović

Smart Home IVR-Based System with South Slavic Language Integration 478
M. Spahić, A. Šečerbegović, V. Mesić, H. Hadžić, A. Hasanbašić, O. Jahić

Software Prototype Based on Augmented Reality for Mastering Vocabulary 482
M. Begić, M. Cirimotić, I. Farkaš, I. Skorić, Ž. Car, I. Rašan, M. Žilak

The Perception of the Fake News Phenomenon on the Internet by Members of Generation Z 488
L. Prelog, Lj. Bakić-Tomić

Smart Cities as an Opportunity and Challenge for People with Disabilities 492
D. Švelec, N. Bjelčić, M. Blažeković

COMPUTING IN TELECOMMUNICATIONS

Cloud Computing Virtualization: A Comprehensive Survey 501
A. Rista, J. Ajdari, X. Zenuni

Algorithms for Computing in Fog Systems: Principles, Algorithms, and Challenges 512
N. Soni, R. Malekian, D.C. Bogatinoska

DNN Placement and Inference in Edge Computing 518
M. Bensalem, J. Dizdarević, A. Jukan

MACHINE LEARNING APPLICATIONS

Predicting and Analyzing Absenteeism at Workplace Using Machine Learning Algorithms 527
A. Rista, J. Ajdari, X. Zenuni

Example of Using Algorithms for Switching Hybrid FSO/RF Systems 533
R. Haluška, Ľ. Ovseník, P. Šuľaj

Crime Analysis and Prediction Using Machine Learning 538
O. Llaha

A Machine Learning Approach for Analysis of Spectrum Availability in Kosovo based on Experimental Measurements 544
Z. Limani Fazliu, H. Maloku, M. Ibrani, M. Limani, B. Gashi

SIGNAL PROCESSING

The Influence of Multipath Propagation of the Signal on the Accuracy of the GNSS Receiver 553
T.B. Iliev, I.S. Stoyanov, S.A. Sokolov, I.H. Beloev

Error Correction with Systematic RLNC in Multi-Channel THz Communication Systems 557
C.V. Phung, A. Engelmann, A. Jukan

Measurement-Based Optimized Propagation Model for Urban, Suburban and Rural Environments for UHF Bands in Kosovo 563
H. Maloku, Z. Limani Fazliu, M. Ibrani, M. Limani, B. Gashi

BUSINESS PROCESSES & DATABASES

Transforming Product Catalogue Relational into Graph Database: a Performance Comparison 571
J. Lorincz, V. Huljić, D. Begušić

Značajka analitičkog pristupa u dizajnu, optimizaciji i održavanju poslovnih procesa 577
D. Markulin, K. Musa, A. Mršić

Komparativna analiza upravljanja obrtnim kapitalom u odjeljcima područja djelatnosti Informacije i komunikacije u Republici Hrvatskoj 582
D. Korent

Pregled performansi djelatnosti J61 Telekomunikacije u Republici Hrvatskoj 588
M. Bubanić

IMPLEMENTATIONS OF STANDARDS & BEST PRACTICES

Implementation of NETCONF Standard by Major Customers in Croatia 597
D. Valenčić

Vendors' Implementation of NETCONF Standard on Routers and Switches 604
D. Valenčić

Dohvat novih ključeva za mobilnu aplikaciju na bankomatu prema ITIL najboljoj praksi 610
D. Muratović, R. Barišić

Uvođenje usluge bežičnog pristupa internetu pomoću ITIL smjernica 616
K. Brodnjak, R. Barišić

COMPUTERS IN EDUCATION

PAPERS

Parents' Attitudes toward Programming in Elementary Schools in City of Osijek 627
A. Papić, I. Ćosić

Artificial Intelligence and Education 633
H. Jaakkola, J. Henno, A. Lahti, J.P. Järvinen, J. Mäkelä

Teaching Physics Using Programming of Simulations 641
R. Repnik, P. Bernad, M. Krašna

Improving the Quality of Entrepreneurial Education by ICT Education of HEI Pedagogical Staff 649
D. Paľová, M. Vejačka

How to Educate Students for the Future? 655
L. Révészová

Game Design Based Learning of Programming for Girls 661
M. Bevčič, J. Rugelj

Gamification Tools Improving University Students' Involvement in the Education Process 666
D. Paľová, M. Vejačka

Continuous Summative Assessment Sessions as a Motivational Tool for STEM students: a Case Study 672
G. Đurović

Increasing the Learning Efficiency in Decision-Making Field Using the Workshop Activity in Moodle 677
N. Kadoić, B. Šlibar

Studying Dynamic Mathematics Software in the Professional Training of Teachers of Computer Science, Mathematics, and IT Specialists 683
I.V. Shishenko, V.H. Shamonia, V.S. Loboda, V.V. Punko, Yu.V. Khvorostina, A.A. Voitenko

The Automated Control of Students Achievements by Using Paper Clicker Plickers 688
M.G. Drushlyak, O.V. Semenikhina, S.M. Kondratiuk, T.M. Kryvosheya, A.V. Vertel, N.M. Pavlushchenko

The Use of Electronic Textbooks in the Learning Process: a Statistical Analysis 693
O.M. Udovychenko, M.M. Ostroha, A.E. Chernysh, O. Kudrina, Yu.A. Bondarenko, A.V. Kurienkova

Influence of Moodle and MS Teams on Teaching-Learning-Studying (TLS) Processes 697
M. Krašna, I. Pesek

The Impact of Distance Learning on Student Success for Electrical Engineering Professional Courses 702
M. Sokele, T. Alajbeg, F. Brkić

Use of Open Training Portals to Host Developed STEM Courses 707
T. Hryhorova, V.P. Lyashenko, I. Hvozdeva, I. Getman

Internationalization and Student Mobility: STEAM Students Case Study 711
P. Cuculić, T. Babić, M. Pavlica

Educating the Educators for Introducing Internet of Things to Primary and Secondary Schools' Curriculums 717
A. Jaklič

An Algorithm for Assessment of Students Using Gamification 721
D. Keremedchiev, D. Borissova, G. Tuparov

Effective Decision Making: the Added Value of Including Humanities in STEM Studies 726
M. Pavlica, T. Babić, P. Cuculić

Using BBC Micro:bit in Primary and Secondary Schools for Creating Simple Smart Home 733
P. Voštinár, J. Knežník

Using mBot Robots for the Motivation of Studying Computer Science 738
P. Voštinár

Investment of European Enterprises in ICT Education of Their Employees: Relationship with Country Competitiveness and Digital Economy Development 743
M. Pejić Bach, J. Zoroja, I. Strugar

Project Based Teaching with Digital Tools in Primary Education 749
J. Mezak, P. Pejić Papak

University Teachers' Needs for the Efficient Use of ICT in the Pedagogical Process 754
N. Špur, E. Kranjec, M. Puhek, K. Breznik, L. Klasinc, S. Frumen

Analysis of Social Networks Usage Among Students 760
M. Lončar, D. Kermek

Digital Technologies in Education of Preschool Children – Preparing for Future 766
J. Gunčaga, E. Severini, B. Kožík Lehotayová, P. Ostradický

Impact of Education on Communication Etiquette 774
G. Kirinić, Lj. Bakić-Tomić

Harmonization of Curriculum with Needs and Requests of Fourth Industrial Revolution: Case of Faculty of Economics and Business Rijeka 779
H. Bezić, D. Balaž, B. Buljat

A Comparative Study of Gamification in Progamming Education in a Croatian High School 785
M. Schatten, M. Schatten

Computer-Based Education in the Course "Digital Electronics" Teaching the Topic "Adders-Subtractors" 790
A.N. Borodzhieva, I.I. Stoev, I.D. Tsvetkova, S.L. Zaharieva, V.A. Mutkov

Low Cost Rapid Control Prototyping – a Useful Method in Control Engineering Education 796
W. Werth, L.M. Faller, H. Liechtenecker, C. Ungermanns

Digital Literacy as a Part of Continuing Education Library Efforts 801
R. Vrana

The Model of Collaborative Terminology Contribution to the Art & Architecture Thesaurus: Application and Improvement of Crowdsourcing Methods in an Educational Context 807
G. Zlodi, T. Ivanjko, P. Štefičar, M. Marochini

ICT for Innovative Education and Science: Smart Environment for Networked Strategies 812
V. Omelyanenko, O. Kudrina, H. Shevtsova, O. Prokopenko, V. Petrenko

Student Social Media Usage and Its Relation to Free-recall Memory Tasks 816
V. Vidaček Hainš, M. Kućar, R. Kovačić

Modernized Courses in Automotive Software Engineering 822
I. Kaštelan, B. Pavković, M. Vranješ, M. Popović

Intuitive and Rational Information Management 826
I. Mikšić, T.Babić, Lj. Bakić-Tomić

Artificial Intelligence – a New Topic in Computer Science Curriculum at Primary and Secondary Schools: Challenges, Opportunities, Tools and Approaches 832
Z. Tkáčová, Ľ. Šnajder, J. Guniš

Inquiry-Based Python Programming at Secondary Schools 835
J. Guniš, Ľ. Šnajder, Z. Tkáčová, V. Gunišová

Sentiment Analysis of Open-Ended Student Feedback 840
T. Hynninen, A. Knutas, M. Hujala

Comparative Analysis of Students' Attitudes on Teaching Quality and its Assessment in Higher Education 845
K. Pavlina, A. Pongrac Pavlina, V. Juričić

Students' Attitudes toward Value-Driven Digital Marketing 849
A.M. Jadanec, T. Babić

The Web-based Lectures as Leverage for Developing the Sense of Belonging in the All-Russian Creative School-Contests 855
O.S. Fomichova, V.A. Fomichov

STEAM Students and Their Expectations from Future Business Life: a Values-driven Workplace 861
Y. Borysiuk, T. Babić

Cryptocurrency as the Currency of the Future: a Case Study among Algebra University College Students 867
A. Knežević, T. Babić, Z. Musa

New Teaching Methods in Higher Education - Management of Information Systems Course 873
K. Aleksić-Maslać, P. Vranešić, B. Debić

Non-Determinism in Nowadays Computing and IT Education 879
J. Henno, H. Jaakkola, J. Mäkelä

An Experiential Learning Approach to Research Methods in Computer Science Based on SMART Goals 887
N. Caporusso

Adaptive Drum Kit Learning System: Impact on Students' Motivation 893
M. Konecki

Impact of Distance Learning on Motivation and Success Rate of Students during the COVID-19 Pandemic 898
M. Konecki

Analysis and Development of a Robotic Disk Mounting Stand 903
F. Koefler, M. Edlinger, N. Jausz, W. Egger, W. Werth, L.M. Faller

SJSU GO 908
R. Fatoohi, S. Ehrman, L. Francesca, C. Chong, J. Khoury, A. Minaise, D. Pham, M. Abhyankar, J. Chen, R. Chen, A. Corona-Sanchez

Online and In-Class Computer Science Teacher Training - Oracle Academy Program Experiences 913
F. Urem, D. Jureković, E. Ban

A Novel System for Automatic, Configurable and Partial Assessment of Student SQL Queries 917
M. Fabijanić, G. Đambić, B. Fulanović

Experience with E-learning in Teaching Combinatorics and Data Processing 923
M. Pokorný

Assessment of the Role of Technology in the Process of Tolerance 928
E. Gürbüzer, G. Dağlı, F. Altinay, Z. Altinay

Distance Learning: Examples of Good Practice, Analysis and Experience 933
M. Mačinko, A. Sović Kržić, I. Mudri

Understanding the Factors that Influence Secondary Teachers' Intention to Use e-Learning Technologies for Teaching after the COVID-19 Pandemic 938
S. Babić, S. Križan Sučić, G. Sinković

Distance Learning Caused by the COVID-19 Pandemic in Croatia: What do Newspaper Portals Actually Deliver to Readers? 944
M. Kuzelj, K. Šamija

The Significance of Transformational Team Members 950
T. Babić, Z. Musa

Digital Transformation as a Process of Using Digital Technologies for Monitoring and Designing the User Experience 956
I. Radoš, T. Babić

The Immigrant Integration Online Training Program in Finland 962
A. Hartikainen, M. Ahola, M. Apiola, E. Sutinen

Assessing the Impact of Mobile Educational Games on Student's Success within Mathematics Subject in Primary Schools 968
M. Fetaji, E. Kajtazi, B. Fetaji, H. Snopce, M. Apostolova

The Teacher's Role in Discovery, Preparation, and Development of Gifted Students in the Field of Informatics 972
G. Atanasova, P. Hristova

Word Cloud Analytics of the Computer Science Research Publications' Titles over the Past Half Century 977
L. Abazi-Bexheti, A. Kadriu, M. Apostolova

Using WebIDE as a Distance Learning Tool for High School Programming 983
M. Mesihović, V. Ljubović, I. Muharemović

Ability of the Information Science Teachers to Teach Programing in the Lower Grades of Primary School 989
J. Žufić, A. Žufić

Perception of Using VAR Technology in Football after Completion of Training and Education and Experiences of Croatian Video Assistant Referees (VARs) and Assistant VARs (AVARs) 995
I. Lucić, S. Babić, D. Vučkov

Implementacija sustava za automatsko praćenje prisutnosti na Visokom učilištu Algebra 1002
B. Fulanović, A. Lacković, D. Bele

Nauči me igrati - igrifikacijom do boljih rezultata 1006
T. Ređep, T. Pavičić Zajec, B. Marčinković

Vrednovanje digitalnih kompetencija putem CRISS projekta 1012
M. Mirković

Implementacija različitih robotskih modela u nastavi tehničke kulture i informatike 1017
P. Dobrić, M. Mačinko, D. Kager

Suvremen pristup učenju putem projektne nastave 1023
N. Boj, A. Tonković, E. Glavaš

Raspberry PI mini računalo kroz nastavnu cjelinu 1028
D. Vrbanc

Sustav upravljanja zadacima iz područja micro:bit 1033
T. Adamović

Aplikacije za fotografsko prepoznavanje i rješavanje matematičkih zadataka 1038
H. Kovač, I. Nađ

Micro:bit-na biometrija 1042
B. Raičković, B. Musović

Glazba2Go, mobilna aplikacija za promoviranje glazbene kulture 1045
I. Britvić, F. Meštrović

Mobilne aplikacije u obrazovnom okruženju 1049
J. Pažanin, M. Rosić

Intermedijalnost autorskih slikovnica Ivane Guljašević Kuman 1056
M. Verdonik, V. Vitas

Predikcija uspjeha u studiranju primjenom Bayesovih mreža 1060
D. Šimić, J. Gusić

Modifikacija arkadnih igara Space Invaders i Super Mario u edukativne inačice za učenje matematike i hrvatskog jezika 1064
I. Franković, M. Ivašić-Kos

Izazovi poučavanja o računalnim mrežama 1070
K. Blažeka

Rudarenje edukacijskih podataka: korištenje klasteriranja za predikciju studentskog uspjeha 1075
K. Kačapor, Z. Lagumdžija

Robotika u funkciji razvoja vještina 21. stoljeća 1081
I. Ružić

Informacijska i komunikacijska tehnologija (ICT) – ključna karika između različitih oblika obrazovanja 1084
M. Božurić, R. Bogut, M. Tretinjak

Laboratorijski 3-osni model realiziran pomoću servo sustava 1090
M. Lučan, I. Vlašić, G. Malčić

Laboratorijska maketa redundantnog sustava upravljanja 1095
A. Omrčen, M. Lučan, G. Malčić

Samoregulacija učenika u virtualnoj učionici nastave matematike 1100
R. Soldo, J. Domac, I. Olujić

Primjena alata Xmpl za brzo pokretanje naredbi u podučavanju Linuxa 1105
I. Krpan, I. Capan, M. Davidović

Održivi razvoj ruralnog područja – Croduino setovi i dronovi 1109
J. Domac

Stavovi studenata o računalnom načinu razmišljanja i vještine rješavanja testa računalnog načina razmišljanja 1112
N. Tatković, P. Radulović, S. Tatković

Percipirana kvaliteta sustava za upravljanje verzijama programskog koda od strane studenata informatike 1118
T. Orehovački, S. Babić, D. Etinger

Usporedba nastave na daljinu i učioničke nastave iz predmeta Informatika i Računalstvo 1124
S. Deljac, M. Berović

COMPUTERS IN TECHNICAL SYSTEMS

PAPERS

ROBOTICS AND AUTOMATION

Application of PandaPower Tool in Evaluating the Potential of Using PV Distributed Generation for Voltage Regulation in Electrical Power Networks 1135
N. Bogunović, S. Vlahinić, D. Franković, V. Komen

Resilience to Cascading Failures: a Complex Network Approach for Analysing the Croatian Power Grid 1141
Z. Sičanica, I. Vujaklija

Analysis of Justification for Using Capacitor Banks in Distribution Network with Low Power Demand 1146
D. Ćetković, S. Vlahinić, D. Franković, V. Komen

AVR and PSS Coordination Strategy by Using Multi-Objective Ant Lion Optimizer 1151
T. Špoljarić, I. Pavić

Integrated Production Optimization at INA d.d. Upstream Supported by Osisoft PI 1157
B. Žeželj, Ž. Frkin, F. Beznea, M. Gojić

Predictive Lane-Keeping System for an Autonomous Vehicle 1161
P. Makarun, Š. Ileš, J. Matuško

Sliding Mode Control of Custom Built Rotary Inverted Pendulum 1166
M. Švec, Š. Ileš, J. Matuško

A Simulator for Training Human Operators of a Remote Controlled Anti-Terrorism Ground Vehicle 1171
J. Fulir, Ž. Mihajlović, M. Seder

INTERNET BASED SYSTEMS, APPLICATIONS AND TECHNOLOGIES

Tools for Analytics and Cognition Framework for a Car-Sharing Use Case 1179
A. Karadimce, D.C. Bogatinoska, M. Sefidanoski, N. Paunkoska Dimoska, N. Marina

Experimenting with Means to Store and Monitor IoT based Measurement Results for Energy Saving 1185
M. Saari, J. Grönman, J. Soini, P. Rantanen, T. Mäkinen

An Open-Source Solution for Mobile Robot based Environmental Sensing 1191
J. Grönman, J. Viljanen, J. Vihervaara, M. Saari

Providing Facilities in Health Care via Brain-Computer Interface and Internet of Things 1196
M. Ullah, A. Hekmatmanesh, D. Savchenko, R. Moioli, P. Nardelli, H. Handroos, H. Wu

The Success Factors of a National Healthcare Ecosystems Maturation: Preliminary Results 1202
M. Trkman, S. Bajrić, R. Malkoč

Interface Digital Twins: Rendering Physical Devices Accessible to People Who are Blind 1207
N. Caporusso, N. Elleman, S.S. Cho

Evaluation of Resource Management System for InfaaS-adaptive Disaster Management Application Platform 1212
Y. Watashiba, Y. Matsui, S. Date

IoT Protocol Selection for Smart Grid Applications: Merging Qualitative and Quantitative Metrics 1218
M. Ullah, S.R. Ullah Kakakhel, T. Westerlund, A. Wolff, D. Carillo, J. Plosila, P. Nardelli

A Survey on Parallel Architectures and Programming Models 1224
B. Pervan, J. Knezović

HPC-Based Parallel Software for Solving Applied Boolean Satisfiability Problems 1231
V.G. Bogdanova, S.A. Gorsky, A.A. Pashinin

SOFTWARE DESIGN AND AUTOMATION IN INDUSTRY

Sample Based Synthesis of Car Engine Noise 1239
D. Miljković

Fault Detection of Aircraft Piston Engine Based on Exhaustive Database Search 1245
D. Miljković

PROFESSIONAL PAPER IN CROATIAN

Automatizacija pročistača otpadnih voda 1253
M. Ilijanić, M. Lučan, G. Malčić

INTELLIGENT SYSTEMS

PAPERS

Automation of a Decision Tree Conversion into a Fuzzy Inference System Using ANTLR 1263
S.S. Sosinskaya, R.S. Dorofeev, A.S. Dorofeev, T.R. Usenko

A Randomized Load Balancing Criteria Using Traffic Flow in SDN 1267
P. Choudhary, P. Thota

Data Cleaning Techniques in Detecting Tendencies in Software Engineering 1272
P.V. Georgieva, E. Nikolova, D. Orozova

Automatic Machine Translation of Poetry and a Low-Resource Language Pair 1278
I. Dunđer, S. Seljan, M. Pavlovski

Human Quality Evaluation of Machine-Translated Poetry 1284
S. Seljan, I. Dunđer, M. Pavlovski

Using Decision Diagrams of Special Kind for Compactification of Conflict Data Bases Generated by CDCL SAT Solvers 1290
V. Kondratiev, I. Otpuschennikov, A. Semenov

Detecting Objects in Drone Imagery: a Brief Overview of Recent Progress 1296
S. Sambolek, M. Ivašić-Kos

Individualization of Anonymous Identities Using Artificial Intelligence (AI) 1302
T. Bronzin, B. Prole, A. Stipić, K. Pap

Towards the Open Ontology for IoT Ecosystem's Security 1308
I. Tomičić, P. Grd

Decision Tree Algorithm for Control of Compressor Multiset in Refrigeration Industry 1314
I. Šulekić, D. Milinković, T. Špoljarić

GIS Analysis of Basketball Courts and Healthy Stores Relationship for Young Population in the City of Skopje 1319
N. Stojanova, R. Vignjevikj, A. Naumoski

Decentralized Trustless Gossip Training of Deep Neural Networks 1324
R. Šajina, N. Tanković, D. Etinger

Discrete Bat Algorithm for Event Planning Optimization 1329
S. Delalić, A. Alihodžić, M. Tuba, E. Selmanović, D. Hasić

A Conceptual Network Analysis of Gamification Practices in Primary and Secondary Education 1335
I. Tomičić, M. Schatten

Adaptive Rolling Window Selection for Minimum Variance Portfolio Estimation Based on Reinforcement Learning 1342
B. Gašperov, F. Šarić, S. Begušić, Z. Kostanjčar

Synthetic Astronomical Image Sequence Generation 1347
A. Gribl, D. Petrinović

Novel Class Detection in Non-stationary Streaming Environment with a Discriminative Classifier 1353
R. Šajina, N. Tanković, D. Etinger

Evaluation of Diatoms Biodiversity Models by Applying Different Discretization on the Class Attribute 1358
A. Naumoski, G. Mirceva, K. Mitreski

Classifying Protein Structures by Using Protein Ray Based Descriptor, KNN and FuzzyKNN Classification Methods 1364
G. Mirceva, A. Naumoski, A. Kulakov

An Analysis of Early Use of Deep Learning Terms in Natural Language Processing 1369
B. Dalbelo Bašić, M.P. di Buono

On Automated Workflow for Fine-Tuning Deepneural Network Models for Table Detection In Document Images 1374
I. Cherepanov, A. Mikhailov, A. Shigarov, V. Paramonov

StimSeqOnt: An Ontology for Formal Description of Multimedia Stimuli Sequences 1378
M. Horvat

Machine Learning Model for Detecting High School Students as Candidates for Drop-Out from a Study Program 1384
Đ. Pašić, D. Kučak

Evaluation of Structural Hyperparameters for Text Classification with LSTM Networks 1389
M. Frković, N. Čerkez, B. Vrdoljak, S. Skansi

On Checking Controllability of Specification Languages for DES 1395
A. Davydov, A. Larionov, N. Nagul

A Note on Geometric Calibration of Multiple Cameras and Projectors 1401
T. Petković, S. Gasparini, T. Pribanić

ROBOTICS TECHNOLOGIES AND APPLICATIONS

PAPERS

Estimating Robot Manipulator End-effector Forces Using Deep Learning 1411
S. Kružić, J. Musić, R. Kamnik, V. Papić

Prediction of the Behavior of a Pneumatic Soft Robot Based on Koopman Operator Theory 1417
E. Kamenar, N. Črnjarić-Žic, D. Haggerty, S. Zelenika, E.W. Hawkes, I. Mezić

Conceptual Design of an Autonomous Rover with Ground Penetrating Radar: Application in Characterizing Soils Using Deep Learning 1422
P. Linna, T. Aaltonen, A. Halla, J. Grönman, N. Narra

Increasing the Accuracy of Robotic Neurosurgical Procedures through Robot Calibration 1428
L. Drobilo, M. Švaco, B. Jerbić

Implementation of a Low-Cost Autonomous Underwater Vehicle Using Open Source ROS Components with Consumer Class Sonar Technologies 1437
T. Aaltonen, M. Saarivirta, T. Kerminen, J. Grönman

Improving Markerless Registration Accuracy by Mapping Facial Deformation 1443
A. Žgaljić, M. Švaco, B. Jerbić

Programiranje robota u učionici budućnosti 1448
D. Vrbanec

INFORMATION SYSTEMS SECURITY

PAPERS

Genetic Algorithm and Artificial Neural Network for Network Forensic Analytics 1457
D. Oreški, D. Andročec

Using FireEye Endpoint Security for Educational Purposes 1463
M. Dujmić, D. Delija, G. Sirovatka, M. Žagar

Constructing a Set of Weak Values for Full-Round MD4 Hash Function 1469
I. Gribanova, A. Semenov

Bug Detection in Embedded Environments by Fuzzing and Symbolic Execution 1475
J. Vijtiuk, L. Perkov, A. Krog

Forensic Analysis of Windows 10 Sandbox 1481
A. Đuranec, S. Gruičić, M. Žagar

Linux Forensic Triage: Overview of Process and Tools 1487
A. Anđelković, K. Hausknecht, G. Sirovatka

Overview of Mac System Security and its Impact on Digital Forensics Process 1493
D. Sladović, D. Topolčić, D. Delija

Red Teams - Pentesters, APTs, or Neither 1499
I. Kovačević, S. Groš

File Fragment Classification with Focus on OLE and OOXML Classes 1507
K. Skračić, F. Rukavina, K. Miličić, J. Petrović, P. Pale

AI Safety: State of the Field through Quantitative Lens 1511
M. Jurić, A. Sandić, M. Brčić

BUSINESS INTELLIGENCE SYSTEMS

INVITED PAPER

The Linked Data Enterprise as Enabler for Both Intra – and Inter-organizational Business Data Integration and Usage 1521
A M. Tjoa

PAPERS

Time Series Model for Sales Predictions in the Wholesale Industry 1527
T. Hlupić, D. Oreščanin, A.M. Petrić

Business Intelligence Approach to Support Decision Making in Publishing Sector 1532
D. Borissova, N. Keremedchieva, D. Keremedchiev

ERP Solutions in Cloud Technologies as a Driver for Digital Transformation of Businesses 1538
S. Križanić, T. Šestanj-Perić, A. Kutnjak

Towards an Agile Framework for Business Intelligence Projects 1544
M. Prouza, S. Brodinová, A M. Tjoa

A Different Approach for Clique and Household Analysis in Synthetic Telecom Data Using Propositional Logic 1550
S. Skansi, K. Šekrst, M. Kardum

Sentiment Analysis of Tweets about COVID-19 Disease during Pandemic 1554
G. Matošević, V. Bevanda

Role of Business Intelligence Systems in Croatian Higher Education Quality Assurance 1560
M. Cvitanušić Brečić

Study of Customer Behavior in Online B2B Shopping 1565
E. Exenberger, J. Bucko

Fruit Firmness Prediction Using Multiple Linear Regression 1570
T. Ivanovski, G. Zhang, T. Jemrić, M. Gulić, M. Matetić

Applying the Decision Tree Method in Identifying Key Indicators of the Digital Economy and Society Index (DESI) 1576
A. Kutnjak, L. Hrustek, S. Križanić

Determining the Location of Postal Centers in B&H Using Machine Learning Clustering Method and GIS 1582
A. Kosovac, E. Muharemović, M. Begović, E. Šimić

Defining ERP System Selection Methodology – Research Carried Out in Small and Medium-Sized Production Companies in Croatia 1587
A. Mahmutović, M. Nikitović

Implementacija ERP sustava iz perspektive konzultanata 1593
I. Kostanjevec, R. Picek

DIGITAL ECONOMY AND DIGITAL SOCIETY

PAPERS

E-Democracy Tools Adoption: Experience of Austria, Croatia, Italy, and Slovenia 1603
V. Roblek, I. Strugar, M. Meško, M. Pejić Bach, B. Jaković

The Importance of Digitizing the Management Processes of Sports Associations in the City of Zadar 1610
J. Miočić

Youth Perception on 5G Networks in Osijek 1615
D.Turkalj, I. Kelić, J. Rašić

Digitalization Impacts on Innovation Networks: Policy and Estimations Issues 1621
O. Kudrina, V. Omelyanenko, H. Shevtsova, V. Samoday, Yu. Mashyna, V. Bilyk

A Literature Review of Digital Transformation in Healthcare 1625
L. Ivančić, Lj. Milanović Glavan, V. Bosilj Vukšić

Creating of Digital Life in Art Museums 1630
D. Ilišević, N. Banović-Ćurguz, S. Vujković

SWOT Analysis of Selected Digital Technologies in Transport Economics 1635
M. Jović, E. Tijan, D. Žgaljić, P. Karanikić

Shared Services Business Model in ICT Environment 1641
T. Žilić

Optimization of Cargo Container Loading on Railway Wagons 1647
S. Aksentijević, E. Tijan, M. Jović, N. Munitić

Applying Adaptive Neuro-Fuzzy Inference System (ANFIS) while Analysing Interdependecies of Tax Burden and Capital Structure of Croatian Hotel Companies 1653
S. Brlečić Valčić, A. Samodol, M. Valčić

Synergy of Innovation Procedures and Communication Skills as a Success Predictor in IT Supported Management 1660
J. Dvorski, A. Bernik, D. Radošević

Cloud-based Services Approach as Accelerator in Empowering Digital Transformation 1664
H. Mydyti, J. Ajdari, X. Zenuni

Performance Analysis of Aruba Wireless Local Network in Croatian Pension Insurance Institute 1671
A. Skendžić, B. Kovačić, L. Ljubičić

A Study of Coordination Challenges in Digital Policy Implementation and Evaluation in Finland 1676
O.C. Osifo

Port Community System Feasibility Analysis – Case Study Split 1684
I. Torlak, E. Tijan, S. Aksentijević, A. Jugović

Digital Business Models in the Logistics Services 1690
A. Agatić, T. Poletan Jugović, E. Tijan, A. Jugović

Using Fuzzy Logic in Analysing and Modelling the Reflection of Monetary and Fiscal Conditions on GDP Per Capita in Croatia 1696
A. Samodol, S. Brlečić Valčić, A. Ostojić

The Role of Perception in the Adoption of Digital Platforms in Agriculture 1703
K. Tomičić-Pupek, I. Pihir, M. Tomičić Furjan

Social Entrepreneurial Intention: Does the Classroom Matter? 1709
I. Kedmenec

Digital Transformation of Monitoring Customer Behaviour in the Cars Sales 1715
I. Radoš, M. Hajnić, I. Radoš

"Financial" Aspects of Spotify Streaming Model 1720
J. Lozić, G. Vojković, M. Milković

Value of Innovation Platforms in Agriculture 1725
M. Tomičić Furjan, L. Hrustek, I. Pihir

Decision Making on Digital Platforms in Agriculture 1731
N. Kadoić, K. Tomičić-Pupek, N. Vrček

Inovacije i tehnološki napredak u poslovanju morskih luka i njihov utjecaj na gospodarstvo 1737
G. Mudronja

Digitalna transformacija pomorskog transporta kao dijela Plave ekonomije 1743
M. Jović, A. Agatić, A. Jugović

Specifičnosti Uberovog modela prijevoza s primjenom u Republici Hrvatskoj 1749
K. Jugović, A. Jugović, S. Hess

INFORMATION AND COMMUNICATION TECHNOLOGY LAW

PAPERS

Application of the General Data Protection Regulation in Schools: A Qualitative Study with Teachers, Professional Associates and Principals 1759
L. Vejmelka, T. Katulić, M. Jurić, M. Lakatoš

IoT Devices and the Need to Inform Utility Users of Collecting, Controlling and Processing of Personal Data 1766
G. Vojković, M. Milenković

Use of Security Settings on Social Networks of Elementary and High School Students in the Split-Dalmatia County 1772
R. Matković, L. Vejmelka, Ž. Ključević

Visual Analysis of Similarity and Relationships between Legal Texts 1778
J. Opiła, T. Pełech-Pilichowski

Privacy Policy Understandability Analysis of Croatian Electronic Publications 1784
M. Alić

ENGINEERING EDUCATION

PAPERS

Closing the Gender Gap in Engineering: Students Role Model Program 1793
E. Vidal, E. Castro, S. Montoya, K. Payihuanca

Integrating Industry Seminars within a Software Engineering Module to Enhance Student Motivation 1797
G.J. Collins

From Framework Programs to Teaching: Integrating Experience from European Research Projects in Teaching Engineering 1803
W. Brenner, N. Adamović

Teamwork Challenges and Solution Strategies of First-Semester Engineering Students 1809
A. Gorup, M. Grzunov, J. Petrović, P. Pale

Differences in the Students' Achievements between Traditional and Project-Based Learning of Basic Engineering Competencies: A Quasi-experimental Study 1814
D. Purković, M. Prihoda

Computer-Based Question and Exam Evaluation in Summative Knowledge Assessment 1820
S. Tomić, V. Paunović, I. Bosnić

Students' Perception of Summative Peer Review Grading 1826
L. Zrnić, L. Korov, J. Petrović, P. Pale

Assessing Students' SQL Knowledge and Skills in Gamification Manner 1831
G. Tuparov, D. Keremedchiev

Changing the Assessment Process in Mathematics for Students in Engineering 1837
P.V. Georgieva, E. Nikolova

Design of Rubrics for Student Outcomes in 2019-2020 ABET Criteria 1843
B. Pejčinović

Electronic Learning Experience Setup: Power Electronics and Electrical Drive Education 1849
P.J. van Duijsen, D.C. Zuidervliet, J.B. Woudstra

Study of Switching Forward Single-ended DC/DC Converter in the Course "Power Supplies" 1855
S.L. Zaharieva, I.I. Stoev, A.N. Borodzhieva, S. Stoyanov

Designing an Interactive Multimedia Application for the Course "Communication Circuits" 1861
A.N. Borodzhieva

Integrated Laboratory Complex 1867
R. Simionov, S. Mollova, R. Dolchinkov

Elastic Collisions Visualization Using OpenCV Object Motion Tracking 1873
M. Hajba, E. Ciriković, M. Pecimotika

Educational Computer Games and Gamification at the Higher Education – Students' Points of View 1879
D. Tuparova, G. Tuparov, D. Orozova

Using Moodle e-Learning Platform in Mechanical Engineering Lectures 1885
P. Tomšič, I. Demšar, T. Finkšt

Creation of Software Platform for Distance Use of Lab Equipment and Data in Virumaa College at Tallinn University of Technology 1891
O. Shvets, K. Murtazin, G. Piho

FPGA Design of Boolean Functions Using a Cascade of Decoders and Logic Gates 1896
A.N. Borodzhieva, I.I. Stoev, I.D. Tsvetkova, S.L. Zaharieva, V.A. Mutkov

Interactive Approach to Digital Logic 1901
M. Špoljarić, M. Hajba, M. Pecimotika

Analysis of Secure Data Deletion and Recovery with Common Digital Forensic Tools and Procedures 1907
S. Žulj, D. Delija, G. Sirovatka

MS Excel-Based Application for Implementing the Cryptographic Algorithm Shamir's Secret Sharing 1911
A.N. Borodzhieva

SOFTWARE AND SYSTEMS ENGINEERING

PAPERS

SOFTWARE SYSTEM ARCHITECTURES

Database Integration Systems 1923
J. Dončević, K. Fertalj

Microservice Performance Degradation Correlation 1929
M. Samardžić, R. Šajina, N. Tanković, T. Galinac Grbac

Robust and Scalable Online Code Execution System 1933
H.Z. Došilović, I. Mekterović

Adapting CERIF for a National CRIS: A Case Study 1939
D. Kremenjaš, P. Udovičić, O. Orel

Knowledge-Based System for Data Modelling Based on Verbalisation – an Architecture Proposal 1945
S. Šuman, A. Jakupović, M. Kaluža

QUALITY ASPECTS

Investigation of the Accessibility of Non-Text Content Published on Websites 1953
K. Kous, S. Kuhar, A. Rajšp, B. Šumak

RESCCUE RAF App – Using Technology to Mitigate Climate Change Urban Impacts 1959
P. Lopes, A. Oliveira, C. Pereira, R.S. Brito, M.A. Cardoso, R. Martins, M. David, J. Gomes, J. Pina

The Advantage of Using SWOT Analysis for Companies with Implemented ITIL Framework Processes 1964
A. Granulo, A. Tanović

TECHNOLOGIES

Smart Contracts as a Diploma Anti-Forgery System in Higher Education - a Pilot Project 1973
D. Čeke, S. Kunosić

A Review on Generating Random Numbers in Decentralised Environments 1979
S.D. Simić, R. Šajina, N. Tanković, D. Etinger

Accessibility Standards and Their Implementation in Custom Data-Driven Maps 1985
I. Serna-Marjanović, A. Tanović, A. Ćerimagić

Improvements of Computer Assisted Virtual Environment (CAVE) 1991
M. Fandáková, K. Zábovská, B. Bučko, M. Zábovský

APPLICATIONS

The Order Batching Concept Implemented In Real Smart Warehouse 1999
S. Delalić, E. Žunić, A. Alihodžić, E. Selmanović

Integration of Photogrammetry within Laser Scanning Approach 2005
P. Kudela, M. Palčák, K. Zábovská, B. Bučko

Some Elements for Assessing the Radiated Heat in Urban Areas 2009
I. Kožar, K. Peša, M. Cuculić, N. Torić Malić

PROFESSIONAL PAPER IN CROATIAN

Unity – 3D i virtualna stvarnost 2015
B. Fulanović, V. Šepl

MIPRO JUNIOR - STUDENT PAPERS

PAPERS

Development of Portable System for Determination of Aircraft Motion 2023
S. Marijan, D. Franjković

Small Piston Engine Aircraft Vibration Measurement and Analysis 2029
F. Juretić, D. Gerhardinger, A. Domitrović, J. Ivošević

Mazzilli Oscillator as a Tool in Education 2035
L. Matić, S. Stojanović, V. Šimović

An Interactive Punch Power Tracker for Heavy Bag Training 2040
A. Vinković, I. Linardić, D. Meštrović, J. Petrović, P. Pale

Simulink Model of Oxygen Distribution in Skeletal Muscle 2044
A. Džuho, A. Aleta, S. Pandža, I. Ramić, N. Mamatnazarov, L. Spahić

Multi-Model Databases - Introducing Polyglot Persistence in the Big Data World 2048
I. Košmerl, K. Rabuzin, M. Šestak

Hot Topic Detection Using Twitter Streaming Data 2054
T. Jagić, Lj. Brkić

Processing and Visualization of Collected Data Based on Open-Source Tools and Principles 2060
S. Grbac Babić, K. Cetina

Web Application Dashboards as a Tool for Data Visualization and Enrichment 2064
M. Holjevac, T. Jakopec

Razvoj sustava za upravljanje rasporedom sati 2070
A. Šturlan, K. Vučković, T. Orehovački

Enhancing Performance of Cloud-based Software Applications with GraalVM and Quarkus 2076
M. Šipek, D. Muharemagić, B. Mihaljević, A. Radovan

Achieving Efficient Structured Concurrency through Lightweight Fibers in Java Virtual Machine 2082
P. Pufek, D. Beronić, B. Mihaljević, A. Radovan

Use of Keystroke Dynamics and a Keystroke-Face Fusion System in the Real World 2088
J. Stavanja, P. Peer, Ž. Emeršič

Digital Forensics Appliance in Corporate Ecosystem Considering Limitations in the EU Legal Framework 2094
V. Rajič, M. Milenković, G. Vojković

Using Convolutional Neural Network for Chest X-ray Image Classification 2101
M. Sorić, D. Pongrac, I. Inza

SMART, SUSTAINABLE AND RESILIENT CITIES AND INFRASTRUCTURE

PAPERS

Creating Sustainable Solutions for Photovoltaics 2111
W. Brenner, N. Adamović

Use of Drone to Improve Healthcare Efficiency and Sustainability 2117
L. Faramondi, G. Oliva, L. Ardito, A. Crescenzi, M. Caricato, M. Tesei, A. Onetti Muda, R. Setola

A Privacy-Oriented Solution for the Improvement of Workers Safety 2123
L. Faramondi, P. Bragatto, C. Fiorevanti, M.G. Gnoni, S. Guarino, R. Setola

LoRa-SDN: Providing Wireless IoT Edge Network Functions via SDN 2129
F. Holik, U. Roedig, N. Race

Analyses of Ecological and Energy Footprint as Indicators of Energy Management in the Transition to Sustainability Using Social Networks 2135
B. Mihajlovski, B. Fetaji, L. Abazi-Bexheti, M. Fetaji

Review of Discrete Simulation Modelling Use in the Context of Smart Cities 2141
M. Jadrić, M. Ćukušić, D. Pavlić

Traffic State Estimation Using Speed Profiles and Convolutional Neural Networks 2147
L. Tišljarić, T. Carić, T. Erdelić, M. Erdelić

Stakeholder Support as Critical Success Factor in Adopting Big Data Technologies for Smart Cities 2153
J. Pivar, N. Vlahović

Enhancing Occupants Comfort and Well-being through a Smart Office Setup 2159
A. Barišić, V. Amaral, M. Challenger

Innovative Predictive Model for Smart City Security Risk Assessment 2165
L. Franchina, A. Socal

Digitalization and Smart Islands in the Kvarner Archipelago 2171
M. Mimica, G. Krajačić, D. Medved, D. Jardas

Autonomous Mobility and User Perception: A Case of City as a Lab in Slovenia 2177
I. Zajc, R. Sernec, G. Lenart, A. Pucihar

The European Infrastructure Simulation and Analysis Centre (EISAC) Initiative and Its Technological Assets 2182
V. Rosato, A. Tofani, A. Di Pietro, M. Pollino, S. Giovinazzi, L. Lavalle, G. D`Agostino

Comparative Analysis of the Selected Practices in the Field of Urban Logistics of the Polish Cities 2186
M. Zysińska

Unleashing the Power of Urban Living Labs to Make our Cities Humanly Smart 2192
F. Molinari, B. Kovačić

The Potential of Mobile Energy Storage in Microgrids 2198
H.H. Abdeltawab, Y.A.I. Mohamed

OPTOELECTRONICS AND PHOTONICS

PAPERS

Analysis of a-Si:H p-i-n Photodiode Detection of HeLa Cells Luminescence 2209
V. Gradišnik

Plasmonic Enhanced Photodetectors for Near Infra-red Light Detection 2214
D. Giubertoni, G. Paternoster, F. Acerbi, X. Borrisé, A. Cian, A. Filippi, A. Gola, A. Guerrero, F. Perez Murano, F. Romanato, E. Scattolo, P. Bellutti

Variable Angle Spectroscopic Ellipsometry Study of Poly(3,4-ethylenedioxythiophene):Polystyrene Sulfonate Thin Films in Contact with Air 2219
G. Pathak, D. Čakara

Radiation Pressure Sensor 2228
M. Karuza, D. Božičević, G. Cantatore, M. Vretenar

Measurement of the Human Cadaver Ossicle Vibration Amplitude by Fiber-Optic Interferometry 2232
Z. Djinovic, M. Tomic, R. Pavelka, G. Sprinzl, H. Traxler

Marine Fiber Optic and Spinning Mass Gyrocompasses 2237
A. Škrobonja, I. Jurdana , I. Panić, N. Wakabayashi

DEW COMPUTING

PAPERS

Microservice Approach to the Qualitative Study of Attractors of Binary Dynamic Systems Based on the Boolean Constraint Method 2247
G.A. Oparin, V.G. Bogdanova, A.A. Pashinin

Impact of Dew Computing on Cyber-Physical Systems and IoT 2253
M. Gusev

The Rainbow through the Lens of Dew 2259
Z. Šojat, K. Skala

A Disaster-Resilient Messaging Protocol Based on Dew Computing 2265
Y. Wang

Ensuring Resource Availability with MRU/FRU Caching: A Dew-Blockcloud Model 2270
C. Chukwuocha, R.K. Thulasiram, P. Thulasiraman, Y. Wang

The Relevance of Blockchain with Dew Computing: a Review 2277
R.N.A. Sosu, C.N. Babu, S.A. Frimpong, J. Essuman

LIST OF PAPER REVIEWERS

Agatić, A.	(Croatia)
Airaksinen, M.	(Finland)
Aksentijević, S.	(Croatia)
Aleksi, I.	(Croatia)
Aleksić, S.	(Germany)
Aleksić-Maslać, K.	(Croatia)
Ašenbrener Katić, M.	(Croatia)
Avbelj, V.	(Slovenia)
Babić, D.	(Croatia)
Babić, J.	(Croatia)
Babić, S.	(Croatia)
Babić, T.	(Croatia)
Bačmaga, J.	(Croatia)
Bagić Babac, M.	(Croatia)
Bajgorić, N.	(Bosnia and Herzegovina)
Balković, M.	(Croatia)
Banek, M.	(Croatia)
Baranović, M.	(Croatia)
Barić, A.	(Croatia)
Bernik, A.	(Croatia)
Betti, T.	(Croatia)
Biondić, I.	(Croatia)
Blašković, B.	(Croatia)
Blažević, Z.	(Croatia)
Boban, M.	(Croatia)
Brčić, M.	(Croatia)
Brčić, M.	(Croatia)
Brezany, P.	(Austria)
Brkić, Lj.	(Croatia)
Budin, A.	(Croatia)
Budin, L.	(Croatia)
Burget, R.	(Czech Republic)
Caporusso, N.	(United States)
Caputo, D.	(Italy)
Car, Ž.	(Croatia)
Cetinić, E.	(Croatia)
Cifrek, M.	(Croatia)
Crneković, M.	(Croatia)
Čakara, D.	(Croatia)
Čandrlić, S.	(Croatia)
Čavrak, I.	(Croatia)
Čeperić, V.	(Croatia)
Čičin-Šain, M.	(Croatia)
Čičin-Šain, M.	(Croatia)
Čubrilo, M.	(Croatia)
Čunko, K.	(Croatia)
Čupić, M.	(Croatia)
Davidović, D.	(Croatia)
Davidović, M.	(Croatia)
Delač, G.	(Croatia)
Delić, V.	(Serbia)
Deljkić, Ž.	(Croatia)
Depolli, M.	(Slovenia)
Derek, A.	(Croatia)
Dešić, S.	(Croatia)
di Buono, M. P.	(Italy)
Đambić, G.	(Croatia)
Etinger, D.	(Croatia)
Fertalj, K.	(Croatia)
Frančić, V.	(Croatia)
Fulir, J.	(Croatia)
Galinac Grbac, T.	(Croatia)
Gamberger, D.	(Croatia)
Glavaš, J.	(Croatia)
Gligora Marković, M.	(Croatia)
Gold, H.	(Croatia)
Golfarelli, M.	(Italy)
Golub, M.	(Croatia)
Golubić, S.	(Croatia)
Gradišnik, V.	(Croatia)
Grazio, S.	(Croatia)
Grbac, N.	(Croatia)
Grd, P.	(Croatia)
Grgić, K.	(Croatia)
Grgurić, A.	(Croatia)
Grgurina, I.	(Croatia)
Grönman, J.	(Finland)
Groš, S.	(Croatia)
Grzinić, T.	(Croatia)
Henno, J.	(Estonia)
Hofman, D.	(Croatia)
Hoić-Božić, N.	(Croatia)
Holenko Dlab, M.	(Croatia)
Horak, H.	(Croatia)
Horvat, M.	(Croatia)
Hudek, V.	(Croatia)
Huljenić, D.	(Croatia)
Humski, L.	(Croatia)
Hyrynsalmi, S.	(Finland)
Ipšić, I.	(Croatia)
Ivašić-Kos, M.	(Croatia)
Jaakkola, H.	(Finland)
Jagušt, T.	(Croatia)
Jakopović, Ž.	(Croatia)
Jakšić, D.	(Croatia)
Jardas, M.	(Croatia)
Jazbinšek, V.	(Slovenia)
Jelenković, L.	(Croatia)
Jerbić, B.	(Croatia)
Jevtić, D.	(Croatia)
Jezić, Z.	(Croatia)
Job, J.	(Croatia)
Jović, M.	(Croatia)
Jugović, A.	(Croatia)

Jukić, O. (Croatia)
Jurdana, I. (Croatia)
Jurić, M. (Croatia)
Juričić, V. (Croatia)
Jurković, I. (Croatia)
Kadoić, N. (Croatia)
Kalafatić, Z. (Croatia)
Kalpić, D. (Croatia)
Karadimce, A. (North Macedonia)
Karaničić, P. (Croatia)
Karuza, M. (Croatia)
Kaštelan, I. (Serbia)
Katulić, T. (Croatia)
Klemenc-Ketiš, Z. (Slovenia)
Knežević, T. (Croatia)
Koceva Rashkovska, A. (Slovenia)
Kocijan, K. (Croatia)
Konecki, M. (Croatia)
Konecki, M. (Croatia)
Koričić, M. (Croatia)
Kosec, G. (Slovenia)
Kostanjčar, Z. (Croatia)
Kovačević, A. (Croatia)
Kovačević, I. (Croatia)
Kovačević, T. (Croatia)
Kovačić, A. (Croatia)
Kovačić, B. (Croatia)
Kozak, D. (Croatia)
Krajčar, S. (Croatia)
Krašna, M. (Slovenia)
Krbot Skorić, M. (Croatia)
Krhen, M. (Croatia)
Krivec, S. (Croatia)
Krivić, P. (Croatia)
Krleža, D. (Croatia)
Krois, I. (Croatia)
Krznar, M. (Croatia)
Kušek, M. (Croatia)
Leitgeb, E. (Austria)
Lenac, K. (Croatia)
Lipic, T. (Croatia)
Lisičar, H. (Croatia)
Lovrenčić, R. (Croatia)
Lučić, D. (Croatia)
Ma, Z. (China)
Magdalenić, I. (Croatia)
Malnar, B. (Croatia)
Mance, D. (Croatia)
Mandić, L. (Croatia)
Mandić, T. (Croatia)
Marčetić, D. (Croatia)
Maričić, S. (Croatia)
Marovic, M. (Croatia)
Mastelic, T. (Croatia)
Matković, R. (Croatia)
Mauša, G. (Croatia)
Medić, Z. (Croatia)
Medved Rogina, B. (Croatia)
Mekterović, I. (Croatia)
Mezak, J. (Croatia)
Mihajlović, Ž. (Croatia)
Mikac, B. (Croatia)
Milanović, I. (Serbia)
Milašinović, B. (Croatia)
Miličević, K. (Croatia)
Mlinarić, H. (Croatia)
Molnar, G. (Croatia)
Mošmondor, M. (Croatia)
Možek, M. (Slovenia)
Mudronja, G. (Croatia)
Načinović Prskalo, L. (Croatia)
Nain, N. (India)
Oreški, D. (Croatia)
Osrečki, Ž. (Croatia)
Pale, P. (Croatia)
Palestri, P. (Italy)
Palova, D. (Slovakia)
Papić, V. (Croatia)
Pečar-Ilić, J. (Croatia)
Pejčinović, B. (United States)
Pejić Bach, M. (Croatia)
Perić Hadžić, A. (Croatia)
Pervan, B. (Croatia)
Pesek, I. (Slovenia)
Petcu, D. (Romania)
Petković, T. (Croatia)
Petrović, I. (Croatia)
Petrović, J. (Croatia)
Pintar, D. (Croatia)
Poletan Jugović, T. (Croatia)
Poljak, M. (Croatia)
Popović, S. (Croatia)
Ptiček, M. (Croatia)
Pušnik, M. (Slovenia)
Rabuzin, K. (Croatia)
Radošević, D. (Croatia)
Rak, K. (Croatia)
Rašan, I. (Croatia)
Rashkovska, A. (Slovenia))
Reljić, I. (Croatia)
Repnik, R. (Slovenia)
Ribarić, S. (Croatia)
Rimac-Drlje, S. (Croatia)
Saari, M. (Finland)
Schatten, M. (Croatia)
Skala, K. (Croatia)
Skansi, S. (Croatia)
Skocir, P. (Croatia)
Skočir, Z. (Croatia)
Slavuj, V. (Croatia)
Slovenec, K. (Croatia)
Sović Kržić, A. (Croatia)

Spasov, G. (Bulgaria)
Spremić, M. (Croatia)
Sruk, V. (Croatia)
Stančin, K. (Croatia)
Stanič, U. (Slovenia)
Stanković, I. (Croatia)
Stipčević, M. (Croatia)
Stojković, N. (Croatia)
Strugar, I. (Croatia)
Suligoj, T. (Croatia)
Svigelj, V. (Slovenia)
Šegvić, S. (Croatia)
Šikić, M. (Croatia)
Šimić, D. (Croatia)
Širaiy, B. (Slovenia)
Škopljanac-Mačina, F. (Croatia)
Škvorc, D. (Croatia)
Šorić, K. (Croatia)
Štruc, V. (Slovenia)
Šumak, B. (Slovenia)
Šuman, S. (Croatia)
Šupić, H. (Bosnia and Herzegovina)
Švaco, M. (Croatia)
Švelec, D. (Croatia)
Tafro, A. (Croatia)
Tan, Z.-H. (Denmark)
Tanković, N. (Croatia)
Tijan, E. (Croatia)
Tomasic, I. (Sweden)
Tomić, D. (Croatia)
Tomicic, I. (Croatia)
Trkman, M. (Slovenia)
Trobec, R. (Slovenia)
Trost, A. (Slovenia)
Ullah, M. (Finland)
Uroda, I. (Croatia)
Vidaček-Hainš, V. (Croatia)
Vidović, K. (Croatia)
Vihervaara, J. (Finland)
Vinko, D. (Croatia)
Vlah, D. (Croatia)
Vlahović, N. (Croatia)
Vojković, G. (Croatia)
Volarić, I. (Croatia)
Vrankić, M. (Croatia)
Vrdoljak, B. (Croatia)
Vukić, Đ. (Croatia)
Vuković, M. (Croatia)
Wang, Y. (Canada)
Watashiba, Y. (Japan)
Zakarija, I. (Croatia)
Zichar, M. (Hungary)
Žilak, M. (Croatia)
Žubrinić, K. (Croatia)
Žufić, J. (Croatia)

AUTHOR INDEX

Aaltonen, T. 1422, 1437
Abazi-Bexheti, L. 977, 2135
Abdeltawab, H.H. 2198
Abhyankar, M. 908
Abugabah, A. 383
Acerbi, F. 2214
Ackovska, N. 333
Adamović, N. 1803, 2111
Adamović, T. 1033
Agatić, A. 1690, 1743
Ahola, M. 962
Ajdaraga Krluku, E. 355
Ajdari, J. 501, 527, 1664
Akagić, A. 135
Aksentijević, S. 1647, 1684
Alajbeg, T. 702
Aleksić, S. 427
Aleksić-Maslać, K. 873
Aleta, A. 2044
Alić, M. 1784
Alihodžić, A. 1329, 1999
Altinay, F. 928
Altinay, Z. 928
Amaral, V. 2159
Anđelković, A. 1487
Andročec, D. 1457
Antunes, A. 379
Apiola, M. 962
Apostolova, M. 968, 977
Arbet, D. 84
Ardito, L. 2117
Arslan, E. 202
Ašenbrener Katić, M. 389
Åslund, M. 281
Assunção, P.A. 306
Åstrand, E. 281
Atanasova, G. 972
Attard, C. 293
Auer, M. 106
Avbelj, V. 345, 350
Babič, M. 265
Babić, D. 5
Babić, S. 938, 995, 1118
Babić, T. 711, 726, 826, 849, 861, 867, 950, 956
Babu, C.N. 2277
Bajrić, S. 1202
Bakić-Tomić, Lj. 488, 774, 826
Balaž, D. 779
Ban, E. 913
Banović-Ćurguz, N. 1630
Barić, A. 78, 90, 96, 100, 111, 153
Barišić, A. 2159
Barišić, R. 610, 616
Baručija, E. 135
Batelić, M. 130
Begić, M. 482
Begović, M. 1582
Begušić, D. 571
Begušić, S. 1342
Bele, D. 1002
Bellutti, P. 2214
Beloev, I.H. 553
Bensalem, M. 518
Beqiri, L. 327
Berdalović, I. 29
Berkmann, F. 24
Bernad, P. 641
Bernik, A. 1660
Beronić, D. 2082
Berović, M. 1124
Bertolan, R. 90
Bevanda, V. 1554
Bevčič, M. 661
Bezić, H. 779
Beznea, F. 1157
Bilandžija, D. 74
Bilyk, V. 1621
Biondić, I. 74
Bjelčić, N. 492
Blažeka, K. 1070
Blažeković, M. 465, 492
Bogatinoska, D.C. 512, 1179
Bogdanova, V.G. 1231, 2247
Bogunović, N. 1135
Bogut, R. 1084
Boj, N. 1023
Bondarenko, Yu.A. 693
Bongard, M. 427
Borissova, D. 721, 1532
Borodzhieva, A.N. 790, 1855, 1861, 1896, 1911
Borrisé, X. 2214
Borysiuk, Y. 861
Bosilj Vukšić, V. 1625
Bosnić, I. 1820
Božek, J. 259
Božičević, D. 2228
Božurić, M. 1084
Bragatto, P. 2123
Brčić, M. 1511
Brečić, P. 193
Brenner, W. 1803, 2111
Breznik, K. 754
Brezovec, I. 78
Brito, R.S. 1959
Britvić, I. 1045
Brkić, F. 702
Brkić, Lj. 2054
Brlečić Valčić, S. 1653, 1696

Brložnik, M. 345, 350
Brodinová, S. 1544
Brodnjak, K. 616
Bronzin, T. 1302
Bubanić, M. 588
Bucko, J. 1565
Bučko, B. 1991, 2005
Buçuku, A. 327
Budimir, M. 213
Buljat, B. 779
Bziuk, W. 439
Cantatore, G. 2228
Capan, I. 1105
Caporusso, N. 887, 1207
Car, Ž. 482
Cardoso, M.A. 1959
Carić, T. 268, 2147
Caricato, M. 2117
Carillo, D. 1218
Castro, E. 1793
Cerin, B. 407
Cesar, I. 198
Cetina, K. 2060
Chahin, T.A. 240
Challenger, M. 2159
Chen, J. 908
Chen, R. 908
Cherepanov, I. 1374
Chernysh, A.E. 693
Cho, S.S. 1207
Chong, C. 908
Choudhary, P. 1267
Chukwuocha, C. 2270
Cian, A. 2214
Cifrek, M. 193, 373
Ciriković, E. 447, 1873
Cirimotić, M. 482
Clausen, C. 24
Collins, G.J. 1797
Corona-Sanchez, A. 908
Crescenzi, A. 2117
Cuculić, M. 2009
Cuculić, P. 711, 726
Cvitanušić Brečić, M. 1560
Čakara, D. 2219
Čaušević, A. 20
Čeke, D. 1973
Čerkez, N. 1389
Črnjarić-Žic, N. 1417
Ćerimagić, A. 1985
Ćetković, D. 1146
Ćosić, I. 627
Ćosić, K. 259
Ćoza, A. 170
Ćukušić, M. 2141
D`Agostino, G. 2182
Dağlı, G. 928
Dalbelo Bašić, B. 1369
Date, S. 1212
David, M. 1959
Davidović, D. 249, 395
Davidović, M. 1105
Davydov, A. 1395
Debić, B. 873
Dekanić, K. 253
Delalić, S. 1329, 1999
Delić, V. 409
Delija, D. 1463, 1493, 1907
Deljac, S. 1124
Demirović, S. 433
Demšar, I. 1885
Depolli, M. 235
Dettling, M.M. 24
di Buono, M.P. 1369
Di Pietro, A. 2182
Dizdarević, J. 518
Djinovic, Z. 2232
Dobrić, P. 1017
Dolchinkov, R. 1867
Domac, J. 1100, 1109
Domanjko Petrič, A. 345
Domitrović, A. 2029
Dončević, J. 1923
Dorofeev, A.S. 1263
Dorofeev, R.S. 1263
Došilović, H.Z. 1933
Drobilo, L. 1428
Drushlyak, M.G. 688
Duh, U. 235
Dujmić, M. 1463
Dunđer, I. 1278, 1284
Dvorski, J. 1660
Džuho, A. 2044
Đambić, G. 917
Đuranec, A. 1481
Đurović, G. 672
Edlinger, M. 903
Egger, W. 903
Ehrman, S. 908
Elleman, N. 1207
Emeršič, Ž. 2088
Engelmann, A. 439, 557
Erdelić, M. 2147
Erdelić, T. 268, 2147
Essuman, J. 2277
Etinger, D. 1118, 1324, 1353, 1979
Exenberger, E. 1565
Fabijanić, M. 917
Faller, L.M. 796, 903
Fandáková, M. 1991
Faramondi, L. 2117, 2123
Farkaš, I. 482
Farrugia, R. 293
Fatoohi, R. 908
Fellner, J. 90, 96, 100
Fertalj, K. 1923

Fetaji, B. 327, 968, 2135
Fetaji, M. 968, 2135
Filippi, A. 2214
Findenig, R. 423
Finkšt, T. 1885
Fiorevanti, C. 2123
Flühr, H. 423
Fomichov, V.A. 855
Fomichova, O.S. 855
Fonseca-Pinto, R. 306, 379
Francesca, L. 908
Franchina, L. 2165
Franjković, D. 2023
Franki, V. 453
Franković, D. 1135, 1146
Franković, I. 1064
Friganović, K. 193, 373
Frimpong, S.A. 2277
Frkin, Ž. 1157
Frković, M. 1389
Frumen, S. 754
Fujimoto, D. 153
Fulanović, B. 917, 1002, 2015
Fulir, J. 1171
Funk, H.S. 20, 24
Galinac Grbac, T. 1929
Gashi, B. 544, 563
Gasparini, S. 1401
Gašperov, B. 1342
Gauci, D. 293
Georgieva, I. 459
Georgieva, P.V. 1272, 1837
Gerhardinger, D. 2029
Getman, I. 707
Giovinazzi, S. 2182
Giubertoni, D. 2214
Glavaš, E. 1023
Glunčić, M. 253
Gnoni, M.G. 2123
Gojić, M. 1157
Gola, A. 2214
Gomes, J. 1959
Gorsky, S.A. 1231
Gorup, A. 1809
Gradišnik, V. 2209
Granulo, A. 1964
Grbac Babić, S. 2060
Grd, P. 1308
Gribanova, I. 1469
Gribl, A. 1347
Grishin, N. 316
Grönman, J. 1185, 1191, 1422, 1437
Groš, S. 1499
Gruber, M. 423
Grubišić, D. 5
Gruičić, S. 1481
Grunert, K. 427
Grzunov, M. 1809
Guarino, S. 2123
Guerrero, A. 2214
Guguieva, K. 15, 20
Gulić, M. 1570
Gunčaga, J. 766
Guniš, J. 832, 835
Gunišová, V. 835
Gürbüzer, E. 928
Gusev, M. 276, 321, 333, 355, 361, 2253
Guseva, E. 321, 333
Gusić, J. 1060
Haberler, M. 106
Hadžić, H. 478
Haggerty, D. 1417
Hajba, M. 1873, 1901
Hajnić, M. 1715
Halla, A. 1422
Haluška, R. 533
Handroos, H. 1196
Hartikainen, A. 962
Hasanbašić, A. 478
Hasić, D. 1329
Hausknecht, K. 1487
Hawkes, E.W. 1417
Hayashi, Y. 153
Heđi, I. 447
Hekmatmanesh, A. 1196
Henno, J. 633, 879
Hess, S. 1749
Hlupić, T. 1527
Holik, F. 2129
Holjevac, M. 2064
Horvat, M. 1378
Hristova, P. 972
Hrustek, L. 1576, 1725
Hryhorova, T. 707
Hujala, M. 840
Huljić, V. 571
Hvozdeva, I. 707
Hynninen, T. 840
Ibrani, M. 544, 563
Ileš, Š. 1161, 1166
Ilić, Ž. 433
Iliev, T.B. 415, 553
Ilijanić, M. 1253
Ilišević, D. 1630
Inza, I. 2101
Ivančić, L. 1625
Ivanjko, T. 807
Ivanoska, I. 219
Ivanova, E.P. 415
Ivanovski, T. 1570
Ivašić-Kos, M. 1064, 1296
Ivković-Kihić, I. 240
Ivošević, J. 2029
Jaakkola, H. 633, 879
Jadanec, A.M. 849
Jadrić, M. 2141

Jagić, T. 2054
Jahić, O. 478
Jajčanin, A. 240
Jaklič, A. 717
Jakopec, T. 2064
Jaković, B. 1603
Jakupović, A. 1945
Jančič, M. 230
Janđel, V. 395
Janković, J. 433
Januário, F. 379
Jardas, D. 2171
Järvinen, J. P. 633
Jausz, N. 903
Jemrić, T. 1570
Jerbić, B. 1428, 1443
Jerković, H. 253
Jović, A. 373
Jović, M. 1635, 1647, 1743
Jugović, A. 1684, 1690, 1743, 1749
Jugović, K. 1749
Jukan, A. 439, 518, 557
Jukić, O. 447
Jurdana, I. 2237
Jureković, D. 913
Juretić, F. 2029
Jurić, M. 1759
Jurić, M. 1511
Jurić, Ž. 135
Juričić, V. 845
Jurišić, D. 170
Kačapor, K. 1075
Kadoić, N. 677, 1731
Kadriu, A. 977
Kadunc Kos, V. 345
Kager, D. 1017
Kainrath, K. 423
Kajtazi, E. 968
Kalajdziski, S. 219
Kaluža, M. 1945
Kamenar, E. 1417
Kamnik, R. 1411
Karadimce, A. 1179
Karanikić, P. 1635
Kardum, M. 1550
Karuza, M. 2228
Kaštelan, I. 822
Katkič, Z. 316
Katulić, T. 1759
Kedmenec, I. 1709
Kelić, I. 1615
Keremedchiev, D. 721, 1532, 1831
Keremedchieva, N. 1532
Kermek, D. 760
Kerminen, T. 1437
Kern, M. 24
Kesedžić, I. 259
Kevrić, J. 310
Khoury, J. 908
Khvorostina, Yu.V. 683
Kinder, I. 193
Kirinić, G. 774
Kiš, P. 433
Klasinc, L. 754
Ključević, Ž. 1772
Knežević, A. 867
Knežević, T. 35
Knežník, J. 733
Knezović, J. 1224
Knutas, A. 840
Koefler, F. 903
Komen, V. 1135, 1146
Kondratiev, V. 1290
Kondratiuk, S.M. 688
Konecki, M. 893
Konecki, M. 898
Korent, D. 582
Koričić, M. 69
Korov, L. 1826
Kos, L. 225
Kosec, G. 230, 235
Košmerl, I. 2048
Kosovac, A. 1582
Kostanjčar, Z. 244, 1342
Kostanjevec, I. 1593
Kostoska, M. 333
Koteska, B. 333
Kous, K. 1953
Kovač, H. 1038
Kováč, M. 84
Kovačević, I. 1499
Kovačević, T. 244
Kovačić, B. 1671, 2192
Kovačić R. 816
Kožar, I. 2009
Kožík Lehotayová, B. 766
Krajačić, G. 2171
Kramberger, R. 198
Kramberger, T. 198
Kranjec, E. 754
Krašna, M. 641, 697
Kremenjaš, D. 1939
Križan Sučić, S. 938
Križanić, S. 1538, 1576
Krog, A. 1475
Krpan, I. 1105
Kružić, S. 1411
Kryvosheya, T.M. 688
Kučak, D. 1384
Kućar, M. 816
Kucukyilmaz, T. 202
Kudela, P. 2005
Kudrina, O. 693, 812, 1621
Kuhar, S. 1953
Kulakov, A. 1364
Kuljak, I. 90

Kundrata, J. 153
Kunosić, S. 1973
Kurienkova, A.V. 693
Kutnjak, A. 1538, 1576
Kuzelj, M. 944
Kuzmanova, J. 276
Lacković, A. 1002
Lagumdžija, Z. 1075
Lahti, A. 633
Lakatoš, M. 1759
Larionov, A. 1395
Lavalle, L. 2182
Leitgeb, E. 423
Lemire, G. 56
Lenart, G. 2177
Liechtenecker, H. 796
Limani Fazliu Z. 544, 563
Limani, M. 544, 563
Linardić, I. 2040
Lindén, M. 367
Linna, P. 1422
Lipić, T. 395
Llaha, O. 538
Loboda, V.S. 683
Loku, L. 327
Lončar, M. 760
Lončarić, S. 208
Lopes, P. 1959
Lorincz, J. 571
Lozić, J. 1720
Lučan, M. 1090, 1095, 1253
Lucić, I. 995
Lyashenko, V.P. 707
Ljubičić, L. 1671
Ljubović, V. 983
Mačinko, M. 933, 1017
Mahmutović, A. 1587
Majstorović, Ž. 268
Makarun, P. 1161
Mäkelä, J. 633, 879
Mäkinen, T. 1185
Malčić, G. 1090, 1095, 1253
Malekian, R. 512
Malkoč, R. 1202
Mallia, S. 293
Maloku, H. 544, 563
Mamatnazarov, N. 2044
Marčinković, B. 1006
Marijan, S. 2023
Marijančić, V. 213
Marina, N. 1179
Marinović, I. 65
Marjanović, M. 198
Marković, L. 35
Marković, T. 61
Markovinović, I. 165
Markulin, D. 577
Marochini, M. 807
Martinho, R. 379
Martinjak, I. 253
Martins, R. 1959
Mashyna, Yu. 1621
Matetić, M. 1570
Matić, L. 2035
Matković, R. 1772
Matošević, G. 1554
Matsui, Y. 1212
Matuško, J. 1161, 1166
Medved, D. 2171
Mekterović, I. 1933
Melinščak, M. 208
Menasalvas, E. 287
Merćep, A. 244
Merdjanovska, E. 339
Mesarić, J. 249, 395
Mesić, V. 478
Mesihović, M. 983
Meško, M. 1603
Meštrović, D. 2040
Meštrović, F. 1045
Mezak, J. 749
Mezić, I. 1417
Mihajlović, Ž. 1171
Mihajlovski, B. 2135
Mihaljević, B. 2076, 2082
Mihaljević, M. 46
Mihaylov, G.Y. 415
Mihoković, V. 473
Mikhailov, A. 1374
Mikšić, I. 826
Mikulić, J. 78, 90, 96, 100, 111
Milanović Glavan, Lj. 1625
Milenković, M. 1766, 2094
Miletić, V. 389
Miličić, K. 1507
Milinković, D. 1314
Milković, M. 1720
Miljković, D. 1239, 1245
Mimica, M. 2171
Minaise, A. 908
Miočić, J. 1610
Mirceva, G. 1358, 1364
Mirković, M. 1012
Mitreski, K. 1358
Mohamed, Y.A.I. 2198
Moioli, R. 1196
Molinari, F. 2192
Mollova, S. 1867
Montoya, S. 1793
Mrčela, L. 244
Mršić, A. 577
Mršić, L. 253
Mudri, I. 933
Mudronja, G. 1737
Muharemagić, D. 2076
Muharemović, E. 1582

Muharemović, I. 983
Mulc, D. 193
Munitić, N. 1647
Muratović, D. 610
Murtazin, K. 1891
Musa, K. 577
Musa, Z. 867, 950
Musić, J. 1411
Musović, B. 1042
Mutkov, V.A. 790, 1896
Mydyti, H. 1664
Nađ, I. 1038
Nagul, N. 1395
Nagy, L. 84
Nardelli, P. 1196, 1218
Narra, N. 1422
Naudi, A. 293
Naumoski, A. 1319, 1358, 1364
Nauwelaers, B. 61
Nikitović, M. 1587
Nikolova, E. 1272, 1837
Nitz Petterson, H. 281
Nizamuddin, N. 383
Njirjak, M. 213
Nuhić, J. 310
Oehme, M. 15, 24
Okanović, V. 240
Oliva, G. 2117
Oliveira, A. 1959
Olujić, I. 1100
Omelyanenko, V. 812, 1621
Omrčen, A. 1095
Onetti Muda, A. 2117
Oparin, G.A. 2247
Opiła, J. 1778
Orehovački, T. 1118, 2070
Orel, O. 1939
Oreščanin, D. 1527
Oreški, D. 1457
Orozova, D. 1272, 1879
Osifo, O.C. 1676
Osmanović, D. 100
Osrečki, Ž. 69
Ostojić, A. 1696
Ostradický, P. 766
Ostroha, M.M. 693
Otović, E. 213
Otpuschennikov, I. 1290
Ovseník, Ľ. 533
Paar, V. 253
Palčák, M. 2005
Pale, P. 1507, 1809, 1826, 2040
Paľová, D. 649, 666
Pandža, S. 2044
Panić, I. 2237
Pap, K. 1302
Papić, A. 627
Papić, V. 1411
Paramonov, V. 1374
Pashinin, A.A. 1231, 2247
Pašić, Đ. 1384
Paternoster, G. 2214
Pathak, G. 2219
Paunkoska Dimoska, N. 1179
Paunović, V. 1820
Pavelka, R. 2232
Pavić, I. 1151
Pavičić Zajec, T. 1006
Pavković, B. 822
Pavlić, D. 2141
Pavlica, M. 711, 726
Pavlina, K. 845
Pavlovski, M. 1278, 1284
Pavlushchenko, N.M. 688
Payihuanca, K. 1793
Pažanin, J. 1049
Pecimotika, M. 1873, 1901
Peer, P. 2088
Pejčinović, B. 56, 1843
Pejić Bach, M. 743, 1603
Pejić Papak, P. 749
Pełech-Pilichowski, T. 1778
Pereira, C. 1959
Perez Murano, F. 2214
Perkov, L. 1475
Pervan, B. 1224
Peša, K. 2009
Pesek, I. 697
Petković, T. 301, 1401
Petrenko, V. 812
Petreska, E. 219
Petrić, A.M. 1527
Petrinović, D. 1347
Petrović, J. 1507, 1809, 1826, 2040
Pham, D. 908
Phung, C.V. 557
Picek, R. 1593
Pihir, I. 1703, 1725
Piho, G. 1891
Pina, J. 1959
Pivar, J. 2153
Plosila, J. 1218
Pokorný, M. 923
Poletan Jugović, T. 1690
Poljak, M. 29, 41, 46, 51
Pollino, M. 2182
Pongrac Pavlina, A. 845
Pongrac, D. 2101
Poposka, L. 321, 333
Popović, M. 822
Popović, S. 259
Povh, J. 265
Prazina, I. 240
Prelog, L. 488
Pribanić, T. 1401
Prihoda, M. 1814

Prokopenko, O. 812
Prole, B. 1302
Prouza, M. 1544
Pucihar, A. 2177
Pufek, P. 2082
Puhek, M. 754
Punko, V.V. 683
Purković, D. 1814
Rabuzin, K. 2048
Race, N. 2129
Radoš, I. 956, 1715
Radoš, I. 1715
Radoš, M. 259
Radošević, D. 1660
Radovan, A. 2076, 2082
Radulović, P. 1112
Raičković, B. 1042
Rajič, V. 2094
Rajšp, A. 1953
Ramić, I. 2044
Rantanen, P. 1185
Rašan, I. 482
Rashkovska, A. 339
Rašić, J. 1615
Ređep, T. 1006
Rehnholm, J. 281
Repnik, R. 641
Révészová, L. 655
Ribić, S. 135
Rijo, R. 379
Rista, A. 501, 527
Rizvić, S. 240
Roblek, V. 1603
Rodriguez González, A. 287
Roedig, U. 2129
Romanato, F. 2214
Rosato, V. 2182
Rosić, M. 1049
Rugelj, J. 661
Rukavina, F. 1507
Ružić, I. 1081
Saari, M. 1185, 1191
Saarivirta, M. 1437
Sagdić, A. 202
Salgueiro, F. 306
Samardžić, M. 1929
Sambolek, S. 1296
Samoday, V. 1621
Samodol, A. 1653, 1696
Sandić, A. 1511
Savchenko, D. 1196
Scattolo, E. 2214
Schatten, M. 785, 1335
Schatten, M. 785
Schatzberger, G. 78, 90, 96, 100, 111
Schulze, J. 15, 20, 24
Schwarz, D. 15, 20, 24
Seder, M. 1171
Sefidanoski, M. 1179
Seljan, S. 1278, 1284
Selmanović, E. 1329, 1999
Semenikhina, O.V. 688
Semenov, A. 1290, 1469
Serna-Marjanović, I. 1985
Sernec, R. 2177
Serov, A.N. 141, 147
Serov, N.A. 147
Seršić, D. 159
Setola, R. 2117, 2123
Sever, K. 159
Severini, E. 766
Shamonia, V.H. 683
Shaqiri, E. 361
Shatokhin, A.A. 147
Shevtsova, H. 812, 1621
Shigarov, A. 1374
Shishenko, I.V. 683
Shvets, O. 1891
Sičanica, Z. 1141
Siegl, I. 106
Silva, E. 379
Simić, S.D. 1979
Simionov, R. 1867
Simjanoska, M. 333
Sinković, G. 938
Sirovatka, G. 1463, 1487, 1907
Skala, K. 249, 395, 2259
Skansi, S. 1389, 1550
Skeledžija, I. 100
Skendžić, A. 1671
Skorić, I. 482
Skračić, K. 1507
Sladović, D. 1493
Slak, J. 230, 235
Slukan, T. 193
Smolčić, T. 301
Smrz, P. 287
Snopce, H. 968
Socal, A. 2165
Soini, J. 1185
Sokele, M. 702
Sokolov, S.A. 553
Soldo, B. 265
Soldo, R. 1100
Soni, N. 512
Sorić, M. 2101
Sosinskaya, S.S. 1263
Sosu, R.N.A. 2277
Sović Kržić, A. 933
Spahić, L. 2044
Spahić, M. 478
Spasov, D. 116, 121, 125, 419
Sprinzl, G. 2232
Stančin, I. 373
Stanič, U.J. 316
Stavanja, J. 2088

Steffan, C. 106
Stipčević, M. 130
Stipić, A. 1302
Stocksreiter, W. 423
Stoev, I.I. 790, 1855, 1896
Stojanova, N. 1319
Stojanović, S. 2035
Stojkoska, B.R. 219
Stojmenski, A. 333
Stopjaková, V. 84
Stoyanov, I.S. 415, 553
Stoyanov, S. 1855
Stresec, I. 51
Strugar, I. 743, 1603
Suligoj, T. 29, 35, 69
Sutinen, E. 962
Svedružić, Ž. 389
Šajina, R. 1324, 1979
Šajina, R. 1353, 1929
Šamija, K. 944
Šarić, F. 1342
Šečerbegović, A. 478
Šekrst, K. 1550
Šepl, V. 2015
Šestak, M. 2048
Šestanj-Perić, T. 1538
Šimić, D. 1060
Šimić, E. 1582
Šimović, V. 2035
Šimunić, D. 453
Šinkovec, M. 367
Šipek, M. 2076
Širaiy, B. 316
Širić, M. 46
Škalic, I. 65
Škokić, S. 301
Škrobonja, A. 2237
Šlibar, B. 677
Šnajder, Ľ. 832, 835
Šojat, Z. 181, 2259
Šovčík, M. 84
Špoljarić, K. 100
Špoljarić, M. 1901
Špoljarić, T. 1151, 1314
Špur, N. 754
Štefičar, P. 807
Šturlan, A. 2070
Šubić, T. 249
Šuľaj, P. 533
Šulekić, I. 1314
Šumak, B. 1953
Šuman, S. 1945
Švaco, M. 1428, 1443
Švec, M. 1166
Švelec, D. 465, 492
Šverko, Z. 165
Tanković, N. 1324, 1353, 1929, 1979
Tanović, A. 1964, 1985
Tatković, N. 1112
Tatković, S. 1112
Tekinbas, C. 202
Tesei, M. 2117
Thota, P. 1267
Thulasiram, R.K. 2270
Thulasiraman, P. 2270
Tijan, E. 1635, 1647, 1684, 1690
Tišljarić, L. 268, 2147
Tjoa, A M. 1521, 1544
Tkáčová Z. 832, 835
Tofani, A. 2182
Tolić, I.P. 111
Tomašić, I. 281
Tomic, M. 2232
Tomić I. 90
Tomić, D. 96
Tomić, D. 100
Tomić, D. 249, 395
Tomić, M. 213
Tomić, S. 1820
Tomičić Furjan, M. 1703, 1725
Tomičić, I. 1308, 1335
Tomičić-Pupek, K. 1703, 1731
Tomšič, P. 225, 1885
Tonković, A. 1023
Topolčić, D. 1493
Torić Malić, N. 2009
Torlak, I. 1684
Traxler, H. 2232
Tretinjak, M. 1084
Trivodaliev, K. 219
Trkman, M. 1202
Trobec, R. 367
Tsvetanov, F. 459
Tsvetkova, I.D. 790, 1896
Tuba, M. 1329
Tuparov, G. 721, 1831, 1879
Tuparova, D. 1879
Turkalj, D. 1615
Udovičić, P. 1939
Udovychenko, O.M. 693
Ullah Kakakhel, S.R. 1218
Ullah, M. 1196, 1218
Ungermanns, C. 796
Urem, F. 913
Usenko, T.R. 1263
Valčić, M. 1653
Valenčić, D. 597, 604
van Duijsen, P.J. 1849

van Slageren, J. 24
Vasileska, I. 225
Vejačka, M. 649, 666
Vejmelka, L. 1759, 1772
Velinov, A. 327
Verdonik, M. 1056
Vertel, A.V. 688
Vidaček Hainš, V. 816
Vidal, E. 1793
Vidović, D. 193
Vignjevikj, R. 1319
Vihervaara, J. 1191
Vijtiuk, J. 1475
Vikström, S. 281
Viljanen, J. 1191
Vinko, D. 74
Vinković, A. 2040
Višković, A. 453
Vitas, V. 1056
Vlah, D. 170
Vlahinić, S. 165, 1135, 1146
Vlahović, H. 213
Vlahović, I. 253
Vlahović, N. 2153
Vlašić, I. 1090
Vlašić, T. 159
Voitenko, A.A. 683
Vojković, G. 1720, 1766, 2094
Voštinár, P. 733, 738
Vrana, R. 801
Vranešić, P. 873
Vranješ, M. 822
Vrankić, M. 165
Vrbanc, D. 1028
Vrbanec, D. 1448
Vrček, N. 1731
Vrdoljak, B. 1389
Vretenar, M. 2228
Vučkov, D. 995
Vučković, K. 2070
Vujaklija, I. 1141
Vujković, S. 1630
Vukojević, J. 193
Wakabayashi, N. 2237
Wang, Y. 2265, 2270
Wanitzek, M. 15
Watashiba, Y. 1212
Weißhaupt, D. 24
Werth, W. 796, 903
Westerlund, T. 1218
Wolff, A. 1218
Woudstra, J.B. 1849
Wu, H. 1196
Zábovská, K. 1991, 2005
Zábovský, M. 1991
Zaharieva, S.L. 790, 1855, 1896
Zajc, I. 2177
Zalović, L. 473
Zalović, V. 473
Zanin, M. 287
Zdraveski, V. 276
Zdravev, Z. 327
Zelenika, S. 1417
Zenuni, X. 501, 527, 1664
Zhang, G. 1570
Zhang, M. 61
Zlodi, G. 807
Zoroja, J. 743
Zrnić, L. 1826
Zuidervliet, D.C. 1849
Zysińska, M. 2186
Žagar, M. 1463, 1481
Žeželj, B. 1157
Žgaljić, A. 1443
Žgaljić, D. 1635
Žilak, J. 69
Žilak, M. 482
Žilić, T. 1641
Žufić, A. 989
Žufić, J. 989
Žulj, S. 1907
Žunić, E. 1999
Župančić, T. 51
Županović, V. 170

FOREWORD

The *43rd International ICT Hybrid Convention MIPRO 2020* was held from September 28th, 2020 to October 2nd, 2020 in Opatija, Croatia. The Convention consisted of thirteen conferences under the titles: *Microelectronics, Electronics and Electronic Technology (MEET), Data Science and Biomedical Engineering (DS-BE), Telecommunications & Information (CTI), Computers in Education (CE), Computers in Technical Systems (CTS), Intelligent Systems (CIS), Robotics Technologies and Applications (RTA), Information Systems Security (ISS), Business Intelligence Systems (miproBIS), Digital Economy and Digital Society (DE-DS), Information and Communication Technology Law (ICTLAW), Engineering Education (EE)* and *Software and Systems Engineering (SSE)*. An additional conference was dedicated to the works of students: *MIPRO Junior-Student Papers (SP)* and there were two special sessions: *Optoelectronics and Photonics (OPHO)* and *Dew Computing (DEWCOM).*

This comprehensive Book of Proceedings includes all positively reviewed and presented papers from each of these conferences. All the papers were reviewed by an International review board and the list of reviewers is contained in the Book of Proceedings. These papers were written by the authors from universities, scientific institutions, educational institutions, industry and other participants. It also includes last year's papers from *42nd International ICT Convention MIPRO 2019* conference *Smart, Sustainable And Resilient Cities And Infrastructure (SSRCI).*

The convention was organized by the Croatian ICT Society MIPRO with the help of numerous patrons and sponsors to whom we owe our sincere thanks. We especially single out our golden sponsors HEP-Croatian Electricity Company and Končar-Electrical Industries along with our bronze sponsor Storm Computers.

To all who helped organizing the 43rd International ICT Hybrid Convention *MIPRO 2020* as well as editing of this Book of Proceedings we extend our heartfelt thanks.

Academician Karolj Skala, PhD
International Program Committee
General Chair

MEET

International Conference on

MICROELECTRONICS, ELECTRONICS AND ELECTRONIC TECHNOLOGY

Steering Committee

Chairs: Marko Koričić, University of Zagreb, Croatia
Željko Butković, University of Zagreb, Croatia

Members: Slavko Amon, University of Ljubljana, Slovenia
Dubravko Babić, SEMICRO, Croatia
Maurizio Ferrari, CNR-IFN, Povo-Trento, Italy
Mile Ivanda, Ruđer Bošković Institute, Zagreb, Croatia
Branimir Pejčinović, Portland State University, USA
Jorg Schulze, University of Stuttgart, Germany
Tomislav Suligoj, University of Zagreb, Croatia
Aleksandar Szabo, IEEE Croatia Section
Davor Vinko, University of Osijek, Croatia

INVITED PAPER

Infrared Quantum Detectors

D. Grubišić and D. Babić
Laser Components DG, Inc., Tempe, Arizona, USA
dgrubisic@laser-components.com dbabic@laser-components.com

Abstract **- Quantum detectors respond to electromagnetic radiation by generating a photocurrent. The detection mechanism involves absorption of photons whose energy exceeds the band gap of the detector material. The absorption generates excited mobile charge carriers (electron-hole pairs) that form the photocurrent. The quantum detectors sensitive in near infrared (NIR) and middle infrared (MIR) wavelength range will be discussed. Different quantum detector types will be outlined including their detection mechanism and performance. Most recent developments in room temperature type-II strained-layer superlattice or digital alloy detectors with sensitivity range between 3 – 5 μm will be included.**

Keywords – quantum detectors, infrared, silicon and InGaAs avalanche photodiodes, lead salts, type II superlattice nBn detectors

I. INTRODUCTION

Quantum detectors generate photocurrent in response to electromagnetic radiation when the photon energy is larger than the band gap of the quantum detector material. The quantum detector types sensitive in near infrared (NIR) and middle infrared (MWIR) wavelength range will be discussed including their detection mechanism and performance. The quantum detector types discussed are silicon avalanche photodiodes, InGaAs avalanche photodiodes, lead salt photodetectors and type II superlattice unipolar barrier detectors. This paper reflects authors' background and interests and is by no means exhaustive and comprehensive review of the literature.

II. SILICON AVALANCHE PHOTODIODES

Silicon avalanche photodiodes (APD) are well established optical quantum detector family, both technically and commercially, due to their outstanding properties [1]. Among the outstanding properties of the silicon APDs are excellent detection sensitivity due to high internal gain and low noise, fabrication compatibility with standard silicon based integrated circuit manufacturing processes and an advantageous cost to performance ratio. The high internal gain is achieved by applying high voltage bias to a suitably designed device that facilitates impact ionization of the charge carriers: electrons and holes. The impact ionization increases current carrying charge by multiplication providing the internal gain. Electrons are much more ionizing than holes in silicon and this enables impact ionization with low noise content supporting the high sensitivity of silicon APDs. Thus, the light absorbing and multiplication regions in silicon APDs must be p-type so that the electrons initiate the impact ionization processes.

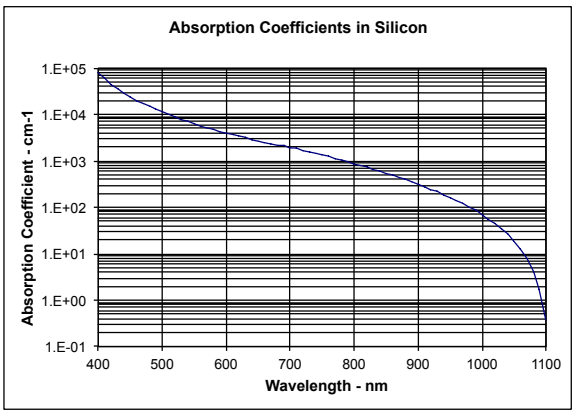


Fig. 1. Absorption coefficient in silicon.

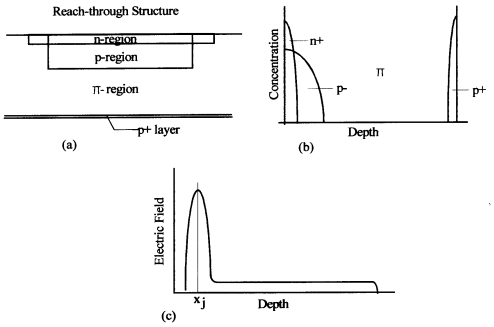


Fig. 2. Silicon reach-through APD: (a) structure, (b) concentration and (c) field profiles.

Silicon absorbs light in a wide wavelength range, from ultraviolet (UV) to near infrared (NIR). Fig. 1 shows the absorption coefficient of silicon as a function of the wavelength in the range 400 to 1100 nm. The absorption coefficient is very high for shorter wavelengths while it is substantially lower for longer NIR wavelengths. Optimum wavelength range of silicon APD operations is between 700 nm and 900 nm where compromise between sensitivity/noise and device characteristics thermal stability is the best. Reflection at the back surface is a "built-in" feature allowing for double pass of the incoming photons through the absorption region and in that way improving device structure while increasing already high quantum efficiency. Applications of the NIR operating silicon APDs are as LIDAR range finders, in analytical instruments for low-level-light detection, spatial light transmission, optical fiber communications, YAG

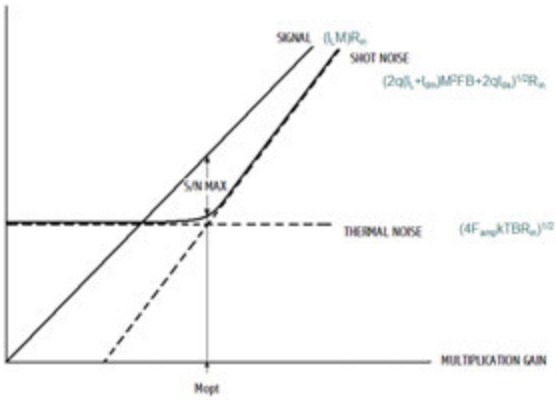


Fig. 3. Optimum operating point location as function of multiplication gain M.

laser detection, and confocal instrumentation in fluorescence measurements.

Silicon avalanche photodiodes are generally designed to have reach-through structure. The basic design of a reach-through APD consists of a narrow high-field region where the multiplication takes place, with a much wider low-field region in which the incoming radiation is absorbed. The design, impurity concentration, and electric field profiles are shown in Fig. 2.

For satisfactory operation of the APD reach-through design of Fig. 2(a), the high resistivity p-type substrate must be fully depleted by the applied bias voltage. Generally, this works well provided the substrate wafer is not too thick, and required response times are not less than about 10 ns or so. Most common silicon APD structure uses an epitaxial version of the design shown in Fig. 2. In this approach, a high-resistivity p-type layer is grown epitaxially on top of a low resistivity p-type substrate. The substrate may be any thickness, and its resistivity is chosen to be low enough that it does not introduce a significant series resistance. The thickness of the p-type epitaxial layer is typically chosen to be in the range 30 to 50 µm, but may be narrower or wider, depending on the requirement of the application. A reach-through structure is achieved by introducing n- and p-layers, as shown in Fig. 2(a). When bias voltage is applied, the depletion layer stops at the interface between the substrate and the epitaxial layer. Where fast response is a requirement, the narrow absorption region of this version of the reach-through APD is normally the best option.

Silicon avalanche photodiodes exhibit outstandingly low multiplication noise [2]. This advantage of the silicon APDs is due to very low impact ionization ratio k of hole to electron impact ionization coefficients whose value is 0.02 for bulk silicon. The low value of k coupled with high achievable values of the multiplication gain gives silicon APDs their superior detection sensitivity. However, since APD noise is increasing as $M^{1.5}$ while signal is directly proportional to the multiplication gain M, there will be an optimum operating bias position for every APD with the highest signal-to-noise ratio (S/N). Increasing the gain by increasing the bias beyond the optimum operation bias would decrease S/N. Fig. 3 shows location of the optimum operating point as a function of the multiplication gain M.

Laser Components DG (LCDG), a leading IR detector vendor, offers near infrared silicon avalanche photodiodes with superb performance characteristics [3]. The SAT series have breakdown voltage above 400 V with quantum efficiency exceeding 40% at 1064 nm. The SAT APDs come in 2 sizes (800 µm and 3 mm diameter) with the dark current of 1-5 nA, capacitance of 2-10 pF, rise time of 1-3 ns and noise of 0.5-1 pA/sqrt(Hz), all values being typical at the operating point and varying on account of the device size.

III. INGAAS AVALANCHE PHOTODIODES

InGaAs *pin* optical detectors, that are lattice matched to InP and have the composition $In_{0.53}Ga_{0.47}As$ (InGaAs from now on), absorb light in the 1 to 1.7 µm wavelength range with the peak at about 1.5 µm. The 1.5 um wavelength falls near a minimum absorption wavelength of the optical fibers and is also safe for human eyes. This is behind wide popularity of infrared InGaAs optical detectors. The *pin* detectors have no internal gain mechanism so their sensitivity is relatively limited. InGaAs avalanche photodiode (APD) would provide internal gain dramatically increasing sensitivity of such detectors.

The InGaAs APDs have been successfully developed and commercialized finding applications in optical communication systems, range finding (LIDAR), optical coherence tomography, low level light detection and a number of other applications. It has to be immediately stated that the InGaAs APDs have very different internal structure than their far better known silicon counterparts [4]-[7]. The reason is a small bandgap of the lattice matched InGaAs that effectively prohibits application of high electric fields necessary for efficient impact ionization and resulting avalanche gain. Application of the high electric fields in InGaAs would cause large tunneling dark currents preventing utilization of such devices as photodetectors. Separate absorption and multiplication (SAM) APD structure addresses the problem. The SAM diodes consist of an InGaAs region for light absorption connected to another lattice matched material with wider band gap where the *pn* junction and high field multiplication regions are located.

InP is the first and most widely established lattice matched wide band gap material used in SAM APDs to form the multiplication layer. Besides the InGaAs absorption layer and InP multiplication layer, the full SAM APD structure also contains a charge (field control) layer whose function is to reduce the high electric field existing in the InP multiplication layer to a much lower value in the InGaAs layer thus limiting the dark tunneling current of the total structure. Both the charge layer and multiplication layer are InP, generating very low tunneling dark current because of the large band gap.

Large band discontinuity exists at the InP/InGaAs interface causing pile up of the carriers that dramatically lowers the speed of the APD. This issue is addressed by introducing a very thin InGaAsP layer with graded band

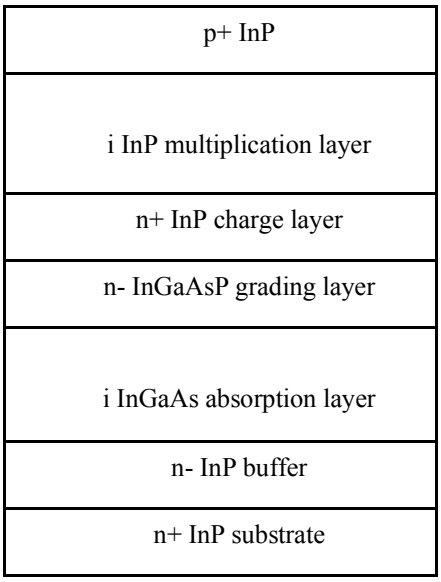


Fig. 4. Schematic cross section diagram of InP/InGaAs APD device structure.

gap, called the grading layer, between the InGaAs and InP in order to smooth the discontinuity and provide for fast carrier transport across the interface. The APD structure has become SAGCM APD meaning separate absorption, grading, charge and multiplication APD. Schematic cross section diagram of an SAGCM InP/InGaAs APD is shown in Fig. 4. The grading layer is lattice matched to InGaAs and InP to avoid crystal defects.

Impact ionization coefficient for holes is larger than the equivalent coefficient for electrons in InP so the SAGCM APDs with an InP multiplication layer are designed so that holes are the carriers that initiate the impact ionization because such arrangement results in lower excess noise. Thus, the InGaAs layer and both multiplication and charge layers are doped n-type with the charge layer being generally highly doped and the other two layers having very low or intrinsic doping. Operation of an SAGCM APD InP/InGaAs device is initiated by photogeneration of electron-hole pairs in the InGaAs absorption layers. The photogenerated minority holes are easily injected into the high field InP through the grading layer where they are accelerated to initiate efficient impact ionization thus producing the internal current gain. Finally, photo and ionization generated holes are collected by an anode.

In InP/InGaAs APDs the anode is usually produced by zinc diffusion from the surface of InP facilitating fabrication of a planar device. Mesa versions of the SAGCM InP/InGaAs APD structure are also popular having the anode doped in situ during the epi layer growth.

Replacement of InP as the multiplication layer by lattice matched InAlAs (actual composition $In_{0.52}Al_{0.48}As$) has attracted considerable attention because application of InAlAs offers reduction of the multiplication noise [5]. The reason is a smaller value of the impact ionization

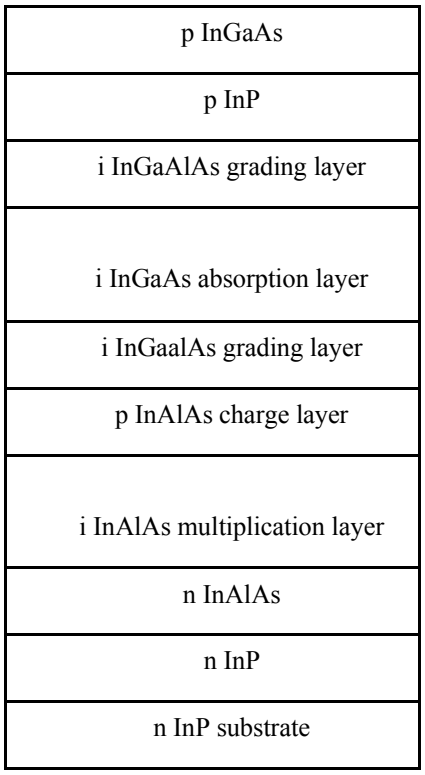


Fig. 5. Schematic cross section diagram of InAlAs/InGaAs APD device structure.

coefficient ratio *k* in InAlAs that leads to lower excess noise per the standard multiplication noise theory. However, the impact ionization coefficient of electrons is larger than the impact ionization coefficient of holes in InAlAs so the electrons have to be made the minority carriers that initiate the impact ionization in an APD structure containing the InAlAs multiplication layer. This means that the doping of the different layers in an SAGCM APD with InAlAs has to be reversed with respect to the corresponding layers in an SAGCM APD with InP, not a big issue for the modern wafer growers. More significant difference seems to be far greater popularity of mesa structured APDs when InAlAs is used. Schematic cross section diagram of an ASGCM InAlAs/InGaAs APD is shown in Fig. 5.

The impact ionization coefficient ratio *k* has the value of 0.4-0.5 in bulk InP (holes initiate ionization) and 0.2-0.3 in bulk InAlAs (electrons initiate ionization). Further reduction of the *k* ratio, required to reduce the excess multiplication noise, has been achieved by applying the impact ionization engineering (I^2E) approach [6]-[7]. The I^2E is based on the nonlocal nature of the impact ionization process as described by the "dead space" concept. In its simplest implementation, the I^2E approach modifies the bulk InAlAs multiplication layer so that it consists of several repeated units. Each unit consists of a very thin layer (tens of nm) of a wide bandgap InAlAs adjacent to a very thin layer (tens of nm) of a narrow band gap InAlGaAs (lattice matched). The ionizing carriers gain energy in the wide bandgap layer but cannot initiate ionization there due to its wide bandgap. They cause ionization quickly upon entering the narrow bandgap

layer. The ionization event in the narrow bandgap material is more spatially localized and deterministic than an ionization event in a bulk material. The increased determinism reduces the excess noise and achieves very low values of the impact ionization ratio. InAlAs/InGaAs APDs with an I^2E multiplication layer have k below 0.1. This improvement in the APD noise characteristics is achieved by substantial increase of complexity and cost of the APD structure and its fabrication.

LCDG offers high quality InGaAs avalanche photodiodes with exceptional performance characteristics [3]. The offered APDs have breakdown voltage in the 45-75 V range. Responsivity of the devices at 1550 nm and unity gain is typically 0.94 A/W. Their dark current is typically below 50 nA at the operating point defined as G=10, while typical excess noise factor is 3.2 at the same point. For 0.2 mm diameter device typical capacitance is 1.7 pF, bandwidth (at G=10) 1 GHz and NEP 0.07 pW/sqrt(Hz).

IV. LEAD SALT PHOTODETECTORS

Lead salt photodetectors are very popular due to their exceptional price to performance ratio and room temperature operation [8]. Lead sulfide (PbS) and lead selenide (PbSe) are commercially available lead salt photodetectors sensitive in the 1-3 μm (PbS) and 1-4.5 μm spectral ranges (PbSe). Their main applications are medical and environmental gas analysis, process control to detect and analyze moisture and other gases and materials, flame and fire detection and optical pyrometry.

Photosensitivity of the lead salts was discovered in the 1930s when Kutzscher discovered photoconductive properties of PbS making it the first infrared semiconductor material. Germany started production of PbS detectors around 1943 for military and commercial applications. Research and development of the lead salt detectors was initiated in U.S. immediately after World War II and lasted until approximately late 1960. The lead salt detectors are polycrystalline thin films of PbS or PbSe with approximately 1 μm thickness, usually deposited on glass or quartz substrate. Although vacuum deposition of the thin films has been investigated, currently available thin films are deposited by chemical bath deposition (CBD). The deposited films have to be sensitized by oxygen to become photosensitive. PbSe films are, in addition, sensitized by iodine. These CBD prepared polycrystalline lead salt thin films have been found to have superior performance to single-crystal films. The CBD technique offers significant cost advantages without compromising performance of the photodetectors, including feasible fabrication of large area (several mm squared) lead salt detectors. Fig. 6. shows a photograph of patterned and metallized PbSe film.

The lead salt photodetectors are p-type photoconductors whose conductivity is increased upon illumination. The high photoconductivity is a result of photogenerated electrons becoming trapped at sites such as oxygen atoms dramatically increasing lifetimes of mobile holes. This model applies to the CBD prepared lead salt photoconductors. Other photoconductivity models have been proposed but are not supported by experimental evidence. A more recent model proposes that the PbSe crystallites consist of p-type cores and n-type shells forming *pn* junction. The *pn* junction spatially separates electrons and holes increasing carrier lifetimes and thus improving detector performance.

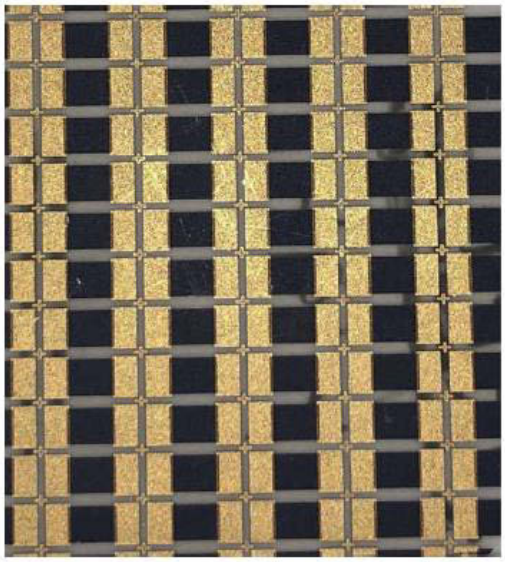


Fig. 6. Photograph of patterned and metallized PbSe film.

PbS and PbSe photodetectors are very sensitive to environmental temperature due to their narrow band gaps. However, their band gaps decrease when they are being cooled contrary to what one generally observes in elemental and III-V semiconductors. No clear explanation of this anomaly has been accepted.

Illumination induced resistance change of lead salt photodetectors is in the range of below or around 0.1% of its dark resistance. Easy measurement of the photoinduced signal is accomplished by modulation of the IR illumination by chopping. The chopping shifts the signal of interest from DC to AC at the chopping frequency enabling easier extraction and amplification of the signal by an AC coupled amplifier. The measured signal is usually a change of voltage across the photoconductor induced by the illumination. The photoconductor is biased via a voltage divider circuit. Simple circuit analysis shows that the AC signal is proportional to a product of the change in the resistance of the photoconductor and to the bias voltage.

Resistance of the lead salt photodetectors tends to slowly and weakly drifts over time following environmental temperature variations. This is another reason for applying the chopped illumination. The measured AC signal of the photodetector is proportional to the bias voltage suggesting application of high voltage bias. However, the high voltage bias causes heating of the detector reducing its dark resistance that also reduces its photoconductivity. The optimum bias voltage has been empirically found to be 50 V/mm of the photodetector's

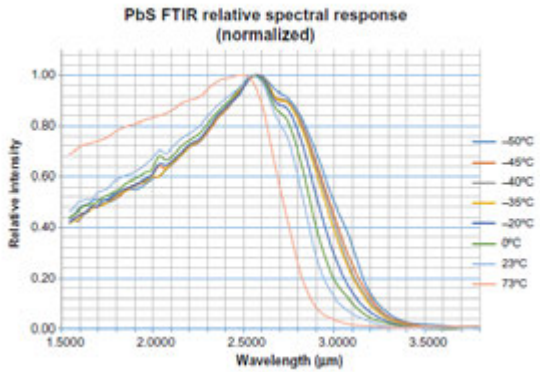


Fig. 7. Spectral response of PbS detector (normalized) measured at different temperatures.

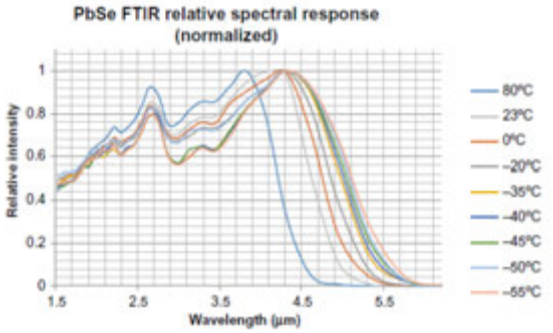


Fig. 8. Spectral response of PbSe detector (normalized) measured at different temperatures.

electrode separation. Need to overcome system noise sets the minimum acceptable bias. As quality of the amplifier electronics has been increasing, lower bias voltages are becoming acceptable.

Generation-recombination and 1/f noise are dominant noise components for the lead salt detectors. The photodetectors have historically been specified at 1 kHz to avoid 1/f noise contribution. However, as the lead salt detectors are frequently used at 90 Hz, LCDG provides the performance data at this frequency in addition to the 1 kHz data [3]. Achievable detectivity D* is typically $1.1 \cdot 10^{11}$ and $1.8 \cdot 10^{10}$ Jones at room temperature for PbS and PbSe, respectively. Prolonged exposure of the lead salt detectors to UV and visible light should be avoided because it can change or degrade their performance.

The anomalous temperature dependence of the lead salt band gap and its high sensitivity to temperature variations has enabled a very sophisticated technique for tuning the peak response wavelength of a lead salt detector by controlled cooling, most frequently by using a thermoelectric cooler. Fig. 7 shows spectral response of a PbS photodetector measured by an FTIR instrument at different temperature. Fig. 8 present spectral response of a PbSe photodetector in the same manner. The curves in the Figs. 7 and 8 illustrate temperature sensitivity of the spectral response, especially of the cut-off wavelength. The spectra shown in the Figures have been measured and collected in an extraordinary wide temperature range in addition to being corrected for the imperfections of the FTIR measurement protocol.

In summary, the lead salt detectors are highly cost effective and sensitive family of infrared detectors whose behavior does not conform to the intuition based on single crystal elemental and III-V semiconductors.

V. MID-WAVE INFRARED TYPE-II SUPERLATTICE UNIPOLAR BARRIER DETECTORS

III-V compound semiconductors and their alloys have received great attention in recent years as a platform to design and fabricate mid-wave and long-wave infrared detectors (MWIR and LWIR) due to their substantial advantages over other competing material systems. These advantages include temporal stability, spatial uniformity, scalability, high operability, producibility and affordability [9].

Unipolar barrier detectors utilizing type-II superlattice (T2SL) as optical absorbing layers are the most promising III-V compound semiconductor detectors because they offer potential of high quantum efficiency and operation at or near room temperature due to reduced dark current [10]. Type-II superlattices for MWIR and LWIR absorbers usually consist of alternating very thin (nanometer scale) sublayers of InAs and GaSb or InAs and InAsSb [9]. Variations of the each sublayer thickness enable adjustment of the absorption cut off wavelength of the absorber. Efficient optical absorption by the T2SL requires that its thickness is of the order of its absorption length. The materials of the sublayers have very similar lattice constant leading to zero or very low strain crystal growth. The materials exhibit type-II band gap alignment. The alignment causes separation of the charge carriers, electrons into the InAs and holes into the other constituent material sublayers. The spatial separation of the charges reduces Auger recombination, but also reduces optical absorption. However, electron wave functions in the InAs sublayers have significant overlap thus forming a miniband. The same is true of the heavy hole wave functions in the other superlattice sublayer material. The band gap between the minibands may be smaller than the band gaps of each material forming the type-II superlattice. Optical absorption of the superlattice involves electron hole transitions between the minibands and is a direct transition in the reciprocal space.

The III-V semiconductor materials used to grow the T2SL belong to the so called 6.1 Å material system because their lattice constants are approximately 6.1 Å. The materials are InAs, GaSb, AlSb and their alloys. Initially, the research tended to focus on InAs/GaSb superlattices because of better lattice match. However, the InAs/GaSb structure exhibits very short minority carrier lifetime attributed to presence of gallium (~ 80 ns). Ga-free InAs/InAsSb T2SL have subsequently become very popular due to their substantially larger minority carrier lifetime (~1-10 µs) and growth and processing advantages. The InAs/InAsSb superlattice are grown to be strain balanced and this places a constraint on their

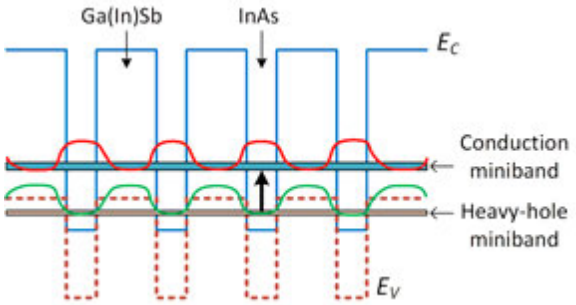


Fig. 9. Schematic illustration of InAs/GaSb T2SL structure.

flexibility. Another disadvantage of the InAs/InAsSb is lower optical absorption compared to the InAs/GaSb. An emerging consensus appears that currently favors InAs/InAsSb for MWIR and InAs/GaSb for LWIR. GaSb wafers are used as the substrates for the epi layer detector growth in both cases. Most detector growth has been performed using molecular beam epitaxy (MBE) although detectors grown by metalorganic chemical vapor deposition (MOCVD) have been reported recently. A schematic illustration of InAs/GaSb type-II superlattice structure is shown in Fig. 9.

The unipolar barrier serves to mitigate large dark current of the T2SL detector structure [11]-[12]. The dark current is large primarily due to the small band gap and consists of diffusion, generation-recombination and surface components. The unipolar barrier blocks majority carriers and passes photogenerated minority carriers thus facilitating operation of the device as a detector with reduced dark current. The barrier can also block the surface leakage current if conductivity type of the current collector on the surface is the same as the conductivity type of the T2SL absorber. The Fermi level at the surface of InAs is pinned in its conduction band pointing to a degenerate n-type InAs surface even in case of p-type doping. Standard generation-recombination process is substantially inactive leading to longer carrier lifetime in InAs. The above considerations favor the n-type doping for the unipolar detector collector and T2SL absorber. The nBn unipolar detector is the most popular barrier detector configuration. In an nBn detector n-type T2SL absorber is separated from the n-type collector by an electron barrier. The electron barrier contains aluminum in addition to other 6.1 Å semiconductor alloys thus being lattice matched and having large band gap. The valence band offset of the barrier heterojuntions is designed to be zero or very small, while the conduction band offset is designed to be large, preventing flow of the electrons. Thus, the holes are the minority carrier in an nBn detector structure.

The nBn barrier reduces the dark current in two ways: (a) it blocks electron surface leakage current flowing on the degenerate n-type surface between metal contacts and (b) it suppresses generation- recombination dark current component. The generation-recombination current is suppressed because the unipolar nBn detector lacks the usual depletion region or, if a depletion region exists, it is located within the large band gap barrier. Theoretical

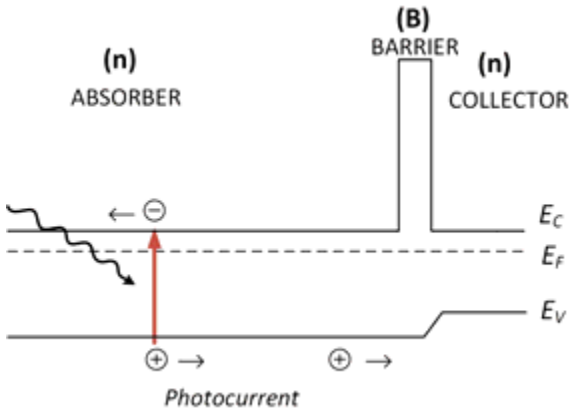


Fig. 10. nBn detector energy band diagram illustration.

models and experiments agree that the dark current of an nBn detector is mainly the diffusion current. Fig. 10 illustrates energy band diagram of an nBn detector with absence of valence band misalignment.

The nBn detector structure does not contain a *pn* junction with the depletion region, but its unipolar barrier serves to block the flow of the majority electrons. The photogenerated minority holes form photocurrent by diffusion. Thus, the nBn detector is a hybrid between a photoconductor and a photodiode. The hole effective mass in the growth direction is very large based on the T2SL band structure. The large hole effective mass is favorable in reducing tunneling, but adversely affects mobility of the holes thus also reducing their diffusion coefficient and diffusion length. However, it has been found that the growth direction effective mass of the holes is not the best descriptor of their transport. The better descriptor is the 'conductivity effective mass' obtained by thermally averaging over the superlattice band structure [13]. The averaging leads to a much lower 'conductivity' effective mass value and faster hole transport.

A small bias in the range from 0 to 1.0 V is frequently applied to a barrier detector to improve its optical absorption, especially its quantum efficiency. The reason is imperfect alignment between the valence bands of different layers of the detector that impede the hole transport. Application of the bias serves to 'smooth' the alignment and so facilitate unimpeded and fast hole transport.

Survey of the literature shows that the recently reported T2SL nBn detectors in the 3-5 µm range achieve dark current of approximately 1 A/cm^2 at room temperature and approximately $3 \cdot 10^{-2}$ A/cm^2 at 222 K. The measured quantum efficiency is in the 40 to 50% range with no anti-reflection (AR) coating. Thus, the D* detectivity of these detectors reaches $2 \cdot 10^{10}$ Jones at 4 µm and 222 K.

VI. CONLUSION

Several different types of infrared quantum detectors such as silicon avalanche photodiodes, InGaAs avalanche photodiodes, lead salt photodetectors and type-II

superlattice unipolar barrier detectors have been discussed reflecting authors' background and interests. The discussion has outlined detection mechanism and performance of the detectors.

ACKNOWLEDGMENT

The authors thank Prof. Tomislav Suligoj for inviting them to present this paper.

REFERENCES

[1] P. P. Webb, R. J. McIntyre, and J. Conradi, "Properties of avalanche photodiodes," RCA Review, vol. 35, June 1974.

[2] R. J. McIntyre, "Multiplication noise in uniform avalanche diodes," IEEE Trans. on Electron Devices, vol. ED 13, pp. 164-168, January 1966.

[3] Laser Components USA, Inc.. [Online]. Available: https://www.lasercomponents.com/us/.

[4] Y.Liu et al., "A planar InP/InGaAs avalanche photodiode with floating guard ring and double diffused junction," J. Lightwave Technol., vol. 10, pp. 182-193, February 1992.

[5] J. C. Campbell, S. Demiguel, et al., "Recent advances in avalanche photodiodes," IEEE J. Sel. Top. Quant. Electron., vol. 10, pp. 777-787, July/August 2004.

[6] J. C. Campbell, "Recent advances in telecommunications avalanche photodiodes," J. Lightwave Technol., vol. 25, pp. 109-121, January 2007.

[7] J. C. Campbell, "Recent advances in avalanche photodiodes," J. Lightwave Technol., vol. 34, pp. 278-285, January 2016.

[8] D. Babic, L. W. Johnson, L. V. Snyder, and J. J. San Roman, "Lead salt photodetectors and their optoelectronic characterization," in Chalcogenide: From 3D to 2D And Beyond, X. Liu, S. Lee, J. K. Furdyna, T. Luo, and Y.-H. Zhang, Eds. Woodhead Publishing, Elsevier, 2019, pp. 182-193.

[9] D. Z.-Y. Ting et al., "Type-II Superlattice Infrared Detectors,", in Semiconductors and Semimetals vol. 84, S. D. Gunapala, D. R. Rhiger, and C. Jagadish Eds., Elsevier, 2011, pp. 1-57.

[10] D. Z. Ting et al., "Mid-wavelength high operating temperature barrier infrared detector and focal point array," Appl. Phys Lett. vol. 113, p. 021101, 2018.

[11] S. Maimon and G. W. Wicks, "*nBn* detector, an infrared detector with reduced dark current and higher operating temperature," Appl. Phys. Lett. vol. 89, p. 151109, 2006.

[12] D. R. Rhiger and E. P. Smith, "Carrier transport in the valence band of nBn III-V superlattice infrared detectors," J. Electron. Mater. vol. 48, pp. 6053-6062, 2019.

[13] D. Z. Ting, "Advances in III-V semiconductor infrared absorbers and detectors," Infrared. Phys. Technol. vol. 97, pp. 210-216, 2019.

PAPERS

Ge-on-Si Avalanche Photodiodes for LIDAR Applications

M. Wanitzek*, M. Oehme, D. Schwarz, K. Guguieva, and J. Schulze *Senior Member IEEE*
Institute of Semiconductor Engineering (IHT), University of Stuttgart,
Pfaffenwaldring 47, 70569 Stuttgart, Germany
*) Corresponding author: st149364@stud.uni-stuttgart.de

***Abstract* – In order to make autonomous driving in cars possible, a precise knowledge of the immediate surroundings is required. One technology for near-field detection is the LIDAR (light detection and ranging) technology. Currently available LIDAR systems operate at a wavelength of 905 nm. However, for wavelengths from 1,300 nm, a significant increase in range resolution is achieved. Here, the receiver side is typically realized as APDs (avalanche photodiodes). In this work the fabrication and characterization of APDs with an absorption region made from Germanium (Ge) are presented. The layer sequence for the APDs are grown directly on Silicon (Si) substrate using a molecular beam epitaxy system. At room temperature the Ge-on-Si-APDs achieve responsivities of up to 6 A/W at a wavelength of 1,310 nm, which corresponds to a gain of 26 compared to conventional Ge photodiodes.**

Keywords – Ge-on-Si; avalanche photodiode; photodetector; lidar

I. INTRODUCTION

In recent years the demand for high-sensitive, low-noise and fast detectors has increased significantly [1] for various emerging applications as quantum key distribution [2], time-of-flight ranging [3], three-dimensional imaging [4] and time resolved spectroscopy [5]. Single photon avalanche photodiodes (SPADs) can fulfill these requirements. SPAD detectors are avalanche photodiodes (APD), which are biased above breakdown voltage in the so-called Geiger-mode. In this mode, a self-sustaining avalanche current can be triggered by a single incident photon [6].

While SPADs made from Silicon (Si) fulfill these requirements at wavelengths below 1,000 nm in large scale complementary metal-oxide-semiconductor (CMOS) arrays [7], efficient detectors for wavelengths greater than this remains unsolved. Despite the fact that SPADs made from Indium gallium arsenide (InGaAs) are commercially available, detectors made from III-IV semiconductors have some drawbacks. On the one hand they are expensive and on the other hand it is hard to CMOS-integrate such detectors in a detector array. A solution to this problem is given by the use of Germanium (Ge), which can be fully integrated within existing Si technology and therefore enable integrated read-out circuits.

Applications as quantum communication in optical fibers require detectors for the wavelengths of 1,310 nm and 1,550 nm. In LIDAR technology SPAD detector arrays are already used although they are fabricated from Si and therefore operate at wavelengths around 900 nm, the so-called infrared(IR)-A region. Using higher wavelengths in the IR-B region brings the positive effect, that significantly more optical power is permitted due to eye safety. Another problem is the glare of the cameras from the sun. This worsens the signal-to-noise ratio. One solution to this is the light irradiation from the backside of a Si wafer. Si acts as an absorber for a large part of the solar spectrum.

The usage of higher wavelengths in the IR-B region will be possible if SPAD detectors made from Ge and integrated on Si are used. Ge shows good absorption for wavelengths up to approx. 1,550 nm [8]. In Fig. 1 the spectral sensitivity of back-site illuminated Ge diodes is shown. Also the solar spectrum is plotted in this figure as the AM1.5 spectrum and idealized as the black body of a temperature of 5800 K. It shows that most of the sunlight can be filtered out by the back lighting. Using Ge as absorber material also gives the option to further increase the sensing wavelength by alloying Tin (Sn) into the active region [10]. As also shown in Fig. 1 an increasing Sn concentration increases responsivities at higher wavelengths.

Ge-on-Si-SPADs with good optical properties were already reported in literature, but they rely on very low temperatures at around $T = 100$ K [11]. It is the goal of this work to achieve these properties at room temperature or at least at temperatures that can be obtain with thermoelectric coolers what enables the usage in automotive applications.

In this work we show Ge-on-Si-APDs operated in linear mode with a gain of 26 compared to reference Ge-pin-photo diode with the same absorption region.

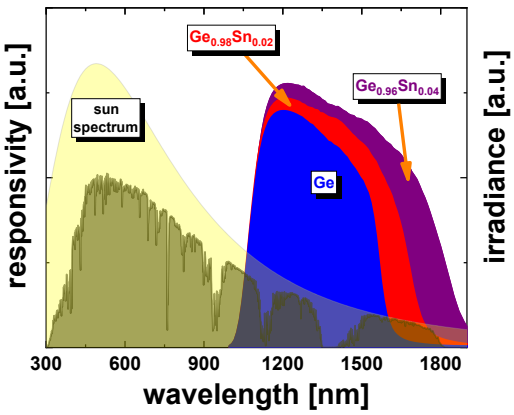


Figure 1: Responsivities of backside illuminated Ge-on-Si diodes and Ge_xSn_{1-x}-on-Si diodes with a Sn concentration of up to 4 %. The solar spectrum is also plotted.

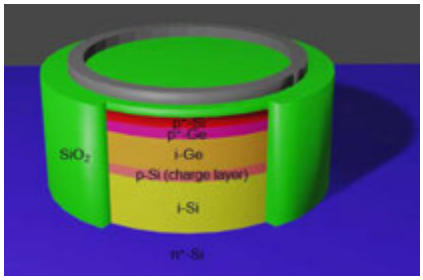


Figure 2: Schematic structure of a Ge-on-Si-APD (sample A)

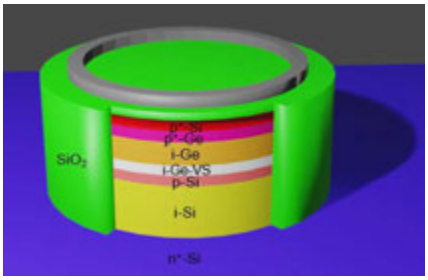


Figure 3: Schematic structure of a Ge-on-Si-APD with Ge-VS (sample B)

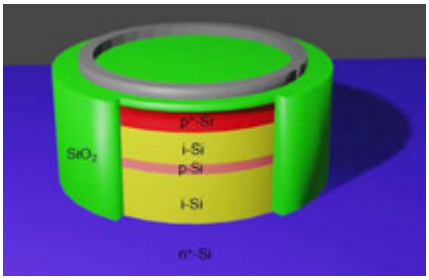


Figure 4: Schematic structure of a Si-APD (sample C)

II. FABRICATION

A. Molecular Beam Epitaxy

The presented structure for the Ge-on-Si-APDs were grown by using molecular beam epitaxy (MBE) using a solid source 6"-MBE system with a base pressure of $P < 10^{-10}$ mbar. The matrix material Si is evaporated by using an electron beam evaporator, whereby the matrix material Ge and the dopant materials Boron (B) and Antimony (Sb) are evaporated by Knudsen-cells, resp. For the Ge-on-Si-APDs we used an Arsenic(As)-doped n^+-Si(100) substrate with a nominal sheet resistance of $\rho < 0.005\ \Omega$ cm, which builds the bottom contact. The growth process starts with an in-situ thermal cleaning step at $T = 900$ °C for $t = 5$ min in order to remove the native Silicon dioxide (SiO_2) [9].

The growth starts with an intrinsic Si layer with a thickness of $d_{i\text{-}Si} = 500$ nm at a growth temperature of $T = 600$ °C. On top of this a p-doped Si layer with a thickness of $d_{p\text{-}Si} = 100$ nm and a nominal doping concentration of $N_A = 1.5 \cdot 10^{17}$ cm^{-3} is grown, which forms the charge layer. After that an intrinsic Ge layer with a thickness of $d_{i\text{-}Ge} = 300$ nm at the temperature of $T = 330$ °C is grown. On top of this a hetero contact layer, which consists of $d_{p+\text{-}Ge} = 100$ nm thick p^+-doped Ge and $d_{p+\text{-}Si} = 100$ nm thick p^+-doped Si is finally deposited. An illustration of this layer stack (sample A) can be seen in Fig. 2.

B. Sample variations

In order to prove the concept and give further information about the growth and fabrication process we processed two additional samples. Sample B was processed with a $d_{Ge\text{-}VS} = 100$ nm thick Ge virtual substrate (Ge-VS) deposited on the before grown p-Si charge layer and i-Si layer. The epitaxial growth of pure Ge on a Si substrate is technologically challenging because of the lattice mismatch of 4 % between Si and Ge. This leads to high threading dislocation densities and bad overall crystal structure if no further growth enhancements are performed. The Ge-VS reduces these threading dislocations [11] and therefore reduces the dark current. In order to create a Ge-VS, a 100 nm Ge layer has to be deposited with a temperature of $T = 330$ °C after which an annealing-step at $T_{anneal} = 810$ °C for $t = 5$ min is performed. Creating such a Ge-VS has the drawback, that the annealing-step leads to intermixing of Ge and Si at the interface. This increases the effective k-ratio of ionization coefficients of electrons and holes [12]. Afterwards another intrinsic layer with a thickness of $d_{i\text{-}Ge} = 200$ nm at the temperature of $T = 330$ °C is grown. Therefore, the Ge absorption region has the same thickness in both samples. On top of this the same hetero contact layer as in sample A gets deposited. The structure of this sample B can be seen in Fig. 3.

In addition to this sample we grow a reference APD made from pure Si (sample C). This sample has the equivalent layer structure and nominal doping concentrations as sample A and was processed in a similar way. The structure is illustrated in Fig. 4.

C. Device fabrication

Following the MBE growth, the 4" wafers were diced to 4 squares with an edge length of $l_{edge} = 35$ mm. A single mesa process was used to fabricate diodes on these samples. The first step in this process is a Ge cleaning. This consists of wet chemically clean with Acetone ($(CH_3)_2CO$) and Isopropyl alcohol ($CH_3CHOHCH_3$) following a plasma ashing with Oxygen (O_2) to remove (in)organic remains. To etch the mesa structure the photoresist AZMIR701 was structured with a 365 nm photolithography technology. After that the mesa etching was performed by Inductive Coupled Plasma (ICP) Reactive Ion Etching (RIE) with Hydrogen bromide (HBr) with an approximated etch rate of 120 nm/s.

The photoresist was removed with the remover P1316 and a O_2 plasma ashing process. Followed by the same wet chemically clean step as at the beginning. Afterwards, SiO_2 ($d_{SiO2} = 300$ nm) was deposited as a passivation and isolation oxide, using Plasma Enhanced Chemical Vapor Deposition (PECVD) with Tetraethyl orthosilicate (TEOS) as a precursor at $T_{sub} = 250$ °C.

In order to make a top contact an oxide window has to be etched, which is performed using RIE with Fluoroform (CHF_3) as etching gas with a subsequently wet etch step using buffered Hydrofluoric acid (bHF) for the last 30 nm of SiO_2 to avoid surface roughening by RIE [12]. The metallization consists of sputtered Aluminum (Al) with a thickness of $d_{Al} = 1300$ nm, which was etched using ICP-RIE dry etch step with HBr and a wet etching step with Phosphoric acid (H_3PO_4).

III. CHARACTERIZATION

A. DC measurements

In order to investigate the electrical behavior of the fabricated diodes, DC measurements, consisting of I(V) and transfer length method (TLM) measurements, were performed after processing at room temperature. These were accomplished on a Keithley 4200SC with a KarlSuss PA200 Manipulator.

Figure 5 shows the I(V) characteristic of the samples A to C. It can be seen that, overall current for sample A is higher than sample B. This proves the improved crystal quality because of the Ge-VS due to lower threading dislocation density. This advantage of sample B over sample A gets smaller for increasing negative applied voltages and thus making them nearly equal for the relevant voltage region. It can further be seen that the overall current of the pure-Si reference APD (sample C) is lower than both Ge samples. This is due to better overall crystal quality in Si than in Ge.

In order to compare the electrical properties, the series resistance R_S and ideality factor n_{id} are calculated using (1) and (2), respectively. These equations can be derived from the Shockley diode equation [13].

$$R_S = \frac{d}{dI}\left(I \cdot \frac{dV}{dI}\right) \tag{1}$$

$$n_{id} = V_T \cdot \left(\frac{d(\ln(I))}{dV_{id}}\right)^{-1} \tag{2}$$

Where I is the measured current, V is the applied voltage, V_T is the thermal voltage and $V_{id} = V - R_S I$ is the voltage applied to the diode, only. The calculated values of R_S are shown in Table 1. It can be seen that the series resistance increases in both sample A and B for decreasing diode radii. This is due to smaller volume of the mesa structure. For sample C this is not the case.

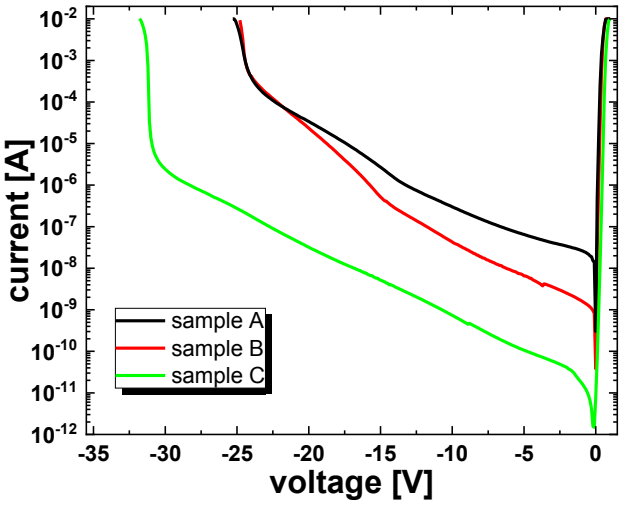


Figure 5: Current voltage characteristics at room temperature for all three samples with an adjusted voltage axis to show the forward region more clearly

TABLE I. SERIES RESISTANCES AS FUNCTION OF THE DEVICE RADIUS FOR ALL THREE SAMPLES

Diode radius	Series resistance of sample A	Series resistance of sample B	Series resistance of sample C
r_{diode} = 80 µm	19.8 Ω	11.9 Ω	14.3 Ω
r_{diode} = 40 µm	28.7 Ω	18.1 Ω	12.0 Ω
r_{diode} = 20 µm	65.9 Ω	30.9 Ω	12.3 Ω
r_{diode} = 10 µm	123.4Ω	77.2 Ω	17.5Ω

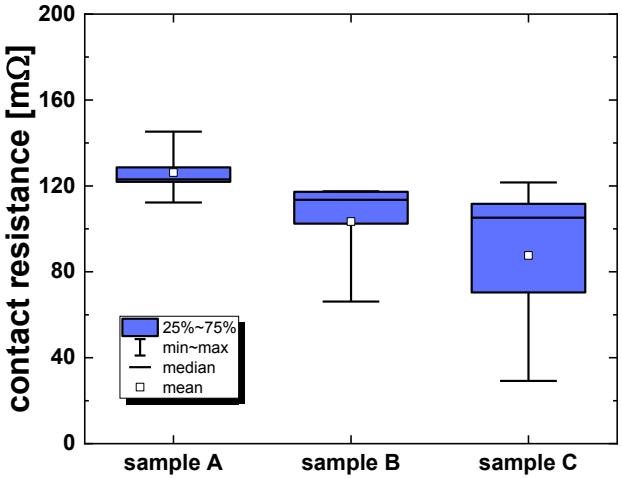


Figure 6: Box-and-whisker plot of the contact resistance of all three samples

The TLM measurements were only performed on the top contact, because of the one mesa process. The contact resistance is calculated using (3).

$$R_C = \frac{R_{TLM}(l=0\mu m)}{2} \tag{3}$$

R_{LTM} is the linear approximated resistance of the measured curve and l is the distance between two contacts. The so calculated contact resistance is shown in Fig. 6.

It can be seen that the overall contact resistance is small, which indicates a good contact of the Al with the Si. The median contact resistance of all three samples is nearly equal at around $R_C \approx 110$ mΩ, even though sample C has a larger spread.

B. Optical measurements

The optical parameters are the key parameters to prove quality of the Ge-on-Si APDs. The optical measurements were performed on a Keithley 2450 with both a continuous wave (CW) laser and a supercontinuum laser (model NKT photonics SuperK).

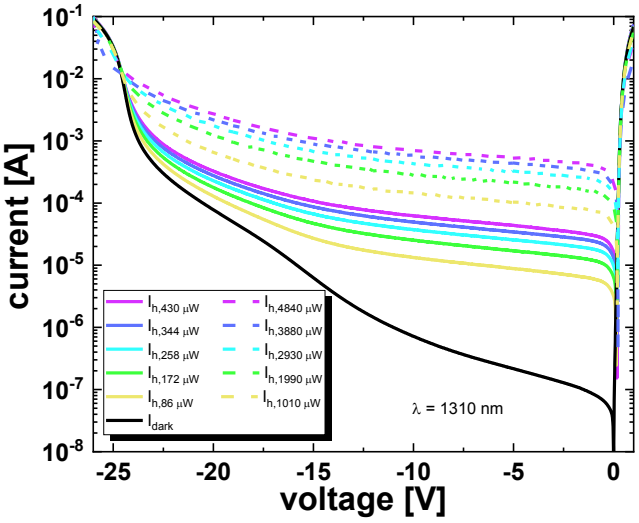


Figure 7: Illuminated and dark current of sample A at a wavelength of 1310 nm as function of different optical powers

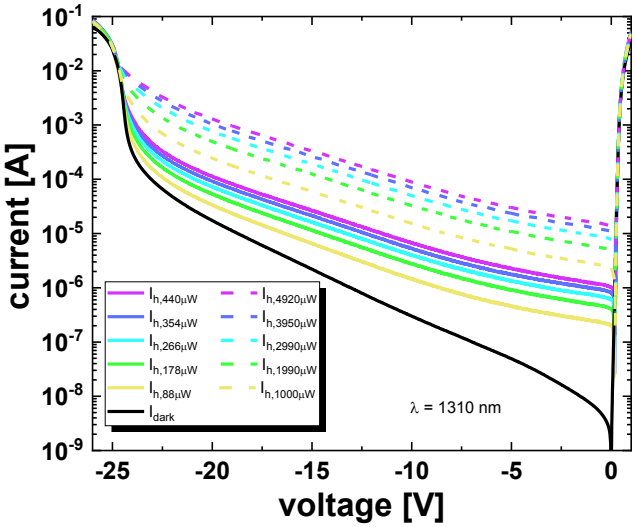


Figure 8: Illuminated and dark current of sample B at a wavelength of 1310 nm as function of different optical powers

In Fig. 7 the illuminated and dark current of the sample A is plotted. The measurement belongs to a diode with a mesa radius of r_{diode} = 80 µm illuminated with radiation of the wavelength λ = 1310 nm. Figure 8 shows the same for sample B.

It is seen that the illumination current and thus the photocurrent is proportional to the power of the illuminating laser.

The responsivity is defined as the optical power divided by the photocurrent as given in (5).

$$R_{opt} = \frac{dI_{phot}}{dP_{opt}} \tag{4}$$

Figure 9 shows that the photocurrent is not linear over the optical power. Due to this we have to calculate the differential quotient.

With the calculated responsivities we can calculate the corresponding gain factor when standardizing it to a comparable Ge-on-Si pin-diode. These pin-diodes achieve a responsivity of up to 0.22 A/W at a reverse bias of V_{bias} = - 0.1 V [10], which leads to a gain factor of up to 26 using (5).

$$G = \frac{R_{opt}}{R_{opt,ref}} = \frac{R_{opt}}{0.22\,\frac{A}{W}} \tag{5}$$

The so calculated gains for different bias voltages are shown in Fig. 10. It can be seen that the gain increases significantly for decreasing optical powers. This is because the Ge-on-Si-APD generates a significant photocurrent with even low optical powers. As stated before the photocurrent is not linear to the optical power. This is due to the fact that the samples are operated near at the avalanche breakdown and therefore there is already a significant avalanche multiplication.

Additionally, Fig. 11 shows the same for sample B. It can be seen that sample B only reaches a lower gain. Therefore, the Ge-VS decreases the optical properties due to increased k-ratio of ionization coefficients of electrons and holes which decreases multiplication factors.

IV. CONLUSION

We reported the successful growth and fabrication of Ge-on-Si APDs with responsivities of up to 6 A/W corresponding to a gain of up to factor 26. These Ge-on-Si APDs can be produced in a CMOS compatible process and therefore enable the integration in large scale detector arrays.

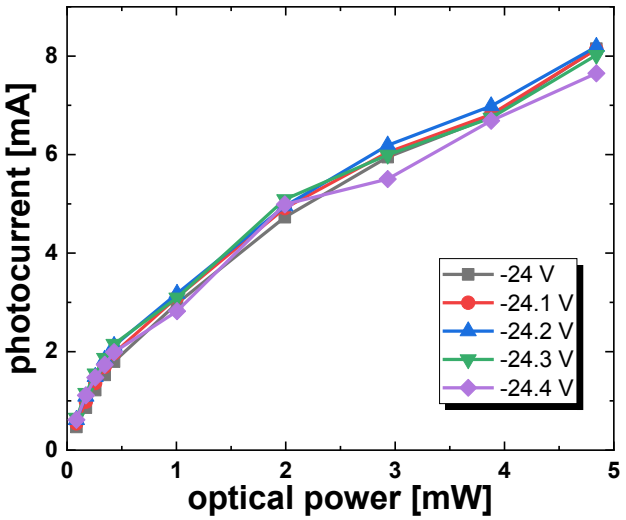


Figure 9: Photocurrent as function of the optical power a different working points from sample A

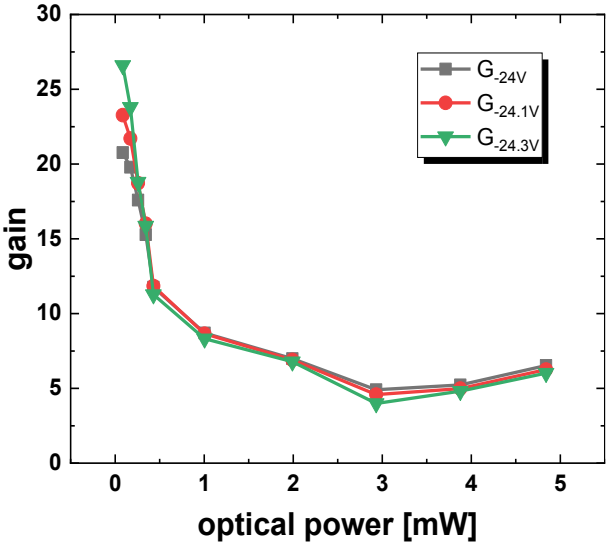


Figure 10: Gain of sample A reaching a gain of up to factor 26 at a low optical power of 86 µW and a wavelength of 1310 nm

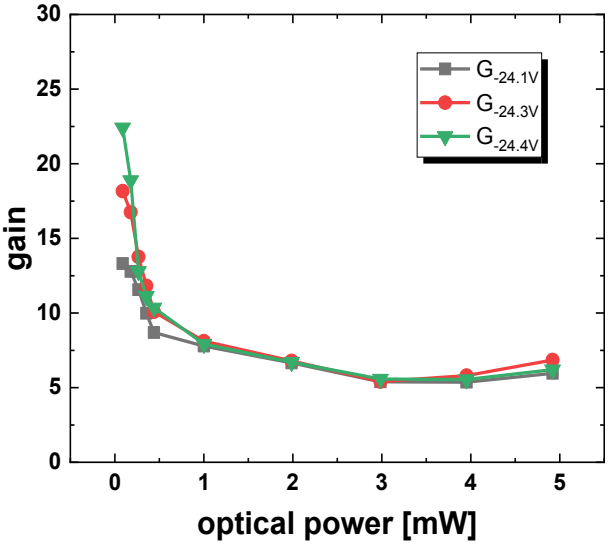


Figure 11: Gain of sample A reaching a gain of up to factor 22 at a low optical power of 88 µW and a wavelength of 1310 nm

REFERENCES

[1] G. S. Buller, R. J. Collins, "Single-photon detection and generation", Meas. Sci. Technol. Vol. 21, 012002, 2010.

[2] H. Takesue, S. W. Nam, Q. Zhang, R. H. Hadfield, T. honjo, K. Tamaki, Y. Yamamoto, "Quantum key distribution over a 40-dB channel loss using sperconducting single-photon detectors", Nat. Phot., Vol. 1, No. 6. pp. 343-348, 2007.

[3] R. E. Warburton, A. McCarthy, A. M. Wallace, S. Hernandez-Marin, R. H. Hadfield, S. W. Nam, G. S. Buller, "Subcentimeter depth resolution using a single –photon counting time-of-flight laser ranging system at 1550 nm wavelength", Opt. Lett., Vol. 32, No. 15, pp. 2266-2268, 2007.

[4] P. Yuan, R. Sudharsan, X. Bai, J. Boisvert, P. McDonald, E. Labois, M. S. Salisbury, G. M. Stuart, H. Danny, A. A. Prtillo, A. B. Roybal, S. Van Duyne, G. Pauls, S. Gaalema, "32x32 Geiger-mode LADAR cameras", Proc. SPIE Int. Soc. Opt. Eng., Vol.7684, p. 76840C, 2010.

[5] W. Becker, A. Bergmann, G. L. Biscotti, A. Rueck, "Advanced time-correlated single photon counting techniques for spectroscopy and imaging in biomedical systems", Proc. SPIE, Vol. 5340, pp. 104-112, 2004.

[6] F. Zappa, S. Tisa, A. Tosi, S. Cova, "Principles and features of single-photon avalanche diode arrays", Sci. Dir, Vol. 140, No. 1, pp. 103-112, 2007.

[7] D. Bronzi, F. Villa, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, W. Brockherde, "100 000Frames/s 64 x 32 single-photon detector array for 2-D imaging and 3-D ranging", IEEE JSTQE, Vol. 20, No. 6, 2014.

[8] W. C. Dash, R. Newman, "Intrinsic optical absorption in singlecrystal germanium and silicon at 77 °K and 300 °K", Phys. Rev., Vol. 99, No. 4, pp. 1151-1155, 1955.

[9] E. Kasper, M. Bauer, M. Oehme, "Quantitative secondary ion mass cleaning", Thin Solid Films, Vol. 321, pp. 148-152, 1998.

[10] M. Oehme, M. Schmid, M. Kaschel, M. Gollhofer, D. Widmann, "GeSn p-i-n detectors integrated on Si with up to 4% Sn", Appl. Phys. Lett. 101, 141110, 2012.

[11] P. Vines, K. Kuzmenko, J. Kirdoda, D. Dumas, M. Mirza, R. Millar, D. Paul, G. Buller, "High performance planar germanium-on-silicon single-photon avalanche diode detectors", Nat. Comm., Vol. 10, No. 1086, 2019.

[12] M. Oehme, J. Werner, M. Kaschel, O. Kirfel, E. Kasper, " Germanium waveguide photodetectors integrated on silicon with MBE", Thin Solid Films, Vol. 517, pp. 137-139, 2008.

[13] Y. Kang, H. Liu, M. Morse, M. J. Paniccia, M. Zadka, S. Litski, G. Sarid, A. pauchard, Y. Kuo, H. Chen, W. S. zaoui, J. E. Bowers, A. beling, D. C. McIntosh, X. Zheng, J. C. Campbell, "Monolithic germanium/silicon avalanche photodiodes with 340 GHz gain-bandwidth product", Nat. Phot., Vol. 3, No. 2, pp. 59-63, 2009.

[14] G. S. Oehrlein, R. G. Schad, M. A. Jaso, "Mechanism of silicon surface roughening by racive ion etching", Surface and Interface Analysis, Vol. 8, pp. 243-246, 1986.

[15] W. Shockley, "The theory of p-n junctions in semiconductors and p-n junction transistors", Bell System Technical Journal, Vol. 28, No. 3, pp. 435-489, 1949.

Processing sequence for a PureB bipolar junction transistor

A. Čaušević*, H. S. Funk, D. Schwarz, K. Guguieva, and J. Schulze *Senior Member IEEE*
Institute of Semiconductor Engineering (IHT), University of Stuttgart
Pfaffenwaldring 47, 70569 Stuttgart, Germany
*) Corresponding author: st149057@stud.uni-stuttgart.de

***Abstract* - Since we have previously shown that the deposition of pure Boron (PureB) on Silicon (Si) forms an almost ideal pn-junction, the next step is to utilize this technology towards a high-speed pnp bipolar junction transistor (BJT) made from Si. Here, the Emitter is made from a PureB. The PureB layer is extremely thin and features an almost defect-free pn-junction, a low off-current and a low series resistance, which makes it feasible to achieve better characteristics than the conventional BJT. However, PureB is very unreactive and therefore hard to etch, which results in a high roughness of the underlying semiconductor after etching. That is why it is challenging to build a three-terminal device such as a BJT. In this work, we present a novel process scheme using a double-layer hardmask of Silicon dioxide (SiO_2) and Aluminum oxide (Al_2O_3) to enable differential epitaxy of PureB. Onto the substrate a layer stack of SiO_2 and Al_2O_3 was deposited and structured, the Al_2O_3 was then recrystallized to be used as an etch stop. PureB was deposited subsequently by differential epitaxy into the SiO_2 windows.**

Keywords – PureB, bipolar junction transistor, molecular beam epitaxy, physical vapor deposition

I. Introduction

Since a layer of pure Boron (PureB) on Silicon (Si) produces a nanometer-shallow p^+n-junction with a low leakage current [1], [2] and was already used in an insulated-gate bipolar transistor (IGBT) [3] and ultraviolet (UV) detectors [4]. The idea is to utilize a PureB layer as an Emitter for a high-speed bipolar junction transistor (BJT). In an optimally designed pnp BJT the Emitter needs to be highly doped and the thickness of the Base should be less than the diffusion length of holes at a specific Base doping concentration [5]. The ratio between the Emitter doping and the Base doping determines the Common-Base Current Gain of a BJT. A higher ratio yields a higher gain. However, the Base cannot be arbitrarily lightly doped, as the Collector needs to be doped even lower than the Base [5]. The best option to improve the gain of a BJT is therefore to increase the doping of the Emitter, e.g. PureB can be used to act as a highly doped emitter.

Boron (B) is an inert material [6] and therefore difficult to etch. B layers exhibit considerable roughness after etching as it is irregularly etched. In case if an extremely thin B layer should be etched, this roughness will be transferred onto the surface of a softer underlying substrate [7]. By irregular etching of the B, some parts of the underlying material are revealed sooner and faster than the B covered parts which leads to a high roughness. Since Si is etched faster than B, this material combination suffers from this problem.

Based on these observations, it is very difficult to etch through B and to contact a thin Base layer of the vertical BJT. The roughness after etching can reach deeper than the Base and contact the Collector instead or short the Base and Collector. Furthermore, the contact quality can suffer as an Aluminum (Al) layer has to be deposited on this rough surface.

Facing these problems, we decided to develop a new process sequence to achieve a precise and smooth contact of the transistor Base after B etching.

II. Processing Sequence

The process sequence is divided into four steps. The epitaxy of the transistor Base, the preparation of the sample for B deposition, the B deposition itself, and further processing for the contact formation. The final structure of the BJT is depicted in Fig. 1.

A. Molecular Beam Epitaxy

For the molecular beam epitaxy (MBE) of the Base, a 6” MBE system was used, where Si acts as matrix material and B and Antimony (Sb) as n-type and p-type dopants, respectively. While Si is evaporated using an e-beam evaporator, Knudsen effusion cells are used for B and Sb. For Si, the growth rate is controlled at R = 1 Å·s^{-1} using a quadrupole mass spectrometer. The transistor is grown on B-doped 4” Si (001) substrates with a specific resistance of 10 Ω·cm ≤ ρ ≤ 20 Ω·cm. A 6” Si adapter ring is used to fit the 4” wafer into the 6” MBE system. The growth process starts with thermal desorption step at a substrate

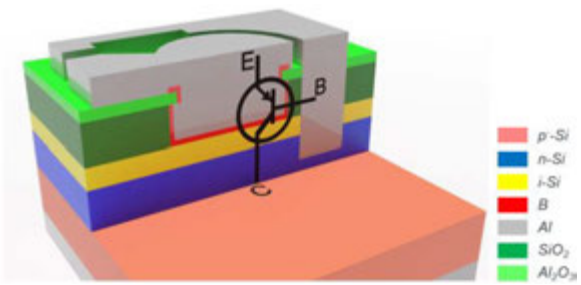


Figure 1. Schematic view of the processed BJT.

temperature $T_S = 900$ °C to remove the native Silicon dioxide (SiO_2) [8]. Afterwards, an Sb pre-deposition of $c_{Sb} = 1.4 \cdot 10^{14}$ cm^{-2} is performed at $T_S = 330$ °C to counter-balance the surface segregation of Sb in Si [9], followed by 500 nm of Sb-doped Si. To consume the segregated Sb, a 30 nm intrinsic Si layer was grown on top without substrate heating. This concludes the epitaxy of the Base. The Base is doped with an estimated doping concentration N_B of $1 \cdot 10^{18}$ cm^{-3} and 500 nm thick, which is determined by the known growth rate. The resulting layer stack is illustrated in Fig. 2.

B. Preparation for Boron Deposition

After this, a 300 nm SiO_2 layer is deposited by plasma-enhanced chemical vapor deposition (PECVD) at T = 250 °C with Tetraethoxysilane (TEOS) and Oxygen (O_2) as precursors. The samples are cleaned for 8 min in Acetone, 3 min in Isopropyl alcohol, rinsed in deionized water. A last O_2 plasma step completes the cleaning. Next, a 15 nm thick Aluminum oxide (Al_2O_3) layer is deposited on top of the SiO_2 layer by atomic layer deposition (ALD) with Trimethyl aluminum (TMA) and O_2 as precursors.

These two oxides are used as a hardmask as well as a passivation. First, oxide windows for the Emitter and Base contacts are wet chemically etched into the Al_2O_3 using an Aluminium etch solution ($H_3PO_4 + CH_3COOH + HNO_3$). Afterwards, the sample is annealed for 5 min at T = 950 °C under an O_2 atmosphere to recrystallize the Al_2O_3 [8] and to oxidize any remaining Al. This recrystallized Al_2O_3 can now act as hardmask. To enable the B deposition to form the Emitter contact first an oxide window is etched into the SiO_2 with reactive-ion etching (RIE) using Fluoroform (CHF_3). During this RIE step is the SiO_2 windows for the emitter contact etched, the base contact is masked. RIE is stopped approximately 50 nm before reaching the substrate. The residual oxide is then removed with buffered Hydrofluoric acid (BHF) to achieve a smooth Si surface for the consecutive epitaxy step. The structure of the sample up to this point is illustrated in Fig. 3.

C. Boron Deposition

B deposition is performed in the same MBE system. After a 10 s HF dip to remove native SiO_2 the sample is directly placed in the MBE-System. The Hydrogen (H_2)-passivation is removed by thermal desorption at $T_S = 700$ °C. B is deposited at a growth temperature of T = 630 °C and a flux rate of $F_B = 1 \cdot 10^{12}$ cm·s^{-1} which leads to a growth rate of approx. 0.14 Å·min^{-1}. The final PureB layer is 3-5 nm thick and is deposited on the whole substrate as seen in Fig. 4.

D. Contacts Formation

A 100 nm Al layer is sputtered immediately after deposition to protect the B layer from environmental influence. The Al layer mainly protects the B layer from physical damage from handling, since B is inert and not affected by atmospheric oxygen or other gasses.

Al and B are simultaneously structured by ICP-RIE using a gas mixture of Hydrogen bromide (HBr) and Helium (He) as etching gases. The etching process is stopped after reaching SiO_2. SiO_2 is again etched by RIE using CHF_3 until 100 nm of it remains. Some roughness in the underlying SiO_2 layer can remain, as seen in Fig.5, and is removed by BHF. In this way, the roughness of the SiO_2 layer caused by B etching is eliminated and the underlying Si maintains a smooth surface.

The intrinsic Si is etched with RIE using CHF_3 to reach the n-type Si region and to contact the Base. Afterwards, 800 nm of Al are sputtered on top and 2 µm on the bottom side of the sample to form the contacts. After structuring of Base and Emitter contacts with ICP-RIE using HBr. The unstructured Al on the backside forms the Collector contact. The final structure of the PureB BJT can be seen again in Fig. 1.

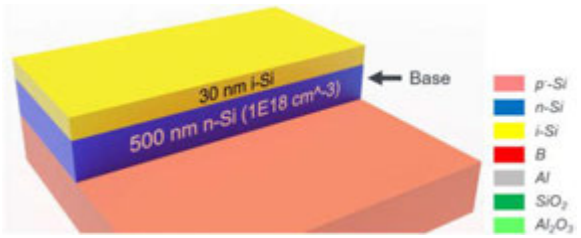


Figure 2. The layer stack after MBE.

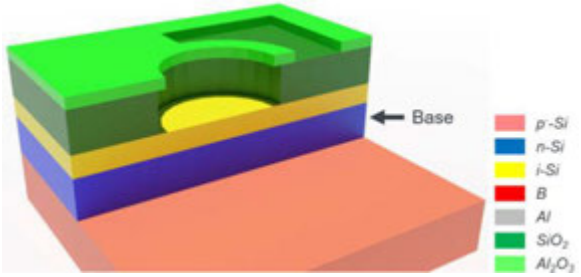


Figure 3. Schematic view of the BJT before B deposition.

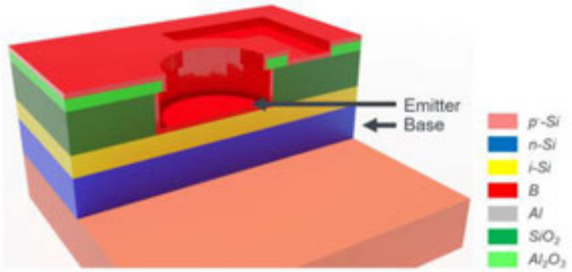


Figure 4. Schematic view of the BJT after B deposition.

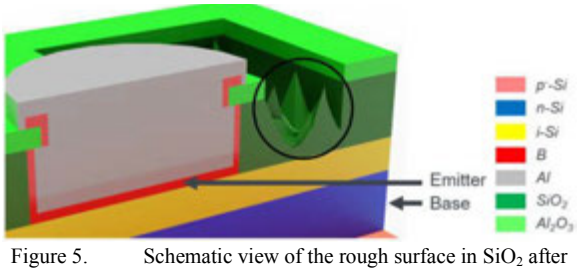


Figure 5. Schematic view of the rough surface in SiO_2 after B etching.

III. RESULTS

Analysis of the BJT cross-section was performed with a scanning electron microscope (SEM). Furthermore, the BJT was characterized electrically.

A. Observations during Processing

BHF was able to etch Al_2O_3 by creeping under the photoresist on top of it. This resulted in a frayed border of the circular window and etched away the isolation between two oxide windows, as seen in Fig. 6. Al_2O_3 becomes harder to structure once it is crystalline [11]. It is therefore annealed before etching with BHF takes place. This is also the reason why both oxide windows, one for the Emitter and one for the Base, need to be opened before annealing.

B. SEM Imaging

Fig. 7 shows a SEM image of a PureB BJT. The Base and Emitter contacts are marked in the image as well as the SiO_2 layer, the Collector contact is not visible. The interface between the Si and the Al on the right side appears smooth, confirming the effectiveness of the SiO_2 as a buffer for the B etching induced roughness.

For the fabrication of the BJTs an available diode mask was used and therefore the processing sequence was adapted to this mask. The gap in Si between emitter and base is created simultaneously with the opening of the oxide windows for contacting the base. However, this gap is not contacted by Al and does not influence the performance of the BJT.

The SiO_2 layer between both top contacts also appears smooth, it is covered by an Al_2O_3 layer and isolates both top contacts. The annealed Al_2O_3 layer was able to shield the SiO_2 from major damages. Al_2O_3 and SiO_2 were etched moderately but not significantly enough to hamper processing.

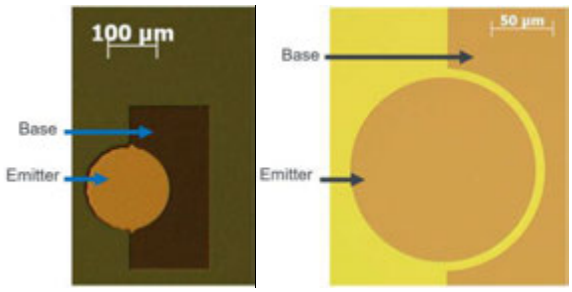


Figure 6. The oxide windows after etching with BHF. Left without annealing, right with annealing. The edge on the left one is frayed, BHF etched Al_2O_3 between both oxide windows. The right one maintained a sharp edge.

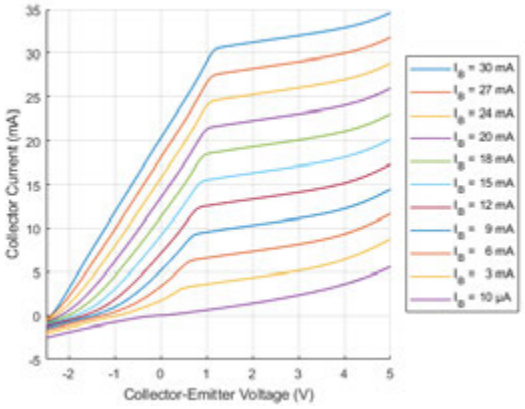


Figure 8. Output characteristics of a PureB BJT with an Emitter surface of $A = \pi \cdot (40\mu m)^2 = 5027\ \mu m^2$ for different Base currents (I_B)

C. Electrical Characterization

The output characteristic of the PureB BJT depicted in Fig. 8 shows BJT like behavior. It has a Common-Emitter Current gain β ranging from

$$\beta = I_C/I_B \approx 32\ mA/30\ mA \approx 1.067 \quad (1)$$

at $V_{CE} = 3$ V to

$$\beta = I_C/I_B \approx 5\ mA/3\ mA \approx 1.667 \quad (2)$$

at $V_{CE} = 3$ V. The gain suffers from the 500 nm thick Base region and the usage of the substrate as the Collector region in this BJT.

Furthermore Fig. 8 shows some leakage which may be a result of the parasitic Schottky diode between the emitter and base contact. Additionally, the collector current does not go to zero for $V_{CE} = 0$ V, as the base current uses the Schottky diode to flow to the emitter. Further refinement of the design can improve the characteristics.

IV. CONCLUSION

A new processing sequence for a PureB BJT was developed. Al_2O_3 can resist BHF etching, if annealed, to keep sharp edges on oxide windows and can resist damages from B etching to protect the underlying material. The drawback is that the Al_2O_3 needs to be completely structured beforehand as it becomes harder later on. By

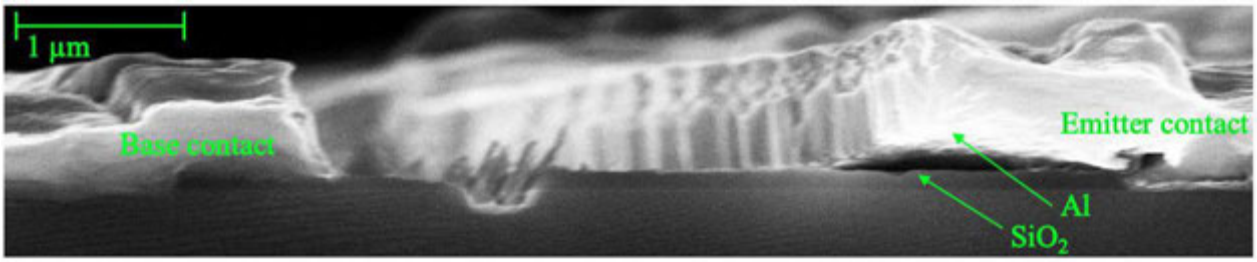


Figure 7. SEM image of the processed PureB BJT

using SiO_2 and Al_2O_3 as double-layer hardmask the underlying substrate can be protected from roughness caused by B etching. It allows for surface smoothness and contact quality. Furthermore, analysis confirms the effectiveness of the processing sequence. SEM images show no signs of roughness in Si and electrical characterization proves the functionality of the PureB BJT and shows a current gain. The low gain may result from the thick base and collector layer.

This proof of concept PureB BJT has demonstrated that fabrication and functionality of these devices is possible. Further improvements can be made by adaption of this processing sequence for different BJT designs.

REFERENCES

[1] V. Mohammadi, L. Shi, U. Kroth, C. Laubis, and S. Nihtianov, "Stability characterization of high-performance PureB Si-photodiodes under aggressive cleaning treatments in industrial applications," in *2015 IEEE International Conference on Industrial Technology (ICIT)*, 2015, pp. 3370–3376

[2] A. Elsayed and J. Schulze, "Characterization of thin Boron layers grown on Silicon utilizing molecular beam epitaxy for ultra-shallow pn-junctions," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 0007–0011

[3] A. Elsayed, I. A. Fischer, Y. Elogail, and J. Schulze, "Novel IGBT Devices Utilizing Ultra-Thin Amorphous Boron Injection Layers."

[4] L. K. Nanver *et al.*, "Robust UV/VUV/EUV PureB Photodiode Detector Technology With High CMOS Compatibility," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 6, pp. 306–316, Nov. 2014

[5] B. El-Kareh and L. N. Hutter, *Silicon Analog Components*, Second Edition. New York, NY: Springer, 2020.

[6] A. W. Laubengayer, D. T. Hurd, A. E. Newkirk, and J. L. Hoard, "Boron. I. Preparation and Properties of Pure Crystalline Boron," *J. Am. Chem. Soc.*, vol. 65, no. 10, pp. 1924–1931, Oct. 1943

[7] N. LiCausi, J. Clinton, Y. Danon, J. J.-Q. Lu, and I. B. Bhat, "Deposition and Etching of Conformal Boron Films for Neutron Detector Applications," *MRS Proc.*, vol. 1307, pp. mrsf10-1307-cc07-02, 2011

[8] E. Kasper, M. Bauer, and M. Oehme, "Quantitative secondary ion mass spectrometry analysis of SiO2 desorption during in situ heat cleaning," *Thin Solid Films*, vol. 321, no. 1, pp. 148–152, May 1998

[9] D. V. Yurasov, M. N. Drozdov, A. V. Murel, M. V. Shaleev, N. D. Zakharov, and A. V. Novikov, "Usage of antimony segregation for selective doping of Si in molecular beam epitaxy," *Journal of Applied Physics*, vol. 109, no. 11, p. 113533, Jun. 2011

[10] S. Jakschik, U. Schroeder, T. Hecht, M. Gutsche, H. Seidl, and J. W. Bartha, "Crystallization behavior of thin ALD-Al2O3 films," *Thin Solid Films*, vol. 425, no. 1, pp. 216–220, Feb. 2003

[11] C. J. McHargue, J. D. Hunn, D. L. Joslin, E. Alves, M. F. da Silva, and J. C. Soares, "Etching of amorphous Al2O3 produced by ion implantation," *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, vol. 127–128, pp. 596–598, May 1997

Carrier mobilities in heavily doped pseudomorphic $Ge_{1-x}Sn_x$-epilayers

M. M. Dettling*, D. Weißhaupt*, H. S. Funk*, M. Kern**, F. Berkmann*, C. Clausen*, M. Oehme*, D. Schwarz*, J. v. Slageren** and J. Schulze* *Senior Member IEEE*

* Institute of Semiconductor Engineering (IHT), University of Stuttgart
Pfaffenwaldring 47, 70569 Stuttgart, Germany
** Institute of Physical Chemistry (IPC), University of Stuttgart
Pfaffenwaldring 55, 70569 Stuttgart, Germany
Corresponding author: st148321@stud.uni-stuttgart.de

***Abstract* - This work reports on the carrier mobilities in heavily doped, pseudomorphically grown $Ge_{1-x}Sn_x$-epilayers with Sn-concentrations up to 9.2 %. For this purpose, $Ge_{1-x}Sn_x$-alloys were grown on a relaxed Ge virtual substrate on top of a commercial Si (100) wafer using molecular beam epitaxy. The crystal structure and quality of the $Ge_{1-x}Sn_x$-epilayers were analyzed by high-resolution X-ray diffraction. In order to extract the carrier mobilities in $Ge_{1-x}Sn_x$, low temperature hall measurements were carried out, using a Van-der-Pauw-geometry. It is shown that with increasing Sn-concentration we find a decrease in carrier mobility, which corresponds to an increasing sheet-resistance.**

Keywords - germanium; tin; low temperature hall-mobility

I. Introduction

Germanium-Tin ($Ge_{1-x}Sn_x$) is a semiconductor alloy of high interest for future generations of semiconductor devices. The direct zero band-gap (E_g = -0.4 eV) [1] and high charge carrier mobility – attributes of Sn [2] – allow for band-gap engineering [3] and opens the possibilities for a new generation of infrared detectors [4] and complementary metal-oxide-semiconductor (CMOS) devices [5] in a Silicon (Si) compatible technology [6]. For low doping concentrations Sn has already been shown to increase charge carrier mobilities in devices made from $Ge_{1-x}Sn_x$ [7, 8]. However, with generational downsizing of semiconductor devices and the accompanied increase of doping concentrations it is prudent to investigate mobility behaviors in heavily doped $Ge_{1-x}Sn_x$ layers. Recent work has shown that a decrease in mobility of holes and electrons may be expected for high doping concentrations in this material composition [9]. This work sets out to further investigate this behavior for a broad range of $Ge_{1-x}Sn_x$-compositions. For this purpose, $Ge_{1-x}Sn_x$-epilayers with Sn-contents of up to 9.2 % and nominal doping concentrations of 1e20 cm^{-3} have been grown using molecular beam epitaxy (MBE), ensuring high crystal quality in the samples. The samples with 0 % Sn act as references. Van-der-Pauw-structures (VdP-structures) where produced using a low thermal budget process to reduce layer degradation during processing. Low temperature hall measurements were performed on the samples to provide a large database for comparison of the material compositions.

II. Experimental Details

A. Low temperature MBE growth

As a substrate for the layer stack, an industry standard Si (001) wafer was used. An initial thermal cleaning step at 900 °C was performed to remove the native oxide from the surface of the wafer [10]. An intrinsic layer (100 nm) of Ge was grown (330 °C) and thermally relaxed (810 °C) to provide a virtual Ge substrate (Ge-VS). Further 200 nm of intrinsic Ge were then grown (330 °C) as a buffer zone. On this buffer the active layer (75 nm) of n(p)-type doped ($N_{D(A)} \sim$ 1e20 cm^{-3}) $Ge_{1-x}Sn_x$ was grown (160 °C). Doping in the layers was realized by using Antimony (Sb) and Boron (B) for n-type and p-type doping, resp. For Sb-doping pre-deposition of pure doping material was performed before the growth of this layer to compensate segregation losses and to achieve an abrupt doping concentration profile [11]. The active and buffer layers suffer from background p-type doping (~1e16 cm^{-3}) originating from the Boron nitride (BN) crucible used for Ge evaporation in the MBE system. The Si wafer also contains a small doping concentration of the opposite type from the active layer, which corresponds to a resistivity of ($10\ \Omega cm \leq \rho \leq 20\ \Omega cm$). Nominal Sn concentrations range from 0 % to 8 % in 2 % increments. Fig. 1 shows a schematic depiction of the layer stack with a n-type active layer and a p-type Si (001) wafer.

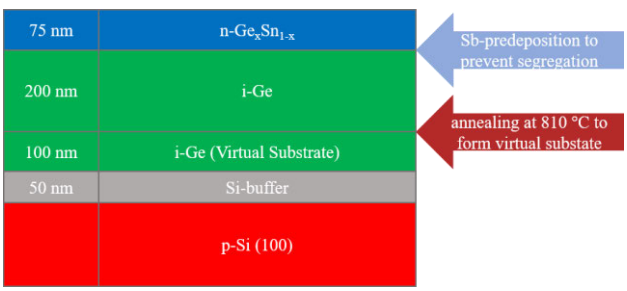


Figure 1. Layer stack for the MBE-growth of n-type active $Ge_{1-x}Sn_x$ (for p-type active $Ge_{1-x}Sn_x$ a n-type Si-substrate is used)

B. *Cleanroom Process*

The process used for fabricating VdP-structures employs standard cleanroom techniques. Patterning was realized using optical contact lithography.

Plasma enhanced chemical vapor deposition (PE-CVD) was used to deposit a SiO_2 layer on the samples for better photoresist adhesion. Tetraethyl orthosilicate (TEOS) and Oxygen (O_2) were used as precursors. After structuring the SiO_2 mask using reactive ion etching (RIE) with Fluoroform (CHF_3) and a short buffered Hydrogen fluoride (BHF) dip a mesa structure (3,5 mm x 3,5 mm) was etched using inductively coupled plasma etching (ICP-RIE) with Hydrogen bromide (HBr) as etching gas. Afterwards the SiO_2 was removed using BHF. Again, a layer of SiO_2 (150 nm) was deposited to passivate the mesa surface using PECVD. After opening contact holes using RIE, Aluminum (Al) contact pads were formed by sputtering followed by structuring of the layer via ICP-RIE.

III. SAMPLE ANALYSIS

A. *X-ray diffraction analysis*

On all samples high resolution X-ray diffraction (HR-XRD) measurements were performed in order to guarantee good crystal quality and well-defined sample conditions going forward. Fig. 2 shows the high resolution 2θ-ω-scan (rocking curve) of the symmetric (004) reflex and Fig. 3 shows the reciprocal space map (RSM) of the asymmetric (-2-24) reflex resulting from XRD measurements on the n-type $Ge_{0.977}Sn_{0.023}$ sample. These results stand exemplary for all samples as measurements have shown similar results. Furthermore, Sn concentrations can be extracted from the RSM with Vegards law as performed in [12] and are shown in Tab. 1. These Sn concentrations exceed the maximum solubility of Sn in Ge which is around 1% [13].

Fig 2. shows the 2θ-ω-scan of n-type $Ge_{0.977}Sn_{0.023}$ in the symmetric (004) reflex. It shows distinct peaks that can be related to the $Ge_{0.977}Sn_{0.023}$-epilayer, the Ge-buffer and the (004) reflex from the Si wafer as indicated in the figure. The peak marked with Ge-VS can be explained by the intermixing occurring during relaxation of the initial Ge-layer due to the annealing. The Ge-VS is relaxed on the Si (001) wafer. The RSM in Fig. 3 shows that the $Ge_{0.977}Sn_{0.023}$-epilayer is strained to match the lattice of Ge in the x-y-plane. Since they have the same in-plane lattice constant the $Ge_{0.977}Sn_{0.023}$ is therefore pseudomorphically grown. The sharp peaks for all layers in Fig.s 2 and 3 indicate high crystal quality in the samples.

We observe that Sn-concentrations appear shifted from their nominal values. This may be due to the large concentration of doping material in the crystal where B and Sb differ greatly in size from Ge.

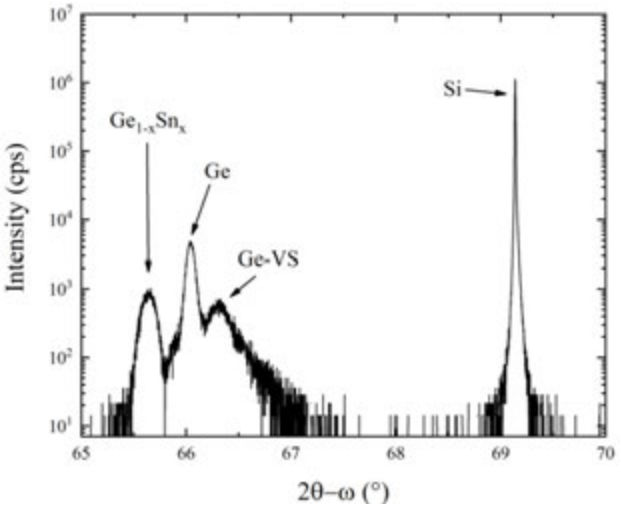


Figure 2. HR-XRD 2θ-ω-scan of the symmetric (004) reflex of the n-$Ge_{0.977}Sn_{0.023}$ sample

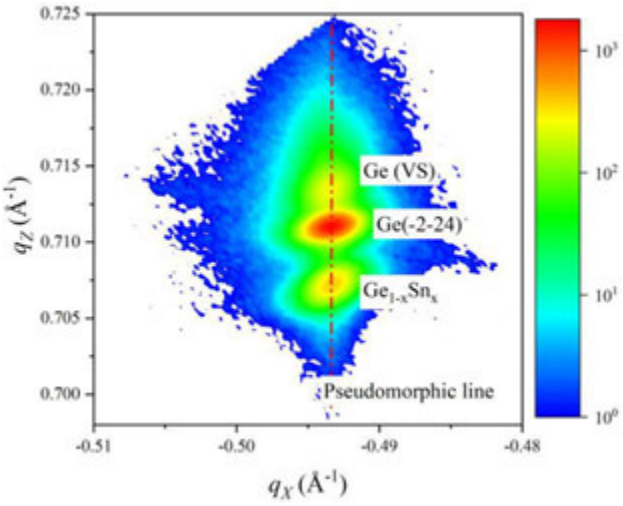


Figure 3. RSM of the asymmetric (-2-24) reflex of the n-$Ge_{0.977}Sn_{0.023}$ sample

TABLE I. SAMPLE NAMES WITH NOMINAL AND FROM XRD MEASUREMENTS DETERMINED SN CONCENTRATIONS

Sample names	Nominal Sn concentration	Sn concentration from XRD
n0	0 %	0 %
n2	2 %	2.3 %
n4	4 %	4.4 %
n6	6 %	6.5 %
n8	8 %	9.2 %
p0	0 %	0 %
p2	2 %	0.7 %
p4	4 %	3.0 %
p6	6 %	5.5 %
p8	8 %	7.2

B. *Layer analysis*

For a correct interpretation of hall measurements taken from the samples it is necessary to consider an electric resistor model of the layer stack beforehand. Fig. 4 illustrates a simple resistor model of all layers including indications for resistivity behavior with temperature. It is necessary to consider all three layers since threading dislocations originating from the Si-Ge-interface are expected to short out the Ge- and $Ge_{1-x}Sn_x$-epilayers. For the n-type $Ge_{1-x}Sn_x$ (Fig. 4.b) this also shorts the pn-junction that is supposed to insulate the Si (001) wafer from the active layer. As a result, three different parallel layers contribute to the resistance at room temperature, as indicated by the three parallel resistors. In case of the p-type active layer (Fig. 4.a) the pn-junction is intact since the threading dislocations originate at the Si-Ge interface and therefore cannot short the depletion region completely. It follows that only the active $Ge_{1-x}Sn_x$-epilayer and the Ge-substrate are electrically connected as indicated by the two resistors representing the layers. Since the active layer is degenerately doped its resistance behavior with temperature is metallic for both n- and p-type $Ge_{1-x}Sn_x$. The wafer substrate and the buffer regions show semiconductor behavior. For low temperatures, we can expect charge carriers in the substrate and buffer layers to freeze out and to be left with electrons and holes in the degenerately doped active $Ge_{1-x}Sn_x$-epilayers. This allows for simplification of the model proposed in Fig. 4. Consequently, low temperature measurements provide property values for the active layer without further need of complicated multilayer analysis.

C. *Hall analysis*

Hall measurements were performed on VdP-structures of all samples at temperatures ranging from 10 K to 300 K. The induced current was set to 100 µA. The applied magnetic field was set at 0.6 T. Fig.s 5 and 6 show carrier concentrations for n- and p-type $Ge_{1-x}Sn_x$. The graphs clearly show the freezing out process of charge carriers in the substrate and buffer materials outlined in section B starting at a temperature of 200 K. Below 100 K carrier concentration is approximately constant. For all temperatures below this point it is assumed that only the $Ge_{1-x}Sn_x$-epilayer is still electrically active. From this data we can conclude that doping concentration in all n-type samples is approximately 1e20 cm^{-3} and in all p-type samples 2e20 cm^{-3}.

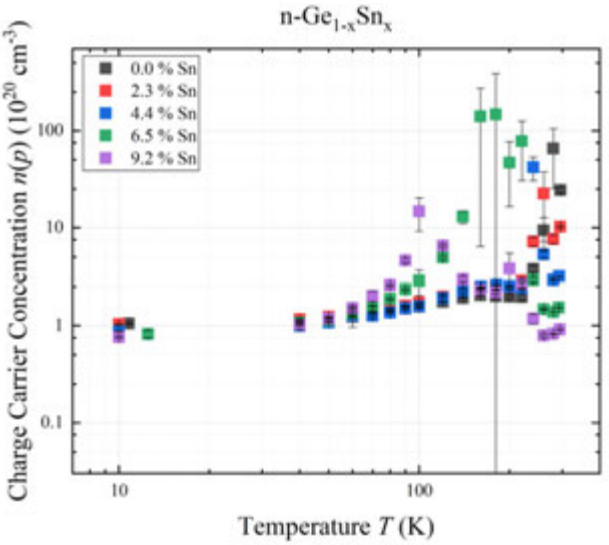


Figure 5. Charge carrier concentrations in all n-type samples resulting from Hall measurements for a temperature range from 10 K to 300 K

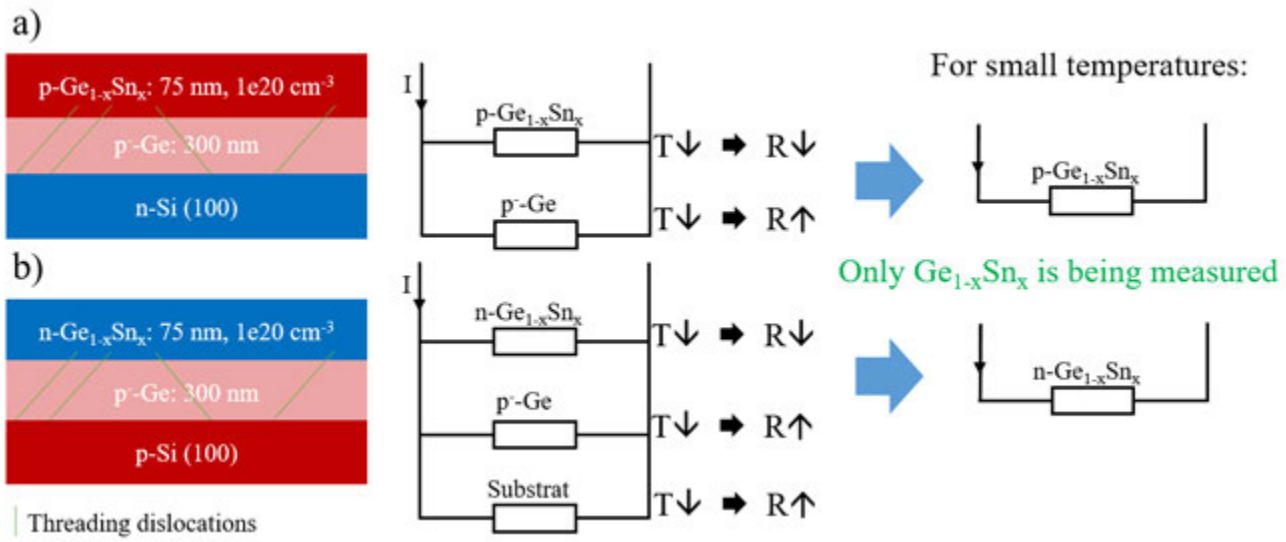


Figure 4. Resistor model of layer stack with indications for resistance behavior with temperature and simplifications resulting from this

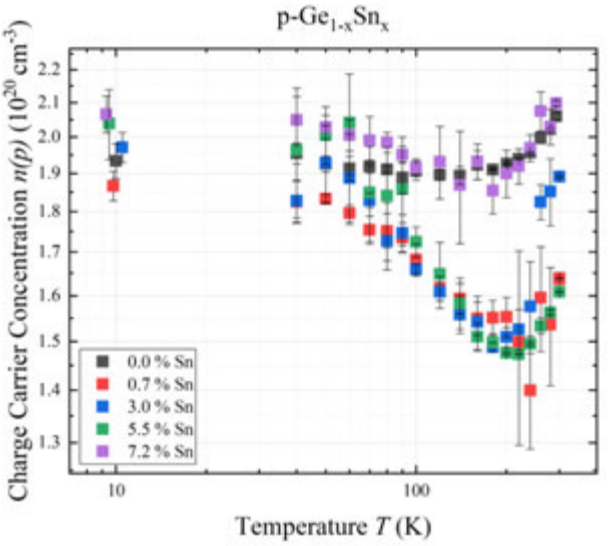


Figure 6. Charge carrier concentrations in all p-type samples resulting from Hall measurements for a temperature range from 10 K to 300 K

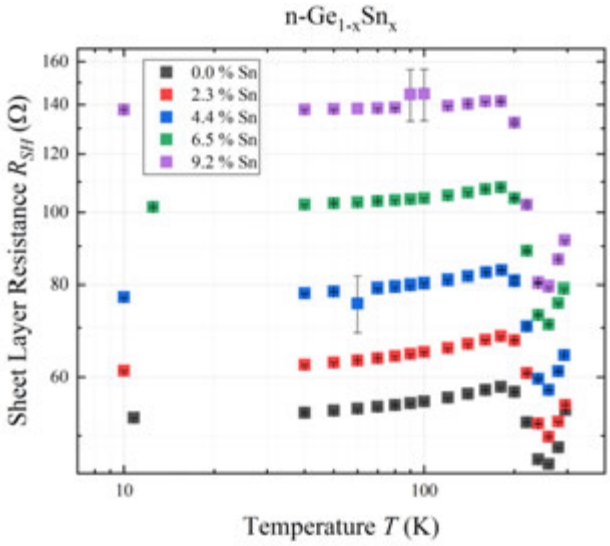


Figure 7. Sheet layer resistances in all n-type samples resulting from Hall measurements for a temperature range from 10 K to 300 K

Fig.s 7 and 8 show the sheet resistance measured at the same temperatures. The disorder of the curves above 200 K is due to the superposition of the layer stack as mentioned before. From these graphs, it can be shown that sheet resistance tends to increase with Sn concentration for both n- and p-type samples.

$$\mu = 1 / (q \cdot R_{SH} \cdot n(p) \cdot d) \qquad (1)$$

Fig.s 9 and 10 show the resulting carrier mobilities in the samples. The values were calculated using (1), where μ is the charge carrier mobility, q the electron charge, R_{SH} the sheet resistance, $n(p)$ the charge carrier concentration and d the sheet thickness, in combination with the data from Fig.s 5, 6, 7, and 8. The mobility curves show a clear negative correlation between Sn-concentration and carrier mobility. One notable exception is sample p8 which shows a lower sheet resistance as well as higher carrier mobility than p6. This stands in contrast to all other samples. Absolute values for charge carrier mobility in the Ge reference samples is comparable to other research [14].

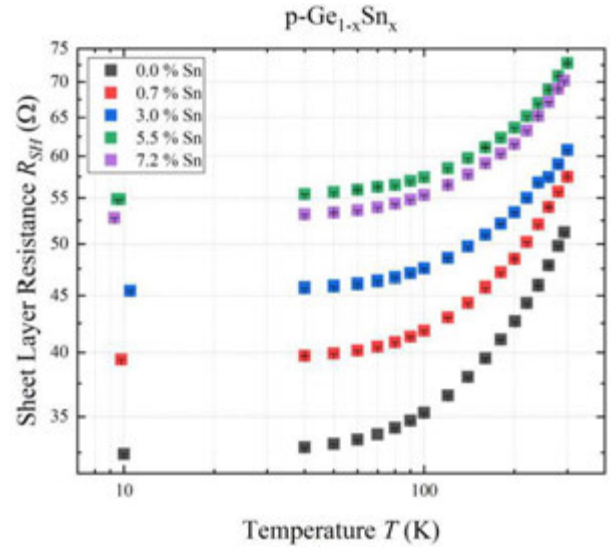


Figure 8. Sheet layer resistances in all p-type samples resulting from Hall measurements for a temperature range from 10 K to 300 K

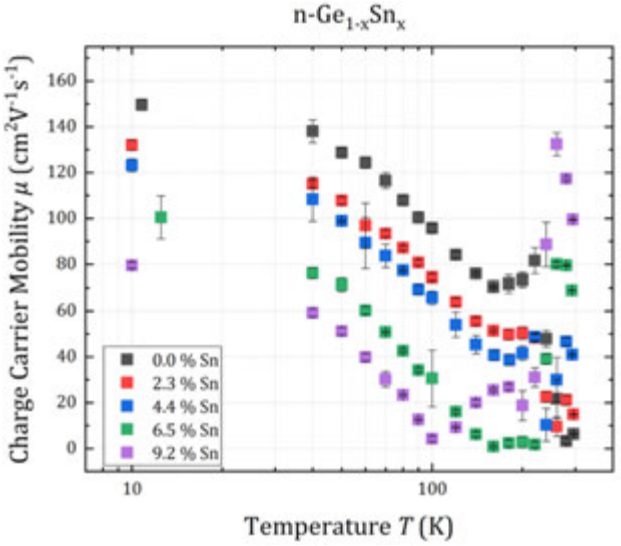


Figure 9. Charge carrier mobilities in all n-type samples resulting from Hall measurements for a temperature range from 10 K to 300 K

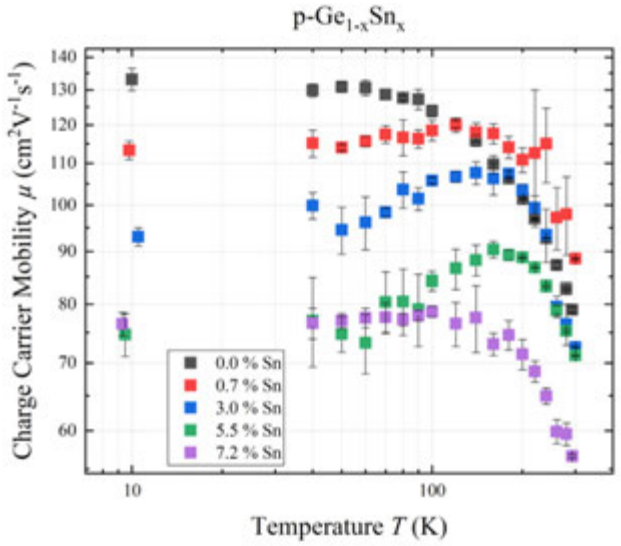


Figure 10. Charge carrier mobilities in all p-type samples resulting from Hall measurements for a temperature range from 10 K to 300 K

D. Data interpretation

The data resulting from the hall measurements clearly show a decrease in mobility with increasing Sn-content supporting existing work on this issue [9]. In contrast to these results, other publications have found an increase in mobility in lightly doped $Ge_{1-x}Sn_x$-epilayers [7, 15]. Since doping concentration is similar across all our samples and XRD-measurements show good crystal quality, coulomb- and impurity-scattering cannot explain this trend. It may be considered, that an increase of alloy scattering due to higher Sn concentrations can lead to the observed decrease of charge carrier mobility in the samples. Further investigation on this theory would clarify this point. Mobility in the samples in generally low due to strong coulomb scattering resulting from high doping concentrations.

IV. CONLUSION

This work has shown a process for fabrication of VdP-structures of $Ge_{1-x}Sn_x$ using MBE for layer growth. An XRD analysis was performed to ensure high crystal quality and to determine Sn-concentrations in the active $Ge_{1-x}Sn_x$-epilayer. Data from Hall measurements clearly show a decrease in charge carrier mobility with increasing Sn-content. This may be due to an increase in scattering at higher Sn-concentrations.

REFERENCES

[1] E. Kasper, J. Werner, M. Oehme, S. Escoubas, N. Burle, J. Schulze "Growth of silicon based germanium tin alloys" in Thin Solid Films 520 (2012).

[2] J. D. Sau and M. L Cohen "Possibility of increased mobility in Ge-Sn alloy systems" in Phys. Rev. B 75, 045208 (2007).

[3] M. R. Bauer, J. Tolle, C. Bungay, A. V. G. Chizmenshya, D. J. Smith, J. Menéndez and J. Kouvetakis "Tunable band structure in diamond-cubic tin-germanium alloys grown on silicon substrates" in Solid State Communications 127 (2003).

[4] S. Wirths, R. Geiger, N. v. d. Driesch, G. Mussler, T. Stoica, S. Mantl, Z. Ikonic, M. Luysberg, S. Chiussi, J. M. Hartmann, H. Sigg, J. Faist, D. Buca and D. Grützmacher "Lasing in direct-bandgap GeSn alloy grown on Si" in nature Photonics 9. 88-92 (2015).

[5] S. Zaima, O. Nakatsuka, Y. Shimuraa, M. Adachi, M. Nakamura, S. Takeuchi, B. Vincent, F. Gencarelli, T. Clarysse, J. Demeulemeester, K. Temst, A. Vantomme, M. Caymax, and R. Loo "GeSn Technology: Impact of Sn on Ge CMOS Applications" in ECS Transactions, 41 (7) 231-238 (2011).

[6] K. P. Homewood and M. A. Lourenço „The rise of the GeSn laser" in Nature Photonics Vol. 9, February 2015.

[7] S. Gupta, R. Chen, B. Magyari-Kope, H. Lin, B. Yang, A. Nainani, Y. Nishi, J. S. Harris and K. C. Saraswat, "GeSn Technology: Extending the Ge Electronics Roadmap" in 2011 International Electron Devices Meeting.

[8] K. Moto, R. Yoshimine, T. Suemasu and K. Toku "Improving carrier mobility of polycrystaline Ge by Sn doping" in Scientific Reports 8:14832 (2018).

[9] B. Vincent, F. Gencarelli, H. Bender, C. Merckling, B. Douhard, D. H. Petersen, O. Hansen, H. H. Henrichsen, J. Meerschaut, W. Vandervorst, M. Heyns, R. Loo and M. Cymax "Undoped and in-situ B doped GeSn epitaxial growth on Ge by atmospheric pressure-chemical vapor deposition" in Appl. Phys. Lett. 99, 152103 (2011).

[10] D. C. Streit and F. G Allen "Thermal and Si-beam assisted desorption of SiO_2 from silicon in ultrahigh vacuum" in J. Appl. Phys. 61 (8) 15. April 1987.

[11] D. V. Yurasov, A. V. Antonov, M. N. Drozdov, V. B. Schmagin, K. E. Spirin and A. V. Novikov "Antimony segregation in Ge and formation of n-type selectively doped Ge films in molecular beam apitaxy" in J. Appl. Phys. 118, 145701 (2015).

[12] M. Oehme, D. Widmann, K. Kostecki, P. Zaumseil, B. Schwartz, M. Gollhofer, R. Koerner, S. Bechler, M. Kittler, E. Kasper and J. Schulze "GeSn/Ge multiquantum well photodetectors on Si substrates" in Optics Letters Vol. 39, No. 16, 4711 (2014).

[13] M. Oehme, K. Kostecki, M. Schmid, F. Oliveira, E. Kasper, J. Schulze "Epitaxial growth of strained and unstrained GeSn alloys up to 25% Sn" in Thin Solid Films 557, 169-172 (2014)

[14] P. J. Newton, R. Mansell, S. N. Holmes, M. Myronov and C. H. W. Barnes "Weak localization and weak antilocalization in doped germanium epilayers" in Appl. Phys. Lett. 110, 062101 (2017)

[15] O. Nakatsuma, N. Tsutsui, Y. Shimura, S. Takeuchi, A. Sakai and S. Zaima "Mobility behaivor of $Ge_{1-x}Sn_x$ Layers Grown on Silicon-on-Insulator Substrates" in Jpn. J. Appl. Phys. 49, 04DA10 (2010).

On the modelling of interface roughness scattering in AlGaN/GaN heterostructures

Ivan Berdalović*, Mirko Poljak*, Tomislav Suligoj*
*Faculty of Electrical Engineering and Computing, University of Zagreb, 10000 Zagreb, Croatia
E-mail: ivan.berdalovic@fer.hr

***Abstract*—AlGaN/GaN heterostructures are attractive for high-power radiofrequency applications due to the wide bandgap of GaN and the high values of spontaneous and piezoelectric polarisation in such material systems, which lead to high breakdown voltages and high 2D carrier concentrations. It is of particular interest to accurately model the 2DEG mobilities in these structures, as different scattering mechanisms can limit the total mobility at different temperatures and carrier concentrations. In this paper, we present a semi-classical modelling framework for low-field mobility in AlGaN/GaN heterostructures, focussing especially on the impact of interface roughness (IFR) scattering on the low-temperature mobility. The framework is validated by comparing calculated mobility values with experimental data, and different ways of modelling the IFR scattering are investigated. A non-linear IFR scattering model is used to obtain the best match to measured data for different Al-contents in the AlGaN layer in case of a low potential barrier at the heterointerface, while the temperature dependence of mobility with a high potential barrier in an AlGaN/AlN/GaN system is best matched with IFR models where the squared scattering matrix elements depend linearly on the roughness power spectrum.**

***Index Terms*—Radiofrequency integrated circuits, power semiconductor devices, quantum well devices, gallium nitride, high electron mobility transistors, heterojunctions, two-dimensional electron gas, semiconductor device modelling, charge carrier mobility**

I. Introduction

In recent years, wide-bandgap III-V semiconductors such as GaN have been gaining popularity in high-power radiofrequency (RF) electronics due to their high breakdown voltage and high thermal conductivity [1]. In addition, heterostructures consisting of GaN and its alloys, most commonly AlGaN, exhibit high spontaneous and piezoelectric polarisation fields stemming from the difference in the spontaneous polarisation of the two polar materials and the mechanical stress between the AlGaN and GaN layers [2]. As a consequence, a bound sheet charge appears at the heterointerface, the sign of which depends on the arrangement and polarity of AlGaN and GaN layers. In the case of a thin AlGaN layer epitaxially grown on Ga-face wurtzite GaN in the [0001] direction, a positive polarisation sheet charge in the order of 10^{13} cm^{-2} is induced at the interface. To compensate this positive charge, electrons are collected in a quantum well on the GaN-side of the interface, creating a two-dimensional electron gas (2DEG). This way, a high carrier concentration can be achieved even without intentional doping in the channel and without applied bias [3], which is the basis of operation for AlGaN/GaN high electron mobility transistors (HEMTs).

Modelling of the electron mobility within such structures has been the topic of a large number of research papers [4]–[8]. Most researchers use simple analytical models describing the electron wavefunctions in the Fang-Howard approximation [9] and taking into account only the ground state within the quantum well. While there is general agreement about the fact that room-temperature 2DEG mobility is limited by scattering due to polar optical phonons, there is still no consensus about the mobility limitations at low temperatures, which have been attributed to alloy scattering in AlGaN, Coulomb scattering at residual impurities or charged centres at the heterointerface as well as scattering at heterointerface roughness.

In this work, we employ a comprehensive numerical modelling framework to analyse the low-temperature mobility in AlGaN/GaN heterostructures. The exact 2DEG wavefunctions, carrier concentrations as well as bound states within the quantum well near the heterointerface are calculated by solving the Schrödinger and Poisson equations self-consistently for a given material system [10]. This is followed by semi-classical mobility calculations based on solving the Boltzmann transport equation (BTE) in the momentum relaxation time approximation (MRTA). The carrier mobilities are calculated from the momentum relaxation times for each scattering mechanism [11]. With all the relevant scattering mechanisms taken into account, we find that scattering at heterointerface roughness is indeed a plausible explanation for the limitation of low-temperature mobility, which is validated by comparing the theoretical results with experimental data. We also explore several different models for interface roughness scattering proposed in literature and find the most suitable one for AlGaN/GaN and AlGaN/AlN/GaN systems.

II. Framework for low-field mobility modelling

Since the 2DEG electrons are confined within a quantum well in the direction perpendicular to the heterointerface,

This work was supported by the Croatian Science Foundation (HRZZ) under the project IP-2018-01-5296.

the single-particle Schrödinger equation is solved for Γ-valley electrons within the effective mass approximation for spherical, non-parabolic energy bands to obtain the electron wavefunctions and bound states needed for the calculation of scattering rates. To obtain the potential profile used in solving the Schrödinger equation, the Poisson equation needs to be solved taking into account the carrier concentrations given by the solution of the Schrödinger equation, and this process is repeated iteratively until convergence is achieved.

The polarisation sheet charge is modelled as a delta-function at the heterointerface using the approach from [12], where the sheet charge density assuming pseudomorphic growth of an $Al_xGa_{1-x}N$ layer on relaxed GaN is given by:

$$\begin{aligned}|\sigma(x)| &= |P_{PE}(\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{N}) + P_{SP}(\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{N}) \\ &- P_{SP}(\mathrm{GaN})| = \left|2\frac{a(0)-a(x)}{a(x)}\left[e_{31}(x) - e_{33}(x)\frac{C_{13}(x)}{C_{33}(x)}\right]\right. \\ &\left. + P_{SP}(x) - P_{SP}(0)\right|,\end{aligned} \tag{1}$$

with P_{PE} and P_{SP} denoting the piezoelectric and spontaneous polarisation, respectively. $a(x)$ is the lattice constant of $Al_xGa_{1-x}N$, while $e(x)$ and $C(x)$ are the relevant piezoelectric and elastic constants, respectively.

The scattering rate for the inelastic transition of a carrier from an initial subband n and wavevector $\mathbf{k}$ to a final subband n' and wavevector $\mathbf{k}'$ is calculated using Fermi's golden rule as:

$$S_{n,n'}(\mathbf{k},\mathbf{k}') = \frac{2\pi}{\hbar}\left|M_{n,n'}(\mathbf{k},\mathbf{k}')\right|^2 \delta[E_n(\mathbf{k}) - E_{n'}(\mathbf{k}') \pm \hbar\omega], \tag{2}$$

where $E(\mathbf{k})$ and $E(\mathbf{k}')$ are the energies of the initial and final state, respectively, $\pm\hbar\omega$ is the phonon energy in the case of phonon absorption/emission and $M_{n,n'}(\mathbf{k},\mathbf{k}')$ is the scattering matrix element calculated from the 2DEG envelope wavefunctions as follows:

$$M_{n,n'}(\mathbf{k},\mathbf{k}') = \int \psi^{\dagger}_{n',\mathbf{k}'}(\mathbf{r}) U_{sc}(\mathbf{r}) \psi_{n,\mathbf{k}}(\mathbf{r}) d\mathbf{r}. \tag{3}$$

Here, U_{sc} is the scattering potential which is dependent on the scattering mechanism in question. In the momentum relaxation time approximation, the MRTs are computed from the scattering rates for each scattering mechanism and each subband as follows [13]:

$$\frac{1}{\tau_i(\mathbf{k})} = \sum_{j,\mathbf{k}'} S_{i,j}(\mathbf{k},\mathbf{k}') \left[\frac{1-f_0(E')}{1-f_0(E)}\right][1-\cos\theta], \tag{4}$$

where f_0 is the Fermi-Dirac distribution and θ is the angle between $\mathbf{k}$ and $\mathbf{k}'$. The energy-dependent MRTs of different scattering mechanisms are then summed up for each subband using Mathiessen's rule, and the subband mobility is calculated as:

$$\mu_i = \frac{e\left\langle\tau_i\right\rangle}{m_c}, \tag{5}$$

where e is the electron charge, m_c the effective conductivity mass, and $\langle\tau_i\rangle$ is given by the Kubo-Greenwood formula for non-parabolic bands:

$$\left\langle\tau_i\right\rangle = \frac{\int_{\varepsilon_i}^{\infty}(E-\varepsilon_i)[1+2\alpha(E-\varepsilon_i)]\tau_i(E)\left|\frac{\partial f_0(E)}{\partial E}\right| dE}{\int_{\varepsilon_i}^{\infty}(E-\varepsilon_i)[1+2\alpha(E-\varepsilon_i)]\left|\frac{\partial f_0(E)}{\partial E}\right| dE}, \tag{6}$$

where ε_i denotes the subband energy and α is the non-parabolicity factor. Finally, the total electron mobility is found by averaging subband mobilities over subband populations:

$$\mu = \sum_i \frac{N_i}{N_{2DEG}}\mu_i, \tag{7}$$

where N_{2DEG} is the total 2DEG charge.

III. Models for interface roughness scattering

Traditionally, interface roughness scattering has been modelled using a squared scattering matrix element which is proportional to the power spectrum $S_R(\mathbf{q})$ of the roughness [13], [14]. The exponential form of the roughness spectrum has been shown the most successful in reproducing HRTEM measurements [15], and is given by:

$$S_R(q) = \frac{\pi\Delta^2\lambda^2}{\left(1+\frac{q^2\lambda^2}{2}\right)^{\frac{3}{2}}}, \tag{8}$$

where Δ is the RMS interface roughness, λ is the roughness correlation length and q is the magnitude of the wavevector redirection due to scattering. The squared matrix element is then given by:

$$\left|M_{n,n'}\right|^2 = \left|\mathcal{M}_{n,n'}\right|^2 \frac{S_R(q)}{A}, \tag{9}$$

where $\mathcal{M}_{n,n'}$ depends on the confining potential in the quantisation direction $V(z)$. If one assumes a high enough step in $V(z)$ at the heterointerface, i.e. approaching an infinite potential barrier Φ_b, where the wavefunctions vanish at the interface, the unscreened intra-subband matrix element for the i-th subband can be expressed as [16], [17]:

$$\mathcal{M}_{i,i} = \frac{\hbar^2}{2m_z}\left(\frac{d\psi_i(0)}{dz}\right)^2, \tag{10}$$

with m_z being the effective quantisation mass. On the other hand, for a finite potential barrier, i.e. taking into account the wavefunction penetration into the AlGaN layer, one can write [14]:

$$\mathcal{M}_{i,i} = \int_0^{\infty} \frac{dV}{dz}\psi_i^2(z)dz. \tag{11}$$

To account for screening by free carriers, following [13], we divide the scattering matrix elements by a static scalar dielectric function $\epsilon_D(q)$ to obtain the screened matrix elements.

In this paper, we also investigate a model proposed in [18], where the matrix element $\mathcal{M}_{n,n'}[\Delta(\mathbf{r})]$ is a non-linear transformation of the roughness spectrum, and for intra-subband transitions reads (neglecting kinetic energy contributions):

$$\mathcal{M}_{i,i}(\Delta) = -\int_{\Delta}^{0} \Phi_b \psi_i^2(z) dz, \quad \Delta < 0, \tag{12a}$$

$$\mathcal{M}_{i,i}(\Delta) = \int_{0}^{\Delta} \Phi_b \psi_i^2(z) dz, \quad \Delta > 0. \tag{12b}$$

The final matrix element $M_{i,i}(q)$ is then obtained as:

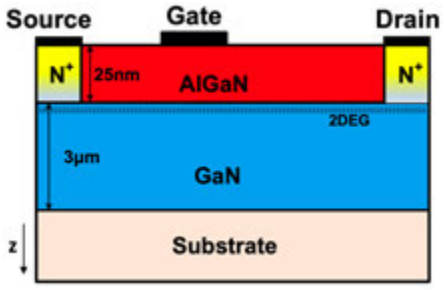
$$|M_{i,i}(q)|^2 = \frac{2\pi}{A} \int_0^{\infty} C_M(r) J_0(qr) r dr, \tag{13}$$

where J_0 is the zero-order Bessel function and C_M is the autocorrelation function of $\mathcal{M}_{i,i}(\Delta)$, as defined in [19]. Mobility calculations for silicon MOSFETs using this non-linear formulation have given the extracted Δ values closest to AFM/TEM measurements on SiO_2/Si and Al_2O_3/InGaAs surfaces [20]. Therefore, it is interesting to compare these findings with AlGaN/GaN interfaces.

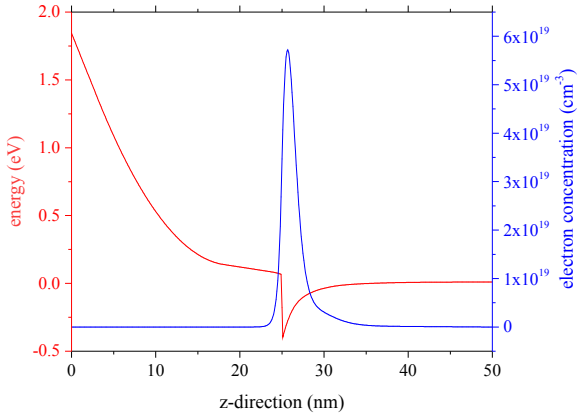
IV. Comparison with Experimental Data

To check the validity of our models, the calculted mobilities have been compared with experimental data from [21]. The structures under study are composed of a 3 µm-thick nominally undoped GaN layer followed by a 7 nm undoped AlGaN barrier layer, a 15 nm Si-doped AlGaN supply layer with a doping density of around 5×10^{18} cm^{-3} and an undoped AlGaN cap layer with a thickness of 3 nm. The Al-content x of the $Al_xGa_{1-x}N$ layers was varied from 0.16 to 0.42. The structure of a basic HEMT utilising the AlGaN/GaN heterostructure described can be seen in Fig. 1a. The simulated conduction band profile in the z-direction as well as the 2DEG carrier distribution of such a structure with an Al-content of 0.31 is shown in Fig. 1b. The bandgap of $Al_xGa_{1-x}N$ was calculated as $E_g = 6.13x - 3.42(1-x) - 0.7x(1-x)$, and the conduction band discontinuity between AlGaN and GaN was taken as $0.7E_g$. With no information on the material of the Schottky gate, the Schottky barrier height was assumed as that of an Ni contact and modelled as an x-dependent quantity according to [22]. The simulated values of 2DEG density as well as the experimental values can be seen in Fig. 2. Note that the applied gate voltage had to be modified to -0.5 V in order to give a good match to the measured 2DEG densities in the investigated range of x, implying that the actual Schottky barrier is higher than assumed.

In the simulations comparing the calculated mobilities with experimental data, we consider the 4 lowest subbands



(a)



(b)

Fig. 1: (a) Basic HEMT structure using the AlGaN/GaN heterostructure under study, (b) conduction band profile and 2DEG concentration under the gate in the z-direction with an Al-content of 0.31 (2DEG density of 1.4×10^{13} cm^{-2}).

within the Γ-valley of GaN, which contain more than 99% of carriers within the 2DEG. For these subbands, we consider scattering by acoustic phonons, including inter-subband deformation potential scattering (ADP) and intra-subband

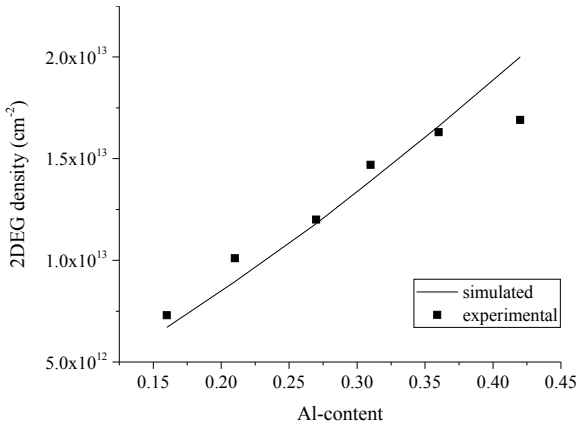


Fig. 2: Simulated and measured 2DEG density as a function of Al-content (experimental data from [21]).

TABLE I: MATERIAL PARAMETERS FOR GaN USED IN SIMULATION.

Parameter	Unit	Value
effective electron mass	m_0	0.21
non-parabolicity factor	eV^{-1}	0.189
mass density	$\mathrm{kg/m}^3$	6.15×10^3
longitudinal sound velocity	m/s	6.56×10^3
acoustic deformation potential	eV	9.1
piezoelectric constant	V/m	4.28×10^9
longitudinal elastic constant	N/m	2.66×10^{11}
transverse elastic constant	N/m	0.62×10^{11}
polar optical phonon energy	meV	99
static dielectric constant	ϵ_0	9.5
high frequency dielectric constant	ϵ_0	5.35
alloy scattering potential	eV	1.8
dislocation density	cm^{-2}	10^6
background impurity concentration	cm^{-3}	10^{16}

piezoelectric scattering (PE), as well as intra-subband scattering by polar optical phonons (POP), alloy disorder (ADO) in the AlGaN barrier layer, interface roughness scattering (IFR), scattering due to dislocations (DIS) and Coulomb scattering (CO) due to remote donors in the AlGaN layer as well as residual impurities in the GaN channel. The material parameters for the GaN layer used in the calculations are listed in Table I.

Note that in comparing the calculated electron mobilities with measured Hall mobilities, the results were corrected by the Hall factor for each scattering mechanism and subband, which was calculated using the ratio [23]:

$$r = \frac{\langle \tau^2 \rangle}{\langle \tau \rangle^2}. \tag{14}$$

The value of this factor ranges between 1 and 1.3 depending on the scattering mechanism and temperature, and needs to be taken into account for a realistic comparison with measured data.

The fact that the RMS interface roughness depends on the strain and hence the Al-content of the AlGaN layer also needs to be taken into account. For simplicity, we adopt the approach from [21] in assuming a linear dependence of $\Delta(x)$ on Al-content and a constant value of λ equal to 3 nm, which is in close agreement with previously reported values [24]. We assume the worst case for the degradation of Δ with increasing Al-content, such that the percentage increase of the RMS roughness Δ at the AlGaN/GaN interface corresponds to the increase of the RMS surface roughness of strained AlGaN layers measured by AFM [21]. Therefore, the increase in interface roughness with strain is modelled entirely by the increase of Δ.

Figure 3 shows the x-dependence of the calculated mobility values for different scattering mechanisms at a temperature of 77 K. It can be seen that the total calculated Hall mobility gives a satisfactory match (within 10%) to the experimental data measured in [21]. The IFR scattering model used to obtain these results is the linear model assuming a finite potential barrier, with Δ values ranging

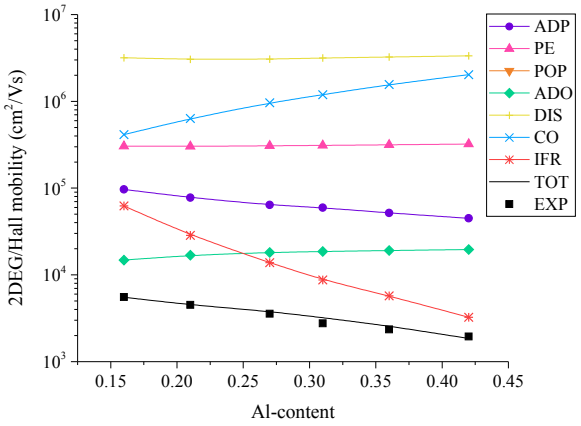


Fig. 3: Calculated 2DEG mobilities limited by different scattering mechanisms together with calculated and measured total Hall mobilities as a function of Al-content at 77 K (experimental data from [21]). At this temperature, the POP-limited mobility is higher than 10^7 cm^2/Vs. The IFR-limited mobility is calculated using the finite-barrier formulation of the linear model.

from 0.18 nm for $x = 0.16$ to 0.32 nm for $x = 0.42$. The decrease in mobility with Al-content is explained not just by increasing Δ values, but also by an increase in 2DEG density. While the trend is the same for all three models, the other two models give vastly different mobility values with the same $\Delta(x)$ values. This is shown in Fig. 4, where the IFR-limited mobility as well as the total Hall mobilities with the three models are compared. It is noticeable that, particularly with the infinite-barrier formulation of the linear IFR model, the calculated mobility values are far from the measured ones.

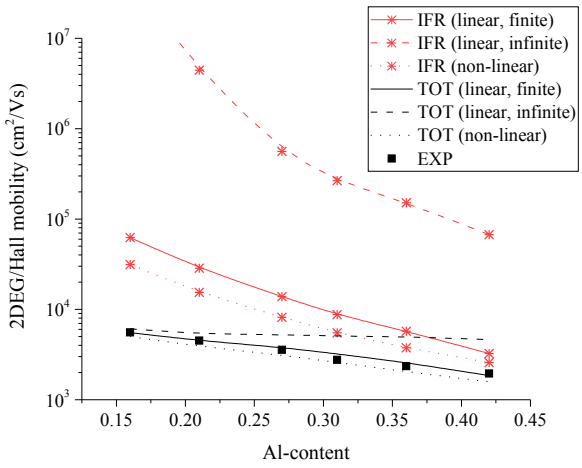


Fig. 4: Calculated IFR-limited 2DEG mobilities as well as calculated and measured total Hall mobilities as a function of Al-content at 77 K (experimental data from [21]), using three different models for the IFR scattering with the same $\Delta(x)$ and λ values. The infinite-barrier linear model severely overestimates the mobility.

Note, however, that all three models can provide a good fit of the experimental data with different values of Δ. Namely, the Δ values giving the best fit range from 0.69 nm for $x = 0.16$ to 1.2 nm for $x = 0.42$ using the infinite-barrier linear model, and from 0.14 nm for $x = 0.16$ to 0.25 nm for $x = 0.42$ using the non-linear model (in both cases considering a linear dependence of Δ on Al-content, as previously). The Δ values needed to obtain a good fit with the linear model assuming an infinite potential barrier are close to 4 times higher than the values used in the other two models and are much larger than the numbers typically reported in literature [24]. Hence, we can conclude that the infinite-barrier formulation of the linear model, which was indeed derived assuming vanishing values of the wavefunctions at the interface, is not applicable in the case of low potential barriers of several tenths of an electronvolt, as is the case in $Al_xGa_{1-x}N$/GaN systems, especially at low values of x. The unphysically high values of interface-roughness-limited mobility with this model using realistic values of Δ, especially at low Al-contents, stem from the fact that the higher subbands are only weakly confined within the shallow potential well. Therefore, the derivative of the envelope functions at the interface in (10) can take exceedingly low values, resulting in a deceptively high mobility in the higher subbands. Even though the higher subbands contain only a low percentage of carriers, the weighted sum in (7) can still result in a gross overestimation of mobility if the mobilities of higher subbands are severely overestimated.

On that note, we can conclude that taking into account the wavefunction penetration due to a finite potential barrier is necessary for IFR-limited mobility calculations with a low potential barrier at the interface (below cca. 1 eV). The non-linear model is also convenient for calculations with these structures, and gives very similar results to the finite-barrier linear model, albeit using slightly lower values of Δ.

V. Interface roughness scattering in an AlGaN/AlN/GaN system

It is interesting to compare the models in the case of a higher potential barrier, where one would expect a closer agreement between the IFR mobility values obtained by assuming a finite or infinite barrier with the linear model. Such a case can easily be engineered by adding an AlN spacer layer between the AlGaN barrier layer and the GaN channel, which is commonly done to improve the confinement of carriers within the 2DEG and to eliminate alloy scattering in AlGaN [25]. The AlN layer is assumed to be thick enough such that the IFR scattering is determined solely by the potential step at the AlN/GaN interface. On the other hand, the AlN layer is assumed to be thin enough not to significantly affect the values of polarisation charge at the interface and hence the 2DEG density, which provides the possibility of directly comparing the scattering rates and mobilities with and without the AlN spacer.

The IFR scattering models for the case of an AlGaN/AlN/GaN system are investigated by comparing the temperature dependence of calculated mobilities with measured Hall mobility data from [26]. These structures consist of a 170 nm thick GaN layer followed by 0.8 nm of AlN spacer, 28 nm of $Al_{0.3}Ga_{0.7}N$ barrier and a 1.2 nm GaN cap layer. All the layers are nominally undoped. The simulated 2DEG density at zero gate bias matches the measured density of 1.2×10^{13} cm^{-2}. Figure 5 shows the evolution of the mobility limited by various scattering mechanisms with changing temperature.

Once again, assuming the same low density of dislocations as before and a low background impurity concentration typically reported in nominally undoped structures [25] (see Table I), IFR scattering is the dominant mechanism at low temperatures. At temperatures above cca. 250 K, polar optical phonon scattering (POP) becomes the dominant mechanism, but IFR scattering still plays a significant role. Therefore, even though IFR scattering is close to being temperature-independent (the slight increase in IFR-limited mobility with temperature is mainly due to a decrease in the population of the lowest subband), the accurate modelling of this scattering mechanism is important to obtain a good match with experimental results over the full temperature range. Once again, the fitting is performed by changing the roughness parameter Δ, while keeping the λ value fixed at 3 nm. The Δ values resulting in the best fitting of the measured data are 0.17 nm, 0.24 nm and 0.09 nm for the linear model with a finite barrier, the linear model with an infinite barrier and the non-linear model, respectively. One can observe a much closer agreement between the two formulations of the linear model, both of which now give realistic values of Δ. This is explained by the fact that

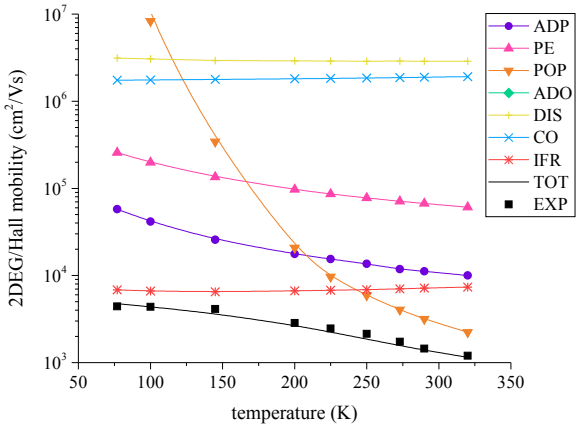


Fig. 5: Temperature dependence of calculated 2DEG mobilities limited by different scattering mechanisms and of calculated and measured total Hall mobilities for the AlGaN/AlN/GaN structure with an Al-content of 0.3 (experimental data from [26], 2DEG density of 1.2×10^{13} cm^{-2}). With the addition of the AlN layer, the ADO-limited mobility is higher than 10^7 cm^2/Vs.

the conduction band discontinuity between AlN and GaN is in the order of 1.8 eV, so several times higher than the discontinuity between AlGaN with a low Al-content and GaN. Hence, with little wavefunction penetration into the AlN spacer, the infinite-barrier formulation gives realistic results. In this case, the non-linear model gives somewhat higher IFR scattering rates compared to the linear model, owing to the stronger dependence of the matrix elements on the barrier height Φ_b. As a result, the fitted Δ value of 0.09 nm is somewhat lower than experimentally observed values.

VI. Conclusion

Reliable modelling of 2DEG mobility in AlGaN/GaN HEMTs is an important step in predicting the overall performance of these devices. We have implemented a comprehensive numerical framework for low-field mobility modelling, which takes into account all the relevant mechanisms for carrier scattering within multiple subbands of the quantum well formed at the interface of an AlGaN/GaN heterostructure. The calculated 2DEG densities and Hall mobilities provide a good match to experimental data over a wide range of temperatures and Al-contents in the AlGaN barrier layer, thus validating the modelling framework as a whole. In particular, we have investigated the use of three different models for interface roughness scattering (IFR), which is confirmed to be the dominant scattering mechanism at low temperatures. The only fitting parameter used in matching the measured data is the RMS interface roughness Δ, which has been measured in a limited number of studies. We find that, in the case of a traditional AlGaN/GaN heterostructure, a linear IFR model assuming an infinite potential barrier overestimates the Δ values needed to match the experiment, while a finite-barrier linear model or a non-linear model can both be used to obtain accurate results with realistic Δ values. On the other hand, in an AlGaN/AlN/GaN structure, where the potential barrier at the heterointerface is significantly higher, all three models give Δ values close to reported ones, with the non-linear model slightly underestimating the mobility.

References

[1] H. Arabshahi and A. Mowlavi, "Low-field electron transport properties in zincblende and wurtzite GaN structures using an iteration model for solving Boltzmann equation," *Mod. Phys. Lett. B*, vol. 23, no. 10, pp. 1359–1366, 2009.

[2] O. Ambacher *et al.*, "Pyroelectric properties of Al(In)GaN/GaN hetero- and quantum well structures," *J. Phys. Condens. Matter*, vol. 14, no. 13, pp. 3399–3434, 2002.

[3] K.-S. Lee, D.-H. Yoon, S.-B. Bae, M.-R. Park, and G.-H. Kim, "Self-consistent subband calculations of AlGaN/GaN single heterojunctions," *ETRI J.*, vol. 24, no. 4, pp. 270–279, 2002.

[4] B. K. Ridley, B. E. Foutz, and L. F. Eastman, "Mobility of electrons in bulk GaN and $Al_xGa_{1-x}N$/GaN heterostructures," *Phys. Rev. B*, vol. 61, pp. 16 862–16 869, 2000.

[5] X. Z. Dang *et al.*, "Measurement of drift mobility in AlGaN/GaN heterostructure field-effect transistor," *Appl. Phys. Lett.*, vol. 74, no. 25, pp. 3890–3892, 1999.

[6] M. N. Gurusinghe, S. K. Davidsson, and T. G. Andersson, "Two-dimensional electron mobility limitation mechanisms in $Al_xGa_{1-x}N$/GaN heterostructures," *Phys. Rev. B*, vol. 72, p. 045316, 2005.

[7] S. B. Lisesivdin, A. Yildiz, N. Balkan, M. Kasap, S. Ozcelik, and E. Ozbay, "Scattering analysis of two-dimensional electrons in AlGaN/GaN with bulk related parameters extracted by simple parallel conduction extraction method," *J. Appl. Phys.*, vol. 108, no. 1, p. 013712, 2010.

[8] K. Abgaryan, I. Mutigullin, and D. Reviznikov, "Theoretical investigation of 2DEG concentration and mobility in the AlGaN/GaN heterostructures with various Al concentrations," *Phys. Stat. Sol. C*, vol. 12, no. 12, pp. 1376–1382, 2015.

[9] F. Fang and W. Howard, "Negative field-effect mobility on (100) Si surfaces," *Phys. Rev. Lett.*, vol. 16, no. 18, pp. 797–799, 1966.

[10] M. Poljak, V. Jovanovic, D. Grgec, and T. Suligoj, "Assessment of electron mobility in ultrathin-body InGaAs-on-insulator MOSFETs using physics-based modeling," *IEEE Trans. Electron Devices*, vol. 59, no. 6, pp. 1636–1643, 2012.

[11] M. Poljak, "Carrier transport in low-dimensional nanoelectronic devices," Ph.D. dissertation, University of Zagreb, Zagreb, 2013.

[12] O. Ambacher *et al.*, "Two dimensional electron gases induced by spontaneous and piezoelectric polarization undoped and doped AlGaN/GaN heterostructures," *J. Appl. Phys.*, vol. 87, no. 1, pp. 334–344, 2000.

[13] D. Esseni, P. Palestri, and L. Selmi, *Nanoscale MOS Transistors: Semi-Classical Transport and Applications*. Cambridge University Press, 2011.

[14] D. Esseni, "On the modeling of surface roughness limited mobility in SOI MOSFETs and its correlation to the transistor effective field," *IEEE Trans. Electron Devices*, vol. 51, no. 3, pp. 394–401, 2004.

[15] S. M. Goodnick, D. K. Ferry, C. W. Wilmsen, Z. Liliental, D. Fathy, and O. L. Krivanek, "Surface roughness at the Si(100)-SiO_2 interface," *Phys. Rev. B*, vol. 32, pp. 8171–8186, 1985.

[16] T. Ando, "Screening effect and quantum transport in a silicon inversion layer in strong magnetic fields," *J. Phys. Soc. Jpn.*, vol. 43, no. 5, pp. 1616–1626, 1977.

[17] T. Nee and R. Prange, "Quantum spectroscopy of the low field oscillations of the surface impedance," *Phys. Lett. A*, vol. 25, no. 8, pp. 582–583, 1967.

[18] D. Lizzit, D. Esseni, P. Palestri, and L. Selmi, "A new formulation for surface roughness limited mobility in bulk and ultra-thin-body metal-oxide-semiconductor transistors," *J. Appl. Phys.*, vol. 116, no. 22, p. 223702, 2014.

[19] R. Baum, "The correlation function of Gaussian noise passed through nonlinear devices," *IEEE Trans. Inf. Theory*, vol. 15, no. 4, pp. 448–456, 1969.

[20] S. Krivec, M. Poljak, and T. Suligoj, "The physical mechanisms behind the strain-induced electron mobility increase in InGaAs-on-InP MOSFETs," *IEEE Trans. Electron Devices*, vol. 65, no. 7, pp. 2784–2789, 2018.

[21] M. Miyoshi, T. Egawa, and H. Ishikawa, "Structural characterization of strained AlGaN layers in different Al content AlGaN/GaN heterostructures and its effect on two-dimensional electron transport properties," *J. Vac. Sci. Technol. B,*, vol. 23, pp. 1527–1531, 2005.

[22] A. F. M. Anwar, E. W. Faraclas, and K. V. Smith, "Schottky barrier height in GaN/AlGaN heterostructures," in *2005 International Semiconductor Device Research Symposium*, Dec. 2005, pp. 221–222.

[23] A. Asgari, S. Babanejad, and L. Faraone, "Electron mobility, Hall scattering factor, and sheet conductivity in AlGaN/AlN/GaN heterostructures," *J. Appl. Phys.*, vol. 110, no. 11, p. 113713, 2011.

[24] J. Zhang, Y. Hao, J. Zhang, and J. Ni, "The mobility of two-dimensional electron gas in AlGaN/GaN heterostructures with varied Al content," *Sci. China Ser. F: Inf. Sci.*, vol. 51, no. 6, pp. 780–789, 2008.

[25] D. Jena, "Polarization induced electron populations in III-V nitride semiconductors: Transport, growth, and device applications," Ph.D. dissertation, University of California, Santa Barbara, 2003.

[26] D. Y. Protasov, T. V. Malin, A. V. Tikhonov, A. F. Tsatsulnikov, and K. S. Zhuravlev, "Electron scattering in AlGaN/GaN heterostructures with a two-dimensional electron gas," *Semiconductors*, vol. 47, no. 1, pp. 33–44, 2013.

Modeling of Electrical Properties of Al-on-Ge-on-Si Schottky Barrier Diode

Lovro Marković*, Tihomir Knežević*, and Tomislav Suligoj*
*University of Zagreb, Faculty of Electrical Engineering and Computing, Micro and Nano Electronics Laboratory
lovro.markovic2@fer.hr

Abstract – In this work, different mechanisms that could cause degradation of the ideality factor in Al/Ge Schottky diodes on Si substrate are examined. Measured *I-V* characteristics of Schottky diodes have been fitted by the model of the diode developed in TCAD environment. The effects of Shockley-Read-Hall recombination parameters of epitaxial Ge on the *I-V* characteristics are simulated. The impact of interface traps on both Al/Ge and Ge/Si interfaces, as well as the effect of a buffer oxide layer on Al/Ge interface are analyzed by simulations providing a possible explanation for a degraded ideality factor in Al/Ge-on-Si Schottky diodes.

***Keywords* – aluminum, germanium, Schottky barriers, semiconductor defects, epitaxial growth, heterojunctions, Ge-on-Si**

I. INTRODUCTION

In contrast to commercially available optoelectronic devices in III-V semiconductor materials, which are not technologically compatible with standard silicon CMOS processes, germanium epitaxy on silicon is offered as an ideal candidate for monolithic integration of both optical and electronic components [1]. Nevertheless, Ge-on-Si thin films could also be used as buffer layers for III-V epitaxy on silicon, which has already enabled the development of GaAs/Ge/Si monolithic tandem solar cells [2]. Various methods and techniques for integrating Ge devices on Si substrate have been developed and are mostly based on epitaxy [3].

Fabrication of stable and reliable metal contacts with n-type Ge has proven to be challenging and has therefore been the subject of numerous studies [4-8]. Most of the metals used for contacting the devices (Au, Cu, Ag, Pb, Ni, Al) form Schottky barriers to n-Ge with low donor concentration of up to 10^{16} cm^{-3} [4]. Most of them form much higher barriers to n-Ge than it would be calculated from metal workfunctions and semiconductor affinities by applying the Schottky-Mott rule [4]. Furthermore, contacting of n-Ge with aluminum has been especially troublesome and remains partly unexplained. Whereas most references report relatively high Schottky barriers calculated from current-voltage characteristics (~0.5 eV – 0.7 eV) which could be attributed to *Fermi level pinning* [6,8], some papers report drastically lower values when calculated from *C-V* characteristics (~0.3 eV) [4]. In addition, time-dependent changes of the Schottky barrier height (Φ_B) have also been measured in [9]. Finally, the value of Al/Ge Schottky diodes ideality factor is often reported to be in the range of 1.2 to 1.6, which should indicate either high values of SRH recombination currents, perimeter effects or the existence of some other physical mechanism which could significantly impact the current in forward bias. It has also been suggested that the increased values of the ideality factor could be explained by the formation of thin oxide on Al/Ge interface [5] or, more recently, by the formation of an inhomogeneous barrier [10]. Moreover, some authors have observed noticable interdiffusion of germanium and silicon atoms upon Ge epitaxy and the existence of defective interface layers up to a few hundred nanometers thick [3], which could also deteriorate optoelectronic device properties. Some results show noticable Al migration through the epitaxial Ge, which could cause pitting of Ge and even Si substrate layers [11].

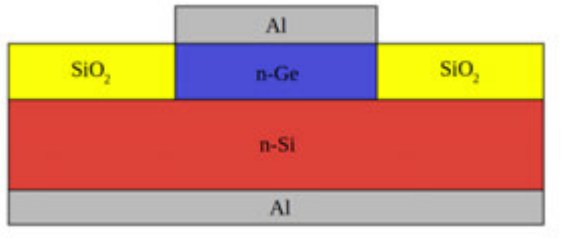


Figure 1: Al/Ge-on-Si Schottky diode structure used in simulations.

Different processes have been applied to improve the performance of Al/Ge contact, to depin Fermi level, pinned by defective interface states, and reduce the Schottky barrier. [6,8,12] However, the deposition of the thin layer of gallium and amorphous boron on the epitaxial germanium before the deposition of an aluminum electrode [3] has been proposed as a technology which could simultaneously provide better contact formation and ultrashallow doping of epitaxial Ge [13]. This method improves the performance of Al/Ge-on-Si photodiodes by significantly enhancing the ideality factor close to one. It has still not been fully explained what is the exact physical cause of ideality factor improvement, but characterization of a similar amorphous-B layer deposited on Si has shown formation of an interface hole layer (IHL) [14] which could directly impact the band structure of the device. In order to broaden the understanding of electrical properties of amorphous-B and Ga layers on the top of Ge, this paper presents an analysis of different mechanisms that could be the potential cause of a degraded ideality factor in Al/Ge-on-Si Schottky diodes.

This work has been fully supported by Croatian Science Foundation under the project IP-2018-01-5296.

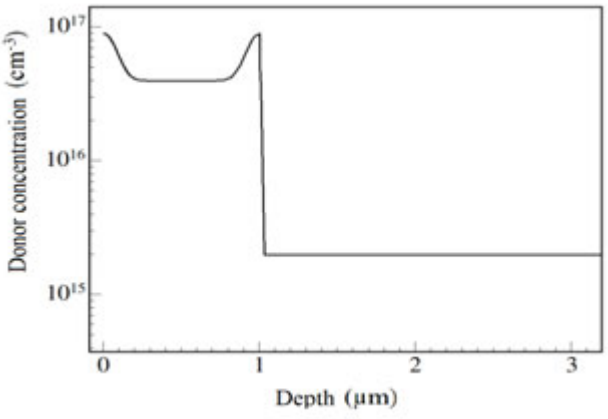


Figure 2: Doping profile of the simulated diode. The top 1 μm region is the Ge layer. Substrate doping is calculated to fit the wafer resistivity, ρ = 2.23 Ωcm

An analysis of Al/n-Ge Schottky barrier diode *I-V* characteristics has been performed by fitting the measured characteristics from [3]. The impact of different defective states on Al/Ge and Ge/Si interfaces in Al/Ge-on-Si Schottky diodes has been modeled in order to explain the contribution of each mentioned effect on current-voltage characteristics of the modeled diode.

II. Examined Structure

The Schottky barrier diode has been modeled using Sentaurus TCAD software. Three different approaches to modeling the Schottky barrier on Al/Ge interfaces have been compared. The basic structure is shown in Fig. 1. Silicon substrate with $2 \cdot 10^{15}$ cm^{-3} doping concentration has been defined, which corresponds to (100) 2-5 Ωcm phosphorous doped n-type Si wafer used in fabricated diodes [3]. The Philips unified mobility model has been used [15] in simulations [16].

On top of the silicon substrate, there is a 1 μm-thick layer of germanium. Germanium is grown inside 40x40 μm^2 oxide windows and doped by As. Material parameters used for simulations of germanium layer are mostly taken from [16]. One exception are mobility parameters where the Philips unified mobility model is used with parameters adjusted to the data found in [17]. Values of maximum carrier lifetime, for low-doped bulk Ge and for temperature of 300 K are $\tau_e = 10^{-3}$ s and $\tau_h = 10^{-4}$ s, for electrons and holes, respectively. Specific lifetime is calculated for every simulation node, depending on local dopant concentration. The other exception are emission velocities for calculation of thermionic current density which had to be calculated for the specific material. The density of thermionic emission current over the Schottky barrier equals to [18]:

$$J = A^* T^2 exp(-q\Phi_B/kT)\,[exp(-qV/kT) - 1], \quad (1)$$

where A^* is the material-specific Richardson constant, Φ_B is the Schottky barrier height and V is the external bias of the diode.The Richardson constant is defined as [18]:

$$A* = 4\pi q m^* k^2 / h^3, \quad (2)$$

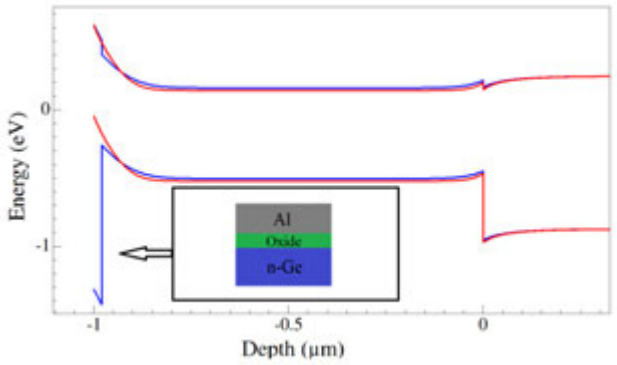


Figure 3: Comparison of band diagrams for simulation of direct Al/Ge Schottky contact and contact with existence of interface oxide layer, shown in red and blue, respectively. Schottky barrier height is set to 0.67 eV. Inset shows 2D structure with interface oxide, which is, in the simulation, modeled to be 20 nm thick.

where m^* is the carrier effective mass used for the density of states (DOS) calculations. The thermionic emission velocity, also called Richardson velocity, is given by [19]:

$$v_R = \sqrt{kT/2\pi m^*}, \quad (3)$$

By combining equations (2) and (3), emission velocity can be calculated from the Richardson constant:

$$v_R = \sqrt{2qk^3T/h^3A^*}. \quad (4)$$

For n-type Ge (100), the Richardson constant equals to 142.8 A/cm^2/K^2 [18] [20]. As electrons are majority carriers, their emission velocity is $2.468 \cdot 10^6$ cm/s.

The doping profile in Ge is shown in Fig. 2. Donor concentration in the central part of the germanium layer is $4 \cdot 10^{16}$ cm^{-3}, with slight uprise toward interfaces, where it reaches a peak of $9 \cdot 10^{16}$ cm^{-3}. It approximately matches the doping profile of epitaxial Ge 40x40 μm^2 Schottky diode from [3]. The aluminum electrode covers the whole area of the diode.

In the basic model of Schottky contact, which assumes a constant Schottky barrier height, the Al electrode is in direct contact with Ge. However, there is also a possibility of germanium and aluminum oxidation upon the deposition of the electrode as shown in [5]. Therefore, another model of the diode, one with a thin layer of interface oxide between Al and Ge, most likely aluminum oxide or germanium oxide, will be examined in this paper [5]. The thickness of the unspecified insulator layer has been set to 2 nm for the purpose of simulation (Fig. 3). In the diode model published in [5], the interface oxide has been modeled as a wide-bandgap semiconductor, so that the conduction mechanisms can be simplified to thermionic emission over the oxide/Ge and Al/oxide heterojunctions. Since the tunneling through the oxide is neglected, the thickness of oxide layer does not significantly impact the current, as it is defined solely by the height of barrier in conduction band. Such a transport model, which implies current flow through the conduction band of an insulator, can be used for fitting the *I-V* characteristics over the interface. Even though other

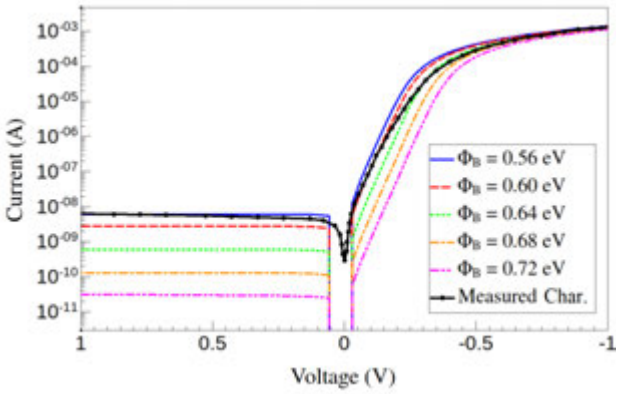


Figure 4: *I-V* characteristics of the simulated diode for different values of the Schottky barrier height (Φ_B).

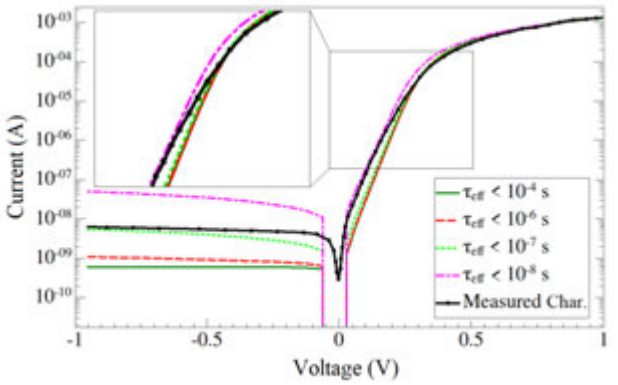


Figure 5: *I-V* characteristics of the simulated diode for different values of the effective carrier lifetime in the Ge layer (τ_{eff})

conduction mechanisms, such as Trap Assisted Tunneling (TAT), can contribute to the total current through the oxide, the model of the total current can be simplified to thermionic emission via extraction of the effective barrier height. This idea comes from the fact that most of the articles reporting barrier heights at Al/Ge interface extracted barrier values from *I-V* characteristics, without modeling possible interface oxide formation. Thus, if any other conduction mechanism on the interface exists, it has already been summed up in the thermionic model defined by the effective height of the potential barrier. In addition to this, the physical properties of the possibly existing oxide layer are unknown, so the simulation model can finally be simplified to direct Al/Ge Schottky with the effective Schottky barrier height, which accounts for all of the mentioned physical effects (Fig. 3). However, the physical model presented in [5] predicts bias-activated deep traps in the interface oxide. Ionization of these deep states captures fixed charge in the oxide, which directly impacts the height of the barrier for electrons that are injected from Ge into the Al electrode. Therefore, the Al/Ge interface and the potential interface oxide have been finally modeled by simple Al/Ge contact, but with the field-dependent, variable Schottky barrier height.

III. RESULTS AND DISCUSSION

The simulated diode is modeled in order to fit *I-V* characteristics from [3]. As TCAD simulation software predicts device currents per unit of the it's length, all characteristic currents in this work have been multiplied to match the area of the measured diode. In this section, the different mechanisms that could possibly impact the diode current are analyzed. These mechanisms include traps at Al/Ge interface, traps at Ge/Si interface, perimeter effects and the formation of a buffer oxide layer between Al and Ge. Each mechanism is discussed as a possible cause of a degraded ideality factor and high values of extracted Schottky barrier heights compared to what is expected in light of the Schottky-Mott rule.

For the purposes of this analysis, the *I-V* characteristics of the diode are divided into three different bias regions. Low-bias characteristics stand for the currents which correspond to bias values of less than 0.3 V. In this operation regime, the current is highly influenced by the Shockley-Read-Hall recombination process through the traps located in the depletion region. Mid-bias regime stands for regime in the vicinity of knee voltage, i.e. 0.3 V - 0.5 V. In this regime, the depletion region of the diode becomes narrower and the current is mostly defined by the barrier height for electrons at the Al/Ge interface, which directly corresponds to the Schottky barrier height. For higher voltages, the current is limited by the series resistance. Due to a very thin layer of epitaxial germanium and the corresponding higher carrier mobilities compared to Si, it can be assumed that the current in this region is mostly limited by the series resistance of a thick Si substrate. To reduce the size of the simulated structure, the area of the bottom-electrode equals the area of the diode. Therefore, silicon layer thickness has been varied in order to account for both contact resistances and collection of carriers by large-area bottom electrode.

A. *Impact of Schottky barrier height*

In the middle part of *I-V* characteristics, the current of a diode is mostly determined by the height of the Schottky barrier on Al/Ge interface (Fig. 4). By increasing the barrier height, the mid-bias current is exponentially reduced, as expected from [1]. However, the Schottky barrier height had a significant impact on modeled characteristics for low forward bias. *I-V* characteristics of the modeled diode for different values of the Schottky barrier height is shown in Fig. 4. The simulated characteristics show an abrupt rise around the origin due to the finite simulation step. Nevertheless, the ideal exponential part of the diode's characteristics is proportionally reduced with a higher barrier. However, the slope remains constant. This implies that the value of the zero-bias Schottky barrier height should not have a significant effect on the diode's ideality factor.

In order to fit the current for mid-forward biases, the modeled diode should have rather high values for barrier height (> 0.65 eV). To give these results some physical meaning, a comparison with literature can be made. As aluminum has a work function of 4.08 eV [21], and germanium an electron affinity of 4.00 eV [17], according to the Schottky-Mott rule, the barrier height in the conduction band should be $\Phi_B \approx 0.08\ eV$.

However, different authors report calculations of Schottky barrier heights from *I-V* characteristics of fabricated diodes to be significantly higher, e.g. $\Phi_B \approx 0.40$ - $0.64\ eV$ in [4], $\Phi_B \approx 0.7\ eV$ in [6] and $\Phi_B \approx 0.6\ eV$ in [8]. Furthermore, when calculated from *C-V* characteristics, the barrier height is

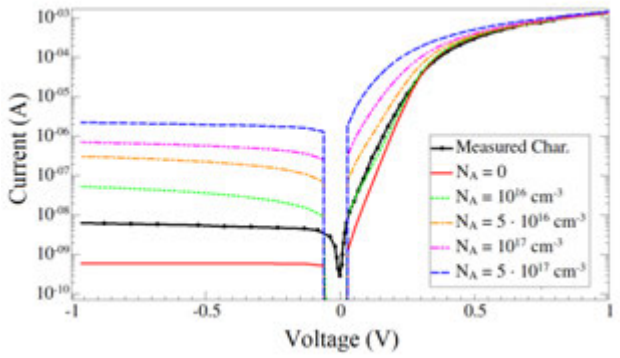


Figure 6: *I-V* characteristics of the simulated diode for different peak concentrations of mid-bandgap acceptor traps on the Al/Ge interface. The spatial distribution of traps is defined as one-sided *Gaussian* with peak on Al/Ge interface and standard deviation of 200 nm.

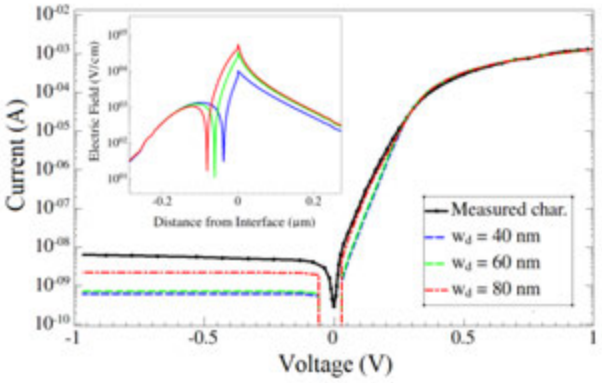


Figure 7: *I-V* characteristics of the simulated diode for different widths of possible depletion on Ge/Si interface. Inset: electric field distribution over the interface is shown for the corresponding *I-V* curves.

reported to be significantly lower, such as $\Phi_B \approx 0.34$ eV in [4]. In addition to this, while the result of *C-V* based barrier height calculations has stable values, the size of the barrier determined from *I-V* curves shows aging properties [4-5].

If the Schottky barrier height was only 0.34 eV, as calculated from the capacitance, the barrier for electrons would barely exist. Thus, for Fermi distribution of carriers at 300 K, the contact would be ohmic and no rectification would be noticed. Interestingly, all of the Schottky diodes from [3] show rectifying behaviour and Schottky characteristics only when there is no final thermal annealing process applied. For those diodes where thermal annealing was performed after the deposition of the Al electrode, characteristics showed ohmic behaviour. This could possibly indicate breaking up of the oxide layer and establishing direct contact of Al and Ge – effective reduction of the Schottky barrier to lower values (F*ermi level depinning*), such as those measured from *C-V* characteristics. However, the extraction of Schottky barrier heights in [3] results in $\Phi_B = 0.65$ eV, which would give a lower reverse saturation current and current for low-biases than measured on actual samples, as shown in Fig. 4.

B. Impact of bulk and surface SRH recombination

In order to increase the current at low biases in simulations and consequently degrade the ideality factor from ideal to the value of 1.24, as measured, the Shockley-Read-Hall doping-dependent recombination has also been included in simulations. If a high concentration of traps is assumed in the epitaxial Ge due to defective epitaxial growth directly on Si and the consequent formation of dislocations in the lattice, recombination in those traps could be the cause of ideality factor degradation. Furthermore, if recombination centers are located in the depletion region, the recombination component of the total current would likely have an ideality factor of up to 2, thus degrading the overall performance of the device. However, only by a significant boost of the SRH current component - by setting the bulk carrier lifetime of carriers to less than 10^{-8} s - the low-bias component of the current could be fitted (Fig. 5). However, for such low carrier lifetimes, the values of the reverse current would be significantly higher than as measured in reality. Furthermore, in the case of carrier lifetime being less than 10^{-8} s, mid-bias forward current is also increased. This effect could be reduced by using even higher values of the Schottky barrier (higher than 0.64 eV), which would inevitably reduce the current in the low-bias region. In addition, the effective carrier lifetime in Ge is reported to be in the range of 10^{-6}s [22] for wedge-cut unpassivated samples to 10^{-3} s for bulk material. [17] Due to the fact that it is not possible to boost bulk SRH recombination current exclusively in forward bias without overestimating the reverse current, and the fact that the effective lifetime values are significantly lower than those reported in the literature, it can be concluded that the modeling of Al/Ge interface with simple Schottky barrier and bulk SRH recombination is not sufficient to fit the measured characteristics.

C. Impact of localized SRH recombination

Another mechanism that could increase the recombination currents without over-estimating bulk Ge recombination parameters, is the introduction of additional recombination centers (traps) in specific regions of the device. Those are for e.g.: lattice dislocations, cracks in the epitaxial layer or grain boundaries, eventual presence of Al atoms deep in Ge layers or fixed charge ions. Simulation of mid-bandgap traps has been performed both at Al/Ge and Ge/Si interfaces. Traps are placed with Gaussian spatial distribution, with peak at the material interface and standard deviation of 200 nm. Characteristics of diode with additional deep-traps in the vicinity of Al/Ge interface is shown in Fig. 6. Simulations give results similar to those for

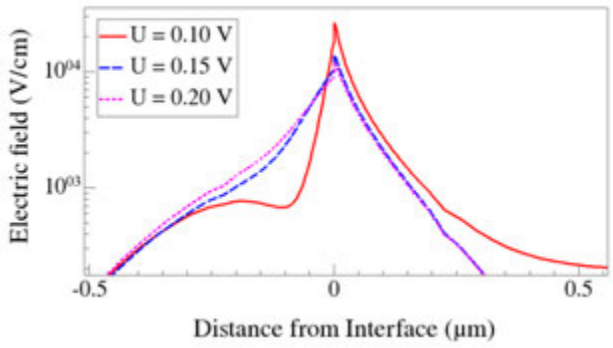


Figure 8: Electric field at Ge/Si interface for different forward bias voltages for Schottky barrier diode with one-sided Gaussian fixed charge distribution at the Ge-side of the interface. Peak concentration of fixed charge is $N_{FIX} = -4 \cdot 10^{16}$cm^{-3}. Spatial deviation of the distribution is σ_{FIX}=200 nm. Interface depletion diminishes for larger biases.

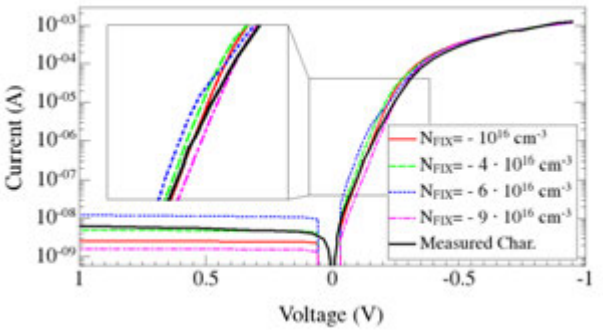


Figure 9: *I-V* characteristics of the simulated diode for different peak concentrations of fixed charges on Ge/Si interface. The spatial distribution of traps is defined as Gaussian with peak on Ge/Si interface and standard deviation of 200 nm.

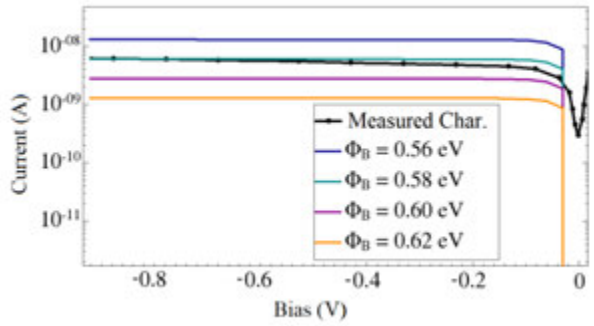


Figure 10: Reverse saturation currents for determination of zero-bias Schottky barrier height in field-dependent barrier model. The model still hasn't been utilized which can be noticed from constant modeled reverse currents. SRH carrier lifetime is doping-dependent, with maximum set to low-doped bulk-Ge values, $\tau_{eff} = 10^{-4}$ s.

simple increase of bulk Ge recombination parameters. The reason is in the fact that most of the deep-traps recombination centers are placed in the depletion region of the Schottky contact. Again, forward bias current is increased in low- and mid-bias regions, but also in reverse-bias region. The third possible cause of ideality factor degradation is recombination at Ge/Si interface. First Ge layers deposited on Si (100), up to 300 nm from Si/Ge interface, can contain dislocations due to significant strain and lattice mismatch of 4.2% between lattice constants of Ge and Si [3, 23]. Furthermore, strain at the interface could induce specific band alignment of Ge and Si, which could differ from expected alignment from heterojunction theory. [1, 24] If either heterojunction band alignments or defective Ge layer in the vicinity of interface could induce depletion at Ge/Si interface, poor ideality factor and also quite low reverse currents could be explained by the recombination in this secondary depletion layer.

The depletion region at the Ge/Si interface is firstly modeled by simple manipulation of material affinities (Fig. 7). It can be observed how low-bias current values are increased with wider depletion region at Ge/Si interface. No additional traps are needed to be simulated at Ge/Si interface, but just the basic depletion region. Increase of depletion region width at the interface increases low-bias forward currents, while at the same time preserves the values of forward mid-bias current. The reason is that the Ge/Si interface depletion region diminishes after forward biases larger than 0.2 V (Fig. 8). For biases under 0.2 V, Ge/Si depletion is dominant recombination center in the device and for larger biases it does not impact the current. The same effect is modeled without manipulation of material affinities, but with modeling of fixed charge mid-gap traps with Gaussian distribution at the Ge/Si interface. Peak values of fixed charge concentration are located at the interface and are marked in the legend of Fig. 9. Standard deviation of the distribution is 200 nm. The same effect of low-bias forward current increase, while preserving the values of mid-bias currents is achieved.

D. Impact of interface oxide formation

The next diode model, assuming the oxide layer at Al/Ge interface. has been simulated by utilizing two different approaches shown on Fig. 3. The first of them models the oxide as wide-bandgap semiconductor and defines the carrier transport as thermionic emission over the barrier in conduction band. The second model simplifies oxide to simple potential barrier at the Al/Ge interface. As suggested in [5], deep defective states in oxide could cause the change in barrier height when a free carrier is released from specific trap and leaves certain amount of space charge fixed in the oxide. This effect could potentially cause time-dependent and field-dependent change of barrier height for electrons [5,9]. In this model, barrier height for zero bias is determined from the reverse saturation current (Fig. 10) and it equals 0.59 eV. The barrier height increases with rise of forward bias, consequently lowering the slope of current characteristics and reducing ideality factor. Furthermore, bias-dependency of Schottky barrier height results with bias-dependent ideality factor. This effect is not observed in the diodes with the same technology parameters, but with Ga and B interlayers between Al and Ge (Fig. 11). By using the field dependent barrier height model, the fitted curve in forward bias finally reaches measured values (Fig. 12). In addition, a part of the applied voltage drops over the interface oxide, which could possibly explain why *C-V* measurements of barrier heights show significantly lower values compared to those extracted from *I-V* curves.

Finally, numerous samples of Al/Ge-on-Si Schottky diodes on the same wafer were characterized in [3]. Most of

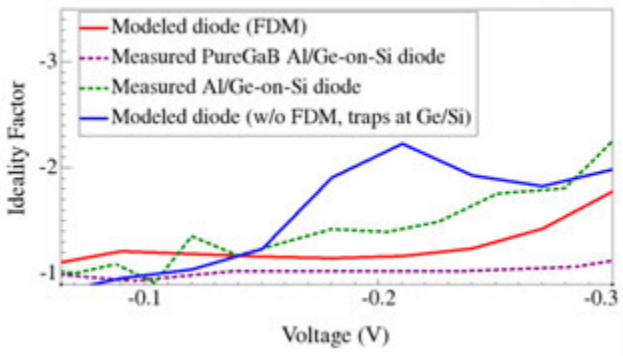


Figure 11: Ideality factors of a modeled diode with field-dependent model of Schottky barrier height (FDM) and another, with fixed charge on Ge/Si interface (peak concentration of -6 · 10^{16} cm^{-3}). Modeled diodes are compared with characterized Al/Ge-on-Si Schottky diode. Ideality factor of the equally processed Schottky diode, with additional Ga and B interlayers before sputtering of the electrode, is given as almost ideal reference (PureGaB). [3] Similar effects of ideality factor degradation can be achieved by modeling of Al/Ge interface oxide (red) or fixed charge on Ge/Si interface (blue).

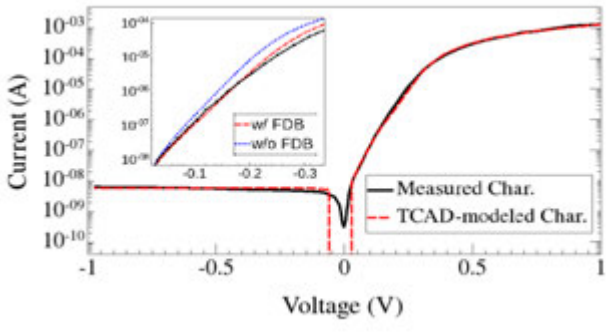


Figure 12: Final fit of the modeled diode to experimental results. Schottky barrier height is set initially to 0.59 eV. Field dependence of barrier height is modeled to reach 0.648 eV for forward bias of 1 V. Fixed charge on Ge/Si interface has peak concentration of $3 \cdot 10^{16}$ cm^{-3} and standard deviation of 200 nm. Inset of the figure shows characteristics with field-dependent barrier (FDB) model and without it.

them fully match in forward bias current and therefore have the same ideality factor. The diodes with area 40x40 μm^2 have ideality factor of 1.24, while those with area 6x6 μm^2, have ideality factor of 1.22. It could be assumed that perimeter effects which impact *I-V* characteristics should have larger effect on smaller diodes, due to higher perimeter to area ratio. In this case, measurements were showing the opposite. Possible explanation comes from the *C-V* curves [3] which show larger concentration of As dopant atoms in smaller diodes, especially in vicinity of Ge/Si interface. If there are defects, cracks in the crystal lattice, which could be modeled by fixed charge on Ge/Si interface, larger doping in epitaxial Ge layer would suppress their impact on current. If this is the case in diodes from [3], the impact of perimeter is presumed not to be dominant source of ideality factor degradation.

IV. CONLUSION

In this work, different simulations of the impact of defective states on *I-V* characteristics of Al/Ge-on-Si Schottky diodes were performed. There are three possible causes of ideality factor degradation. The first one are acceptor traps in Ge near the interface with Al. Aluminum deposition could cause p-type doping of the Ge substrate or introduce other defects in the Ge lattice. The second possible cause could be the defective states at the Si/Ge interface, which induce the depletion region at the interface. This can be explained by strain due to a mismatch between Si and Ge lattice constants, which could introduce dislocations in the Ge lattice, up to a few hundred nanometers from the Si/Ge interface. Also, the presence of Al atoms deep in the Ge layer could be introduced during processing. The final explanation for performance degradation is the formation of an interface oxide between Al and Ge, which could induce voltage dependence of the barrier height and the ideality factor. Thermal annealing could possibly enable direct contact of Al and Ge, which would result in the reduction of the barrier height and the formation of ideal ohmic contact.

REFERENCES

[1] J. Liu et al., "Ge-on-Si optoelectronics", *Thin Solid Films*, vol. 520, no. 8, pp. 3354-3360, 2012.

[2] W. Kuo, H. Hsieh, W. Chih-Hung, H. Wen-Hsiang, C. Lee and J. Chang, "High Quality GaAs Epilayers Grown on Si Substrate Using 100 nm Ge Buffer Layer", *International Journal of Photoenergy*, vol. 2016, pp. 1-5, 2016.

[3] A. Sammak, L. Qi, W. de Boer and L. Nanver, "PureGaB p+n Ge diodes grown in large windows to Si with a sub-300nm transition region", *Solid-State Electronics*, vol. 74, pp. 126-133, 2012.

[4] A. Thanailakis and D. Northrop, "Metal-germanium Schottky barriers", *Solid-State Electronics*, vol. 16, no. 12,pp.1383-1389, 1973.

[5] A. Thanailakis, D. Chan and D. Northrop, "Activation energy for the slow states in the aluminium-germanium contact", *Journal of Physics D: Applied Physics*, vol. 5, no. 10, pp. 1930-1936, 1972.

[6] D. Gajula, P. Baine, M. Modreanu, P. Hurley, B. Armstrong and D. McNeill, "Fermi level de-pinning of aluminium contacts to n-type germanium using thin atomic layer deposited layers", *Applied Physics Letters*, vol. 104, no. 1, p. 012102, 2014.

[7] S. Baek et al., "Alleviation of fermi-level pinning effect at metal/germanium interface by the insertion of graphene layers", *Applied Physics Letters*, vol. 105, no. 7, p. 073508, 2014.

[8] R. Lieten, V. Afanas'ev, N. Thoan, S. Degroote, W. Walukiewicz and G. Borghs, "Mechanisms of Schottky Barrier Control on n-Type Germanium Using Ge3N4 Interlayers", *Journal of The Electrochemical Society*, vol. 158, no. 4, p. H358, 2011.

[9] A. Thanailakis and D. Northrop, "Slow voltage-induced barrier height changes in metal-germanium surface barriers", *Journal of Physics D: Applied Physics*, vol. 4, no. 11, pp. 1776-1784, 1971.

[10] R. Tung, "The physics and chemistry of the Schottky barrier height", *Applied Physics Reviews*, vol. 1, no. 1, p. 011304, 2014.

[11] A. Sammak, L. Qi and L. Nanver, "Restricted-Access Al-Mediated Material Transport in Al Contacting of PureGaB Ge-on-Si p + n Diodes", *Journal of Electronic Materials*, vol. 44, no. 12, pp. 4676-4683, 2015.

[12] T. Nishimura, K. Kita and A. Toriumi, "A Significant Shift of Schottky Barrier Heights at Strongly Pinned Metal/Germanium Interface by Inserting an Ultra-Thin Insulating Film", *Applied Physics Express*, vol. 1, p. 051406, 2008.

[13] A. Sammak, M. Aminian, L. Qi, W. de Boer, E. Charbon and L. Nanver, "(Invited) Fabrication of Pure-GaB Ge-on-Si Photodiodes for Well-Controlled 100-pA-Level Dark Currents", *ECS Transactions*, vol. 64, no. 6, pp. 737-745, 2014.

[14] T. Knezevic, X. Liu, E. Hardeveld, T. Suligoj and L. Nanver, "Limits on Thinning of Boron Layers With/Without Metal Contacting in PureB Si (Photo)Diodes", *IEEE Electron Device Letters*, vol. 40, no. 6, pp. 858-861, 2019.

[15] D. Klaassen, "A unified mobility model for device simulation", *International Technical Digest on Electron Devices*, 1990.

[16] Synopsys, *Sentaurus Device User Guide Version* N-2017.09. Mountain View, CA, USA: Synopsys, 2017.

[17] "Ge - Germanium", *Ioffe.ru*, 2020. [Online]. Available: http://www.ioffe.ru/SVA/NSM/Semicond/Ge/

[18] C. Crowell, "The Richardson constant for thermionic emission in Schottky barrier diodes", *Solid-State Electronics*, vol. 8, no. 4, pp. 395-399, 1965.

[19] B. Van Zeghbroeck, "Schottky diode current", *Ecee.colorado.edu*, 2011. [Online]. Available: https://ecee.colorado.edu/~bart/book/book/chapter3/ch3_4.htm#3_4_2.

[20] S. Sze and K. Ng, *Physics of semiconductor devices*. John Wiley & Sons, 2006.

[21] "Work Functions for Photoelectric Effect", *Hyperphysics.phy-astr.gsu.edu*, [Online]. Available: http://hyperphysics.phy-astr.gsu.edu/hbase/Tables/photoelec.html

[22] E. Gaubas and J. Vanhellemont, "Comparative Study of Carrier Lifetime Dependence on Dopant Concentration in Silicon and Germanium", *Journal of The Electrochemical Society*, vol. 154, no. 3, p. H231, 2007.

[23] J. Michel, J. Liu and L. Kimerling, "High-performance Ge-on-Si photodetectors", *Nature Photonics*, vol. 4, no. 8, pp. 527-534, 2010.

[24] O. Nayfeh, C. Chleirigh, J. Hennessy, L. Gomez, J. Hoyt and D. Antoniadis, "Design of Tunneling Field-Effect Transistors Using Strained-Silicon/Strained-Germanium Type-II Staggered Heterojunctions", *IEEE Electron Device Letters*, vol. 29, no. 9, pp. 1074-1077, 2008.

Material and device properties of bismuthene nanoribbons from multi-orbital quantum transport simulations

M. Poljak*
Computational Nanoelectronics Group
Micro and Nano Electronics Laboratory, Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, HR 10000, Croatia
*E-mail: mirko.poljak@fer.hr

***Abstract*—A multi-orbital tight-binding Hamiltonian is employed to study the electronic, transport and ballistic device properties of bismuthene nanoribbons (BiNRs) within the non-equilibrium Green's function (NEGF) formalism. We report the bandgap-width dependence that demonstrates an increase from 0.57 eV to 0.84 eV when the BiNR width is downscaled from 10.3 nm to 1.1 nm. The relevance of BiNRs for logic device applications at the nanoscale is further analyzed by studying the width-dependence of the ON and OFF state conductance and their ratio that reaches ~10^7 in 1.1 nm-wide BiNRs. Moreover, BiNR FET simulations reveal ballistic ON-state drain currents of up to 5.1 mA/μm, surpassing those of graphene nanoribbon FETs by 38% at 0.5 V supply voltage.**

***Keywords*—bismuthene; nanoribbon; quantum transport; atomistic simulation; bandstructure; FET**

I. INTRODUCTION

Two-dimensional (2D) materials have gained a tremendous interest due to their unique material properties, including superior flexibility, absence of dangling bonds, adjustable bandgap, atomic thickness, etc. [1]. These features are promising for numerous potential applications ranging from energy harvesting, over photodetectors, to nanoscale field-effect transistors (FETs) [2], [3]. Unlike 2D nanosheets based on group-IVA materials (e.g. graphene and silicene), group-VA elements result in 2D materials with considerable bandgaps, which makes them promising for nanoscale electronic device applications [4]. Bismuthene, a monolayer of hexagonally arranged Bi atoms (Fig. 1), exhibits a bandgap of ≈0.5 eV both in the free-standing buckled geometry and in the on-substrate planar geometry [5]. The latter, achievable experimentally in bismuthene-on-SiC heterostructures [6], is additionally promising because it could enable a quantum spin Hall (QSH) system thus paving the way towards spin-based nanodevices at room temperature [5].

An additional possibility for adjusting the material and device properties at the nanoscale is available by patterning bismuthene sheet into bismuthene nanoribbons (BiNRs), using similar processing as reported for graphene nanoribbons (GNRs) [7], [8] (illustration provided in Fig. 1). In these nanostructures, quantum confinement along the nanoribbon width direction is expected to considerably alter its electronic, transport and device properties [9]–[11]. However, while the nanoribbon geometry is well-studied for graphene, phosphorene [12], [13], silicene and germanene [14], [15], very little is known about BiNR characteristics relevant for applications in nanoelectronics.

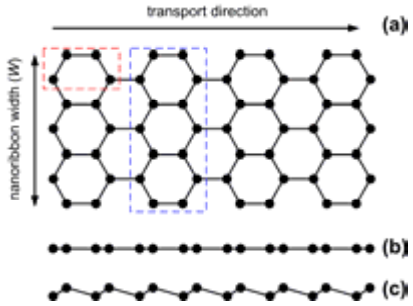


Fig. 1. (a) Top-view of an armchair nanoribbon of a 2D material with a hexagonal crystal lattice. Side-views of (b) flat graphene and (c) buckled bismuthene.

In this paper, we employ multi-orbital Hamiltonians and atomistic quantum transport simulations in order to study the bandstructure and transport properties of BiNRs, and ballistic device characteristics of BiNR FETs (illustrated in Fig. 2a). At several points in this work, results obtained for BiNRs are compared to GNRs with the same bandgap, which allows a direct performance comparison. Most importantly, we find that width-adjustment in BiNRs can lead to a bandgap of up to 0.84 eV, and that BiNR FETs outperform GNR FETs by almost 40% in terms of the ON-state current driving capabilities.

II. MODELING AND SIMULATION APPROACH

A. Nanoribbon Hamiltonian

Bismuthene exhibits a (buckled or planar) honeycomb crystal lattice in which a 2-atom (A and B) unit cell can be identified and used to construct the total BiNR Hamiltonian (H). In this work, the σ-band model from [5] is used that includes p_x and p_y orbitals, with the following basis functions

$$\left|p_{x\uparrow}^{A}\right\rangle,\left|p_{y\uparrow}^{A}\right\rangle,\left|p_{x\uparrow}^{B}\right\rangle,\left|p_{y\uparrow}^{B}\right\rangle;$$
$$\left|p_{x\downarrow}^{A}\right\rangle,\left|p_{y\downarrow}^{A}\right\rangle,\left|p_{x\downarrow}^{B}\right\rangle,\left|p_{y\downarrow}^{B}\right\rangle, \quad (1)$$

i.e. spin is taken into account in order to include the on-site (atomic) intrinsic spin-orbit coupling (SOC). The effective Hamiltonian for the bismuthene-on-SiC system is given by

$$H_{eff}^{\sigma\sigma} = \begin{bmatrix} H_{\uparrow\uparrow}^{\sigma\sigma} & H_{\uparrow\downarrow}^{\sigma\sigma} \\ H_{\downarrow\uparrow}^{\sigma\sigma} & H_{\downarrow\downarrow}^{\sigma\sigma} \end{bmatrix}, \quad (2)$$

where $H_{\uparrow\downarrow}^{\sigma\sigma} = (H_{\downarrow\uparrow}^{\sigma\sigma})^{\dagger}$ is a null-matrix because Rashba SOC is assumed to be negligible, while

$$H_{\uparrow\uparrow(\downarrow\downarrow)}^{\sigma\sigma} = \begin{bmatrix} 0 & \mp i\lambda_{SO} & h_{xx}^{AB} & h_{xy}^{AB} \\ \dagger & 0 & h_{yx}^{AB} & h_{yy}^{AB} \\ \dagger & \dagger & 0 & \mp i\lambda_{SO} \\ \dagger & \dagger & \dagger & 0 \end{bmatrix}, \quad (3)$$

where h parameters are obtained from Slater-Koster integrals depending on the high-symmetry point for which the material is analyzed, and λ_{SO} defines SOC strength [5]. In this work, the Hamiltonian is found at the K point since a fundamental gap opening in bismuthene-on-SiC system occurs at this point.

The 2-atom unit cell Hamiltonian in (2) is therefore an 8×8 matrix, which is employed to construct the Hamiltonian of the entire nanoribbon. We choose a 4-atom-cell as a unit cell (red rectangle in Fig. 1a) to construct one super-cell along BiNR width (blue rectangle in Fig. 1a) that contains NW 4-cells. The super-cell Hamiltonian is designated H_W, while the interaction between neighboring super-cells along BiNR length is designated H_L. For a given BiNR width, its bandstructure (i.e. E-k diagram, where k is the wave-vector in the transport direction) is obtained by calculating the k-dependent eigen-values of

$$H_{E-k}(k) = H_W + H_L \exp(ika) + H_L^{\dagger} \exp(-ika), \quad (4)$$

for each wave-number k in $[-\pi/a, +\pi/a]$ [16]. The total Hamiltonian of the BiNR is obtained by repeating the column super-cells H_W along the BiNR length NL times, so the final total nanoribbon H is of the size $(16\times NW\times NL) \times (16\times NW\times NL)$.

B. Quantum Transport

Non-equilibrium Green's function (NEGF) formalism [16], [17] is used to directly solve the Schrödinger's equation with open boundary conditions (OBCs). Retarded Green's function (GF) of the device is

$$G^R(E) = \left[\left(E + i0^+\right)I - H - \Sigma_S^R(E) - \Sigma_D^R(E)\right]^{-1}, \quad (5)$$

where H is the total BiNR Hamiltonian, and Σ matrices designate the retarded contact self-energies that account for OBCs to the two contacts (source and drain). Sancho-Rubio method is implemented to find these retarded self-energies [18]. The retarded and advanced GF of the device (with $G^A = (G^R)^{\dagger}$) are then used to find the transmission function, $T_{SD}(E)$, and density of states, $D(E)$

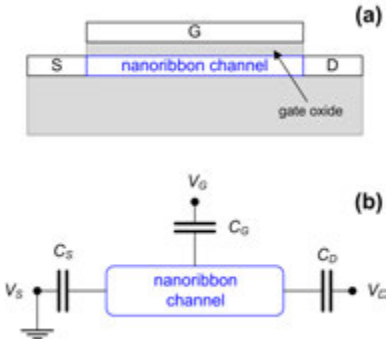


Fig. 2. (a) Illustration of a BiNR/GNR FET in which the channel is a monolayer nanoribbon. (b) Capacitances used in the Top-of-the-Barrier (ToB) FET model.

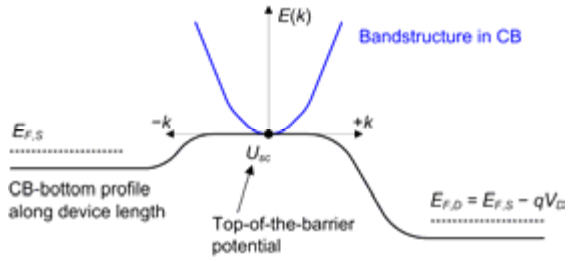


Fig. 3. Profile of the conduction band along the channel length with quasi Fermi levels in the source and drain. The E-k diagram is superimposed at the ToB to illustrate the inclusion of bandstructure in FET simulation.

[12], [14], [19]. The conductance of the nanoribbon at $T = 300$ K is calculated using

$$G = G_0 \int_0^{\infty} T_{SD}(E)\left(-\partial f(E - E_F)/\partial E\right) dE, \quad (6)$$

where $f(E - E_F)$ is the Fermi-Dirac function, $G_0 = q^2/h$ (factor 1 for BiNRs, and 2 for GNRs), and h is the Planck's constant, with E_F of 0 eV and 1 eV for the OFF- (G_{OFF}) and ON-state (G_{ON}) conductance, respectively.

C. Top-of-the-Barrier FET Model

Top-of-the-barrier (ToB) model is a simple approach for obtaining ballistic FET characteristics from quantum-transport simulations [20]–[22]. Within ToB, the focus is on a single point in the channel, i.e. top-of-the-barrier potential between the source and drain regions (U_{sc}), while the model still includes an atomistic material resolution and its full bandstructure. The ToB potential is found from

$$U_{sc} = U_{sc0} - q\left(\alpha_G V_G + \alpha_S V_S + \alpha_D V_D\right) + \frac{q}{C_{ox}}\left(Q_{ch} - Q_{ch}(V_G = 0)\right), \quad (7)$$

where α parameters describe capacitive coupling strengths to the channel (see illustration in Fig. 2b). For example, $\alpha_G = C_G/(C_S + C_D + C_G)$, and usually $\alpha_G = 1$ for a well-tempered FET. In (7), the C_{ox} is the gate oxide capacitance, and Q_{ch} is the inversion (mobile) charge in the channel. The Q_{ch} depends on the positions of U_{sc} and quasi Fermi levels in source and drain ($E_{F,S}$ and $E_{F,D}$,

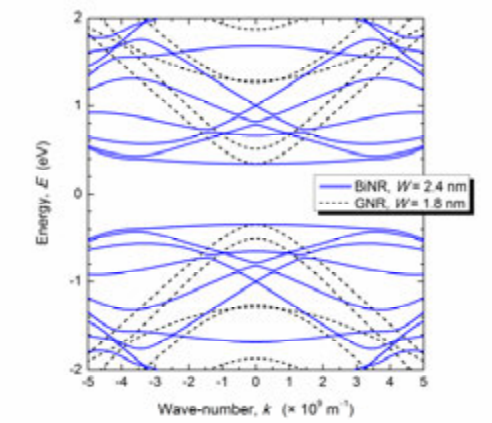


Fig. 4. Bandstructure of 2.4 nm-wide BiNR and 1.8 nm-wide GNR showing the same bandgap value.

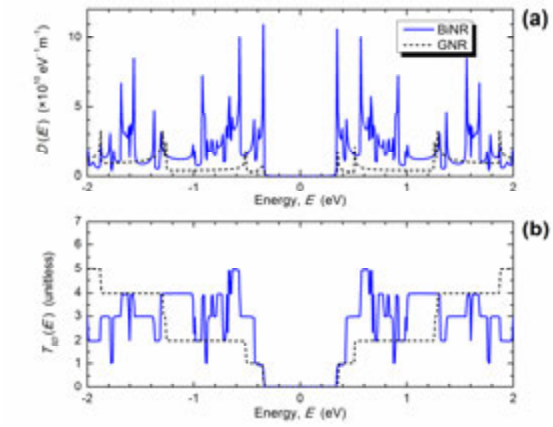


Fig. 5. (a) Density of states and (b) transmission function for 2.4 nm-wide BiNR and 1.8 nm-wide GNR.

respectively), as illustrated in Fig. 3, and is found using $D(E)$ obtained in NEGF simulations, i.e.

$$Q_{ch} = Q_{ch,S} + Q_{ch,D}, \tag{8}$$

$$Q_{ch,S(D)} = \frac{1}{2}\int_{-\infty}^{+\infty} D(E-U_{sc}) f(E-E_{F,S(D)})dE, \tag{9}$$

with $E_{F,D} = E_{F,S} - qV_D$. After convergence is achieved for U_{sc} and Q_{ch}, the drain current (I_D) is calculated using Landauer formula [16] and $T_{SD}(E)$ found in NEGF simulations. The ToB model does not include any kind of tunneling, therefore, its relevance is limited only to the above-threshold operation region, i.e. to the assessment of the ON-state FET performance [20], [22].

III. Results and Discussions

A. Electronic and Transport Properties

Figure 4 reports the dispersion of 2.4 nm-wide BiNR and 1.8 nm-wide GNR that exhibit the same E_G of 0.67 eV. The E-k diagram is plotted against the wave-number k in the transport direction, and gives information about the electronic properties and transport effective mass. The GNR bandstructure contains simple parabolic bands, with an effective mass of $m^*/m_0 \approx 0.1$ in the lowest subband of the CB. In contrast, the BiNR bandstructure is

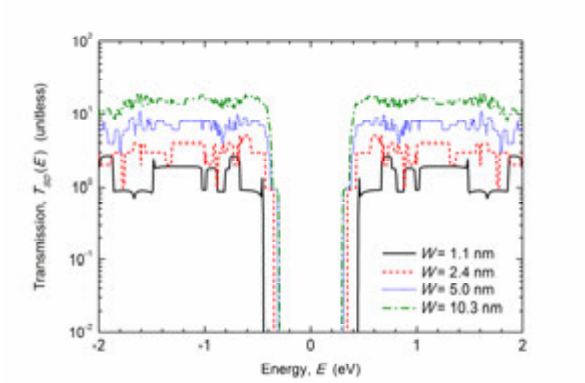


Fig. 6. Transmission function versus energy in 20 nm-long BiNRs of various widths (1.1 nm, 2.4 nm, 5.0 nm, and 10.3 nm).

complex with intertwined regions of positive- and negative-velocity subbands, which should result in a non-trivial energy-dependence of the transmission function. Moreover, $m^*/m_0 \approx 10$ (≈1) in the lowest (second lowest) BiNR CB subband is much heavier than in the GNR with the same E_G.

This difference in E-k diagrams leads to significant differences in both the available DOS and transport properties between BiNRs and GNRs, as reported in Fig. 5. Figure 5a demonstrates that a heavier m^* and a greater number of bands leads to a much higher DOS in BiNRs, which is expected to translate into higher inversion charge densities in the channel for the same voltage. In addition, higher number of positive-velocity bands in BiNRs, especially for energies between ≈0.4 eV ≈1.2 eV, leads to a significantly higher transmission in BiNRs in comparison to GNRs, as shown in Fig. 5b. These bandstructure results indicate the following: FETs based on BiNRs are expected to have a lower threshold voltage, higher charge densities, and higher drain currents that their GNR FET counterparts. This topic is discussed in Section III.B, while the rest of Section III.A is dedicated to width-dependent properties of BiNRs.

The impact of width-scaling on the transmission in BiNRs is reported in Fig. 6 for a nanoribbon width in the range from 1.1 nm to 10.3 nm. As the width decreases, transmission generally decreases due to reduced number of conducting channels/subbands available in the nanoribbon, and bandgap increases. The former effect should negatively impact the ON-state FET performance since W-scaling decreases transmission, whereas both the decreasing transmission and increasing bandgap should improve the OFF state performance.

In order to assess the potential device performance from BiNR material properties, in Fig. 7 we report the width-dependence of bandgap (Fig. 7a), G_{ON} (Fig. 7b), G_{OFF} (Fig. 7c), and ON-OFF conductance ratio (Fig. 7d) as a figure of merit for intrinsic capability for digital switching. As reported in Fig. 7a, E_G increases from 0.57 eV to 0.84 eV as the width decreases from 10.3 nm to 1.1 nm, indicating that the examined BiNRs show appropriate bandgap values for FET applications. Furthermore, these results show that BiNR with $W > 5$ nm are resilient to quantum confinement effects in terms of the bandgap. In contrast, G_{ON} monotonically decreases in

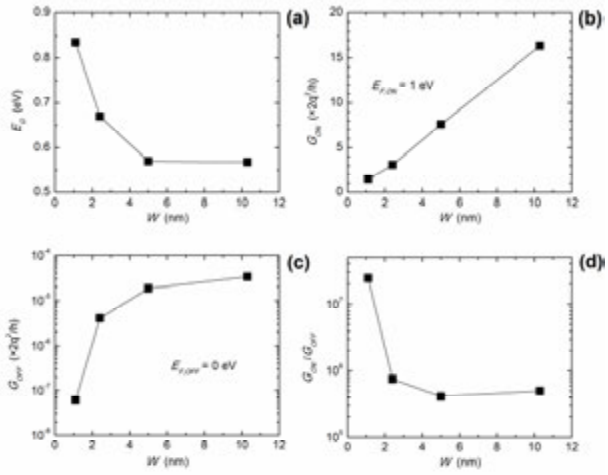


Fig. 7. Nanoribbon width dependence of (a) bandgap, (b) ON-state conductance, (c) OFF-state conductance, and (d) ON-OFF conductance ratio in 20 nm-long BiNRs. $E_{F,ON} = 1$ eV, $E_{F,OFF} = 0$ eV.

the given W range, from $16.4G_0$ down to $1.5G_0$, demonstrating a significantly reduced ON-state performance in narrower BiNRs (Fig. 7b). Similarly, as shown in Fig. 7c, the decrease of G_{OFF} is even more pronounced as it changes by approximately two orders of magnitude. Specifically, when the nanoribbon width is scaled from 10.3 nm down to 1.1 nm, G_{OFF} decreases from 3.3×10^{-5} to 6×10^{-8}. At the same time, G_{ON}/G_{OFF} reported in Fig. 7d increases in narrower BiNRs due to strong G_{OFF} suppression. The G_{ON}/G_{OFF} ratio values increase from $\approx5\times10^{5}$ to $\approx2\times10^{7}$, indicating the ability of BiNRs to appropriately switch between the ON and OFF state in digital logic device applications.

B. Ballistic BiNR FET Performance

The ToB model is applied on atomistic NEGF results in order to obtain ballistic FET characteristics. Figure 3 illustrates the CB-bottom profile along the FET length in the source, channel and drain regions. The NEGF provides $D(E)$ that is used for all three regions, while $T_{SD}(E)$ is used for the nanoribbon channel region and drain current calculation. Gate oxide is a 1 nm-thick layer of SiO_2 ($\varepsilon_r = 3.9$). The S/D doping equals 0.001 molar fraction of the areal density of the crystal (i.e. $\approx2\times10^{12}$ cm^{-2} for BiNR, and $\approx6\times10^{12}$ cm^{-2} for GNR), whereas the channel is undoped. The $E_{F,S}$ level is set such that charge neutrality in S/D regions is achieved self-consistently. Within the ToB model, we set $\alpha_S = 0$, $\alpha_G = 0.97$, $\alpha_D = 0.03$, in order to introduce a small contribution (perturbation of the barrier) from the drain electrode. Gate and drain voltage is biased from 0 V up to 0.5 V, where the latter is foreseen as a supply voltage level for future extremely scaled CMOS technology generations.

Figure 8a reports the V_G-dependence of mobile electron density in the channel at $V_D = 0.5$ V for the 2.4 nm-wide BiNR and 1.8 nm-wide GNR (with the same bandgap of 0.67 eV, see Fig. 4). The BiNR delivers a significantly higher amount of Q_{ch} in comparison to the GNR, due to much higher DOS in the BiNR for the same E_G. While a maximum Q_{ch} of 2.2×10^{13} cm^{-2} is achievable in GNRs, BiNR can provide up to 8.2×10^{13} cm^{-2} of

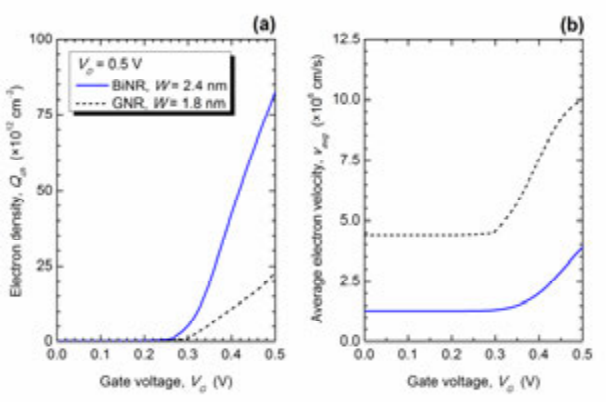


Fig. 8. (a) Electron density in the channel and (b) average electron velocity at the ToB versus gate voltage. $V_D = 0.5$ V.

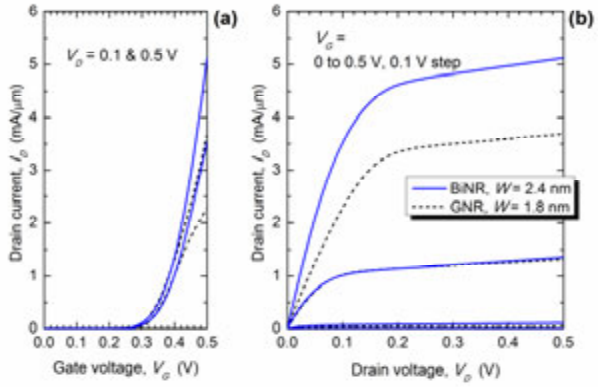


Fig. 9. Comparison of (a) transfer and (b) output characteristics of BiNR and GNR FETs.

electron charge for the same bias, i.e. ≈3.7× more than the GNR. Another consequence of the lower DOS in GNRs is that GNR FET has a higher threshold voltage than BiNR FET, i.e. 0.30 V and 0.26 V, respectively. While BiNRs are beneficial in terms of the amount of mobile charge that can be generated in the channel, the GNRs exhibit higher average velocities, as reported in Fig. 8b. Maximum achievable v_{avg} equals 1×10^{7} cm/s in the GNR and 3.9×10^{6} cm/s in the BiNR (≈2.6× higher in GNRs). The higher v_{avg} is caused by the lighter effective mass in GNRs, which is discussed briefly when the E-k diagrams in Fig. 4 are analyzed.

The net result of different properties of Q_{ch} and v_{avg} in BiNRs and GNRs is the higher drain current in BiNR FETs, as reported in the transfer (Fig. 9a) and output (Fig. 9b) current-voltage (I-V) characteristics. The ON state current I_{ON} (I_D at $V_D = V_G = 0.5$ V) equals 5.1 mA/μm in the BiNR FET and 3.7 mA/μm in the GNR FET, i.e. BiNR FET outperforms the GNR counterpart by ≈38%. The high milli-ampere driving currents are a consequence of normalization by the nanoribbon width, which is on the nanometer scale. The absolute I_{ON} values are quite modest, 12.3 μA and 6.8 μA for the BiNR and GNR FET, respectively, with corresponding maximum trans-conductance (g_m) of only 112 μA/V and 55 μA/V. As for

the output *I-V* characteristics, GNR FETs exhibit less impact by increasing V_D in the saturation region, leading to a higher output resistance (r_{out}) of 577 kΩ in the GNR FET than in the BiNR FET where $r_{out} = 286$ kΩ. The combined differences between g_m and r_{out} in these two FET structures lead to an approximately the same intrinsic gain, $A_i = g_m \times r_{out}$, that equals ≈32.

IV. Conlusions

We employ atomistic quantum transport simulations in combination with multi-orbital Hamiltonians to study the electronic, transport and ballistic device properties of BiNRs and compare them with those of GNRs with the same bandgap ($E_G = 0.67$ eV, i.e. 2.4 nm-wide BiNR and 1.8 nm-wide GNR). The BiNR gives approx. 3.7× more inversion charge and approx. 2.6× lower average velocity at the same ON-state bias when compared to the GNR. These differences lead to a lower threshold voltage (0.26 V vs. 0.30 V) and higher ON-state drain current (5.1 mA/μm vs. 3.7 mA/μm) in the BiNR FET than in the GNR FET. While the BiNR FET exhibits a considerably larger transconductance, GNR FET has a significantly higher output resistance, leading to approx. equal intrinsic gains of ≈32 in both nanoribbon-based transistors. Therefore, BiNRs with their peculiar bandstructure seem promising for nanoscale device applications, while further work is needed to properly address the OFF-state operation by including direct S-D tunneling in the ToB model.

Acknowledgments

This work is sponsored by the Croatian Science Foundation (HRZZ) under the project CONAN2D (UIP-2019-04-3493).

References

[1] N. Briggs *et al.*, "A roadmap for electronic grade 2D materials," *2D Mater.*, vol. 6, no. 2, p. 022001, Jan. 2019, doi: 10.1088/2053-1583/aaf836.

[2] F. Schwierz, "Graphene transistors," *Nature Nanotechnology*, vol. 5, no. 7, pp. 487–496, May 2010, doi: 10.1038/nnano.2010.89.

[3] F. Schwierz, J. Pezoldt, and R. Granzner, "Two-dimensional materials and their prospects in transistor electronics," *Nanoscale*, vol. 7, no. 18, pp. 8261–8283, Apr. 2015, doi: 10.1039/C5NR01052G.

[4] S. Zhang *et al.*, "Recent progress in 2D group-VA semiconductors: from theory to experiment," *Chem. Soc. Rev.*, vol. 47, no. 3, pp. 982–1021, Feb. 2018, doi: 10.1039/C7CS00125H.

[5] G. Li *et al.*, "Theoretical paradigm for the quantum spin Hall effect at high temperatures," *Phys. Rev. B*, vol. 98, no. 16, p. 165146, Oct. 2018, doi: 10.1103/PhysRevB.98.165146.

[6] F. Reis *et al.*, "Bismuthene on a SiC substrate: A candidate for a high-temperature quantum spin Hall material," *Science*, vol. 357, no. 6348, pp. 287–290, Jul. 2017, doi: 10.1126/science.aai8142.

[7] M. Y. Han, B. Özyilmaz, Y. Zhang, and P. Kim, "Energy Band-Gap Engineering of Graphene Nanoribbons," *Phys. Rev. Lett.*, vol. 98, no. 20, p. 206805, May 2007, doi: 10.1103/PhysRevLett.98.206805.

[8] X. Li, X. Wang, L. Zhang, S. Lee, and H. Dai, "Chemically Derived, Ultrasmooth Graphene Nanoribbon Semiconductors," *Science*, vol. 319, no. 5867, pp. 1229–1232, Feb. 2008, doi: 10.1126/science.1150878.

[9] D. Basu, M. J. Gilbert, L. F. Register, S. K. Banerjee, and A. H. MacDonald, "Effect of edge roughness on electronic transport in graphene nanoribbon channel metal-oxide-semiconductor field-effect transistors," *Applied Physics Letters*, vol. 92, no. 4, pp. 042114-042114–3, Jan. 2008, doi: doi:10.1063/1.2839330.

[10] A. Betti, G. Fiori, and G. Iannaccone, "Atomistic Investigation of Low-Field Mobility in Graphene Nanoribbons," *IEEE Transactions on Electron Devices*, vol. 58, no. 9, pp. 2824–2830, Sep. 2011, doi: 10.1109/TED.2010.2100045.

[11] M. Poljak, T. Suligoj, and K. L. Wang, "Influence of substrate type and quality on carrier mobility in graphene nanoribbons," *Journal of Applied Physics*, vol. 114, no. 5, p. 053701, Aug. 2013, doi: 10.1063/1.4817077.

[12] M. Poljak and T. Suligoj, "Immunity of electronic and transport properties of phosphorene nanoribbons to edge defects," *Nano Res.*, vol. 9, no. 6, pp. 1723–1734, Jun. 2016, doi: 10.1007/s12274-016-1066-1.

[13] M. Poljak and T. Suligoj, "The Potential of Phosphorene Nanoribbons as Channel Material for Ultrascaled Transistors," *IEEE Trans. Electron Devices*, vol. 65, no. 1, pp. 290–294, Jan. 2018, doi: 10.1109/TED.2017.2771345.

[14] M. Poljak, "Intrinsic Capabilities for Digital Switching of Silicene Nanoribbons With Edge Defects," *IEEE Transactions on Electron Devices*, vol. 67, no. 1, pp. 354–359, Jan. 2020, doi: 10.1109/TED.2019.2950967.

[15] M. Poljak, "Electron Mobility in Defective Nanoribbons of Monoelemental 2D Materials," *IEEE Electron Device Letters*, vol. 41, no. 1, pp. 151–154, Jan. 2020, doi: 10.1109/LED.2019.2952661.

[16] S. Datta, *Quantum Transport: Atom to Transistor*, 2nd ed. New York, NY, USA: Cambridge University Press, 2005.

[17] M. Luisier, "Quantum transport beyond the effective mass approximation," Doctoral Thesis, ETH Zurich, 2007.

[18] M. P. L. Sancho, J. M. L. Sancho, J. M. L. Sancho, and J. Rubio, "Highly convergent schemes for the calculation of bulk and surface Green functions," *J. Phys. F: Met. Phys.*, vol. 15, no. 4, p. 851, 1985, doi: 10.1088/0305-4608/15/4/009.

[19] M. Poljak, E. B. Song, M. Wang, T. Suligoj, and K. L. Wang, "Influence of Edge Defects, Vacancies, and Potential Fluctuations on Transport Properties of Extremely Scaled Graphene Nanoribbons," *IEEE Trans. Electron Devices*, vol. 59, no. 12, pp. 3231–3238, Dec. 2012, doi: 10.1109/TED.2012.2217969.

[20] A. Rahman, Jing Guo, S. Datta, and M. S. Lundstrom, "Theory of ballistic nanotransistors," *IEEE Transactions on Electron Devices*, vol. 50, no. 9, pp. 1853–1864, Sep. 2003, doi: 10.1109/TED.2003.815366.

[21] G. Liang, N. Neophytou, D. E. Nikonov, and M. S. Lundstrom, "Performance Projections for Ballistic Graphene Nanoribbon Field-Effect Transistors," *IEEE Transactions on Electron Devices*, vol. 54, no. 4, pp. 677–682, Apr. 2007, doi: 10.1109/TED.2007.891872.

[22] S. Kaneko, H. Tsuchiya, Y. Kamakura, N. Mori, and M. Ogawa, "Theoretical performance estimation of silicene, germanene, and graphene nanoribbon field-effect transistors under ballistic transport," *Appl. Phys. Express*, vol. 7, no. 3, p. 035102, Mar. 2014, doi: 10.7567/APEX.7.035102.

Single-band quantum transport study of resonant tunneling diodes based on silicene nanoribbons

M. Mihaljević, M. Širić, and M. Poljak*
Computational Nanoelectronics Group
Micro and Nano Electronics Laboratory, Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, HR 10000, Croatia
*Corresponding author: mirko.poljak@fer.hr

***Abstract*—The 1D quantum transport simulations based on a single-band effective-mass Hamiltonian are employed to study resonant tunneling diodes (RTDs) based on width-engineered silicene nanoribbons (SiNRs). Numerical analysis is used to assess the peak and valley current and voltage, peak-to-valley current ratio (PVCR), mean negative differential conductance (NDC), maximum output power (P_{MAX}) and intrinsic cut-off frequency (f_c), for different dimensions of device regions. We find that a typical nanoscale SiNR-based RTD could potentially offer a peak current of ~8 mA/μm^2, PVCR of 1.6, NDC of ~30 mS/μm^2, P_{MAX} of ~100 μW/μm^2, and f_c of ~3 THz.**

Keywords—silicene nanoribbon; resonant tunneling diode; quantum transport; NEGF

I. INTRODUCTION

Resonant tunneling diode (RTD) is an electronic element with a negative differential conductance region in its current-voltage (*I-V*) characteristics. Its most significant feature is the ability to provide terahertz (THz) oscillations, which could be used in many areas such as high-speed wireless communications, medical imaging, detection of hidden objects, etc. [1]–[3]. The RTD structure contains two potential barriers and one quantum well, in which the applied bias modifies the potential in such way that resonant tunneling can occur through the quasi-bound states (QBSs) in the well. The RTDs are mainly fabricated as vertical heterostructures, in which the usage of different materials results in creating barriers and quantum wells due to different bandstructure properties [4]–[6]. Another method of producing RTDs is by employing a lateral structure of a single material [7]–[10]. In this case, two-dimensional (2D) materials such as graphene are especially promising due to excellent transport properties, atomically-thin structure, and the possibility of patterning the material into nanoribbons where its properties can be adjusted by width-engineering [11]–[14]. Silicene is a 2D material similar to graphene with an added benefit of being compatible with the existing silicon-based semiconductor electronics industry [15]–[18]. While RTDs based on graphene nanoribbons (GNRs) have been studied previously [8], [19], [20], little is known about the potential performance of RTD based on silicene nanoribbons (SiNRs).

In this work, we evaluated the performance of SiNR-based RTDs using quantum transport simulations based on the non-equilibrium Green's function (NEGF) formalism. We analyzed the impact of different barrier and quantum well lengths on RTD figures of merit (FOMs) such as peak and valley current (I_P, I_V), peak-to-valley current ratio (PVCR), peak and valley voltage (V_P, V_V), mean negative differential conductance (G_{AVG}), and their influence on the maximum output power (P_{MAX}) and intrinsic cut-off frequency (f_c). We demonstrate that ~mA/μm^2 peak currents, ~100 μW/μm^2 output power, and ~THz frequency limits are achievable in RTDs based on width-engineered SiNRs.

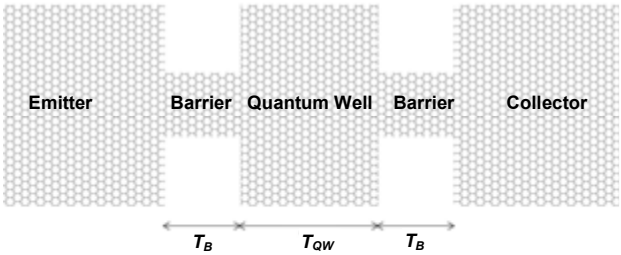


Fig. 1. Illustration of the SiNR-based RTD with two barriers and a single quantum well.

II. MODELING AND SIMULATION

A. Quantum transport and Poisson equation

A single-band effective-mass Hamiltonian and 1D quantum transport modeling based on the NEGF formalism are employed to study the performance of SiNR RTDs. Within this approach, SiNRs of various widths are used for the contact, barrier and quantum well regions of the RTD (see illustration in Fig. 1). In turn, various RTD regions are represented by different bandgap (E_G) and effective mass (m^*) values achievable in SiNRs [21]. Although a fully atomistic approach is necessary to capture all relevant physical effects, this study serves as a preliminary 1D analysis that explores the possible design space of lateral SiNR-based RTDs before undergoing more complex 2D atomistic NEGF simulations in future work. The barrier region is a 1.5 nm-wide SiNR with $E_G = 0.6$ eV and $m^* = 0.12 \times m_0$ (its length is designated T_B). The contact and quantum-well regions are a 5.5 nm-wide SiNR with $E_G = 0.2$ eV and $m^* = 0.16 \times m_0$ (quantum well length is designated T_{QW}), resulting in a tunnel barrier of 0.2 eV. Contact regions are 20 nm long.

The central quantity in the NEGF formalism [22]–[24] is the retarded Green's function calculated as

$$G^R(E) = \left[\left(E + i0^+\right)I - H - \Sigma_1^R(E) - \Sigma_2^R(E)\right]^{-1}, \quad (1)$$

where H is the position-dependent effective-mass Hamiltonian expressed in the finite difference basis, and Σ matrices designate the retarded contact self-energies that account for the two contacts (1 and 2, emitter and collector). The diode current is found from the Landauer formula

$$I = \frac{2q}{h}\int_{-\infty}^{+\infty} T(E)\left[f_1(E) - f_2(E)\right]dE, \quad (2)$$

where q is the unit charge, h is the Planck's constant, $T(E)$ is the transmission function obtained by using the retarded and advanced Green's functions, and $f_1(E)$ and $f_2(E)$ are Fermi-Dirac functions in the emitter and collector contacts, respectively. More details can be found in [23].

The electrostatic potential profile is obtained by solving the Poisson equation within the Thomas-Fermi (semi-classical) approximation, while the influence of the substrate and the charge inside the quantum well are neglected. This approach speeds-up the simulation considerably. However, Poisson-NEGF loop is not solved self-consistently since the Hartree model with quantum charge is not used here. Using the Hartree approach results in very poor convergence when NEGF charge density is inserted directly into the Poisson equation. We use an internal damping mechanism of converting the charge into a quasi-Fermi level. This approach allows us to arrive at a potential profile in the RTD with a very good accuracy. The resulting non-linear Poisson equation itself is solved iteratively using the Newton-Raphson method until convergence is achieved [25]. In the 1D description, contact regions are doped by 10^{18} cm^{-3}, and Fermi levels in the two contact regions are set such that charge neutrality is preserved.

B. RTD figures of merit

The parameters I_P, I_V, PVCR, V_P, and V_V are directly extracted from RTD *I-V* characteristics, while the following FOM definitions are taken from [26]. The NDC is defined as

$$G_{AVG} = \frac{3}{2}\frac{\Delta I}{\Delta V}, \quad (3)$$

where $\Delta I = I_P - I_V$, and $\Delta V = V_V - V_P$. According to [26], in the cubic polynomial approximation of RTD *I-V* curve, the maximum output power can be expressed as

$$P_{MAX} = \frac{3}{16}\Delta I \Delta V. \quad (4)$$

The theoretical maximum oscillation frequency, when quantum inductance is neglected, can be found from

$$f_{MAX} = \frac{G_{AVG}}{2\pi C_D}\sqrt{\frac{1}{R_S G_{AVG}} - 1}, \quad (5)$$

where R_S is the series resistance and C_D is the capacitance of the RTD. Moreover, the output power in relation to signal frequency is calculated by

$$P_{MAX}(f) = \frac{3}{16}\Delta I \Delta V \cos(2\pi f \tau_T), \quad (6)$$

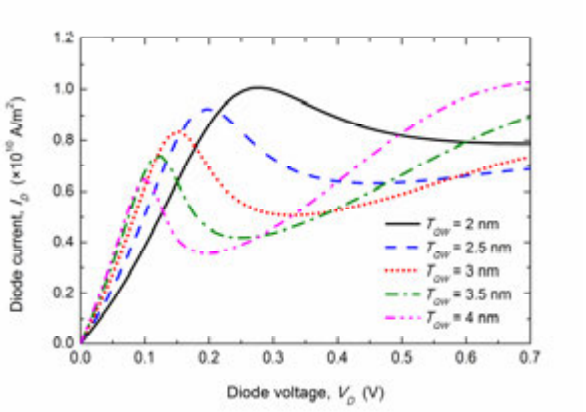


Fig. 2. Simulated current-voltage characteristics of SiNR RTDs with different quantum well lengths. T_B = 1 nm.

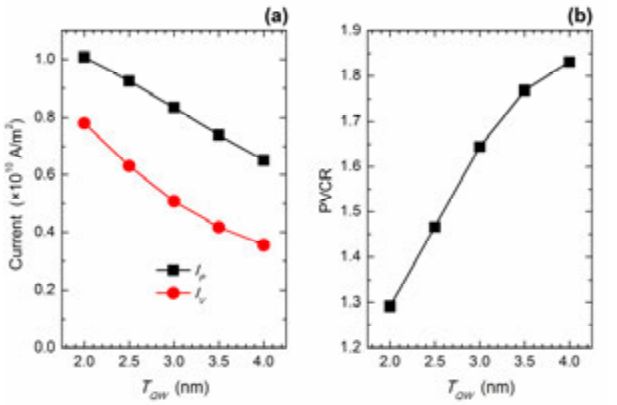


Fig. 3. Dependence of (a) peak and valley current densities, and (b) PVCR on the quantum well length in SiNR RTDs with T_B = 1 nm.

where τ_T is the total intrinsic delay, calculated as

$$\tau_T = \tau_{rtd} + \frac{\tau_{dep}}{2}, \quad (7)$$

where τ_{rtd} represents the carrier tunneling time (i.e. life-time of the QBS in the quantum well) and τ_{dep} represents the carrier transit time in the depletion region. The intrinsic response frequency limit (f_c) represents the frequency where $P_{MAX}(f)$ equals zero, which leads to

$$f_c = \frac{1}{2(2\tau_{rtd} + \tau_{dep})}, \quad (8)$$

since $\cos(2\pi f_c \tau_T) = 0$.

III. Results and Discussions

A. Impact of increasing T_{QW}

In this section, we analyze the performance of SiNR-based RTDs with different T_{QW} for a constant T_B = 1 nm. Figure 2 shows the *I-V* characteristics of the RTD for 5 different quantum well lengths ranging from 2 nm to 4 nm. Each curve contains one negative differential conductance region, which is essential for gaining the wanted device performance. When T_{QW} increases, the shape of the curves changes significantly, with the reduction of V_V and I_V being the most prominent effects.

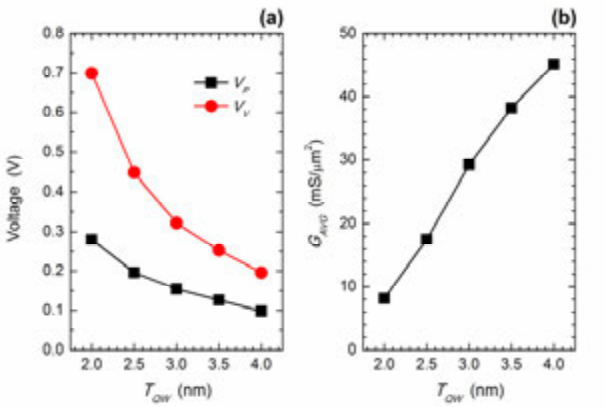


Fig. 4. Dependence of (a) peak and valley voltages, and (b) NDC on the quantum well length in SiNR RTDs with T_B = 1 nm.

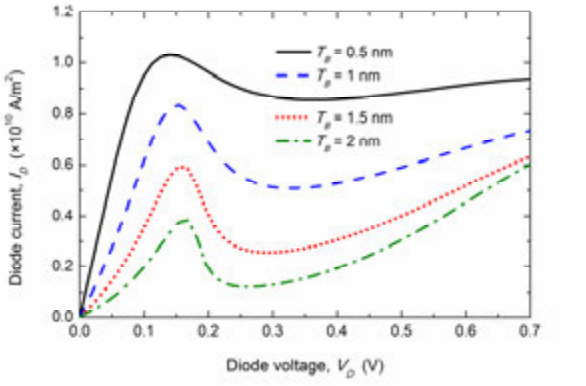


Fig. 5. Simulated current-voltage characteristics of SiNR RTDs with different barrier lengths. T_{QW} = 3 nm.

Figure 3a reports the dependence of I_P and I_V on T_{QW}. By increasing the length of the quantum well, the peak and valley values of the diode current are decreasing. The I_P decreases from 10.1×10^9 A/m^2 down to 6.5×10^9 A/m^2, when T_{QW} is increased from 2 nm to 4 nm. Similarly, I_V deteriorates from 7.9×10^9 A/m^2 down to 3.6×10^9 A/m^2 for the same T_{QW} range. It is noticeable that the peak current decreases slower in comparison to the valley current, which results in the increase of the PVCR. As shown in Fig. 3b, PVCR increases from 1.28 to 1.83, when T_{QW} increases from 2 nm to 4 nm. Therefore, a wider quantum well is beneficial in terms of PVCR since it provides a 43% higher PVCR for T_{QW} = 4 nm than for T_{QW} = 2 nm.

Figure 4a shows the dependence of V_V and V_P on T_{QW}. Similar to the peak and valley current values, the V_V and V_P also decrease when T_{QW} increases. The V_P decreases from 0.280 V down to 0.098 V, when T_{QW} changes from 2 nm to 4 nm, while the V_V decreases from ~0.7 V to 0.196 V for the same T_{QW} range. Unlike the peak and valley current densities, the peak and valley voltages are approaching each other, resulting in a smaller difference. Such behavior of V_V and V_P is not desirable since it reduces the maximum output power of this SiNR-based RTD according to (4). Contrary to that, such behavior of V_V and V_P also results in the increase of G_{AVG}, as reported in Fig. 4b. Specifically, G_{AVG} increases from 7.88 mS/μm^2 to 45.16 mS/μm^2, i.e. by 473% for the case of T_{QW} = 4 nm in comparison to T_{QW} = 2 nm. This G_{AVG} improvement is beneficial for the device in terms of high frequency operation. Namely, as shown in (5), f_{MAX} is proportional to G_{AVG}, which means that we need to maximize G_{AVG} (and reduce the influence of parasitic capacities and resistances) in order to maximize the maximum oscillation frequency. Nevertheless, this benefit comes at the expense of reduced P_{MAX}, as discussed in the previous paragraph.

B. Impact of increasing T_B

In this section, we evaluated RTD *I-V* curves and respective FOMs for different T_B values with a constant T_{QW} fixed at 3 nm. Figure 5 reports the *I–V* characteristics of RTD for 4 different quantum barrier lengths ranging from 0.5 nm to 2 nm. Every curve contains a single negative differential conductance region. With increasing T_B, the diode current is generally decreasing. The curves

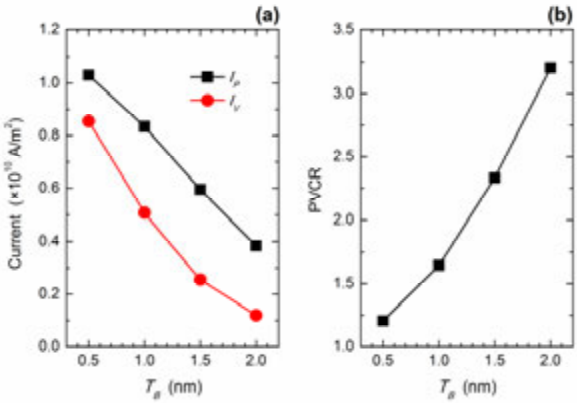


Fig. 6. Impact of barrier length on (a) peak and valley current densities, and (b) PVCR in SiNR RTDs with T_{QW} = 3 nm.

obtained for different T_B values exhibit a similar form, but the slope of the NDC region is increasing significantly when T_B increases, which means that G_{AVG} is also increasing. Another qualitative difference between the curves in Fig. 5 is a noticeable increase of the PVCR.

We observed that the peak and valley current densities decrease with the increase of T_B, which is shown in the Fig. 6a. The I_P decreases from 10.3×10^9 A/m^2 to 3.8×10^9 A/m^2 when T_B rises from 0.5 nm to 2 nm. Similarly, I_V decreases from 8.6×10^9 A/m^2 to 1.2×10^9 A/m^2 for the same change of the barrier length. Accordingly, the PVCR reported in Fig. 6b is increasing from 1.21 to 3.20, which improves the PVCR by 166% for T_B = 2 nm than in SiNR RTDs with 0.5 nm-long barriers. Such behavior is potentially beneficial for the RTD performance in terms of the output power, however, we need to assess the ΔV before drawing any conclusions.

Figure 7a shows the change in V_P and V_V by increasing T_B. It is noticeable that the peak voltage is only slightly increasing from 0.140 V to 0.168 V, when T_B increases from 0.5 nm to 2 nm. While the V_P increases, the valley voltage decreases from 0.364 V to 0.266 V for the same T_B range. Such behavior results in a rapid reduction of the peak to valley voltage difference (i.e. ΔV), which is not desirable since P_{MAX} deteriorates when ΔV decreases according to (4). In contrast, the G_{AVG} increases from 11.74 mS/μm^2 to 40.42 mS/μm^2, i.e. by 244% when the

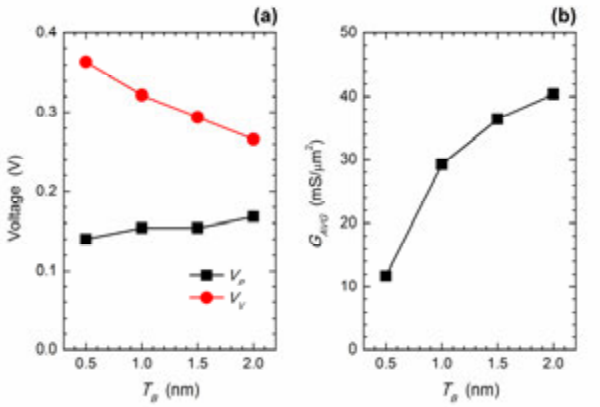


Fig. 7. Impact of barrier length on (a) peak and valley voltages, and (b) NDC in SiNR RTDs with T_{QW} = 3 nm.

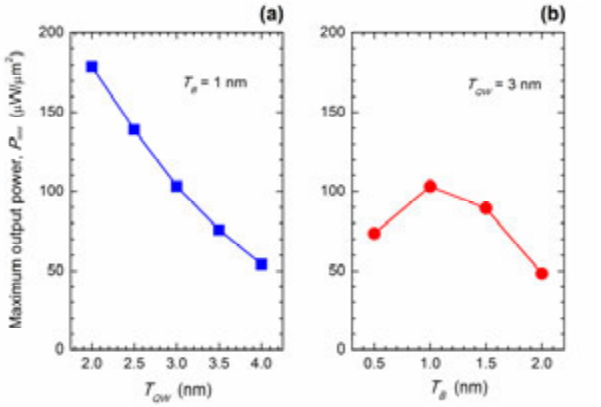


Fig. 8. Maximum output power versus (a) quantum well length (fixed T_B = 1 nm), and (b) barrier length (fixed T_{QW} = 3 nm).

barrier length increases from 0.5 nm to 2 nm, which is shown in Fig. 7b. As mentioned earlier, the increase of G_{AVG} should lead to f_{MAX} improvement according to (5), however, this parameter needs to be co-optimized with the output power that is expected to decrease when increasing barrier length.

C. *Output power and upper frequency limits*

Since the device performance of RTD is measured by its output power and intrinsic response frequency limit, it is important to analyze these two parameters for practical reasons. Regarding the power output, we assess the frequency-dependent $P_{MAX}(f)$ defined in (6), whereas the upper frequency limit is assessed by investigating f_c defined in (8).

Figure 8a illustrates the dependence of P_{MAX} on the quantum well length for T_B = 1 nm. As T_{QW} increases from 2 nm to 4 nm, P_{MAX} monotonically decreases from 179 μW/μm^2 to 54 μW/μm^2. Therefore, the advantageous properties of PVCR and G_{AVG} shown in Figs. 3b and 4b come at the expense of reduced P_{MAX} when the length of the quantum well increases. Figure 8b reports the dependence of P_{MAX} on the barrier length for a fixed T_{QW} = 3 nm. This dependence is not monotonic, i.e. we observe a local maximum of P_{MAX} for T_B = 1 nm. For this barrier length the maximum output power equals 103 μW/μm^2, while for the largest T_B value (2 nm) the P_{MAX} drops below 50 μW/μm^2.

The frequency dependent P_{MAX} for different quantum well lengths, calculated by (6), is shown in Fig 9. While increasing T_{QW}, the shape of the curve does not change significantly, except for the magnitude, which is falling as seen previously in Fig. 8a. For frequencies under ~500 GHz the $P_{MAX}(f)$ is constant and equal to P_{MAX}. When the frequency increases further, the $P_{MAX}(f)$ deteriorates. The $P_{MAX}(f)$ reaches zero at the intrinsic response frequency limit (f_c), as defined in (8). From Fig. 9 we extract the f_c values and observe that the intrinsic response frequency limit decreases from ~3.8 THz to ~2.5 THz when T_{QW} increases from 2 nm to 4 nm.

Similarly, Fig. 10 shows the frequency dependent P_{MAX} for different quantum barrier lengths, with a fixed T_{QW} of 3 nm. For lower frequencies up to approx. 500 GHz, the

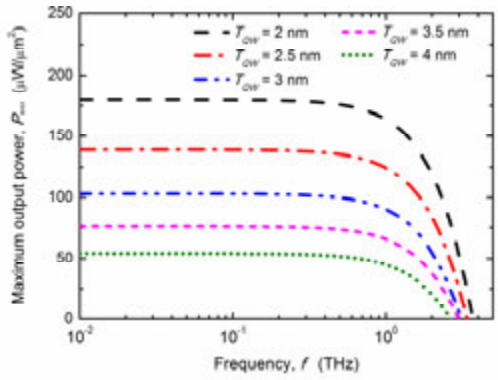


Fig. 9. Frequency-dependence of the calculated P_{MAX} for different quantum well lengths. T_B = 1 nm.

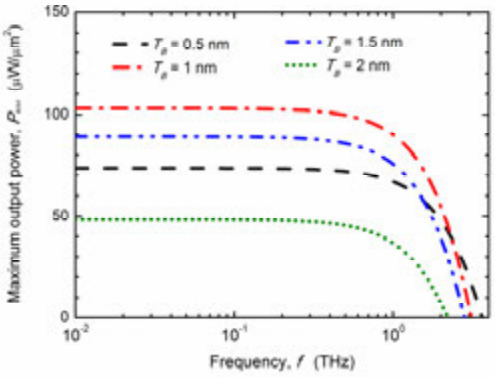


Fig. 10. Frequency-dependence of the calculated P_{MAX} for different barrier lengths. T_{QW} = 3 nm.

frequency dependent maximum output power is equal to P_{MAX} from (4), whereas it decreases towards zero at ~THz frequencies. While the P_{MAX} dependence on T_B is non-monotonic (also seen in Fig. 8b), the impact of T_B increase on f_c is monotonic. Namely, the intrinsic response frequency limit decreases from ~3.6 THz down to ~2.2 THz when T_B increases from 0.5 nm to 2 nm.

IV. CONCLUSIONS

In this study, we assessed the performance of RTDs based on width-engineered SiNRs by using 1D effective-mass quantum transport simulations based on the NEGF formalism. The device design space is analyzed by simulating SiNR-based RTDs with different lengths of the quantum well and barrier regions. By analyzing the DC characteristics, we found that increasing T_{QW} from 2 nm to 4 nm (with T_B fixed at 1 nm) causes a decrease of I_P, I_V, V_P and V_V, while PVCR increases up to 1.83 and G_{AVG} increases up to ~45 mS/μm^2 for T_{QW} = 4 nm. The increase of T_B from 0.5nm to 2nm (with T_{QW} fixed at 3 nm) results in similar qualitative behavior except for V_P that increases with increasing T_B. For T_B-scaling, the maximum achievable PVCR is 3.2 and maximum G_{AVG} is ~40 mS/μm^2 for T_B = 2 nm. The RTD power output is assessed by analyzing the frequency-dependent $P_{MAX}(f)$, while the upper oscillation limit is assessed by exploring the f_c behavior. We found that P_{MAX} and f_c decrease monotonically with increasing T_{QW}, while P_{MAX} exhibits a maximum for T_B = 1 nm when T_B-scaling is analyzed. A SiNR RTD with T_{QW} = 3 nm and T_B = 1 nm exhibits P_{MAX} of ~100 μW/μm^2, f_c of ~3 THz, I_P of ~8×10^9 A/m^2 (=8 mA/μm^2), NDC of ~30 mS/μm^2, and PVCR of 1.6, demonstrating promising results and the motivation for further more advanced atomistic NEGF simulations and analysis of SiNR-based RTDs.

ACKNOWLEDGMENTS

This work is sponsored by the Croatian Science Foundation (HRZZ) under the project CONAN2D (UIP-2019-04-3493).

REFERENCES

[1] J. P. Sun *et al*., "Resonant tunneling diodes: models and properties," *Proceedings of the IEEE*, vol. 86, no. 4, pp. 641–660, Apr. 1998, doi: 10.1109/5.663541.

[2] M. Asada *et al*., "Resonant Tunneling Diodes for Sub-Terahertz and Terahertz Oscillators," *Jpn. J. Appl. Phys.*, vol. 47, no. 6R, p. 4375, Jun. 2008, doi: 10.1143/JJAP.47.4375.

[3] S. Suzuki *et al*., "High-Power Operation of Terahertz Oscillators With Resonant Tunneling Diodes Using Impedance-Matched Antennas and Array Configuration," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, no. 1, pp. 8500108–8500108, Jan. 2013, doi: 10.1109/JSTQE.2012.2215017.

[4] S.-Y. Chung *et al.*, "Monolithic vertical integration of Si/SiGe HBT and Si-based resonant interband tunneling diode demonstrating latching operation and adjustable peak-to-valley current ratios," in *IEEE International Electron Devices Meeting 2003*, 2003, pp. 12.2.1-12.2.4, doi: 10.1109/IEDM.2003.1269283.

[5] N. Deng *et al*., "Ge quantum dot memory realized with vertical Si/SiGe resonant tunneling structure," in *The Fourth International Workshop on Junction Technology 2004.*, 2004, pp. 256–258, doi: 10.1109/IWJT.2004.1306848.

[6] M. A. M. Zawawi *et al*., "Design and fabrication of low power GaAs/AlAs resonant tunneling diodes," in *2016 5th International Symposium on Next-Generation Electronics (ISNE)*, 2016, pp. 1–2, doi: 10.1109/ISNE.2016.7543315.

[7] K. J. P. Jacobs *et al*., "A Dual-Pass High Current Density Resonant Tunneling Diode for Terahertz Wave Applications," *IEEE Electron Device Letters*, vol. 36, no. 12, pp. 1295–1298, Dec. 2015, doi: 10.1109/LED.2015.2491339.

[8] F. M. Yami *et al*., "A New Graphene Nanoribbon based Resonant Tunneling Diodes using BN Quantum Well," in *27th Iranian Conference on Electrical Engineering (ICEE)*, 2019, pp. 315–317, doi: 10.1109/IranianCEE.2019.8786464.

[9] V. Moskaliuk *et al*., "Compact models of the double-barrier resonant tunneling diode," in *IEEE 35th International Conference on Electronics and Nanotechnology (ELNANO)*, 2015, pp. 177–180, doi: 10.1109/ELNANO.2015.7146866.

[10] T. P. E. Broekaert *et al*., "Resonant tunneling in InP/InGaAs lateral double barrier heterostructures," in *53rd Annual Device Research Conference Digest*, 1995, pp. 56–57, doi: 10.1109/DRC.1995.496269.

[11] K. S. Novoselov *et al.*, "Electric Field Effect in Atomically Thin Carbon Films," *Science*, vol. 306, no. 5696, pp. 666–669, Oct. 2004, doi: 10.1126/science.1102896.

[12] M. Y. Han *et al*., "Energy Band-Gap Engineering of Graphene Nanoribbons," *Phys. Rev. Lett.*, vol. 98, no. 20, p. 206805, May 2007, doi: 10.1103/PhysRevLett.98.206805.

[13] V. Barone *et al*., "Electronic Structure and Stability of Semiconducting Graphene Nanoribbons," *Nano Lett.*, vol. 6, no. 12, pp. 2748–2754, 2006, doi: 10.1021/nl0617033.

[14] M. Bresciani *et al*., "Low-field mobility and high-field drift velocity in graphene nanoribbons and graphene bilayers," in *IEEE International Electron Devices Meeting (IEDM)*, 2010, pp. 32.1.1-32.1.4, doi: 10.1109/IEDM.2010.5703461.

[15] P. Vogt *et al.*, "Silicene: Compelling Experimental Evidence for Graphenelike Two-Dimensional Silicon," *Phys. Rev. Lett.*, vol. 108, no. 15, p. 155501, Apr. 2012, doi: 10.1103/PhysRevLett.108.155501.

[16] C. Clendennen *et al*., "Non-equilibrium Green function simulations of graphene, silicene, and germanene nanoribbon field-effect transistors," *Journal of Advanced Simulation in Science and Engineering*, vol. 2, no. 1, pp. 171–177, 2015, doi: 10.15748/jasse.2.171.

[17] M. Poljak, "Intrinsic Capabilities for Digital Switching of Silicene Nanoribbons With Edge Defects," *IEEE Transactions on Electron Devices*, vol. 67, no. 1, pp. 354–359, Jan. 2020, doi: 10.1109/TED.2019.2950967.

[18] M. Poljak, "Electron Mobility in Defective Nanoribbons of Monoelemental 2D Materials," *IEEE Electron Device Letters*, vol. 41, no. 1, pp. 151–154, Jan. 2020, doi: 10.1109/LED.2019.2952661.

[19] H. Teong *et al*., "A Computational Study on the Device Performance of Graphene Nanoribbon Resonant Tunneling Diodes," *Jpn. J. Appl. Phys.*, vol. 48, no. 4S, p. 04C156, Apr. 2009, doi: 10.1143/JJAP.48.04C156.

[20] A. Yazdanpanah Goharrizi *et al*., "Armchair Graphene Nanoribbon Resonant Tunneling Diodes Using Antidote and BN Doping," *IEEE Transactions on Electron Devices*, vol. 63, no. 9, pp. 3761–3768, Sep. 2016, doi: 10.1109/TED.2016.2586459.

[21] G. Wang, "Do silicene nanoribbons have high carrier mobilities?," *Europhysics Letters*, vol. 101, no. 2, p. 27005, Jan. 2013, doi: 10.1209/0295-5075/101/27005.

[22] S. Datta, *Electronic Transport in Mesoscopic Systems*. Cambridge University Press, 1997.

[23] S. Datta, *Quantum Transport: Atom to Transistor*, 2nd ed. New York, NY, USA: Cambridge University Press, 2005.

[24] M. Luisier, "Quantum transport beyond the effective mass approximation," Doctoral Thesis, ETH Zurich, 2007.

[25] D. Vasileska, S. M. Goodnick, and G. Klimeck, *Computational Electronics: Semiclassical and Quantum Device Modeling and Simulation*, 1 edition. Boca Raton: CRC Press, 2010.

[26] C. P. Allford *et al*., "Strain Compensated InGaAs/AlAs Triple Barrier Resonant Tunneling Structures for THz Applications," *IEEE Transactions on Terahertz Science and Technology*, vol. 7, no. 6, pp. 772–779, Nov. 2017, doi: 10.1109/TTHZ.2017.2758266.

Predicting the transport properties of silicene nanoribbons using a neural network

T. Župančić, I. Stresec, and M. Poljak*
Computational Nanoelectronics Group
Micro and Nano Electronics Laboratory, Faculty of Electrical Engineering and Computing
University of Zagreb, Zagreb, HR 10000, Croatia
*Corresponding author: mirko.poljak@fer.hr

***Abstract*—Atomistic quantum transport simulations are used to generate the electronic and transport properties of 10,000 realistic silicene nanoribbons (SiNRs) with edge defects. This ensemble of 20 nm-long and 2.1 nm-wide SiNRs is divided into the training and inference set for the artificial neural network (ANN) employed for the prediction of edge-defect-limited carrier mobility from the known values of bandgap and nanoribbon conductance. We find that an optimized ANN with 3 hidden layers can predict SiNR mobility values and variability histograms with acceptable accuracy, thus providing a useful supplement to atomistic quantum transport simulations that take several hours or days for large device ensemble sizes.**

Keywords—quantum transport; silicene nanoribbon; mobility; bandgap; neural network; TensorFlow; Keras

I. INTRODUCTION

The traditional approach of semiconductor research was in large part based on laboratory experiments, leading to an expensive development process, primarily because of the high cost of clean room maintenance. With the development of computer-aided design (CAD) software, the research and development process today includes material and device simulation, which give researchers a possibility to gain deeper physical insights, and to design and optimize materials and devices without the need for costly fabrication, especially at the nanoscale [1], [2]. This paradigm has greatly decreased research costs, as well as reduced the total development time at different parts of the semiconductor electronics industry.

Over the last decade, quantum transport formalism for nanodevices has gained a tremendous interest due to its ability to capture atomistic resolution of the material used in the device, and due to its ability to incorporate the effects of real-world contacts. Among these, the non-equilibrium Green's function (NEGF) formalism, in its matrix formulation, has become the de-facto standard due to its relative simplicity and versatility [3], [4]. Atomistic NEGF presents a bottom-up approach in the device research community, and it is able to tackle different problems from classical FETs [5], nanotube and nanoribbon devices [6], to molecular sensors. However, computational burden increases significantly when realistically-sized nanostructures (containing several thousands of atoms) are simulated within the NEGF approach [7]. This problem becomes even more pronounced when statistical simulations of a large number of devices are needed for proper assessment of device variability.

Recently, artificial neural networks (ANNs) have been proven useful in device analysis in terms of statistical variability of e.g. Si junctionless nanowire transistors [8]. The ANNs are collections of nodes called neurons that are interconnected via synapses and designed to mimic a (very small part of a) biological brain. The unique design of ANNs has proven to be successful in solving various problems, being especially effective in pattern recognition, classification and image processing [9], [10]. Recent advancements in ANN design suggest a great potential of their large-scale application, both in consumer software business [11]–[13], but also directly in scientific research in various fields [14]–[16].

In this paper, we present a machine learning approach for predicting a chosen transport parameter from the known several electronic and transport parameters of realistic silicene nanoribbons (SiNRs) with edge defects. The ANN is trained on the data obtained from 10,000 atomistic quantum transport simulations of randomly-defective SiNRs. The analysis provided herein gives a detailed assessment of ANN prediction performance (i.e. comparing predicted vs. true values). We show that a relatively simple ANN with 3 hidden layers can predict averaged values and variability properties of nanoscale SiNRs with a reasonable accuracy.

II. MODELING AND SIMULATION APPROACH

A. Quantum Transport

For the SiNRs illustrated in Fig. 1, we employ an atomistic tight-binding (TB) Hamiltonian that accounts for the nearest-neighbor interactions

$$H = \sum_i \varepsilon_i c_i^\dagger c_i + t \sum_{i,j} \left(c_i^\dagger c_j + c_j^\dagger c_i \right), \tag{1}$$

where ε_i is the on-site energy, and t is the hopping parameter between neighboring Si atoms. Using $t = 1.03$ eV for internal Si-Si bonds and a larger $t' = 1.15$ eV for edge bonds reproduces the electronic structure of hydrogen-terminated SiNRs obtained by *ab initio* simulations [17]. Edge defects are implemented by randomly removing single atoms from the top and bottom edges of the nanoribbon shown in Fig. 1. Edge atoms are removed in a given percentage (P_{def}) which equals 10% in this work. The total Hamiltonian is then adjusted by

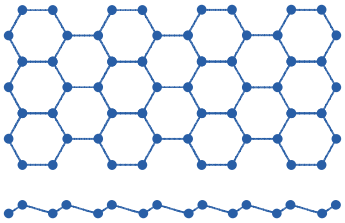


Fig. 1. Top and side view of a monolayer silicene armchair nanoribbon. The side view shows that silicene is not flat but periodically buckled.

setting the interaction parameters of defects (i.e. missing atoms) to zero.

The Schrödinger's equation with open boundary conditions is solved directly by employing the NEGF formalism [4], [18]. Retarded Green's function of the device is calculated with

$$G^R(E) = \left[\left(E + i0^+\right)I - H - \Sigma_1^R(E) - \Sigma_2^R(E)\right]^{-1}, \quad (2)$$

where H is the device Hamiltonian defined in (1), and Σ matrices designate the retarded contact self-energies that account for the coupling of the device to the two contacts (1 and 2, or source and drain). The retarded and advanced Green's functions of the device are then used to find the transmission function, extract the bandgap (E_G), calculate the ON- and OFF-state conductance (G_{ON}, G_{OFF}) and their ratio ($R = G_{ON}/G_{OFF}$) [19], and edge-defect limited mobility (μ_{ED}), according to the following approach. Total resistance of a SiNR can be expressed as $R_{TOT} = R_B + R_D$, where R_B is the length-independent ballistic resistance, and R_D is the length-dependent diffusive resistance. Since the latter can be found from $R_D = L/(q\mu_{ED}n_{1D}) = R_{TOT} - R_B$ [20]–[22], we can extract μ_{ED} using

$$\mu_{ED} = \frac{L}{qn_{1D}\left(R_{TOT} - R_B\right)}, \quad (3)$$

for a given line charge density (n_{1D}).

Figure 2 reports the transmission functions for $N = 10{,}000$ defective 2.1 nm-wide and 20 nm-long SiNRs, and also contains the transmission of an ideal (defect-less) SiNR and an averaged transmission. We note the decrease of transmission, leading to the possibility of extraction of R_D and μ_{ED}, and the increase of the transport energy gap in

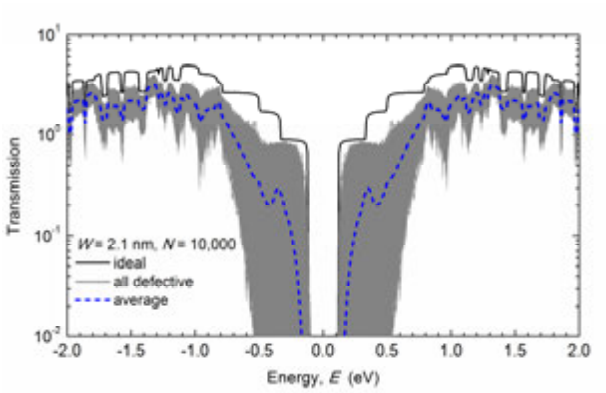


Fig. 2. Transmission for 10,000 defective SiNRs with 10% of edge defects. For comparison, averaged transmission and the transmission of an ideal SiNR are also inserted. $W = 2.1$ nm, $L = 20$ nm.

defective SiNRs. The data reported in Fig. 2 serve as the training set for the ANN described in the following Subsection.

B. Neural Network

In order to obtain meaningful data about the electronic and transport properties of realistic SiNRs with edge defects we need to simulate hundreds or thousands of nanoribbons to study their average properties and variability of their parameters caused by random defects. Due to numerical complexity of atomistic NEGF simulations, these simulations exhibit wall-times on the order of ~hours or ~days, even for 20 nm-long and 2.1 nm-wide SiNRs that contain only 720 atoms [23]. Therefore, this work aims to explore the applicability of ANNs for the prediction of certain SiNRs properties such as μ_{ED}, thus alleviating the need for time-consuming quantum transport simulations.

The feed forward regression NN is implemented in Python using Keras, a high-level NN API built on TensorFlow, and Pandas and NumPy libraries for data manipulation. The neural network, illustrated in Fig. 3, is fully connected and has three inputs: E_G, G_{ON} and $R = G_{ON}/G_{OFF}$; three hidden layers of 32, 64, and 32 neurons, respectively; and one output: edge-defect-limited electron mobility (μ_{ED}). Hence, the 3-layer NN is used to predict μ_{ED} values from the known triplets of E_G, G_{ON} and $R = G_{ON}/G_{OFF}$. The activation function used in all hidden

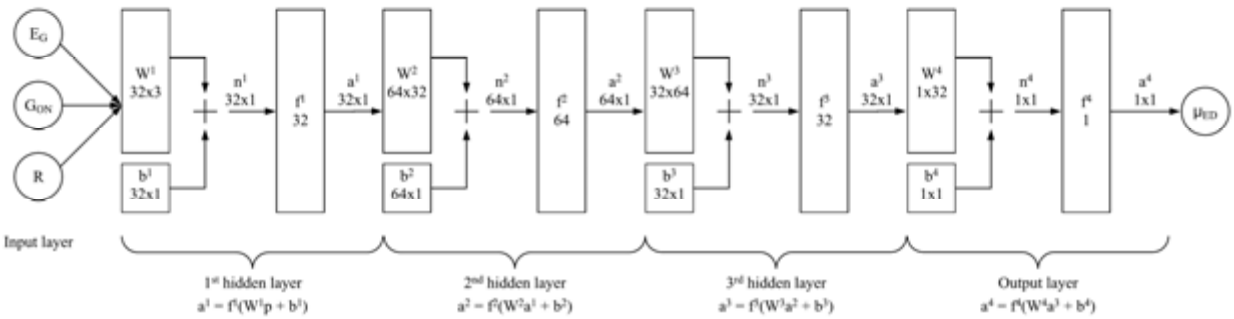


Fig. 3. Model of the artificial neural network used for the prediction of edge-defect-limited electron mobility. The weight and the bias functions are depicted by W^i and b^i, respectively, f^i is the activation function (ReLU), while the output of each layer is annotated by a^i.

layers is the rectified linear unit function (ReLU) and back propagation is used for training. For the optimization of the network we used the ADADELTA method [24] with a starting learning rate of 0.1 and a decay factor (ρ) of 0.95. All other parameters of the network were left to their default values [25]. This NN-based approach has two goals: (1) for a given SiNR size, demonstrate the possibility of mobility prediction from the large existing set of SiNR data for E_G, G_{ON} and $R = G_{ON}/G_{OFF}$, and (2) demonstrate the same possibility for another SiNR size using the same NN found in the previous step without additional training or any parameter adjustment.

III. Results and Discussions

A. NN training and mobility prediction for W = 2.1 nm

The data used in training of the network consists of 9524 sets of E_G, G_{ON}, R, and μ_{ED} values from a Matlab simulation of 2.1 nm-wide and 20 nm-long SiNRs with 10% edge defects. The original 10,000 sets simulated in Matlab is filtered to remove physically unrealistic mobility values. Statistical parameters for this set are summarized in Table I, reporting the mean and median value, and standard deviation for each SiNR parameter. The data is then split as follows: 85% for learning (of which 85% was used for training and 15% for training validation) and 15% for testing.

Optimization with the ADADELTA method demonstrated greater training success compared to Adam, RMSProp and SGD optimizers. These results can be attributed to the adaptive learning rate approach which seems to be the best fit for our data set, presumably due to its comparatively unstable parameters. In general, stated optimizers can serve as good alternatives for training, with small performance differences which in our case suggest the possibility of them outperforming ADADELTA on similar data sets.

The NN architecture used in this work can be considered relatively simple. With only three hidden layers, the network takes significantly less time to train compared to more complex networks. Furthermore, a smaller size of the network did not show significant performance decrease compared to larger ones, rendering more complex NNs as non-cost-effective.

Mean absolute error (MAE) and mean square error (MSE) were used as training metrics, with MSE as the loss function. As mentioned earlier, three inputs (E_G, G_{ON}, R) are used in order to predict the mobility (μ_{ED}). Fig. 4 shows the MAE of training and validation sets through 400 training steps (epochs). MAE decreases quickly in the first few epochs, later stabilizing just above 20 cm^2/Vs. As expected, the validation MAE is slightly higher than that of the training, indicating a successful learning process.

The results of training can be seen in Fig. 5, which shows the distribution of prediction errors (i.e. differences between predicted and true mobility values). The prediction error exhibits a normal (Gaussian) distribution with a mean value close to the ideal 0 cm^2/Vs (actual value is 2.1 cm^2/Vs) and a standard deviation of 27.6 cm^2/Vs, which is considerably lower than the standard deviation of the original data for μ_{ED} (i.e. 40.2 cm^2/Vs) reported in Table I.

Table I. Statistical parameters for W = 2.1 nm (N = 9524)

Stat. values	SiNR parameters			
	E_G	G_{ON}	R	μ_{ED}
std	0.072	0.172	249654	40.2
median	0.336	1.912	35031	31.8
mean	0.339	1.924	117498	45.2

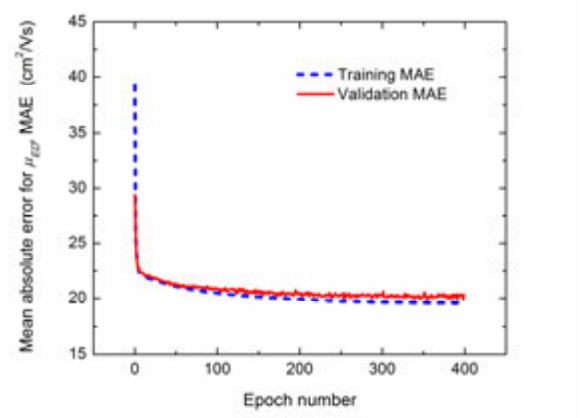


Fig. 4. Mean absolute error of training and training validation versus training epoch number.

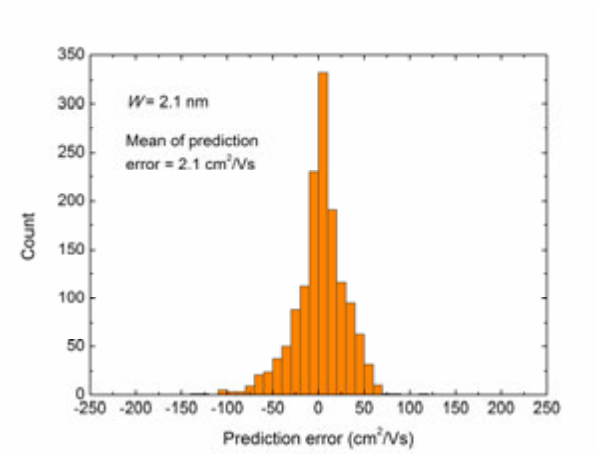


Fig. 5. Distribution of prediction errors for 2.1 nm-wide SiNR.

Furthermore, the NN's tendency to predict values closer to the mean can be seen in Fig. 6, which show the distribution of true (Fig. 6a) and predicted (Fig. 6b) edge-defect-limited mobility values. True values seem to exhibit a Poisson distribution, while predicted values follow a more Gaussian-like distribution. Mean values in both cases are almost the same, with the predicted mean being less than 5% higher than the true mean. Lower number of predicted values in the high-mobility region and the region near 0 cm^2/Vs results in standard deviation of predicted values being 24% lower than the standard deviation of true values.

The success of the NN-based predictions is demonstrated in Fig. 7 that reports the predicted-true value scatter plot for 2.1 nm-wide SiNRs. The plot contains a

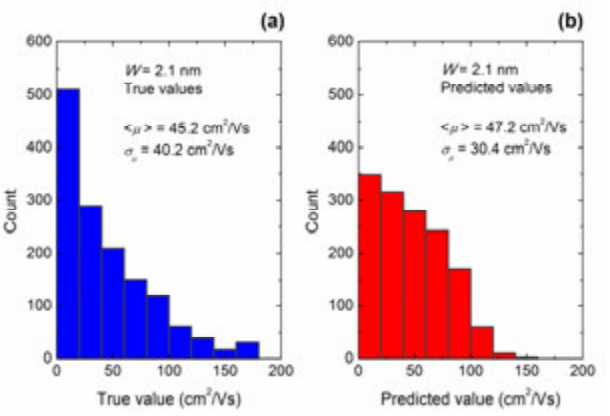


Fig. 6. Distribution of (a) true μ_{ED} values, and (b) predicted μ_{ED} values for 2.1 nm-wide SiNRs.

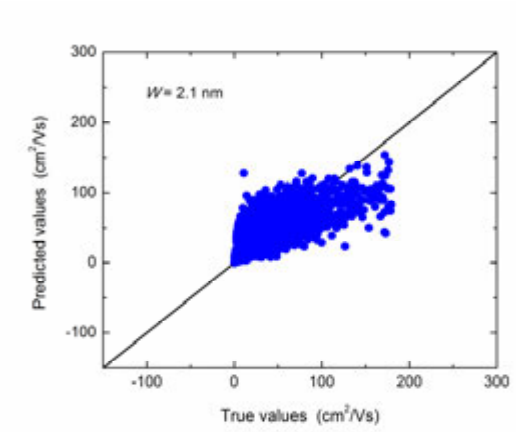


Fig. 7. Scatter graph of predicted and true μ_{ED} values for 2.1 nm-wide SiNRs.

$y = x$ line, meaning the prediction quality can be evaluated by the proximity of blue dots to the diagonal line. In a perfect scenario, where every prediction is equal to the real value, all dots would be placed on the line, however in our case results close to the perfect scenario are not possible due to defect-induced stochastic nature of the data. As expected, the NN predicts values on which it was predominantly trained most accurately, while for values above 100 cm^2/Vs, which were fewer in number in the training dataset, the NN prediction exhibits greater errors. This results in a tighter scatter plot in the lower-mobility region, and a looser scatter plot in the higher-mobility region, as noticeable in Fig. 7.

The training takes around 65.5 s on an Intel i7-8750H CPU and the prediction itself only 36 ms, as calculated using Python's `time` module.

B. NN mobility prediction for W = 3.2 nm

To further test the capabilities of the NN in predicting mobility values, 200 sets of data were simulated in Matlab for 3.2 nm-wide and 20 nm-long defective SiNRs with other parameters unchanged. Statistical parameters of E_G, G_{ON}, $R = G_{ON}/G_{OFF}$, and μ_{ED} for 3.2 nm-wide SiNRs are listed in Table II. For these wider nanoribbons, μ_{ED} values

TABLE II. Statistical parameters for W = 3.2 nm (N = 200)

Stat. values	SiNR parameters			
	E_G	G_{ON}	R	μ_{ED}
std	0.052	0.195	8167	82.8
median	0.194	3.626	3216	67.2
mean	0.204	3.624	5672	89.2

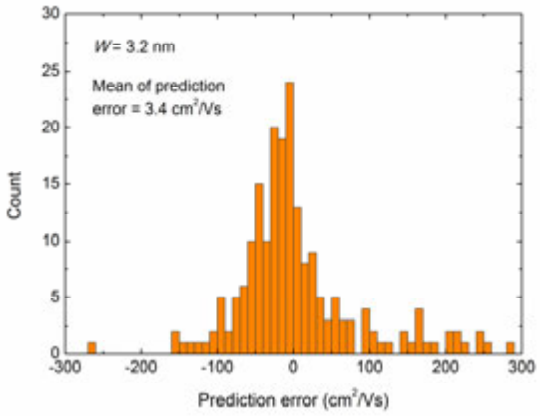


Fig. 8. Distribution of prediction errors for the 3.2 nm-wide SiNRs.

are predicted using the previously trained NN (trained on the data belonging to 2.1 nm-wide SiNRs).

As reported in Table II, the wider defective SiNRs exhibit a lower average bandgap and a higher average μ_{ED} value (average of 89.2 cm^2/Vs, and a standard deviation of 82.8 cm^2/Vs). With a larger MAE of 55 cm^2/Vs, the distribution of which can be seen in Fig. 8, the NN shows promising results and predicts μ_{ED} values surprisingly well despite high defect-induced variability in the original set, and despite being trained on data obtained for a different SiNR width. Just like with the 2.1 nm-wide SiNRs, the prediction error graph exhibits a normal (Gaussian) distribution. The prediction error mean is close to ideal 0 cm^2/Vs, being only slightly higher at 3.4 cm^2/Vs. In this case, the standard deviation of the prediction error is 95.9 cm^2/Vs and is slightly higher than the standard deviation of the true μ_{ED} values (see Table II).

The promising prediction abilities of the neural network are further affirmed by the predicted mean of 92.6 cm^2/Vs, which is less than 5% different from the real value of 89.2 cm^2/Vs. Figure 9 shows the scatter graph of the predicted and true μ_{ED} values for 3.2 nm-wide SiNRs, which acts as expected: namely, mobility values near the mean of the original data are being predicted most successfully, and those further away less so. Nevertheless, the predicted mean and variability of μ_{ED} in the case of W = 3.2 nm demonstrate the suitability of our optimized 3-hidden-layer NN for predicting the transport physics in defective SiNRs irrespective of their size.

The prediction of the sets took about 4 ms, which corresponds to a smaller data set, roughly 8 times smaller than the previous one (for 2.1 nm-wide SiNRs).

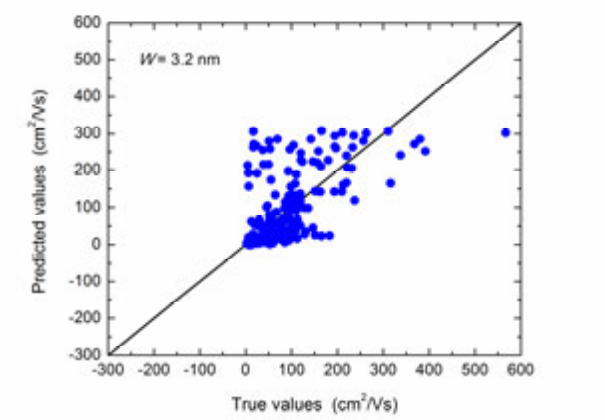


Fig. 9. Scatter graph of predicted and true μ_{ED} values in the case of 3.2 nm-wide SiNRs.

IV. CONLUSIONS

We have demonstrated the potential of neural network usage for predicting the transport properties of defective SiNRs. The NN with relatively simple architecture (three inputs, three hidden layers with 32, 64, and 32 neurons, respectively, and one output) has been constructed and has predicted average μ_{ED} values and its variability within 5% difference from the true values, with average errors significantly smaller than the standard deviation of the data set. We have also shown that the 3-hidden-layer NN, trained on smaller nanoribbons, can also be used for the prediction of device parameters for larger SiNRs. This finding is quite useful since NN training and validation take less than a minute, whereas atomistic quantum transport simulations of a large set of SiNRs take ~hours or even ~days, depending on nanoribbon dimensions and ensemble size. Therefore, this work shows that NNs can be a useful supplement to time-consuming NEGF simulations of nanoscale 2D material-based devices.

ACKNOWLEDGMENTS

This work is sponsored by the Croatian Science Foundation (HRZZ) under the project CONAN2D (UIP-2019-04-3493).

REFERENCES

[1] A. R. Newton and A. L. Sangiovanni-Vincentelli, "CAD tools for ASIC design," *Proceedings of the IEEE*, vol. 75, no. 6, pp. 765–776, Jun. 1987, doi: 10.1109/PROC.1987.13798.

[2] M. Ciappa, "Semiconductor Devices & Processes: TCAD Modeling and Simulation," presented at the IEEE Summer School 2017. Nanoelectronic technologies & devices - From basic principles to highly reliable applications, 2017.

[3] S. Datta, "Nanoscale device modeling: the Green's function method," *Superlattices and Microstructures*, vol. 28, no. 4, pp. 253–278, Oct. 2000, doi: 10.1006/spmi.2000.0920.

[4] S. Datta, *Quantum Transport: Atom to Transistor*, 2nd ed. New York, NY, USA: Cambridge University Press, 2005.

[5] M. Luisier, A. Schenk, W. Fichtner, and G. Klimeck, "Atomistic simulation of nanowires in the sp3d5s* tight-binding formalism: From boundary conditions to strain calculations," *Phys. Rev. B*, vol. 74, no. 20, p. 205323, Nov. 2006, doi: 10.1103/PhysRevB.74.205323.

[6] G. Fiori and G. Iannaccone, "Simulation of Graphene Nanoribbon Field-Effect Transistors," *IEEE Electron Device Letters*, vol. 28, no. 8, pp. 760–762, Aug. 2007, doi: 10.1109/LED.2007.901680.

[7] M. Calderara, S. Brück, A. Pedersen, M. H. Bani-Hashemian, J. VandeVondele, and M. Luisier, "Pushing back the limit of ab-initio quantum transport simulations on hybrid supercomputers," in *SC15: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2015, pp. 1–12, doi: 10.1145/2807591.2807673.

[8] H. Carrillo-Nuñez, N. Dimitrova, A. Asenov, and V. Georgiev, "Machine Learning Approach for Predicting the Effect of Statistical Variability in Si Junctionless Nanowire Transistors," *IEEE Electron Device Letters*, vol. 40, no. 9, pp. 1366–1369, Sep. 2019, doi: 10.1109/LED.2019.2931839.

[9] M. T. Hagan, H. B. Demuth, M. H. Beale, and O. D. Jesús, *Neural Network Design*. Martin Hagan, 2014.

[10] Z. Zhao, P. Zheng, S. Xu and X. Wu, "Object Detection With Deep Learning: A Review," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212-3232, Nov. 2019. doi: 10.1109/TNNLS.2018.2876865.

[11] A. Kusiak, "Convolutional and generative adversarial neural networks in manufacturing," *International Journal of Production Research*, vol. 58, no. 5, pp. 1594–1604, Mar. 2020, doi: 10.1080/00207543.2019.1662133.

[12] M. Tkáč and R. Verner, "Artificial neural networks in business: Two decades of research," *Applied Soft Computing*, vol. 38, pp. 788–804, Jan. 2016, doi: 10.1016/j.asoc.2015.09.040.

[13] N. Shahid, T. Rappon, and W. Berta, "Applications of artificial neural networks in health care organizational decision-making: A scoping review," *PLoS ONE*, vol. 14, no. 2, p. e0212356, Feb. 2019, doi: 10.1371/journal.pone.0212356.

[14] R. Poplin *et al.*, "Predicting Cardiovascular Risk Factors in Retinal Fundus Photographs using Deep Learning," *Nature Biomedical Engineering*, 2018.

[15] A. Esposito, S. Bassis, and F. C. Morabito, "Recent Advances of Neural Networks Models and Applications: An Introduction," in *Advances in Neural Networks: Computational and Theoretical Issues*, S. Bassis, A. Esposito, and F. C. Morabito, Eds. Cham: Springer International Publishing, 2015, pp. 3–8.

[16] "Object Detection With Deep Learning: A Review - IEEE Journals & Magazine." [Online]. Available: https://ieeexplore.ieee.org/document/8627998. [Accessed: 20-Feb-2020].

[17] S. Kaneko, H. Tsuchiya, Y. Kamakura, N. Mori, and M. Ogawa, "Theoretical performance estimation of silicene, germanene, and graphene nanoribbon field-effect transistors under ballistic transport," *Appl. Phys. Express*, vol. 7, no. 3, p. 035102, Mar. 2014, doi: 10.7567/APEX.7.035102.

[18] M. Poljak, E. B. Song, M. Wang, T. Suligoj, and K. L. Wang, "Influence of Edge Defects, Vacancies, and Potential Fluctuations on Transport Properties of Extremely Scaled Graphene Nanoribbons," *IEEE Trans. Electron Devices*, vol. 59, no. 12, pp. 3231–3238, Dec. 2012, doi: 10.1109/TED.2012.2217969.

[19] M. Poljak, "Intrinsic Capabilities for Digital Switching of Silicene Nanoribbons With Edge Defects," *IEEE Transactions on Electron Devices*, vol. 67, no. 1, pp. 354–359, Jan. 2020, doi: 10.1109/TED.2019.2950967.

[20] Y.-M. Niquet, V.-H. Nguyen, F. Triozon, I. Duchemin, O. Nier, and D. Rideau, "Quantum calculations of the carrier mobility: Methodology, Matthiessen's rule, and comparison with semi-classical approaches," *Journal of Applied Physics*, vol. 115, no. 5, p. 054512, Feb. 2014, doi: 10.1063/1.4864376.

[21] M. Luisier *et al.*, "First-principles simulations of 2-D semiconductor devices: Mobility, I-V characteristics, and contact resistance," in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 123–126, doi: 10.1109/IEDM.2016.7838353.

[22] M. Poljak, "Electron Mobility in Defective Nanoribbons of Monoelemental 2D Materials," *IEEE Electron Device Letters*, vol. 41, no. 1, pp. 151–154, Jan. 2020, doi: 10.1109/LED.2019.2952661.

[23] M. Poljak, M. Glavan, and S. Kuzmić, "Accelerating simulation of nanodevices based on 2D materials by hybrid CPU-GPU parallel computing," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2019, pp. 51–56.

[24] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," *arXiv:1212.5701 [cs]*, Dec. 2012.

[25] "Home - Keras Documentation." [Online]. Available: https://keras.io/. [Accessed: 20-Feb-2020].

Introduction to the Design and Simulation of Reflectionless Filters

G. Lemire* and B. Pejcinovic*
* Portland State University, Portland, OR, USA
glemire@pdx.edu, pejcinb@pdx.edu

***Abstract* - One novel approach to improving filter performance is to use so-called reflectionless filters which have recently been proposed. In this paper we demonstrate how to design, simulate and manufacture a reference reflectionless low-pass filter with the corner frequency of 188 MHz. We have found very good agreement between measurements, simulations, and published data, which validates our implementation of the design procedure. We have also designed a comparable diplexer so that we can examine the relative advantages and disadvantages of the reflectionless filters vs. diplexers. Measurements, simulations and published data, all agree very well in this case as well. This agreement extends into multi-GHz range, far into the filter stopband, which illustrates the high quality of the modeling tools used. The reflectionless filter has been verified to provide low return loss for all ports in the stopband which enables cascading. It also has a much simpler bill-of-materials and uses only three unique components, whereas the diplexer uses 32 unique components. On the other hand, the diplexer exhibits steeper initial fall-off after the 3-dB corner frequency and somewhat improved attenuation in the stopband.**

Keywords – reflectionless filters; filter design; diplexer design; simulation.

I. INTRODUCTION

Maximizing useful signal power from the source to the load has been a fundamental design goal in electrical networks since the dawn of modern electronics in the early 20th century [1]-[3]. The mechanism for this matching source to load impedance was usually through some intermediary circuit located between the source and the load. The impedance matching network or more generally "filter" can be as simple as a single termination resistor placed near the load, to an active multitap adaptive broadband filter to quarter wavelength microstrip stubs. Regardless of complexity or target application in which the filter is required, these structures can have one or several design parameters to be optimized such as: maximum power transfer, selective filtering (passband) for a frequency range of interest, and controlling reflection of the signal from the load in an unwanted (stopband) frequency range of interest

Depending on the application and filter topology chosen, in general the design becomes more difficult and complex the more design parameters that need to be optimized. Additionally, many of the current solutions have problems or shortcomings which include inherently limited bandwidth (e.g. quarter wavelength stubs), excessive passband loss, reflective peaking in the transition region and can have a large physical topology which makes implementation difficult for space limited applications. Often the shortcomings of the filter design necessitate additional elements that need to be inserted into the system design to mitigate the unwanted effects.

A subclass of filters called absorptive filters allows for optimization of both passband and stopband characteristics as compared to conventional reflective-type filters. As the name suggests, absorptive filters attenuate the unwanted out of band incident signals by absorption through resistive elements rather than reflection. The concept of absorptive filtering has been around for many decades, with one of the earliest and simplest form of this type of filter is a leaky wall filter which couples unwanted energy to an auxiliary signal path with an absorptive load [1]. No comprehensive list of the different types of absorptive filters was found in the literature, but for the purposes of this work Bulja et al. [4] suggested that absorptive filters can be categorized into transmissive-type absorptive filters and reflective-type absorption filters. Transmission-type absorptive filters can come in several different topologies. The first provides two or more signal paths, of which the sideband path(s) are used to attenuate the stop band signals of the transmission path. The second topology contains the absorptive elements in the body of the filter.

A classical approach to designing a filter that is matched in both the passband and the stopband is to use a diplexer (or multiplexer for more than two frequency bands) and then terminating all but the main path with matched loads. A diplexer is a three-port network that splits the incoming signal from the common port into two frequency dependent signal paths. Diplexers can be realized in admittance complimentary or impedance complimentary architectures [8], [9].

To address the out of band reflective problem a novel methodology for absorptive filters has been recently proposed in [5]-[7]. The authors term these filters reflectionless rather than absorptive because theoretically they have reflection coefficients of zero at all ports for all frequencies for ideal elements [10]. To follow their terminology, these filters will be designated henceforth as reflectionless. This filter has the similar out of band properties as a diplexer but with a simpler design process and more importantly, all ports are matched due to symmetry. This allows for cascading of reflectionless filter sections without the debilitating inter-stage interaction found with many filters.

II. REFLECTIONLESS FILTER DESIGN

The proposed methodology makes use of even-odd mode analysis, symmetry, duality and couples these basic concepts with Cauer ladder network topologies for filters to realize a unique topology for reflectionless filters. Symmetry follows from even-odd mode analysis and this allows the derived filter to have impedance matching at all ports, which implies that the cascading of these structures is at least theoretically possible without interactions between stages.

A. Even-Odd Mode Analysis

Even-mode excitation is defined as impressing two signals which are equal in both magnitude and phase on each port simultaneously. Odd-mode excitation is defined as impressing two signals equal in magnitude but opposite in phase on each port simultaneously. Even and Odd mode equivalent half circuits are sufficient to completely describe the behavior of a symmetric two-port network.

In Figure 1a a symmetric 2-port network is shown. In Figure 1b, the network is divided on its symmetry plane and pulled apart exposing wires representing the central nodes of the circuit. First, even-mode excitation is applied to the two ports and because of symmetry no current will flow in any of the exposed wires, thus creating an open circuit which can be represented in Figure 1c. Similarly, when odd mode excitation is applied to the two ports the wires will be at virtual ground and the circuit can be represented in Figure 1d.

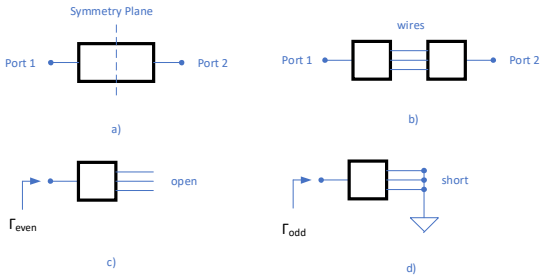


Figure 1. Two port representation of even and odd modes

For the even and odd mode equivalent circuits [5] defines reflection coefficients Γ_{even} and Γ_{odd} . For the even-mode circuit

$$\Gamma_{even} = S_{11} + S_{12} \text{ and by symmetry } \Gamma_{even} = S_{22} + S_{21}$$

Similarly, for the odd-mode circuit

$$\Gamma_{odd} = S_{11} - S_{12} \text{ and by symmetry } \Gamma_{odd} = S_{22} - S_{21}$$

By rearranging and setting $S_{11} = S_{22} = 0$ it can be seen that

$$\Gamma_{even} = -\Gamma_{odd}$$

and

$$S_{11} = \frac{z_{even} - 1}{z_{even} + 1} = -\frac{z_{odd} - 1}{z_{odd} + 1} = -\frac{1/y_{odd} - 1}{1/y_{odd} + 1} = \frac{y_{odd} - 1}{y_{odd} + 1}$$

By inspection it can be seen that

$$Z_{even} = y_{odd} \text{ and } S_{21} = \Gamma_{even}$$

These two relationships are key to realizing reflectionless filters. A general procedure for realizing a reflectionless filter is presented next. To help illustrate the procedure a reflectionless low-pass filter will be designed using the above two key relationships.

Steps 1 and 2: First, design the even mode equivalent half-circuit. The left side of the Figure 2 represents the even mode equivalent circuit of a low pass filter using $S_{21} = \Gamma_{even}$. Second, construct the odd mode dual using $Z_{even} = y_{odd}$, as shown on the right side of the Figure 2.

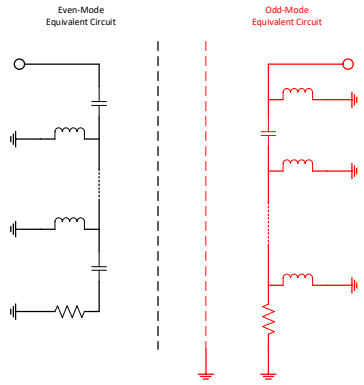


Figure 2. Design of low-pass prototype: even-mode section on the left, and odd-mode section on the right.

Step 3: Symmetrizing the circuit in two steps:

a) redrawing the circuit into more convenient form, as shown in Figure 3. Redrawn components are in red.

b) adding components to make the overall circuit symmetric, as shown in Figure 4. New components are in red. Note that they are superfluous as they do not affect circuit function because they either have a short-circuit in parallel or end in an open-circuit.

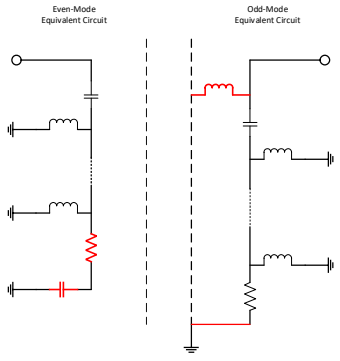


Figure 3. First step in symmetrizing the design: redraw circuit from Figure 2. Redrawn components are in red.

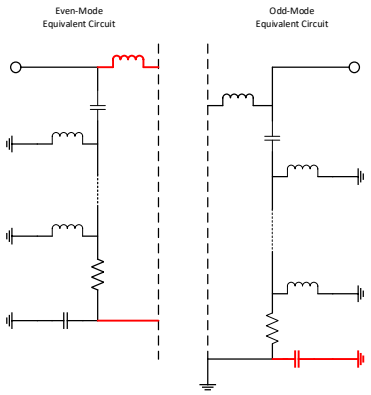


Figure 4. Second step in symmetrizing: adding superfluous components (in red).

Step 4: even and odd sections are connected to form the full filter, as shown in Figure 5.

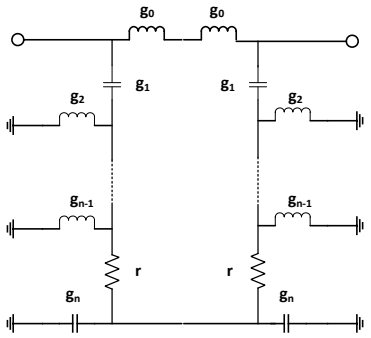


Figure 5. Final low-pass filter prototype of arbitrary order n and normalized component values.

With the topology of the filter established, normalized component values can now be determined, with the values being guided by both symmetry and duality constraints and can be generalized for any characteristic impedance. Duality implies that inductor g_0 must be the dual of the capacitor g_1 (step 2). Determining component values to satisfy the duality condition we have for the first two components

$$L_0 = Z_o^2 C_1$$

and substituting the impedance of both inductor and capacitor elements into the above equality gives

$$\frac{g_0 Z_0}{\omega_c} = Z_o^2 \frac{g_1 Y_0}{\omega_c}$$

By inspection $g_0 = g_1$, and similarly, using duality, it can be seen that g_1 must be the dual of g_2 and using this same reasoning we can infer $g_0 = g_1 \ldots = \cdots g_k$ and r=1.

B. Ideal Filter Simulation

As explained in [5], design of a single filter stage is limited to odd orders and the third order filter stage gives the best compromise among various requirements. To obtain the required characteristics, it is usually necessary to cascade several filter sections, as explained later.

First, we will examine ideal filter behavior using simulation program ADS [12]. Filtering performance of a low-pass filter (LPF) consisting of a cascade of four stages where each stage is of the third order is shown in Figure 6. Magnitude of S_{21} clearly illustrates where passband and stopband regions are and 3-dB cutoff frequency for this filter is 188 MHz. For the magnitude of S_{11}, also known as return loss, we present two plots. One plot is for circuit simulation using elements with four significant digits and the other for using six significant digits. Obviously, the more significant digits we use the closer we are to ideal behavior. However, practical implementations will have much worse tolerances, and therefore fewer significant digits. Practical components are specified between 1% and 5% of the nominal value. Note that the return loss for ideal filters is very small across all frequencies.

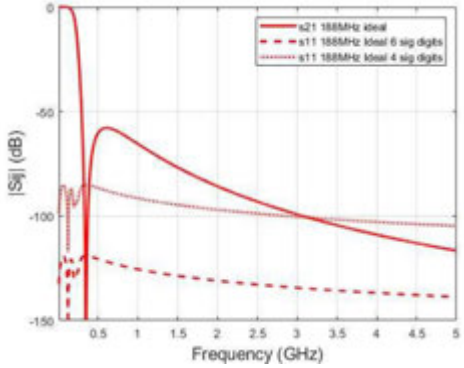


Figure 6. Ideal low-pass filter performance. The amount of return loss is presented for the component accuracy of four and six significant digits.

Next, we will examine reflectionless filters which use non-ideal components using simulations as well as measurements of the manufactured filter.

III. FILTER IMPLEMENTATION

The PCB stack-up and material used for this work is well beyond what is required for both the frequency of operation as well as the complexity of the circuit. The specific materials were chosen because the stack-up is familiar, well controlled and PCB panels are regularly run at Synopsys which allows the opportunity to include the test cards in a standard panel. The stack up is 12-layer low-loss Tachyon-100G material and the material properties are shown in Table I [13].

TABLE I. PRINTED CIRCUIT BOARD PROPERTIES, TACHYON-100G

No. of layers	Dielectric thickness (mils)	Dk	Df	Copper thickness (mils)
12	3.25 to 6	3.02	0.0021	0.6 or 2

Obviously, we cannot purchase SMD components with precision of six significant digits, as given in Table II. Instead, we decided to use components within 5% tolerance with values closest to the ideal ones, as shown in Table II.

TABLE II. IDEAL AND ACTUAL SMD COMPONENTS SELECTED FOR FILTER IMPLEMENATION. COMPONENT TOLERANCES ARE 5%.

	L (nH)	*C (pF)*	*R (Ohms)*
Ideal	22.7176	9.08706	50
Non-ideal (5%)	23	9.1	49.9

A four-stage filter, equivalent to a 12-th order traditional filter, was designed and implemented in this technology. Photo of the final PCB is shown in Figure 7. PCB has dimensions of 3" x 1".

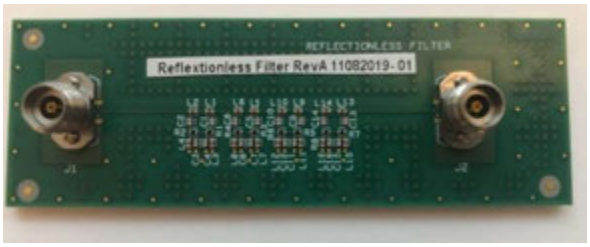


Figure 7. PCB implementation of reflectionless LPF with four cascaded sections and equivalent filter order n=12 with corner frequency 188 MHz.

A. Simulated and Measured LPF Results

Before the LPF was manufactured, we examined its performance relative to the ideal case by using ADS simulation. Results are shown in Figure 8 which demonstrates that S_{21} behaves almost ideally up to 1.5 GHz. After that frequency we suspect that some interactions among parasitics in components and layout cause significant differences. Return loss (S_{11}) is very far from ideal but it is less than 20 dB up to 1.5 GHz. This would indicate that S_{11} is more sensitive than S_{21} to departures from ideal component values.

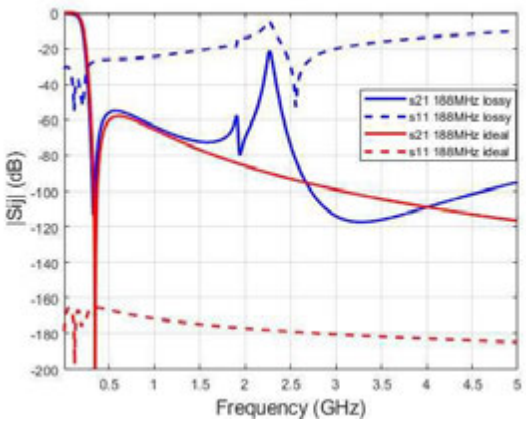


Figure 8. Simulated S_{11} and S_{21} for reflectionless filter with ideal and realistic (lossy) components.

Final test of any design is a comparison between simulation and measurements. The circuit shown in Figure 7 was carefully measured on a 40 GHz VNA. Effects of the fixture were removed by de-embedding using AFR (Automatic Fixture Removal). As shown in Figure 9, we observed excellent agreement between simulation and measurements. We attribute this to very accurate component models [11] and careful circuit simulation which includes most external parasitics, such as transmission line effects. This gives us confidence that more challenging designs, e.g. those using higher frequencies, and circuit optimization will also be successfully simulated and manufactured.

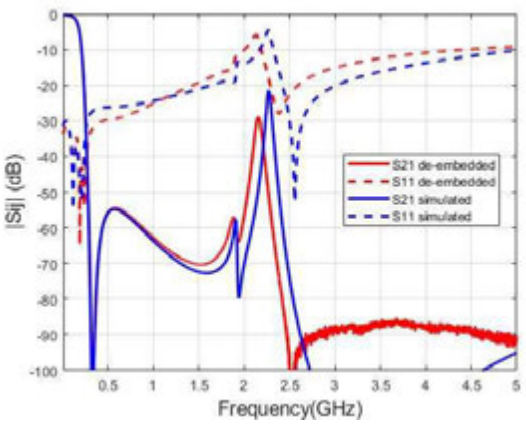


Figure 9. Simulated and measured S_{11} and S_{21} for the reflectionless LPF.

IV. DIPLEXER DESIGN AND IMPLEMENTATION

To evaluate the quality of the reflectionless filters return loss, a conventional filter with similar low reflection out of band qualities was chosen to be used as a comparison. A dual diplexer formulation [8], [9] is considered where one constructs low pass and high pass filters that have complimentary input admittances. The filters are connected in parallel at a common port with the input admittance of the diplexer given as $Y_o = Y_p + Y_s$ where Y_p and Y_s represent the input admittances of the low-pass and high-pass filters respectively. A matched filter requires that the input admittance Y_o be real and constant for all frequencies on the common port. Filter pairs that satisfy this condition are termed complimentary. The physical implementation of the diplexer in Figure 10 shows the three-port diplexer on the left with the separate two port de-embedding structure on the right. The diplexer common port is on the far left with the low-pass port on the upper center and the high-pass port on the lower center. For this work, and following [6], an order n=16 Butterworth diplexer was chosen as a comparison.



Figure 10. Diplexer implementation of LPF on the left. Structure on the right is used for deembedding.

The comparison of the reflectionless and the diplexer LPFs are shown in Figure 11. The reflectionless LPF displays similar responses for S_{21} except the transition

region is initially less steep. More importantly, the S_{11} response for reflectionless LPF is superior to the diplexer across almost the entire frequency range.

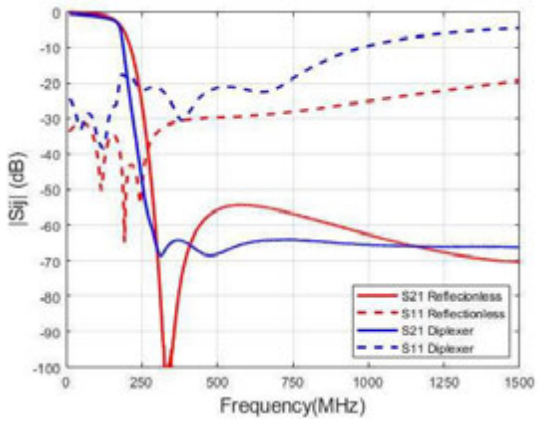


Figure 11. Comparison of measured S_{11} and S_{21} for reflectionless and diplexer implementation of LPF.

Another important feature of the reflectionless filter is the symmetry property. The filter is a two-port device and because of symmetry the return loss at both ports is theoretically identical. This allows these filters to be cascaded with other filters or with other sensitive 50-ohm components with little interaction. Figure 12 demonstrates that the reflectionless filter has very close to identical return loss responses S_{11} and S_{22}. The diplexer is matched for S_{11} (common port) relatively well but S_{22} shows almost unity reflection in the stop band. Any noise above the passband frequency on diplexer port 2 will tend to reflect from the filter rather than being absorbed with possible resonance ramifications between the filter and any connected circuits.

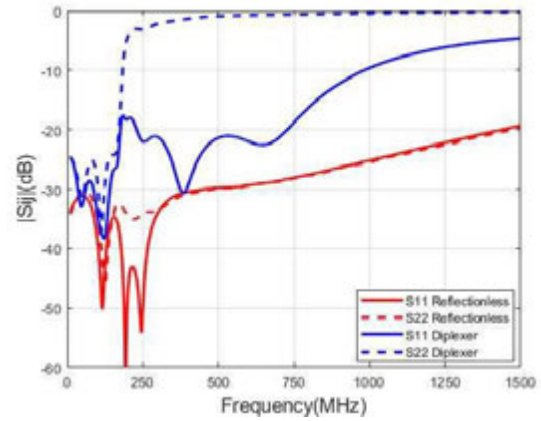


Figure 12. Comparison of measured return loss on both ports of reflectionless and diplexer implementation of LPF.

V. Conclusion

We have demonstrated and verified the novel reflectionless filter design procedure outlined in Morgan and Boyd's papers. A low pass reflectionless filter prototype was realized. A classical Butterworth diplexer prototype satisfying the same requirements was also realized and used to evaluate the quality of the return loss characteristics for the reflectionless filter. The prototype measurement results obtained for reflectionless filter and equivalent diplexer show that the reflectionless filter has improved return loss characteristics over a wide frequency range. We have also observed very good matching between simulation and measured response for reflectionless filters. Return loss of both ports on reflectionless filter was superior to diplexer design. Diplexer has an advantage of steeper fall-off in the stopband albeit in a limited frequency range close to the corner frequency.

Acknowledgment

We would like to thank Synopsys for use of their lab and for manufacturing circuits, Keysight for providing ADS simulation software, and to Modelithics for providing SMD component models for ADS.

References

[1] V. Met, "Absorptive Filters for Microwave Harmonic Power", Proc, IRE, vol. 47, no. 10, pp. 1762-1769, Oct. 1959. DOI 11.1109/JRPROC.1959.287111

[2] H.W. Bode, "Network Analysis and Feedback Amplifier Design" New York: D. Van Nostrand Company, 1945.

[3] G. Matthaei, L. Young, and E. Jones, "Microwave Filters, Impedance Matching Networks, and Coupling Structures". Norwood, MA: Artech House, 1980

[4] S. Bulja, A. Grebennikov, and P. Rulikowski, "Theory, analysis and design of high order reflective, absorptive filters", IET Microwaves, Antennas and Propagation, March 2017. DOI 10.1049/iet-map.2016.0431

[5] M. A. Morgan, "Think Outside the Band: Design and Miniaturization of Absorptive Filters," in *IEEE Microwave Magazine*, vol. 19, no. 7, pp. 54-62, Nov.-Dec. 2018. DOI: 10.1109/MMM.2018.2862541

[6] M.A. Morgan and T.A. Boyd, "Technical and experimental study of a new class of reflectionless filter", *IEEE Trans. Microwave Theory and Techniques*, vol. 59, no. 5, pp 1214-1221, May 2011. DOI 10.1109/TMTT.2011.2113189

[7] M. A. Morgan and T. A. Boyd, "Reflectionless Filter Structures," IEEE Transactions on Microwave Theory and Techniques, vol. 63, no. 4, pp. 1263–1271, Apr. 2015, doi: 10.1109/TMTT.2015.2403841.

[8] G .L. Matthaei and E. G. Cristal, "Theory and Design of Diplexers and Multiplexers", Advances in Microwaves, vol. 2, pp 237-326, 1967.

[9] R. G. Veltrop and R. B. Wilds, "Modified Tables for the Design of Optimum Diplexers", Electronic Defense Laboratories, Technical Memorandum, No EDL-M559, 3 July 1963

[10] M. A. Morgan, "Reflectionless Filters", Norwood MA: Artech House, 2017

[11] Modelithics Exemplar Library [online] Available: https://www.modelithics.com/

[12] Advanced Design System (ADS), Keysight, [online], Available: https://www.keysight.com/en/pd-2998595/pathwave-advanced-design-system-ads-2020?cc=US&lc=eng

[13] Isola group, Tachyon-100G data sheet, [online], Available: http://www.isola-group.com/wp-content/uploads/Tachyon-100G-Laminate-and-Prepreg-Data-Sheet-0417.pdf

Microwave Interferometry Measurements of Yeast Cell Suspension and Sediment Process

M. Zhang*, T. Markovic** and B. Nauwelaers*
* Department of Electrical Engineering, University of Leuven, Leuven, Belgium
** imec, Heverlee, Belgium
meng.zhang@esat.kuleuven.be

***Abstract* – A microwave interferometry based coplanar waveguide sensor is proposed for cell dynamic process monitoring applications. Combining the ease of microfluidics integration and impedance modeling from coplanar waveguide with high sensitivity and wide frequency range from interferometer structure, the designed setup has proved its potential for biological applications. Based on characteristic of interferometer and sensor modeling, the algorithm for extraction of the material under test complex permittivity from the S-parameters changes has been developed. Measurement and calculation of 2 mg/ml yeast cell suspension have been carried out to validate the setup, which agreed well with reference data in literature. Furthermore, monitoring of yeast cell sediment process was performed to lay the groundwork for future cell growth monitoring.**

Keywords - dielectric spectroscopy, interferometry sensor, millimeter wave, microfluidics, biogenic liquid

I. INTRODUCTION

Cell suspension measurement and cell growth monitoring are of great importance for both academic and industries in biology, biochemistry, and pharmaceutical fields [1], [2]. Among the fast developing and widely used bio-sensing techniques, such as optical, mechanical and chemical methods, electrical detection and property extraction of cell cultures have raised significant interests for label-free, noninvasive and high-sensitivity biosensor designs [3], [4]. One major method is to characterize the material-under-test (MUT) with its frequency dependent permittivity, a physical representation of (di)electric polarization [5]. Biogenic liquid's permittivity contains valuable biological information, especially for dielectric dispersions at microwave and millimeter wave frequencies. However, most of the work studying cell dynamic properties so far mainly focus on kHz frequency range [6], [7] and are often limited on sensitivities.

Many efforts, for example, resonator structures, have been explored to improve electrical biosensor's sensitivity [8], but its frequency range is confined to the resonance point. In order to measure broad frequency range complex permittivity of MUT with high sensitivity, interferometric dielectric spectroscopy has proved its potential for bio-material characterization in recent years [9]. Due to its flexibility regarding sensing structures, many microwave topologies, e.g. coplanar waveguide (CPW), with easy integration of miniaturized microfluidics have led to broad frequency range and high sensitivity measurements of liquid mixtures [10-12], cell cultures [13-14], single cells [15] and even microwave heating at the micro-meter scale [11], [16-17]. And yet, few researches have been carried out on the cell culture monitor, especially, the sediment process.

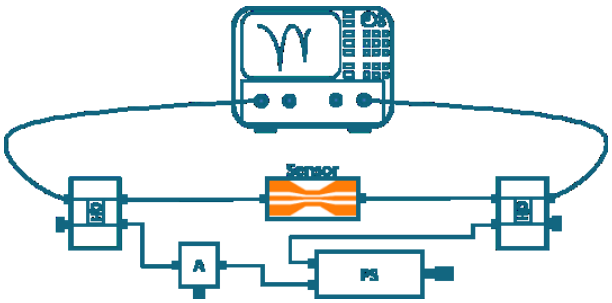


Figure 1. Schematic of measurement setup.

As a capability demonstration of microwave interferometry techniques in measuring and monitoring cell behavior in suspensions, this work illustrates the application of a CPW-based biosensor to extract the dielectric dispersion of yeast cell suspensions and their sediment process at 12-18 GHz. Section II describes the measurement setup, the sensing principle with CPW sensor design, and the permittivity extraction method. In section III, the measurements are conducted with yeast cell to water concentration of 1 g / 50 ml and 1 g / 500 ml. Finally, the paper is concluded in the last section.

II. MEASUREMENT SETUP AND SENSING PRINCIPLE

A. Measurement Setup

Schematic of the system is shown in Fig. 1. Two 90-degree couplers (Quadrature Hybrid (QH), SigaTek, SQ16506) split and combine the sensing signal (e.g. 15 GHz) from Vector Network Analyzer (VNA) into two branches: the sensor channel and the reference/tuning channel with 180-degree phase difference, aiming to form destructive interference. The reference channel is connected to mechanically controlled Attenuator (Narda 4799) to compensate the attenuation and Phase Shifter (PS, ARRA 9428A) to finely tune the phase difference between these two branches. The sensor channel consists of a tapered CPW sensing structure to adapt to miniaturized microfluidics (Fig. 2). Interference null is achieved as S_{21}

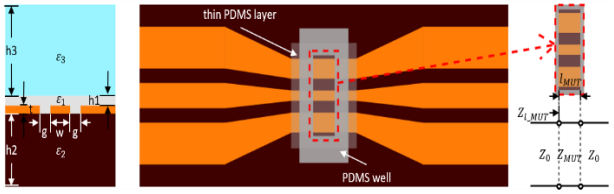


Figure 2. The cross-section (left), top views (middle) and impedance model (right) of the CPW sensing structure, where w = 0.6 mm, g = 0.4 mm, t = 18 µm (copper), h_1 = 20 µm (thin PDMS layer), h_2 = 1.6 mm, h_3 = 5 mm (MUT), ε_1 = 2.69 (PDMS), ε_2 = 2.2 (Duroid 5880), ε_3 = permittivity of MUT.

on VNA, when these two channels are turned to similar attenuation but around 180-degree phase difference.

B. *Sensing Principle and Sensor Design*

Fig. 2 illustrates the cross-section (left), top views (middle) and impedance model (right) of the CPW sensing structure for the measurements. The CPW is made of Duroid 5880 with copper thickness of 18 µm. And to make it biocompatible, a 20 µm polydimethylsiloxane (PDMS) layer is applied in the sensing area, which is indicated with red dashed rectangular. At last, the PDMS microfluidic well used to hold MUT is attached on top. CPW-based sensor was chosen because of its ease of fabrication and connection to the measurement system and high sensitivity to changes of MUT above. The depth of microfluidic well on top of the CPW sensing area are designed relatively deep (h_1 = 5 mm) compared to the gap of CPW (g = 4 mm) to guarantee sufficient height of MUT, so that the detected signal change is caused mainly by MUT, and also reduce errors introduced by manual filling and evaporation. The width of well, l_{MUT}, is 1 mm, resulting an effective volume of 7 µl.

The attenuator and PS are tuned to obtain small S_{21}, inducing better destructive interference, at a certain frequency f_0 = 15 GHz. The smaller S_{21} was tuned, the higher Q-factor the system will achieve, thus, larger sensitivity. Fig. 3 displays the system's S-parameters in frequency domain after tuning S_{21} to below -70 dB and its Q-factor has reached 10^4.

Similar to the permittivity extraction described in [18], with corresponding S_{11} and f_0, the complex permittivity of MUT,

$$\varepsilon = \varepsilon' - j\varepsilon'' \tag{1}$$

can be calculated from the signal propagation constant:

$$\gamma_{MUT} = (\alpha_c + \alpha_c) + j\beta \tag{2}$$

of the CPW sensing area filled with MUT. Specifically, here we apply:

$$\frac{S_{11(MUT)}}{S_{11(CAL)}} = \frac{Z_{l_MUT} - Z_0}{Z_{l_MUT} + Z_0} \times \frac{Z_{l_CAL} + Z_0}{Z_{l_CAL} - Z_0} \tag{3}$$

$$\frac{S_{21(MUT)}}{S_{21(CAL)}} = \frac{\exp(-\gamma_{MUT} \times l_{MUT})}{\exp(-\gamma_{CAL} \times l_{CAL})} \tag{4}$$

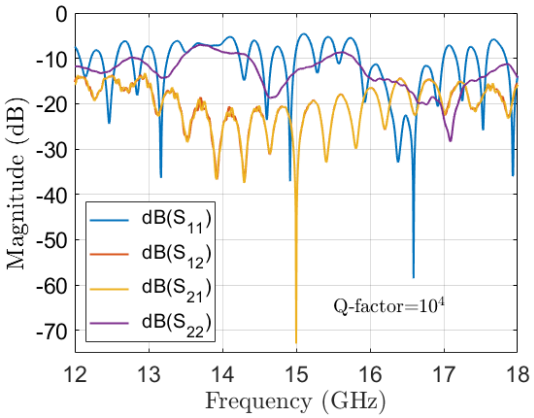


Figure 3. S-parameters of the system with S_{21} (15 GHz) tuned below -70 dB, Q-factor = 10^4.

among which the subscripts MUT and CAL, refer to "material under test" and "calibration liquid" respectively; And $l_{MUT/CAL}$ is the physical length of the sensing area sections; Z_{l_MUT} and Z_0 are impedances marked in Fig.2. Deionized water is used as calibration liquid to help eliminate the signal transfer functions of all other sensor and system components except for the sensing area section.

With the obtained signal propagation constant of the sensing area filled with MUT (γ_{MUT}), described in [19], $Z_{l_MUT/CAL}$ is determined by the term "tanh(-γ_{MUT} × l_{MUT})". Thus, the real and imaginary part of MUT permittivity can be acquired. Then, repeating the same calculations for each frequency point.

III. Measurements and Discussions

A. *Yeast Cell Suspension Measurement*

To validate the design, yeast cell suspension with concentration of 2 mg/ml is tested first. The measurement was conducted under room temperature of 16.7°C with setup schematic illustrated in Fig. 1 and guided by the following steps:

- Firstly, after connecting all the components and sensor, deionized water was injected into the PDMS well with a pipette.
- Secondly, tuning attenuator and PS alternately at 15 GHz in frequency domain sweep mode until S_{21} reaches below -70 dB for higher sensitivity and recording the S-parameters.
- Thirdly, mixing the Bruggeman dry bakery yeast with deionized water and rotating the container for 5 mins to fully suspend the yeast cells.
- Finally, carefully and completely removing deionized water inside the well and refilling it with 2 mg/ml yeast cell suspension and recording the S-parameters again.

Fig. 4 shows the extracted real and imaginary permittivity of 2 mg/ml yeast cell suspension with error bars. The extracted permittivity correlates well with reference data in [20], validating the sensor's capability of sensitive and accurate permittivity measurements.

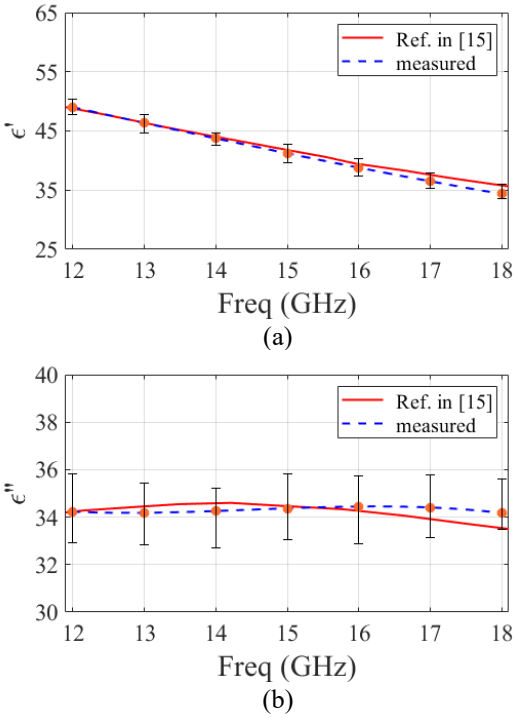


Figure 4. Extracted permittivity of 2 mg/ml yeast cell suspension: (a) real part, (b) imaginary part.

B. *Yeast Cell Sediment Monitoring*

Since the density of yeast cell is slightly larger than deionized water, the sediment process will happen right after filling the PDMS well with yeast cell suspension. Accumulating at the bottom of CPW sensor, the extracted permittivity will decrease for both real and imaginary parts (Fig. 5). From the sediment monitoring measurement, we can see that for 2 mg/ml yeast cell suspension, the cells will settle down in about 23 minutes with a volume of 7 µl.

C. *Discussions*

The error of extracted permittivity is considered relatively large especially for imaginary part. A few methods can be applied to improve its accuracy; firstly, making a better fixation of the system to eliminate the environmental noises, since the system is very sensitive to mechanical movements; secondly, providing temperature control of the MUT to reduce the error caused by temperature difference, noted that the measurement temperature in [18] is at 18.5°C while this work is carried out at 16.7°C; and finally, coming up with more accurate permittivity extraction algorithm, taking the reflection of signal between different components into consideration. Our next step also includes cell growth monitoring on thin layer of cell culture mediums, such as luria agar, lysogeny broth and collagen gel, broadening the horizon of microwave sensors on biological applications.

IV. CONLUSION

This paper presents a broadband, high sensitivity microwave measurement setup for monitoring cell dynamic properties. The system is based on microwave interferometry structure and combined with CPW sensor to enable adaptation to microfluidics and extraction of permittivity. The setup was validated by 2 mg/ml yeast cell suspension measurement with good agreement compared to reference data in [20]. Then, it was applied to sediment process monitoring, providing valuable information about cell settlement for future cell growth monitoring experiments.

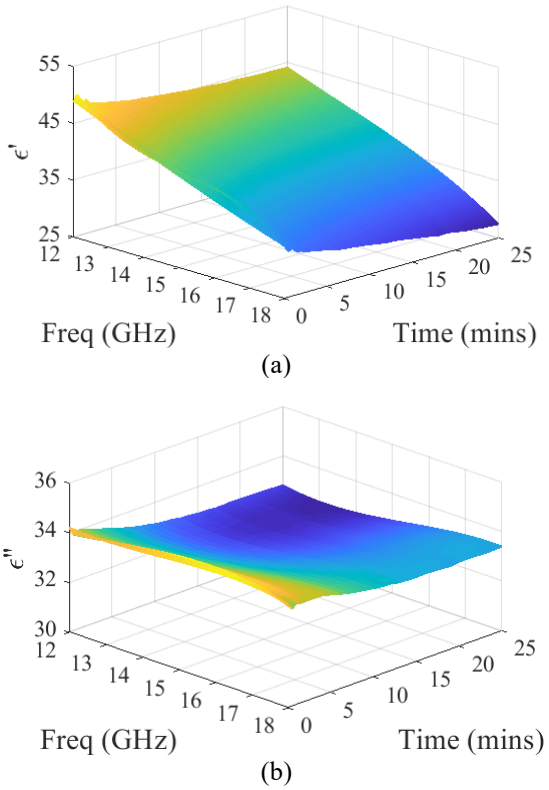


Figure 5. Sediment monitoring of 2 mg/ml yeast cell suspension for 25 minutes: (a) real part, (b) imaginary part.

ACKNOWLEDGMENT

This work was supported in part by the KU Leuven under project C24/15/015 Microwave Microbiology (µ²BIO), and in part by Chinese Scholarship Council for Meng Zhang.

REFERENCES

[1] A. K. Laird, "Dynamics of tumour growth: comparison of growth rates and extrapolation of growth curve to one cell," Br. J. Cancer, vol. 19, no. 2, pp. 278-291, 1965.

[2] M. Mir et al., "Optical measurement of cycle-dependent cell growth," Proc. Natl. Acad. Sci. U.S.A., vol. 108, no. 32, pp. 13124-13129, 2011.

[3] T. Sun and H. Morgan, "Single-cell microfluidic impedance cytometry: a review," Microfluid. Nanofluid., vol. 8, pp. 423-443, 2010.

[4] Q. Liu, C. Wu, H. Cai, N. Hu, J. Zhou, and P. Wang, "Cell-based biosensors and their application in biomedicine," Chem. Rev., vol. 114, no. 12, pp. 6423-6461, 2014.

[5] F. S. Barnes and B. Greenebaum, Biological and medical aspects of electromagnetic fields, third edition, CRC press, Taylor & Francis Group, 2007.

[6] I. Giaever and C. R. Keese, "Use of electric fields to monitor the dynamical aspect of cell behavior in tissue culture," IEEE Trans. Biomed. Eng., vol. BME-33, no. 2, pp. 242-247, 1986.

[7] I. O. K'Owino and O. A. Sadik, "Impedance spectroscopy: a powerful tool for rapid biomolecular screening and cell culture

monitoring," Electroanalysis, vol. 17, no. 23, pp. 2101–2113, October 2005.

[8] D. Wood, S.-H. Oh, S.-H. Lee, H. Soh, and A. Cleland, "High-bandwidth radio frequency Coulter counter," Appl. Phys. Lett., vol. 87, pp. 184106, 2005.

[9] Y. Yang et al., "Distinguishing the viability of a single yeast cell with an ultra-sensitive radio frequency sensor," Lab Chip, vol. 10, pp. 553-555, 2010.

[10] S. Liu et al., "Broadband dielectric spectroscopy calibration using calibration liquids with unknown permittivity," 84th ARFTG Microwave Measurement Conference, Boulder, CO, 2014, pp. 1-5.

[11] T. Markovic et al., "An interdigital capacitor for microwave heating at 25 GHz and wideband dielectric sensing of nL volumes in continuous microfluidics," Sensors, vol. 19, no. 3, pp. 715, 2019.

[12] J. Bao et al., "A 20-GHz microwave miniaturized ring resonator for nL microfluidic sensing applications," IEEE Sensors Letters, vol. 3, no. 6, pp. 1-4, 2019.

[13] S. Liu, I. Ocket, M. Cauwe, D. Schreurs and B. Nauwelaers, "Sensitivity analysis of broadband on-wafer dielectric spectroscopy of yeast cell suspensions up to 110 GHz," IEEE Microw. Wireless Compon. Lett., vol. 25, no. 3, pp. 199–201, 2015.

[14] T. Chen, D. Dubuc, M. Poupot, J-J. Fourni, and K. Grenier, "Accurate nanoliter liquid characterization up to 40 GHz for biomedical applications: toward noninvasive living cells monitoring," IEEE Trans. Microw. Theory Techn., vol. 60, no. 12, pp. 4171–4177, 2012.

[15] H. Li, X. Ma, Y. Ning, X. Cheng, and J. C. M. Hwang, "Correlation between morphology change and microwave property during single-cell electroporation," 2017 IEEE MTT-S Int. Microw. Sym. (IMS) Dig., Honololu, HI, 2017, pp. 869–871.

[16] T. Markovic, I. Ocket, B. Jones and B. Nauwelaers, "Characterization of a novel microwave heater for continuous flow microfluidics fabricated on high-resistivity silicon," 2016 IEEE MTT-S Int. Microw. Symp. (IMS), San Francisco, CA, 2016, pp. 1-4.

[17] T. Markovic, S. Liu, I. Ocket and B. Nauwelaers, "A 20 GHz microwave heater for digital microfluidic", Int. J. Microw. Wirel. T., vol. 9, no. 08, pp. 1591-1596, 2017.

[18] M. Zhang et al., "Broadband interferometric dielectric spectroscopy for aqueous solutions," 2017 IEEE MTT-S International Microwave Workshop Series on Advanced Materials and Processes for RF and THz Applications (IMWS-AMP), Pavia, 2017, pp. 1-3.

[19] Y. Cui, J. Sun, Y. He, Z. Wang, and P. Wang, "A simple, tunable, and highly sensitive radio-frequency sensor," Appl. Phys. Lett., vol. 103, no. 6, pp. 062906, 2013.

[20] S. Liu, I. Ocket, P. Barmuta, A. Lewandowski, D. Schreurs and B. Nauwelaers, "Broadband dielectric spectroscopy calibration for microliter samples of biogenic liquid," 2014 44th European Microwave Conference, Rome, 2014, pp. 279-282.

Energy Harvesting on Power Amplifiers Based on Application of Thermoelectric Generators

Ivan Škalic[1], Ivan Marinović[1]
[1] University of Split, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Split, Croatia
iskalic@fesb.hr, imarin@fesb.hr

***Abstract* – This paper proposes a system for energy harvesting or converting wasted energy (heat) into electric energy using solid state thermoelectric generators (TEGs). As a source of heat, wasted energy on output power transistors of PA (power amplifier) in AB-class were used. Comparative performance analysis of two different commercially available TEGs was made in terms of thermal and internal electrical resistance, generated voltage and power. The main drawback which restricts wider usage of TEGs are their limits on the power conversion efficiency. Obtained result shows only 2.23% of successfully converted energy. In addition to measurements, further improvements of the harvesting setup were proposed.**

Keywords – thermoelectric generators; energy harvesting; solid-state devices

I. INTRODUCTION

TEGs are solid state devices which convert a heat into electric energy. Their main advantages are extended maintenance-free durability and noiseless operation without moving parts. Performances of TEGs are strongly influenced by temperature distribution inside the materials and thermal gradient which can be achieved between the hot and cold junctions of thermocouples [1]. Despite their low power conversion efficiency, TEGs find their application in the field of low-power electronics as alternative power supply. To obtain temperature gradient (and thus power) it's essential to heat the hot junction of module continuously and to dissipate the heat from the cold side of module at the same time [2]. Proposed cooling system consists of air-cooled heat sink and a PA in AB-class was used as the source of heat.

In this paper two different commercially available generators were used for measurements and comparison. Chinese origin generator in the future text will be labeled as TEG „A“ and Russian as TEG „B“.

A. Related works

There are numerous research papers with topics on parameters, applications and performances of TEGs in different conditions. Most similiar application of TEGs compared to the topic of this paper can be found in [3] and [4], where the TEGs are used with PA in class E and F, respectively. In [5] the numerical background of PA energy conversion using TEGs is emphasized.

B. Objective of proposed work

Two main goals are underlined in this paper. First is to measure the amount of successfully converted heat energy produced by output transistors of PA. Second goal is to establish same conditions and compare performances of TEGs produced by different manucfacturers.

II. EXPERIMENTAL SETUP

Proposed thermoelectric energy conversion setup is shown on Fig. 1 and 2. The system consists of 100W/4Ω AB-class PA with it's complementary output transistors Toshiba 2SC5200 and 2SA1943 (source of heat) (1) mounted side by side on aluminium plate (2) (dimensions 43x52.5x5mm). Plate needs to be made of materials with high thermal conductivity and good mechanical properties, since it acts like uniform heat distribution spreader [2]. Top and bottom layers of TEGs are made of ceramics to provide protection for semiconductor materials inside. TEG (3) (dimensions 40x40x3.5mm) is „sandwiched“ between aluminium plate and heat sink (4). Purpose of heat sink is to transfer heat from the cold side of module into ambient, preventing the temperature of cold side from increasing. Size of selected heat sink depends on the available cooling solutions. For our purposes the only requirement in choosing size of the heat sink was to have a cooling fan (5) of adequate power which can effectively cool down the heat sink in a short period of time. Using small heat sinks is not preferable since the high temperature difference is not achievable.

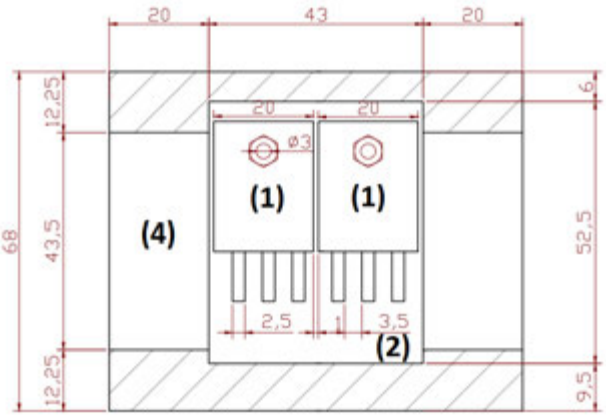


Figure 1: Schematic of output transistors attached to metal plate

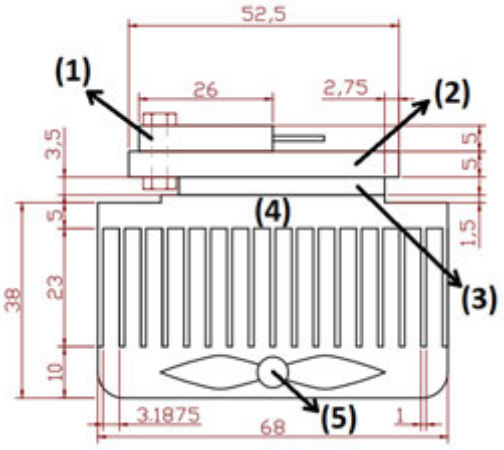


Figure 2: Schematic side view of proposed energy harvesting setup

All heat transfering surfaces of the system are coated with thermoconducting paste in order to reduce overall thermal resistance. There is also significant importance of having uniform pressure distribution on TEG surfaces in order to reduce thermal resistance between components of the system [6]. TEGs used in the setup are shown on Fig. 3 while experimental setup is on Fig. 4.

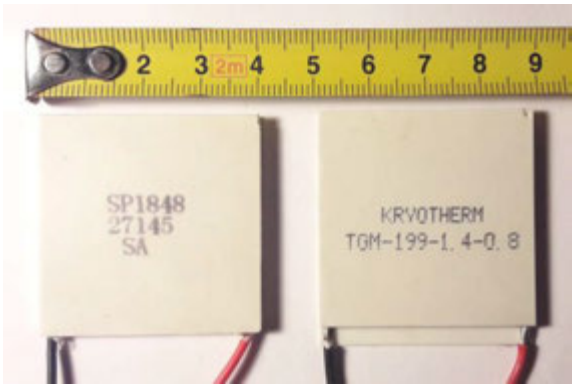


Figure 3: TEGs used in measurements (TEG „A“ and TEG „B“)

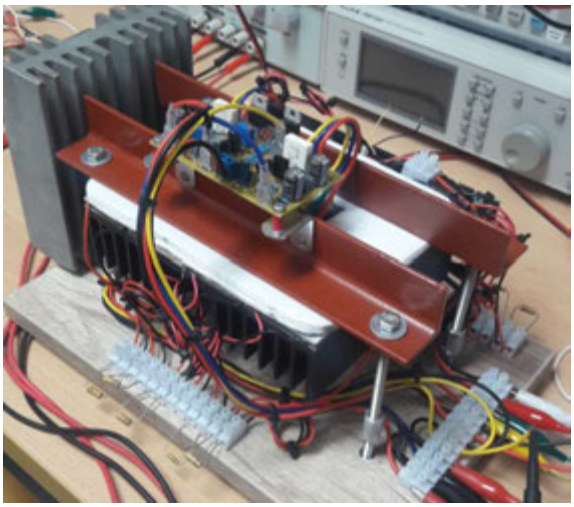


Figure 4: Proposed experimental setup for energy harvesting

III. TEGs Characteristics

Performances of each TEG were measured under the same conditions. Parameters of interest for comparing are achieved temperature difference (ΔT) versus time, open circuit voltage (U_{OC}) and internal electrical resistance (R_{IN}) versus achieved gradient between junctions. Each TEG has same dimensions (as mentioned in Chapter 2) and consists of thermocouples made of the same material, Bismuth telluride (Bi_2Te_3).

Results obtained from Fig. 5 and Fig. 6 show higher thermal resistance of TEG „A“ (ΔT of 58°C achieved in 10:30min) in comparison to TEG „B“ (ΔT of 38°C achieved in 14:38min). Graphs end when the temperature T_{HOT} reaches the value of 150°C (the maximum operating temperature of TEGs and output transistors). Thermal resistance is the parameter which has significant impact to all other obtained graphs. For setup used in this paper, higher thermal resistance of TEG is preferable feature.

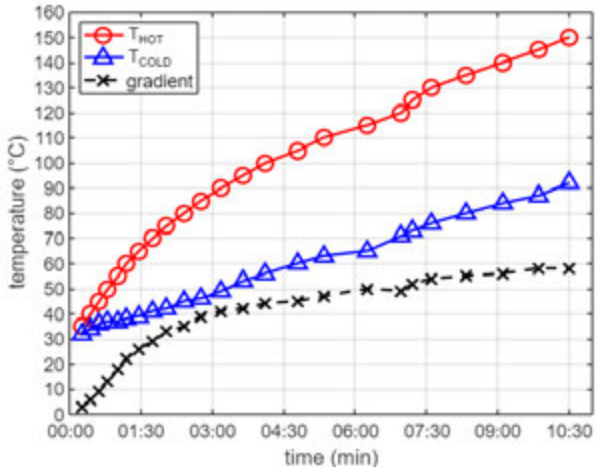


Figure 5: Temperature variations as a function of time for TEG „A“

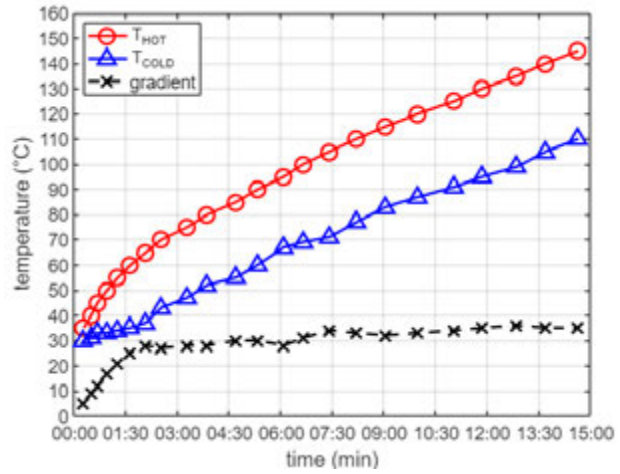


Figure 6: Temperature variations as a function of time for TEG „B“

Results obtained from Fig. 7 show better performance of TEG „A“ (generates 3.77V for ΔT of 58°C) compared to TEG „B“ (generates 2.11V for ΔT of 35°C) in terms of generated open-circuit voltage U_{OC}.

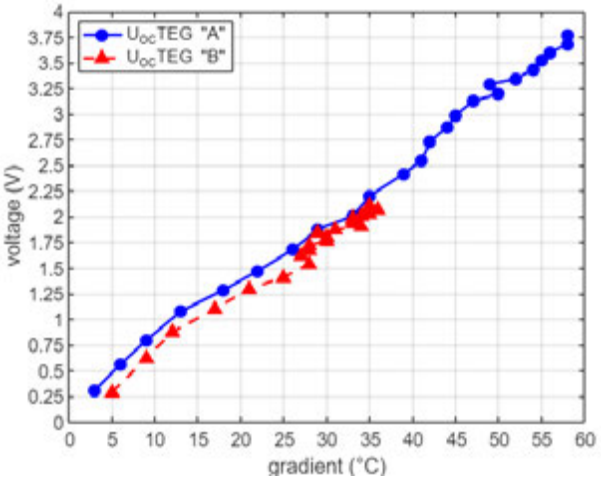


Figure 7: Open-circuit voltage as a function of achieved gradient

Maximum power is generated when the load resistance is equal to the internal resistance of the source. Since the internal resistance of generators changes according to changes in temperature gradient, the usage of a maximum power point tracker (MPPT) to increase the electrical power output is recommended [7]. Calculated internal resistances of each TEG plotted as a function of achieved temperature gradient are shown on Fig. 8. Increase of ΔT causes increase of R_{IN}. Values of internal resistance go up to 5.75Ω and 4.75Ω at the peak of gradient for TEG „A“ and „B“, respectively.

In this paper maximum power transfer didn't occur since all measurements were made using load resistor of fixed value (5.1Ω).

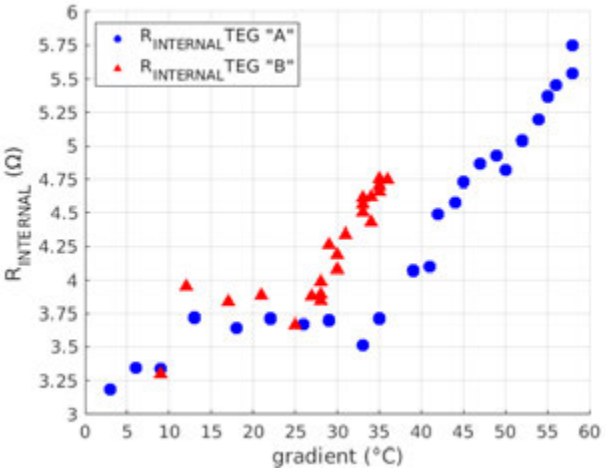


Figure 8: Dependence of internal resistance on the achieved gradient

IV. Heating And Cooling Cycles

Ideal situation for generating power out of TEG is to establish constant amount of heat flow over time. In such case constant amount of power would be generated, but it requires more complex numerical calculations in terms of selecting suitable heat sink. Those calculations are not subject of this work, so the most used practical method to achieve highest temperature gradient is to use cooling and heating cycles.

Cycling regime is set up so that the system is heated up until the hot junction of the module reaches temperature of 140°C, not going any further to avoid risk of damaging system components, altough some researchers concluded that the module tolerates temperature peaks well [8]. After the predetermined temperature value of hot side is reached, the cooling fan turns „on“ and draws air from the heat sink, so it conductes heat away from the cold side of module. Cooling fan stays „on“ until the generated voltage falls into saturation and then it turns „off“. Saturation voltage (marked with green) presents peak voltage value, when voltage rise slope becomes almost horizontal. Time during cooling fan „on“ state is called cooling cycle and during „off“ state heating cycle.

Cycling regime for each generator is shown on Fig. 9, graph presents generated voltages across resistor (U_R) as a function of time. It is obvious that TEG „A“ has better performance compared to TEG „B“ in terms of generated voltages and cycle frequency. Difference in frequency of cycles are associated with thermal resistance of each TEG. For lower value of thermal resistance, more heat pass through TEG to heat sink and therefore more time is needed to heat up aluminium plate (hot side of TEG) to temperature of 140°C, value when cooling cycle start.

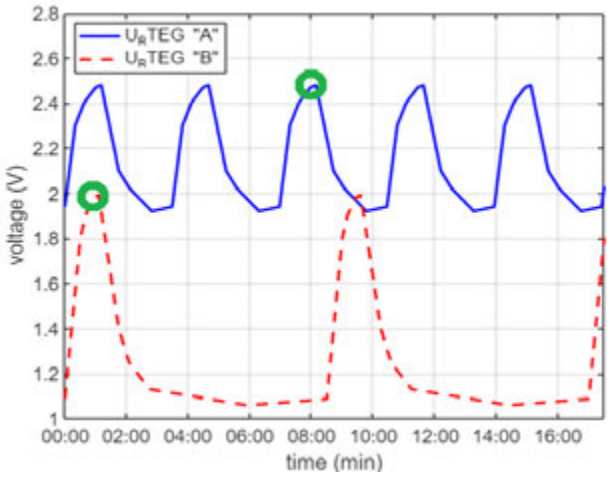


Figure 9: Generated voltage across the resistor as a function of time

Using equation derived from Joule's law, generated power was calculated and displayed as a function of time for each generator (Fig. 10 and 11). Obtained temperature differences during thermal cycles were also measured and displayed on the same graph. Mean value of generated power across resistor (P_{AVG}) is marked with dashed line. Results obtained from graphs show better performance of TEG „A“ in terms of generated power (generates P_{AVG} of 0.91W) compared to TEG „B“ (P_{AVG} equal to 0.31W).

Used PA consumes 140.8W from laboratory DC power supply in order to maintain effective value of 100W on the output load resistor (aluminium shell wire-wound 4Ω resistor). Assuming that all wasted energy appears on output transistors in the form of heat, using TEG „A“ it is possible to successfully convert 2.23% of wasted energy.

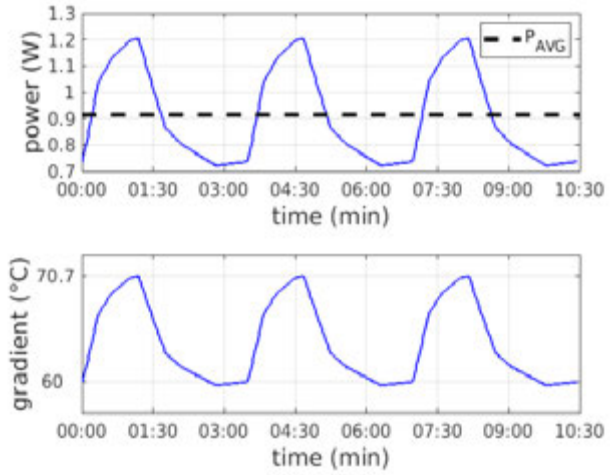


Figure 10: Generated power and temperature difference as a function of time for TEG „A“ during thermal cycles

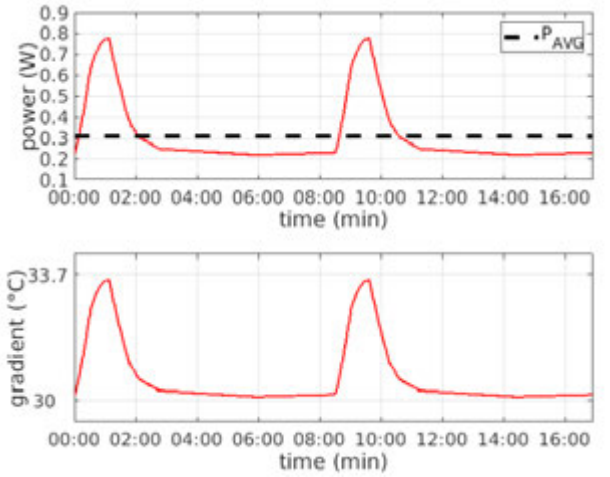


Figure 11: Generated power and temperature difference as a function of time for TEG „B“ during thermal cycles

Further improvement of the system is to measure the exact amount of wasted energy dissipated by output transistors, since in this paper the power consumption of other components in the circuit was neglected compared to wasted heat on output transistors.

V. CONCLUSION

An experimental harvesting set-up for wasted thermal energy has been established. Key elements of the system are solid state thermoelectric generators (TEGs). They convert thermal energy into electric energy according to established temperature difference (ΔT) between hot and cold junctions. Comparative performance analysis of two different comercially available TEGs was made. They were compared in terms of their thermal and internal electrical resistance, generated voltage and power. Results differ due to different thermal resistance of each TEG. As a source of heat, output transistors of 100W/4Ω power amplifier in AB-class were used. Cold side of the module was attached to the air cooled heat sink for dissipating heat from cold side into ambient. In order to achieve highest temperature gradient, thermal cycles (heating and cooling) were used. Experimental results show that the generator is able to successfully convert 2.23% of wasted energy into electric energy. Assumption that total wasted energy appears on output transistors in form of heat is considered.

The research is ongoing and the future focus will be on improving cooling system, implementing MPPT and using larger number of TEGs electrically and thermally connected in serial and parallel combinations.

REFERENCES

[1] C. Veri, M. Pasca and others „A flexible thermoelectric generator with a fully electrical, low startup voltage and high efficiency DC-DC converter“, 6th International Workshop on Advances in Sensors and Interfaces (IWASI), 2015., Gallipoli, Italy

[2] J. G. Haidar, J. I. Ghojel „Optimization of the thermal regime of thermoelectric generators in waste heat recovery applications“, 21st International Conference on Thermoelectrics. Proceedings ICT ’02, 2002., Long Beach, California, USA

[3] K. Niotaki, A. Georgiadis, A. Collado „Thermal energy harvesting for power amplifiers“, IEEE Radio and Wireless Symposium, 2013., Austin, Texas, USA

[4] T. Qi, S. He „Further efficiency improvement of power amplifiers using thermal energy harvesting“, IEEE Transactions on Industrial Electronics (Volume: 66, Issue: 12, Dec. 2019)

[5] K. J. Kim, M. Hodes „Thermoelectric energy scavenging from waste heat of power amplifier transistors“, 15th International Workshop on Thermal Investigations of ICs and Systems, 2009., Leuven, Belgium

[6] T. Kaktus „Design and construction of high temperature thermoelectric power generator module characterisation system“, thesis, Institute for Superconducting and Electronic Materials, University of Wollongong, 2015.

[7] N. D. Trip, A. Şchiop, A. Burca „Considerations on the use of a MPPT circuit for a thermoelectric generator“, 14th International Conference on Engineering of Modern Electric Systems (EMES), 2017., Oradea, Romania

[8] K. Ivanov, A. Aleksandrov „Method of study of thermoelectric generators", 16th Conference on Electrical Machines, Drives and Power Systems (ELMA), 2019., Varna, Bulgaria

Doherty Power Amplifier in Horizontal Current Bipolar Transistor (HCBT) Technology

Željko Osrečki*, Josip Žilak*, Marko Koričić*, and Tomislav Suligoj*
*Department of Electronics, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
zeljko.osrecki@fer.hr

***Abstract* — The first RF power amplifier (PA) in low-cost Horizontal Current Bipolar Transistor (HCBT) technology is demonstrated in this paper. Doherty PA is designed for the modern wireless communication devices at 2.4 GHz operating with high Peak-To-Average-Power-Ratio (PAPR) signals. Microstrip technology is used for the design of matching circuits, phase shift lines, and bias networks, with very few discrete components for RF signal coupling and low-frequency bypassing. The HCBT Doherty PA provides output power of up to 22 dBm with collector efficiency of 45% over 4 dB of output power back-off (OPBO). Furthermore, the efficiency higher than 41.5% and output power of 19 dBm +/- 1.5 dB is achieved at 4-dB OPBO over the bandwidth of 100 MHz.**

***Keywords* — Horizontal Current Bipolar Transistor (HCBT), wireless communications, power amplifiers, Doherty.**

I. Introduction

As recent developments in wireless communication standards push radiofrequency (RF) circuits into the milimeter-wave frequency range, most of the wireless communication remains in the sub-6 GHz range. RF front-ends need to satisfy extreme performance goals dictated by modern wireless communications [1], the most difficult being the transmission of low distortion signals with high efficiency.

RF power amplifiers are the components consuming the highest amount of power both in mobile handsets and base stations, in the former determining the life span of a battery while in the latter increasing the cost of the cooling system. The most important figure-of-merit of an RF power amplifier used in modern communications is efficiency. The goal is to amplify the signals with high fidelity while operating with high efficiency, in order to increase the life span of the battery and/or reduce the cooling costs. With the advent of wideband communication standards, PAs need to maintain high efficiency over the entire signal modulation bandwidth, and, if possible, over more than one frequency band. The main culprit is the large Peak-to-Average-Power-Ratio (PAPR) of a wideband signal to be amplified, since PAs operate under maximum efficiency only for maximum output power. In order to achieve high efficiency even in the power back-off (PBO) range, PA architectures, which can maintain high efficiency over large PBO range, such as Doherty [2], are becoming increasingly popular, both for handset and base station PA design.

In order to optimize the PA characteristics, high-performance devices must be employed for their implementation, such as SiGe HBTs, III-V HEMTs, or advanced down-scaled MOSFETs resulting in a high cost of implementation. In this paper, a low-cost Horizontal Current Bipolar Transistor (HCBT) technology is used for the implementation of Doherty PA for the first time. The large-signal performance of HCBT devices is demonstrated at 2.4 GHz and the designed PA is aimed to be used as a low-cost alternative to the existing Si or SiGe PAs. The adopted design methodology relies on empirical data gathered using load-pull setup rather than the use of nonlinear transistor simulation model and is presented in Section II. In Section III, the measurement results are presented verifying the design methodology.

This work has been fully supported by Croatian Science Foundation under the project IP-2018-01-5296.

II. Design Methodology

A. Horizontal Current Bipolar Transistor (HCBT)

HCBT is a state-of-the-art low-cost silicon bipolar transistor technology integrated with 180 nm CMOS by adding minimal process complexity [3]. Due to a very compact structure and optimized doping profiles, HCBT outperforms all other silicon lateral transistors and can be used as a low-cost, high-performance alternative to conventional BiCMOS technologies. The optimized HCBT has a peak cut-off frequency (f_T) and maximum frequency of oscillations (f_{max}) of 51 and 61 GHz, respectively, whereas the $f_T BV_{CEO}$ product equals 173 GHz·V, which is among the highest-performance Si BJTs. Also, devices with high breakdown voltages have been successfully demonstrated [4] thus broadening the application scope of the technology to high-voltage circuits. High-performance RF circuits are already demonstrated using HCBT technology, such as down-converting mixers and divide-by-2 static frequency dividers [5]. Therefore, HCBT technology allows PA design with the possibility of low-cost integration with CMOS, enabling fully integrated RF front-ends for sub-6 GHz frequency range. In this paper, discrete HCBTs packaged in a 4-pin SOT343 package are used.

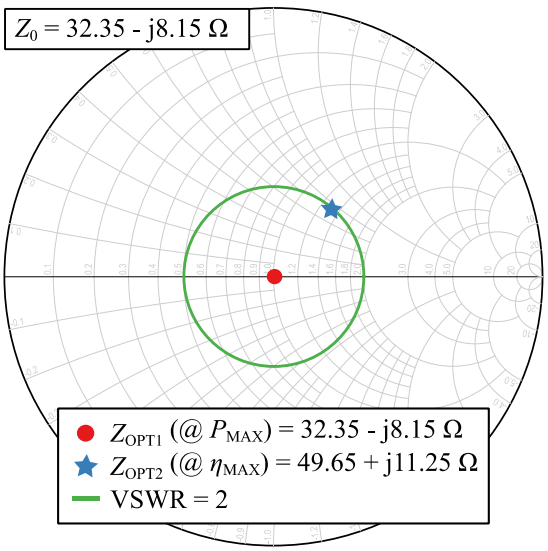


Fig. 1. Matching impedances for maximum output power and maximum collector efficiency obtained using load-pull setup. The load impedance is transformed from Z_{OPT1} at full power to Z_{OPT2} at 6 dB OPBO. The impedance for maximum efficiency is chosen on the VSWR = 2 circle centered on Z_{OPT1}.

B. *Doherty PA and Load-Pull Analysis*

Doherty PA consists of two amplifiers, Carrier PA and Peaking PA, wherein the Peaking PA actively modulates the load impedance of Carrier PA in order to keep it optimally matched in the back-off power range [6]. Carrier PA is biased in Class AB mode and Peaking PA is biased in Class C mode, whereas the input signal is equally split among Carrier and Peaking PAs (3-dB split). At full output power, both PAs contribute equal power to the load, with Carrier PA being terminated with optimal load impedance. At 6 dB OPBO, only Carrier PA is active, and it is again matched to optimal load impedance but with Peaking PA being turned off due to Class C biasing. Therefore, there are two distinct points of operation in Doherty PA, namely, full output power and 6 dB back-off, for which the optimal load impedances need to be found.

Load-pull setup, employing automated mechanical impedance tuners [7], is used to find the optimal output matching impedances for the HCBT device at 2.4 GHz. Optimal impedances at both full power and 6 dB OPBO are shown in Fig. 1. At full output power, both PAs work at maximum power. Load-pull analysis is performed for Class AB biased HCBT, with collector current and collector-emitter voltage of 20 mA and 2.87 V, respectively. A maximum output power of P_{MAX} = 19.5 dBm is measured for Z_{OPT1}, with a collector efficiency of 46%.

At 6 dB OPBO, Peaking PA is turned off and Carrier PA provides P_{MAX} – 3 dBm output power. At the same time, efficiency should be as high as possible, with the peak at 6 dB OPBO. Therefore, the load-pull is performed again to find the impedance for maximum collector efficiency with P_{MAX} – 3 = 16.5 dBm output power.

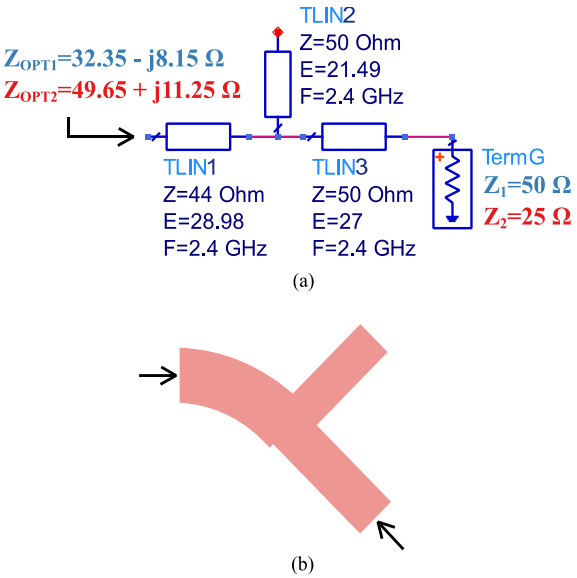


Fig. 2. Output matching network of Carrier PA. (a) $TLIN_1$ and $TLIN_2$ transform the load impedance at full output power (Z_1 = 50 Ω) to Z_{OPT1}. With added $TLIN_3$, the load impedance Z_2 = 25 Ω is transformed to Z_{OPT2} at 6 dB OPBO. (b) Layout of the output matching network with curvature added to facilitate power combining at the output.

C. *Carrier PA Output Matching Network*

The role of Peaking PA is to modulate the load impedance presented to Carrier PA. This modulation sets the load impedance to 50 Ω at full drive, where both PAs contribute equal output power. Thus, Carrier PA's output matching network (OMN) must transform 50 Ω to Z_{OPT1}. The matching network is designed in Advanced Design System (ADS) using ideal transmission lines (Fig. 2a). The necessary transformation from 50 Ω to Z_{OPT1} is achieved with lines $TLIN_1$ and $TLIN_2$. On the other hand, at 6 dB OPBO, the load impedance is 25 Ω since the load is not modulated by the Peaking PA, which is turned off at this point. The system impedance of 50 Ω is transformed to 25 Ω using quarter-wave line with Z_0 = 35.35 Ω. In order to keep Carrier PA at maximum possible efficiency at back-off, $TLIN_3$ is added to match the output of the Carrier PA to Z_{OPT2} at 6 dB OPBO. Since the characteristic impedance of $TLIN_3$ is 50 Ω, Carrier PA will not mismatch the device at full power since the load impedance at full power is also 50 Ω. The network is optimized in electromagnetic (EM) simulator. The layout of the Carrier PA's OMN is shown in Fig. 2b.

D. *Peaking PA Output Matching Network*

At full output power, Peaking PA provides the same power as Carrier PA, thus it should be matched to its optimal impedance for P_{MAX} output power. Regardless, output matching network is the same as the one for Carrier PA. Such arrangement provides equal phase shift at the output in both Carrier and Peaking branches of the Doherty PA. Furthermore, input matching network is

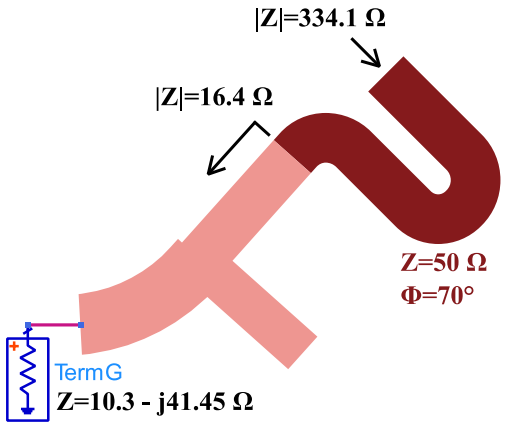


Fig. 3. Output matching network of Peaking PA. Matching network is the same as for Carrier PA except the 50-Ω line used for impedance transformation.

designed as a complex-conjugate match and is the same for both PAs.

For output power up to 6 dB OPBO, only Carrier PA provides output power while Peaking PA is turned off. Maximum efficiency in this power region is achieved at 6 dB OPBO with Carrier PA matched with Z_{OPT2}. Since the PAs are connected in parallel at the output, the impedance of Peaking PA should be as high as possible so as not to change the load impedance (25 Ω). For that purpose, the small-signal output impedance is extracted from *S*-parameter measurements of HCBT in Class C bias, which is set by base-emitter voltage, collector current, and collector-emitter voltage of 0 V, 0 mA, and 2.87 V, respectively. Measured output impedance at 2.4 GHz is 10.3 – j41.45 Ω, the absolute value of which equals 42.7 Ω. With the output matching network, impedance is very low, equalling 16.4 Ω (Fig. 3). Therefore, the output impedance of Peaking PA is low at the frequency of operation and would lower the 25 Ω at the output significantly thus degrading the efficiency at 6 dB OPBO. To maximize the output impedance of the Peaking PA, a 50-Ω phase shift line is placed after the OMN to transform the impedance to the maximum possible value, which resulted in 334.1 Ω. Because of the added line, additional phase shift of 70° is introduced at the output of the Peaking PA branch.

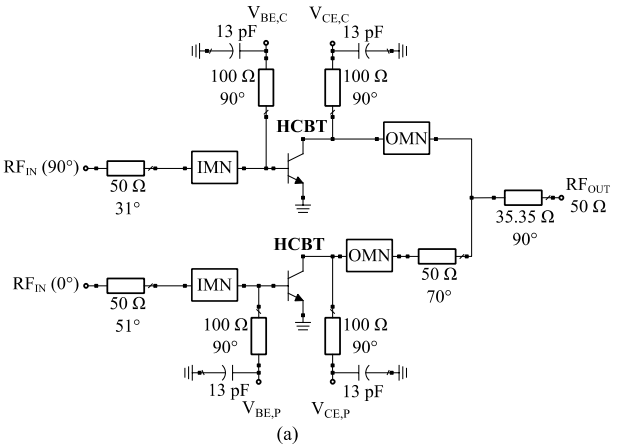
E. Power splitting and input phase shift

The schematic of the Doherty PA is shown in Fig. 4a. Input power is split using 3-dB hybrid coupler with 90° phase shift between the output signals, where the 90° signal is brought to the input of Carrier PA. Since the signals from Carrier PA and Peaking PA must be in-phase at the combining point, additional phase shift, introduced by the impedance transformation at the output of the Peaking PA, is countered by input delay lines. Phase shift (Z_0 = 50 Ω) lines are placed at the input of Carrier and Peaking branches, with 31° and 51° phase shift, respectively, thus equalizing the phase shift in both branches and enabling in-phase addition of signals at the combining point. During the design, phase shifts of Class AB and Class C PAs are taken to be equal so that same input and output matching networks can be used for both Carrier and Peaking PAs. Equal matching networks introduce equal phase shift at the operating frequency.

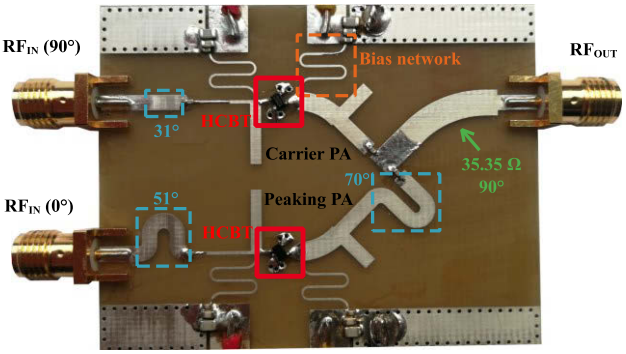
III. Implementation and Measurements

A. Implementation

The HCBT Doherty PA is implemented on a 1-mm-thick FR4 substrate (Fig. 4b). Two emitter pins of the transistors in Carrier and Peaking PAs are grounded using plated through vias. Matching networks, bias networks, and phase shift lines are designed in microstrip technology and the performance is optimized using EM simulator. Bias networks are implemented with quarter-wave lines at the base and collector with 13-pF capacitors providing RF short. Coupling capacitors, employed at the input and at the combining point, are used at the series resonant frequency to ensure a low impedance at the operating



(a)



(b)

Fig. 4. The schematic (a) and fabricated Doherty PA (b) using HCBTs as active devices. Input power is split using hybrid coupler with 90° phase difference. Input delay lines bring signals in Carrier and Peaking branches in-phase at the combining point. Input and output matching networks, IMN and OMN, are the same for the two PAs. Bias networks are implemented using quarter-wave microstrip lines terminated with 13-pF capacitors which provide RF short. Additional capacitors, larger in value, are also used for bypassing the lower frequencies.

frequency. Base and collector bias points of the Carrier and Peaking PAs are separated to enable performance optimization by varying the bias of the Peaking PA during testing.

B. *Collector efficiency and output power*

From the previous analysis, the expected output power at maximum input drive is 22.5 dBm, since both PAs provide 19.5 dBm at their outputs, with collector efficiency of 46%. At 6 dB OPBO, expected output power is 16.5 dBm with collector efficiency of 42%. Collector efficiency as a function of output power is shown in Fig. 5. The amplifier provides maximum efficiency of up to 46%. By varying the base bias of the Peaking PA, $V_{BE,PEAKING}$, the efficiency at full output power can be optimized. The optimum is found for $V_{BE,PEAKING}$ = 0.4 V, where the maximum efficiency range extends from 18 up to 22 dBm. Therefore, the PA is suitable for signals having peak to average power ratio (PAPR) of around 4 dB, whereas a 6 dB OPBO was envisioned during the design. The difference arises mainly due to the bias dependent AM/PM characteristic of the Peaking PA. Since the design was carried out for Class C bias of $V_{BE,PEAKING}$ = 0 V, the length of the output phase shift line of Peaking PA, shown in Fig. 3, should be optimized so that the output impedance is maximum for $V_{BE,PEAKING}$ = 0.4 V, while working at 6 dB OPBO. On the other hand, a flat efficiency characteristic is achieved, with up to 45% collector efficiency over the 4 dB OPBO.

C. *Collector efficiency vs. frequency*

Since the optimum input and output impedances of the HCBT devices are close to 50 Ω, a wideband design is achievable, employing simple microstrip matching networks. The frequency dependence of collector efficiency and output power at 4-dB OPBO is shown in Fig. 6. The collector efficiency is maintained above 41.5% over the bandwidth of 100 MHz. Achieved output power equals 19 dBm +/- 1.5 dB over the entire bandwidth. The amplifier is tested with 5-MHz-wide, WCDMA modulated signal with 3.5-dB PAPR. For output power of 16.5 dBm,

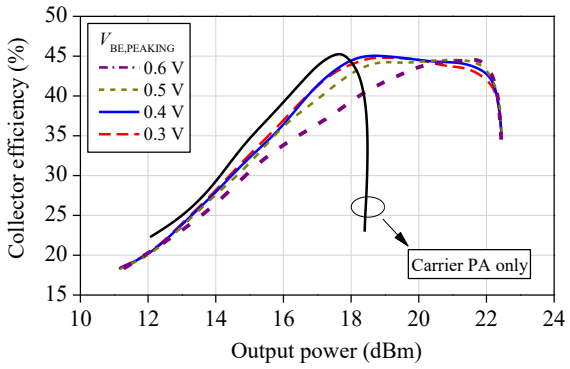


Fig. 5. Measured collector efficiency as a function of output power. Collector efficiency is measured for different base voltages of the Peaking PA with the optimum voltage of 0.4 V, providing a flat efficiency characteristic over 4 dB OPBO.

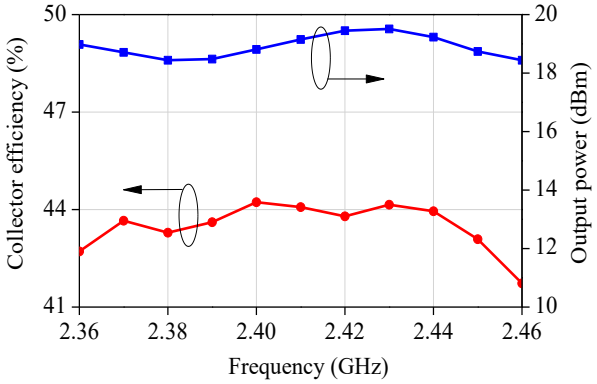


Fig. 6. Measured collector efficiency and output power at 4-dB OPBO as a function of frequency for $V_{BE,PEAKING}$ = 0.4 V.

the collector efficiency of 38.7% is achieved. The performance is comparable with state-of-the-art handset PAs [8].

IV. Conclusion

The Doherty PA is designed at the frequency of 2.4 GHz using high-performance, pure-silicon Horizontal Current Bipolar Transistor (HCBT) technology. The design is based on measured load-pull data of the HCBTs employed in the PA. Implementation is done in microstrip technology on a FR4 substrate with very few discrete components for RF coupling and bypassing. The HCBT Doherty PA provides near 45% collector efficiency over 4 dB of output power back-off range with maximum output power of 22 dBm. The efficiency profile in the back-off power range is easily optimized by varying the base bias of the Peaking PA. Additionally, the bandwidth of 100 MHz is achieved with the collector efficiency above 41.5% and output power of 19 dBm +/- 1.5 dB at 4-dB OPBO. For a 3.5-dB-PAPR WCDMA signal, the efficiency of 38.7% is achieved for output power of 16.5 dBm. Thus, HCBT enables efficient, wideband PA design for high-PAPR modern wireless communication signals.

References

[1] IEEE Standard for Information technology--Telecommunications and information exchange between systems Local and metropolitan area networks--Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE Std 802.11-2016, 2016.

[2] W. H. Doherty, "A New High Efficiency Power Amplifier for Modulated Waves," in *Proceedings of the Institute of Radio Engineers*, vol. 24, no. 9, pp. 1163-1182, Sept. 1936.

[3] T. Suligoj, M. Koričić, H. Mochizuki, S. i. Morita, K. Shinomura and H. Imai, "Horizontal Current Bipolar Transistor With a Single Polysilicon Region for Improved High-Frequency Performance of BiCMOS ICs," in *IEEE Electron Device Lett.*, vol. 31, no. 6, pp. 534-536, June 2010.

[4] M. Koričić, T. Suligoj, H. Mochizuki, S. i. Morita, K. Shinomura and H. Imai, "Double-Emitter HCBT Structure - A High-Voltage Bipolar Transistor for BiCMOS Integration," in *IEEE Trans. Electron Devices*, vol. 59, no. 12, pp. 3647-3650, Dec. 2012.

[5] J. Žilak, M. Koričić, T. Suligoj, H. Mochizuki and S. Morita, "A low-cost 180 nm BiCMOS technology with Horizontal Current

Bipolar Transistor (HCBT) for wireless communication ICs," in 2016 11th European Microwave Integrated Circuits Conference (EuMIC), London, 2016, pp. 373-376.
[6] S. Cripps, "RF Power Amplifiers for Wireless Communications," 2nd ed., Artech House, 2011.
[7] Ž. Osrečki, J. Žilak, M. Koričić and T. Suligoj, "Large-signal characterization of horizontal current bipolar transistor (HCBT) by load-pull measurements," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2018, pp. 0072-0077.
[8] C. Lee and C. Park, “A 2.4 - GHz CMOS power amplifier using a gain and stability enhancement technique for IEEE 802.11N WLAN applications,” in *Microw. Opt. Technol. Lett.*, 58, pp. 2265-2268, June 2016.

Design Considerations for Mid-Power Receiver in Resonant Wireless Power Transfer System

D. Vinko, D. Bilandžija and I. Biondić
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Osijek, Croatia
davor.vinko@ferit.hr

***Abstract* - Resonant wireless power transfer offers high degree of freedom for the receiver positioning, but it also has higher restraints on the design of the receiver circuit. This paper discusses design considerations for the design of the receiver circuit for applications in mid-power range (10-100 W) using resonant wireless power transfer system. Wireless power receiver circuit consists of three main sub-circuits: resonant tank, rectifier circuit and DC-DC converter. The impact of each sub-circuit is analyzed with respect to efficiency and overall performance of the receiver circuit. Results from the mathematical models and simulations are compared with measurements. The final design of the receiver circuit is implemented in LCD TV and evaluated through laboratory measurements.**

Keywords – wireless power transfer; receiver circuit design; efficiency

I. INTRODUCTION

The discussion presented in this paper is based on wireless power transfer system [1] with parallel-serial topology [2], Fig. 1. WPT system consists of transmitter (V, C1, L1) and receiver (L2, C2, R). Transmitter and receiver are magnetically coupled through coils L1 and L2 [3]. The strength of magnetic coupling is given by coupling coefficient (k). The transmitter uses parallel resonant tank (L1, C1), while receiver circuit uses serial resonant tank (L2, C2), hence the name parallel-serial topology of the WPT system [4]. Parallel LC tank at the transmitter side minimizes the supply current from the voltage source (V). At the receiver side, the serial resonant tank is used because it results with lower Thevenin impedance seen by the load. This has a two-fold benefit for mid-power applications. Fig. 2 shows the simulation of the voltage-current and power-current characteristics for both parallel and serial LC resonant tank of the receiver. Maximal voltage is significantly higher for parallel LC tank (dashed blue line) than for serial LC tank (solid red line). Also, the maximal voltage in parallel resonant tank is affected by quality factor of reactive components. This is a problem since the voltage across the load should be constant, i.e. it should vary as little as possible. By using WPT system topology with serial resonant tank at receiver side, the quality factor does not affect the maximal voltage and it has a higher current drivability compared to parallel resonant tank.

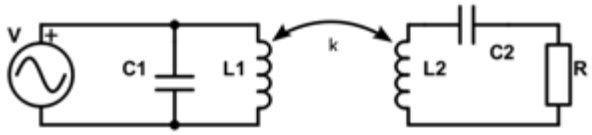


Figure 1. WPT system with parallel-serial topology

When observing maximal output power from the receiver, both topologies have similar output values for a given quality factor, but for serial LC tank the maximal power point is placed at higher current value compared to parallel LC tank, which makes serial LC tank at the receivers end more appropriate for mid-power applications.

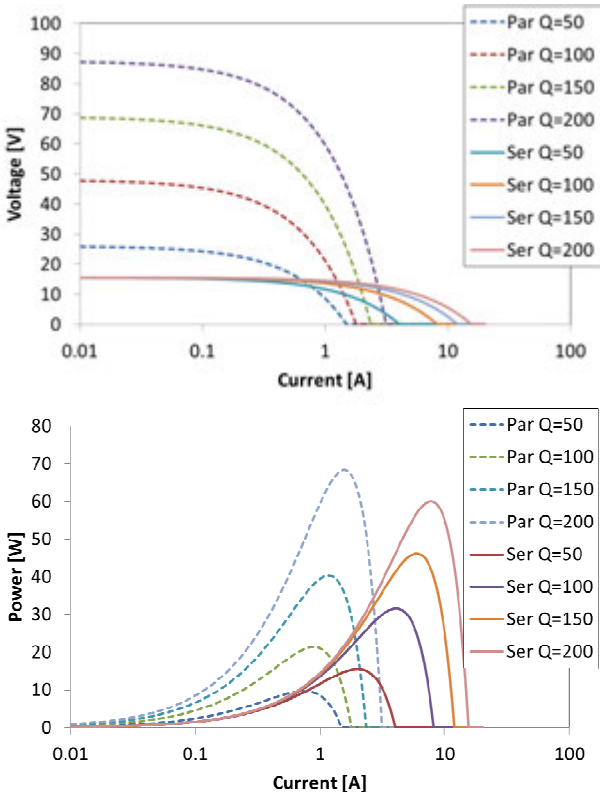


Figure 2. Simulation of voltage-current and power-current characteristics of parallel and serial receiver resonant tank for different quality factors, L_1=L_2=102 µH, C_1=C_2=11 nF, k=0.1, V=150V.

This work has been supported in part by Croatian Science Foundation under the project "Efficient Wireless Power Supply" (UIP-2017-05-5373).

II. Design Considerations for Mid-Power Receiver

There are three main design aspects to consider for a mid-power WPT receiver: resonant tank, rectifier circuit and DC-DC converter. As explained in Introduction, the serial resonant tank is used in receiver design. Resonant magnetic WPT system is characterized by low magnetic coupling between transmitter and receiver [5]. If the coupling coefficient k is under 0.1, the resonant frequency of both the transmitter and the receiver can be tuned to the same resonant frequency [4]. Fig. 3 shows the voltage-current characteristics of serial resonant tank in receiver for different values of coupling coefficient k. There is a linear relationship between maximal voltage and coupling coefficient. To have a constant voltage at the receiver, the coupling coefficient must also be kept constant. The most efficient way to do so is to have uniform magnetic field generated by the transmitter coil [6-9]. With coupling coefficient close to constant, the design of the receiver circuit is significantly simplified. Also, a uniform magnetic field ensures that the value of the coupling coefficient k is dictated by the size of the transmitting and receiving coil and not by their relative position. Such WPT system has a high degree of freedom for receiver positioning with respect to transmitter coil [9-11].

The proposed structure of the receiver circuit is given in Fig. 4. It consists of a resonant tank, followed by rectifier circuit and DC-DC converter. In WPT system, the transmitter coil generates AC magnetic field. The receiver is placed in this field, and then the voltage induces across a receiver coil (L2 in Fig. 4). A capacitor C2 in resonant tank is used to increase efficiency, i.e. to lower the impedance seen by the load. With AC magnetic field, induced voltage is also AC voltage. To accommodate a DC load, induced voltage must be rectified and converted to needed DC level.

Three most common rectifier types are half-wave rectifier, full-wave bridge rectifier and full-wave rectifier with center-tapped coil, Fig. 5. For a mid-power receiver design, only full-wave rectifier topology is suitable. The half-wave rectifier topology does not allow the current flow in both directions through serial LC tank, which is needed to achieve resonance. Both full-wave topologies have pros and cons. Bridge topology has higher voltage drop on the diodes, and center-taped topology needs coil with twice as much windings. But, the main criteria is if the given rectifier topology is compatible with serial resonant tank (L2,C2). In full-wave rectifier topology with

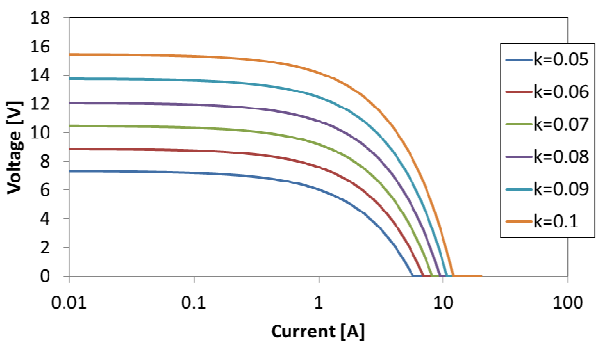


Figure 3. Voltage-current characteristics of serial resonant tank for different values of coupling coefficient k

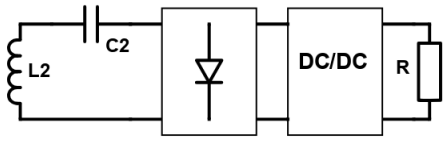


Figure 4. Proposed structure of the receiver circuit

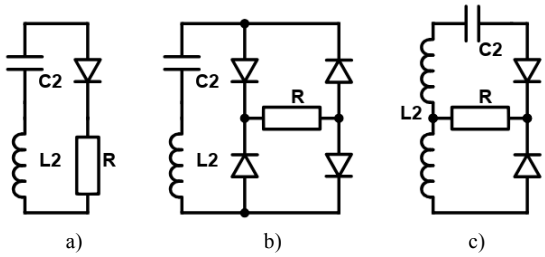


Figure 5. Common rectifier topologies: a) half-wave rectifier, b) full-wave bridge rectifier, c) full-wave rectifier with center-tapped coil

center-taped coil the current through the capacitor C2 is a pulsed DC current, and not AC current as it should be. Therefore, the only suitable rectifier topology is a full-wave bridge topology. With the operating frequency of the WPT system in 100-200 kHz region, the diodes used for the bridge rectifier have to be appropriately selected.

The final building block of the receiver circuit is DC-DC converter. There is a wide range of commercially available DC-DC converters that achieve high efficiency (up to 90%). Different converter topologies (buck, boost, buck-boost, Sepic, Ćuk) are available for different applications and purposes. For this work a commercially available 150 W DC-DC step-up converter is selected, Fig. 6. With serial resonant tank, inductor L2 requires a fairly large number of turns to achieve desired induced voltage. Therefore, a step-up converter is used to boost the rectified voltage. Table I gives the rated values for the DC-DC converter used.

The proposed WPT receiver circuit is designed for a LCD TV with DC power supply (19V, 1.6A). DC-DC converter has an adjustable output voltage (via trimmer resistor) that ranges from 12 V to 32 V and can be set to 19 V with sufficient accuracy. To have a constant 19 V at the output port, the input voltage value (voltage from rectifier) must range from 10 V to 19 V. 10 V is the lowest value of input voltage supported by selected DC-DC converter, and since the step-up converter cannot generate output voltage lower than the input voltage, the maximal input voltage should not exceed 19 V.



V in	10-32 V
V out	12-32 V
I in	10 A (max)
I out	6 A (max)
Size [mm]	65x56.5x23

Figure 6. 150 W DC-DC converter used for receiver circuit

III. MEASUREMENTS

Receiver circuit was built using design considerations given in section II and implemented in LCD TV, Fig.7. Receiver coil is added on the base of the TV stand, and the rest of the receiver circuit is mounted on the back-side of the TV. Receiver coil has trapezoid shape with 15 turns of Litz wire which consists of 100 strands of insulated wire. Each wire strand has a diameter of 0.2mm. The inductance of receiver coil is 102 μH with quality factor of 84 at 100 kHz. Serial capacitors used to form receiver resonant tank are WIMA film capacitors with total capacitance of 11 nF. Resonant frequency of the receiver is then 150 kHz. The transmitter resonant tank is tuned to the same frequency. In Fig.7 the transmitter coil can be seen integrated in plywood stand (white sections) on which the TV is placed. The rectifier is full-wave bridge rectifier with Schottky rectifier diodes rated for 2A continuous forward current. To prevent overheating three diodes connected in parallel are used in each branch of the bridge rectifier.

The efficiency of the resonant WPT system with receiver circuit implemented in LCD TV was evaluated through laboratory measurements. The block diagram of the measured WPT system is given by Fig. 8. The WPT system consists of a transmitter and receiver. Transmitter is powered by a DC power supply V1, and it consists of a current-mode class-D amplifier [12] which is driving parallel resonant tank L1C1. Receiver consists of a serial resonant tank L2C2, full bridge rectifier, DC/DC step-up converter and LCD TV. Four measuring points are marked on Fig. 8. Total power consumption of whole WPT system is measured by measuring voltage and current of DC voltage supply V1. Voltage is measured between points A-A' and the current is measured at point A. The same methodology is used to measure input voltage and input current of rectifier circuit (B, B'), DC-DC converter (C, C') and LCD TV (D, D').

Table I shows the measured values for three different power consumption levels of LCD TV: stand-by, min and max. Stand-by refers to a TV which is turned off but still connected to power supply. Min and max values of power consumption refer to minimal and maximal power consumption levels during normal TV operation. Based on the measurement results given in Table I, the efficiency of each sub-circuit is calculated and the values are given in Table II.



Figure 7. WPT receiver implemented in LCD TV

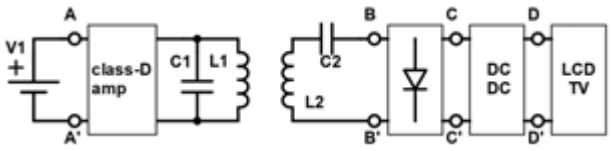


Figure 8. Block diagram of measured WPT system

TABLE I. MEASUREMENT RESULTS

LCD TV power		DC supply A, A'	Rectifier B, B'	DC-DC C, C'	LCD TV D, D'
stand-by	Voltage [V]	60	18.5	17.86	19.07
	Current [A]	32.2 mA	23.9 mA	23.9 mA	7.60 mA
	Power [W]	1.938	0.442	0.427	0.145
min	Voltage [V]	60	15.62	14.68	19
	Current [A]	0.35	1.16	1.16	0.83
	Power [W]	21	18.12	17.03	15.77
max	Voltage [V]	60	14.50	13.42	18.97
	Current [A]	0.55	1.88	1.88	1.22
	Power [W]	33	27.26	25.23	23.14

TABLE II. EFFICIENCY

LCD TV power	η_WPT $(U_{BB'}I_B)/(U_{AA'}I_A)$	η_Rectifier $(U_{BB'}I_B)/(U_{AA'}I_A)$	η_DC/DC $(U_{BB'}I_B)/(U_{AA'}I_A)$	η_Total $(U_{BB'}I_B)/(U_{AA'}I_A)$
stand-by	0.228	0.965	0.340	0.075
min	0.863	0.940	0.926	0.751
max	0.826	0.926	0.917	0.701

TABLE III. POWER CONSUMPTION [W]

DC supply	WPT	Rectifier	DC/DC	TV
1.94	1.50	0.015	0.28	0.145
21	2.88	1.09	1.26	15.77
33	5.74	2.03	2.09	23.14

There are 4 efficiency values given in Table II. WPT efficiency is the efficiency of the wireless power transfer from the DC power supply (A, A') to the rectifier (B, B'). It consists of current-mode class-D amp and resonant tanks of transmitter and receiver. Rectifier and DC-DC efficiencies are measured efficiency values for these specific sub-circuits. Total efficiency is ratio of load power (D, D') and DC supply power (A, A'). Total efficiency during normal LCD TV operation with power ranging from 15 W (min) to 25 W (max) is over 70 %, with maximal measured value of 75.1 %. Both rectifier circuit and DC-DC converter have efficiency values higher than 90%, and the combined efficiency of transmitter circuit and receiver resonant tank ranges from 82.6 % to 86.3 %. The input power to the receiver's resonant tank L2C2 cannot be measured, and therefore the efficiency of resonant tank L2C2 could not be measured. The receiver efficiency (omitting the resonant tank) is between 85% and 87%.

All sub-circuits of the WPT system have idle power consumption even during the stand-by, which is the reason for the low efficiency value when the LCD TV is on stand-by. Then the idle power consumption of the WPT system is significantly higher than the power consumption of the load. Table III gives the power consumption values for each sub-circuit. It can be seen that during stand-by, the idle power consumption of whole WPT system is below 2 W. With increased power consumption of the TV, the power losses also increase for each sub-circuit.

IV. CONCLUSION

A mid-power receiver circuit for resonant WPT system is proposed. The proposed circuit consists of resonant tank, rectifier circuit and DC-DC converter. The analysis of the resonant tank and rectifier circuit topologies has shown that the only suitable configuration is serial resonant LC tank followed by full-wave bridge rectifier. DC-DC converter used in this paper is step-up converter, but different converter types can be used. The proposed receiver circuit is build and implemented in LCD TV. Used LCD TV as rated maximal power of 30.4 W. Laboratory measurements were carried out and the measured power consumption of the LCD TV ranged from 15 W to 25 W depending on the TV operation. LCD TV with implemented receiver for resonant WPT system is successfully tested. Uninterrupted wireless power transfer is achieved by placing the LCD TV anywhere on the transmitter coil (surface with size 100x30 cm). The position of the TV on the transmitter coil can be changed during operation without any impact on performance. The maximal achieved efficiency of the whole wireless power transfer system is measured at 75.1% and the maximal efficiency of the receiver circuit without the losses in resonant tank is 87%.

REFERENCES

[1] AM. Jawad, R. Nordin, SK. Gharghan, HM. Jawad, M. Ismail, "Opportunities and Challenges for Near-Field Wireless Power Transfer: A Review," in *Energies*, vol. 10, no. 7, 2017. doi: 10.3390/en10071022

[2] I. Adam, K. A. Kadir, S. Khan, A. Nurashikin and H. Mansor, "Inductive resonant power transfer and topology consideration," *2017 IEEE 3rd International Conference on Engineering Technologies and Social Sciences (ICETSS)*, Bangkok, 2017, pp. 1-5, doi: 10.1109/ICETSS.2017.8324197

[3] A. Kurs, A. Karalis, R. Moffatt, JD. Joannopoulos, P. Fisher, M. Soljačić, "Wireless power transfer via strongly coupled magentic resonances," in *Science*, vol. 317, no. 5834, pp. 83-86, 2007. doi: 10.1126/science.1143254

[4] D. Vinko, I. Biondić and D. Bilandžija, "Impact of Receiver Power and Coupling Coefficient on Resonant Frequency in Wireless Power Transfer System," *2019 International Symposium ELMAR*, Zadar, Croatia, 2019, pp. 207-210. doi: 10.1109/ELMAR.2019.8918807

[5] D. Ahn and S. Hong, "Effect of Coupling Between Multiple Transmitters or Multiple Receivers on Wireless Power Transfer," in *IEEE Transactions on Industrial Electronics*, vol. 60, no. 7, pp. 2602-2613, July 2013. doi: 10.1109/TIE.2012.2196902

[6] X. Liu and S. Y. Hui, "Optimal Design of a Hybrid Winding Structure for Planar Contactless Battery Charging Platform," in *IEEE Transactions on Power Electronics*, vol. 23, no. 1, pp. 455-463, Jan. 2008. doi: 10.1109/TPEL.2007.911844

[7] WS. Lee, HL. Lee, KS. Oh, JW. Yu, "Uniform magnetic field distribution of a spatially structured resonant coil for wireless power transfer," in *Applied Physics Letters*, vol. 100, no. 21, 2012. doi: 10.1063/1.4719585

[8] D. Bilandžija, D. Vinko and I. Biondić, "Achieving Uniform Magnetic Field with Rectangular Coil in Wireless Power Transmission System," *2019 International Symposium ELMAR*, Zadar, Croatia, 2019, pp. 179-182. doi: 10.1109/ELMAR.2019.8918673

[9] M. R. Basar, M. Y. Ahmad, F. Ibrahim and J. Cho, "Resonant inductive power transfer system for freely moving capsule endoscope with highly uniform magnetic field," *2016 IEEE Industrial Electronics and Applications Conference (IEACon)*, Kota Kinabalu, 2016, pp. 393-397. doi: 10.1109/IEACON.2016.8067412

[10] T. Yeo, D. Kim, S. C. Chae, S. Khang and J. Yu, "Design of free-positioning wireless power charging system for AAA rechargeable battery," *2016 46th European Microwave Conference (EuMC)*, London, 2016, pp. 759-762. doi: 10.1109/EuMC.2016.7824454I.

[11] F. Jolani, Y. Yu and Z. Chen, "A planar positioning-free magnetically-coupled resonant wireless power transfer," *2015 IEEE Wireless Power Transfer Conference (WPTC)*, Boulder, CO, 2015, pp. 1-3. doi: 10.1109/WPT.2015.7140176

[12] D. Bilandžija, I. Biondić and D. Vinko, "Impact of Choke Selection on Performance of Current-Mode Class-D Amplifier in WPT System," *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2019, pp. 92-97. doi: 10.23919/MIPRO.2019.8757203

Semi-Analytical Estimation of On-Chip Intertwined Rectangular Transformer Parameters in 180 nm CMOS Technology

Ivan Brezovec*, Josip Mikulić*, Gregor Schatzberger* and Adrijan Barić†
*ams AG, Tobelbader Strasse 30, Premstaetten 8141, Austria
†University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, Zagreb, Croatia
ivan.brezovec@ams.com

***Abstract*—This paper presents the design methodology of an on-chip intertwined rectangular transformer in 180 nm CMOS technology intended for fully integrated LC oscillators. A simple model based on the physical interpretation of the transformer is presented. A method for fast estimation of inductance and magnetic coupling which takes into account the spacing between the turns is presented. Estimation of other parameters is given based on a combination of simple formulas, extraction from electromagnetic simulations and a fitting procedure. Four transformers with different turn ratios are simulated in HFSS and the results of the simulations to the fitted data of the lumped elements model are compared. The model's relative error of S-parameters compared to the simulations is less than 5% for inductors and 10% for the transformer in a high frequency range.**

***Index Terms*—integrated circuits, CMOS, inductor, transformer, inductance, mutual inductance, coupling**

I. Introduction

Last years have shown a need for better on-chip transformer models because some of the recent LC oscillator architectures use them instead of the classical inductive components [1] [2] [3]. The older transformer models proposed in [4] [5] are not completely symmetrical, as shown in Figure 1. The model that will be presented in this paper is a rectangular, intertwined differential transformer.

The modelling of transformers is usually achieved through calculating the parameters of the model [6], extraction of the parameters [7], or a combination of the former methods [8]. The work presented in this paper will be a combination of the before mentioned methods, prioritizing a fast procedure that is able to achieve good fitting between simulated data and the lumped elements model.

A problem in the transformer modelling is the issue of mutual coupling. It can be calculated as in [6], where a method similar to [9] is used, where all mutual inductances are calculated and summed up. A simpler approach is using the formulas developed in [10] on the transformers themselves to estimate the total mutual inductance, as it will be shown in the next section.

Section II covers the inductance and mutual inductance calculation. Section III covers the estimation of the model parameters. Section IV covers the fitting procedure and the comparison between the simulated data and the extracted lumped elements. Section V presents the conclusions.

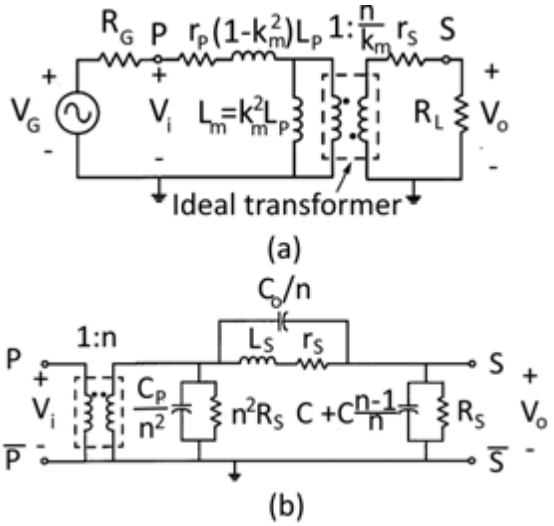


Fig. 1. Circuit schematic of previously common transformer model [5].

II. Inductance Calculation

Typical structures of on-chip inductors are rectangular, octagonal, hexagonal or round. The inductors used for transformers in this work are rectangular and are similar to the inductor in Fig. 2. By intertwining such inductors, transformers similar to the one shown in Fig. 3 can be built.

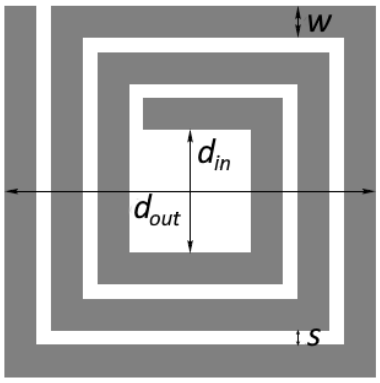


Fig. 2. Geometrical parameters of an inductor [10].

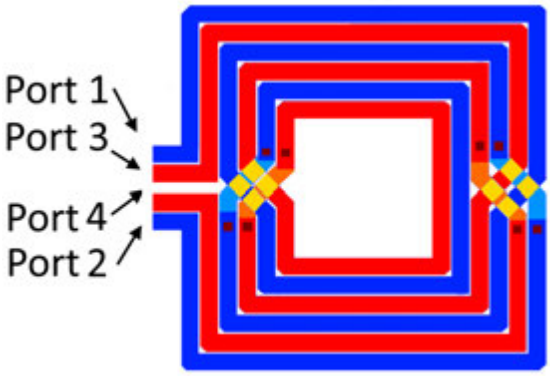


Fig. 3. Example of a rectangular intertwined transformer.

Important parameters for defining the inductor are: the outer diameter d_{out}, the line width w, the spacing between the lines s, and the number of turns n. All of these have to be defined for both the primary and secondary inductor. Two other variables are also introduced for the simplification reasons, namely the average diameter d_{avg} and the fill ratio ρ, defined as:

$$d_{avg} = \frac{d_{out} + d_{in}}{2}, \tag{1}$$

$$\rho = \frac{d_{out} - d_{in}}{d_{out} + d_{in}}. \tag{2}$$

Since the number of turns n is greater than one, the lines have to be intertwined at each turn of the spiral. In this work, the width and spacing of both the primary and secondary inductor will be kept the same. There are multiple ways to calculate the inductance quickly and precisely. A commonly used method for quick estimation, presented in [10], is the Expression based on Current Sheet Approximation. This formula is based on approximating sides of the spirals by symmetrical current sheets of equivalent current densities. The accuracy of this formula is better for $s \leq w$. The inductance can then be calculated by:

$$L = \frac{\mu n^2 d_{avg} c_1}{2} \left(\ln\left(\frac{c_2}{\rho}\right) + c_3\rho + c_4\rho^2 \right). \tag{3}$$

Here, μ is the magnetic permeability and the constants c_1, c_2, c_3 and c_4 are defined by the used shape. The coefficients for rectangular, octagonal and round shapes are shown in Table I [10].

Since (3) is valid for $s \leq w$, the formula is not precise enough to calculate the inductor inductance and all the mutual inductances. That is because the spacing between the turns of an inductor used in a transformer is larger than the width

TABLE I. Coefficients of Current Sheet Expression [10].

shape	c_1	c_2	c_3	c_4
square	1.27	2.07	0.18	0.13
octagon	1.07	2.29	0.00	0.19
circle	1.00	2.46	0.00	0.20

for a transformer with the same primary and secondary inductor widths. This work features only rectangular inductors and in [10], a formula with space correction for rectangular inductors has also been developed:

$$\begin{aligned} L = & \frac{2\mu n^2 d_{avg}}{\pi} \left[\ln\left(\frac{2.067}{\rho}\right) + 0.178\,\rho + 0.125\,\rho^2 + \right. \\ & + 0.5\,\frac{(n-1)\,s^2}{(\rho d_{avg})^2} + 0.178\,\frac{(n-1)\,s}{n d_{avg}} + \\ & \left. + 0.0833\,\frac{(n-1)\,s\,(s+w)}{d_{avg}^2} + \frac{1}{n}\ln\left(\frac{w+t}{w}\right) \right]. \end{aligned} \tag{4}$$

By using (4), the correct inductances of the primary and secondary inductor can be quickly estimated. The quick estimation of the mutual coupling can also be achieved using (4) for common symmetrical cases where the ratio of turns is $1:n$, $n:1$, $n:n$, $n:(n+1)$, or $(n+1):n$. This is done by inserting the average diameter of the transformer and the sum of turns of both the primary and secondary inductor in (4). By doing this, the total mutual inductance of the structure is determined. The total inductance of the structure is the sum of the primary and secondary inductance as well as the double mutual coupling between the primary and secondary inductor:

$$L_{TOT} = L_P + L_S + 2M. \tag{5}$$

Since both the total, primary, and secondary inductance can be calculated using (4), the mutual inductance is easy to determine from (5):

$$M = \frac{1}{2}\left(L_{TOT} - L_P - L_S\right). \tag{6}$$

The coupling coefficient between the primary and secondary inductor, k_m, can then also be determined:

$$k_m = \frac{M}{\sqrt{L_P L_S}}. \tag{7}$$

III. Estimation of Other Parameters

The model for a transformer from Fig. 3 is shown in Fig. 4. It is important to note that the calculated inductances in the previous section are DC inductances. The first estimation of the DC resitance is:

$$R_S = \frac{\rho l}{wt}, \tag{8}$$

Here, ρ is the resistivity $(1/\sigma)$, l is the inductor length, w is the metal width and t is the metal thickness. Skin effect has to be taken into account, and this can be modelled using RL ladders as shown in Figure 4. In the model, L_{Ps}, L_{Ss}, R_{Ps} and R_{Ss} are defined as (X denotes the primary P, or the secondary S):

$$L_{Xs} = K_{LX} L_X, \tag{9}$$

$$R_{Xs} = K_{RX} R_X. \tag{10}$$

Using this, the calculated values of the inductances in the previous section and the DC resistances are equal to:

$$L_{XDC} = L_X + \frac{R_X}{R_X + R_{Xs}} L_{Xs}, \tag{11}$$

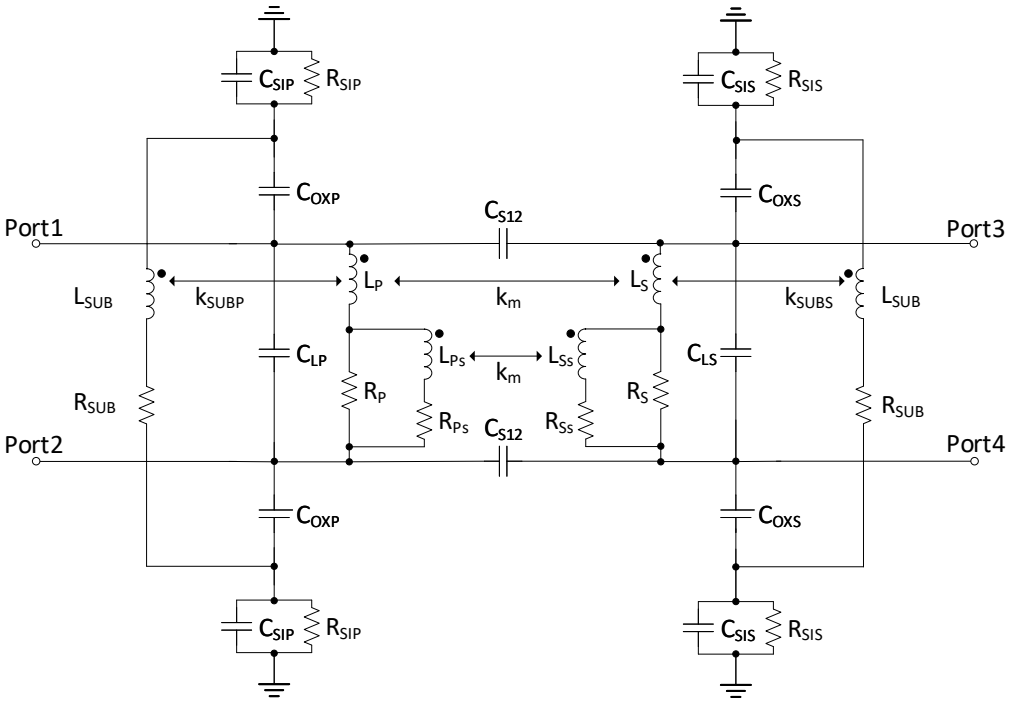


Fig. 4. Circuit schematic of the transformer model.

$$R_{XDC} = R_X || R_{Xs}. \tag{12}$$

The parameters K_{LX} and K_{RX} are hard to estimate and are fitted in this work.

Other parameters also have to be given physical meaning. C_L is approximated as the parallel-plate capacitance between the spiral and the center-tap underpass, C_{OX} is the capacitance between the spiral and substrate, C_{SI} is the parasitic substrate capacitance and R_{SI} is the parasitic substrate resistivity, L_{SUB} and R_{SUB} are the inductance and resistance induced by the eddy currents. The expression for quick estimation of C_L is [10]:

$$C_L = N w^2 \frac{\epsilon_{OX}}{t_{OX}}, \tag{13}$$

Here, N is the number of the overlaps, ϵ_{OX} is the oxide dielectric permittivity, and t_{OX} is the oxide thickness.

With the assumption of a π - model, where Y_1 and Y_2 are side admittances, and Y_3 is the central admittance, Y_1 is defined with the equation:

$$Y_1 = y_{11} + \frac{1}{2}\left(y_{12} + y_{21}\right). \tag{14}$$

Parameters C_{OX}, C_{SI}, and R_{SI} are extracted similar to the work in [8]:

$$C_{OX} = \lim_{\omega \to 0} \frac{\mathrm{Im}(Y_1)}{\omega}, \tag{15}$$

$$R_{SI} = \lim_{\omega \to 0} \mathrm{Re}\left(\frac{1}{Y_1}\right), \tag{16}$$

$$C_{SI} = \frac{\lim_{\omega \to \infty} \frac{\mathrm{Im}(Y_1)}{\omega} C_{OX}}{\omega C_{OX} - \lim_{\omega \to \infty} \frac{\mathrm{Im}(Y_1)}{\omega}}. \tag{17}$$

Regarding the modelling of the eddy currents L_{SUB}, is assumed to be in the $\pm 10\%$ range of the total mutual inductance from (5). The resistance R_{SUB} and coupling k_{SUBX} used to model the eddy currents are additionally fitted after first estimation of other parameters. The parasitic capacitance between the turns C_{S12} is also fitted.

IV. Simulation Results

Four test cases, shown in Fig. 5, are simulated to confirm the model from Fig. 4 and the coupling estimation. The width of the lines in the cases is kept 14 μm, the spacing between the turns 2 μm, and the outer diameter 304 μm. The ratios between the turns of the primary and secondary inductor are: 1:1, 1:2, 1:3, 3:3.

Fully 3-D electromagnetic simulations of the inductors and transformers are performed in HFSS. After first estimation of the lumped elements models, the parameters are fitted using the optimization features in ADS. Ports of the lumped elements model are 1-4, and ports of the electromagnetic simulations are 5-8. The primary ports are 1, 2, 5, and 6, and the secondary ports are 3, 4, 7, and 8. The relative error conditions used to match the S-parameters are:

$$\left|\frac{\mathrm{Re}(S_{i,j}) - \mathrm{Re}(S_{i+4,j+4})}{\mathrm{Re}(S_{i+4,j+4})}\right| < \mathrm{relError}, 1 \leqslant i,j \leqslant 4, \tag{18}$$

$$\left|\frac{\mathrm{Im}(S_{i,j}) - \mathrm{Im}(S_{i+4,j+4})}{\mathrm{Im}(S_{i+4,j+4})}\right| < \mathrm{relError}, 1 \leqslant i,j \leqslant 4. \tag{19}$$

Separate simulations of the primary and secondary inductor are performed in HFSS (the secondary inductor was not drawn when the primary inductors was simulated and vice-versa).

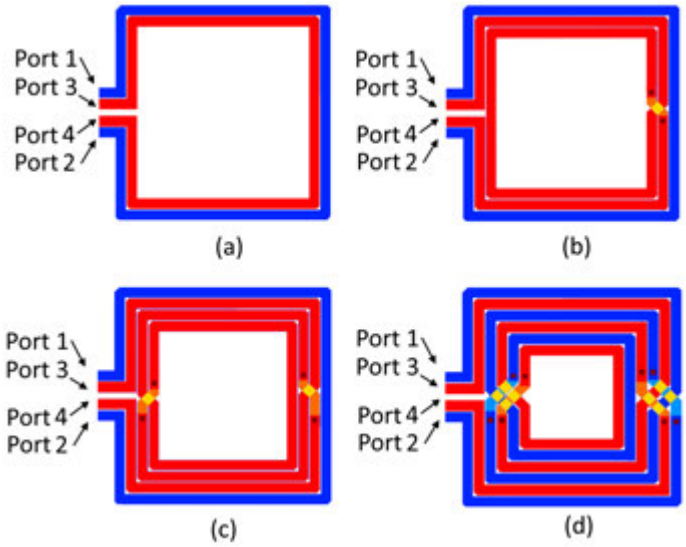


Fig. 5. Tested transformer structures imported with different turn ratios (primary coil is blue, secondary coil is red): (a) primary coil 1 turn, secondary coil 1 turn, (b) primary coil 1 turn, secondary coil 2 turns, (c) primary coil 1 turn, secondary coil 3 turns, (d) primary coil 3 turns, secondary coil 3 turns.

The model parameters of the primary and secondary inductor are extracted based on these separate EM simulations. Also, additional conditions for fitting are used:

$$L = \frac{\mathrm{Im}\left(\frac{1}{y_{11}}\right)}{\omega}, \tag{20}$$

$$Q = \frac{\mathrm{Im}\left(\frac{1}{y_{11}}\right)}{\mathrm{Re}\left(\frac{1}{y_{11}}\right)}, \tag{21}$$

$$\left|\frac{L_{MODEL} - L_{EM}}{L_{EM}}\right| < \mathrm{relError}, \tag{22}$$

$$\left|\frac{Q_{MODEL} - Q_{EM}}{Q_{EM}}\right| < \mathrm{relError}. \tag{23}$$

The achieved relative error for the separate inductors is 5%. After the inductors are fitted, the parameters are inserted into the transformer model. The complete transformer is also separately simulated in HFSS. The optimized parameters are k_m, k_{SUB}, C_{S12}, L_{SUB}, R_{SUB} , K_{LP}, and K_{RP}. The achieved relative error is 10% on a lower frequency range, depending on the transformer. The values of the elements are shown in Table II. At the end of the table, the maximum fitting frequency is specified separately for the inductors and transformers, namely, f_P for the primary inductor, f_S for the secondary inductor, and f_T for the transformer. Since there is a lot of generated graphs, only results of choice are shown. In Fig. 6 the results are shown for the 1:1 transformer, in Fig. 7 for the 1:2 transformer, in Fig. 8 for the 1:3 transformer, and in in Fig. 9 for the 3:3 transformer.

The primary and secondary inductor frequency are fitted almost to the resonance frequency of the inductor, which for inductors having a lower number turns, is a higher frequency

TABLE II. Lumped elements model values.

	1:1	1:2	1:3	3:3
L_P[nH]	0.81	0.81	0.81	2.34
L_S[nH]	0.70	1.73	3.31	2.02
R_P[kΩ]	8.86	8.92	8.92	46.34
R_S[kΩ]	8.86	49.67	38.04	28.14
C_P[fF]	5.17	5.17	5.17	38.36
C_S[fF]	6.10	39.68	58.20	31.53
C_{OXP}[fF]	8.69	8.69	8.69	14.42
C_{OXS}[fF]	8.16	5.87	6.93	9.52
C_{SIP}[fF]	11.94	11.94	11.94	14.42
C_{SIS}[fF]	6.04	9.53	11.23	6.01
R_{RIP}[kΩ]	0.95	0.95	0.95	1.10
R_{RIS}[kΩ]	1.67	1.92	1.95	1.93
K_{LP}	0.10	0.06	0.10	0.38
K_{LS}	0.10	0.22	0.25	0.36
K_{RP}	0.57	0.66	0.65	0.17
K_{RS}	0.57	0.14	0.27	0.32
k_m	0.65	0.60	0.55	0.79
L_{SUB}[nH]	2.87	4.79	6.64	9.42
R_{SUB}[kΩ]	0.75	1.30	1.50	1.54
k_{SUBP}	0.46	0.51	0.50	0.05
k_{SUBS}	0.43	0.34	0.12	0.05
C_{S12}[fF]	3.90	3.62	3.45	4.15
f_P[GHz]	50.0	50.0	50.0	13.5
f_S[GHz]	47.0	16.5	9.0	13.0
f_T[GHz]	26.0	10.0	7.0	7.5

than for higher number turns inductors. For the $1:n$ transformers, the mutual magnetic coupling k_m drops with increase of n. This is because the magnetic flux is lowered due to the lowering of the inner area through which the flux flows with secondary turn increase. In the 3:3 transformer case, the

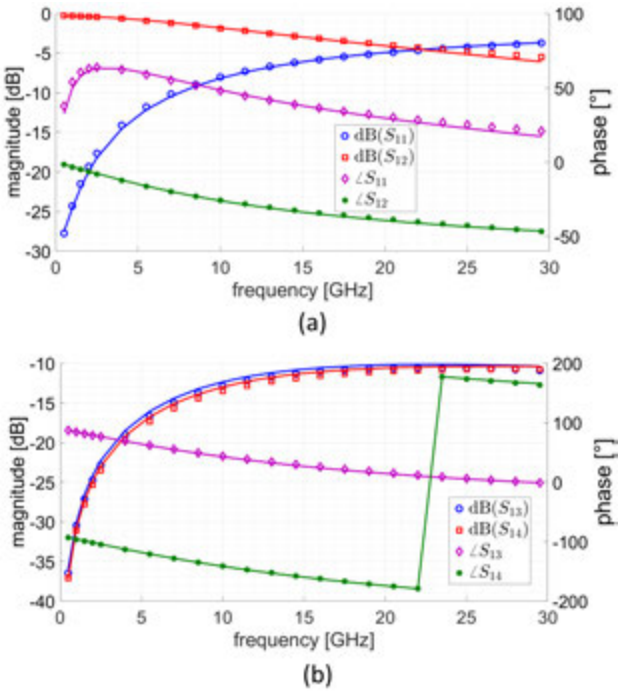


Fig. 6. Rectangular transformer with turn ratio 1:1 (points represent simulated data) (a) Magnitude and phase of S_{11} and S_{12}, (b) Magnitude and phase of S_{13} and S_{14}.

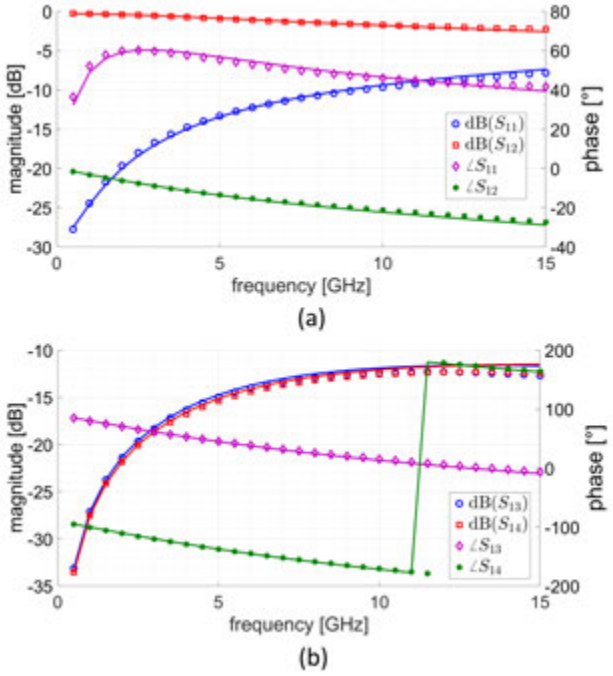


Fig. 7. Rectangular transformer with turn ratio 1:2 (points represent simulated data) (a) Magnitude and phase of S_{11} and S_{12}, (b) Magnitude and phase of S_{13} and S_{14}.

mutual magnetic coupling increases since the primary and secondary turns are intertwined more densely. Also, there is more coupling between the turns because the number of the primary and secondary inductor segments is larger.

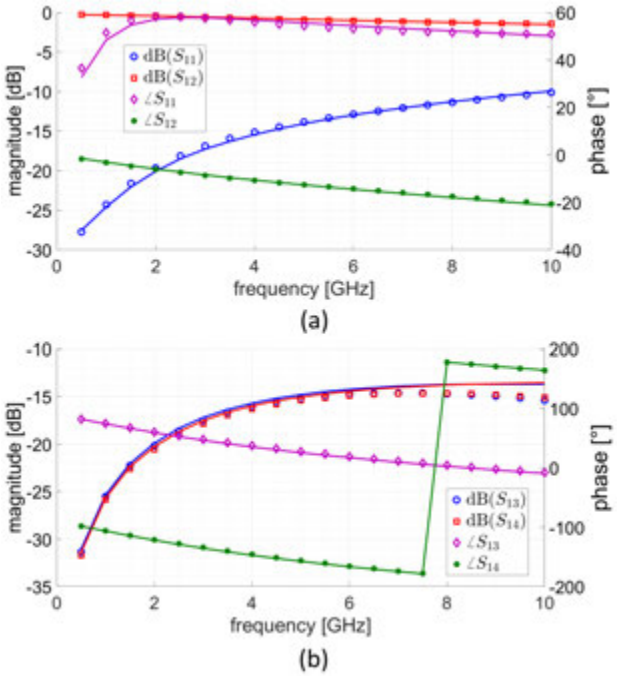


Fig. 8. Rectangular transformer with turn ratio 1:3 (points represent simulated data) (a) Magnitude and phase of S_{11} and S_{12}, (b) Magnitude and phase of S_{13} and S_{14}.

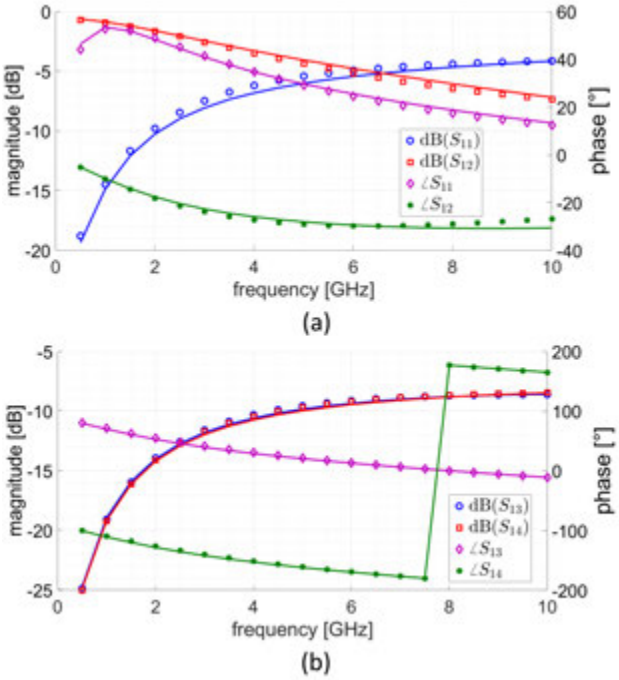


Fig. 9. Rectangular transformer with turn ratio 3:3 (points represent simulated data) (a) Magnitude and phase of S_{11} and S_{12}, (b) Magnitude and phase of S_{13} and S_{14}.

V. Conclusion

One of the critical parameters while modelling transformers is the mutual inductance between the primary and secondary inductor. This works developed a simple methodology to quickly obtain the coupling coefficient between the inductors for rectangular, intertwined transformers, as well as the model of the whole transformer. The methodology is valid for the transformers with common turn ratios, namely $1:n$, $n:1$, $n:n$, $n:(n+1)$, or $(n+1):n$. The first estimation of the model parameters is achieved through a combination of calculations and extraction from 3-D electromagnetic simulations. After the first estimation, the S-parameters of the model are fitted in ADS to the S-parameters of the simulation results obtained in HFSS. Four test cases are analyzed. The width of the lines in the cases is kept 14 μm, the spacing between the turns 2 μm, and the outer diameter 304 μm. The ratios between the turns in the cases are 1:1, 1:2, 1:3, 3:3. The results for the inductors show less than 5% error in a high frequency range. The transformer has less than 10% error in a smaller frequency range than the inductor.

References

[1] M. Babaie, R. B. Staszewski, "A class-F CMOS oscillator", IEEE Journal of Solid-State Circuits, 48(12), 3120–3133., 2013., https://doi.org/10.1109/JSSC.2013.2273823

[2] C. C. Lim, H. Ramiah, J. Yin, P. I. Mak, R. P. Martins, "An inverse-class-F CMOS oscillator with intrinsic-high-Q first harmonic and second harmonic resonances", IEEE Journal of Solid-State Circuits, 53(12), 3528–3539., 2018., https://doi.org/10.1109/JSSC.2018.2875099

[3] F. Wang, H. Wang, "A noise circulating oscillator", IEEE Journal of Solid-State Circuits, 54(3), 696–708., 2019., https://doi.org/10.1109/JSSC.2018.2886321

[4] D. T. S. Cheung, J. R. Long, R. A. Hadaway, "Monolithic Transformers for Silicon RF IC Design", Proceedings of the 1998 Bipolar/BiCMOS Circuits and Technology Meeting (Cat. No.98CH36198), 105–108., 1998., https://doi.org/10.1109/BIPOL.1998.741890
[5] J. R. Long, "Monolithic transformers for silicon RF IC design", IEEE Journal of Solid-State Circuits, 35(9), 1368–1382., 2000., https://doi.org/10.1109/4.868049
[6] W. Gao, C. Jiao, T. Liu, S. Member, Z. Yu, S. Member, "Scalable Compact Circuit Model for Differential Spiral Transformers in CMOS RFICs", 53(9), 2187–2194., 2006.
[7] Z. Gao, K. Kang, C. Zhao, Y. Wu, Y. Ban, K. Sun, L., … Q. Xue, "A Broadband and Equivalent-Circuit Model for Millimeter-Wave On-Chip M:N Six-Port Transformers and Baluns", IEEE Transactions on Microwave Theory and Techniques, 63(10), 3109–3121., 2015., https://doi.org/10.1109/TMTT.2015.2466549
[8] J. Kang, L. Sun, J. Wen, J., M. Zhao, "An equivalent lumped-circuit model for on-chip symmetric intertwined transformer", ASICON 2009 - Proceedings 2009 8th IEEE International Conference on ASIC, 674–677., 2009., https://doi.org/10.1109/ASICON.2009.5351336
[9] H. Greenhouse, "Design of Planar Rectangular Microelectronic Inductors", IEEE Transactions on Parts, Hybrids, and Packaging, 10(2), 101–109., 1974., https://doi.org/10.1109/TPHP.1974.1134841
[10] S. Mohan, "The Design , Modeling and Optimization of on-Chip Inductor and Transformer Circuits", PhD Thesis, (December), 193., 1999.

Multi-Topology DC-DC Converter for Low-Voltage Energy Harvesting Systems

Martin Kováč, Daniel Arbet, Lukáš Nagy, Michal Šovčík and Viera Stopjaková
Department of IC Design and Test,
Institute of Electronics and Photonics,
Slovak University of Technology, Bratislava, Slovakia
e-mail: martin_kovac@stuba.sk

***Abstract*—This paper deals with the design and analysis of a multi-topology DC/DC converter (MT-DC/DC), suitable for ultra-low voltage energy harvesters, such as thermoelectric, photovoltaic or biofuel-cell generators. MT-DC/DC was proposed and investigated as proof of concept with ultra-low voltage cold-start capability (deep sub-hundred mV) and implemented in a standard** $130\ nm$ **CMOS fabrication process. By the use of multiple operating modes, such as the self-oscillation mode, inductive boost-mode based on transformer reuse technique, and tapped-inductor boost-mode, the MT-DC/DC achieves coverage of a wide range of the input voltages. The design also includes a maximum power point tracking control loop, specifically designed for energy harvesting applications, and is capable to deliver more than** $2.5\ mW$ **of power while maintaining a regulated output voltage of** $1.2\ V$**.**

I. Introduction

As indicated by scientific community, the recent trends in development of modern embedded systems such as medical, industrial and personal sensor nodes are strongly focused on the system energy autonomy (also called self-powered or energy–autonomous systems), where the low-power design approaches are necessary to fulfil the challenging energy requirements [1]–[3]. This goes hand-in-hand with the development of variety of energy harvesters (EH), which are identified especially by the nature of the output voltage (DC or AC), intrinsic impedance ($m\Omega$-$M\Omega$), power density (nW/cm^2-W/cm^2, or per cm^3) and condition for the maximum power extraction (Fig. 1). Because of existence of some characteristic similarities, some of them can be group together to design unified power management circuitry (PMC). However, one has to be careful, since also the same EHs can have significantly different properties. For example, a thermoelectric generator (TEG) with the power densit of $100\ \mu W/mm^2$ has the internal impedance R_{IN} of $400\ \Omega$ [4] compared to the value of units of Ω in case of a miniature TEG labeled as 1MD02-024-03TEG. The similar situation can also be found in case of other low-voltage (LV) generators (objects of our interest), as for example, photovoltaic gererators (PVGs) [5]. The fuel-cells (FCGs) with promising high power density usually work at higher temperature, where room temperature operating ones like bio-fuel cells are characterized only by hundreds of ($\mu W/cm^2$) power density that corresponds to the internal resistance from $1\ k\Omega$ to a few $10\ k\Omega$ [6], [7].

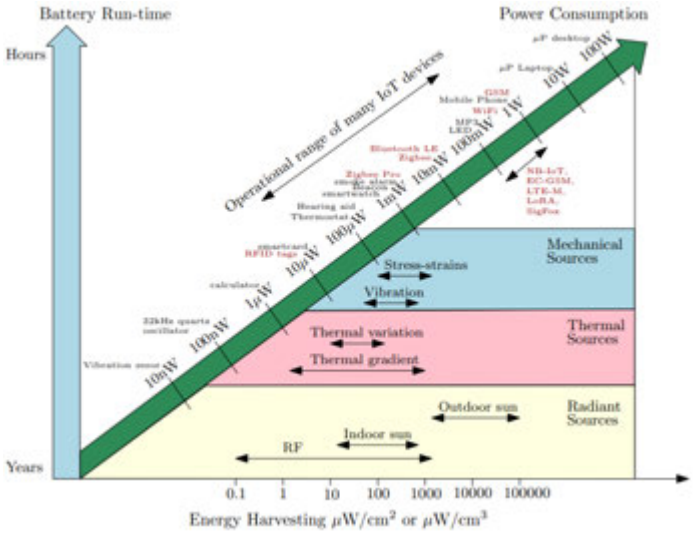


Fig. 1. Power consumption and power densities for various applications and energy sources respectively [1].

In limited space applications, the knowing about value of R_{IN} introduced very important factor into low-voltage PMC design, not only in terms of ultra low-power self-consumption requirements. Additionally, low-voltage design has to be guaranteed to ensure the controllable operational state of the system, starting from the lowest voltage possible (well below $100\ mV$ in case of TEG). Special care must be given to the latter because the open voltage can be insufficient to feed the basic circuit elements under worse conditions. This leads to definition of a special group of LV EH.

Generally, the DC-DC boost type converters (published in the literature) can be categorized from the topology point of view in three basic group. The first one is inductor-based boost converters (IBC) where the simple inductor is used [3], [8]–[10]. The main issue of IBC is self-starting capability from low value of input voltage V_{IN} where auxiliary start-up circuit is usually employed. In order to improve input (start-up) voltage $V_{IN,SU}$, the transformer-based boost converters (TBC) with the coupled inductors should be used [11], [12]. The TBC topology is able to start-up with input voltage in order of tens of mV [12]. The last one is transformer-reuse boost converters, which use TBCs topology for self-start-up and IBCs topology

for normal operation [12]–[14]. From the efficiency point of view, some reported converters can achieve the efficiency up to $80\ \%$ [10], [13], but only in the narrow range of the input voltage. Therefore, in order to achieve the high efficiency in the wide range of input voltage, looking for hybrid topology that combine advantages of various topologies (like IBC and TBC) is required [13].

To deal with this issue, reconfigurable LV EH is proposed, where design requirements and implementation of overall system are described in Section II and II-A. In the Section III verification of proposed concept is presented and discussed. In the last section, the achieved results and findings are concluded.

II. Design Requirements and Implementation

The proposed concept of LV EH called as Multi-Topology DC/DC Converter (MT-DC/DC) is developed as part of ECSEL JU project CONNECT (Innovative smart components, modules and appliances for a truly connected, efficient and secure smart grid), where the requirement has been established according Table I in terms of expected implementation (mainly as start-up circuit for main AC/DC converter [15] in presence of a TEG). The proposed solution has been inspiried by [12], where great partial degradation of the efficiency (valley) in the lower input voltage range was observed. To overcome this phenomenon, the three separate modes were determined for this purpose (three physically reconfigurable topologies): TOM (Transformer-based Oscillation Mode), IBM (Inductor-based Boost Mode), and TIBM (Tapped Inductor-based Boost Mode). This approach is similar to another recently published solution [13] where quasi-reconfigurable or also called "dual stage' topology was also employed.

The top level block diagram of MT-DC/DC is shown in Fig. 2, where the transformer T_1 is a part of the untouched TOM and IBM mode, and added auxiliary branch consisting of the T_2 is a part of new TIBM, based on tapped inductor-based boost converter topology. This topology is characterized by higher voltage conversion ratio in CCM (Continuous Conductivity Mode) than commonly used simple-inductor boost topology and the same ratio is reported in DCM (Discontinuous Conductivity Mode) mode (Table III). Compared to the original design, the added branch brings new degree of freedom in the case of T_2 selection. However proposed MT-DC/DC requires several external off-the-shelf components such as already mentioned transformers (or coupled inductors) with compact physical dimensions ($\leq \quad 12.7 \times 10.2\ mm^2$), an input/output filter capacitor, capacitors for internal charge pumps (positive and negative), auxiliary and storage capacitors and also resistors for sensing the V_{OUT}. Summary of all needed passive components (also used in simulation) are summarized in Table II. Because the transformer represents dominant source of conduction losses and also belongs to the critical parts in the start-up process, it has been carefully modelled as 2-port network including resistive, capacitive, and coupling parasitics taken from the datasheet information. All the off-the-shelf capacitors with the capacitance value above $0.5\ nF$ were managed as ideal components. Capacitors with the capacitance value well below $0.5\ nF$ (that can be potentially used for integration) were taken from the IC foundry library as area-effective MOM (metal-oxid-metal) structure.

The MT-DC/DC contains a mode arbiter (MA) with integrated logic (MAL), which evaluates the status of the MT-DC/DC based on the input data and select the corresponding mode through a local control logic CL_X. The MT-DC/DC has a low-voltage domain (LVD, $500 - 510\ mV$) generated by a low dropout linear regulator directly from the output voltage V_{OUT}. This is part of the PMU and LVD takes care of reducing the power consumption of digital controllers. Furthermore, LVD also provides a supply voltage for a thermally stabilized oscillator and two charge pumps producing the positive and negative voltage rails with sufficient levels to attain full on/off state of individual transistors during each modes. Overvoltage protection that is particularly dangerous in the TOM mode (high oscillation amplitude in combination with voltage multipliers, denoted as NVG and PVG in Fig. 2) is ensured across a built-in active limiter. Active realization of the voltage limiter uses adaptive biasing and in comparison to passive (diode) solution contributes not only to robust reliability but also to the power reduction. In other words, the adaptive biasing was implemented in terms of rapid response capability irrespective of the actual power consumption during action (excess current becomes part of the self-consumption).

The PMU also contains temperature-independent voltage and current references, designed for sub-$500\ mV$ supply voltage and for the power consumption of tens nW under nominal conditions. Generally speaking, the PMU was designed to minimize self-consumption, which is kept unloaded at sub-μA level, making the solution competitive with many state-of-the-art solutions of comparable system requirements [3], [8], [11], [12].

A. Start-up Issue

To ensure a successful reconfiguration together with reliable cold start without significant interference to the degradation of the inverter efficiency, the following points should be kept in mind:

- operating voltage of all components in the initial starting phase is limited to $0 - 50\ mV$ range in the worst case, depending on project requirements (Table I).
- oscillation condition in the TOM mode depends on current leakage outside of the main loop and parasitic capacities referenced to ground that should be minimized.
- all transistors must have strictly defined starting condition and have to ensure reliable on/off state in the respective operational mode (special care must be taken in cold-start phase).
- additional switches increase non only the conduction but also the switching losses, which can be significantly reflected into the degradation of the converter efficiency.

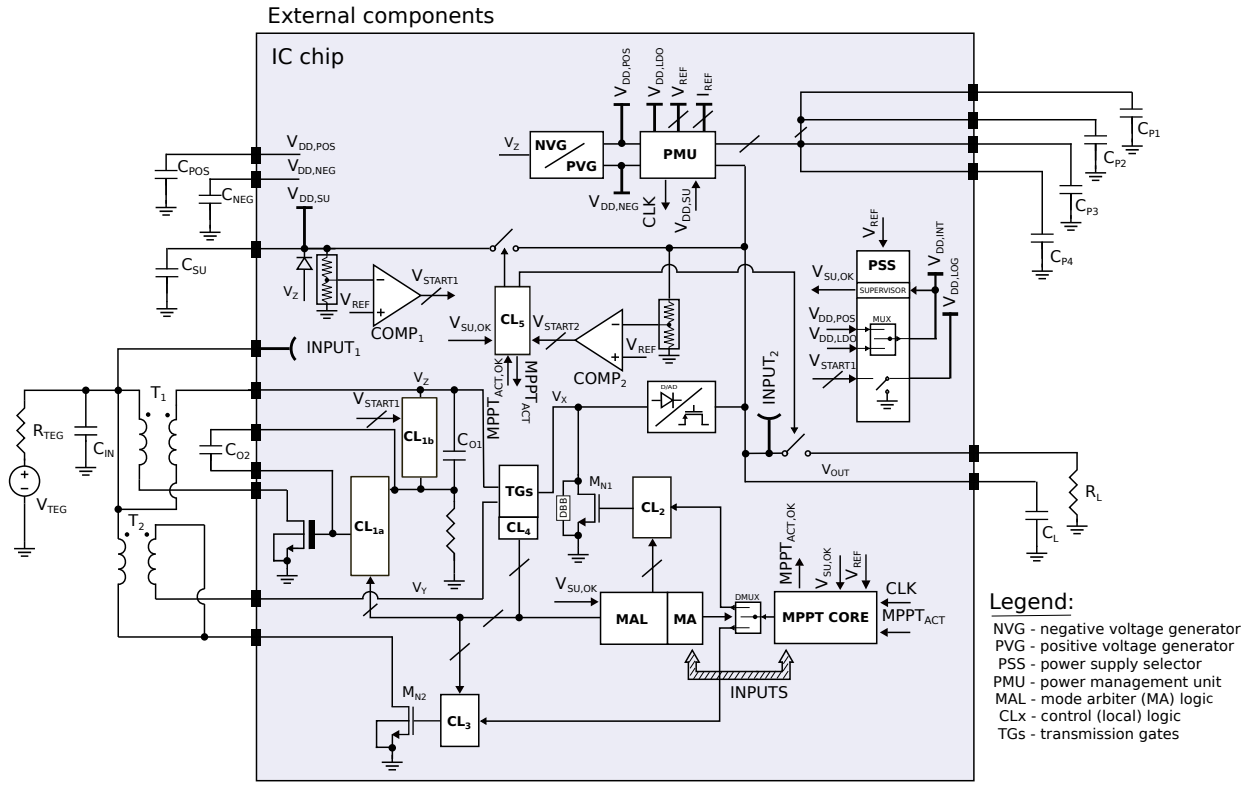


Fig. 2. Block diagram of the proposed Multi-Topology DC/DC Converter (MT-DC/DC)(to maintain clarity, supply voltage domains of individuals block are not depicted).

TABLE I
EXPECTED OPERATION CONDITIONS AND REQUIREMENTS.

Technology	**Source**	$^{*}\mathbf{V_{IN,MIN}/V_{IN,MAX}}$	$\mathbf{V_{OUT,MIN}/V_{OUT,MAX}}$	$\mathbf{P_{OUT,MAX}}$	$\eta_{\mathbf{e-e}}$
"native" and multi-V_{TH} $3.3V$ triple-well MOS transistors	TEG, PV,G FCG	$\leq 50mV/\leq 1V$	$\leq 0.6V/\geq 1.8V$	$\geq 2.5mW$	TOM:cold-start TIBM: $\geq 40\%$ IBM: $\geq 60\%$

* also corresponds to minimum start-up voltage $V_{IN,SU}$

Project requirements set application of self-start-up capability with the minimum input voltage $\leq$ 50 mV. For this purpose, the proposed MT-DC/DC converter utilize the TOM mode (similar like in [12]) and ensures reliable start-up from very low value of the input voltages (in current configuration $\approx$ 40 mV) without the presence of any other external power source (cold-start). The start-up principle using a transformer is very advantageous in terms of minimizing the start-up voltage $V_{IN,SU}$ but at the cost of a higher minimum input power $P_{IN,SU}$ while maintaining the necessary voltage amplification (conversion ratio). In other words, the TOM implementation is suitable for alternative sources with an relative small internal impedance (ordinarily well below several hundreds Ω). On the other hand, this type of start-up suffers for great conduction (high resistive parasitic at secondary side of windings) and switching looses (raised from presence of 90° RC phase shifter at secondary side of the transformer and Miller effect applied on interwinding parasitic capacities) [11], [12]. For these reasons, some concerns are growing especially if the starter circuit is directly a part of the main power transfer branch between the input and output terminals, as in the investigated design.

In order to ensure a controllable start, the TOM mode has been modified to a two-stage form compared to the mentioned competitive solution, however, at the cost of the considerably extended start-up phase of the MT-DC/DC. This effect is shown in more detail in Fig. 3. In the initialization phase called self-oscillations or TOM mode, the main EH branches and the load are cut from the input source by TGs transmission gates (Fig. 2), whereby the harvested energy is continuously stored in the storage tank C_{SU} and powers control circuits. The status is monitored by a comparator $COMP_1$ which repeatedly activates the free-run mode with a fixed defined duty cycle. Since the MT-DC/DC operates in deep DCM mode at this start-up phase and the load remains disconnected, the converter has still high voltage conversion ratio and is able

TABLE II
SUMMARY OF OFF-THE-SHELF COMPONENTS.

TOM/IBM	**TIBM**	**TEG**		**Source/Load**		**Storages**		**Charge pumps**	***Voltage divider**
transformer		$V_{TEG}[mV]$	$R_{TEG}[\Omega]$	$C_{IN}[\mu F]$	$C_L[\mu F]$	$C_{POS/NEG}[nF]$	$C_{SU}[nF]$	$C_{Px}[nF]$	$R_{1/2}$
EPCOS Series B78302A 1:60	CoilCraft T6650-DL 1:5	40-340	5	0.1-1	10	0.5-2	200	0.5	in order of $M\Omega$

* is not depicted in Figure 2

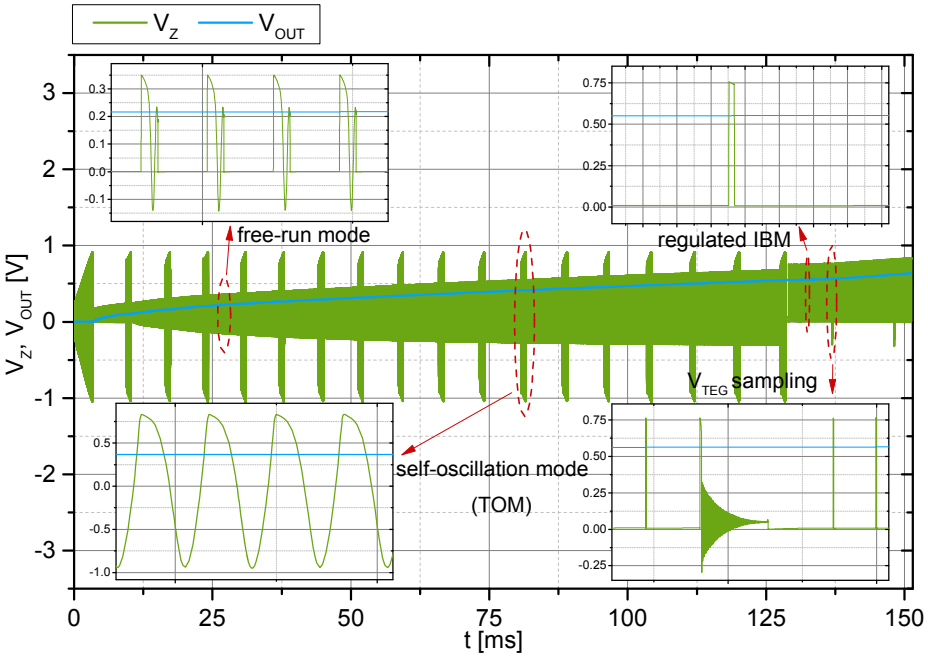


Fig. 3. Demonstration of two-stage start-up phase.

to continuously accumulate energy in the output capacitor C_L. The output node V_{OUT} is monitored by a comparator $COMP_2$, which then in corresponding IBM/TIBM mode activates feedback control based on the Maximum power point tracking (MPPT) algorithm. Just implementation of the MPPT control is the main feature that distinguishes the EH from conventional converters and therefore, their performance is characterized by end-to-end efficiency η_{e-e} rather than the converter efficiency itself. The η_{e-e} reflect effectiveness of energy extraction and transfer from a source to load. In other words, it expresses the converter efficiency under fulfilment of the maximum power extraction condition, which means under the MPPT algorithm. In this context, progress of the proposed solution beyond the state-of-the-art is demonstrated in the following section that should be managed as a proof of concept.

III. CONCEPT VERIFICATION AND DISCUSSION

The proof of concept and possible competitiveness with actual works have been carried out through prelayout simulations, where a standard 130 nm CMOS process with triple-well option and multi-voltage transistors was chosen as suitable candidate. In simulations, transistor-level models are employed in all parts of the system with one exception of the MPPT algorithm implementation that will be explained later in this section. Fig. 4 shows the most important parameter, namely end-to-end converter efficiency η_{e-e} that can directly reflects the performance and compare individual converters (energy harvesters) with similar properties [3], [8], [11], [12] under the MPPT operation, where η_{e-e} gives relevant information about the EH performance. Voltage signals in the legend correspond to voltage values where the maximum of η_{e-e} has been observed, and R_{TEG} and V_{REG} correspond to conditions of the internal impedance of TEG and the output regulated voltage, respectively.

Two red curves reflect the performance of two individual modes, i.e. IBM and TIBM modes. It is obvious that the proposed MT-DC/DC solution brings a promising results especially under ultra low-voltage conditions. Presence of the regulated operation from low values of V_{TEG} (can be roughly taken as $V_{IN} \approx 0.5V_{TEG}$ under MPPT condition at the EH input), directly reflects $2-3$ times improvement over [12],

TABLE III
BASIC PROPERTIES OF IDEAL CONVERTERS.

Converter/Properties	IB	TIB	where
CCM	$M_{IB,CCM} = \frac{1}{1-D}$	$M_{TIB,CCM} = (1+nD)M_{IB,CCM}$	$K_{IB} = \frac{2L_{sec}f_{sw}}{R_L}$
DCM	$M_{IB,DCM} = \frac{1+\sqrt{1+\frac{4D^2}{K_{IB}}}}{2}$	$M_{TIB,DCM} = M^*_{IB,DCM}$	$K_{TIB} = \frac{2L_{prim}f_{sw}}{R_L}$
Boundary Condition	$K_{IB} < \overbrace{D(1-D)^2}^{K_{crit,IB}}$	$K_{TIB} < K_{crit,IB}\left(\frac{n+2}{n+1}\frac{1}{(1+nD)^2}\right)$	
Input DC Impedance	$R_{IN,IB,CCM} = R_L(1-D)^2$ $R_{IN,IB,DCM} \approx \frac{2L_{sec}f_{sw}}{D^2}$	$R_{IN,TIB,CCM} = \frac{R_{IN,IB,CCM}}{(1+nD)^2}$ $R_{IN,TIB,DCM} \approx R^*_{IN,IB,DCM}$	

* the L_{sec} must be replaced by L_{prim} of T_2

D - duty cycle, L_{prim} - inductance of primary winding of T_2, L_{sec} - inductance of secondary winding of T_1, n - turn ratio of T_2 windings, f_{sw} - switching frequency

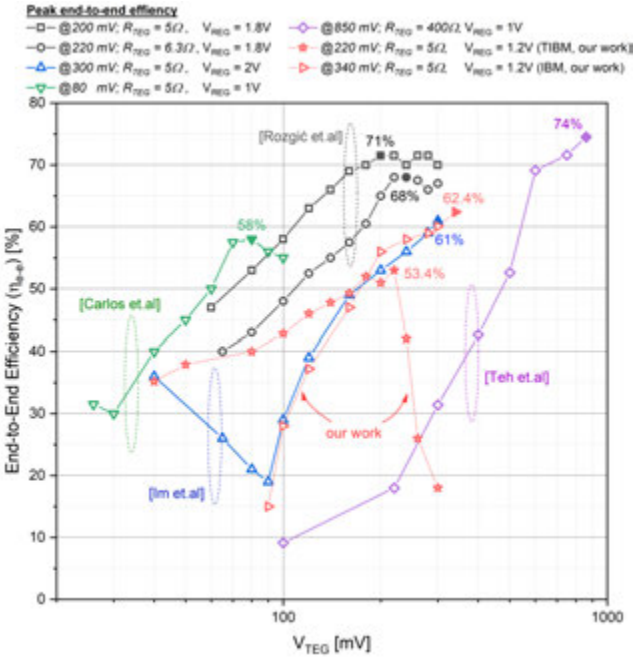


Fig. 4. State-of-the-art comparison through end-to-end converter efficiency.

where benefiting from the TIBM mode implementation and taking advantage of the higher voltage conversion ratio. TIBM mode is also masking valley originated from TOM and IBM transitions from the original work (blue line) that was the primary goal of the proposed solution. The observed reduction of η_{e-e} in TIBM mode is mainly caused by the two following negative effects:

- a negative voltage value of $-nV_{IN}$ is observed at D terminal of transistor M_{N1} , where n is the turn ratio of coil windings, which generates additional power dissipation when charging significant parasitic capacities at this node (node V_X). In addition, this phenomenon is accompanied by an undesirable change in the state of the respective transistor (OFF⇒ON) caused by insufficient negative voltage at its gate terminal G.
- the parasitic series resistance of the transformer secondary winding T_2 is 4-5 times higher than T_1, which is the dominant degradation factor in the conduction losses (the effect of conduction losses increases with V_{TEG} because the available input power also increases). Thus, here arises a question of using coupled inductors with a lower winding ratio, series parasitic resistances, or inductances while maintains its favourable form factor.

If V_{TEG} increases, IBM takes over the work resulting in close performance reproduction of [12] due to using the same transformer part T_1 and converter topology at respective mode. Because SMD/SMT transformer has greater parasitic serial resistance than the inductor from [3], conduction losses deteriorate efficiency, and the increasing gap referenced to [3] (black line) is an inevitable consequence. This phenomenon can be reduced using area-hungry components with reduced parasitics. We have to note that the simulated η_{e-e} was performed under emulation of a MPPT algorithm (behavioral ideal MPPT model in the closed-loop control was used), i.e. physical implementation of MPPT was missing, however its effect was considered (start-up issue demonstrated in Section II-A already includes physical implementation of a MPPT regulation loop). This captures additional conduction losses resulting from raising tendency to switch into the CCM mode due to significant transformer inductances (hundreds of μH, see Table III) and employed of the MPPT algorithm that is similar with one presented in [3], [12]. The used MPPT algorithm is characterized by low complexity however, for the price of existing compromise between switching frequency (directly reflect into converter efficiency across switching looses and indirectly across tendency to switch into the CCM) and the speed of events in the open voltage sampling process all managed by input capacitor C_{IN}. This is in contrary to the recommendation of pushing the EH converter into preferable DCM mode [13]. All these phenomena are involved in effecting the η_{e-e} in both modes. Finally, it important to note that numerical verification at the top-level scheme was carried out under nominal conditions.

The presented results show promising starting point for the

next design steps dedicated to the design of the MA with the relevant functional blocks (close loop decision about suitable mode based on Fig. 4 is not yet a part of the design), to solve the trade-off between area requirements and efficiency η_{e-e}, the verification of design robustness to variation of fabrication process, etc. Despite the fact that except the mentioned block, the presented results taks into account all dominant losses (switching and conduction losses of the converter topology, losses of PMU, MA+MAL and CLx including driver losses), no significant deviation of η_{e-e} from the presented result is expected in the final implementation. Fig. 4 shows not only encouraging results in the case of end-to-end efficiency (especially in relationship to original design from [12]) but also expanding the input/output voltage range for ultra-low voltage (unregulated TOM mode is fully replaced by regulated one - TIBM) and also for high voltage conditions (standard IBM mode). Since $3.3\ V$ transistors are used in design supported by controlled generation of positive/negative voltage rails and aggressive active limiter, the range of regulated output voltage V_{OUT} up to $1.8\ V$ can be ensured. Controlled generation of positive/negative voltage rails in conjunction with LV domain (LVD, 500-510 mV) also supports demand for lower band of V_{OUT} ($0.6\ V$) through ensuring the sufficient overdrive voltage on the main and TGs switches. However, the price of added complexity brings higher number of external components resulting in larger footprint in comparison to other solutions.

IV. Conclusion

The paper deals with a proof of concept and discussion about design of special multi-purpose DC/DC converter also called EH, mainly in terms of its operation modes reflected to the positive effect on end-to-end efficiency η_{e-e}. The proposed MT-DC/DC solution can cover a wide range of the input/output voltage as a result of LVD and active limiter implementation and can deliver at $300\ mV$ V_{TEG} up to $2.7\ mW$ under the regulated output voltage of $1.2\ V$ that correspond to $60\ \%$ end-to-end efficiency. Based on η_{e-e} parameter, it has been proven that the proposed concept can overcome problems originated from the referenced solution [12]. On the other hand, few considerable drawbacks have been also revealed in the paper discussion. Thus, a reduced competitiveness to the most efficient solution presented in Section I is expected. However, in our best knowledge, unconventional high-voltage conversion tapped-inductor boost topology (TIBM) in EH research area has been implemented for the first time, and its performance analysis is the main part of the presented work.

Acknowledgment

This work was supported in part by ECSEL JU under project CONNECT (737434), and the Ministry of Education, Science, Research and Sport of the Slovak Republic under grant VEGA 1/0905/17.

References

[1] M. Shirvanimoghaddam, K. Shirvanimoghaddam, M. M. Abolhasani, M. Farhangi, V. Zahiri Barsari, H. Liu, M. Dohler, and M. Naebe, "Towards a Green and Self-Powered Internet of Things Using Piezoelectric Energy Harvesting," *IEEE Access*, vol. 7, pp. 94 533–94 556, 2019.

[2] Z. Lou, L. Li, L. Wang, and G. Shen, "Recent Progress of Self-Powered Sensing Systems for Wearable Electronics," *Small*, vol. 13, no. 45, p. 1701791, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/smll.201701791

[3] D. Rozgić and D. Marković, "A miniaturized 0.78-mw/cm2 autonomous thermoelectric energy-harvesting platform for biomedical sensors," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 11, no. 4, pp. 773–783, Aug 2017.

[4] H. Bottner, J. Nurnus, A. Schubert, and F. Volkert, "New high density micro structured thermogenerators for stand alone sensor systems," in *2007 26th International Conference on Thermoelectrics*, June 2007, pp. 306–309.

[5] S. K. Sharma and K. Ali, *Solar Cells - From Materials to Device Technology*. Springer International Publishing, 2020.

[6] A. B. F. Barbir and T. N. Veziroglu, *Compendium of Hydrogen Energy*. Woodhead Publishing, 2016, vol. Volume 3: Hydrogen Energy Conversion.

[7] K. J. Kim, B. H. Park, S. J. Kim, Y. Lee, H. Bae, and G. M. Choi, "Micro solid oxide fuel cell fabricated on porous stainless steel: a new strategy for enhanced thermal cycling ability," *Scientific reports*, vol. 6, no. 1, pp. 1–8, 2016.

[8] E. J. Carlson, K. Strunz, and B. P. Otis, "A 20 mV Input Boost Converter With Efficient Digital Control for Thermoelectric Energy Harvesting," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 4, pp. 741–750, April 2010.

[9] A. Das, Y. Gao, and T. T. Kim, "A 220-mv power-on-reset based self-starter with 2-nw quiescent power for thermoelectric energy harvesting systems," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 1, pp. 217–226, Jan 2017.

[10] J. Katić, S. Rodriguez, and A. Rusu, "A dual-output thermoelectric energy harvesting interface with 86.6control power of 160 nw," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 8, pp. 1928–1937, Aug 2016.

[11] Y. Teh and P. K. T. Mok, "Design of Transformer-Based Boost Converter for High Internal Resistance Energy Harvesting Sources With 21 mV Self-Startup Voltage and 74% Power Efficiency," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 11, pp. 2694–2704, Nov 2014.

[12] J. Im, S. Wang, S. Ryu, and G. Cho, "A 40 mV Transformer-Reuse Self-Startup Boost Converter With MPPT Control for Thermoelectric Energy Harvesting," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 12, pp. 3055–3067, Dec 2012.

[13] S. C. Chandrarathna and J. Lee, "A dual-stage boost converter using two-dimensional adaptive input-sampling mppt for thermoelectric energy harvesting," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 12, pp. 4888–4900, Dec 2019.

[14] N. V. Desai, Y. K. Ramadass, and A. P. Chandrakasan, "A bipolar ±40 mV self-starting boost converter with transformer reuse for thermoelectric energy harvesting," in *2014 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, Aug 2014, pp. 221–226.

[15] M. Potočný, J. Brenkuš, and V. Stopjaková, "High side power mosfet switch driver for a low-power ac/dc converter," in *2019 IEEE 22nd International Symposium on Design and Diagnostics of Electronic Circuits Systems (DDECS)*, April 2019, pp. 1–6.

Design and measurements of low power 32-kHz oscillators and a test interface in 180-nm CMOS technology

Ivan Kuljak†, Ivan Tomić†, Roman Bertolan†,
Josip Mikulić⋆, Gregor Schatzberger⋆, Johannes Fellner⋆, and Adrijan Barić†

†University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
†Email: adrijan.baric@fer.hr
⋆ams AG, Tobelbader Straße 30, 8141 Premstätten, Austria
⋆Email: josip.mikulic@ams.com

***Abstract*—This paper presents the design and measurements of two oscillators used as oscillator cores in the 1-Hz oscillator circuit. The output frequency drift is obtained by parametric analyses varying the supply voltage from 1.1 to 1.3 V, referent current from 70 to 140 nA, bias current from 35 to 70 nA and temperature from -40°C to 80°C. The oscillator chip is processed and measured. The frequency characteristics are extracted from measurements using the test interface. The built-in test interface is supplied by an external power supply whose nominal voltage is 1.8 V. The voltage buffer within the test interface allows for indirect measurement of signals on high impedance nodes. The voltage buffer output is capable of driving standard digital multimeter (DMM) input loads.**

***Index Terms*—electronics, microelectronics, CMOS technology, 180-nm, oscillator, test interface, low power, real time clock, temperature compensated**

I. Introduction

In an integrated circuit that relies on a precise time reference it is crucial to be able to characterise and produce a well-defined time reference, with a guaranteed stability of its output frequency over temperature and time [1] - [8].

This work presents the design and measurement of two oscillator topologies, together with a test interface of this integrated circuit (IC), which have been implemented in the 180-nm TSMC CMOS techology. The output of the two cores is later used by other circuits to generate a low power, temperature compensated 1-Hz oscillator, which may be used as an real time clock (RTC) reference. To verify the performance of each oscillator core, a test fixture has been developed, and each core has been characterised for a temperature range from -40°C to 80°C in a temperature chamber.

The test fixture is a test PCB placed in the temperature chamber and all signals for a particular test mode have been fed to a data acquisition system outside the temperature chamber through SMB connectors.

Section II presents the cross-coupled comparator ring oscillator architecture and the measurements of the oscillator block. Section III describes the sawtooth relaxation oscillator architecture and presents the measurements. Section IV describes the test interface module topology. The conclusions are given in Section V.

II. Cross coupled comparator ring oscillator core

A. Circuit design

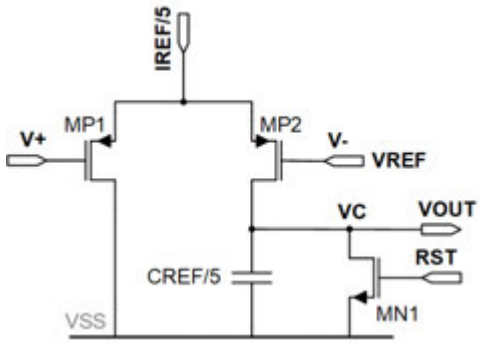


Fig. 1. Basic integrator schematic.

The cross-coupled relaxation oscillator (Fig. 2) receives the referent voltage V_{REF} = 0.35 V and the referent current $I_{REF}/5$ = 20 nA [1], [2], [3].

The oscillator consists of 5 identical integrators (Fig. 1) connected in the ring structure. The fraction of the referent current, $I_{REF}/5$, is integrated on the capacitor having the value of $C_{REF}/5$, resulting in the output voltages V_{C1} to V_{C5}. Inside each integrator, the integrated voltage V_C from the previous stage is compared with the reference voltage V_{REF}. The output voltage of one stage then controls the integration cycle of the following stage and at the same time resets one of the other four stages. In this way, the cross-coupled scheme maintains the oscillations.

Additionally, the integrated voltages V_{C1} to V_{C5} generate the pulse comb at the output of the pulse generator. The pulse generator is a 5-input rising edge detector and OR gate. In combination with a D flip-flop, the output frequency having the nominal value f_{OSCC0} = 32 kHz is generated.

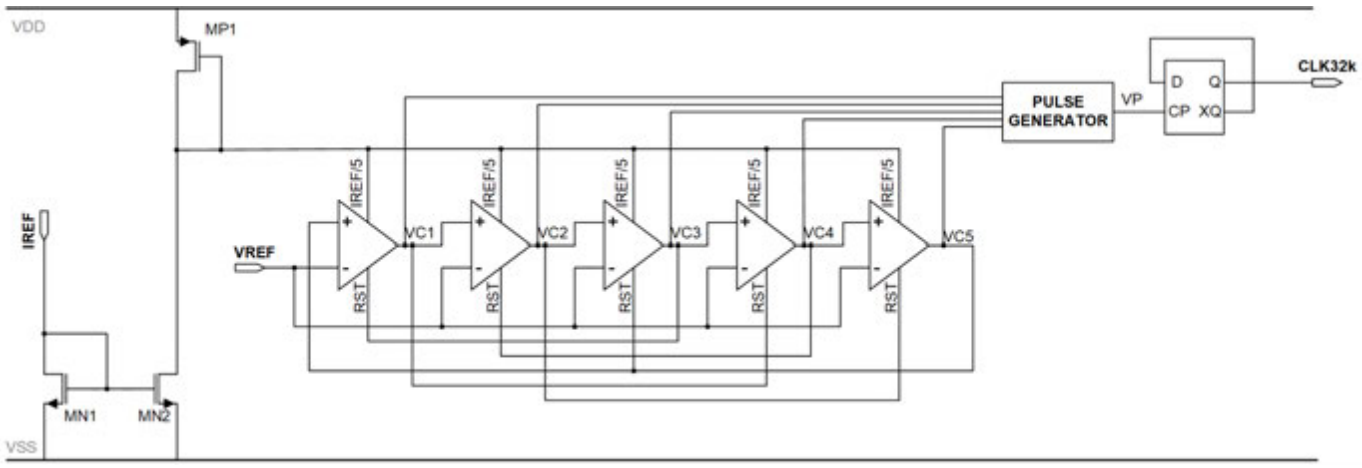


Fig. 2. Cross-coupled relaxation oscillator.

B. Measurement results

The oscillator is one of the blocks of the 1-Hz oscillator chip. Therefore, all the other chip blocks had to be turned off to test only the 32-kHz oscillator. It is achieved through the test interface.

The measurements are performed in a laboratory environment using the following equipment: source measure unit (SMU U2722A), power supply (E3646A), temperature chamber, digital multimeter (34410A), oscilloscope (RTO2044), and frequency counter (33522B).

First, the output frequency dependence on the change of V_{DD} is measured for 10 different chips. The comparison of the simulations and the measurements is shown in Fig. 3 where the solid lines represent the measurements and the dashed line represents the result obtained by simulations. The simulated frequency is 32.0 kHz @ $V_{DD} = 1.2$ V, while the average value of the measured frequencies at $V_{DD} = 1.2$ V is 23.42 kHz which is 26.8% less compared to simulations. The left and right axes have the same span of 0.8 kHz and the results show that the variation of the frequency for $V_{DD} \in$ [1.1 V, 1.3 V] is very similar, except for the shift of the nominal frequency of -26.8%.

The dependence of the output frequency on I_{REF} is measured next. The measurements are performed on 10 different chips. The comparison of the simulations and the measurements is shown in Fig. 4 where the solid lines represent results obtained by measurements and the dashed line represents the results obtained by simulations. The simulated frequency is 32.0 kHz @ $I_{REF} = 100$ nA, while the average value of the measured frequencies at $I_{REF} = 100$ nA is 23.42 kHz which is 26.8% less compared to simulations. The left and right axes have the same span of 20 kHz and the results show that the variation of the frequency for $I_{REF} \in$ [70 nA, 140 nA] is slightly better for the measurements, except for the shift of the nominal frequency of -26.8%. The slope obtained by simulations is 242 Hz/nA, while the measurements result in the average slope of 178 Hz/nA. It is obvious that the variation of the frequency on I_{REF} is much larger than the deviation caused by the change of V_{DD}.

The final measurements are performed in the temperature chamber. Since the whole chip has to be inside the temperature chamber, these measurements are performed with all chip blocks turned on, except the frequency divider.

The change of temperature in the temperature chamber and the output frequency vs. time are shown in Fig. 5.

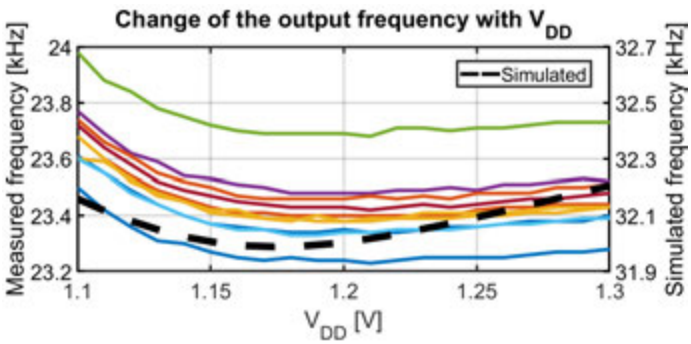


Fig. 3. Frequency change with respect to V_{DD}: the measurements of 10 chip samples (solid lines, left y-axis), simulation results, right y-axis).

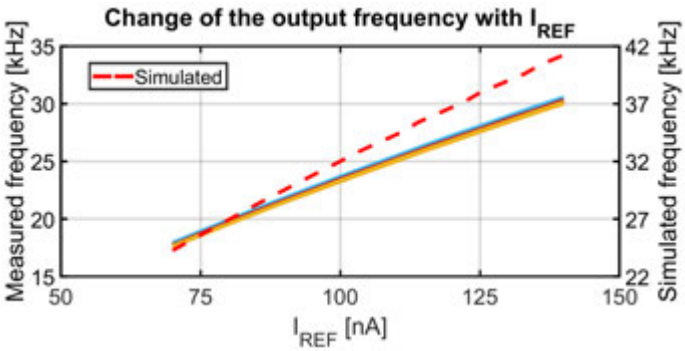


Fig. 4. Frequency change with respect to I_{REF}: the measurements of 10 chip samples (solid lines, left y-axis), simulation results, right y-axis).

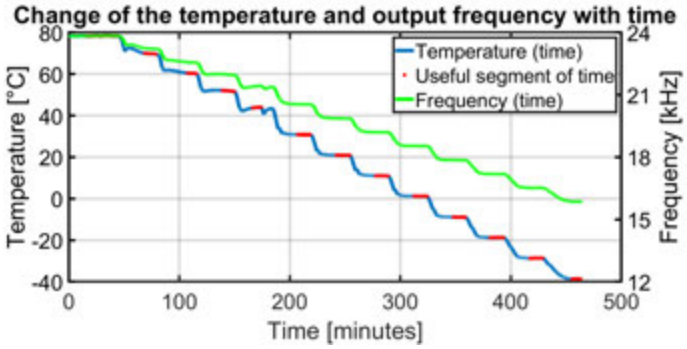


Fig. 5. Change of temperature and output frequency in time.

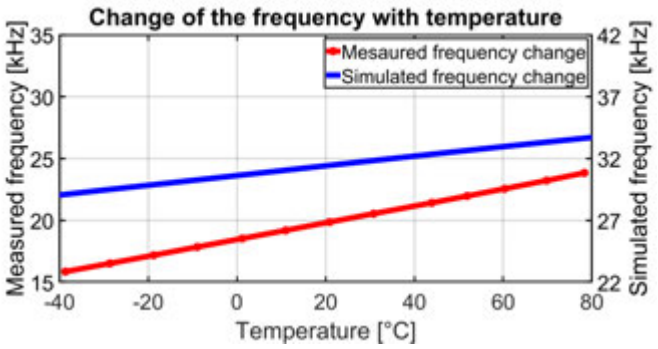


Fig. 6. Frequency vs. temperature.

The section highlighted in red in Fig. 5 represents the part of the time where the temperature is stable and which is used for calculating the average frequency at a certain temperature.

The output frequency change is shown in Fig. 6 where the red trace represents the linear regression of the measured output frequency change with temperature in steps of approximately 10°C from -40°C to 80°C, and the blue line represents the simulated frequency change. The marked red points represent the measurements. The temperature coefficient obtained by simulations is 37.8 Hz/°C, while the measurements result in the temperature coefficient of 68.2 Hz/°C. The current consumption of the cross-coupled oscillator core, obtained by the simulations, is 121.4 nA in the nominal conditions (146 nW) and it varies from 101 nA to 220 nA for a temperature range of [-40 °C, 125 °C].

As can be seen in Fig. 6, there is a discrepancy between the simulations and the results of the measurements carried out on the cross-coupled oscillator core. The simulations were conducted for the nominal transistor parameters and for the corner cases. The extraction of the parasitic elements from the layout was not made, therefore, it could be one of the reasons for the discrepancy. The nominal frequency of the oscillator core is 32 kHz which is transformed to 1 Hz in the frequency divider which is implemented on the same chip. The frequency divider can trim the frequencies in the range from 22 kHz to 48 kHz to enable the chip to still output 1 Hz. Therefore, even since the cross-coupled oscillator core has a discrepancy between the simulations and the measurements, the frequency divider can still trim the output frequency to be 1 Hz.

III. Sawtooth Relaxation Oscillator Core

A. Circuit description

A sawtooth relaxation oscillator (OSCC1) provides the square-wave on its output, having the nominal frequency of 32 kHz. A simplified schematic of the circuit is shown in Fig. 7. It receives the referent voltage V_{REF}, having the nominal value of 350 mV, and the referent current I_{REF}, having the nominal value of 100 nA. It is also supplied by the current I_{BC}, having the nominal value of 50 nA, which ensures proper operation of the comparator [1], [2], [4], [5], [6].

The current I_{REF} charges the capacitor C_{REF} causing a linear growth of the voltage V_C. The comparator coupled with

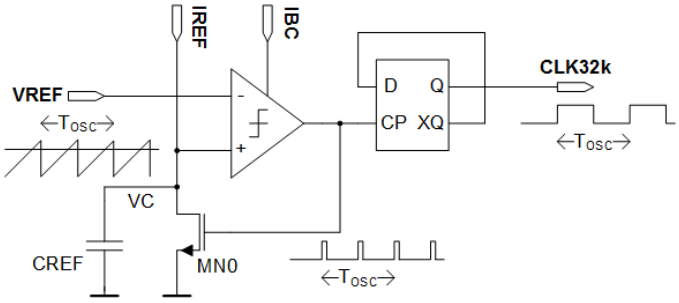


Fig. 7. Simplified schematic of the sawtooth relaxation oscillator.

the transistor M_{N0} performs the function of a resetable integrator. When the capacitor voltage V_C surpasses the referent voltage V_{REF}, the output of the comparator switches to a high level, causing the discharge of C_{REF} through M_{N0}. The capacitor voltage becomes smaller then the V_{REF} voltage, which causes the comparator output to switch back to a low level. The described pulse is also used to change state a of D flip-flop, whose output CLK_{32k} is also the output of the relaxation oscillator.

B. Timing analysis

The signal waveforms of the sawtooth relaxation oscillator are shown in Fig. 8. The period of the signal V_C is approximately equal to one half of the period of the output. The period of V_C is determined from Fig. 9. The first part of this period is the capacitor charging time. This is the time that takes the constant current I_{REF} to charge the capacitor C_{REF} to the voltage V_{REF}. The second part is the delay time of the integrator (t_d). This is the time between the moment when V_C reaches level V_{REF} and the moment when the linear growth of V_C voltage starts over. The output period can be written as:

$$T_{OSC} = \frac{1}{f_{OSC}} = 2\left(\frac{C_{REF}V_{REF}}{I_{REF}} + t_d\right) \qquad (1)$$

C. Simulations and measurements

The sawtooth relaxation oscillator (Fig. 7) is designed and processed in the 180-nm TSMC CMOS technology. The voltage and current references (V_{REF}, I_{REF} and I_{BC}) are sourced from the reference generator and the voltage to current

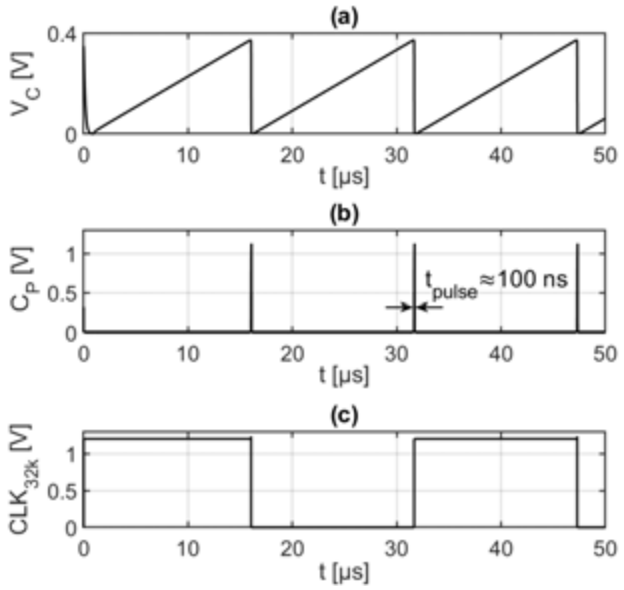


Fig. 8. Signal waveforms of the sawtooth relaxation oscillator: (a) Capacitor voltage V_C, (b) Comparator output C_P, (c) Oscillator output CLK_{32k}.

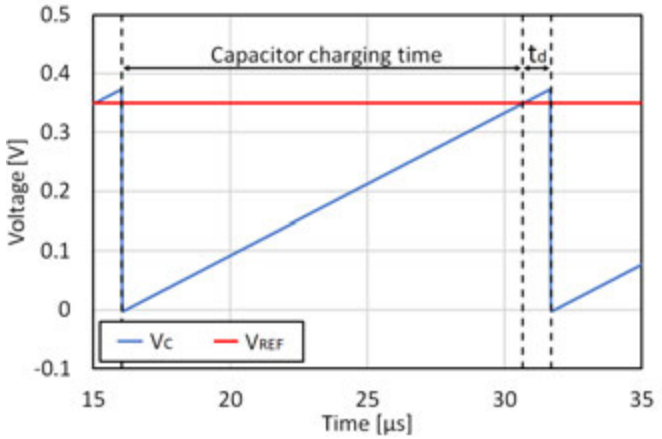


Fig. 9. Period of sawtooth voltage V_C.

converter which are implemented on the same chip, but not described in this paper.

Proper operation of the oscillator is verified in the laboratory by measuring the output frequency when the supply voltage and ambient temperature are changing according to the specifications.

a) Frequency shift with respect to the supply voltage: The frequency is measured with the oscilloscope while the supply voltage varied from 1.1 V to 1.3 V with a 20 mV step. For every supply voltage, the frequency is measured 1000 times and the average value is determined. This procedure is repeated for ten test chips. The results of the measurements compared with the simulations are shown in Fig. 10 and they demonstrate a very good agreement between simulations and measurements. Since the frequency divider can operate properly in the frequency range from 22 to 48 kHz, frequency deviations due to the supply voltage variations are not causing any critical problems.

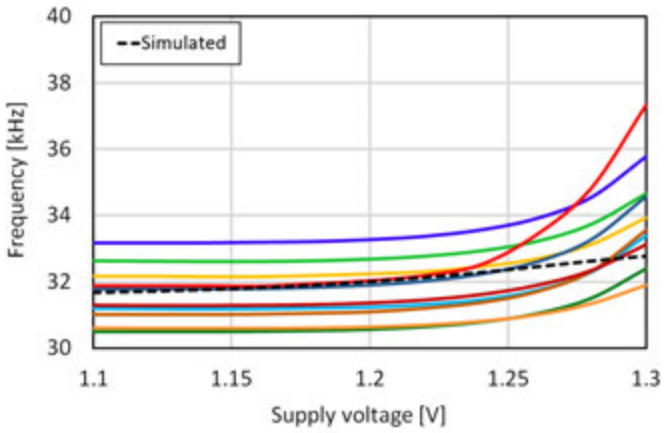


Fig. 10. Measured frequency for 10 chip samples (solid lines) and simulated frequency (dashed line) vs. supply voltage.

b) Frequency shift with respect to temperature: Temperature measurements are made on one test chip only. The frequency is measured using the period counter. The temperature varied from -40°C to 80°C with a 10°C step. Every temperature is maintained for 50 minutes and the transitions between temperatures lasted 10 minutes. The frequency and temperature are measured every 5 seconds, so there are approximately 600 measuring points for each temperature. For each temperature an interval is determined in which both temperature and frequency are stable. The results of measurements are compared with the simulations for the nominal conditions and also with the worst-case corner simulations (Fig 11). The current consumption of the sawtooth oscillator core is 198 nA in the nominal conditions (238 nW) and it varies from 120 nA to 350 nA for a temperature range of [-40 °C, 125 °C].

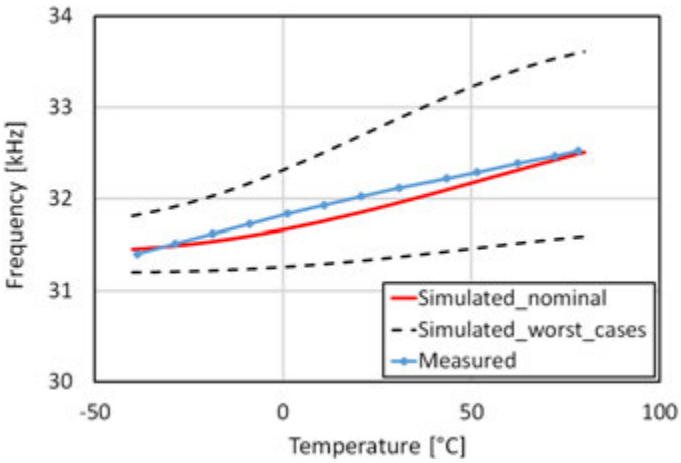


Fig. 11. Measured and simulated frequency vs. temperature (V_{DD} = 1.2 V).

IV. Test interface

The test interface consists of several switches realised as transmission gates, a voltage buffer and an input test vector decoder [7], [8]. A transmission gate is made up of a pair of pMOS and nMOS transistors which are controlled by a common enable signal, which is internally inverted (Fig. 13). These transmission gates multiplex internal signals to the outside world or bring an external signal to a desired circuit module. The circuit diagram is shown in Fig. 12.

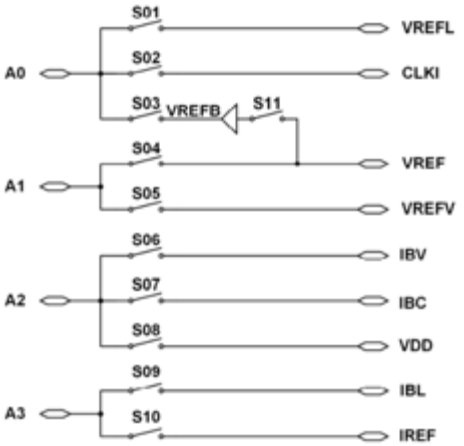


Fig. 12. Test interface analog multiplexer.

The input test vector decoder is implemented by using predefined digital cells. It is used to convert a 4-bit input test word into the appropriate transmission gate enable signals for the specific test case. Only vectors `0x9` to `0xF` are used for testing the internal modules, while vectors `0x0` to `0x8` are used to check the performance of each oscillator core and the frequency division circuit.

A. Transmission gate

The transmission gate (Fig. 13) consists of an nMOS and a pMOS pair (Q_3 and Q_4 respectively) which are connected in parallel. The parallel connection assures that the current flow is bidirectional and that the transmission gate operates well when the input voltage is close to ground or close to the power supply voltage.

The transmission gate cell has an integrated inverter, which is used to minimise the number of traces connecting the decoder and each transmission gate. It is implemented as a standard single stage inverter (consisting of only 2 transistors, nMOS and pMOS), where the pMOS has twice the W/L ratio of the nMOS device, which is necessary because of the lower majority charge carrier mobility in the p-type substrate. Usually, the ratio should be around 4 times bigger, but this is halved to reduce the transmission gate footprint in silicon and to minimise the parasitic capacitance seen by the decoder.

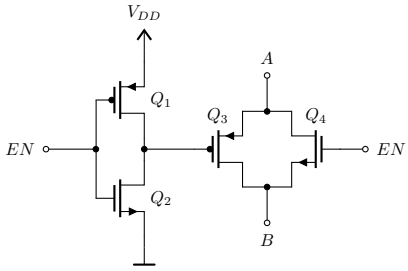


Fig. 13. Transmission gate schematic.

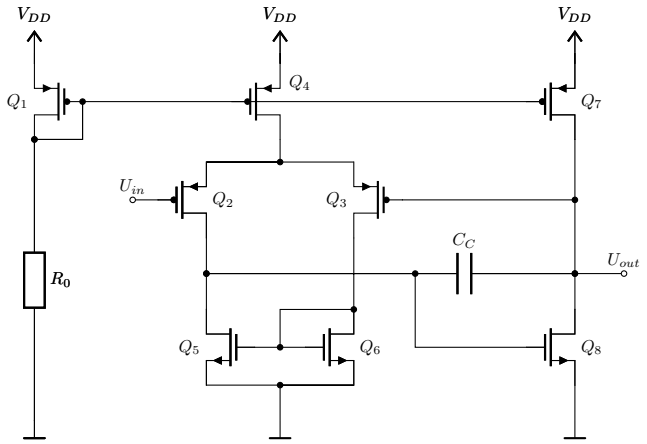


Fig. 14. Voltage buffer schematic.

B. Voltage buffer

The voltage buffer (Fig. 14) is a two stage, frequency compensated amplifier with unity gain. The transistor Q_1 and the resistor R_0 set the bias current for both amplifier stages. Both the transistor and the resistor are temperature dependent, which means that the whole amplifier is temperature dependent, mostly due to this bias network Q_1/R_0. The second source of variation is due to the power supply voltage variation, and this too determines the circuit behaviour (generally, higher V_{DD} implies higher current consumption, which in turn gives a better performance). The input stage consists of a differential pair (Q_2 and Q_3) whose sources are connected to the constant current source Q_4. The drains of the differential pair are connected to an active load (Q_5 and Q_6). The drain of Q_2 is the output of the first stage. The output stage consists of the output transistor Q_8 and its active load Q_7, as well as the compensating capacitor C_C which sets the dominant pole of the amplifier and this assures that the amplifier is stable even for unity-gain configuration. The output transistor Q_8 works in a common source configuration. A relatively large output stage allows for measuring some of the high impedance nodes of the whole circuit, since a load of $1\,\mathrm{M\Omega}\,||\,10\,\mathrm{pF}$ is considered to be quite demanding. Such a load is a typical equivalent circuit of a measurement setup (for example of a bench DMM). Therefore this voltage buffer makes sure that the high impedance nodes are not affected as much and that the measurement can be performed.

V. Conclusion

This paper presents the design and measurements of two different oscillator cores and the test circuit module. The two oscillator cores are essential for the operation of the 1-Hz oscillator circuit. Their accuracy is directly affected by the accuracy of the reference generators which generate the referent bias currents and voltages for the operation of the oscillator cores presented here. The two oscillator cores show good results when compared to what has been simulated. The sawtooth oscillator shows a frequency deviation of around ± 0.5 kHz for the entire temperature range. A variation from f_0 of +5/-2 kHz is observed in the output frequency as the supply voltage is varied. The cross-coupled oscillator shows a frequency deviation of around 800 Hz for the specified V_{DD} range both in simulations and in measurements. The slope of this deviation is 242 Hz/nA for simulations and 178 Hz/nA measured, for a linear change of the output frequency with respect to I_{REF}. The simulated temperature coefficient is 37.75 Hz/°C and the measured temperature coefficient is 68.21 Hz/°C. The test interface module allows for testing individual sub-circuits in the chip.

Acknowledgment

This work is partially supported by ams AG, Premstätten, Austria.

References

[1] Y. Tokunaga, S. Sakiyama, A. Matsumoto, and S. Dosho, "An on-chip CMOS relaxation oscillator with voltage averaging feedback," *IEEE J. Solid-State Circuits*, vol. 45, no. 6, pp. 1150–1158, 2010.

[2] Y. Cao, P. Leroux, W. D. Cock, and M. Steyaert, "A 63000 Q-factor relaxation oscillator with switched-capacitor integrated error feedback," *Solid-State Circuits Conference Digest of Technical Papers (ISSCC) IEEE International*, pp. 186–187, 2013.

[3] T. Tokairin, K. Nose, K. Takeda, K. Noguchi, T. Maeda, K. Kawai, and M. Mizuno, "A 280nW, 100kHz, 1-cycle start-up time, on-chip CMOS relaxation oscillator employing a feedforward period control scheme," pp. 16–17, 2012.

[4] S. L. J. Gierkink and A. J. M. V. Tuijl, "A coupled sawtooth oscillator combining low jitter and high control linearity," *IEEE J. Solid-State Circuits*, vol. 37, no. 6, pp. 702–710, 2002.

[5] U. Denier, "Analysis and Design of an Ultralow-Power CMOS Relaxation Oscillator," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 8, pp. 1973–1982, 2010.

[6] A. Paidimarri, D. Griffith, A. Wang, G. Burra, and A. Chandrakasan, "An RC Oscillator With Comparator Offset Cancellation," *IEEE Journal of Solid-State Circuits*, vol. 51, pp. 1866–1877, 2016.

[7] J. Mikulic, G. Schatzberger, and A. Baric, "A 1-MHz On-Chip Relaxation Oscillator with Comparator Delay Cancelation," *IEEE European Solid-State Circuits Conference (ESSCIRC)*, pp. 95–98, 2017.

[8] ——, "A 1-MHz Relaxation Oscillator Core Employing a Self-Compensating Chopped Comparator Pair," *Circuits and Systems (ISCAS) IEEE International Symposium on*, pp. 1–4, 2018.

Programmable low-frequency divider in 180-nm CMOS technology

Dubravko Tomic[+], Josip Mikulic[*], Gregor Schatzberger[*], Johannes Fellner[*], and Adrijan Baric[+]

[+]Unitversity of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia ,
[+]Email: adrijan.baric@fer.hr
[*]amsAG, Tobelbader Straße 30, 8141 Premstätten, Austria
[*]Email: josip.mikulic@ams.com

***Abstract*—This paper presents the design and measurements of a programmable frequency divider that converts the input clock of 32 kHz to 1 Hz. The circuit is designed, simulated and the layout is drawn in Cadence simulation environment. The binary counter used in the circuit design is synchronous counter with asynchronous reset. The output frequency drift is measured and compared with simulation results for the varying supply voltage from 1.1 to 1.3 V and temperature from -40°C to 80°C.**

***Index Terms*—frequency divider, 180-nm CMOS technology, microelectronics, chip-design**

I. Introduction

The frequency dividers are essential blocks in the design of the real time clock (RTC) circuits and phase lock loop (PLL) circuits. They are often used in PLL circuits to scale down the voltage controlled oscillator (VCO) frequency so it becomes equal to the reference frequency. In the RTC circuits they are used to scale down the VCO frequency to the wanted output frequency, and by doing so the jitter of the VCO is reduced [1] [2] [3].

The frequency divider is a circuit that transforms one input frequency into a lower frequency. It counts up to a predefined value n, where $n \in \mathbb{N}$. At that moment, the output of the circuit changes the state from 0 to 1 or from 1 to 0. The frequency divider transforms a square wave signal with frequency f to a square wave signal with the frequency $f/2n$.

Section II describes the design of the programmable frequency divider and the selection of the comparison constant. Section III presents the measurements and compares the measured data with the simulations. The conclusions are given in Section IV.

II. Programmable frequency divider

Fig. 1 presents a general schematic of a programmable frequency divider. The main components of the frequency divider are: binary counter, binary comparator, generator of a comparison constant (binary adder) and some digital circuits used to form the signal.

The binary comparator is a digital circuit that compares two binary numbers and in the case of equality, it generates a short impulse which is used to reset the binary counter to the initial state. The first binary number is generated by the counter. The second binary number is defined by the data C1<14:0> from the binary adder. The short impulse generated at the output of the comparator is also used to change the state of the D flip flop, and by doing so the output signal having the desired frequency is formed. To avoid the clock ripple problem the synchronous counter with asynchronous reset is used [2] [3].

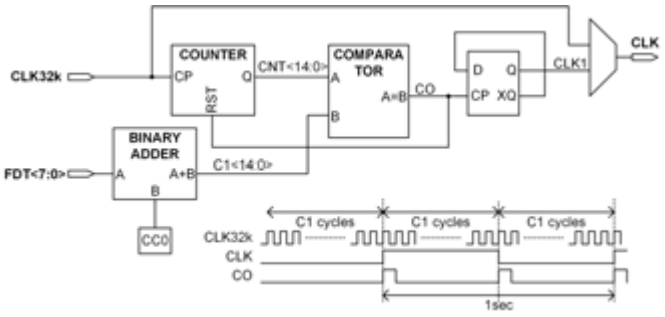


Fig. 1. Block design of frequency divider.

The binary adder is used as the generator of the comparison constant C1<14:0>. It has two inputs: the input B is used to apply the binary data CC0 which is a fixed (hard-wired) constant, and the input A is set by the external input FDT<7:0> (Fig. 1) and it is an 8-bit vector used to trim the comparison constant, thus allowing for the programmability of the output frequency.

A. Selection of the comparison constant

The hard-wired constant CC0 is the input B of the binary adder. The output of the binary adder C1 determines the output of the frequency divider ($f_{out} = f_{in}/(2 \cdot C1)$). In order to generate the frequency of 1 Hz for the nominal input frequency of 32 kHz, C1 has to be equal to $C1_{nominal} = 16000_{(10)} = 0011\ 1110\ 1000\ 0000_{(2)}$. The hard-wired constant CC0 must be selected to be as close as possible to $16000_{(10)}$.

The specified supported range of the input frequencies is from 22 kHz to 48 kHz and to achieve the output of 1 Hz the value C1 has to be in the range from $C1_{min} = 11000_{(10)} = 0010\ 1010\ 1111\ 1000_{(2)}$ to $C1_{max} = 24000_{(10)} = 0101\ 1101\ 1100\ 0000_{(2)}$. Thus the output of the binary adder must have 15 bits. The most significant bit of C1, bit C1<14>, can be 1 as in the case for $C1_{max} = 24000_{(10)}$ or zero.

The output of the binary adder has 15 bits and that means that the inputs also have to have 15 bits, which implies that an overflow will be ignored if it occurs. For example, if the inputs and the output are 3-bit long, for $B = CC0 = 4_{(10)} = 100_{(2)}$ representing a hard-wired constant and A representing a positive number that is used to increase the value of n, e.g. $A = 2_{(19)} = 010_{(2)}$, the result of addition will be regular and equal to $C1 = 6_{(10)} = 110_{(2)}$. In the case the value of C1 has to be decreased below $CC0 = 4_{(10)}$, the 2's complement representation will be used and the subtraction by $2_{(10)} = 010_{(2)}$ becomes the addition of $101_{(2)} + 1_{(2)} = 110_{(2)}$, which results in $C1 = 100_{(2)} + 110_{(2)} = 010_{(2)} = 2_{(10)}$, obtained by neglecting the overflow.

A reasonable choice for the hard-wired constant CC0 is between $11000_{(10)}$ and $24000_{(10)}$. For this design, the value of CC0 is chosen to be $16768_{(10)} = 100\ 0001\ 1000\ 0000_{(2)}$. The maximum positive number that can be represented by $C1 = 111\ 1111\ 1111\ 1111_{(2)}$ is $32767_{(10)}$, which means that the input defined by FDT<7:0> should not be larger than $32768–16768 = 15999_{(10)}$. In this design the maximum value of C1 is $24000_{(10)}$, which is larger than $16768_{(10)}$ only by $7232_{(10)}$ and there is no risk that the maximum value defined by FDT will be exceeded. The exact definition of FDT<7:0 > and the calculation of its contribution to CC0 will be given shortly.

With respect to the minimum value of $C1 = 11000_{(10)}$, it is now clear that it is necessary to subtract $5768_{(10)}$ from the hard-wired CC0 to reach $11000_{(10)}$. The value that has to be subtracted has to be represented in the 2's complement format and in this case it is $5768_{(10)} = 001\ 0110\ 1000\ 1000_{(2)}$, which translates into the 2's complement representation as $-5768_{(10)} = 110\ 1001\ 0111\ 0111_{(2)} + 1 = 110\ 1001\ 0111\ 1000_{(2)}$.

The frequency divider vector FDT<7:0> has only 8 bits and that would not be sufficient to trim the constant CC0 enough if used to change the lower digits of the constant CC0. Thus, the vector FDT<7:0> is shifted 7 bits to the left (Fig. 2). For every value of the vector FDT<6:0> having the bit FDT<7> equal to 1 the reduction of CC0 will be obtained, and for the bit FDT<7> equal to 0 the addition to CC0 will be performed. The $C1_{min}$ and $C1_{max}$ achievable by using this configuration are $C1_{min} = 384_{(10)}$ for FDT = $1000\ 0000_{(2)}$, which is equivalent to the value of the input $A = 100\ 0000\ 0000\ 0000_{(2)}$ or the real number of -16384, and $C1_{max} = 32640_{(10)}$ with FDT = $0111\ 1100_{(2)}$ and $A = 011\ 1110\ 0000\ 0000_{(2)}$. Fig. 2 depicts the corresponding bits of the binary adder's inputs.

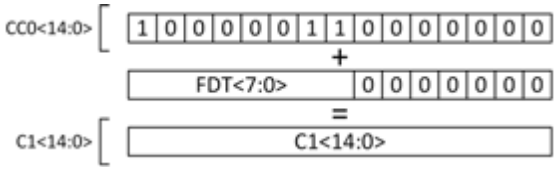


Fig. 2. Inputs (CC0 and shifted FDT vector) and output (C1) of the binary adder.

III. Simulations and measurements

The frequency divider presented in Fig. 1 is designed and manufactured in the 180-nm TSMC CMOS technology for the nominal input frequency of 32 kHz and nominal supply voltage of V_{DD} = 1.2V. Because of the very long simulation time the simulations have to be done with minimal precision and for parametric simulations few steps are made. The simulated characteristic in Fig. 4 is obtained by interpolating results of the simulations for voltages 1.1 V, 1.2 V and 1.3 V and the characteristics in Fig. 9 by interpolating results of the simulations for temperatures -40°C, 0°C, 40°C and 80°C.

The measurement system presented in Fig. 3 consists of the following instruments: Keysight Source/Measure Unit U2722A (SMU) [4], Keysight E3646A (Voltage source) [5], Keysight 33522B (Waveform generator) [6] and Keysight 53220A (Frequency counter) [7]. The SMU is used to vary the V_{DD} voltage from 1.1 to 1.3 V. The waveform generator is set up to generate square wave signal having the amplitude of 1.2 V (V_0 = 0 V, V_1 = 1.2 V) and the frequency of 32 kHz.

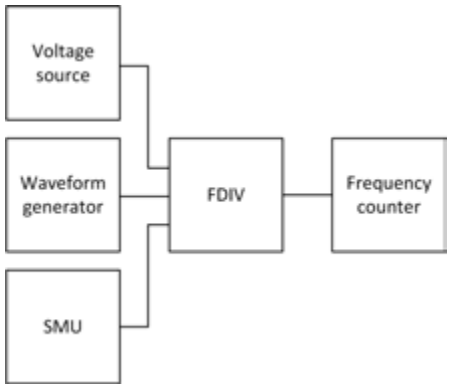


Fig. 3. Measuring setup for frequency drift measurement with respect to supply voltage.

A. Frequency drift with respect to supply voltage

The frequency is measured by the frequency counter while the supply voltage is changed from 1.1 V to 1.3 V in a 10 mV step. Fig. 4 presents the output frequency measured 150 times at each step of the supply voltage compared to the simulated results. The red markers in Fig. 4 present the results which are interpolated. Fig. 5 shows the drift of the output frequency with respect to the supply voltage of multiple chips. The calibration vector FDT<7:0> was the same for the measurements of all chips with the nominal input frequency since all chips generate the signals having very similar frequencies, as it is shown in Fig. 5. The calibration vector FDT<7:0> is equal to $11111010_{(2)}$. These frequencies are measured with the step of 20 mV, and for each supply voltage 50 measurements are conducted. To plot these results the data is interpolated using the spline interpolation method. The maximum measured deviation of the output frequency from 1 Hz is -0.0503 ppm. The maximum measured standard

deviation with the change of the supply voltage V_{DD} is 2.13×10^{-9} Hz and the minimal measured standard deviation is 2.47×10^{-10} Hz.

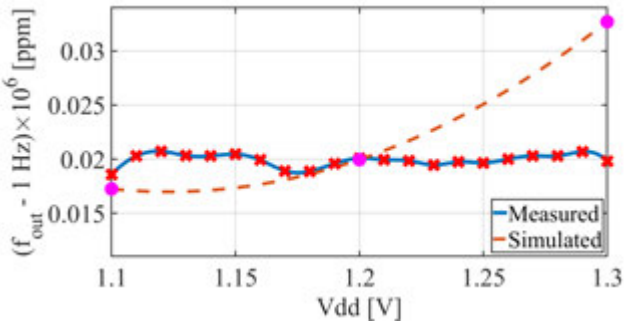


Fig. 4. Measured frequency deviation from 1 Hz as a function of the supply voltage compared to simulated results.

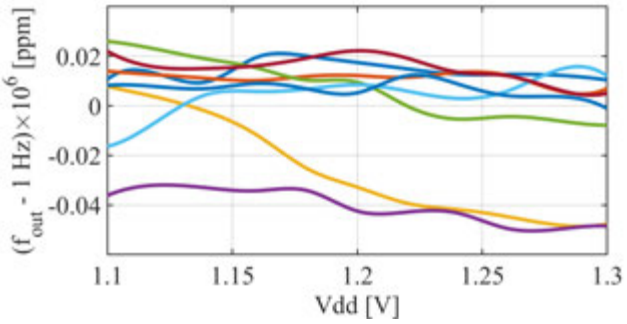


Fig. 5. Measured frequency deviation from 1 Hz as a function of the supply voltage measured on ten different chip samples.

B. Frequency drift with respect to input frequency

A predefined supported range of the input frequencies is from 22 kHz to 48 kHz, with the nominal input frequency of 32 kHz. There is a requirement that the precision of 0.5% must be achievable for every frequency in the specified range. To verify this functionality, the output frequency is measured while changing the input frequency with a step of 2 kHz. For each step of the input frequency the circuit is trimmed by the FDT vector to achieve the output frequency closest to 1 Hz and that output frequency is measured 50 times by the measurement setup in Fig. 3. The results of the measurements are shown in Fig. 6. The results show the supported range of the input frequencies for which the output will deviate less than 0.5% is from 18 kHz to 64 kHz.

C. Frequency drift with respect to the temperature

The measurement system presented in Fig. 7 is used to measure the frequency drift with respect to the temperature. The waveform generator is used for the input frequency of 32 kHz with the amplitude of 1.2 V ($V_0 = 0$ V, $V_1 = 1.2$ V). The frequency divider is trimmed by the vector FDT so that the output frequency is as close as possible to 1 Hz. The temperature is being varied from -40°C to 80°C. The temperature

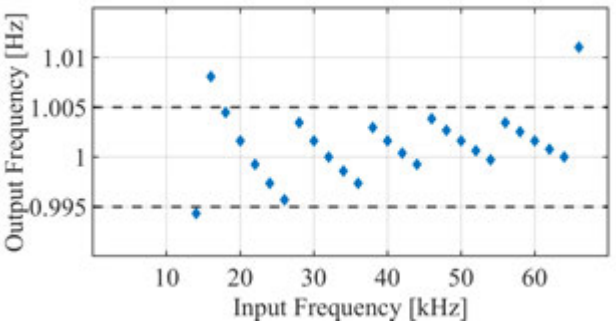


Fig. 6. The best achievable output frequency for different input frequencies.

in the chamber is measured with a Keysight 34410A digital multimeter [8] and a PT100 resistor [9]. The temperature chamber is first heated to 80°C and held at that temperature for 60 minutes, and after that the temperature is reduced by 10°C linearly in time over 10 minutes and held at the new temperature for 50 minutes (Fig. 8). The output frequency and temperature are measured every 5 seconds. Red markers in Fig. 9 show the average values of the output frequency measured on the stable temperature interval from Fig. 8, and these values are interpolated by the spline method. The measured standard deviation with the change of temperature is 2.5498×10^{-9} Hz for the measured chip sample. Only one sample is measured.

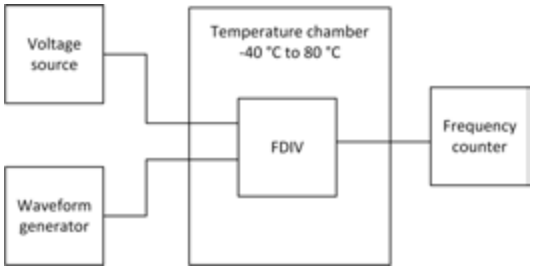


Fig. 7. Measuring setup for frequency drift measurement with respect to temperature change.

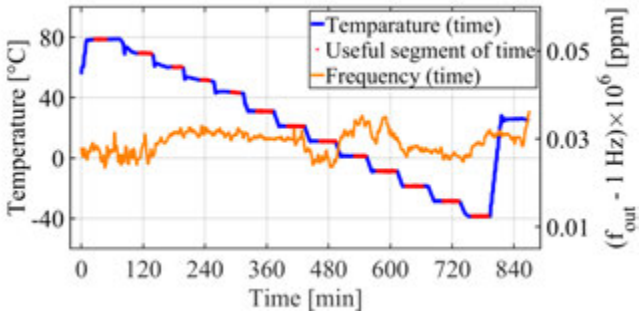


Fig. 8. Change of the temperature and measured output frequency with time.

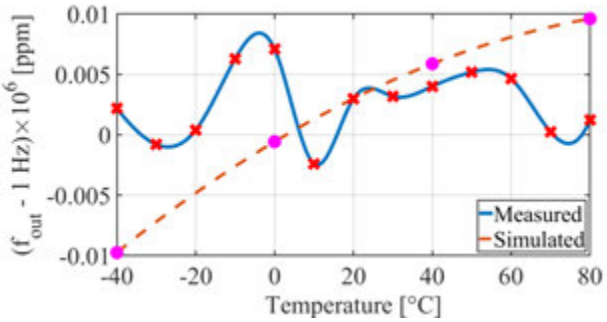


Fig. 9. Measured and simulated frequency deviation from 1 Hz as a function of the temperature.

IV. Conclusion

This paper presents the design and measurements of a programmable frequency divider that converts the input clock frequency from 32 kHz to 1 Hz. The largest measured deviation of the output frequency from 1 Hz is -0.05 ppm for the change of the supply voltage V_{DD} from 1.1 V to 1.3 V. The maximum measured standard deviation, on a single chip, with the change of the supply voltage V_{DD} is 2.13×10^{-9} Hz. The measured standard deviation with the change of temperature from -40°C to 80°C is 2.5498×10^{-9} Hz. The specified range of the supported input frequency is -10/+16 kHz around 32 kHz, while this design achieves the range of -16/+32 kHz with the ability to adjust the frequency to 1 Hz $\pm$ 0.05%.

Acknowledgment

This work is partially supported by ams AG, Premstätten, Austria.

References

[1] Behzad Razavi, *Design of CMOS Phase-Locked Loops: From Circuit Level to Architecture Level*. Cambridge University Press, 2020.
[2] Marko Cupic, *Digitalna elektronika i digitalna logika*. Kigen, 1 ed., 2006.
[3] Anim Arifah Ahmad, Sawal Hamid Md Ali, Noorfazila Kamal, Siti Raudzah Abdul Rahman, and Masuri Othman, "Design of Phase Frequency Detector (PFD),Charge Pump (CP) and Programmable Frequency Divider for PLL in 0.18um CMOS Technology," 2018.
[4] Keysight Technologies, *Keysight U2722A/U2723A USB Modular Source Measure Units*, 8 ed., 2017.
[5] Keysight Technologies, *Keysight E364xA Dual OutputDC Power Supplies*, 11 ed., 2014.
[6] Keysight Technologies, *Keysight Trueform Series Waveform Generator*, 3 ed., 2020.
[7] Keysight Technologies, *Keysight 53220A/53230A 350 MHz Universal Frequency Counter/Timer*, 3 ed., 2017.
[8] Keysight Technologies, *Keysight 34410A/11A 6 ½ Digit Multimeter*, 7 ed., 2014.
[9] Labfacility, *General Purpose Pt100 Probe with Teflon insulated cable*.

Design of a tunable temperature coefficient voltage reference with low-dropout voltage regulator in 180-nm CMOS technology

David Osmanović*, Ivan Skeledžija*, Krešimir Špoljarić*, Domagoj Tomić*,
Josip Mikulić†, Gregor Schatzberger†, Johannes Fellner† and Adrijan Barić*
*University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
Email: adrijan.baric@fer.hr
†ams AG, Tobelbader Strasse 30, Premstaetten 8141, Austria
Email: josip.mikulic@ams.com

***Abstract*—This paper presents a voltage reference having a tunable temperature coefficient and a low-dropout voltage supply. The circuit has been designed using the Cadence Virtuoso software package and afterwards processed and measured. It is possible to arbitrarily set the first-order temperature coefficient to achieve compensation of the oscillator resultant temperature coefficient. The low-dropout voltage regulator provides a nominal output voltage of 1.2 V with a maximum output current of 2 μA. It requires a 600-mV voltage reference and 50-nA current reference for proper operation. As a result of the analysis, it is concluded that the voltage reference has a default temperature coefficient of -37.7 ppm/°C. The voltage reference is buffered and additionally compensated using the tunable temperature coefficient resistor with the step of 20 ppm/°C controlled by a 4-bit input vector and temperature coefficient range of ±150 ppm/°C.**

***Index Terms*—electronics, CMOS technology, chip design, integrated circuit design, clock generator, reference resistor, voltage reference, current reference, bandgap, temperature compensation, LDO**

I. General introduction of the topic

The presented system consists of four main blocks: low-dropout voltage regulator (LDO), reference generator (REFG), voltage to current converter (V2I) and reference resistor (RREF). It is part of a real-time clock (RTC) chip manufactured in the 180-nm TSMC CMOS technology [1]–[8]. These blocks are presented in Sections II to V respectively. The purpose of this work is to verify the operation of each block, examine its performance in a wide temperature range and compare results with simulations conducted using the Cadence Virtuoso software package.

II. Circuit architecture

The low-dropout voltage regulator provides a regulated and temperature stable supply voltage for the circuit using references provided by the reference generator. The voltage to current converter buffers reference generator outputs. The reference resistor is used at the V2I output to enable fine-tuning of the temperature coefficient.

III. Test environment

The test environment consists of several measurement instruments: E3646A power supply, U2722A source measure unit, Pt100 temperature probe, 34465A and 34410A digital multimeters and a temperature chamber capable of operation in the temperature range from -40°C to 150°C. The chip itself is placed on a test PCB which allows for measurements of the chip outputs and setting configuration vectors. The outputs are connected to measurement instruments using coaxial cables with SMB connectors.

IV. Low-dropout voltage regulator

A. Voltage regulator architecture

The designed low-dropout voltage regulator provides a nominal output voltage V_{DD} = 1.2 V with a maximum output current of 2 μA. The nominal input voltage is V_{DD0} = 1.8 V. The regulator is a two stage Miller compensated amplifier with a differential NMOS pair in the input stage [2], [7].

The reference voltage V_{REFL} = 600 mV and reference current I_{BL} = 50 nA are provided by the reference generator circuit.

B. Line and load regulation

Line and load regulation are measured for ten low-dropout regulators. Other parts of the chip are disabled during measure-

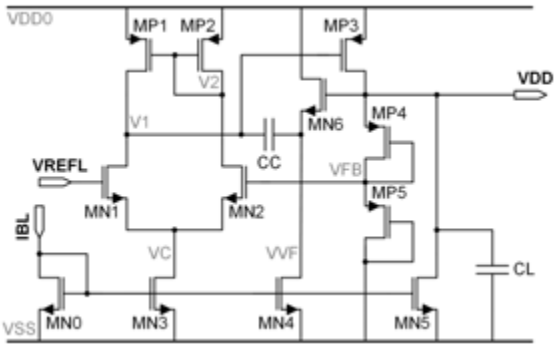


Fig. 1. The schematics of the low-dropout voltage regulator (LDO).

ment. The reference voltage V_{REFL} and reference current I_{BL} are sourced from the SMU (Source Measure Unit) during these measurements. In addition, the SMU is used as a current sink for the LDO regulator output current. The results are shown in Fig. 2 and Fig. 3. For the line regulation measurement, the input voltage is swept from 1 V to 2 V in 10 mV steps. Only the results from 1.2 V to 2 V are shown to better see the behavior at 1.8 V and 1.2 V to 1.98 V which is the expected operating range. It can be seen that the dropout voltage is around 20 mV because the output voltage V_{DD} is slightly above 1.18 V for a set input voltage V_{DD0} = 1.20 V. As the input voltage increases above 1.8 V, there is a noticeable rise in the output voltage.

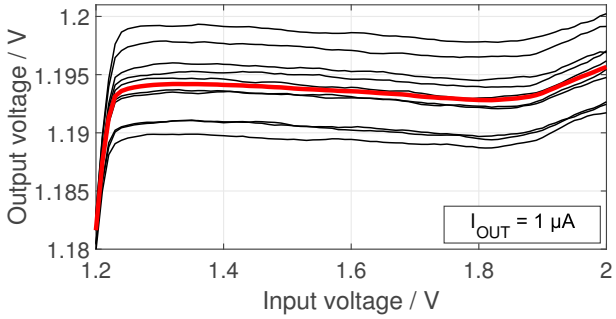


Fig. 2. Line regulation of the LDO regulator measured on a sample of ten chips. The average of all measurements is depicted by the red line.

As it can be expected, the output voltage drops linearly by increasing the output current. All ten manufactured regulators are performing well within the specifications under static conditions, since the regulated output voltage is between 1.1 V and 1.3 V in all observed scenarios.

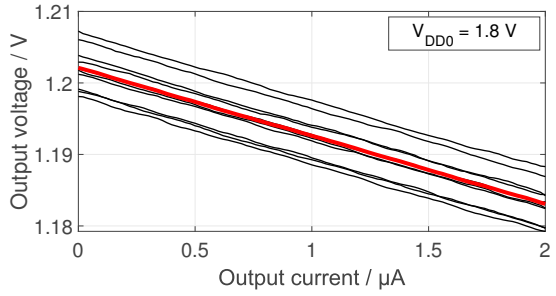


Fig. 3. Load regulation of the LDO regulator measured on a sample of ten chips. The average of all measurements is depicted by the red line.

C. Static temperature characteristics

The output voltage is measured in the temperature range from -40°C to 80°C for a fixed input voltage V_{DD0} = 1.8 V and the output current I_{DD} = 1 μA. A segment of the used temperature profile along with the corresponding voltage measurement is shown in Fig. 4. The calculated results along with a third order polynomial fit of the obtained measurement data is shown in Fig. 5.

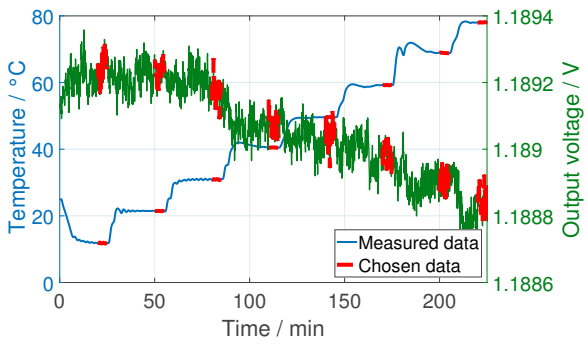


Fig. 4. Temperature profile from 10°C to 80°C and response of the output voltage V_{DD} to the temperature sweep.

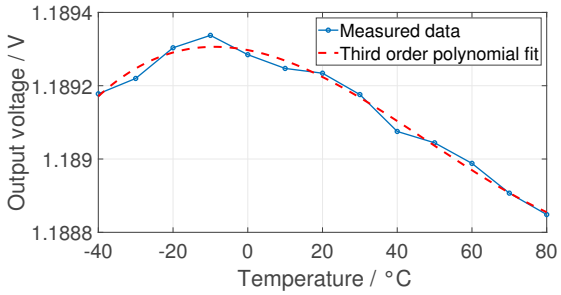


Fig. 5. Temperature dependence of the LDO regulator output voltage.

V. Reference Generator

A. Reference generator architecture

The basic architecture of the reference voltage and current generator (REFG) is similar to the circuit know as the Widlar's bandgap voltage reference, which is usually implemented in bipolar technology because of the favourable exponential characteristic of a pn-junction. A good replacement for bipolar transistors is recognized in MOSFET transistors which work in weak inversion, also known as sub-threshold conduction (exponential characteristics (1)).

$$I_D \approx I_S \frac{W}{L} e^{\frac{V_{GS}-V_{TH}}{nV_T}} \quad (1)$$

Weak inversion occurs when the voltage V_{GS} of a MOSFET transistor is smaller than the threshold voltage V_{TH}. When in weak inversion, a transistor can work with very small currents (dozens of nanoamperes), which makes it suitable for low power applications [6], [8].

Temperature compensation of the outputs is achieved by connecting PTAT (proportional to absolute temperature) resistor and CTAT (complementary to absolute temperature) diode-connected NMOS voltages in series. The main feature of this design is stable output regardless of power supply, temperature and current consumption variations.

At the level of the whole Real-Time Clock (RTC) circuit, the specifications of the REFG circuit are defined and shown in Table I.

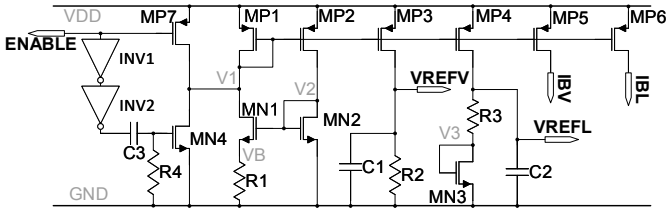


Fig. 6. The schematics of the reference generator (REFG).

TABLE I
AGREED SPECIFICATIONS OF THE REFG CIRCUIT

Parameter	Description	Min	Nom	Max	Unit
V_{DD}	Input voltage	1.1	1.2	1.3	V
I_{DD}	Current consumption	150	250	400	nA
V_{REFV}	PTAT referent voltage	0.28	0.35	0.45	V
V_{REFL}	CWT referent voltage	0.55	0.6	0.65	V
I_{BV}	PTAT current	25	50	75	nA
I_{BL}	PTAT current	25	50	75	nA
$t_{settling}$	Settling time			~32	μs

B. Line regulation

The architecture of the REFG has inherently excellent load regulation. In theory, the output current and voltages should not depend on the supply voltage. This follows from the expression for the voltage V_B and current through this branch (this current is mirrored in all other branches), which does not depend on the supply voltage.

Line regulation of V_{REFV} and V_{REFL} output voltages is measured for 10 different chips and a large supply voltage span. The results are shown in Fig. 7 and Fig. 8. The voltage V_{REFV} is in the range from 274 mV to 293 mV and the voltage V_{REFL} is in the range from 600 mV to 618 mV for the value of V_{DD} from 1.1 V to 1.3 V. The relative change of both voltages is below 20 mV which fits well in the specification. The average measured value of the voltage V_{REFL} is 609.7 mV and is close to the specification (600 mV). The average measured value of the voltage V_{REFV} is 283.9 mV, this value is 18.88% lower than the specification (350 mV). It is speculated that this is due to the layout and tolerances of the resistor $R2$. The corner analysis showed that this case is possible.

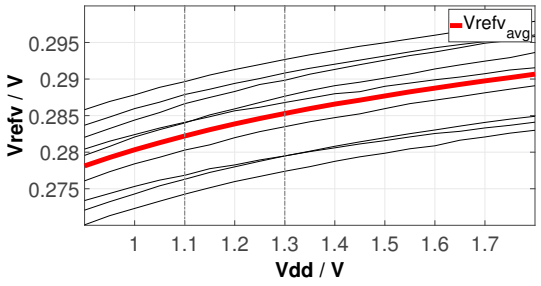


Fig. 7. Line regulation of the V_{REFV} output voltage measured on a sample of ten chips. The average of all measurements is depicted by the red line.

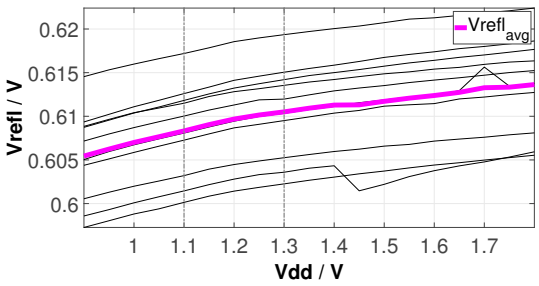


Fig. 8. Line regulation of the V_{REFL} output voltage measured on a sample of ten chips. The average of all measurements is depicted by the magenta line.

C. Static temperature characteristics

In order to generate PTAT voltage in the REFG circuit the voltage difference between the gate-source voltages V_{GS} of transistors M_{N1} and M_{N2} is used (2).

$$V_B = V_{GS2} - V_{GS1} \approx nV_T \ln\left(\frac{W_{MN1}}{W_{MN2}}\right) \quad (2)$$

The CWT (constant with temperature) voltage V_{REFL} is achieved by combining generated PTAT current and inherently CTAT transistor voltage.

The measured and simulated temperature characteristics are shown in Fig. 9 and 10. The characteristics are approximated using the second order Taylor series (3) around the nominal temperature of T_0 = 27 °C.

$$V(T) = V(T_0) \cdot (1 + TC_1 \cdot (T - T_0) + TC_2 \cdot (T - T_0)^2) \quad (3)$$

For the voltage V_{REFV} the coefficients are:

- $TC1_{measured}$ = 2941 ppm/°C
- $TC1_{simulation}$ = 2926 ppm/°C
- $TC2_{measured}$ = 1.545 ppm/°C^2
- $TC2_{simulation}$ = 3.246 ppm/°C^2

For the voltage V_{REFL} the coefficients are:

- $TC1_{measured}$ = -36.45 ppm/°C
- $TC1_{simulation}$ = -37.67 ppm/°C
- $TC2_{measured}$ = 0.7863 ppm/°C^2
- $TC2_{simulation}$ = 1.908 ppm/°C^2

The measurement results are very similar to the Cadence Virtuoso simulation results and the circuit is behaving as theory predicted.

D. Transient and AC analysis

Without the start-up circuit, the reference generator tends to turn on very slowly or not to turn on at all. The start-up circuit solves the problem using two inverters, a pair of CMOS transistors, a resistor and a capacitor as shown in the schematic in Fig. 6. The voltage V_{REFL} settles after 16 μs and the voltage V_{REFV} after 40 μs from the 50% of the rising edge of the *ENABLE* signal.

The AC analysis shows that the circuit has a high Power Supply Rejection Ratio (PSRR). The attenuation of voltage disturbances between the V_{DD} pin and output is in the range from -15 dB to -40 dB.

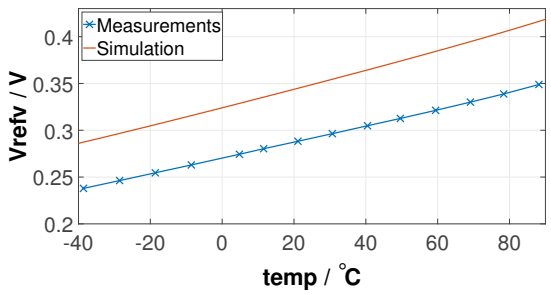


Fig. 9. Temperature dependence of the V_{REFV} output voltage.

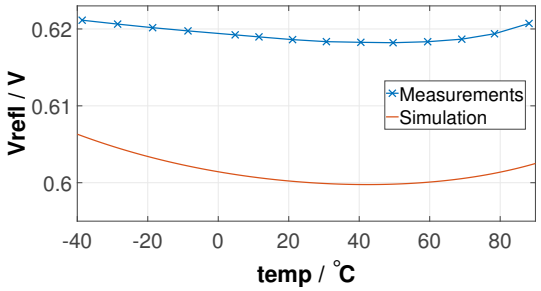


Fig. 10. Temperature dependence of the V_{REFL} output voltage.

VI. Voltage to Current Converter

The voltage to current converter (V2I) is a two-stage Miller amplifier. The transistor M_{P9} is added to the circuit to reduce the impact of a positive zero introduced by C_C. The goal was to design a low power voltage to current converter that consumes up to 400 nA in nominal conditions at a supply voltage of 1.2 V [3], [4]. A gain-bandwidth GBW of at least 64 kHz and a phase margin PM of at least 70° should be provided for at nominal conditions.

A. Description of circuit operation

Fig. 11 shows the schematic of the voltage to current converter. The role of a voltage to current converter in the 1-Hz real-time clock (1-Hz RTC) electronic system is to receive a reference voltage V_{REFV} from a reference voltage generator REFG, and to propagate it through a buffer amplifier to its output V_{REF}. In addition, using the current I_{BV} supplied from the same REFG generator, the output currents I_{REF} and I_{BC} should be generated using the current mirror and the reference resistor R_{REF}.

B. Measurements

The measurements are made in such a way that one parameter is changed while the other two are kept at a constant nominal value in order to see the dependence of the output voltage V_{REF} on the parameters: input current I_{BV}, input voltage V_{REFV} and supply voltage V_{DD}. Table II shows the nominal values of the parameters.

Fig. 12 shows the dependence of the output voltage V_{REF} on the input voltage V_{REFV}. Fig. 13 shows the dependence

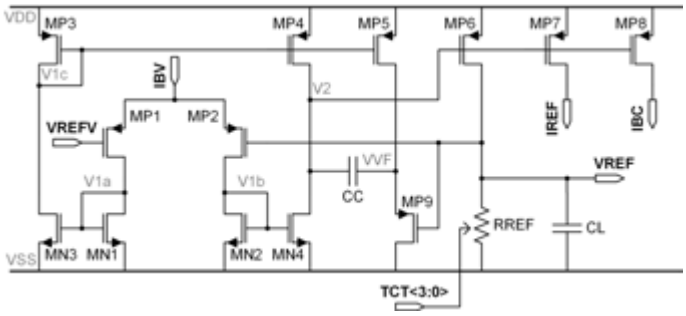


Fig. 11. Schematic of voltage to current converter.

TABLE II
Nominal Parameter Values

Parameter	Nominal value	Unit
V_{DD}	1.2	V
I_{BV}	50	nA
V_{REFV}	0.35	V
V_{REF}	0.35	V
I_{REF}	100	nA
I_{BC}	50	nA

of the output voltage V_{REF} on the input current I_{BV}, and Fig. 14 shows the dependence of the output voltage V_{REF} on the supply voltage V_{DD}. The measurements are performed on 10 chips so that the maximum deviations of parameters could be observed through statistical analysis.

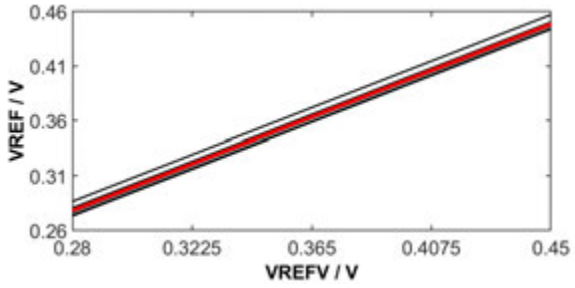


Fig. 12. Output voltage V_{REF} with changing input voltage V_{REFV}. Measured on a sample of 10 chips. The average of all measurements is depicted by the red line. The voltage gain is 1.0003 V/V.

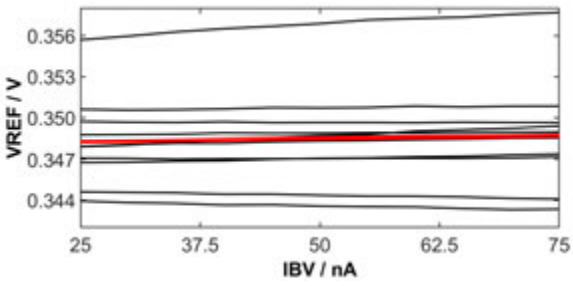


Fig. 13. Output voltage V_{REF} with changing input current I_{BV}. Measured on a sample of 10 chips. The average of all measurements is depicted by the red line. The transimpedance is 8.06 μV/nA.

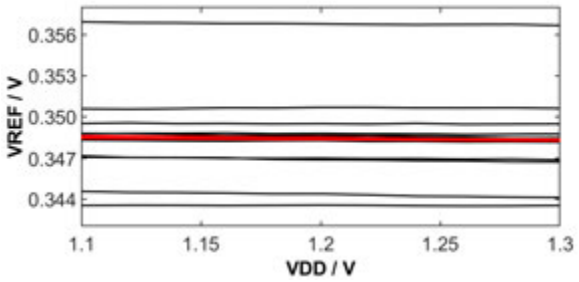


Fig. 14. Output voltage V_{REF} with changing supply voltage V_{DD}. Measured on a sample of 10 chips. The average of all measurements is depicted by the red line. The gain variation with supply voltage is -911 μV/V.

VII. Reference resistor

The reference resistor module has an important role in the overall functionality of the RTC. It allows for a simple temperature compensation of the whole circuit by the means of a 4-bit digital interface with two's complement input (TCT trim vector) [1], [5].

A. Reference resistor architecture

The resultant circuit architecture consists of a series of resistors with negative and a series of resistors with positive temperature coefficient (Fig. 15). The total nominal resistance is constant and equal to 3.5 MΩ. By combining the resistors R_N with negative and the resistors R_P with positive temperature coefficient, it is possible to adjust the temperature coefficient of the whole reference resistor R_{REF} to zero in a narrow temperature range, regardless of variations in process parameters and by keeping the same value of R_{REF}. Switching between R_N and R_P resistors is achieved by implementing an analog CMOS switch matrix. These switches are controlled by the 4-to-16 decoder (Fig. 16). At a specific time, only one analog switch out of 16 switches is turned on.

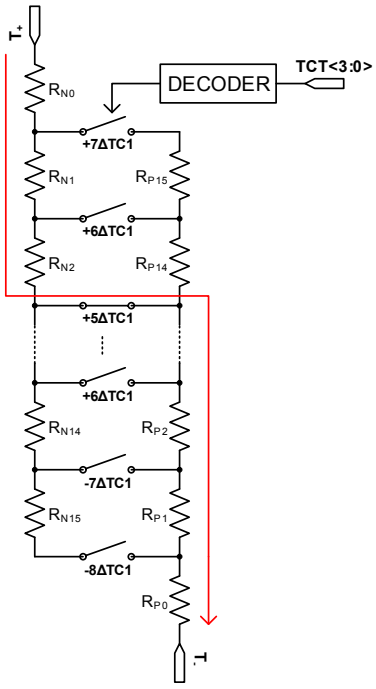


Fig. 15. Reference resistor functional schematic.

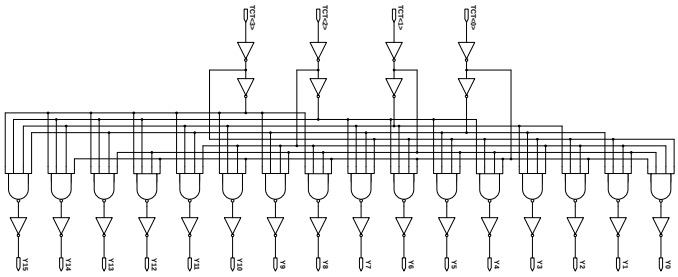


Fig. 16. Decoder schematic.

B. Description of temperature compensation

The oscillator core frequency is determined by the reference resistance R_{REF}, reference capacitance C_{REF} and propagation time delay t_d according to the following relation:

$$f_{oscc} = \frac{1}{2 \cdot (R_{REF} \cdot C_{REF} + t_d)} . \quad (4)$$

From this expression, we can clearly see the direct dependence of the output frequency as the function of the time constant $\tau = R_{REF} \cdot C_{REF}$, hence the total first order temperature coefficient of the output frequency of the oscillator $TC1_{OSC}$ is determined by the sum of the first order temperature coefficients of the reference resistor $TC1_{RREF}$, reference capacitor $TC1_{CREF}$ and residual oscillator core contribution $TC1_{OSCC}$:

$$TC1_{OSC} = TC1_{RREF} + TC1_{CREF} + TC1_{OSCC} . \quad (5)$$

By adjusting $TC1_{RREF}$ it is possible to minimize $|TC1_{OSC}|$ and therefore achieve temperature compensation of the circuit output.

C. Static temperature characteristics

In the general case, the integrated resistors (such as poly and diffusion resistors) are modelled as being both temperature and voltage dependent. But given that we switch between the two identical strings of resistors connected in series, which have the same constant current passing through them, we can ignore the voltage dependence and model the total change in the network resistance by considering only the temperature dependence $R_{REF} = f(T)$:

$$R_{REF}(T) = R_{REF0} \cdot \left(1 + TC1 \cdot \delta T + TC2 \cdot \delta T^2\right) \quad (6)$$

where $\delta T = T - 25\,[^\circ\mathrm{C}]$. From (6), the formula for the total nominal first order temperature coefficient is derived:

$$TC1 = \Bigg((TC1_{N0} - TC1_P) \cdot R_{N0} + (TC1_{P0} - TC1_P) \cdot R_{P0} + (TC1_N - TC1_P) \cdot \sum_{i=1}^{k} R_{Ni} + TC1_P \cdot R_{REF0} \Bigg) \cdot \frac{1}{R_{REF0}} \quad (7)$$

where $k = 1, 2, 3, \ldots, 15$ denotes the number of R_N of the resistor in the middle part and is determined by the decoder output, i.e. the trim vector input. Using the expression for $TC1$ in this form, the initial parameters for R_{N0}, R_{P0} are defined, as well as the rest of the resistor ladder, which are further optimised with the help of simulations. The measurement results for the final design are shown for the $R_{REF}(T)$ characteristics for three ICs and three input TCT trim vectors in Fig. 17, 18 and 19. Each measurement point is marked by an X and the complete characteristic is then interpolated by the third order polynomial function.

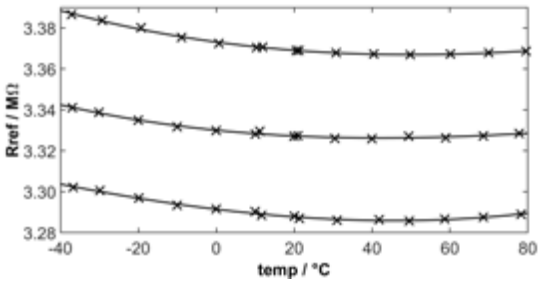


Fig. 17. Measurements of the temperature dependence of the reference resistor with the TCT input code of value 1000 for three different IC samples.

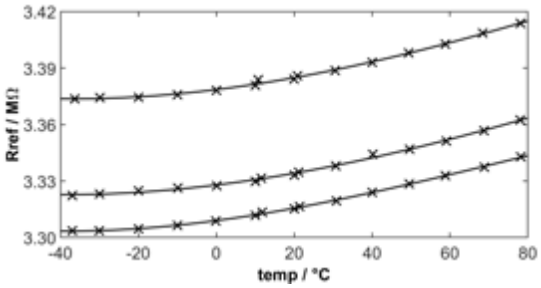


Fig. 18. Measurements of the temperature dependence of the reference resistor with the TCT input code of value 0000 for three different IC samples.

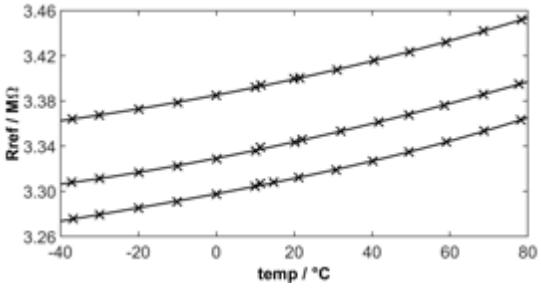


Fig. 19. Measurements of the temperature dependence of the reference resistor with the TCT input code of value 0111 for three different IC samples.

The average resistance (at T_0 = 25 °C) calculated from this set of measurements (i.e. R_{REF} for three different ICs and three different TCT values at nominal temperature) is R_{REF0} = 3.3438 MΩ with the minimum R_{REFmin} = 3.2869 MΩ and the maximum value R_{REFmax} = 3.4044 MΩ. The standard deviation is σ_{RREF} = 0.0374 MΩ. The first order temperature coefficients calculated for three TCT vectors as an average of the characteristics of three IC samples are:

- For TCT = 1000 (minimum): $TC1$ = -68.23 ppm/°C
- For TCT = 0000 (nominal): $TC1$ = 78.51 ppm/°C
- For TCT = 0111 (maximum): $TC1$ = 196.45 ppm/°C

VIII. Conclusion

The low-dropout voltage regulator, reference generator, voltage to current converter and reference resistor, which together constitute the tunable temperature coefficient voltage reference, are designed and measured. The line regulation of the low-dropout voltage regulator is 5.19 mV/V and the load regulation is -9.51 mV/μA. The line regulation of the reference generator is 11 mV/V and the first order temperature coefficient of the constant with temperature output voltage is -36.45 ppm/°C. The voltage gain of the voltage to current converter is 1.0003 V/V and transimpedance is 8.06 μV/nA. The reference resistor has the trimmable first order temperature coefficient with the nominal resistance of 3.3438 MΩ. The minimum value of the temperature coefficient is -68.23 ppm/°C, the nominal value is 78.51 ppm/°C and the maximum value is 196.45 ppm/°C. The most important feature of the system as a whole is the proper behaviour at different temperatures. The completed integrated circuit design is suitable for usage in a temperature compensated RTC system.

Acknowledgment

This work is partially supported by ams AG, Premstaetten, Austria.

References

[1] Y. Tsai and L. Lu, "A 51.3-MHz 21.8-ppm/°C CMOS relaxation oscillator with temperature compensation," EEE Trans. Circuits Syst. II Express Briefs, vol. 64, no. 5, pp. 490–494, 2017.

[2] J. Wang and W. L. Goh, "A 13.5-MHz relaxation oscillator with ±0.5% temperature stability for RFID Applicaton," Circuits and Systems (ISCAS) IEEE International Symposium on, pp. 2431–2434, 2016.

[3] F. Sebastiano, L. J. Breems, K. Makinwa, S. Drago, D. Leenaerts, and B. Nauta, "A low-voltage mobility-based frequency reference for crystal-less ULP radios," IEEE J. Solid-State Circuits, vol. 44, no. 7, pp. 2002–2009, 2009.

[4] K. Tsubaki, T. Hirose, N. Kuroki, and M. Numa, "A 32.55-kHz 472-nW 120ppm/°C fully on-chip variation tolerant CMOS relaxation oscillator for a real-time clock application," IEEE European Solid-State Circuits Conference (ESSCIRC), pp. 315–318, 2013.

[5] A. V. Boas and A. Olmos, "A temperature compensated digitally trimmable on-chip IC oscillator with low voltage inhibit capability," Circuits and Systems (ISCAS) IEEE International Symposium on, vol. 1, pp. 501–504, 2004.

[6] Y. H. Chiang and S. I. Liu, "Nanopower CMOS relaxation oscillators with sub-100ppm/°C temperature coefficient," IEEE Trans. Circuits Syst. II Express Briefs, vol. 61, no. 9, pp. 661–665, 2014.

[7] C. Yilmaz, L. Heiss, C. Werner, and D. Schmitt-Landsiedel, "Modeling of NBTI-recovery effects in analog CMOS circuits," Reliability Physics Symposium (IRPS) IEEE International, pp. 1–4, 2013.

[8] I. Fakharyan and M. Ehsanian, "A sub-1V nanowatt CMOS bandgap voltage reference with temperature coefficient of 13ppm/°C," 2015 23rd Iranian Conference on Electrical Engineering, pp. 1129–1132, 2015.

Mismatch Reduction Techniques for Current-Mirror Based Potentiostats

Markus Haberler*†, Inge Siegl*†, Christoph Steffan*, and Mario Auer†
*Infineon Technologies Austria AG, Graz, Austria
†Graz University of Technology, Institute of Electronics, Graz, Austria
markus.haberler@infineon.com / markus.haberler@student.tugraz.at

***Abstract*—Current-mirror based potentiostats suffer from systematic and random errors that mainly appear due to mismatch in the current-mirror. To limit the standard deviation of the error to less than 1.1 % over a dynamic range of four decades, either a high overdrive voltage or comparable big transistor dimensions are required. In this work, an optimized design approach for the current-mirror is presented. Based on the given accuracy and dynamic range requirements, an optimized programmable current-mirror is implemented. An accuracy driven design approach is used to achieve the required accuracy over the wide range of sensor currents, while keeping the area consumption low. Compared to a linearly scaled current-mirror based potentiostat design, an output stage area reduction by a factor of four is achieved. Additionally, the voltage headroom consumption is kept at a minimum. The potentiostat is implemented in a 130 nm CMOS technology and consumes an area of only 0.14 mm^2. The simulation results show that even in the worst case the specified accuracy is achieved over the complete dynamic range of sensor currents from nA to mA.**

***Index Terms*—Point-of-care (PoC) diagnostics, potentiostat, current-mirror, dynamic element matching, wide dynamic range**

I. Introduction

Point-of-care (PoC) diagnostics for biomedical applications have received increasing attention in recent years. The continuously rising technological connectivity among the population suggests to replace traditional optical PoC diagnostics by ones that are based on electrochemical readout principles. These electrochemical diagnose platforms promise to provide highly-accurate and fast results in non-laboratory environments, while keeping the medical costs at a minimum [1].

Diagnostics that are based on electrochemical detection mostly use amperometric measurement methods to extract information from the biosensor. These methods capture the current that results from a redox reaction at the sensor electrodes [2]. Amperometric measurement principles are usually carried out by three-electrode potentiostats. A potentiostat is a sensor front-end circuit, which maintains a desired potential difference between the working electrode (WE) and the reference electrode (RE) by adjusting the counter electrode (CE) according to the current that arises due to the electrochemical reaction [2].

This research is part of a project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 761167 (IMPETUS).

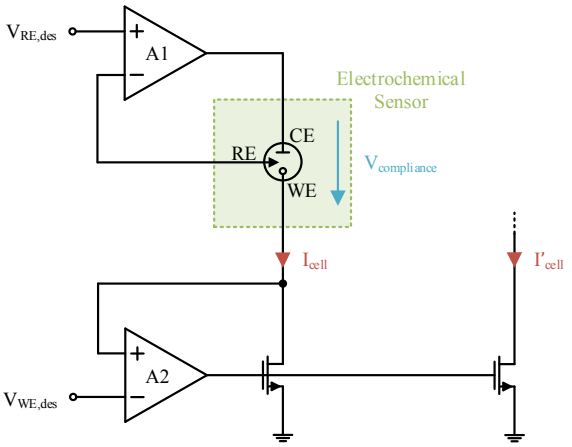


Fig. 1. Conceptual schematic of a current-mirror based potentiostat. The voltages $V_{RE,des}$ and $V_{WE,des}$ are used to drive the sensor. The resulting current is measured.

There are several topologies of potentiostats available, among which the amplifier that drives the WE and extracts the information of the biosensor is of special design interest. Topologies consisting of transimpedance amplifier (TIA) based designs [3], [4], current-conveyor based designs [5]–[8], and current-mirror based designs [9]–[11] were recently presented. TIA based designs and current-conveyor based designs suffer from a high voltage headroom requirement, thereby limiting the possible compliance voltage range for the sensor. The current-mirror based approach on the other hand, as shown in Fig. 1, just requires one transistor saturation voltage for proper operation. This makes this topology especially interesting for low voltage applications.

The advantage of the low voltage headroom requirement of current-mirror based designs comes at the expense of an additional error source due to the current-mirror. The systematic and random errors in the circuit lead to gain, offset and linearity errors in the sensor front-end. Conceivable ways to minimize these undesired effects are to increase the overdrive voltage of the current-mirror or to increase the active area of the corresponding transistors. Both of these possible countermeasures have their drawbacks. The former approach restricts the compliance voltage range, whereas the latter approach leads to a big area consumption, especially when designing for a wide dynamic range of sensor currents. In this work,

an optimized implementation is shown, which achieves high accuracy over a wide range of sensor currents, while the area and voltage headroom consumption is kept at a minimum.

The paper is structured in the following way. Section II describes the fundamental design considerations for the current-mirror based potentiostat. In Section III, the optimized design approach of the current-mirror is described and a comparison to a linearly scaled design is given. Section IV shows the achieved performance of the implementation. Finally, Section V concludes the paper by summarizing the main findings of the research work.

II. System Architecture of the Current-Mirror Based Potentiostat

In a current-mirror based potentiostat, the voltage across the electrochemical sensor is set by two operational amplifiers that are operated in a feedback-configuration, as shown in Fig. 1. The sensor current, which flows through the NMOS transistor controlled by amplifier $A2$, is replicated by a current-mirror structure. This replica can then be processed in the additional branch without any influence on the available compliance voltage range of the potentiostat.

However, adding a current-mirror adds an additional potential error source to the sensor interface. To minimize the systematic and random errors in the current-mirror, optimizations of the current-mirror based potentiostat were implemented. These optimizations include the use of a regulated-cascode, the implementation of an area-optimized programmable current-mirror and the use of dynamic element matching (DEM). A simplified schematic of the implemented topology is shown in Fig. 2.

The regulated-cascode keeps the difference between the drain-source voltages of the current-mirror transistors low and reduces errors related to the finite output impedance of the transistors. A symmetrical OTA with a PMOS input pair is used to perform this task. The choice of a PMOS input pair allows the potentiostat to use WE potentials close to ground. Thus, the addition of the regulated-cascode has no influence on the compliance voltage range of the topology.

To achieve a wide dynamic range in the sensor front-end, a programmable current-mirror was implemented. A 2-bit adjustable current-mirror for current down-scaling is used to process sensor currents up to the mA range. Using scaling factors of 1, 10, 100, and 1 000, a down-scaling of the sensor current to a range of 0 to 1 µA is possible. The down-scaling minimizes the power consumption of the front-end and reduces the constraints for the post-processing circuity, i.e. the analog-to-digital converter (ADC). An ADC with a fixed input full-scale range is then used to process the sensor current information. Depending on the result of the analog-to-digital conversion, the programmable current-mirror is adjusted to achieve optimum performance.

However, the area consumption of a high scaling factor current-mirror strongly depends on the required mirror accuracy. In the particular application of amperometric

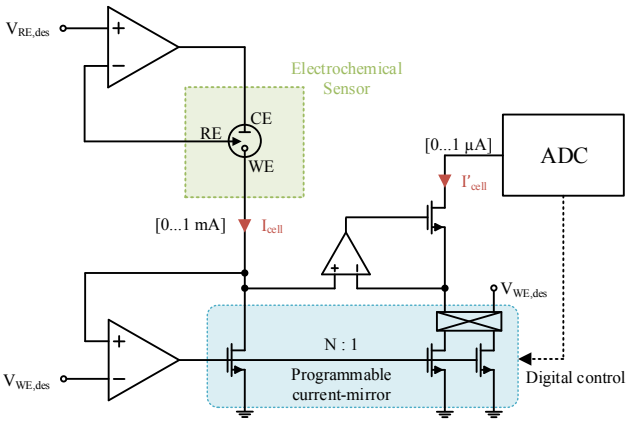


Fig. 2. Simplified schematic of the implemented current-mirror based potentiostat. A regulated-cascode is used to minimize the systematic error due to channel-length modulation. An area-optimized programmable current-mirror is used to achieve high accuracy over a wide dynamic range. Dynamic element matching on the mirror transistor side is used to further reduce the area consumption of the high current scaling current-mirrors.

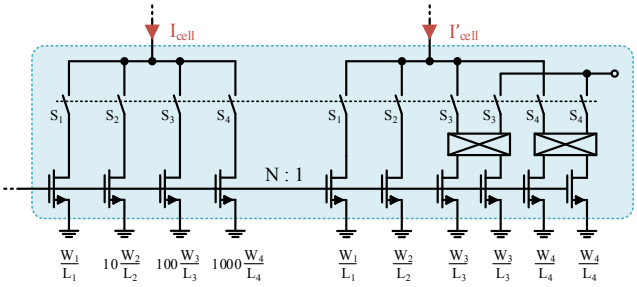


Fig. 3. Conceptual schematic of the programmable current-mirror. Different reference transistor dimensions enable a ratio dependent optimization.

measurements, a high processing accuracy is especially required for the lower current region. Thus, current-mirror scaling by adding additional parallel unit transistor elements on the input side soon leads to a considerable silicon area consumption. The respective optimum depends on the mirror scaling factor, therefore a programmable current-mirror was used to enable a ratio dependent optimization. In this optimized design, the four selectable regions of the current-mirror differ not only by the number of parallel input transistors, they are also different in terms of reference transistor dimension. This is illustrated in Fig. 3. The additional use of DEM results in a further reduction in area. A more detailed description of the realized optimizations in the current-mirror and the thereby achieved improvements are given in Section III.

III. Programmable Current-Mirror Design

Systematic errors are minimized by the use of a regulated-cascode structure. Random errors require proper transistor sizing. Using dynamic techniques area consumption can be reduced. The theoretical background of the design approach and the advantages compared to a linearly scaled design are given in the first part of this section. Based on the findings, the implementation of the optimized

current-mirror potentiostat is presented in the second part of the section.

A. Analytical Description of the Design Approach

The transistor is described using the square-law model [12]. The current factor is abbreviated by β ($= \mu C_{ox} W/L$, where μ is the channel mobility, C_{ox} the gate-oxide capacitance per unit area, W and L the transistor width and length).

While the influence of β mismatch in a current-mirror can only be addressed by increasing the transistor dimensions, threshold voltage mismatch can also be reduced by using a high overdrive voltage for the current-mirror transistors [13]. Assuming an optimal layout implementation and equal drain-source voltages at the transistors, the standard deviation of the input-referred relative current mismatch of a current-mirror, $\sigma(I_{err})/I_{cell}$, operated in saturation region, is found to be

$$\frac{\sigma(I_{err})}{I_{cell}} = \sqrt{\frac{{A_K}^2}{W\,L} + \frac{2}{(V_{gs}-V_{th})^2}\frac{{A_{V_T}}^2}{W\,L}} \cdot \sqrt{\frac{N+1}{N}}, \tag{1}$$

using variance propagation [13]. In (1), I_{err} is the input-referred current mismatch of a current-mirror with an input current of I_{cell} and a scaling factor of N. Furthermore, $V_{gs}-V_{th}$ is the overdrive voltage of the current-mirror, A_K and A_{V_T} the β and threshold voltage mismatch parameter of the transistors, and W and L the reference transistor width and length. As observable in (1), an increasing scaling factor N leads to a decreasing value of the term $\sqrt{(N+1)/N}$, which soon approaches a value of 1 for increasing mirror factors. Compared to a 1:1 current-mirror, a linearly scaled sizing with parallel unit transistor elements thus leads to an accuracy improvement by a factor of $\sqrt{2}$ for $N \gg 1$.

Without taking advantage of the increased accuracy, the area consumption of the current-mirror would scale by N. However, considering this property, a lower area requirement is obtained. In this area-optimized design with regard to a certain accuracy requirement, the left side of (1) is kept constant, and the required reference transistor area depending on the current scaling factor N is determined. The corresponding reordering of (1) leads to

$$W\,L = \left({A_K}^2 + \frac{2}{(V_{gs}-V_{th})^2}\cdot {A_{V_T}}^2\right) \cdot \frac{N+1}{N} \cdot \left(\frac{I_{cell}}{\sigma(I_{err})}\right)^2. \tag{2}$$

For increasing current scaling factors, the reference transistor area decreases in this design approach. Because of the term $(N+1)/N$ in (2), high scaling factor current-mirrors require a reference transistor area that is just half of the area of a low current scaling factor mirror. In other words, only the reference transistor is responsible for the inaccuracy of the current processing at high scaling factors.

Figs. 4 to 6 visualize the relationship between current-mirror accuracy and required active transistor area as a function of the current-mirror scaling factor. In these plots, the dashed trace shows the relationships for the linear current-mirror sizing, whereas the dotted trace shows the relationships for the proposed ratio dependent sizing. The visualizations are based on the analytical descriptions (1) and (2). As observable in Fig. 6, the biggest area saving due to the ratio dependent optimization is achieved in the high current scaling region. A current-mirror area reduction by a factor of 2 is obtained.

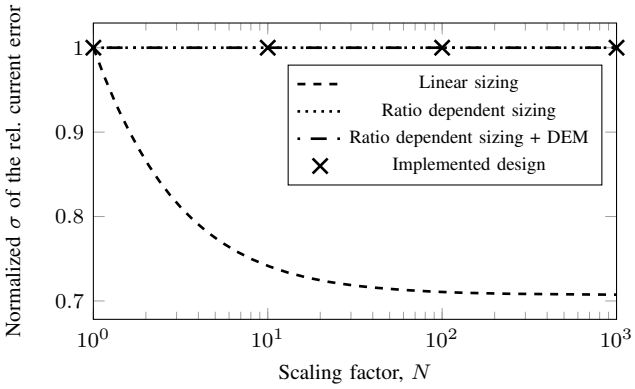


Fig. 4. Standard deviation of the relative current error depending on the current-mirror scaling factor for the different design approaches. Normalized plot to the standard deviation of a 1:1 current-mirror.

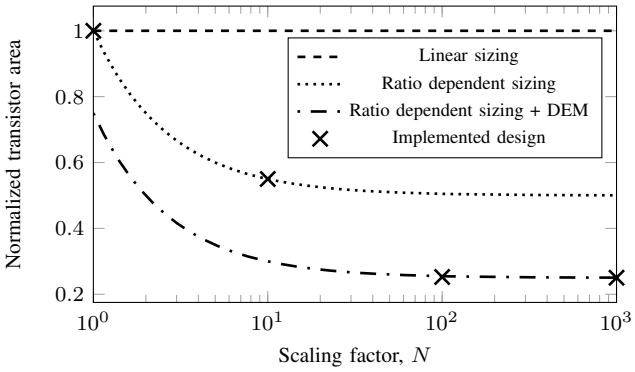


Fig. 5. Reference transistor area depending on the current-mirror scaling factor for the different design approaches. Normalized plot to the reference transistor area of a 1:1 current-mirror.

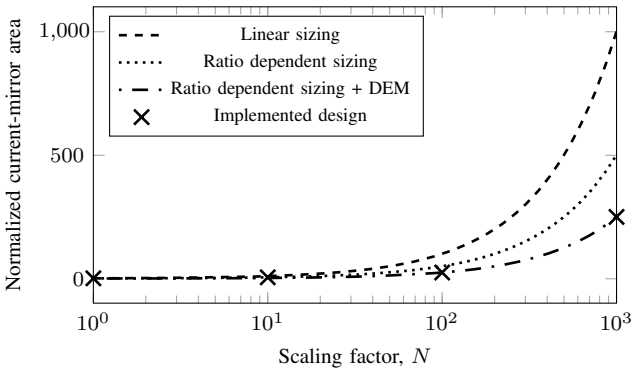


Fig. 6. Total current-mirror area consumption depending on the current-mirror scaling factor for the different design approaches. Normalized plot to the area consumption of a 1:1 current-mirror.

However, the area consumption of the high scaling factor current-mirrors is still dominating the overall area consumption. To address this problem, an additional optimization using DEM techniques was done in the circuit level implementation of the current-mirror.

B. Implementation of the Design Approach

As discussed in Section II, the wide dynamic range in the sensor front-end is reached by using a programmable current-mirror. One of the design limitations of the current-mirror would be flicker noise, which can be reduced by increasing the size of the transistors. To limit the current error to less than $1.1\,\%$ and the overdrive voltage to $200\,\mathrm{mV}$, a minimum transistor area of $80\,\mu\mathrm{m}^2$ with a W/L ratio of 0.2 was determined by simulation for the 1:1 current-mirror. Hence, the area of the transistors is already sufficiently high to keep the flicker noise at an acceptable level. According to the optimized design approach shown above, current-mirror areas of $160\,\mu\mathrm{m}^2$, $484\,\mu\mathrm{m}^2$, $4\,080\,\mu\mathrm{m}^2$, and $40\,080\,\mu\mathrm{m}^2$ result for the scaling factors 1, 10, 100, and 1 000, respectively. Apparently, the main area is consumed by the current-mirrors with scaling factors of 100 and 1 000.

To further optimize the implementation with regard to area consumption, DEM techniques were applied. As pointed out earlier, for the high scaling factors only the reference device transistor on the mirror side has a significant contribution to the current mismatch. Therefore, a dynamic switching between two reference transistors on the mirror side was implemented to increase the accuracy, as shown in Fig. 2. Using DEM only on the mirror side has the additional advantage of keeping the electrode connection away from a potential influence due to high-frequency switching, which is essential to perform accurate electrochemical analyses. The improvement due to DEM allows a further reduction of the transistor areas by a factor of up to 2, while achieving a similar accuracy in the current-mirror as without DEM. This is illustrated by dash-dotted traces in Figs. 4 to 6. Again, the biggest area savings are achieved in the high current regions. For this particular design, this leads to an active area reduction to roughly $2\,080\,\mu\mathrm{m}^2$ and $20\,080\,\mu\mathrm{m}^2$ for the current-mirrors with scaling factors of 100 and 1 000, respectively.

The relationship between accuracy and area consumption of the finally implemented design approach is additionally visualized by crosses in Figs. 4 to 6. Since the area consumption of the current-mirrors with scaling factors of 1 and 10 is comparable low, they have been implemented without DEM. The overall advantage in area consumption compared to a current-mirror design with linearly sized transistors is apparent. With focus on the application of low-cost PoC diagnostics, this results in significant cost savings in the production of the electrochemical read-out device.

IV. Potentiostat Performance Evaluation

The current processing accuracy of the implemented sensor front-end was characterized by Monte Carlo analysis. In this analysis, local as well as global process and mismatch parameter variation were taken into account. Depending on the sensor current, the appropriate range was selected.

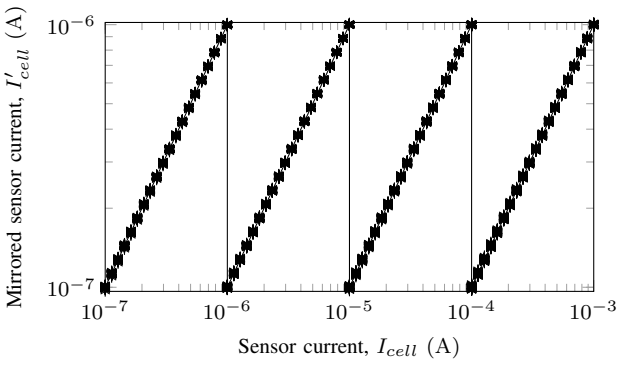


Fig. 7. Transfer characteristic of the implemented current-mirror based potentiostat. The solid line shows the nominal simulation run, whereas the stars show the result of the Monte Carlo simulation.

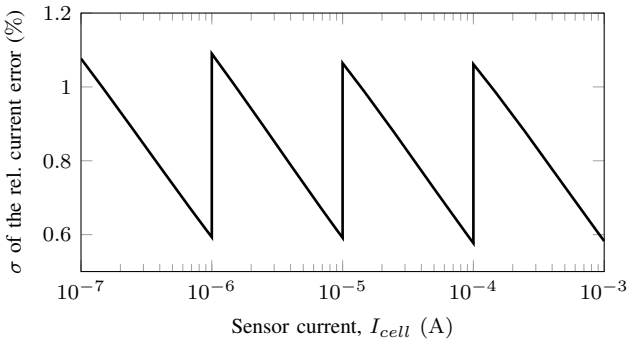


Fig. 8. Standard deviation of the relative current error of the implemented current-mirror based potentiostat.

Fig. 7 shows the resulting transfer characteristic of the current-mirror based potentiostat. The four regions of the programmable current-mirror and the corresponding scaling of the sensor current are clearly observable. The standard deviation of the relative current error depending on the sensor current is shown in Fig. 8. As can be seen, the current error in each region of the programmable current-mirror decreases with increasing sensor current. This is due to a rising overdrive voltage in the current-mirror for higher currents. The standard deviation of the worst-case current error is less than $1.1\,\%$ over the complete dynamic range of sensor currents. This allows the current-mirror potentiostat to accurately distinguish between biochemical parameters in the application of amperometric measurements.

The circuit was realized in a 130 nm standard CMOS process as a main building block of a wirelessly-operated electrochemical measurement platform IC. Fig. 9 shows a layout picture of a test-chip that is currently in production. The chip contains a slightly modified version of the presented current-mirror based potentiostat. Compared to the discussed implementation, the area of the high scaling factor output stage is reduced due to an application related requirement, leading to a standard deviation of the worst-

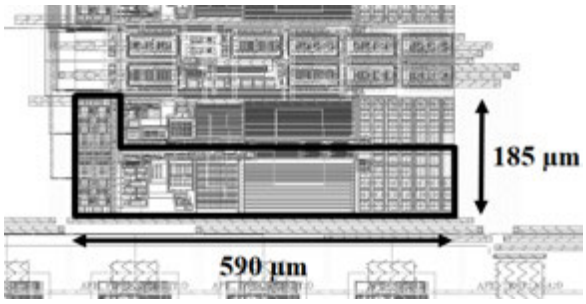


Fig. 9. Part of the test-chip layout containing the current-mirror based potentiostat.

case error of roughly $2.2\,\%$ in this region. Taking the additional area consumption of this stage into account, the total area of the presented implementation can be estimated to a value of less than $0.14\,\mathrm{mm}^2$.

V. Conclusion

Accuracy and area consumption trade-offs in current-mirror based potentiostats were presented. A design methodology for the current-mirror optimization using a reconfigurable topology was proposed. The accuracy driven design approach leads to a reduction in the area consumption of wide dynamic range current-mirrors due to a mismatch-optimized transistor sizing. Analytically derived design rules were verified by simulation in a 130 nm CMOS technology. The achieved accuracy over the complete dynamic range proves that the interface is suitable for amperometric measurement methods. The low area consumption and the high compliance voltage range of the design makes it a potential candidate for low-cost PoC diagnostics in biomedical applications.

References

[1] M. Zarei, "Portable biosensing devices for point-of-care diagnostics: Recent developments and applications," *TrAC Trends in Analytical Chemistry*, vol. 91, pp. 26–41, 2017.

[2] A. J. Bard and L. R. Faulkner, *Electrochemical methods: Fundamentals and applications*, 2nd ed. New York and Chichester: John Wiley, 2000.

[3] M. Punjiya, P. Mostafalu, and S. Sonkusale, "Low-cost paper-based electrochemical sensors with CMOS readout IC," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*. Piscataway, NJ: IEEE, 2014, pp. 324–327.

[4] H. M. Jafari and R. Genov, "Chopper-stabilized bidirectional current acquisition circuits for electrochemical amperometric biosensors," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 5, pp. 1149–1157, 2013.

[5] R. Genov, M. Stanacevic, M. Naware, G. Cauwenberghs, and N. Thakor, "16-channel integrated potentiostat for distributed neurochemical sensing," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 11, pp. 2371–2376, 2006.

[6] M. H. Nazari, H. Mazhab-Jafari, L. Leng, A. Guenther, and R. Genov, "CMOS neurotransmitter microarray: 96-channel integrated potentiostat with on-die microsensors," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 7, no. 3, pp. 338–348, 2013.

[7] H. Li, C. S. Boling, and A. J. Mason, "CMOS amperometric ADC with high sensitivity, dynamic range and power efficiency for air quality monitoring," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 4, pp. 817–827, 2016.

[8] D. de Venuto, M. D. Torre, C. Boero, S. Carrara, and G. de Micheli, "A novel multi-working electrode potentiostat for electrochemical detection of metabolites," in *2010 IEEE Sensors*. Piscataway NJ: IEEE, 2010, pp. 1572–1577.

[9] M. M. Ahmadi and G. A. Jullien, "A wireless-implantable microsystem for continuous blood glucose monitoring," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 3, no. 3, pp. 169–180, 2009.

[10] M. Razzaghpour, S. Rodriguez, E. Alarcon, and A. Rusu, "A highly-accurate low-power CMOS potentiostat for implantable biosensors," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*. Piscataway, NJ: IEEE, 2011, pp. 5–8.

[11] W.-S. Wang, W.-T. Kuo, H.-Y. Huang, and C.-H. Luo, "Wide dynamic range CMOS potentiostat for amperometric chemical sensor," *Sensors (Basel, Switzerland)*, vol. 10, no. 3, pp. 1782–1797, 2010.

[12] H. Shichman and D. A. Hodges, "Modeling and simulation of insulated-gate field-effect transistor switching circuits," *IEEE Journal of Solid-State Circuits*, vol. 3, no. 3, pp. 285–289, 1968.

[13] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors," *IEEE Journal of Solid-State Circuits*, vol. 24, no. 5, pp. 1433–1439, 1989.

Design of CMOS Temperature Sensors Based on Ring Oscillators in 180-nm and 110-nm technology

Ivan Porin Tolić*, Josip Mikulić*, Gregor Schatzberger* and Adrijan Barić†
*ams AG, Tobelbader Strasse 30, Premstätten 8141, Austria
†University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, Zagreb, Croatia
ivanporin.tolic@ams.com

***Abstract*—In this work two on-chip smart temperature sensors are presented. Both designs are based on two ring oscillators operating complementary. The readout circuit is achieved using time-to-digital converter and digital counter. The first sensor is designed in 180-nm technology, and the other is designed in 110-nm. The simulation results from both sensors are presented in the temperature range from -40 to 125°C. The first design achieves an accuracy of +4.9/-0.2°C, while the other achieves an accuracy of ±3.8°C.**

***Index Terms*—CMOS integrated circuit, smart temperature sensor, ring oscillator, time-to-digital converter, digital counter, mixed-signal design**

I. Introduction

The development of modern electronic devices mostly relies on the improvement of the integrated circuit technology. This improvement enabled the design of high-performance circuits, such as systems-on-chip and microprocessors. Those complex integrated circuits are required to perform high number of operations. In order to maximize the number of functions executed per unit of time, high clock frequencies are used in digital circuits. Consequently, fast transistor switching can lead to a significant increase of junction temperatures. The poor and degraded performance of integrated circuits can be avoided using thermal management. Therefore, an on-chip temperature sensor detecting temperature variations is often incorporated as a part of larger design. Thermal management can include multiple temperature sensors distributed all over the chip or only one sensor for calibration of a critical IP block. Considering design requirements, compact and low-power sensors are preferred, especially for the case of distributed sensors. The accuracy of ±1°C is often sufficient for temperature monitoring of IP blocks. Improved performance and better control of the hotspots are ensured using this approach. Smart temperature sensors provide temperature information as an N-bit digital output. In order to achieve this, it is necessary to exploit designs compatible with CMOS digital circuits as this additionally enables the easier scaling of the design. Pushing a design more towards digital or even all-digital implementation can result in better performance with scaling.

BJT based temperature sensors have drawn a lot of attention, as they show very good performance in terms of accuracy over the broad temperature range [1]–[4], [6]. However, to achieve excellent accuracy and resolution, batch-calibration combined with other techniques are often used [2], [5], namely the trimmable current sources and DAC, dynamic element matching and chopping. Moreover, BJT based sensors occupy large area while consuming considerable amount of power, mostly due to the ADC (Analog-to-Digital Converter) needed as a part of the readout. Temperature dependence of the p-n junction is also exploited in various references using bulk diode instead of bipolar transistors [7], [8]. The usage of bulk diode has advantages over using BJT devices in FinFET technology. The design based on diode is rather robust and suitable for technology scaling and combining with digital circuits. Moreover, it is suitable for the case when the supply voltage variations are present and critical [8]. Sensors based on thermal diffusivity are improving their performance with scaling [9], [10], and consequently gaining more attention. Still, this type of sensors is not the first choice for the design as the long measurement time and rather large power consumption dominate over high resolution. The similar argument applies for the resistor based temperature sensors [11]–[14]. CMOS based temperature sensors [15]–[23], [26] are suitable for integration in larger design, and they are, in general, compact while consuming small amount of power. They achieve moderate accuracy, but sometimes they need 2-point batch-calibration to achieve very small error of about ±0.2°C. Another drawback is that they are highly sensitive to process variations and voltage supply variations. Nevertheless, several approaches were presented, mostly based on sensing a temperature dependent MOSFET drain current. The drain currents of the inverter transistors change with respect to the temperature, causing the temperature dependence of the inverter delay. Ring oscillator based sensors [15]–[18], [20], [22], [23], [26] use inverter chains, combined with a digital counter and a TDC (Time-to-Digital Converter) [25], [26] instead of large ADC as a part of readout. Moreover, this type of sensors is often implemented in FPGA [16], [17], as the all-digital implementation is possible. This means that the easier portability of the design through nodes is enabled. Better results are achieved with the ratiometric measurements, as the first order errors are diminished that way. The measurement error can be reduced by 1-point or 2-point trimming, but with self-calibration the trimming effort could be reduced. Self-calibration is included in design [18], which uses two ring oscillators with PTAT (Proportional to Absolute Temperature) and CTAT (Complemetary to Absolute Temperature) frequencies, with the use of several different bias voltages.

The design presented in [20] exploits two ring oscillators with different temperature sensitivities. The supply sensitivity is reduced and no reference frequency is used. The ratio of the frequency follows the PTAT characteristics. Several low power designs are presented [21]–[24]. They rely on the exponential temperature dependence of the weak inversion current. In [22], the ratio of the saturation current and the triode region current, both in weak inversion, gives a linear function of the temperature. Low power design from [23] reports a low power supply sensitivity, achieved by canceling the supply sensitivities of leakage current with the sensitivity of the voltage swing for charging.

This work presents two sensors based on the ring oscillators. The first design consists of two identical ring oscillators working complementary. One sensor is designed in 180-nm, and the other in 110-nm technology. This paper is organized as follows. Section II describes the detailed description of the first sensor design. Section III discusses the implementation details and results of the second sensor. Section IV concludes this work.

II. Complementary Ring Oscillators Based Sensor - First Design

The first design based on ring oscillators is presented in Fig. 1. This architecture exploits two identical ring oscillators, each having 16 buffers (32 inverters), and operating complementary. The outputs of the buffers in the first ring oscillator are marked as $phase[i]$, while the outputs of the buffers in the second ring oscillator are marked as $phase_c[i]$ ($i \in \{0, 1, ..., 15\}$). In order to achieve complementary operation, first oscillator uses NAND gate, the second uses NOR gate, and the oscillators are triggered with complementary signals EN and $\overline{EN}$. The dimensions of inverters are optimized so that the delay of one buffer is equal to 700 ps for the temperature of 27°C, and the nominal supply voltage of 1.8 V.

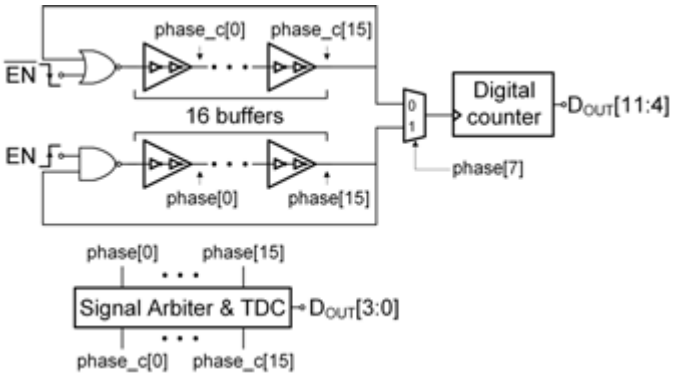


Fig. 1. Ring oscillator based temperature sensor. This architecture uses two complementary ring oscillators, signal arbiter, TDC, and a digital counter. Temperature is represented as the 12-bit binary number D_{OUT}.

A. Operation Principle and Block Description

First, a reference pulse is generated by using a reference clock with the frequency equal to 20 MHz. The pulse duration is equal to the predefined number of reference clock period. The length of the reference pulse, which represents the enable signal EN, is chosen so that good resolution can be achieved (in the range of 0.1°C), avoiding long measurement time. Therefore, a measurement time of 4.5 μs is chosen, which is equivalent to the 90 reference cycles. During this time, rising edges of signals $phase[i]$ and $phase_c[i]$ are counted. This is done alternately, first by counting rising edges of $phase_c[i]$, and then $phase[i]$. In order to achieve this, the 4-bit TDC, and an arbiter circuit are used (Fig. 2). The timing diagram is shown in Fig. 3.

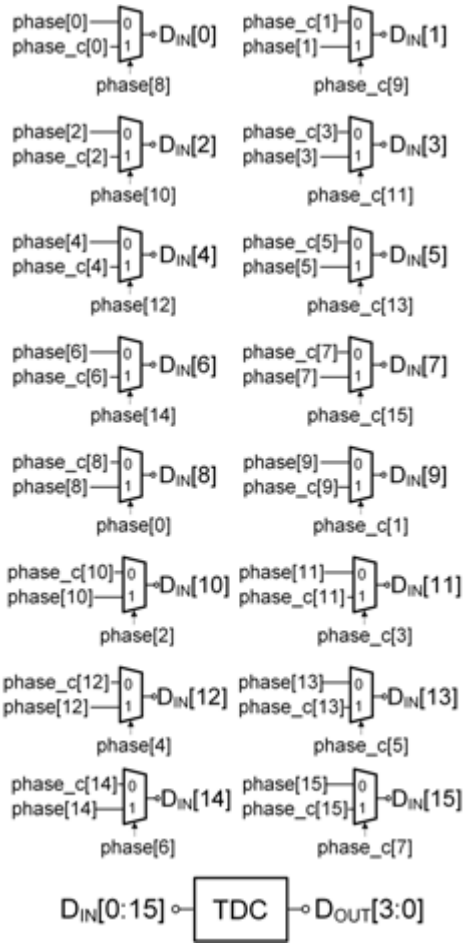


Fig. 2. Arbiter and 4-bit flash TDC. The arbiter is used in order to achieve alternate counting of rising edges coming from the two ring oscillators.

The counting is done by using the 4-bit flash TDC as a fine counter and the 8-bit digital counter as a coarse counter. Additionally, one counting overflow happens. The fine counter consists of 16 DFF-s, corresponding to 4 bits when thermometer code is converted to binary code. They are connected serially, which means that the output of i-th DFF is connected to the input of the (i+1)-th DFF. The input of the first DFF is the logical '1', and the i-th output of the arbiter is connected to the clock pin of the i-th DFF. This is shown in Fig. 4. The output of the system is the 12-bit digital number from which the temperature can be extracted.

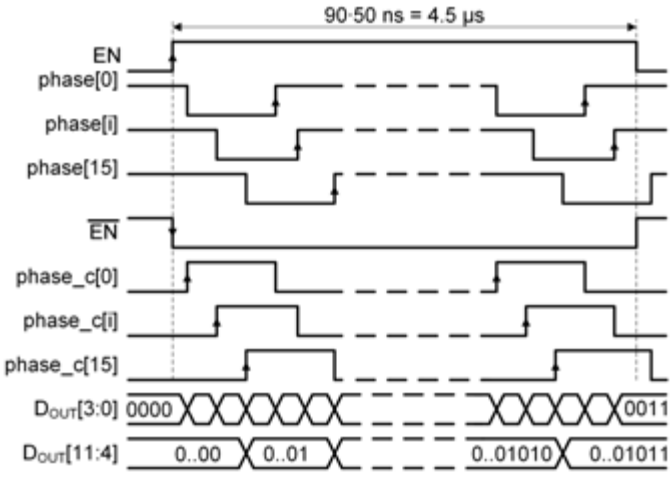


Fig. 3. Simplified timing diagram of sensor operation.

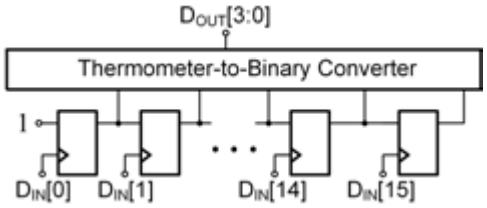


Fig. 4. 4-bit flash TDC needed for fine counting. 16 DFF-s are used to obtain 4-bit fine code, after conversion to binary code. The outputs of the arbiter are applied to the clock pins of serially connected DFF-s.

B. Simulation Results

In order to get the formula for the delay, the temperature dependence of the buffer is simulated on the circuit level. It is concluded that the delay can be approximated by:

$$t_{buffer}[ps] \approx 1.9567 \cdot temp + 647.32, \tag{1}$$

where the temperature is given in Celsius (°C). Using (1), the temperature can be calculated from the digital output word W as follows:

$$temp_{calc} = \frac{1}{1.9567}\left(\frac{N_{ref}t_{ref}}{W + 2^{12} - C} - 647.32\right), \tag{2}$$

where N_{ref} is the number of the reference clock cycles ($N_{ref} = 90$ for this circuit), t_{ref} is the period of the reference clock (in ps), and C is the fitting constant. The optimal value of C is calculated to be 96. The output digital code W with respect to the temperature is shown in Fig. 5.

The error ϵ can be defined as a difference between the temperature calculated by using the proposed circuit and the real value:

$$\epsilon = temp_{calc} - temp_{real}. \tag{3}$$

The error for the temperature range $-40 \sim 125$°C is shown in Fig. 6. It is possible to see that the largest error is +4.9/-0.2°C, without any trimming or calibration.

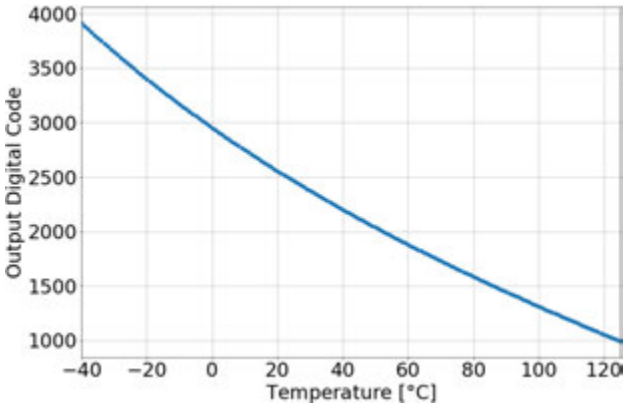


Fig. 5. The output digital code W with respect to the temperature (180-nm technology).

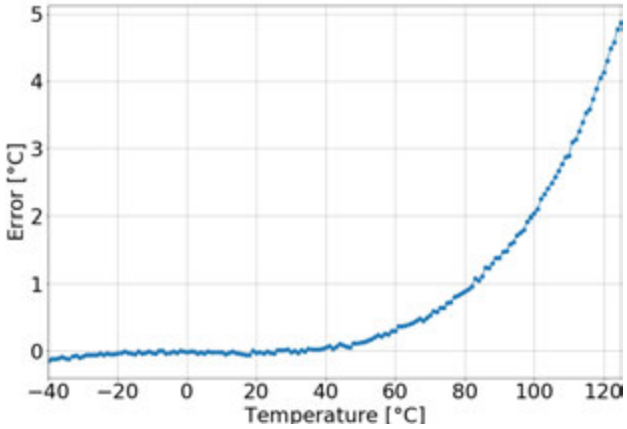


Fig. 6. The accuracy of the proposed sensor within the temperature range $-40 \sim 125$°C (180-nm technology).

III. Complementary Ring Oscillators Based Sensor - Second Design

The second design uses the same principle as the first one, but it is designed in 110-nm. Therefore, the previous analysis and block description apply for this sensor.

A. Simulation Results

The temperature dependence, concluding from the circuit level simulations, is described by:

$$t_{buffer}[ps] \approx 1.58 \cdot temp + 755.2, \tag{4}$$

where the temperature is given in Celsius (°C). The buffers are designed to have larger delay than in the first case. Using (4), the temperature can be calculated from the digital output word W as follows:

$$temp_{calc} = \frac{1}{1.58}\left(\frac{N_{ref}t_{ref}}{W + 2 \cdot 2^{12} - C} - 755.2\right), \tag{5}$$

where N_{ref} is the number of reference clock cycles ($N_{ref} = 160$ for this case), t_{ref} is the period of the reference clock (in ps), and C is the fitting constant ($C = 180$ for this case). The number $2 \cdot 2^{12}$ represents two overflows taking place during

the temperature measurement. The output digital code W with respect to the temperature is shown in Fig. 7.

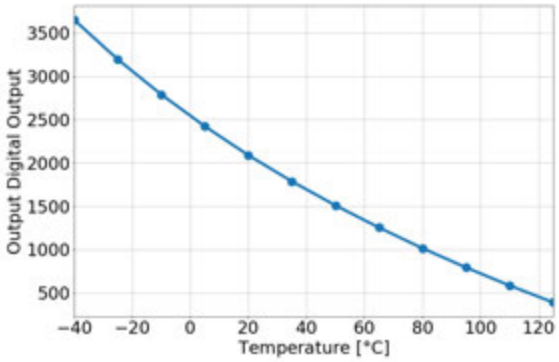


Fig. 7. The output digital code W with respect to the temperature (110-nm technology).

The error, defined as (3) for the temperature range $-40 \sim 125$°C is shown in Fig. 8. It is possible to see that the maximum error is ± 3.8°C, without any trimming or calibration.

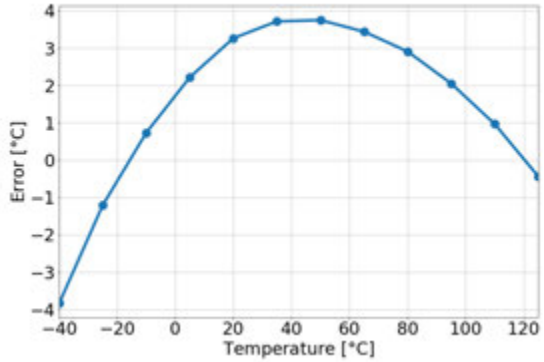


Fig. 8. The accuracy of the proposed sensor within the temperature range $-40 \sim 125$°C (110-nm technology).

B. Temperature Error Comparison

The main difference in the results can be seen by comparing Fig. 8 with Fig. 6. More specifically, the shape of the error curves are different, This is mainly due to the fact that the delay defined by (1) is a better approximation of the delay for the first design than (4) is for the delay in the second case. The linear approximation is not sufficient for the second design and the additional term, namely the quadratic term, should be taken into account during calculation. However, this further complicates the temperature calculation.

IV. Conclusion

Two ring-oscillator based temperature sensors are described and analyzed in this work. Both sensors are all-CMOS sensor based on two ring oscillators operating alternately. Regarding the first design, the error for the typical case is measured to be +4.9/-0.2°C within the temperature range $-40 \sim 125$°C. The second architecture achieves the accuracy of ± 3.8°C within the temperature range $-40 \sim 125$°C for the typical case. It is exepcted that the accuracy can be further improved by adapting the architecture, as well as by using some calibration techniques.

References

[1] K. A. A. Makinwa, "Temperature Sensor Performance Survey," [Online]. Available: http://ei.ewi.tudelft.nl/docs/TSensor_survey.xls

[2] M. A. P. Pertijs, K. A. A. Makinwa, and J. H. Huijsing, "A CMOS Smart Temperature Sensor With a 3σ Inaccuracy of ± 0.1°C From -55°C to 125°C," IEEE J. Solid-State Circuits, vol. 40, no. 12, pp. 2805-2815, 2005.

[3] A. L. Aita, M. A. P. Pertijs, K. A. A. Makinwa, and J. H. Huijsing, "A CMOS smart temperature sensor with a batch-calibrated inaccuracy of ± 0.25°C (3σ) from -70°C to 130°C ," Dig. Tech. Pap. - IEEE Int. Solid-State Circuits Conf., pp. 342-344, 2009.

[4] K. Souri, Y. Chae, and K. A. A. Makinwa, "A CMOS Temperature Sensor With a Voltage-Calibrated Inaccuracy of ± 0.15°C (3σ) From -55°C to 125°C," IEEE J. Solid-State Circuits, vol. 48, no. 1, pp. 292-301, 2013.

[5] M. Chuang, C. Tai, Y. Hsu, A. Roth, and E. Soenen, "A Temperature Sensor with a 3 Sigma Inaccuracy of ± 2°C without Trimming from -50°C to 150°C in a 16 nm FinFET Process," ESSCIRC Conf. 2015 - 41st Eur. Solid-State Circuits Conf., pp. 271-274, 2015.

[6] Z. Tang, Y. Fang, X.-P. Yu, Z. Shi, and N. Tan, "A CMOS Temperature Sensor With Versatile Readout Scheme and High Accuracy for Multi-Sensor Systems," IEEE Trans. Circuits Syst. I Regul. Pap., vol. 65, no. 11, pp. 3821-3829, 2018.

[7] G. Chowdhury and A. Hassibi, "An On-Chip Temperature Sensor With a Self-Discharging Diode in 32-nm SOI CMOS," IEEE Trans. Circuits Syst. II Express Briefs, vol. 59, no. 9, pp. 568-572, 2012.

[8] M. Eberlein and H. Pretl, "A No-Trim, Scaling-Friendly Thermal Sensor in 16 nm FinFET Using Bulk Diodes as Sensing Elements," IEEE Solid-State Circuits Lett., vol. 2, no. 9, pp. 63-66, 2019.

[9] U. Sönmez, F. Sebastiano, and K. A. A. Makinwa, "Compact Thermal-Diffusivity-Based Temperature Sensors in 40-nm CMOS for SoC Thermal Monitoring," IEEE J. Solid-State Circuits, vol. 52, no. 3, pp. 834–843, 2017.

[10] C. Van Vroonhoven, D. D. Aquino, and K. Makinwa, "A ± 0.4°C (3σ) -70 to 200°C Time-Domain Temperature Sensor Based on Heat Diffusion," 2012 IEEE Int. Solid-State Circuits Conf., vol. 10, pp. 204-206, 2012.

[11] C.-H. Weng, C.-K. Wu, and T.-H. Lin, "A CMOS Thermistor-Embedded Continuous-Time Delta-Sigma Temperature Sensor With a Resolution FoM of 0.65 pJ °C^2," IEEE J. Solid-State Circuits, vol. 50, no. 11, pp. 2491-2500, 2015.

[12] H. Xin, M. Andraud, P. Baltus, E. Cantatore, and P. Harpe, "A 174 pW-488.3 nW 1 S/s-100 kS/s All-Dynamic Resistive Temperature Sensor With Speed/Resolution/Resistance Adaptability," IEEE Solid-State Circuits Lett., vol. 1, no. 3, pp. 70-73, 2018.

[13] Y. Lee, W. Choi, T. Kim, S. Song, K. A. A. Makinwa, and Y. Chae, "A 5800 μm^2 Resistor-Based Temperature Sensor With a One-Point Trimmed Inaccuracy of ± 1.2°C (3σ) From -50°C to 105°C in 65-nm CMOS," IEEE Solid-State Circuits Lett., vol. 2, no. 9, pp. 67-70, 2019.

[14] S. Pan and K. A. A. Makinwa, "A Wheatstone Bridge Temperature Sensor with a Resolution FoM of 20 fJK^2," 2019 IEEE International Solid- State Circuits Conference - (ISSCC), San Francisco, CA, USA, 2019, pp. 186–188.

[15] P. Chen, C. Chen, Y. Peng, K. Wang, and Y. Wang, "A Time-Domain SAR Smart Temperature Sensor With Curvature Compensation and a 3σ Inaccuracy of -0.4°C $\sim$ +0.6°C Over a 0°C to 90°C Range," IEEE J. Solid-State Circuits, vol. 45, no. 3, pp. 600-609, 2010.

[16] C. Chen, C. Chen, Y. Lin, and S. You, "An All-Digital Time-Domain Smart Temperature Sensor With a Cost-Efficient Curvature Correction," IEEE Trans. Very Large Scale Integr. Syst., vol. 27, no. 1, pp. 29-36, 2019.

[17] S. Xie and W. T. Ng, "Delay-line Temperature Sensors and VLSI Thermal Management Demonstrated on a 60nm FPGA," 2014 IEEE Int. Symp. Circuits Syst., pp. 2571-2574, 2014.

[18] Y. Lo and Y. Chiu, "A High-Accuracy, High-Resolution, and Low-Cost All-Digital Temperature Sensor Using a Voltage Compensation Ring Oscillator," vol. 16, no. 1, pp. 43-52, 2016.
[19] S. Kim and M. Seok, "A 30.1μm^2, $< \pm$1.1°C-3σ-error, 0.4-to-1.0V temperature sensor based on direct threshold-voltage sensing for on-chip dense thermal monitoring," 2015 IEEE Custom Integrated Circuits Conference (CICC), pp. 4-7, 2015.
[20] T. Anand, K. A. A. Makinwa, and P. K. Hanumolu, "A VCO Based Highly Digital Temperature Sensor," IEEE J. Solid-State Circuits, vol. 51, no. 11, pp. 2651-2663, 2016.
[21] C. Ku and T. Liu, "A Voltage-Scalable Low-Power All-Digital Temperature Sensor for On-Chip Thermal Monitoring," IEEE Trans. Circuits Syst. II Express Briefs, vol. 66, no. 10, pp. 1658-1662, 2019.
[22] T. Someya, A. K. M. M. Islam, T. Sakurai and M. Takamiya, "An 11-nW CMOS Temperature-to-Digital Converter Utilizing Sub-Threshold Current at Sub-Thermal Drain Voltage," 2019 IEEE Journal of Solid-State Circuits, vol. 54, no. 3, pp. 613–622, March 2019.
[23] Z. Tang et al., "A 1770μm^2 Leakage-Based Digital Temperature Sensor With Supply Sensitivity Suppression," vol. 55, no. 3, pp. 781-793, 2020.
[24] W. Song, J. Lee, N. Cho, and J. Burm, "An Ultralow Power Time-Domain Temperature Sensor With Time-Domain Delta Sigma TDC," vol. 64, no. 10, pp. 1117-1121, 2017.
[25] G. W. Roberts and M. Ali-bakhshian, "A Brief Introduction to Time-to-Digital and Digital-to-Time Converters," IEEE Trans. Circuits Syst. II Express Briefs, vol. 57, no. 3, pp. 153-157, 2010.
[26] P. Chen, C. Chen, C. Tsai, and W. Lu, "A Time-to-Digital-Converter-Based CMOS Smart Temperature Sensor," IEEE J. Solid-State Circuits, vol. 40, no. 8, pp. 1642-1648, 2005.

A Circuit for Identifying Oldest Ready Instructions in Reservation Stations

Dejan Spasov
Faculty of Computer Science and Engineering,
Skopje, Republic of North Macedonia
dejan.spasov@finki.ukim.mk

Abstract **– Microprocessors use reservation stations to host instructions that are waiting to be sent to the execution units. Reservation stations may include selection logic configured to select oldest ready instructions for execution. Conventional selection logic uses age matrices to track relative age among instructions. In this paper, we propose an alternative circuit for tracking the relative age among instructions in reservation stations. The proposed selection logic uses wrap bits and reorder buffer indexes to track relative age among instructions. We compare the proposed selection logic with the conventional selection logic and discuss advantages.**

Keywords – superscalar processor; scheduler; issue queue; reservation stations; oldest-ready instructions.

I. INTRODUCTION

Reservation stations are buffering stations between the front end and the back end of the microprocessors. They were introduced in 1967 [1] as a means to host instructions that are waiting for source operands to become ready. Among the hosted ready-for-issue instructions, reservation stations are configured to select the oldest instruction to be issued to the execution units. Selecting oldest-ready instruction is a time-critical operation that affects the clock cycle of the microprocessor.

Fig. 1 shows the microarchitecture of a superscalar core. The in-order front end consists of a fetch unit, a decode unit, a renaming unit, and a dispatch unit [2]. The back end of the core may include, in parallel, one or more execution slices [3].

The fetch unit is configured to fetch instructions from memory or L1 cache. The decode unit is configured to decode fetched instructions into micro-operations, which may be executed in the back end. The core may be configured to execute speculatively fetched instructions. Instructions may be executed out of program order. The core may include a plurality of physical registers organized in one or more physical register files (PRF). Physical registers of the core store speculative results and architecturally visible results. The core may employ stand-alone addressing scheme for the physical registers, or physical registers may use the addressing scheme of the reorder buffer entries. For each instruction that may write

Funded by a project from the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University.

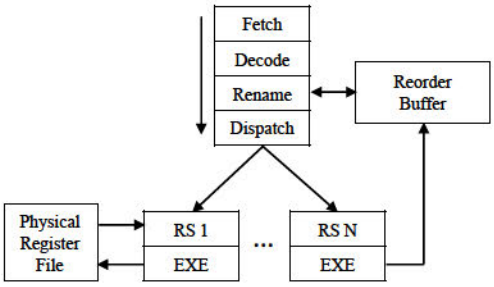


Figure 1. Microarchitecture of superscalar microprocessor

to destination operand, the renaming unit may allocate a physical register, where the speculative result may be stored. Additionally, the renaming unit is configured to rename source operands of instructions consumers of a result to the physical register allocated to the instruction producer of the result. A source operand of an instruction is renamed to the physical register most recently allocated to instruction with destination operand equal to the source operand. The dispatch unit may be configured to send the renamed micro-operations to the one or more reservation stations.

The back end of the core may include one or more reservation stations RS 1 to RS N, where each reservation station is coupled to one or more execution units (EXE). Instructions in the reservation stations may be waiting for source operands to become ready, designated execution unit to become available, through-memory data dependencies to be resolved, etc. Execution units (EXE) may include any number and type of execution units, e.g. integer unit, floating point unit, load/store unit, branch unit, etc., configured to execute instructions.

The core may include a reorder buffer to support out-of-order execution. For each instruction the reorder buffer allocates an entry. Index of the allocated entry is assigned to the instruction and stored in the reservation stations together with the instruction. Reorder buffer entries are organized as circular buffer with a tail pointer and a head pointer. Instructions enter at one end of the buffer and retire at the other end. Thus, instructions can be executed out of program order but retired in program order.

The core may include other components and interfaces not shown on fig. 1 to simplify the microarchitecture. The core may support multiple instruction issue, switching

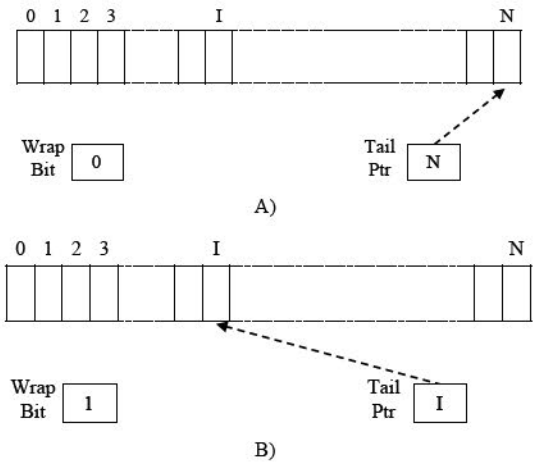


Figure 2. Principle of operation of the reorder buffer

between in-order and out-of-order execution, and simultaneous multithreading, where instructions from up to two threads may simultaneously be processed.

In this paper we propose a circuitry for selecting oldest-ready instructions from the reservation stations. The circuitry is configured to use wrap bits and reorder buffer indexes to determine relative age of instructions. The rest of the paper is organized as follows: principle of operation of the reorder buffer is explained in Section II; conventional select logic is described in Section III; proposed selection logic is described in detail in Section IV; related work is outlined in Section V; experimental results are given in Section VI; and section VII concludes the paper.

II. Reorder Buffer

Figs. 2 A and 2 B demonstrate the principle of operation of the reorder buffer. The reorder buffer is a circular buffer with N+1 entries. Entries of the reorder buffer, denoted with 0 to N, are configured to host in-flight instructions in the microprocessor. Reorder buffer entry with index 0 is considered successor of the entry with index N. Instructions in the reorder buffer are considered to be ordered in program order. Thus, the reorder buffer supports out-of-order execution and in-order retirement of the instructions. For each instruction entering the core (fig. 1), the reorder buffer allocates an entry. Index of the allocated entry is assigned to the instruction when it enters the reservation stations.

The reorder buffer includes a tail pointer to point to the next-in-line for allocation entry. For an instruction entering the core, the reorder buffer allocates the next-in-line for allocation entry and increments the tail pointer. The reorder buffer may include a head pointer (not shown) to point to the oldest in-flight instruction. An instruction may be retired only if the head pointer points to the reorder buffer entry allocated to the instruction.

The reorder buffer may be configured to maintain a wrap bit with the tail pointer. The wrap bit is toggled each time the tail pointer wraps-around from the entry N to the entry 0. Fig. 2 A) shows the wrap bit holding logical 0. After the entry with index N is allocated, the wrap bit is

	0	1	2	3	4	5	6
0	X	0	0	0	1	0	0
1	1	X	0	0	1	0	0
2	1	1	X	1	1	1	1
3	1	1	0	X	1	0	1
4	0	0	0	0	X	0	0
5	1	1	0	1	1	X	1
6	1	1	0	0	1	0	X

Figure 3. Age Matrix

toggled and holds logical 1 (fig. 2 B). The wrap bit maintains logical 1 for the duration of the entire allocation cycle from 0 to N. The reorder buffer may be configured to maintain another wrap bit with the head pointer, which is toggled each time the head pointer wraps-around.

One application of the wrap bits may be to determine whether the reorder buffer is full or empty [4]. If the tail pointer and the head pointer point to the same entry and respective wrap bits are equal, the reorder buffer is empty. If the tail pointer and the head pointer point to the same entry and respective wrap bits are not equal, the reorder buffer is full.

Another application of the wrap bits may be to determine relative age between two instructions [5]. The wrap bit is assigned to each instruction when the reorder buffer index is assigned to the instruction. If wrap bits of two instructions are equal, then instruction with smaller index is older. If wrap bits of two instructions are different, then instruction with larger index is older.

III. Age Matrix

A reservation station with M entries may host up to M instructions. The reservation station may include a selection logic configured to issue the oldest ready instruction. The selection logic may include an age matrix of size MxM to maintain the relative age among instructions. Fig. 3 shows a 6x6 age matrix for a reservation station with M=6 entries. Each row and column in the age matrix corresponds to an instruction in reservation stations. For example, the row with index 3 and the column with index 3 correspond to the relative age

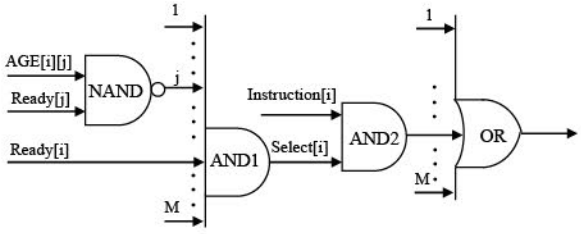


Figure 4. Conventional Selection Logic

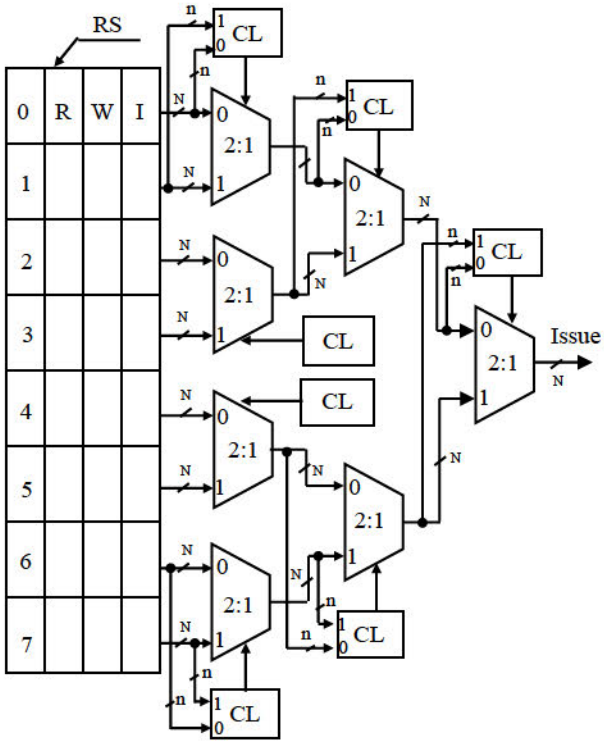


Figure 5. Proposed Selection Logic

of an instruction stored in an entry in the reservation station that is denoted with the index 3.

The age matrix is an array of one-bit memory cells, where a memory cell A[i][j] holds the relative age between the instruction in the entry i and the instruction in the entry j. For example, A[3][5]=0 denotes that instruction in the entry 3 is older than the instruction in the entry 5; A[5][3]=1 denotes that instruction in the entry 5 is younger than the instruction in the entry 5. Memory cells on the main diagonal are not used, i.e. A[i][i]=X. When an instruction is stored in the reservation station, a vector of logical 1s is written in the corresponding row, and a vector of logical 0s is written in the corresponding column. For example, row 2 is a row of logical 1s, and column 2 is a column of logical 0s. This indicated that the instruction in the entry 2 in the reservation station is the youngest instruction. The oldest instruction is the instruction in the entry 4, as indicated by the all-0 row 4 and all-1 column 4.

Fig. 4 shows the conventional selection logic that utilizes age matrix for tracking the relative age. For each instruction in the reservation station, the conventional selection logic maintains a ready bit, denoted with Ready[0] to Ready[M-1]. Ready bits indicate if the corresponding instruction is ready to be issued to the execution units, i.e. source operands are ready, no data dependencies through memory exists, designated execution unit is available, etc. The selection logic includes M^2-M NAND gates. Each NAND gate is coupled to a ready bit Ready[j] and to a memory cell A[i][j]. If the output of the NAND1 gate is 0, instruction in the entry j is older than the instruction in the entry i. M-1 NAND gates are coupled to the memory cells of the row i of the age

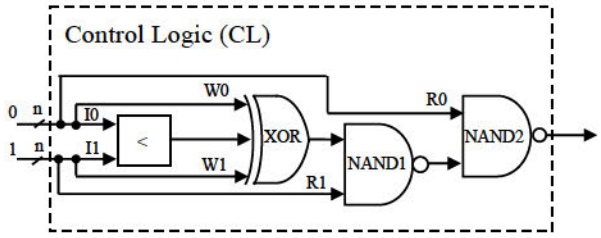


Figure 6. Control logic for the muxes

matrix to determine if the instruction in the entry i is the oldest. If all M-1 NAND gates output logical 1, and the instruction i is ready (Ready[i]=1), the AND1 gate outputs Select[i]=1, thus indicating that the instruction i is the oldest ready instruction. If Select[i]=1, AND2 and OR gates are configured to mux-out the Instruction[i].

Age matrix of size MxM implies that the hardware complexity of the conventional selection logic is quadratic, i.e. proportional to M^2. Moreover, if we want to insert two or more instructions to the reservation station, row-vectors and column-vectors that need to be inserted in the age matrix are not all-1s and all-0s vectors. Row vectors and column vectors that need to be inserted need to be computed with a complex serial circuit.

IV. Proposed Selection Logic

Fig. 5 shows a reservation station (RS) with M=8 entries coupled to the proposed selection logic. Reservation station entries are indexed with 0 to 7 and configured to host instructions. Instructions hosted in the reservation station entries comprise ready-to-issue bit R, wrap bit W, and index I. The R bit indicates if the instruction may be issued to the execution units. The wrap bit W and the index I are assigned by the reorder buffer allocation circuit. A selection logic is coupled to the reservation station (RS) and configured to issue the oldest ready instruction. The selection logic is configured to group instructions in pairs and to select older ready instruction from each pair. The selection logic may be configured to keep pairing and selecting instructions until one instruction remains.

In general, for an M-entry reservation station, the selection logic includes M 2-to-1 multiplexers coupled as a binary tree. Multiplexers are internal vertices of the binary tree, while reservation station entries are leaves of the tree. Each multiplexer may be coupled to receive two instructions and to forward older ready instruction. The root of the binary tree issues the oldest ready instruction from the reservation station. In general, an instruction may be considered N bits wide. The size N depends on the architecture of the reservation station. For example, some reservation stations may include source-operand values, while other may include physical register where the value of the source operand is kept. Thus, bus lines coupled to the inputs of the multiplexers are also N bits wide.

A control logic (CL) is coupled to the selection input on each mux. The control logic is configured to identify the older ready instruction and to direct the mux to forward the older ready instruction. The control logic is

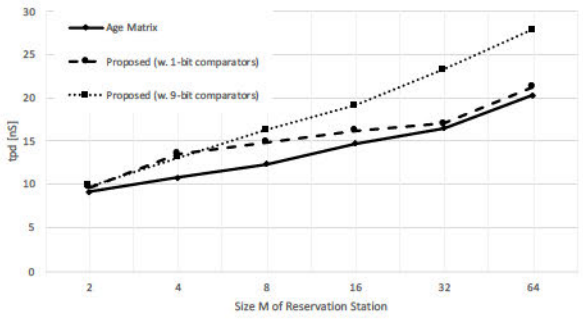


Figure 7. Delay comparison for various sizes of the reservation station

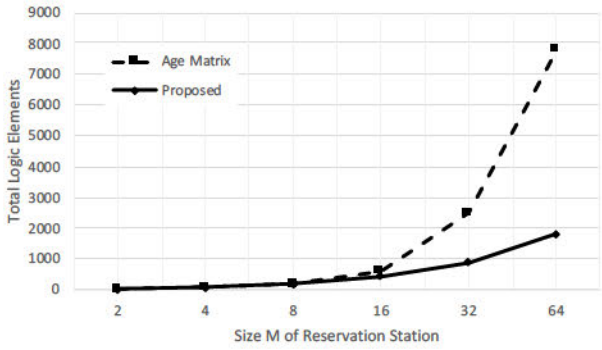


Figure 8. Number of logic elements for various sizes of the reservation station

coupled to receive n bits (n<N) from each of the instructions that are coupled to the mux. These n bits consist of the R bit, the W bit and (part of) the index I.

Fig. 5, shows the control logic (CL) that is coupled to each mux. An instruction In0 is provided to the input denoted with 0, and an instruction In1 is provided to the input denoted with 1. On each input, the control logic receives R bit, W bit, and (part of) the index I. Both instructions In0 and In1 are provided to a mux that uses the output of the control logic as selection control. If In0 is not ready to be issued (R0=0), the NAND2 gate outputs 1, which directs the mux to forward In1. If In0 is ready to be issued (R0=1), but In2 is not ready to be issued (R1=0) the NAND2 gate outputs 0 and the mux forwards the In0.

If both In0 and In1 are ready to be issued (R0=R1=1), the output is determined by the XOR gate and the comparator. The comparator is configured to output logical 0 if I0<I1, or logical 1 otherwise. If W0=W1 and I0<I1, or if W0≠W1 and I0>I1, the control logic directs the mux to forward In0 as older ready. If W0=W1 and I0>I1, or if W0≠W1 and I0<I1, the control logic directs the mux to forward In1 as older ready.

For a reservation station with M entries, the proposed selection logic requires M multiplexers and M control logic circuits. Therefore, we can conclude that the hardware complexity of the proposed selection logic is linear with respect to the number of reservation station entries.

V. Related Work

Reservation stations can be implemented as shifting FIFO-like structures, where instructions enter at one end of a queue, while the search for the oldest ready instruction starts at the other end. The position of an instruction with respect to the exiting end of the queue is indicator of the relative age of the instruction. Shifting reservation stations may be implemented with collapsing buffers. When an instruction is issued, collapsing buffers move younger instructions to the exiting end of the queue; thus, making space for new instructions to enter the bottom of the queue. An example of such a reservation station may be found in the Alpha 21264 microprocessor [6]. Collapsing buffers require a lot of hardware to squeeze empty slots in the reservation stations. Therefore, some microarchitectures organize the reservation stations as FIFO structures without the collapsing buffers. Issued instruction leaves a hole in the buffer that is destroyed when the instruction reaches the exiting end of the buffer. Pentium 4 processors are implemented as such FIFO structures [7], [8]. Leaving a hole when an instruction is issued degrades the performance of the processor, which can be compensated with a larger buffer.

Reservation stations can be implemented as non-shifting structures, where instructions are kept in one entry until their issuance to the execution units. The MIPS R10000 superscalar microprocessor is one of the first microprocessors to use non-shifting reservation stations [9]. However, in R10000 instructions are selected randomly, i.e. without selection logic. Non-shifting reservations with age matrices are used in most of today's processors [10], [11]. Some microarchitectures use hybrid schemes: a combination of non-shifting and shifting reservation stations [12].

VI. Experimental Results

We have designed a Verilog description of the conventional selection logic and the proposed selection logic and tested them using the Quartus software targeting Cyclone FPGA family.

Fig. 7 compares propagation delays of the proposed selection control logic and the conventional selection logic with respect to the number of entries of the reservation station. We simulate two configurations of the proposed selection logic. In the first configuration we use 9-bit comparators in the control logic (CL). In the second configuration we use 2-bit comparators in the control logic (CL).

Modern microprocessors include reorder buffers with more that 256 entries, but less than 512. Hence, 9 bits may be used to represent the address space of the modern reorder buffers. This explains the use of 9-bit comparators in the first configuration of the proposed circuit.

On the other hand, it is not necessary to find the oldest instruction in the reservation stations. We can issue arbitrary instruction at the cost of decreased instruction-per-cycle rate. Hence, we can use sub-optimal selection

logic to improve the propagation delay. In the sub-optimal selection logic, we compare only parts of the indexes I. For example, we can compare the k least important bits of the indexes I. Thus, we can use smaller k-bit comparators to improve the propagation delay. In an extreme situation, we can use k=1 and compare only the least important bits of the indexes I. Thus, we can use 1-bit comparators in the control logic (CL). This explains the use of 1-bit comparators in the second configuration of the proposed circuit. The propagation delay of the first configuration corresponds to maximal propagation delay. The propagation delay of the second configuration corresponds to minimal propagation delay. The second configuration is sub-optimal selection logic.

From the fig. 7, it can be observed that the conventional selection logic shows better performance in terms of the propagation delay than the proposed selection logic. This is somewhat expectable result, since the critical path in the proposed selection logic is of depth $\sim\log_2(M)$.

Fig. 8 compares total number of logic elements synthetized for the proposed selection control logic and for the conventional selection logic with respect to the number of entries of the reservation station. Here the total number of elements refers to the total number of combinatorial elements plus the total num number of registers. From fig. 8 it can be observed that the total number of elements in the conventional selection logic grows quadratically with respect to the number of entries in the reservation station. This is expected due to the quadratic hardware complexity of the age matrix. A floorplan of the conventional selection logic is shown in [11], where it can be observed that the age matrix occupies half of the selection logic area. Fig. 8 shows that the total number of elements in the proposed selection logic grows linearly with respect to the number of entries in the reservation station.

VII. Conlusion

We have proposed a circuit designed to issue the oldest ready instruction from the reservation stations in a superscalar microprocessor. Reservation stations in everyday microprocessors track relative age with internally built tracking mechanism. Our design tracks relative age with a mechanism that is external to the reservation stations. Therefore, our design simplifies the hardware complexity of the age-tracking mechanisms. Improved hardware complexity implies improved die area and less power consumption. Thus, although simulation results show weaker propagation delay, we believe that proposed selection logic may improve operation of modern microprocessors.

References

[1] R. M. Tomasulo, "An Efficient Algorithm for Exploiting Multiple Arithmetic Units," IBM Journal of Research and Development, Issue 11, pp. 25–33, 1967.

[2] J.E. Smith, and G.S. Sohi, "The microarchitecture of superscalar processors," Proceedings of the IEEE, Vol. 83, Issue 12, pp. 1609 – 1624, 1995.

[3] S. K. Sadasivam, B. W. Thompto, R. Kalla, W. J. Starke, "IBM Power9 Processor Architecture," IEEE Micro, Vol. 37, Iss. 2, pp. 40-51, 2017.

[4] D. B. Papworth, et. all, "Entry allocation in a circular buffer using wrap bits indicating whether a queue of the circular buffer has been traversed," US Patent #5584038, 1996.

[5] D. B. Papworth, et. all, "Entry allocation in a circular buffer using wrap bits indicating whether a queue of the circular buffer has been traversed," US Patent #5584037, 1995.

[6] R. Kessler, "The Alpha 21264 microprocessor," IEEE Micro, vol. 19, no. 2, pp. 24–36, 1999.

[7] G. Hinton, D. Sager, M. Upton, D. Boggs, D. Carmean, A. Kyker, P. Roussel, "The Microarchitecture of the Pentium® 4 Processor," Intel Technology Journal, 2001.

[8] J. Sokes, "Inside the machine: an illustrated introduction to microprocessors and computer architecture," No Starch Press, 2007.

[9] K. C. Yeager, "The MIPS R10000 superscalar microprocessor," IEEE Micro, vol. 16, issue 2, pp. 28–41, 1996.

[10] P. G. Sassone, J. Rupley II, E. Brekelbaum, G. H. Loh, and B. Black, "Matrix Scheduler Reloaded," In Proceedings of the 34th Annual International Symposium on Computer Architecture, pp. 335–346, 2007.

[11] M. Golden, S. Arekapudi, and J. Vinh, "40-Entry Unified Out-of-Order Scheduler and Integer Execution Unit for the AMD Bulldozer x86-64 Core," In 2011 IEEE International Solid-State Circuits Conference, Digest of Technical Papers, pp. 80–82, 2011.

[12] B. Sinharoy, et al. "IBM POWER8 processor core microarchitecture," IBM Journal of Research and Development, Vol. 59, Iss. 1, pp. 1-21, 2015.

An Improvement in the Convergence of Superscalar Processors

Dejan Spasov
Faculty of Computer Science and Engineering, Skopje, Republic of North Macedonia
dejan.spasov@finki.ukim.mk

***Abstract* – Modern processors include reservation stations to host instructions that are waiting to be sent to the execution units. Responsive to exception event, instructions in reservation stations that are younger than the instruction executed with exception need to be flushed. In this paper, we propose a mechanism for flushing instructions from reservation stations. The proposed mechanism determines relative age between instructions in reservation stations and instruction executed with exception by comparing reorder buffer indexes assigned to the instructions in reservation stations with the tail pointer of the reorder buffer and with the reorder buffer index of the instruction executed with exception. Instructions younger than the instruction executed with exception are flushed from the reservation stations.**

Keywords – superscalar processor; branch miss-prediction; issue queue; exception.

I. INTRODUCTION

Instructions in modern microprocessors arrive at the execution units through a pipeline made of several stages. In order to keep the pipeline full, microprocessors need to speculate and to insert in the pipeline instructions that are speculated to represent the program order. In the case of miss-speculation, speculatively inserted instructions need to be flushed from the microprocessor and instructions from the correct program order need to be fetched and inserted.

Fig. 1 shows a microarchitecture of a superscalar core. The core may include fetch unit, decode unit, renaming unit, reorder buffer, one or more physical register files, dispatch unit, reservation stations, and execution units [1]. The core may include other components and interfaces not shown on fig. 1. The core may support multiple instruction issue, in-order and out-of-order execution, and simultaneous multithreading, where instructions from up to two threads may simultaneously be processed.

The fetch unit is configured to fetch instructions from memory or L1 cache. Instructions may be from any instruction set architecture, e.g. PowerPC™, ARM™, SPARC™, x86™, etc. The fetch unit may include a branch predictor to predict the result of the execution of branch instructions. Responsive to a branch prediction, the fetch unit may start fetching instructions from the predicted branch. The decode unit is configured receive

Funded by a project from the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University.

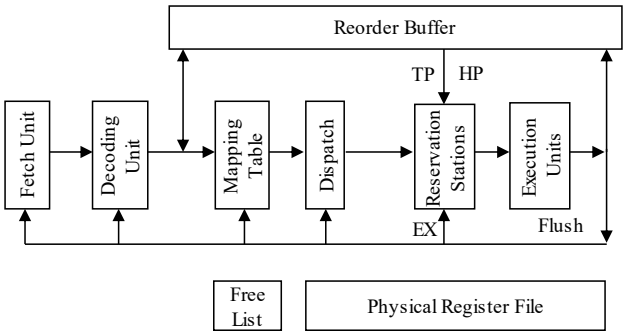


Figure 1. Microarchitecture of superscalar microprocessor

instructions from the fetch unit and to output decoded instructions or micro-operations, which may be executed in the execution units. The free list is a list of physical registers that may be allocated to instructions with destination operands. For each instruction with destination operand, the free list is configured to allocate a physical register. The mapping table is configured to rename (map) source operands of instructions consumers of a result to the physical register allocated to the instruction producer of the result. A source operand of an instruction is renamed to the physical register most recently allocated to an instruction with destination operand equal to the source operand.

The reorder buffer maintains in-order retirement of the instructions. For each instruction the reorder buffer allocates an entry. Reorder buffer entries are organized as circular buffer with a tail pointer (TP) and a head pointer (HP). Instructions enter at one end of the buffer and retire at the other end. Thus, instructions can be executed out of program order but retired in program order. The core may include a plurality of physical registers organized in one or more physical register files (PRF). Physical registers of the core store speculative results and architecturally visible results. The core may employ stand-alone addressing scheme for the physical registers, or physical registers may use the address scheme of the reorder buffer entries.

Execution units (EU) may include any number and type of execution units, e.g. integer unit, floating point unit, load/store unit, branch unit, etc., configured to execute instructions. Instructions may be executed in-order or out-of-order. One or more reservation stations (RS) may be included in each slice to host instructions

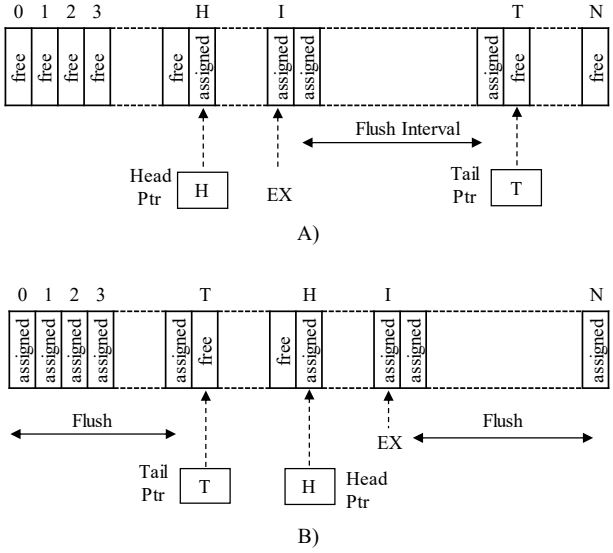


Figure 2. Principle of operation of the reorder buffer

waiting to be issued to the execution units (EU). The dispatch unit is configured to dispatch instructions to the reservation stations.

An instruction may be executed with exception. The core handles the exception event by stalling the pipeline, flushing the miss-speculated instructions, and restoring to the state prior to the exception. According to [2], 49% of all fetched instructions are flushed. The term convergence of the microprocessors refers to the process of flushing miss-speculated instructions and fetching instructions from the correct program order. During the stall, physical registers allocated to miss-speculated instructions are returned to the free list; the mapping table is restored to the state prior to the exception (restoration of the mapping table may be done with checkpoints [3] or with a walk on the reorder buffer); then, the reorder buffer releases all entries allocated to miss-speculated instructions. However, the pipeline remains stalled until reservation stations flush all instructions younger than the instruction executed with exception. Therefore, it is important for microprocessors reservation stations to be flushed as soon as possible.

In this paper we propose a circuitry for selectively flushing instructions from reservation stations. Responsive to exception event, the circuitry immediately flushes from reservation stations miss-speculated instructions. Execution units are coupled to the reservation stations to provide the reorder buffer index EX of the instruction executed with exception. The reorder buffer is coupled to the reservation stations to provide the tail pointer TP or the head pointer HP. The reservation stations are configured to compare reorder buffer indexes of the hosted instructions with the index EX and the index TP (or HP) and to determine relative age of each of the hosted instructions with the instruction executed with exception. Younger instructions are instantaneously flushed from reservation stations. Older instructions are left to be executed.

The remainder of the paper is organized as follows: principle of operation of the reorder buffer is given in Section II, a mechanism for flushing instructions from reservation stations is proposed in Section III, prior work is discussed in Section IV, experimental results that support instantaneous flush of the reservation stations are given in Section V. Section VI concludes the paper.

II. THE REORDER BUFFER

Principle of operation of the reorder buffer is demonstrated on fig. 2. The reorder buffer is a circular buffer with entries denoted with indexes 0 to N, where the entry with index 0 is considered successor of the entry with index N. For each instruction entering the reservation stations in program order, the reorder buffer allocates an entry and returns (assigns) the index of the allocated entry. Instructions with the assigned indexes are stored in the reservation stations.

An entry may be allocated or free for allocation. The reorder buffer includes a head pointer and a tail pointer to maintain the beginning and the end of the consecutive interval of allocated entries. The head pointer points to the entry allocated to the oldest in-flight instruction. The tail pointer points to the next-in-line for allocation entry. The tail pointer is incremented with each allocation in wrapped around increasing fashion. The head pointer is incremented with retirement of the instructions also in wrapped around increasing fashion. The core is stalled when the reorder buffer is full.

The head pointer or the tail pointer may be used as reference points in determining the relative age between instructions in the reservation stations, i.e. the closer, in wrapped around incrementing order, the assigned index of an instruction to the tail pointer is, the younger the instruction is.

An instruction, e.g. with reorder buffer index I, may be executed with exception. Reservation stations are configured flush instructions that are younger, in program order, than the instruction executed with exception. Reorder buffer indexes of instructions that are younger than the instruction executed with exception are in the increasing interval from the index I to the index T, which is stored in the tail pointer. Reservation stations may be configured to flush instructions with reorder buffer indexes in the increasing interval from the index I to the index T. If the index I is larger than the index T (Fig. 2 B), the increasing interval wraps around from the index N to the index 0. Since reorder buffer entries with indexes in the increasing interval from the tail pointer to the head pointer are not associated with in-flight instructions, the flush interval may be defined from the index I to the index H, which is stored in the head pointer.

Reservation stations may be configured to compare the index I with the tail pointer (the index T). If $I<T$ (fig. 2 A), reservation stations may be configured to flush instructions with reorder buffer indexes that are larger than the index I and smaller than the index T. If $I \geqslant T$ (fig. 2 B), reservation stations may be configured to flush instructions with reorder buffer indexes that are larger than the index I or smaller than the index T.

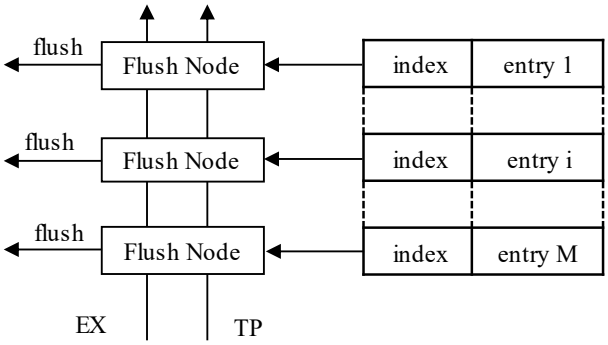


Figure 3. Reservation station

III. THE RESERVATION STATIONS

Fig. 3 shows a reservation station with the circuitry for instantaneous flushing instructions from the reservation station. The reservation station includes m entries, denoted with entry 1 to entry M. Each entry is configured to store one instruction. The reservation station includes flush nodes coupled to the entries. Each flush node is configured to flush the instruction hosted in the coupled entry. Each entry includes index field, where the reorder buffer index assigned to the hosted instruction may be stored.

The reorder buffer and the execution units may be coupled to the flush nodes. Responsive to exception event that requires selectively flushing instructions from the reservation station execution units may provide the reorder buffer index EX of the instruction executed with exception and the reorder buffer may provide reorder the buffer index TP stored in the tail pointer (or the reorder buffer index HP stored in the head pointer). Flush nodes are configured to flush entries with values in the Index field in the increasing interval from the index EX to the index TP (or HP).

Flush nodes are configured to compare the index EX with the index TP. If EX < TP, each flush node is configured to compare the value in the Index field with the index EX and with the index TP, and to assert flush signal for the entry i if the value in the Index field is larger than the index EX and smaller than the index TP. If EX $\geqslant$ TP, each flush node is configured to compare the value in the Index field with the index EX and with the index TP, and to assert flush signal for the entry i if the value in the Index field is larger than the index EX or smaller than the index TP.

Fig. 4, shows a flush node. The flush node includes comparators and combinatorial logic. A reservation station entry, e.g. entry i, is coupled to the flush node to provide the value stored in the Index field. The flush node is configured to receive the index EX and the index TP and to output a flush signal, which may flush the instruction hosted in the reservation station entry i. Comparators are configured to output logical 1 if Index>EX, EX < TP, and Index<TP, respectively. And gates are configured to output logical 1 if the Index is in a flush interval. Thus, AND1 outputs 1 if the Index is in the right flush interval

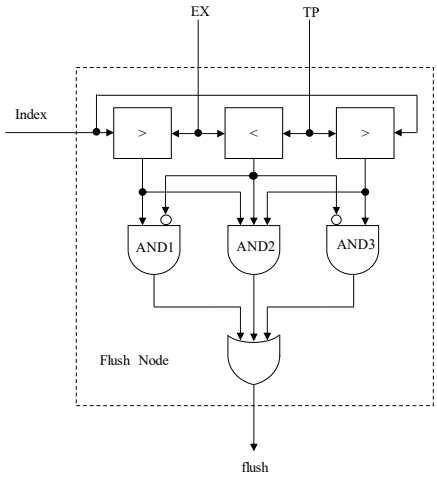


Figure 4. Flush node

on fig. 2 b; AND2 outputs 1 if the Index is in the flush interval on fig. 2 a; AND3 outputs 1 if the Index is in the left flush interval on fig. 2 b. The OR gate outputs logical 1 if the Index is in any of the flush intervals on fig. 2.

For a reservation station with n entries, the proposed circuit requires 2n+1 magnitude comparators and 4n logic gates. In addition, circuits with fan-out of n may be coupled the flush nodes to provide EX and TP. Large fan-out circuits may be considered as constraint for single-cycle implementation of the flush logic.

IV. RELATED WORK

Reservation stations can be implemented as shifting FIFO-like structures, where instructions enter at one end of a queue, while the search for the oldest ready instruction starts at the other end. The Alpha 21264 microprocessors are such an example, where reservation stations are implemented with collapsing buffers [4]. The position of an instruction with respect to the exiting end of the queue is indicator of the relative age of the instruction. This approach simplifies flushing due to misprediction. However, a constant shift of instructions is needed in order to make space for new instructions. Constantly shifting instructions is expensive operation in terms of gates, die area, wiring, and power consumption.

Due to the complexity of the collapsing buffers, most modern reservation stations are implemented as non-shifting structures, where instructions are kept in one entry until their issuance to the execution units. Flushing instructions from the reservation stations may be implemented by assigning branch masks to the hosted instructions. Each bit in the branch mask indicates dependence with prior branch instruction that may initiate flush. The MIPS R10000 superscalar microprocessor, for example, uses four-bit branch masks [5]. Hence, the MIPS R10K microprocessor can support at most four branch instructions in the reservation stations. A fifth instruction will cause a stall. Therefore, branch masks are complex design solution that does not scale well with the size of the branch mask. Keeping the branch mask as small as possible introduces stalls at runtime. Moreover, the

TABLE I. SIMULATED PROCESSOR

Processor Frequency	1 GHz
Instruction Set Architecture	x86
Reservation Station Size	64 entries
Reorder Buffer Size	192 entries
Branch Predictor	Tournament
Main Memory	512 MB, DDR3
L1 I-cache	16 kB, assoc. 2
L1 D-cache	64 kB, assoc. 2
L2 Cache	256 kB, assoc. 8

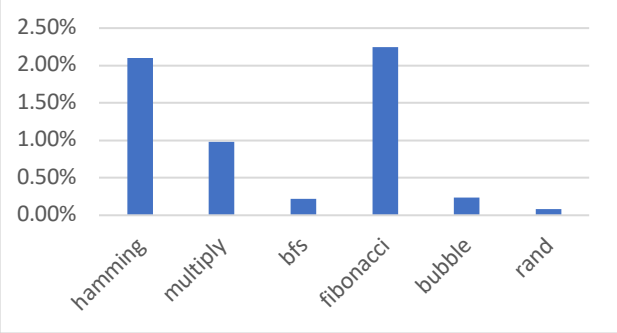


Figure 5. Improvement in execution time

number and the type of instructions (e.g. branches) that may initiate flush operation is predetermined at the design stage of the processor.

POWER8 and older microprocessors use a variant of the branch masks, which are called flush masks [6]. A core of the POWER8 microprocessor supports 224 in-flight instructions divided in 28 groups, where each group consists of up to 8 instructions. Branch instructions are always the last instructions in the group. For each instruction in the reservation stations a 28-bit flush mask is assigned. Flush masks do not introduce stalls at runtime; however, they may leave numerous holes in the 224 slots for instructions due to the grouping of instructions. To alleviate this, POWER8 microprocessors support two branches per group, and partial flush of a group. Partial flush further complicates the design of the reservation stations, and the processor in general. Therefore, starting from POWER9 [7], flush masks are not used for flushing instructions from reservation stations.

Another way to flush instructions from the reservation stations is to wait for the head pointer of the reorder buffer to point to the instruction executed with exception. Then, the reservation stations may be flushed, and normal operation may resume. We believe that this approach is favored among the producers of microprocessors. The reason for this is: contemporary predictor circuits can predict the correct program order with accuracy above 99%; exception events are becoming rare; thus, waiting for an instruction executed with exception to retire does not affect much the instructions-per-cycle (IPC) rate.

V. EXPERIMENTAL RESULTS

We use GEM5 simulator [8] in our experimental work. Table I summarizes the parameters of the simulated processor. We have used a reservation station with 64 entries, which is comparable with reservation stations in the commercial processors. The remaining CPU parameters are selected randomly. Our goal was to show improvement. Exact magnitude of the improvement would be different for the commercial microprocessors. The GEM5 simulator was modified so that reservation stations are immediately flushed following a branch misprediction. Several programs (Hamming codes, matrix multiplication, breadth first search, Fibonacci series, bubble sort, and random numbers) were compiled with gcc 8.3.0 compiler and run on the simulated processor. Results on fig. 5 show reduction in clock cycles between 2% and 0.2% when reservation stations perform instantaneous flush following branch mispredictions.

VI. CONLUSION

We have proposed a circuit designed to flush instructions from reservation stations, following an exception event. The circuit is designed so that any instruction of any type executed with exception may initiate flush operation on younger instructions, not only at-design-stage predetermined type and number of instructions. We have shown that a circuit that immediately flushes instructions from reservation stations improves execution time of programs. Although improvement is around 2%, we believe that the proposed circuitry needs to be implemented in microprocessors. Recently, a vulnerability known as side channel attack has been discovered [9]. Side channel attacks exploit speculatively fetched instructions to execute malicious instructions. We believe that, following an exception event, the proposed circuit can kill all not-executed malicious instructions; thus, improving the security of the microprocessors.

REFERENCES

[1] J.E. Smith, and G.S. Sohi, “The microarchitecture of superscalar processors,” Proceedings of the IEEE, Vol. 83, Issue 12, pp. 1609 – 1624, 1995.

[2] V. R. K. Naresh, R. Sheikh, A. Perais, H. W. Cain, “SPF: Selective Pipeline Flush,” IEEE 36th International Conference on Computer Design (ICCD), Oc. 2018, Orlando, FL.

[3] H. Akkary, R. Rajwar, S.T. Srinivasan, “Checkpoint processing and recovery: towards scalable large instruction window processors,” Proceedings of the 36th Annual IEEE/ACM International Symposium on Microarchitecture, MICRO-36, 2003.

[4] R. Kessler, “The Alpha 21264 microprocessor,” IEEE Micro, vol. 19, no. 2, pp. 24–36, 1999.

[5] K. C. Yeager, “The MIPS R10000 superscalar microprocessor,” IEEE Micro, vol. 16, issue 2, pp. 28–41, 1996.

[6] B. Sinharoy, et al. “IBM POWER8 processor core microarchitecture,” IBM Journal of Research and Development, Vol. 59, Iss. 1, pp. 1-21, 2015.

[7] S. K. Sadasivam, B. W. Thompto, R. Kalla, W. J. Starke, “IBM Power9 Processor Architecture,” IEEE Micro, vol. 37, Iss. 2, pp. 40-51, 2017.

[8] N. Binkert et al.,“The gem5 simulator,” ACM SIGARCH Comput. Architecture News, vol. 39, pp. 1–7, 2011.

[9] J. Horn, “Reading privileged memory with a side-channel,” Available at: googleprojectzero.blogspot.com, January 2018.

Sequential Register Renaming

Dejan Spasov
Faculty of Computer Science and Engineering, Skopje, Republic of North Macedonia
dejan.spasov@finki.ukim.mk

***Abstract* – Register renaming unit is a bottleneck in the superscalar cores because it limits the number of instructions and the number of threads that may concurrently be processed. We propose a register renaming unit with linear complexity with respect to the number of instructions simultaneously renamed. The proposed renaming unit renames source operands in a sequential manner following the program order of the instructions. We show that in worst case sequential register renaming may follow contemporary trends with respect to the number of instructions and the number of threads that may be simultaneously renamed.**

Keywords – register renaming; superscalar microprocessor; microarchitecture; register alias table; pipelining.

I. INTRODUCTION

In order to improve the instructions-per-cycle (IPC) rate, modern superscalar cores are processing more than one instruction per clock cycle at any stage of computation. Instructions may be executed out of program order and results may be stored in physical registers of the core. This approach eliminates false data dependencies (WAR and WAW) among instructions and allows execution of instructions out of program order. A register renaming unit is needed to maintain true (RAW) data dependencies between instructions waiting for execution and results stored in the physical registers.

Fig. 1 shows a microarchitecture of a core of a superscalar microprocessor. The core may include fetch unit, decode unit, renaming unit, reorder buffer, one or more physical register files (PRF), dispatch unit, reservation stations (RS), and execution units (EU) [1]. The core may include other components and interfaces not shown on fig. 1. The core may support multiple instruction issue, in-order and out-of-order execution, and simultaneous multithreading, where instructions from one or more threads may simultaneously be processed.

The fetch unit is configured to fetch instructions from memory or L1 cache. Instructions may be from any instruction set architecture, e.g. PowerPC™, ARM™, SPARC™, x86™, etc. Instructions comprise source and destination operands. Instruction set architectures define a set of architectural registers, which are used to denote source and destination operands of the instructions. The compiler uses architectural registers to maintain true data dependencies between the instruction producer of a result and instructions consumers of the result. The fetch unit may include a branch predictor to predict the branch that will be taken when a branch instruction is executed. Responsive to a branch prediction, the fetch unit may speculatively fetch instructions from the predicted branch.

Funded by a project from the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University.

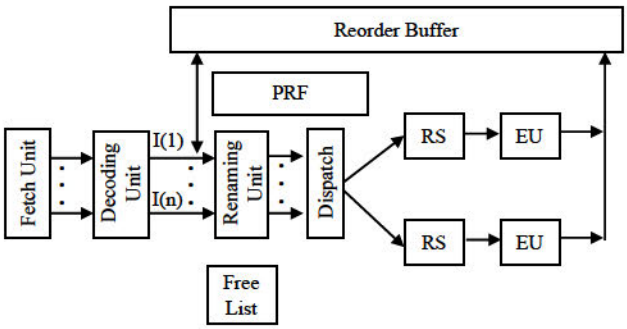


Figure 1. Microarchitecture of superscalar core

The decode unit may be configured to output decoded instructions or instruction (micro-)operations. These decoded instructions may be executed by the execution units. In superscalar cores, decode units are configured to simultaneously output a renaming group of n instructions denoted with I(1) to I(n). Outputted instructions I(1) to I(n) are considered to be in program order, where I(1) is the oldest instruction in the group. Instructions include source operands (SOP) and destination operands (DOP) identified with architectural registers 0, 1, …, L, such that source operands of instructions consumers of a result are equal to a destination operand of an instruction producer of the result.

The reorder buffer maintains in-order retirement of the instructions; thus, supporting out of (program) order execution of instructions. For each instruction the reorder buffer allocates an entry. Instructions enter at one end of the buffer and exit (retire) at the other end. Instructions may be executed out of program order but retired in program order. Reorder buffer entries are organized as circular buffer, where adjacent, in program order, instructions allocate entries with adjacent addresses.

The core may include a plurality of physical registers organized in one or more physical register files (PRF). Physical registers of the core store speculative results and architecturally visible results. The core may employ stand-alone addressing scheme for the physical registers, or physical registers may adopt the addressing scheme of the reorder buffer.

The free list maintains a list of physical registers that may be allocated to instructions with destination operands. For each instruction with destination operand, the free list

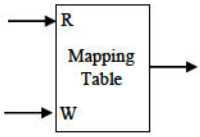


Figure 2. Register renaming unit of a scalar microprocessor

is configured to allocate a physical register. This way destination operands are renamed and false data dependencies among instructions are removed.

The renaming unit is configured to rename (map) source operands of instructions consumers of a result to the physical register allocated to the instruction producer of the result. A source operand of an instruction is renamed to the physical register most recently allocated to instruction with destination operand equal to the source operand.

Execution units (EU) may include any number and type of execution units, e.g. integer units, floating-point units, load/store unit, branch units, etc., configured to execute instructions. Instructions may be executed in-order or out-of-order. One or more reservation stations (RS) may be included within the execution units (EU), where instructions may be waiting to be sent to the execution units. The dispatch unit is configured to dispatch instructions to the reservation stations.

The remainder of the paper is organized as follows: register renaming is introduced in Section II, limitations of contemporary register renaming units are outlined in Section III, novel register renaming techniques that overcome these limitations are proposed in Section IV, hardware complexity of the proposed register renaming units is given in Section V, delay estimates are computed in Section VI, related work is presented in Section VII. The paper concludes with Section VIII.

II. Register Renaming

Superscalar cores include a renaming unit where source operands of instructions are renamed to physical registers. Source and destination operands are architectural registers, such that source operands of instructions consumers of a result must be equal to the destination operand of the instruction producer of the result. For each instruction with destination operand, a physical register is allocated. A source operand of an instruction is renamed to the physical register most recently allocated to instruction with destination operand equal to the source operand.

Most recently allocated physical registers, indexed by the architectural registers, are known as architectural to physical register mappings. Architectural to physical register mappings may be stored in a mapping table (fig. 2). The mapping table includes one read port for each source operand and one write port for each destination operand. In a scalar core, source operands of an instruction are renamed to physical registers from the mapping table at indexes provided by the source operands. After source operands of the instruction are renamed, physical register allocated to the instruction is stored in the mapping table.

The mapping table may be implemented with content addressable memory (CAM), where mapping table entries

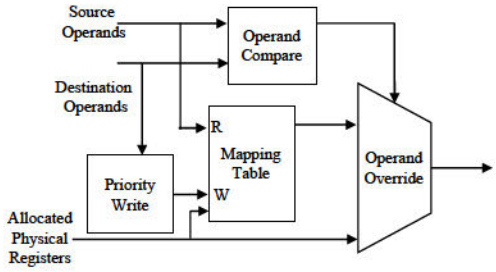


Figure 3. Register renaming unit of a superscalar microprocessor

are indexed with physical register identifiers. Examples of microprocessors with CAM-based mapping tables are the Alpha 21264 microprocessor [2] and the Power4 microprocessor [3].

The mapping table may be implemented with SRAM cells. Entries of the SRAM-based mapping table are indexed with architectural registers. Physical registers may share same address scheme with reservation station entries, as in the System/360 [4]. Physical registers may share same address scheme with reorder buffer entries, as in the Pentium III processors, or may use stand-alone address scheme, as in the Netburst microarchitecture [5]. Modern microarchitectures employ stand-alone address scheme for the physical registers.

Fig. 3 shows basic building blocks of a register renaming unit in a superscalar microprocessor. The register renaming unit is configured to simultaneously rename source and destination operands in a renaming group of two or more instructions. The renaming unit includes mapping table, operand compare unit, operand override unit, and priority write logic. When renaming source operands of instructions in a group, an instruction may be data-dependent on another instruction from the same group (intra-group dependency) or from another group (inter-group dependency).

The mapping table maintains architectural to physical register mappings from prior group of instructions; thus, handling the intra-group dependencies. The operand compare unit and operand-override handle inter-group dependencies. The priority write unit handles priority writes to the mapping table.

A renaming group of n instructions I(1), I(2), …, I(n), may be provided to the register renaming unit (fig. 3). An instruction I(i) may be considered older than an instruction I(j) if i<j. Each instruction, e.g. I(i), may be considered to include a destination operand DOP(i), and a source operand SOP(i). For each instruction I(i) with destination operand DOP(i), a physical register PR(i) is allocated. The mapping table comprises plurality of read ports, coupled to receive source operands SOP(i). For each source operand SOP(i), the mapping unit is configured to output physical register stored at index SOP(i). The operand compare unit is configured to compare each source operand SOP(i) with destination operands of older instructions DOP(1), …, DOP(i-1). If a match is found, e.g. SOP(i)=DOP(j), the operand override unit is configured to output PR(j); thus, renaming the source operand SOP(i). If a match is not found, the operand override unit is configured to output

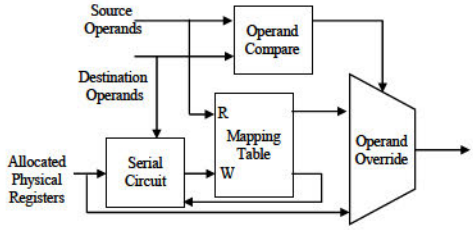


Figure 4. Register renaming unit with one write port and 2n+1 read ports to the mapping table

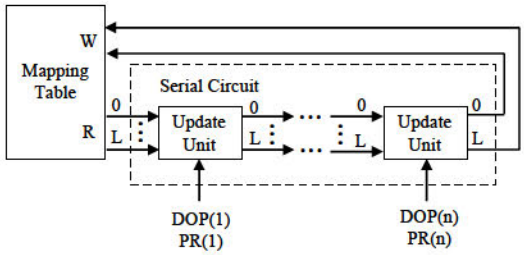


Figure 5. Serial ciruit

physical register stored in the mapping table at index SOP(i).

The priority write unit is configured to compare each destination operand DOP(i) with destination operands of older instructions DOP(1), …, DOP(i-1). If a match is not found, PR(i) is stored in the mapping table at index DOP(i).

III. Problem Statement

For a renaming group of n instructions, the register renaming unit may include n write ports and 2n (or even 3n) read ports to the mapping table. Each read port is coupled to receive a source operand from the group. Each write port is coupled to receive a destination operand from the group. The mapping table may be implemented as a multi-ported SRAM. Die area of multi-ported SRAMs grows quadratically with the number of ports. Power consumption and power density of the SRAM memory are proportional to the die area.

The priority write unit may include $\frac{1}{2}n(n-1)$ comparators to compare each destination operand with destination operands of older instructions from the group. The operand compare unit may include $n(n-1)$ comparators to compare each source operand with destination operands of older instructions in the group.

Therefore, it may be concluded that die area, wiring complexity, power consumption, and power density of the register renaming unit on fig. 3 grow quadratically with respect to the number of instructions in the renaming group. Moreover, in simultaneous multithreading, a separate register renaming unit is needed for each thread.

IV. Proposed Solution

Fig. 4 shows a register renaming unit comprising a mapping table, operand compare unit, operand override unit, and a serial circuit. The register renaming unit includes one write port and 2n+1 read ports to the mapping table. The serial circuit replaces the priority write logic and n-1 write ports to the mapping table.

Fig. 5 shows the principle of operation of the serial circuit. The serial circuit is coupled to a read port of the mapping table to receive architectural to physical register mappings. The serial circuit comprises n update units coupled in a chain. Architectural to physical register mappings propagate through the chain of update units over bus lines denoted with architectural registers 0 to L. A bus line denoted with architectural register I may be considered to propagate physical register allocated to instruction with destination operand I. Each update unit is coupled to receive destination operand DOP(i) and allocated physical register PR(i) of an instruction I(i). Each update unit is configured to update architectural to physical register mappings with the mapping DOP(i) to PR(i). The first update unit, coupled to the mapping table, is configured to output PR(1) on a bus line denoted with DOP(1). The second update unit, coupled to the first update unit, is configured to output PR(2) on a bus line denoted with DOP(2), etc.

The chain of update units sequentially, in program order, inserts allocated physical registers PR(1), PR(2), …, PR(n) on bus lines denoted with DOP(1), DOP(2), …, DOP(n), respectively. The last update unit is coupled to the mapping table to write updated architectural to physical register mappings.

Fig. 6 shows an update unit. The update unit is coupled to receive architectural to physical register mappings on the bus lines denoted with 0 to L. Bus lines 0 to L are terminated at 2-in-1 multiplexers. The other input on the 2-in-1 multiplexers is coupled to receive a physical register PR allocated to an instruction with destination operand DOP. A decoder is coupled to receive destination operand DOP. Output signal lines of the decoder are coupled as selection control to the 2-in-1 multiplexers. Multiplexers are configured to output architectural to physical register mappings on the bus lines denoted as 0' to L'. The update unit is configured to output PR on a bus line denoted with DOP. The remaining bus lines are configured to output physical register received from the coupled input bus line.

Fig. 7 shows further improved register renaming unit. The register renaming unit includes a mapping table, a serial circuit, and renaming multiplexers. The mapping table is operated as register with one write port and one read port. A serial circuit comprising a chain of update units is coupled to the mapping table. Architectural to physical register mappings propagate from the mapping table through the update units over dedicated bus lines. Rename multiplexers are coupled to the chain of update units.

An update unit may be configured to output PR(i-1) on a bus line denoted with DOP(i-1). A rename multiplexer may be coupled to the update unit. Source operand SOP(i) of an instruction I(i) may be provided as selection control to the multiplexer. The multiplexers is configured to output physical registers propagating on the bus lines denoted with SOP(i); thus, renaming the source operand SOP(i) to a physical register.

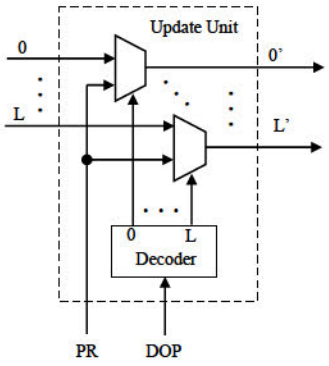


Figure 6. Update unit

V. Hardware Complexity

Hardware complexity of the register rename unit on fig. 4 is quadratic $O(n^2)$ with respect to the size n of the renaming group. The serial circuit on fig. 5 is configured to update architectural to physical mappings in a sequential manner. Therefore, the register rename unit on fig. 4 does not need the priority write unit with $\frac{1}{2}n(n-1)$ comparators. However, this register renaming unit still maintains the operand override unit with a quadratic number of comparators and 2n+1 read ports to the mapping table also with quadratic complexity.

Hardware complexity of the register rename unit on fig. 7 is linear $O(n)$ with respect to the size of the renaming group n. For each instruction in the renaming group, this register renaming unit requires one update unit and two rename multiplexers.

The register rename unit on fig. 7 also does not need the priority write unit with $\frac{1}{2}n(n-1)$ comparators. Moreover, this register rename unit does not need the operand compare unit and the operand override unit. The mapping table is operated as register with one read port and one write port. Using a register instead of a table reduces wiring complexity (bit lines), the number of circuitries for pre-charge, the number of sense amplifiers, and the number of write drivers. Therefore, none of the building blocks of the register rename unit on fig. 7 is with quadratic complexity.

VI. Delay Estimate

For a renaming group of n instructions, the proposed renaming units consists of a chain of n update units consisting of 2-in-1 multiplexers. Therefore, the critical path of the proposed renaming units is a chain of n 2-in-1 multiplexers. The 2-in-1 multiplexers may be implemented with transmission gates [6]. Hence, the critical path in the serial circuit is a chain of n transmission gates. Assuming ideal bus lines, the delay of a chain of n identical transmission gates can be estimated using the Elmore approximation

$$t_p = 0.69CR_{eq}\frac{n(n+1)}{2} \qquad (1)$$

In [7], it has been reported that for a chain of transmission gates Elmore approximation (1) corresponds

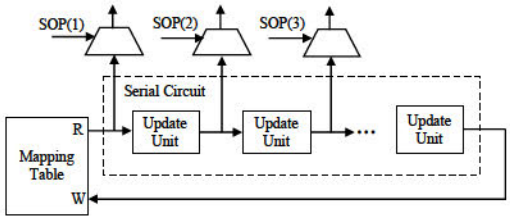


Figure 7. Register renaming unit with one write port and one read port to the mapping table

to simulated delay with SPICE simulators. Moreover, considering ideal bus lines and rename multiplexers implemented with a decoder and transmission gates coupled to the bus lines, the Elmore delay approximation (1) can be applied even to the renaming unit on fig. 7.

Table I summarizes estimated delays for different process technologies. Values for the resistance and the capacitance are taken from [6] and [7]. Modern superscalar cores are configured to simultaneously rename up to five or six instructions. Table I suggests that proposed renaming units can reach the renaming capacity of five or six instructions, while maintaining the delay requirement for high clock frequencies.

It may be observed that the proposed renaming units can be configured to rename more than six instructions. For example, using different process technology (e.g. 7nm), using repeaters every three or four gates [6], using tapered transmission gates [8], etc. However, the size of the renaming group is determined not only by the capacity of the renaming unit, but also by design choices in the fetch unit, the decode unit, and the dispatch unit. Thus, any increase of the renaming group requires careful redesign of the entire core. Transmission gates (or even pass transistors) in this context offer smallest die area and power saving.

VII. Related Work

Complexity, power density, and power consumption are major weaknesses of the register renaming units. In order to overcome these weaknesses, several techniques have been proposed. In general, these techniques focus on reducing the number of ports to the mapping table and adding additional memory (caches) to relieve the pressure on the mapping table.

In [9], Moshovos studies the average number of operands in the SPEC2000 benchmark programs and in multimedia applications. Moshovos proposed a reduction on the number of read ports at the cost of a negligent slowdown of the instruction-per-cycle performance.

In [10], a mechanism is proposed to reduce power consumption. In particular, if a source operand depends on a destination operand from the same renaming group, read port for that source operand is turned off. Moreover, the authors propose buffering of the renamed destination operands in a small number of latches; thus, reducing the pressure on the mapping table.

In [11], E. Safi, et al. propose two-stage pipelined register renaming technique. Priority write unit, operand

TABLE I. ESTIMATED DELAY FOR PROCESS TECHNOLOGIES

Process tech. [nm]	250	65
R_{eq} [kΩ]	8	5
C [fF]	3.6	0.2
Num. of stages [n]	5	6
t_p [pS]	298	14

compare unit and operand override unit are placed in one of the pipeline stages while reading and writing to the mapping table in the other pipeline stage.

In [12], S. Petit et al. use RAM-based and CAM-based tables to perform what they call hybrid renaming at the expense of increased power dissipation.

In [13], [14], techniques for caching renamed operands have been proposed.

VIII. CONLUSION

Modern state-of-the-art microprocessor architectures are designed to simultaneously rename up to five or six instructions per cycle. We have proposed register renaming units that can rename five to six instructions, while maintaining delay requirements for high clock frequencies.

Proposed renaming units are configured to rename source operands in sequential manner. Sequential manner requires less hardware. Less hardware implies smaller die area, reduced power consumption, and power density. Moreover, the proposed renaming units can be implemented with transmission gates or pass transistors to further reduce die area and power consumption.

We believe that the proposed register renaming unit may give designers opportunity to increase the width n of the renaming group, or to increase the number of threads per core in simultaneous multi-threading, or to reduce the number of pipeline stages. Ultimately, this may improve the instructions per cycle (IPC) rate of the microprocessors.

REFERENCES

[1] J.E. Smith, and G.S. Sohi, "The microarchitecture of superscalar processors," Proceedings of the IEEE, Vol. 83, Issue 12, pp. 1609 – 1624, 1995.

[2] R. Kessler, "The Alpha 21264 microprocessor," IEEE Micro, vol. 19, no. 2, pp. 24–36, 1999.

[3] T. N. Buti, R. G. McDonald, Z. Khwaja, A. Ambekar, H. Q. Le, W. E. Burky, B. Williams, "Organization and implementation of the register-renaming mapper for out-of-order IBM POWER4 processors," IBM Journal of Research and Development, Vol. 49, Issue 1, pp. 167 - 188, 2005.

[4] R. M. Tomasulo, "An efficient algorithm for exploiting multiple arithmetic units," IBM Journal of Research and Development, Vol. 11, No. 1, pp. 25-33, 1967.

[5] G. Hinton, D. Sager, M. Upton, D. Boggs, D. Carmean, A. Kyker, P. Roussel, "The Microarchitecture of the Pentium® 4 Processor," Intel Technology Journal, 2001.

[6] N. H. E. Weste, and D. M. Harris, "CMOS VLSI Design: A Circuits and Systems Perspective," 4th Edition, Pearson/Addison-Wesley, 2011.

[7] J. M. Rabaey, A. Chandrakasan, B. Nikolic, "Digital Integrated Circuits," 2nd edition, Pearson, 2003.

[8] B. D. Andreev, E. Titlebaum, and E. G. Friedman, "Tapered transmission gate chains for improved carry propagation," in The 2002 45th Midwest Symposium on Circuits and Systems, pp. 449-452, Tulsa, OK, USA, 2002.

[9] A. Moshovos, "Power-aware Register Renaming," Technical Report, Computer Engineering Group, University of Toronto, 2002.

[10] G. Kucuk, O. Ergin, D. Ponomarev, and K. Ghose, "Reducing power dissipation of register alias tables in high-performance processors," IEE Proc.-Comput. Digit. Tech., Vol. 152, No. 6, November 2005, pp. 739-746.

[11] E. Safi, A. Moshovos, and A. Veneris, "Two-Stage, Pipelined Register Renaming," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 19, Iss. 10, pp. 1926 – 1931, 2011.

[12] S. Petit, R. Ubal, J. Sahuquillo, P. López, "Efficient Register Renaming and Recovery for High-Performance Processors," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 22, Iss. 7, pp. 1506 – 1514, 2014.

[13] S. Vajapeyam and T. Mitra, "Improving superscalar instruction dispatch and issue by exploiting dynamic code sequences," in Proc. of the Int. Symp. on Computer Architecture, 1997, pp. 1–12.

[14] R. Shioya, and H. Ando, "Energy Efficiency Improvement of Renamed Trace Cache through the Reduction of Dependent Path Length," 2014 IEEE 32nd International Conference on Computer Design (ICCD), 2014.

Improved circuits for a random pulse computer

Mateja Batelić[1,2], Mario Stipčević[1]
[1]Photonics and Quantum Optics, Center of Excellence for Advanced Materials and Sensing Devices, Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia
[2]Department of Physics, Faculty of Science, University of Zagreb, Bijenička cesta 32, 10000 Zagreb, Croatia
E-mail: mateja.batelic096@gmail.com
E-mail: mario.stipcevic@irb.hr

***Abstract*— We present improved circuits intended for building a universal computer based on Random Pulse Computing (RPC) paradigm, a biologically-inspired way of computation in which variable is represented by a frequency of a Random Pulse Train (RPT) rather than a logic state. The RPC we mention here is also known as "stochastic unipolar computation" in newer literature. Unlike in previous art, where randomness is obtained from electronics noise or a pseudorandom shift register while processing circuitry is deterministic, in our approach both variable generation and signal processing rely on the random flip-flop (RFF), whose randomness is derived from a fundamentally random quantum process. This offers better conceptual simplicity and may lead to better precision in chained operations due to the higher randomness of output RTPs.**

***Keywords*— entropy, random pulse computing, randomness, random pulse train, stochastic computing**

I. INTRODUCTION

Today, the digital computing paradigm (DC), based on Turing machine theoretical model, is almost exclusively used. Implemented in electronics logic circuitry, that executes basic logic operations, and realized in solid-state chips, it allows very fast computation with an arbitrary precision. Since DC is incapable of generating randomness, a version enriched by a (single) random number generator, so called "randomized Turing machine", offers large speedup of certain tasks by use of randomized algorithms, for example testing primality of (big) numbers by Soloway-Strassen algorithm [1-2].

A radically new, quantum computation (QC) paradigm has been proposed by Feynman in 1981 [3]. It makes use of strong correlations of quantum entanglement and superposition principle to reach an exponential speed-up over DC of a small set of algorithms of a great practical importance. However, QC has proven difficult to realize in practice.

The Random Pulse Computing (RPC) paradigm, proposed in a seminal work of John von Neumann [4] in 1956, shortly flashed in 1960s only to be run-over by the digital computation that flourished in 1970s, after which RPC fell into oblivion. Re-born in mid 2010s, RPC can be thought of as a third computational paradigm. RPC uses *random pulse train* (RPT), a sequence of logic pulses of fixed (and irrelevant) width that appear randomly in time, wherein each pulse can be thought of as being a Poissonian random event. The only parameter describing RPT is its average pulse rate which is used as a variable (number) in this type of calculation. RPC paradigm features a set of unique and peculiar characteristics that qualitatively resemble some of the characteristics of living beings, such as ability to perform complex operations with a few hardware, immunity to errors in hardware components, massive parallelism and streamline computation (the output is formed continuously from the input), to name just a few. The main drive behind recent revival of the RPC is hope that it could efficiently (in terms of execution time, hardware and energy) solve problems that are currently difficult and/or energy consuming for the DC, but apparently easy for living beings, such as: learning, recognition and autonomous behavior. One has now a much better technological platform(s) for its realization than back in 1960s. In our approach, we use FPGA and the quantum RFF [5] as a single additional circuit required to produce variables, random numbers, make random decisions and perform mathematical operations. By using a true, quantum randomness, the RPC may be seen as a bridge between DC and QC.

In this work we present five improved or new circuits that perform basic algebraic functions and compare their transfer functions and errors to the existing art.

II. RANDOM PULSE TRAIN

In its original appearance [6-7], the RPT is a time-wise random sequence of logic pulses of constant height and width, as shown in Fig. 1, wherein each positive-going

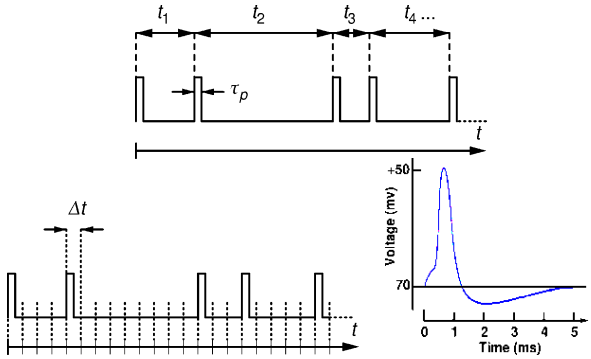


Figure 1. RPT in which each pulse corresponds to a Poissonian random event with frequency $1/\langle t_i \rangle$ (up); Time-digitized RPT wherein appearance of a pulse in a given time bin of width Δt is an independent binomial random event with probability p (down left); one full cycle of a biological nerve pulse in mammalians that lasts about 5 ms (down right).

edge corresponds to one Poissonian random event.

Such events can be obtained from certain types of electronics noise passing a constant threshold [8], decay of a nuclei, etc. As mentioned before, a p.d.f. of waiting times is an exponential distribution with a single parameter: average frequency of pulses. As will be shortly discussed, the idea of RPC includes performing basic logic operations between two or more RPTs. This can be conveniently realized in field programmable gate array (FPGA) chips.

In order to avoid problems with randomly appearing signals in FPGA, we digitize the timeline to non-overlapping bins of duration Δt, as shown in Fig. 1. Technically, a "clock" of frequency $f_{\mathrm{CLK}} = 1 / \Delta t$ is now defining time. Each time bin can hold a pulse or no-pulse. A probability $p \in [0,1]$ of a pulse to appear in a time bin, rather than frequency, is now the numerical information carried by the RPT. Note that pulses will still appear with a definite frequency, namely $p / \Delta t$, but time digitization ensures that pulses from various RPTs appear neatly aligned in time so that they can be unambiguously combined by logic circuits.

III. Improved Basic Operations Circuits

Basic arithmetic circuits perform elementary arithmetic operations between two variables: addition, subtraction, multiplication and division, and are a necessity towards building a universal RPC-based computer. We realize all circuits mentioned in this section using a DE0-nano board of Terrasic with Cyclone IV FPGA chip. In graphical representation of experimental data, transfer functions and calculation errors of these circuits, each curve is made of 72 points connected with a straight line, while each of these points is obtained by measuring 80 million time intervals (bits) to achieve statistical error of 0.0001.

A. Multiplication and summation circuits

Multiplication is the simplest operation in RPC: it can be done with a single AND gate, as shown in Fig. 2. Because pulses from RPTs p_0 and p_1 are independent, a probability to have two such pulses in the same time bin is just $p_0 \cdot p_1$, therefore this circuit gives the exact result. Note that both input and output variables have span in the interval $[0,1]$, which means that this circuit is optimal in terms of dynamic range (no scaling needed).

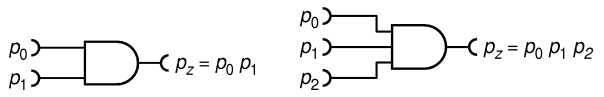


Figure 2. Circuits for exact multiplication of two numbers (a); and three numbers (b).

This principle can be upgraded to multiply three or more numbers *at the same time* with a minimum additional hardware requirement, just by adding the required number of inputs to the AND gate, as shown in Fig 2b.

Approximate addition of two RPTs can be done with a single OR gate, as shown in Fig. 3a. Any input pulse passes through the OR gate, but if the two pulses coincide, only one pulse will be formed at the output.

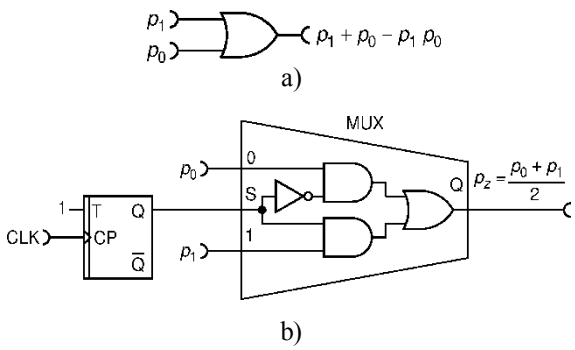


Figure 3. Addition circuits: approximate one with OR gate (a); exact one with a multiplexer (MUX) circuit (b).

Following the discussion from the previous section, it is plain to see that operation done by the OR gate is:

$$p_z = p_0 + p_1 - p_0 p_1. \tag{2}$$

It is interesting that $p_z \in [0,1]$, which can be seen immediately from equality:

$$p_z = 1 - (1 - p_0)(1 - p_1). \tag{3}$$

Thus, this circuit is dynamically optimal, albeit it does not compute addition but something else. For small p_0 and p_1, this circuit performs an approximate addition because then the multiplicative term in Eq. (2) may become negligible. This circuit can be generalized to perform a simultaneous approximate summation of 3 or more numbers by just adding inputs to the OR gate.

Exact addition with a dynamically optimal scaling can be done via a multiplexer (MUX) circuit, shown in Fig. 3b. This circuit samples randomly, with equal probability, pulses from the two inputs. In previous art, thus, the resulting operation is:

$$p_z = \frac{p_0 + p_1}{2} \tag{5}$$

and the circuit is dynamically optimal. The extra factor 2 is not a caveat because it can be counted in a calculation.

This is illustration of one of the key differences between DC and RPC. While calculation in DC is sequential, i.e. multiplication of several numbers assumes first multiplying two of them, then multiplying the result with the third one etc. until numbers are exhausted, in RPC, the multiplication is done simultaneously. Furthermore, there is no idle time during which no information is present at the output.

B. Division circuits

Division is a difficult operation in RPC and only approximate methods are known. Let us imagine we want to calculate $p_z = p_0/p_1$. The idea is to start with a guessed value (probability) of p_z (say 0.5), multiply it with p_1 (we do know how to multiply exactly!) and compare result to p_0. If the result is smaller than p_0, we enlarge p_z. On the contrary, we take a smaller p_z until the right solution is reached. A circuit that does that,

published in [10] with a small error (corrected here), is shown in Fig. 4.

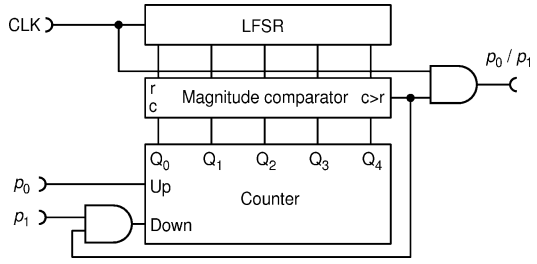
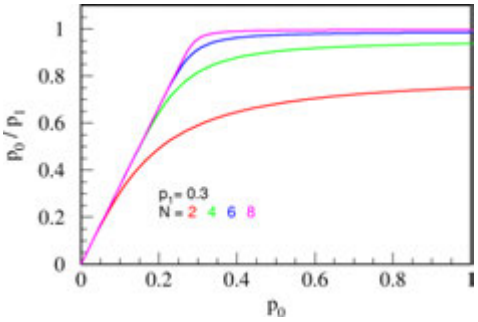


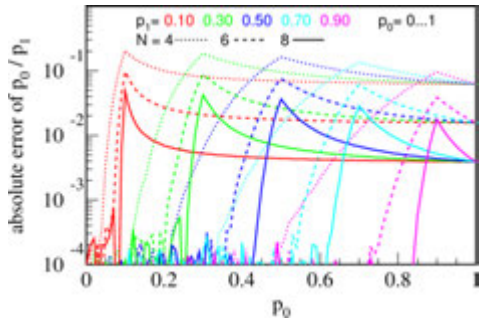
Figure 4. Feedback based division circuit with LFSR.

The transfer function and errors for a certain space of input parameters p_0, p_1 and the counter bit of length N, for the divider in Fig. 4, are shown in Fig. 5. When $p_1 > p_0$, the function levels off at the maximum possible value of 1 as shown in Fig. 5a. This behavior is accounted for when calculating errors shown in Fig 5b. We see that error is largest when $p_1 \approx p_0$. The error becomes smaller with enlarging counter's number of Nbits, but it remains rather large even for an 8-bit counter.

A problem with this circuit is that, due to the use of a LFSR which shifts by one bit with each CLK pulse, it tends to cause long sequences of pulses and long periods of absence of pulses at the output p_z, thus deforming the desired exponential p.d.f. This will have a deteriorating impact to precision of calculations in which output of this circuit is used as an input.



a)



b)

Figure 5. The transfer function (a); and errors from the ideal division calculation (b) for the division circuit shown in Fig. 4.

In order to improve on that, we propose to use true randomness implemented via RFFs, as shown in Fig. 6. According to [5], a T-type RFF (TRFF) acts as an ordinary TFF with provision that its clock input (Cp) acts with probability of 0.5, randomly. Thus, if T input is held high, a TRFF generates an independent random bit upon each clock pulse. Indeed, we find that this circuit performs better in terms of randomness of the output than circuit in Fig 4, but its precision differs insignificantly form the circuit in Fig 4.

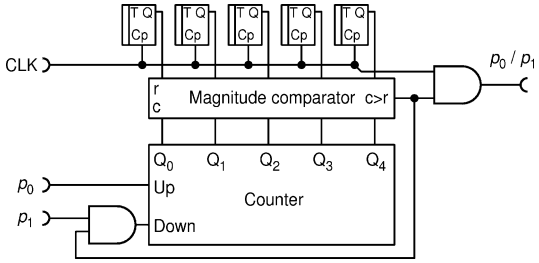


Figure 6. T-type random flip flop (a); and its use in the improved division circuit (b). If not connected, input T is assumed being HIGH.

To improve greatly on precision and simplicity, we further propose circuit shown in Fig 7.

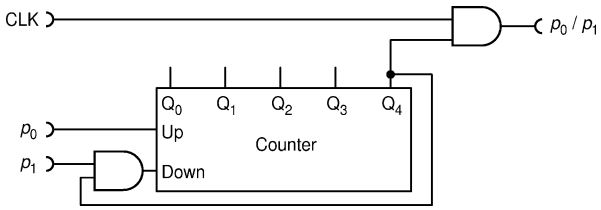
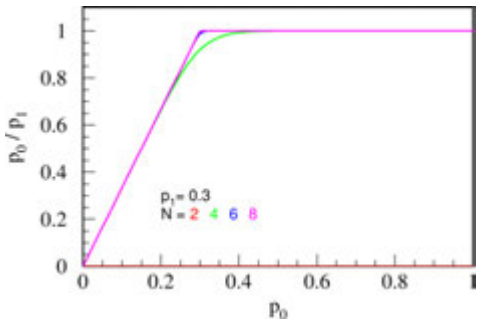
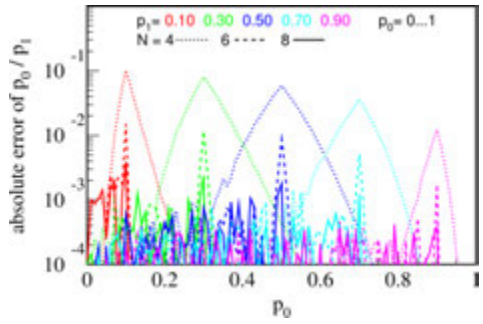


Figure 7. A simple and precise division circuit.

This circuit is much more sensitive to the output error and therefore yields better precision. Its transfer function and errors are shown in Fig. 8. Because of its precision, circuit in Fig. 7 allows us to use a smaller counter, which further simplifies the division circuit.



a)



b)

Figure 8. The transfer function (a); and errors from the ideal division calculation (b) for the circuit shown in Fig. 7.

Note that it has no comparator nor sources of randomness, only the counter, but, because the AND gate pass through or stops high frequency pulses from CLK, it tends to generate long sequences of pulses and absence of pulses, thus being the worst in terms of deviation from the exponential distribution.

C. *Subtraction circuits*

Somewhat surprisingly, subtraction turns out to be the most complicated of all basic arithmetic functions in RPC paradigm. This can be understood by noticing the following expression:

$$p_1 - p_0 = \left(1 - \frac{p_0}{p_1}\right) p_1, \tag{7}$$

which indicates that subtraction relies on division. Multiplication and subtraction from 1 can be easily done by logic circuits as shown in Fig 9a. By inserting the division circuit shown in Fig. 4, one arrives to the subtraction circuit in Fig. 9b [10].

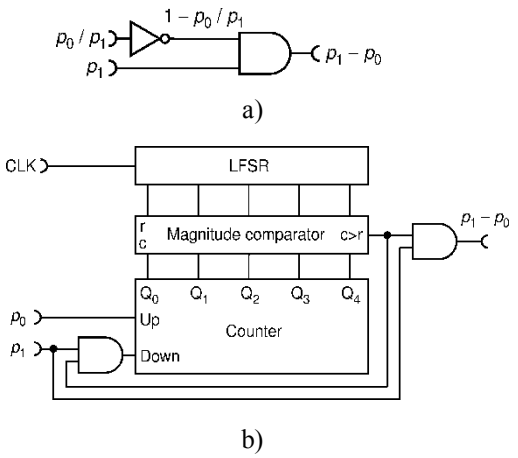


Figure 9. Subtraction via division principle (a); realization of a subtraction circuit utilizing a division circuit (b).

Regarding precision, and to some extent randomness of output RPT, this circuit inherits properties of the division circuit that is used within. Again, improvement in randomness and hardware reduction, without a gain in precision, is obtained by substituting LFSR with TRFF, as shown in Fig. 10.

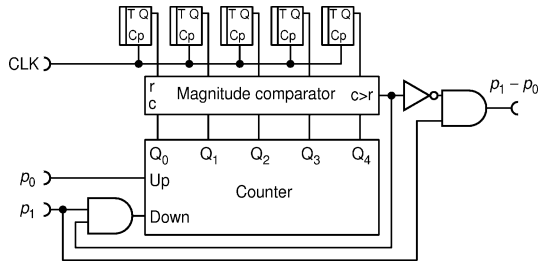


Figure 10. A subtraction circuit with an improved output randomness.

As with the division circuits, there is no difference in transfer function and errors between circuits in Fig. 9 and Fig. 10 except that the TRFF version shows better randomness. Results for both circuits are shown in Fig. 11.

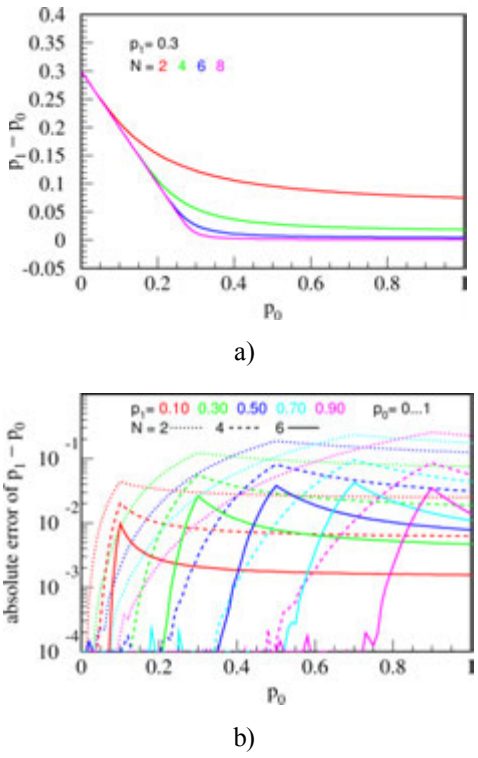


Figure 11. The transfer function (a); and errors from the ideal division calculation (b) for the circuit shown in Fig. 7.

We see that the transfer function in Fig. 11a contains two linear parts. A non-sharp joint of the two indicates computation error around $p_0 = p_1$, the problem inherited from the division circuit used herein and explained for graphs in Fig. 5 and circuit in Fig. 4a.

To significantly improve on precision, albeit with worsening of the randomness of the output, we propose the circuit in Fig. 12.

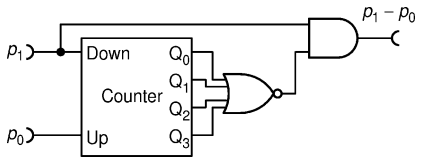
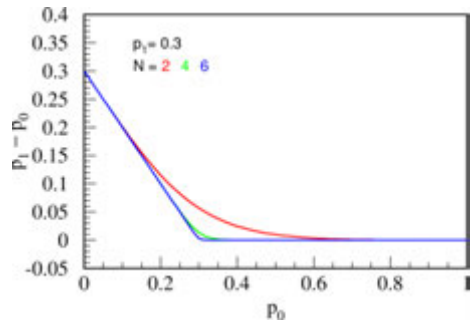
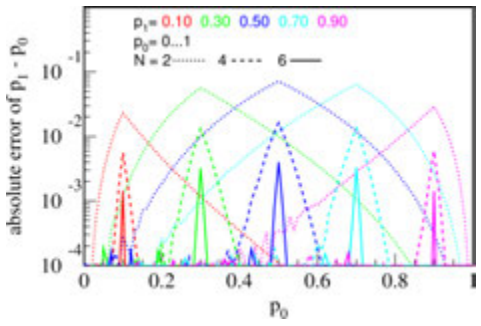


Figure 12. A simple and precise subtraction circuit with counter of $N = 4$ bits.

The circuit functions in the following way. The counter counts pulses from p_0 and inhibits exactly as many pulses from p_1 by means of the output AND gate. A source of error is when counter does not have enough capacity to count pulses from p_0 before a pulse from p_1 arrives and goes into saturation. The chance of this happening is highest when $p_0 \approx p_1$ and can be lowered by taking a bigger counter. Even though this subtractor is not derived from a divider, the transfer function and errors, shown in Fig. 13, show close resemblance to patterns seen in division circuit in Fig. 7.



a)



b)

Figure 13. The transfer function (a); and errors from the ideal division calculation (b) for the circuit shown in Fig. 12.

Again, we see there is a sharp edge between linear parts of the transfer curve, indicating small errors. In fact, explanation of the transfer function and errors is quite similar of the division circuit in Fig. 7.

IV. Conclusion

In this work we have introduced five improved or new circuits for RPC computer: the MUX-based addition, iterative division and subtraction, all using RFF, and deterministic iterative division and subtraction. The RFF-based circuits have better output randomness than LFSR-based circuits but their precision *per se* is virtually the same. We hypothesize that this is important for precision of subsequent operations that use output of these circuits as their input, which needs to be investigated further.

However, precision of the RFF-based circuits remains essentially the same as of LFSR-based ones. We therefore develop two fully deterministic circuits for division and subtraction which are perhaps the most precise possible, but with not-so-great output randomness. It may be that there is an inevitable trade-off between precision and randomness of the output. These are questions open for further investigation.

Acknowledgments

Funded by Croatian Ministry of Science, Education and Sports, grants 533-19-14-0009 and KK.01.1.1.01.0001.

References

[1] Solovay, Robert M.; Strassen, Volker, "A fast Monte-Carlo test for primality". SIAM Journal on Computing. **6,** 84–85 (1977).

[2] Solovay, Robert M.; Strassen, Volker, "Erratum: A fast Monte-Carlo test for primality". SIAM Journal on Computing. **7**, 118 (1978).

[3] R. P. Feynman, "Simulating Physics with Computers", Int. J. Theor. Phys., **21**, 467-488 (1982).

[4] J. von Neumann, "Probabilistic logics and the synthesis of reliable organisms from unreliable components". In Bródy, F.; Vámos, Tibor (eds.). *The Neumann Compendium*. World Scientific. pp. 567–616 (1995).

[5] M. Stipčević, "Quantum random flip-flop and its applications in random frequency synthesis and true random number generation", Rev. Sci. Instrum. **87**, 035113 (2016).

[6] S. T. Ribeiro, "Random-Pulse Machines" IEEE Transactions on Electronic Computers, **16**, 261-276 (1967).

[7] B. R. Gaines, "Stochastic Computing Systems", Advances in Information Systems Science, **2**, 37-72 (1969) Ed. T. T. Julius, Springer US, Boston, MA.

[8] R. C Lawlor, "Computer utilizing random pulse trains", patent US3612845A, priority date Oct. 12, 1971.

[9] A. Alaghi, J. P. Hayes, "Survey of stochastic computing", ACM Trans. Embed. Comput. Syst., 12, 92:1-92:19 (2013).

[10] Y. Liu and K. K. Parhi. 2017. "Computing Polynomials Using Unipolar Stochastic Logic.", J. Emerg. Technol. Comput. Syst. **13**, 3, Article 42 (April 2017).

[11] B. Widrow, Y. Kim, D. Park, and J. Krause Perin, "Nature's Learning Rule: The Hebbian-LMS Algorithm", appears in Artificial Intelligence in the Age of Neural Networks and Brain Computing, Chaper 1, Pages 1-30, Eds.:Robert Kozma, Cesare Alippi, Yoonsuck Choe, Francesco Carlo Morabito, Academic Press, 2019.

Two approaches in solving Rubik's cube with Hardware-Software Co-design

Emir Barucija, Amila Akagic, Samir Ribic, Zeljko Juric
Department of Computer Science and Informatics,
Faculty of Electrical Engineering,
University of Sarajevo
Email: {ebarucija1, aakagic, sribic, zjuric}@etf.unsa.ba

***Abstract*—Algorithms for solving Rubik's cube have been an active research area since the first appearance of the cube in 1974. The challenge posed when solving the cube is to choose an algorithm that solves the cube for the minimum number of steps. Many algorithms are already implemented in software, but not many are tested with modern hardware-software methodologies, such as hardware-software co-design. Here, the challenge is to take into consideration limiting factors of hardware and implement the most efficient solution. In this paper, the hardware/software co-design is used to solve the random configuration of Rubik's cube. Two algorithms are used: the Basic algorithm and the *Kociemba* algorithm. The Basic algorithm is easy to understand and implement but requires many more steps to solve the cube than the *Kociemba* algorithm. The *Kociemba* algorithm requires some pre-processing tasks, such as depth-first search and pruning trees, but can solve the cube in about 25 moves. Both algorithms are implemented and tested on a custom-made robot with mechanical parts, actuators, grippers and Intel's DE1-SoC for drive control and implementation of solving algorithms. The robot successfully solved a number of random configurations. Performances (running time, number of moves needed for solving the cube) of both algorithms are measured and compared.**

***Keywords*-Rubik's cube, Kociemba algorithm, System on Chip, Hardware/Software Co-design, Image Processing**

I. Introduction

The Rubik's cube was invented in 1974 by the Hungarian architecture professor Ernö Rubik. Over the years it has captured the imagination of millions of people and quickly became one of the world's best selling toys ever [1] and the most famous combinatorial puzzle of its time [2]. It is remarkable that so many years after its invention there are still debates about the underlying mathematical principles of the Rubik's cube. Its popularity can be attributed to the simplicity of its design, which is combined with the mathematical complexity of the puzzle. Solving Rubik's cube is a challenge for everyone who likes solving puzzles and it is as old as the cube itself. The original-size Rubik's cube (3x3x3) has 43 quintillion possible states, of which only one is desired - the solved state.

First algorithms for solving the cube were simple and had only one task - to solve the cube successfully. During the past three decades, as the cube gained popularity, better and better algorithms were arising. Most of the algorithms are designed in a way suitable for manipulation with human hands. However, as technology advanced, researchers were able to design very sophisticated robots for solving the cube. Principles of real cube solving using a robot include three specific phases:

- Detect the cube's initial state.
- Generate the solution.
- Perform particular moves on a robot to solve the cube.

There are many research papers in the literature that address issues with specific phases or provide complete solutions.

The process of designing the Rubik's cube solver robot has an educational purpose, and that knowledge can be further applied to a number of problems, such as manipulating different objects and performing different human tasks. Many aspects have to be considered when building a Rubik's cube solver. Some questions to consider are: which mechanical parts, actuators, grippers, servo controllers to choose? and how to coordinate these parts with the solving algorithm? The speed of solving the cube is directly related to the quality of the parts used, and consequently, this is related to the cost of the robot. Several robots that can very quickly solve the Rubik's cube already exist. Currently, the fastest solving robot is MIT Rubik's solving robot [3] with time 0.38s. Other examples of high-performance robots include very costly Shadow Dexterous E Series Hand which can move faster than a human hand.

In this paper, we present two approaches in solving Rubik's cube with off-the-shelf inexpensive components with Hardware/Software Co-design methodology. We use Terasic's DE1-SoC hardware design platform built around the Altera's System-on-Chip (SoC) FPGA [4]. We implement all phases needed for solving the cube and build a robot platform to demonstrate the process. First, we detect the initial state of the cube by analyzing images of cube's sides with a simple color recognition algorithm. Second, we generate the solution with two algorithms: the Basic algorithm (defined by David Singmaster) and the Kociemba algorithm. Both algorithms are implemented on an ARM-based hard processor system (HPS), and servo motors are controlled by a custom-made servo controller designed in RTL (Register Transfer Level) on the FPGA [5]. These systems are integrated with Altera's Qsys integration tool. Third, we present the design of the in-expensive platform that can be obtained for three orders of magnitude less than state-of-the-art platforms. Contributions of this paper

are: 1) a new system design based on hardware/software co-design and 2) comparison of two known algorithms on the proposed system.

Results indicate that the platform can successfully solve any random configuration of the cube. We present results from ten different random configurations of the cube solved by two different algorithms. Both algorithms have advantages in some regard, but we can conclude that the Kociemba algorithm has proven to be more than 5 times better than the Basic algorithm when considering the number of moves and solving time. The robot is not as fast as the state-of-the-art and this is due to low-performance servo motors used in the system. By replacing these motors with high precision pneumatic manipulators, the system could be executing much faster and guarantee the operating accuracy of every step.

The paper is organized as follows: related work is briefly reviewed in Section II. In Section III, we describe the new system for solving a random configuration of the cube. In Section IV we briefly overview two algorithms for solving the cube. Results are presented in Section V, and in Section VI we conclude the paper with some remarks.

II. Related Work

The methodology for solving the Rubik's Cube usually includes some or all of the following phases:

- Detection of the initial state (may include color recognition).
- Generation of the solution for solving a randomly scrambled cube.
- Implementation of the moves on the robot to solve the cube.

In the literature, many papers present results of a single phase, a combination of phases or all of the phases. Here, we present some works in a systematic manner.

The following works address the problem of color recognition of the initial state of the cube and test the solution on the existing robots[1]. Lu et al. [7] presented a Rubik's Cube robot based on a mechatronic system with a simple structure, with all the manipulators driven by pressurized air. High precision pneumatic manipulators are used to guarantee the operating accuracy of every step. The color recognition of the cube's initial state is analyzed in the HSV color space. The robot can solve a random configuration of the cube in less than 60 seconds. In [8], a vision-based algorithm for 2D Rubik's cube state detection is presented. The paper addresses the problem when a long-term fixed camera-cube position is not possible. The Hough transform and advanced clustering functions are used for the cube position detection, then a color detection algorithm is applied to detect the state of the cube. In [9], the color drifting problem of color recognition of the cube is addressed. A method "Scatter balance & extreme learning machine" (SB-ELM) is proposed to illustrate the efficiency of the training based method.

[1]The list of all Rubik's Cube solving robots is available at [6].

The following works address the problem in simulation mode and test the solutions on a multi-fingered dexterous hand. Li et al. [10] presented an interesting approach to solving the 2x2x2 Rubik's cube with a multi-fingered dexterous hand with a 24-DoF. They designed a model-based cube solver to find optimal move sequence with a hierarchical deep reinforcement learning method that separates planning and manipulation. Akkaya et al. [11] demonstrated that models trained only in simulation can be used to solve a manipulation problem on a real robot. They presented a novel algorithm called "Automatic domain randomization" and a new robot platform built for machine learning. The platform is based on the costly Shadow Dexterous E Series Hand (E3M5R) which can move faster than a human hand.

The strategy of many existing algorithms for solving the cube is to find a solution that requires the fewest number of moves. This strategy is often referred to as the "God's algorithm". There are a number of algorithms that can solve the cube very quickly. The most popular algorithms are: Kociemba's Two-Phase algorithm [12], Korf algorithm [2], and Rokicki algorithm [13], but none of them has been proved as God's algorithm [14]. In [15], an exhaustive computer search has shown every Rubik's Cube is solvable in 20 moves in half-turn metrics, or in 26 moves in the quarter-turn metrics.

At the moment, MIT Rubik's Cube solving robot developed by Ben Katz and Jared Di Carlo is the world record holder in category Fastest Rubik's cube solving robot with time 0.38s (2018). The robot uses software to identify the colors of each part within the cube to determine the cube's initial state. A set of cameras is firstly used to detect the initial state of each side of the cube. Then, the custom-built electronics and controls are used to control the high-speed motors. The software used to instruct the robot is based on the Kociemba's algorithm. The fastest human record is held by Australian Felix Zemdegs with a time 4.22 seconds (2018).

III. System Overview

A. Conceptual design of the system

Our system's flow diagram is shown in Fig. 1 and block diagram is in Fig. 2.

In the beginning, the initial state of the cube is detected by taking the image of each side of the cube. This process is semi-automated, since an operator is required to turn the cube six times over. The cube has six faces/sides (Up, Down, Left, Right, Back, Front). We use a simple USB web camera to take six images and then perform image analysis for color recognition of each cubie[2] in the image (the setup is presented in Fig 3.a). A simple algorithm implemented in C++ is developed for image analysis in RBG color space. The output of image analysis is a text file with six lines identifying the colors of cubies for each face. This process is shown in Fig. 3. Generated file is then forwarded to the algorithm for solving the cube.

[2]The word "cubie" is used to refer to one out of nine individual cubes on one side of the Rubik's cube [16].

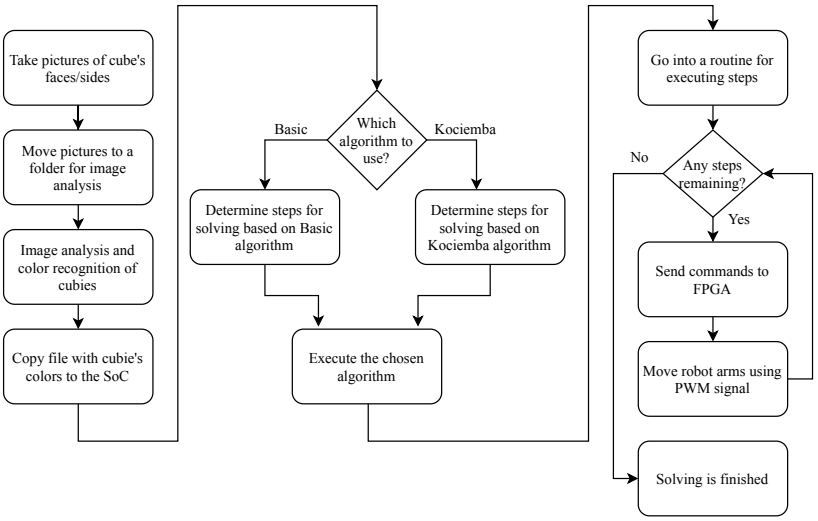


Fig. 1: The system's flow diagram.

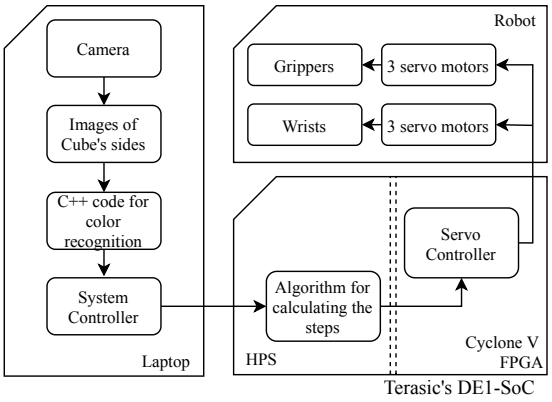


Fig. 2: The system's block diagram. The system includes basic principles needed for solving the real cube: a) detection of the cube's initial state implemented on the external laptop, b) generation of the solution for a randomly scrambled cube, which is executed on DE1-SoC HPS and c) performing moves on the robot, which is controlled by servo controller implemented in RTL on Cyclone V FPGA and executed on the actual robot platform.

Second, we implement two existing algorithms for solving the cube: 1) the Basic algorithm and 2) the Kociemba algorithm. In Section IV, we briefly overview these algorithms. Algorithms are implemented in C++ and cross-compiled for ARM-based hard processor system (HPS). The goal of the algorithm is to provide necessary steps for solving the cube.

And third, the steps are forwarded to the robot to start solving the cube. Steps are sent in a sequence to the servo controller which generates PWM signals and sends them to servo motors. Motors rotate wrists and grippers of the robot to manipulate the cube.

B. DE1-SoC development board

We used Terasic's DE1-SoC development board as a prototype board. The board combines ARM-based dual-core Cortex-A9 embedded core operating on 800MHz with FPGA fabric using a high-bandwidth interconnected backbone. FPGA device is Cyclone V SoC 5CSEMA5F31C6N with 85K logic elements. DE1-SoC includes 1GB (2x256Mx16) DDR3 on HPS and 64MB (32Mx16) SDRAM on FPGA. FPGA is connected to HPS through specialized bridges.

The proposed system is designed with Altera's Qsys integration tool, which is used to design digital hardware systems by combining various components (processors, memories, I/O interfaces, timers, etc). The tool helped us to connect HPS with custom-made RTL design into an operational system. This system enabled us to run Rubik's cube solving algorithms, control servo motors and communicate to and from the HPS. FPGA receives and sends data to and from the HPS using AXI (Advanced eXtensible Interface) buses, bridges, and Avalon master-slave devices. HPS and FPGA components each have their own pins and peripherals and they are not freely shared between the HPS and FPGA fabric.

Once the system was built, we installed the Linux Operating System Console to operate the HPS. The FPGA is registered as a connected programmable device in the OS. In our experiments, the Basic and Kociemba algorithms are cross-compiled for the ARM processor on an external laptop using the GNU compiler collection (GCC) tool-chain and then run on the HPS. Six servo motors are connected to FPGA's GPIO, and the servo controller is implemented in RTL and run on FPGA [17].

Additionally, the board was connected to a laptop via SSH protocol in order to send/receive information to/from the device.

C. Servo motors - rotating the cube

We used six Hitec HS-422 servo motors to manipulate the cube (Fig. 4 and Fig. 5). Motors are connected to Cyclone V FPGA's GPIO and controlled by the servo controller implemented in VHDL. A simple C program implemented in HPS tells the controller which wrist/gripper to move. Communication between the HPS and FPGA is implemented with



(a) The web camera taking an image of a cube's face/side.



(b) What the camera sees.

```
RYRGWYRRW
WYOOYGOBB
YOYGORYWG
WWBGRBGBW
BYBBGOYWR
ORGOBWGRO
```

(c) An example text file created after the image analysis.

Fig. 3: The process of detection of initial state of the cube. At first, USB web camera takes an image of every cube's face/side (a). Then, image by image is analyzed for color recognition of cubies colors. The image (b) shows an example of what the camera sees. Output of the analysis is a text file with 6 lines which is forwarded to the solving algorithm as the initial state of the cube.

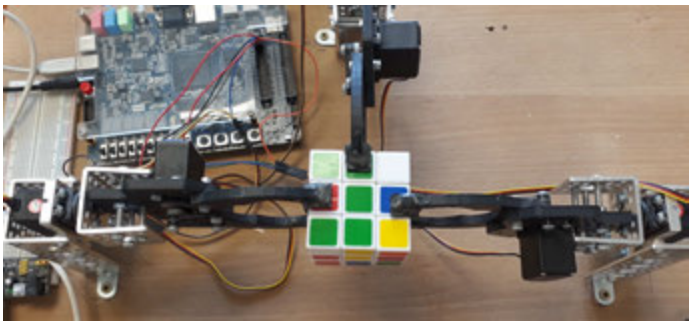


Fig. 4: The setup of the system. The system comprises of the Terasic's DE1-SoC, three grippers and three wrists. The gripper can open or close which makes possible the rotation of one side of the cube. The wrist performs actual rotation. Implementation details are provided in Fig. 5.

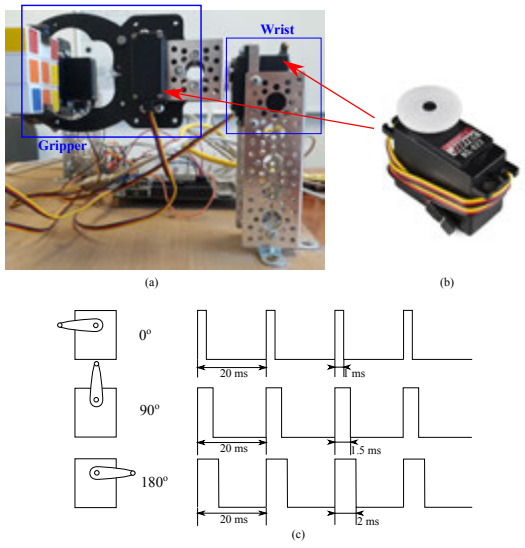


Fig. 5: The setup of one gripper and one wrist on the robot is presented in (a). Each component is driven by one servo motor shown in (b). Motors are controlled by the PWM signal, where positions of motor are related to the length of the PWM signal, as shown in (c).

PIO (Parallel Input Output) which defines memory locations for writing/reading to/from HPS or FPGA. In our system, a simple C program writes commands (i.e. which wrist/gripper to move) to predefined memory locations, and these values are then read by the servo controller on every clock cycle. The controller generates necessary PWM signals to control the motors based on the values read from memory locations.

Three different positions of servo motors are used, as shown in Fig. 5(c). The total time to complete one PWM cycle is 20 milliseconds. The signal is in high (on) state for 1 millisecond if the motor should turn left, 1.5 milliseconds if the motor should be centered, and 2 milliseconds if the motor should turn right. The signal is generated by using counters in VHDL code. Since the frequency of the clock on SoC is 50 MHz, the value to which the counter should count is calculated as a ratio of the clock frequency and frequency of the needed signal.

By controlling servo motors we can rotate wrists and grippers, and they manipulate the cube.

D. The robot platform

The overall look of the system is shown in Fig 4. Besides servo motors, wrists and grippers, we use servo plates and aluminum channels to connect all parts and fix the robot to a table.

We use three grippers and three wrists with a total of six servo motors. It is important to point out that, due to the stability of the cube, at least two grippers should hold the cube at any moment. If only one gripper holds the cube, it could easily slip from grippers. A breadboard is used for easier connection of servo motors to FPGA's GPIO. Servo motors were powered up from the breadboard, which was connected to a laptop via USB cable.

We use the Terasic DE1-SoC development board to implement the system. The LAN cable is used to connect the SoC board to a laptop, so that the SoC board can get an IP address, and that the file with colors of cubies can be transferred to the SoC board. There is also a UART-to-USB cable, which is used for serial communication between the SoC board and the laptop. The UART side was plugged into the SoC board, while the USB side is plugged into a laptop. This is used so that the algorithm (which runs on the ARM processor) can be started by the laptop. The standard USB camera is used for

taking pictures of cube's sides, which are later processed in an image analysis algorithm.

These are the main components (besides the board itself and the laptop) that were used to run the algorithm and solve the cube.

IV. Brief overview of algorithms used in our robot

Two algorithms were used in this paper for solving Rubik's cube: 1) the Basic algorithm and 2) the Kociemba algorithm. In this section, we briefly describe basic concepts of these algorithms.

A. The Basic algorithm

This algorithm was first discovered in 1979 by an early adopter of the Rubik's cube, David Singmaster [18]. It consists of 6 phases. The expected output after each phase is the following:

- Form a cross on the top face of the cube (i.e. white cross).
- Correctly position top face corner pieces (i.e. white corners).
- Correctly place edge pieces on the "middle" layer of the cube (middle layer).
- Form a cross on the new top face of the cube (yellow corners)
- Place yellow corners on right spots
- Solve yellow corners

The state of the cube, as it goes through these phases, is shown in Fig. 6.

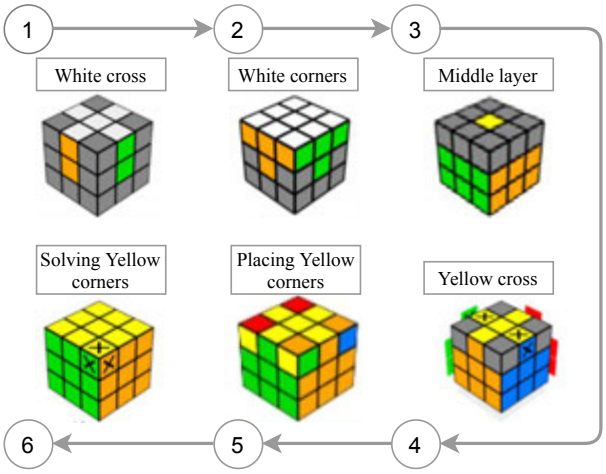


Fig. 6: Illustration of the Basic algorithm's phases. Six major milestones are represented, where the names of phases represent the shapes of cubies. Many more steps are required in-between phases.

B. The Kociemba algorithm

This algorithm is an improvement of Thistlethwaite's algorithm created by Herbert Kociemba in 1992. The basic idea is to divide the possible configurations of the cube into groups. Kociemba defined three groups:

- G0 = {U, D, L, R, F, B}
- G1 = {U, D, Lx2, Rx2, Fx2, Bx2}
- G2 = {1}

Letters U, D, L, R, F and B represent moves that can be executed on the cube. U stands for Up, i.e. moving the upper side of the cube in clockwise direction. Other moves are: D - down side, L - left side, R - right side, F - front side, B - back side.

The principle of this algorithm is to start with group G0, and then move some target cubies to their expected positions by using only moves from this group. When the cube is in group G1, then cubies are moved using only allowed moves from group G1, to arrive to the final state, group G2.

G0 group contains all possible configurations of the cube. G1 group contains all configurations to which we can arrive if we take solved cube, and apply only moves U, D, Lx2, Rx2, Fx2, Bx2. Group G2 has only one state, solved cube.

V. Results

In Table I we show results in terms of: a) the number of steps needed to solve the cube and b) the total execution time of the robot for solving the cube. The results are presented for eight random and two selected cube configurations (No. 1 and No. 10). The results are presented for the two algorithms explained in Section IV. It is important to note that many more configurations are tested, but here we present only ten configurations.

The ratio of an average number of moves between the Basic and the Kociemba algorithm is 5.42, and the ratio of execution times is 5.34, both in favor of the Kociemba algorithm, as it was expected. It can be said that the Kociemba algorithm on average has 5.42 times fewer moves than the Basic algorithm, and the Basic algorithm needs 5.34 times more time to solve the cube than the Kociemba algorithm.

The Basic algorithm gives on average 144.3 steps for solving the cube, and the robot needs 14 minutes and 9 seconds to solve the cube using this algorithm. The Kociemba algorithm, on the other side, gives an average of 26.6 steps for solving the cube, and the robot needs 2 minutes and 39 seconds to solve the cube.

The result was expected, but the question was how much indeed was the Basic algorithm slower than the Kociemba algorithm? Now, the numeric results have been obtained. On average, it can be said that the Kociemba algorithm is 5 to 5.5 times faster than the Basic algorithm.

In the first combination (No. 1 in Table I), the first layer (white cross and corners) was completely solved, so the Basic algorithm started from the third phase. Here, the Kociemba algorithm showed that this fact does not have significance for the algorithm. The number of steps for the Kociemba algorithm was close to the average number of steps.

The last configuration (No. 10 in Table I) is also different from others. In this configuration, the solved cube was taken, and 10 random moves have been made. The cube looked as it was randomly twisted. It can be seen that the Kociemba algorithm successfully identified these 10 moves, and for solving the cube, just returned those moves in reverse order.

	a) Number of steps		b) Total execution time	
No.	Basic	Kociemba	Basic	Kociemba
1	81	25	08:22	02:40
2	141	33	12:29	03:26
3	134	30	11:51	02:36
4	152	32	15:07	03:11
5	177	27	17:03	02:57
6	162	28	17:39	02:12
7	175	30	17:10	03:06
8	152	20	16:00	01:38
9	103	31	09:00	03:27
10	166	10	16:45	01:16
Average	144.3	26.6	14:09	02:39
Ratio	5.42		5.34	

TABLE I: The table shows a) the number of steps the algorithm needs to solve the configuration of the cube and b) the total execution time for solving the cube for each algorithm. We present measurements for ten different configurations.

On the other hand, the Basic algorithm was quite worse here, because it needed 166 moves to solve the cube.

The robot uses three grippers for solving the cube (Fig. 4). Two side grippers are used for Left and Right moves and the middle gripper is used for all other moves (U, D, F, B). So, for both algorithms, most time is consumed when there are sequences like U $\rightarrow$ D, or F $\rightarrow$ B and vice versa. Because then, the robot needs first to set up the proper side, and then do the turn. Both algorithms spend quite a lot of time on these side setups.

The advantage of the Kociemba algorithm is that usually, the last 10 steps are 5 pairs of same steps, i.e. U $\rightarrow$ U $\rightarrow$ R $\rightarrow$ R $\rightarrow$ F $\rightarrow$ F, etc. This means that the side needs to be set up only for the first step from the pair, and for the next step, everything is already prepared.

The advantage of the Basic algorithm is that, in some phases, it has a sequence such as: R $\rightarrow$ D $\rightarrow$ R' $\rightarrow$ D' $\rightarrow$ R $\rightarrow$ D $\rightarrow$ R' $\rightarrow$ D' etc. For these kinds of sequences, the side needs to be set up on the middle hand only for the first D move. For all other moves, the D side is already in the proper place.

To summarize, both algorithms have advantages in some points regarding this robot configuration, but we can say that, on average, the Kociemba algorithm has proven to be a bit more than 5 times better than the Basic algorithm, when considering both number of moves and solving time.

VI. Conclusion

Advancement of the FPGA technology has led to the development of the FPGA-based System on Chip, including FPGA-based embedded processor SoCs. These devices offer hundreds of GFOPs with maximum power efficiency, and this is why they became implementation platform of choice between hardware and software. When an algorithm is too complex to be implemented in RTL, a decision has to be made about what should be implement in RTL, and what should be executed on the HPS. Systems can be seemingly integrated with efficient circuit switching interconnections, thus offering hardware-like performance and software-like flexibility. In recent years, the knowledge about both hardware and software design is becoming a necessity for system designers, and this is where the Hardware/Software Co-design comes into existence.

In this paper, we present two approaches for solving Rubik's cube with Hardware/Software Co-design. Our goal was to design an in-expensive educational board for the demonstration of different concepts and principles of Hardware/Software Co-design. We built a robot platform that supports all the phases needed to solve a random configuration of the cube. We tested the platform with two different algorithms that were implemented on HPS, and we implemented a custom-made servo controller in RTL. Our tests indicate that the platform can successfully solve random configurations of the cube that we tested. We can conclude that the Kociemba algorithm is about 5 times better than the Basic algorithm when considering the number of moves and total execution time.

References

[1] E. McManus, "Mathematical understandings of a Rubik's Cube," *Honors Theses, Salem State University*, 2018.

[2] R. E. Korf, "Finding optimal solutions to Rubik's Cube using pattern databases," in *AAAI/IAAI*, pp. 700–705, 1997.

[3] "Solving a Rubik's Cube in record time." MIT News Office, March 2018.

[4] A. Akagic, E. Buza, R. Turcinhodzic, H. Haseljic, N. Hiroyuki, and H. Amano, "Superpixel accelerator for computer vision applications on arria 10 soc," in *2018 IEEE 21st International Symposium on Design and Diagnostics of Electronic Circuits & Systems (DDECS)*, pp. 55–60, IEEE, 2018.

[5] A. Akagic and H. Amano, "A study of adaptable co-processors for cyclic redundancy check on an fpga," in *2012 International Conference on Field-Programmable Technology*, pp. 119–124, IEEE, 2012.

[6] "List of cube solving robots." speedsolving.com Wiki, accessed in February 2020.

[7] S. L. Lu, M. Huang, and F. R. Kong, "The design of a Rubik's Cube robot," in *Advanced Materials Research*, vol. 709, pp. 432–435, Trans Tech Publ, 2013.

[8] F. Ďurovskỳ, "Robust Rubik's Cube detection using Hough transform and advanced clustering functions," in *Applied Mechanics and Materials*, vol. 613, pp. 253–264, Trans Tech Publ, 2014.

[9] S. Liu, D. Jiang, L. Feng, F. Wang, Z. Feng, X. Liu, S. Guo, B. Li, and Y. Cong, "Color recognition for Rubik's Cube robot," *arXiv preprint arXiv:1901.03470*, 2019.

[10] T. Li, W. Xi, M. Fang, J. Xu, and M. Q.-H. Meng, "Learning to solve a Rubik's Cube with a dexterous hand," *arXiv preprint arXiv:1907.11388*, 2019.

[11] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, *et al.*, "Solving Rubik's Cube with a robot hand," *arXiv preprint arXiv:1910.07113*, 2019.

[12] H. Kociemba, "Close to God's algorithm," *Cubism for Fun*, vol. 28, no. April, pp. 10–13, 1992.

[13] T. Rokicki, "Twenty-five moves suffice for Rubik's Cube," *CoRR*, vol. abs/0803.3435, 2008.

[14] C. Aksop, "God's algorithm for Rubik's Cube an integer programming approach," in *International Mathematical Forum*, vol. 4, pp. 2217–2222, Citeseer, 2009.

[15] T. Rokicki, H. Kociemba, M. Davidson, and J. Dethridge, "God's number is 20, 2010," *URL: http://www.cube20.org*, 2015.

[16] K. M. Dismuke, "Mathematics of the Rubik's Cube," *Honors Thesis, Loyola Marymount University*, 2016.

[17] A. Akagic and H. Amano, "An FPGA implementation of CRC slicing-by-N algorithms," *RECONF: IEICE technical report*, vol. 110, no. 319, pp. 19–24, 2010.

[18] D. Singmaster, *Notes on Rubik's magic cube*. Hillside, New Yersey, USA: Enslow Publishers, 1981.

A Methodology for Frequency-Measurement Characterization Based on Increment of Input Signal Phase

Andrey N. Serov
National Research University "Moscow Power Engineering Institute",
Diagnostic Information Technologies Department, Moscow, Russian Federation
SerovAN@mpei.ru

Abstract - **It is known that the signals of modern power networks are non-sinusoidal, and in this case frequency measurement methods for sinusoidal signals cannot be effectively applied. The widely applied measurement method, based on cross-zero technique, requires a long measurement time. In a number of problems of power quality parameters measurement, this drawback is significant and necessitates the application of a faster measurement method, which allowing measurements for nonsinusoidal input signals. In this paper discusses a frequency measurement method based on determination of increment of input signal phase. The phase value is determined by applying of the discrete Fourier transform (DFT). An analysis of influence of additive, multiplicative and nonlinear components of the samples error to the frequency measurement error for the considered measurement method is performed. The influence of the measurement system parameters on the algorithmic error component is investigated. Recommendations to significantly reduce of the error have been developed. Analytical relationships that allows to evaluate the algorithmic component of the measurement error are obtained. A simulation model of the considerate measurement method is developed by Matlab software package. The reliability of the analytical relationships is confirmed by the coincidence of the results which are obtained by analytical expressions and the results of simulation modeling at check points.**

Keywords - fast Fourier transform, phase spectrum, simulation modeling, measurement error, frequency deviation, polyharmonic signal

I. INTRODUCTION

Currently, there are a many digital methods for the frequency measurement of power network signals. All measurement methods can be divided into several categories. The main distinguishing feature is a type of waveform of the input signal. Real power networks have a number of features, including: the presence of harmonics, the presence of noise, the presence of a sinusoidal or rectangular flicker.

A number of measurement methods [1]-[2] was developed for the case of sinusoidal input signals, which limits it application in real power networks. For the successful application of such methods, a low-pass (LPF) or band-pass filter is required, which allows to eliminate the above features and transform the input polyharmonic signal to a sinusoidal form. But this significantly increases the dynamic component of the error. The cross-zero technique method [1] and a method based on spectral analysis [4]-[5] can be applied for nonsinusoidal input signals. However, a significant measurement time is required to obtain a measurement result.

One of the promising measurement methods is a method based on determining the phase increment of a signal (MBDIP) [1]-[3]. The essence of this method consists in sequentially determining the phase spectrum of the signal and calculating its increment in time, which is proportional to the input signal frequency. The values of the phase spectrum can be obtained by applying the spectral analysis. The main idea of this method is presented in [1]. It has been shown in [1] that this method can be used to measure the frequency of both sinusoidal and polyharmonic signals. However, the issues of the influence of noise, flicker on the measurement error have been researched insufficiently. The problem of the influence of the instrumental components (additive, multiplicative, nonlinearity and quantization) of the samples error and the "spectrum leakage" effect on the frequency measurement error also needs additional research.

II. METHOD OF THE FREQUENCY MEASUREMENT, BASED ON DETERMINATION OF THE PHASE INCREMENT

A. Description

This method is based on determination of the phase increment [1] of the input signal during the measurement process. It is assumed that the phase and frequency values of the input signal are proportional. For the case of the input sinusoidal signal, the phase increment can be obtained by determining the samples of the same signal, delayed relative to each other by a certain number of samples:

$$\begin{cases} u[n] = U_m \sin\left(2\pi n f_1 T_S + \varphi_U\right) = U_m \sin\left(\varphi[n]\right), \\ u[n-L] = U_m \sin\left(\varphi[n-L]\right) = U_m \sin\left(\varphi[n] - \Delta\varphi[n]\right); \end{cases} \quad (1)$$

where $u[n]$ – input signal; U_m – amplitude; T_S – sampling time; L – integer, number of delay of the samples.

Expressions (1) are analytical dependences for the same signal (input voltage), but with a time delay of L samples. The value of the phase increment is determined by the expression:

$$\varphi[n,L] = \varphi[n] - \varphi[n-L] = 2\pi L f_1 T_S. \quad (2)$$

The input signal frequency f_1 can be obtained from the following expression:

$$f_1[n] = \varphi[n,L]/(2\pi L T_S). \quad (3)$$

To determine the phase increment in practice, the following approaches are applied:

– application of orthogonal filters;

– application of discrete Fourier transform (DFT).

The number of samples L should be selected from the condition of ensuring the minimum measurement error. To reduce the random measurement error component, the frequency instantaneous values obtained by expression (3) can be processed by a low-pass filter, by polynomial approximation function [1] or by the least squares technique [1].

B. Application of Orthogonal Filters

Orthogonal filters [2]-[3] are a pair of filters with identical magnitude characteristics and phase characteristics with difference equal to $\pi/2$. The output signals of the orthogonal filter differ in phase by $\pi/2$. For the case of a sinusoidal signal with a frequency appropriate to the filter passband, the filter output:

$$\begin{cases} u_S[n] = U_m K_S(f_1)\sin(2\pi f_1 T_S n + \alpha + \varphi_S(f_1)), \\ u_C[n] = U_m K_C(f_1)\sin(2\pi f_1 T_S n + \alpha + \varphi_C(f_1) + 0.5\pi); \end{cases} \quad (4)$$

where $u_S[n]$, $u_C[n]$ – orthogonal filter output signals "S" and "C"; $K_S(f)$, $K_C(f)$ – magnitude response of the orthogonal filter for frequency f; $\varphi_S(f)$, $\varphi_C(f)$ – orthogonal filter phase response for frequency f.

Orthogonal filter parameters are selected so that it magnitude response are identical and the phase response are differ by $\pi/2$:

$$\begin{cases} K_S(f) = K_C(f), \\ \varphi_S(f) = \varphi_C(f) - 0.5\pi. \end{cases} \quad (5)$$

In the case of sinusoidal input signals, the filter bandwidth is selected based on the range of the input signal frequency deviation. For a polyharmonic signal, the selection of the filter bandwidth is by a similar way. Incomplete suppression of harmonics, as well as failure to fulfill conditions (5), results to the appearance of an additional measurement error.

As a result, the phase can be obtained by determining the ratio of the output signals of the orthogonal filter:

$$\varphi[n] = \arctan(-u_S[n]/u_C[n]). \quad (6)$$

The application of orthogonal filters is justified only for measuring frequency transducers at which DFT is not implemented and its separate implementation is required. For example, for power quality measurement devices, DFT already implemented to determine most of the signal parameters and the implementation of the frequency measurement method does not require a separate implementation of DFT.

C. Application of DFT

The DFT technique allows to determine the values of both amplitude and phase spectrum. In the problem under consideration, only one spectral component is required – corresponding to the fundamental spectral component. Since an array of phase measurement samples are required, the so-called sliding DFT can be applied. The sliding DFT allows to obtain a samples sequence of the input signal complex spectrum by applying a DFT of a fixed length (N samples), with a sequential increase in the number of the initial signal sample:

$$\dot{U}[n,k] = \sum_{i=n-N+1}^{n} u[i] e^{-j\frac{2\pi nk}{N}}. \quad (7)$$

As a result of applying the sliding DFT, the spectrum samples corresponding to the fundamental spectral component have a phase shift determined by expression (2). Expression (2) is applicable if the window for performing the DFT is shifted by a fixed number of L samples. If the duration of the DFT window (the duration of the measurement) is not a multiple of the input signal period, the "spectrum leakage" effect occurs, resulting to the occurrence of measurement error.

The value of the phase spectrum from the complex spectrum can be obtained by applying the expression:

$$\varphi[n,k] = \arctan\left(\frac{\mathrm{Im}(\dot{U}[n,k])}{\mathrm{Re}(\dot{U}[n,k])}\right). \quad (8)$$

The analysis of expression (8) from the point of view of the measurement error of the phase spectrum is difficult. To simplify obtaining the final expression of the measurement error of the phase spectrum, let's represent the DFT (see expression (7)) as the sequential execution of the operations of multiplication by a complex exponential signal ($e^{-j2\pi nk/N}$) and the sequential application of a moving average filter (averaging filter).

III. Methodical Measurement Error of the Phase in Case of Application of DFT

For the case of a sinusoidal input signal, the result of multiplying of the input signal and the complex exponent:

$$u[n]r[n] = 0.5U_m\left(e^{j(\Delta\omega n + \alpha - 0.5\pi)} - e^{-j((\omega_1+\omega_0)n + \alpha + 0.5\pi)}\right), \quad (9)$$

where $r[n] = e^{-j2\pi nk/N} = e^{-jn\omega_0}$ – reference (complex) signal; $\omega_0 = 2\pi k/N$ – angular frequency of the exponential signal, relative to the sampling frequency; ω_1 – angular frequency of the input sine wave relative to the sampling frequency.

The application of low-pass filter (moving average filter) in the DFT operation allows to pick out the low-frequency spectral component of the signal $u[n]r[n]$, which is proportional to the component of the complex spectrum of the input signal with frequency ω_1:

$$\dot{y}[n] = 0.5U_m\left(H_1 e^{j\eta_1} - H_2 e^{-j\eta_2}\right), \quad (10)$$

where H_1, H_2 – the magnitude response of the moving average filter for $(\omega_1 - \omega_0)$ and $(\omega_1 + \omega_0)$ frequencies; $\eta_1 = \Delta\omega n + \alpha - 0.5\pi + \varphi_1$; $\eta_2 = (\omega_1 + \omega_0)n + \alpha + 0.5\pi + \varphi_2$; φ_1, φ_2 – filter phase response for $(\omega_1 - \omega_0)$ and $(\omega_1 + \omega_0)$.

The magnitude response of the moving average filter:

$$H(\omega) = |H(j\omega)| = \frac{1}{N}\frac{\sin(0.5N\omega)}{\sin(0.5\omega)}, \quad (11)$$

where $N = round(T_A / T_S)$ – filter order.

By simplifying expression (10), we obtain:

$$\dot{y}[n] = 0.5U_m H_1 e^{j\eta_1}\left(1 - (H_2 / H_1)e^{-j(\eta_2+\eta_1)}\right). \quad (12)$$

For brevity, we introduce the notation $\xi = H_2 / H_1$ and present in the last expression a complex number in the trigonometric form:

$$\dot{y}[n] = \frac{U_m H_1 e^{j\eta_1}}{2}\left(1 - \xi\cos(\eta_2 + \eta_1) + j\xi\sin(\eta_2 + \eta_1)\right) \quad (13)$$

As a result, we'll get:

$$\arg(\dot{y}[n]) = \eta_1 + \arctan\left(\frac{\xi\sin(\eta_2 + \eta_1)}{1 - \xi\cos(\eta_2 + \eta_1)}\right). \quad (14)$$

Given that $\xi << 1$, as well as expressions for η_1 and η_2, imagine (14) in a more convenient form for analysis:

$$\arg(\dot{y}[n]) = \Delta\omega n + \alpha - 0.5\pi + \varphi_1 + \xi\sin(2\omega_1 n + \theta), \quad (15)$$

where $\theta = 2\alpha + \varphi_1 + \varphi_2$.

As noted earlier, the phase increment of the input signal is of interest, which is determined from (14) as:

$$\varphi[n, L] = \arg(\dot{y}[n]) - \arg(\dot{y}[n - L]) - \Delta\varphi[n, L], \quad (16)$$

where $\Delta\varphi[n, L]$ – the absolute error of the phase increment due to the finite stopband attenuation of LPF:

$$\Delta\varphi[n, L] = 2\xi\sin(\varphi[n, L])\cos(2\omega_1 n + L\omega_1 + \theta). \quad (17)$$

It can be seen from formula (17) that the phase increment error has a time-variable component and does not contain a constant component. The phase increment error oscillates at twice the input signal frequency ($2\omega_1$). This error component can be reduced by applying an additional LPF.

IV. Instrumental Error of Phase Measurement in Case of Application of DFT

A. General

In addition to the algorithmic error of phase and phase increment, there are error components associated with the imperfection of the chips and implementation components. For the signal samples, there are additive, multiplicative, quantization errors and nonlinearity [6]:

$$u[n] = u_{id}[n] + \Delta_{ADD} + u_{id}[n]\delta_{MUL} + \Delta_L[n] + \Delta_Q[n], \quad (18)$$

where Δ_{ADD} – additive error component; δ_{MUL} – multiplicative error component; $\Delta_L[n]$ – nonlinearity; $\Delta_Q[n]$ – quantization error; $u_{id}[n]$ – input signal true value.

Each of the considered error components affects to the complex spectrum of the signal $u[n]$. This results to errors in both the amplitude and phase spectrum.

B. Additive Component Impact

The additive component of the error is a DC signal. In the frequency domain, its influence is manifested in the appearance of an additional spectral component at zero frequency. Since in the problem to be solved, we are interested in the phase increment of the fundamental spectral component, we can conclude that the additive error component does not affect to the frequency measurement error by the considerate method.

C. Multiplicative Component Impact

The multiplicative error component is proportional to the value of the input signal. The influence of this component on the complex spectrum is expressed in a change in the values of nonzero spectral components. Additional (imaginary) spectral components do not appear because the signal waveform does not change. The multiplicative error results to an measurement error of the amplitude spectrum, which is seen when substituting expression (18) for the multiplicative component into expression (7). The relative measurement error of the amplitude spectrum can be calculated by expression:

$$\delta|\dot{U}[n, k]| = \delta_{MUL}. \quad (18)$$

The error of the phase spectrum for the multiplicative error component is equal to zero, because the signal waveform does not change as a result of the influence of this component. When substituting expression (18) for the multiplicative component into expression (7), it can be seen that the real and imaginary spectral components change by the same coefficient. Therefore, when determining the ratio of the imaginary part to the real part in determining the phase, the influence of the multiplicative part is eliminated.

D. Nonlinearity Component Impact

The presence of a nonlinearity (linearity error) results to a change in the signal waveform, which results in the frequency domain in deviation of amplitude values of nonzero spectral components and the appearance of

additional (imaginary) spectral components. The forms of ADC nonlinearity are very diverse, but can be divided into two categories: relatively smooth [7]-[8] (can be approximated by a relatively low order polynomial function) and pseudo-random (random functions are well-suit for the mathematical description) [9]-[10].

A relatively smooth form of nonlinearity is specific of Sigma-Delta ADC architecture [6]. The pseudo-random nonlinearity form is specific for sequential approximation register (SAR) and pipeline ADC [10]. Simulation results shows that the largest deviations of the amplitude and RMS are observed for the following nonlinearity form (the "worst case" representation) [8]-[9]:

$$\begin{cases} \Delta_L(t) = INL \cdot q, \ for\ u(t) \geq 0, \\ \Delta_L(t) = -INL \cdot q, \ for\ u(t) < 0; \end{cases} \quad (19)$$

where *INL* – ADC integral nonlinearity; q – ADC least significant bit (LSB).

Figure 1 shows the dependence of the ADC nonlinearity for the "worst case" representation and for the ADC nonlinearity representation as a pseudorandom function. Figure 2 shows the dependence of the phase measurement error on the sampling frequency of the input signal for the case of nonlinearity which is represented by a pseudo-random function and the "worst case" method (see Fig. 1 and expression (19)). The case of the absence of the "spectrum leakage" effect [9] is considered: when the measurement time (equals to an integer number of input signal samples) is selected as a multiple of the input signal period. The results of Fig 1 are obtained by the simulation modeling performed in Matlab package.

When obtaining the dependence on Fig. 2, an 1000 different forms of ADC nonlinearity were considered, the result shown in Fig. 2 is the worst in terms of phase error. The input signal is selected sinusoidal with a unit amplitude, frequency – 50 Hz. The measurement time is equal to 0.02 sec (one nominal period of the input signal). The integral nonlinearity is 1 LSB. The ADC resolution is 16 bits. Figure 2 shows that the phase error in the nonlinearity representation by a pseudo-random function is negligible. In the case of nonlinearity being represented by dependence (19), the phase measurement error in time remains constant; therefore, the phase increment error is also equal to zero. The influence of the ADC nonlinearity is manifested only in conjunction with the "spectrum leakage" effect. Figure 3 shows the dependence of the phase measurement error on the sampling frequency for the case when the "spectrum leakage" effect occurs. It can be seen that for the pseudo-random representation of nonlinearity, the error takes even larger values than for the "worst case" method. The simulation parameters are selected similar to used to obtain the results of Fig. 2.

An approach based on the application of a polynomial function can be applied to estimate the error of phase measurement. Simulation results shows that the measurement error is proportional to the *INL* value, regardless of the sampling frequency. For this reason, the following approximation polynomial can be used for phase error estimation:

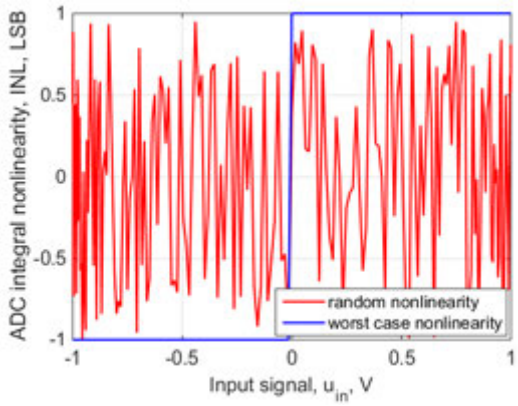


Figure 1. The dependence of the ADC nonlinearity on the input signal level for the nonlinearity representeon by "worst case" method and pseudo-random function

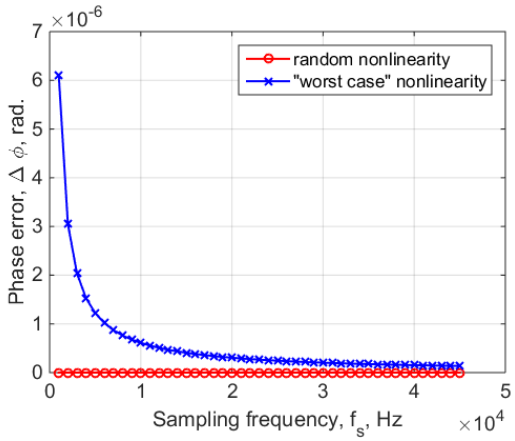


Figure 2. The dependence of the phase measurement error on the sampling frequency. The error is caused by the ADC nonlinearity. Nonlinearity is represented by the "worst case" method and pseudo-random function. "Spectrum leakage" effect is absent

$$\Delta\varphi_L(f_S) = q \cdot INL \cdot \left(A_0 + A_1 f_{S,1}^{-1} + A_2 f_S\right), \quad (20)$$

where A_0, A_1, A_2 – approximation polynomial coefficients; q – ADC least significant bit (*LSB*).

Dependence (20) and the coefficients A_0, A_1, A_2 of the approximating polynomial were obtained by performing simulation modeling in Matlab. To obtain the values of the coefficients A_0, A_1, A_2, a solution of a system of three equations is required. The experimental data obtained for the sampling frequencies at the beginning, middle, and end of considered sampling frequency range were used:

$$\begin{cases} q \cdot INL \cdot \left(A_0 + A_1 f_{S,1}^{-1} + A_2 f_{S,1}\right) = L_1 \cdot \Delta\varphi_L(f_{S,1}), \\ q \cdot INL \cdot \left(A_0 + A_1 f_{S,2}^{-1} + A_2 f_{S,2}\right) = L_2 \cdot \Delta\varphi_L(f_{S,2}), \\ q \cdot INL \cdot \left(A_0 + A_1 f_{S,3}^{-1} + A_2 f_{S,3}\right) = L_3 \cdot \Delta\varphi_L(f_{S,3}); \end{cases} \quad (21)$$

where $f_{S,1}$, $f_{S,2}$, $f_{S,3}$ – sampling frequencies used to define reference points for determining the coefficients of the approximating polynomial (20); $\Delta\varphi_L(f_{S,1})$, $\Delta\varphi_L(f_{S,2})$, $\Delta\varphi_L(f_{S,3})$ – maximal values of phase measurement error, simulation results; L_1 = 1.1, L_2 = 1.2, L_3 = 1.0 – accordingly, the values of the coefficients used to ensure metrological reserve.

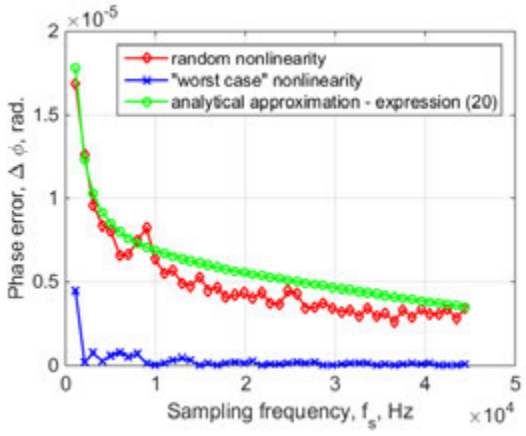


Figure 3. The dependence of the phase measurement error on the sampling frequency. The error is caused by the ADC nonlinearity. Nonlinearity is represented by the "worst case" method and pseudo-random function. "Spectrum leakage" effect is present

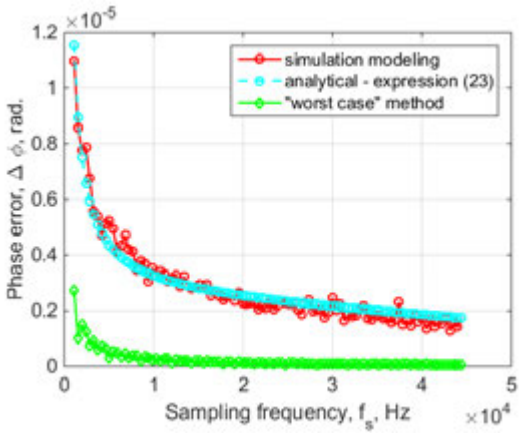


Figure 4. The dependence of the phase measurement error on the sampling frequency. Measurement error is a result of ADC quantization noise. Quantization noise is represented by the "worst case" method and pseudo-random function. "Spectrum leakage" effect is present

The values of approximation coefficients A_0, A_1, A_2 were obtained for the sampling frequency range of 1 kHz ÷ 45 kHz:

$$\begin{cases} A_0 = 0.0205886760684043, \\ A_1 = 42.2429778157319, \\ A_2 = -2.25709666496865 \cdot 10^{-7}. \end{cases} \quad (22)$$

In the case of estimating the phase increment error, the doubled value of the error which is obtained by expression (20) can be used.

E. The Quantization Error Impact

Figure 4 shows the dependence of the phase measurement error on the sampling frequency as a result of ADC quantization error. The dependence is based on the results of simulation modeling performed in Matlab software package. When formation of the dependence, 1000 different initial phases of the input signal were considered and the result shown in Fig. 4 is the worst in terms of phase measurement error. The simulation parameters are selected similar to used to obtain the results of Fig. 2.

As in the case of nonlinearity, the phase measurement error occurs only in the case of appearance of the "spectrum leakage" effect. To evaluate the phase measurement error, the same approach can be used as in the case of ADC nonlinearity:

$$\Delta\varphi_Q = 0.5 \cdot q \cdot \left(B_0 + B_1 f_{S,1}^{-1} + B_2 f_S\right), \quad (23)$$

where B_0, B_1, B_2 – approximation polynomial coefficients.

As for nonlinearity, dependence (23) and the coefficients of the approximating polynomial B_0, B_1, B_2 were obtained by performing simulation modeling in Matlab software. To determine the value of the coefficients, an equation similar to (21) was applied for the simulation results obtained for the case of quantization error. Instead of the *INL* parameter, 0.5 should be used. The values of approximation coefficients were obtained for the sampling frequency range of 1 kHz ÷ 45 kHz:

$$\begin{cases} B_0 = 0.00808216446557064, \\ B_1 = 33.1282715412236, \\ B_2 = -6.96065393079313 \cdot 10^{-8}. \end{cases} \quad (24)$$

In the case of estimating the error of the phase increment, the doubled error value which is obtained by expression (24) can be used.

V. Influence of Harmonics, Noise and Flicker on the Frequency Measurement Error

A. Harmonics Impact

The influence of the harmonics of the input signal on the frequency measurement error is caused by the "spectrum leakage" effect. To assess the effect of a single harmonic on the phase measurement error, expressions (10) and (11) can be used when replacing the parameter ω_1 to the angular frequency of considered spectral component. The analysis of the measurement error in an analytical form for the case of a polyharmonic (containing a large number of harmonics) signal is difficult to obtaining the final expression and its further practical application. Table 1 shows the frequency measurement error obtained as a result of simulation modeling in Matlab. A polyharmonic voltage is considered, the harmonic coefficients of which correspond to the maximal values represented in EN 50160:2010. The use of a post-filter for frequency measurement can significantly reduce the measurement error. Fourth order Butterworth IIR filter is applied as a post-filter. The stopband frequency is 90 Hz, the minimum stopband attenuation is 60 dBl. The total measurement time is 0.2 sec.

B. Noise Impact

When performing simulation, the white noise was considered, with RMS equal to 10 % of the input signal RMS. The effect of noise on the frequency measurement error is similar to the influence of harmonics. The application of a post-filter can significantly reduce the frequency measurement error by the considered method. To simulate white noise, the built-in "*randn*" function of the Matlab software package was used.

TABLE I. COMPARATIVE ANALYSIS OF FREQUENCY MEASUREMENT METHODS (MEASUREMENT ERROR IS REPRESENTED IN PERCENT)

Measurement method	Relative frequency deviation, δ_f, Hz								
	-0.06	-0.045	-0.03	-0.015	0.00	0.015	0.03	0.045	0.06
	Sinusoidal signal								
MBDIP[a]	6.4	4.5	3.0	1.5	0.0	1.5	2.9	4.3	6.0
MBDIP+LPF (Butterworth 4-th order)	**$4.2\ 10^{-2}$**	**$3.4\ 10^{-2}$**	**$2.4\ 10^{-2}$**	**$1.4\ 10^{-2}$**	**$4.8\ 10^{-3}$**	**$1.2\ 10^{-2}$**	**$1.5\ 10^{-2}$**	**$2.7\ 10^{-2}$**	**$3.5\ 10^{-2}$**
MBDIP+least squares	0.28	0.58	0.46	0.24	0.0	0.30	0.36	1.1	1.5
	Polyharmonic signal								
MBDIP	6.4	15	14	1.6	0.00	13	2.9	4.3	14
MBDIP+LPF (Butterworth 4-th order)	**$4.5\ 10^{-2}$**	**$3.0\ 10^{-2}$**	**$2.1\ 10^{-2}$**	**$1.2\ 10^{-2}$**	**$4.8\ 10^{-3}$**	**$1.4\ 10^{-2}$**	**$2.0\ 10^{-2}$**	**$2.8\ 10^{-2}$**	**$3.6\ 10^{-2}$**
MBDIP+least squares	0.60	0.47	0.38	0.19	0.00	0.40	0.25	1.0	1.5
	Sinusoidal signal+noise								
MBDIP	8.6	15	11	7.4	7.4	12	9.7	11	12
MBDIP+LPF (Butterworth 4-th order)	**$9.5\ 10^{-2}$**	**0.16**	**$1.2\ 10^{-2}$**	**$5.6\ 10^{-2}$**	**$7.8\ 10^{-2}$**	**0.11**	**0.15**	**$4.5\ 10^{-2}$**	**0.22**
MBDIP+least squares	0.29	1.4	0.36	0.02	0.29	0.81	0.70	1.3	2.1
	Sinusoidal signal+flicker (sinusoidal)								
MBDIP	6.4	4.5	3.0	1.5	0.2	1.5	2.9	4.3	6.0
MBDIP+LPF (Butterworth 4-th order)	**$3.8\ 10^{-2}$**	**$3.6\ 10^{-2}$**	**$2.6\ 10^{-2}$**	**$1.6\ 10^{-2}$**	**$5.7\ 10^{-3}$**	**$1.1\ 10^{-2}$**	**$1.4\ 10^{-2}$**	**$2.5\ 10^{-2}$**	**$3.1\ 10^{-2}$**
MBDIP+least squares	0.28	0.58	0.45	0.24	$1.6\ 10^{-2}$	0.30	0.38	1.1	1.5
	Sinusoidal signal+flicker (rectangular)								
MBDIP	6.4	4.5	3.0	1.6	0.3	1.5	2.9	4.3	6.0
MBDIP+LPF (Butterworth 4-th order)	**$3.3\ 10^{-2}$**	**$3.4\ 10^{-2}$**	**$2.1\ 10^{-2}$**	**$2.7\ 10^{-2}$**	**$7.5\ 10^{-3}$**	**$2.7\ 10^{-2}$**	**$2.0\ 10^{-2}$**	**$4.1\ 10^{-2}$**	**$4.4\ 10^{-2}$**
MBDIP+least squares	0.38	0.48	0.36	0.17	$8.6\ 10^{-2}$	0.35	0.40	1.2	1.6

a. Measurement method based on the determination of the increment of the input signal phase (MBDIP)

C. *Flicker Impact*

When performing simulation, a flicker of sinusoidal and polyharmonic forms was considered. In both cases, the flicker frequency was chosen equal to 8 Hz, the modulation depth – 5%. As in the case of harmonics and noise, the use of a post-filter can significantly reduce the measurement error.

VI. SIMULATION MODELING

Simulation modeling is performed by Matlab software package. The results of the frequency measurement error by the considered method are presented in Table 1. For a sinusoidal signal, the parameters values are correspond to the values accepted during the obtaining results of Fig. 2. The parameters of harmonics, noise and flicker are represented in the previous section.

VII. CONCLUSION

Frequency measurement method based on determination of the phase increment (MBDIP) allows to measure the frequency of real electrical networks under conditions of frequency deviation, harmonics, random noise and sinusoidal and rectangular flicker. The frequency measurement error of considered measurement method is a result of "spectrum leakage" effect when applying the Fourier transform; an analytical expression (17) is obtained for calculation of error for the case of a sinusoidal input signal.

It is shown that the additive and multiplicative instrumental errors of measurement channel do not affect to the frequency measurement error of the considered method. It was found that the ADC nonlinearity and ADC quantization error affects to the frequency measurement error only in the case of the presence of the "spectrum leakage" effect; a method is proposed (see expressions (20) and (22)) for assessing the influence of these error components to the frequency measurement error.

REFERENCES

[1] M.M. Begovic, P.M. Djuric, S. Dunlap and A.G. Phadke. "Frequency tracking in power networks in the presence of harmonics," IEEE Trans. on Power Delivery, 8(2), 1993, pp. 480-486.

[2] P.J. Moore, J.H. Allmeling and A.T. Johns, "Frequency relaying based on instantaneous frequency measurement power systems," IEEE Tran. on Power Delivery, 11(4), 1996, pp. 1737-1742.

[3] T.S. Sidhu, "Accurate measurement of power system frequency using a digital signal processing technique," IEEE Trans. on Instr. and Meas., 48(1), 1999, pp. 75-81.

[4] Alberto Carboni; Alessandro Ferrero. "A Fourier Transform-Based Frequency Estimation Algorithm," IEEE Transactions on Instrumentation and Measurement, 2018, vol. 67, issue 7, pp. 1722–1728.

[5] D. Agrez, "Power system frequency estimation in the shortened measurement time," Proceedings of the 18th IEEE Instr. and Meas. Technology Conference, vol. 2, 2001, pp. 1094-1098.

[6] IEEE Standard for Terminology and Test Methods for Analog-to-Digital Converters IEEE Std 1241-2010 (Revision of IEEE Std 1241-2000), pp.: 1 – 139, 2011.

[7] P. Suchanek P., D. Slepicka, V. Haasz "Several approaches to ADC transfer function approximation and their application for ADC nonlinearity correction," Metrology and Measurement Systems, vol. 15, no. 4, pp. 501–511, 2008.

[8] Andey N. Serov, Nikolay A. Serov, Petr K. Makarychev. "Evaluation of the Effect of Nonlinearity of the Successive Approximation ADC to the Measurement Error of RMS," International Symposium on Industrial Electronics, 2018, pp. 1-5.

[9] Andrey N. Serov; Nikolay A. Serov; Alexander A. Shatokhin. "Evaluation of the Effect of Successive Approximation ADC Nonlinearity on the Measurement Error of the Phase Spectrum," 2018 Renewable Energies, Power Systems & Green Inclusive Economy (REPS-GIE), 2018, pp. 1-6.

[10] S. Medawar, P. Handel, N. Bjorsell, M. Jansson "Postcorrection of Pipelined Analog–Digital Converters Based on Input-Dependent Integral Nonlinearity Modeling," IEEE Trans. on Instr. and Meas., vol. 60, issue 10, 2011, pp. 3342-3350.

Comparative Analysis Of the Active Power Measurement Methods in Time Domain

Andrey N. Serov, Alexander A. Shatokhin, Nikolay A. Serov
National Research University "Moscow Power Engineering Institute",
Diagnostic Information Technologies Department, Moscow, Russian Federation
SerovAN@mpei.ru, ShatokhinAA@mpei.ru, SerovNA@mpei.ru

***Abstract* - Active power is one of the most informative parameters of the power grids signals. Signals of present power grids are non-sinusoidal. There are harmonics, noise and flicker (sinusoidal or rectangular). For this reason, the methods used to measure active power must take into account the above signal features. The article discusses two of the most popular measurement methods of the active power in the time domain: the method based on averaging the instantaneous power and the method of low-pass filtration of instantaneous power. For these methods, analytical expressions are obtained that allows to estimate the measurement error of active power in the case of a polyharmonic input signal. Simulation modeling was performed to evaluate the metrological characteristics of the considerate methods in Matlab. The research of influence of the signal parameters on the measurement error is performed. The implementation features, advantages and disadvantages of each method are considered. Approaches to improve the accuracy of the active power measurements when performing additional measurements of power frequency are proposed. It is shown that in the case of using a moving average (single-stage) filter, both considered methods are equivalent. The reliability of analytical expressions is confirmed by comparing of the analytical and simulation modeling results at check points.**

Keywords* - *fast Fourier transform; digital filteration; simulation modeling; measurement error; frequency deviation; polyharmonic signal

I. INTRODUCTION

Under the active power of a periodic signal is understood the value which is determined by the following expression [1]-[2]:

$$P_{id} = \frac{1}{T_A}\int_0^{T_A} u(t)i(t)dt\,, \qquad (1)$$

where $u(t)$, $i(t)$ – voltage and current; T_A – measurement time which is multiple of the input signal period.

For the case of polyharmonic signals, the value of active power can be calculated by the following expression [1]-[2]:

$$P_{id} = \sum_{k=1}^{K} U_k I_k \cos(\varphi_k), \qquad (2)$$

where U_k, I_k – RMS of k-th spectral component of voltage and current; φ_k – phase shift between k-th spectral components of voltage and current; K – number of harmonics taken into account.

Currently, mainly digital methods are applied to measure active power. All digital methods can be separated into several categories according to various criteria. According to the field of signal processing, all methods can be divided into analysis in the time domain [3]-[7] and frequency domain [8]-[12]. By the type of input signals, there are methods for the measurement of sinusoidal and polyharmonic signals.

Both the current and voltage of real electric networks are nonsinusoidal. There are harmonics, noises, flicker of sinusoidal or rectangular forms. In addition, the frequency of electrical networks is deviate over the time. The results of simulation and physical experiments shows that the application of methods which are designed to measure sinusoidal signals results to a significant errors.

This article discusses the most popular methods for the active power measurement: a method based on averaging of instantaneous power and low-pass filtration method. The methods are designed to measure polyharmonic input signals. The principles of active power measurement by applying these methods are described in [3], [5]. Analytical expressions are obtained for estimating the error in the case of sinusoidal signals. However, the issues of polyharmonic signal measurement taking into account the above features are outlined in insufficient volume. In addition, the task of a comparative analysis of these methods from the point of view of measurement error and implementation complexity seems to be relevant.

II. MEASUREMENT METHOD BASED ON THE AVERAGING OF INSTANTANEOUS POWER

A. Description

When applying digital measurement methods, voltage and current signals are represented by discrete analogs and the direct application of expression (1) is impossible. In this method, discrete voltage and current signals are restored by applying a zero-order polynomial. As a result, the formula of the active power calculation:

$$P_{id} = \frac{1}{N}\sum_{n=0}^{N-1} u[n]i[n], \qquad (3)$$

where $u[n] = u(nT_S)$, $i[n] = i(nT_S)$ – discrete current and voltage signals; N – number of samples corresponding to the total measurement time T_A (a multiple of the input signal period).

The number of averaged samples corresponds to the total measurement time and is determined as:

$$N = round(T_A / T_S) = round(mT_1 / T_S), \qquad (4)$$

where T_S – sample time; m – number of a measured signal periods; T_1 – input signal period.

B. Measurement Error of the Sinusoidal Signals

In [3]-[4], an expression was obtained for calculation the active power measurement error in case of sinusoidal signals:

$$\delta_P = -\frac{\cos((N-1)\omega + 2\alpha + \varphi)\sin(\omega N)}{N\sin(\omega)\cos(\varphi)}, \qquad (5)$$

where α – initial voltage phase; φ – phase shift of current relative to voltage; ω – angular frequency (true value) of signals relative to the sampling frequency $f_S = 1 / T_S$.

Fig. 1 shows the dependence of the active power measurement error on the frequency of input sinusoidal signal. From the expression (5) it is seen that the measurement error is equal to zero when the conditions:

– factor $\sin(\omega N)$ equals to zero;

– factor $\cos((N\text{-}1)\omega + 2\alpha + \varphi)$ equals to zero.

The first condition can be provided by adjusting the number of samples to the input signal frequency. Thus, if the measurement time is equal to an integer number of the input signal periods, the active power measurement error is equal to zero (see expression (4)). It is possible to ensure the fulfillment of the second condition by adjusting the initial phase of the voltage. In this case, to ensure that the factor is equal to zero, the initial phase of the voltage:

$$\alpha_{opt} = \alpha = 0.25\pi - 0.5(N-1)\omega - \varphi \pm 0.5\pi k. \qquad (6)$$

This approach, in addition to measurement the frequency value, also requires determining the phase shift between voltage and current, which significantly limits practical application of this approach.

C. Measurement Error of Polyharmonic Signals

In the case of input polyharmonic signals, the discrete analogs can be represented by the following form:

$$\begin{cases} u[n] = \sum_{k=1}^{K} U_k \sqrt{2}\sin(\omega_k n + \alpha_k), \\ i[n] = \sum_{k=1}^{K} I_k \sqrt{2}\sin(\omega_k n + \alpha_k + \varphi_k); \end{cases} \qquad (7)$$

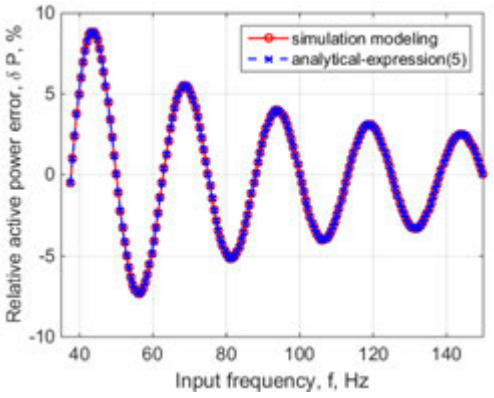


Figure 1. The dependence of the active power measurement error by the input signal frequency

where ω_k – angular frequency of k-th harmonic relative to the sample rate; α_k – initial phase of k-th voltage harmonic.

The instantaneous power signal can be represented as:

$$\begin{aligned} p[n] = &\sum_{k=1}^{K} U_k I_k (\cos(\varphi_k) - \cos(2\omega_k n + 2\alpha_k + \varphi_k)) + \\ &+ \sum_{j=1}^{K}\sum_{k=1,\,k\neq j}^{K} U_j I_k \cos((\omega_j - \omega_k)n + (\alpha_j - \alpha_k) + \varphi_{jk}) - \\ &- \sum_{j=1}^{K}\sum_{k=1,\,k\neq j}^{K} U_j I_k \cos((\omega_j + \omega_k)n + (\alpha_j + \alpha_k) + \varphi_{jk}). \end{aligned} \qquad (8)$$

Then, for the case of polyharmonic input signals (expression (7)), the value of active power takes the form (see expressions (3), (5), (8)):

$$\begin{aligned} P = P_{id} &- \sum_{k=1}^{K} \frac{U_k I_k \cos((N-1)\omega_k + 2\alpha_k + \varphi_k)}{N\sin(\omega_k)\sin^{-1}(\omega_k N)} + \\ &+ \sum_{k=1}^{K}\sum_{j=1,k\neq j}^{K} \frac{U_k I_j \cos(0.5(N-1)(\omega_k - \omega_j) + \zeta_{kj})}{N\sin\left(\frac{\omega_k - \omega_j}{2}\right)\sin^{-1}\left(\left(\frac{\omega_k - \omega_j}{2}\right)N\right)} - \\ &- \sum_{k=1}^{K}\sum_{j=1,k\neq j}^{K} \frac{U_k I_j \cos(0.5(N-1)(\omega_k + \omega_j) + \psi_{kj})}{N\sin\left(\frac{\omega_k + \omega_j}{2}\right)\sin^{-1}\left(\left(\frac{\omega_k + \omega_j}{2}\right)N\right)}, \end{aligned} \qquad (9)$$

where $\zeta_{kj} = \alpha_k - \alpha_j - \varphi_{kj}$; $\psi_{kj} = \alpha_k + \alpha_j + \varphi_{kj}$.

Due to the application of the averaging operation, it is possible to suppress time-dependent components of expression (8). It can be seen from expression (9) that the active power measurement error for the case of a polyharmonic signal can be eliminated by adjusting the number of samples in exact accordance with the period T_1 of the fundamental spectral component:

$$N = round(T_1 / T_S). \qquad (10)$$

Under condition (9), the factors $\sin(\omega_k N)$, $\sin((\omega_k - \omega_j)N)$ and $\sin((\omega_k + \omega_j)N)$ equals to zero for any combination of indices j and k. For arbitrary values of the initial phases and phase shifts, the maximum error value:

$$\Delta P = \sum_{k=1}^{K} \frac{U_k I_k \sin(\omega_k N)}{N \sin(\omega_k)} + \sum_{k=1}^{K} \sum_{j=1, k\neq j}^{K} \frac{U_k I_j \sin((\omega_k - \omega_j)N)}{N \sin(\omega_k - \omega_j)} + \sum_{k=1}^{K} \sum_{j=1, k\neq j}^{K} \frac{U_k I_j \sin((\omega_k + \omega_j)N)}{N \sin(\omega_k + \omega_j)}. \quad (11)$$

III. Method of Low-Pass Filtration of Instantanius Power

A. Description

The measurement method based on the application of low-pass filtration consists in sequentially determining of the instantaneous power and suppressing the time-variable components by application of a low-pass filter:

$$P[n] = F_{LP}(p[n]). \quad (12)$$

For the case of polyharmonic input signals, the instantaneous power is determined by expression (8). Then the filter output signal:

$$P[n] = F_{LP}(p[n]) = \sum_{k=1}^{K} U_k I_k \left(H_1 \cos(\varphi_k) - H(2\omega_k) \cos\left(2\omega_k n + \varphi_{LP1,k}\right)\right) + \sum_{j=1}^{K} \sum_{k=1,\, k\neq j}^{K} U_j I_k H(\omega_j - \omega_k) \cos\left(\left(\omega_j - \omega_k\right)n + \varphi_{LP2,jk}\right) - \sum_{j=1}^{K} \sum_{k=1,\, k\neq j}^{K} U_j I_k H(\omega_j + \omega_k) \cos\left(\left(\omega_j + \omega_k\right)n + \varphi_{LP3,jk}\right), \quad (13)$$

where $\varphi_{LP1,k} = 2\alpha_k + \varphi_k + \varphi(2\omega_k)$, $\varphi_{LP2,jk} = \alpha_j - \alpha_k + \varphi_k + \varphi(\omega_j - \omega_k)$, $\varphi_{LP3,jk} = \alpha_j + \alpha_k + \varphi_k + \varphi(\omega_j + \omega_k)$ – phase values of the spectral components of the filter output signal; H_1 – filter gain.

B. Measurement Error

The value of the active power error (see (8), (13)):

$$\Delta P = -\sum_{k=1}^{K} U_k I_k \left(H(2\omega_k) \cos\left(2\omega_k n + \varphi_{LP1,k}\right)\right) + \sum_{j=1}^{K} \sum_{k=1,\, k\neq j}^{K} U_j I_k H(\omega_j - \omega_k) \cos\left(\left(\omega_j - \omega_k\right)n + \varphi_{LP2,jk}\right) - \sum_{j=1}^{K} \sum_{k=1,\, k\neq j}^{K} U_j I_k H(\omega_j + \omega_k) \cos\left(\left(\omega_j + \omega_k\right)n + \varphi_{LP\backslash 3,jk}\right). \quad (14)$$

The maximum value of the error (for arbitrary initial phases and phase shifts):

$$\Delta P = \sum_{k=1}^{K} U_k I_k H(2\omega_k) + \sum_{j=1}^{K} \sum_{k=1,\, k\neq j}^{K} U_j I_k H(\omega_j + \omega_k) + \sum_{j=1}^{K} \sum_{k=1,\, k\neq j}^{K} U_j I_k H(\omega_j - \omega_k). \quad (15)$$

C. Selection of Low-pass Filter Parameters

It can be seen from expressions (14) and (15) that the active power measurement error is determined by the filter suppression coefficient for the time-variable instantaneous power components (8). For the problem being solved An ideal filter should have the following characteristics:

$$\begin{cases} H(\omega) = H_1 = 1, \ \omega < \omega_P, \\ H(\omega) = H_2 = 0, \ \omega > \omega_{NOM} + \Delta\omega; \end{cases} \quad (16)$$

where ω_P – maximum frequency deviation of the useful component (active power); $\Delta\omega$ – maximum deviation of the signal frequency from the nominal value ω_{NOM}.

In the expression (16), all spectral components of the instantaneous power are taken into account. The frequency range of the spectral components to be suppressed is $\omega_{NOM} \div 2N\omega_{NOM}$. Expression (16) is written for the case of an ideal filter. A real filter has non-zero passband attenuation and finite stopband suppression. For the problem to be solved, the filter must have the following characteristics:

– passband limit $\omega_{pass} = \Delta\omega$;

– stopband limit $\omega_{stop} = \omega_{stop} - \Delta\omega$;

– passband attenuation $A_{pass} = 0.1$ dBl;

– stopband suppression $A_{stop} = 60$ dBl.

By using expression (15), we can obtain the ratio for estimating the minimum attenuation of the filter stop-band suppression to provide the required active power measurement error:

$$H_2 \leq \frac{\Delta P_{\max}}{\sum_{k=1}^{K} U_k I_k + 2\sum_{j=1}^{K} \sum_{k=1,\, k\neq j}^{K} U_j I_k}, \quad (17)$$

where ΔP_{max} – maximum value of the active power measurement error.

D. Moving Average Filter

The moving average filter is the easiest solution to implement. The frequency response of such filter has the following form [7], [13]:

$$H(\omega) = \left|H(j\omega)\right| = \left(\frac{1}{L} \frac{\sin(0.5L\omega)}{\sin(0.5\omega)}\right)^M, \quad (18)$$

where L – filter order; M – stage number (number of applied cascades) of the filter.

The order of the moving average filter is determined by the following relation:

$$L = round(2\pi / \omega_{MAF}), \quad (19)$$

where ω_{MAF} – normalized angular frequency of the first pole of the moving average filter magnitude response.

In the problem to be solved, one should choose:

$$\omega_{MAF} = \omega = 2\pi f / f_S , \qquad (20)$$

where ω – nominal or measurement value of the input signal frequency.

Comparing expressions (15), (18) for the low-pass filtration method and expression (11) for the instantaneous power averaging method, we can conclude that these methods are equivalent when applying a moving average filter for low-pass filtration method implementation.

If the frequency of the input signal is time-nonvariable and equal to its nominal value, then the use of relations (19) and (20) will ensure complete suppression of all variable spectral components of instantaneous power. For real electrical networks, there is a finite frequency deviation. As a result of this, deviations of the frequencies of the instantaneous power spectral components with respect to the frequencies at which the filter magnitude response is equal to zero. This results to an finite value of the active power measurement error. This error can be reduced by increasing the filter stage number (parameter *M*, see expression (18)) or by performing an additional measurement of the frequency of the input signal.

The influence of the parameter *M* on the magnitude response of the moving average filter is shown in Fig. 2. It is seen that with increasing of *M*, the magnitude response in stopband tends to decrease. However, the non-uniformity of the filter in the passband increases, as well as the overall duration of the transition process. The total settling time can be calculated by using the following formula:

$$t_{trans} = N \cdot T_S \cdot M . \qquad (21)$$

By additionally frequency measurement and adjusting the filter order, it is possible to reduce the influence of frequency deviation on the measurement error of active power. In this case, in expression (20), instead of the nominal value of the signal frequency, the measurement result must be used. In this case, due to the finite measurement error of the frequency and the discrete nature of the parameter *L*, a deviation of the calculation result from the optimal value *L* occurs at which a zero measurement error is provided:

$$\Delta L = 0.5 + \Delta T / T_S , \qquad (22)$$

where ΔT – frequency measurement error.

The influence of deviations on the frequency response of the filter can be expressed by the following formula:

$$H(\omega) \cong \left(\frac{1}{L + \Delta L} \frac{0.5\Delta L\omega}{\sin(0.5\omega)} \right)^M \cong \left(\frac{1}{L} \frac{0.5\Delta L\omega}{\sin(0.5\omega)} \right)^M . \qquad (23)$$

Expression (23) is obtained from expression (18) by expanding in a Taylor series up to two terms by the variable ΔL. The use of this expression allows to calculate maximum frequency measurement error to measure the

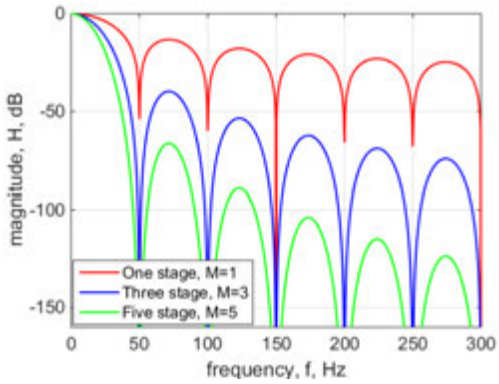


Figure 2. Magnitude response of the mouving average filters of different stages (cascades number, parameter *M*)

active power with an error not exceeding ΔP_{max}. For this, the calculation of the minimum stopband attenuation is performed according to the formula (17). Further, the maximum value of the frequency measurement error is determined by using relations (23) and (22):

$$\Delta T \le \left(\frac{\sqrt[M]{H_2}\, L \sin(0.5\omega)}{0.5\omega} - 0.5 \right) T_S . \qquad (24)$$

The moving average filter can be implemented by FIR or IIR structure. Implementation of the IIR-structure for a filter of the one stage requires one feedback; for implementation of the FIR-structure, the filter order corresponds to the value of parameter *L* from expression (19). In the case of adjusting the number of averaged samples, the implementation of the IIR structure is preferable, since it is easier to control the filter order.

E. Application of the IIR Filters

Digital filters with an infinite impulse response (IIR) have several advantages: smaller order compared to finite impulse response (FIR) filters; ease of development, analysis and modeling. Filters of this type are constructed by an analog filter prototype and bilinear z-transform technique. In this work, to develop filters of this type, the Matlab functions were used. Butterworth, Chebyshev I type, Chebyshev II type, elliptic filters were considered. Fig. 3 shows the magnitude response of these filters. Table 1 presents the main characteristics of these filters. In addition to the order of the filter, which affects the complexity of the implementation, an important characteristic is the settling time, which determines the value of dynamic component of the measurement error.

F. Application of the FIR Filters

Filters with a finite impulse response have not feedback from the filter output to its input, which determines its main advantage. The presence of this feature results to smaller errors associated with the discreteness of the samples of the input signal and the possibility of a filter implementation on integer arithmetic. The most popular type of digital FIR filter is a filter based on the Kaiser window. The frequency response of such a filter represented in Fig. 3, the main characteristics are shown in Table 1.

G. Application of the "Combined" Filter

Note that for all considered IIR and FIR filters, the exact fulfillment of relation (16) is impossible. An exception is the digital moving average filter. Cascading a moving average filter and any IIR or FIR filter will increase the stopband suppression (reducing the static measurement error by decreasing H_2). However, the total filter settling time will increase, which results to increase of error dynamic component. Also, for the purpose of increasing stopband suppression, a filter-compensator [7], [13] can be applied. The total transfer coefficient of the combined filter is determined by:

$$H_{COMB}(j\omega) = H_{MAF}(j\omega)H_C(j\omega), \qquad (25)$$

where $H_{MAF}(j\omega)$ – moving average filter gain; $H_C(j\omega)$ – an additional FIR or IIR filter or filter-compensator gain.

IV. Frequency Measurement

For both considered methods of active power measurement, performing an additional frequency measurement of the input signal results to a decrease of the measurement error. As can be seen from expressions (10), (11) and (19), (20), (23), the frequency measurement error affects the accuracy of tuning the number of averaged samples and results to a active power measurement error. When choosing a measurement method, it should be noted that the input signals are polyharmonic. For this reason, the use of methods for sinusoidal signals measurement can result to the appearance of a significant additional error. Input signal frequency deviations are a slowly process. According to the publication [14], the rate of change of the frequency of the fundamental component does not exceed 0.02% / sec. For this reason, frequency measurement methods that require a significant measurement time can be applied.

Methods of the frequency measurement of a polyharmonic signal that are resistant to flicker, noise, and suprarmonics are [15]-[16]:

– zero-crossing technique;

– method based on spectral analysis with determination of the fundamental spectral component position;

– method based on spectral analysis followed by approximation of the phase spectrum.

The implementation of the first method is much simpler. Simulation results shows that the frequency measurement error in real power networks does not exceed 0.01 %.

V. Simulation Modeling

Based on the results of the research done, simulation modeling of the considered active power measurement methods was performed. For a method based on averaging the instantaneous power (AIP), a construction option with an additional adjustment of the number of averaged samples and without adjustment is considered. For the method of the low-pass filtration of instantaneous power

TABLE I. Comparative Analysis of the Digital Filters

No	Filter type	Parameter		Matlab functions
		Order	*Tras. time, sec*	
1.	Butterworth	5	0.23	buttord, butter
2.	Chebyshev, I type	4	0.4	cheb1ord, cheby1
3.	Chebyshev, II type	4	0.18	cheb2ord, cheby2
4.	elliptic	3	0.25	ellipord, ellip
5.	Kiser	829	0.08	kaiserord, fir1

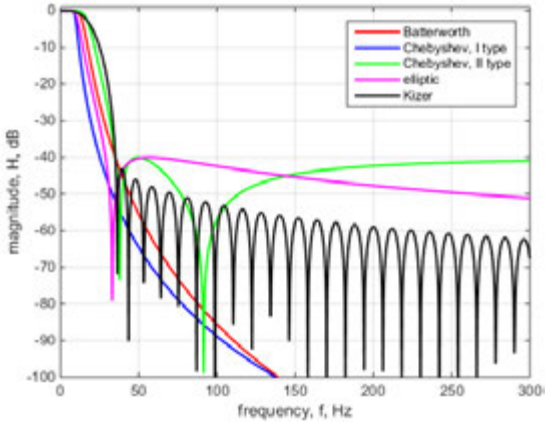


Figure 3. IIR and FIR filters magnitude response

(LPF), a digital moving average filter, Butterworth, Chebyshev of the first and second type, elliptic, FIR-filter based on the Kaiser window and a combined filter are considered. Butterworth IIR filter and moving average filter (MAF, without adjusting of the filter order) were selected as a combined filter. The case of input polyharmonic signals, and sinusoidal signals distorted by white noise are considered. The filter characteristics are shown in Table 1. The total simulation time was chosen equal to 0.2 sec. Sampling frequency – 10 kHz. RMS noise value is selected equal to 2 % of the input signal RMS. The harmonic coefficients of voltage and current correspond to the limit values in EN 50160:2010 and [17], respectively. A sinusoidal flicker with a modulation depth of 10 % and a frequency of 8 Hz was selected. The simulation results are presented in Table 2.

VI. Conclusion

Based on the results of the research done, the following conclusions can be drawn:

– for a method based on averaging of the instantaneous power, analytical expressions (5) and (11) are obtained that allows to estimate the measurement error in the case of polyharmonic input signals;

– simulation modeling was performed in Matlab and Simulink packages, which allows to evaluate the effect of harmonics, noise and flicker of signals on the active power measurement error; according to the simulation results, a measurement method based on the application of a combined filter has the best metrological characteristics for polyharmonic input signals;

TABLE II. COMPARATIVE ANALYSIS OF THE ACTIVE POWER MEASUREMENT METHODS IN THE TIME DOMAIN (ACTIVE POWER RELATIVE MEASUREMENT ERRORS ARE REPRESENTED IN PERSENTS)

Measurement method	Frequency deviation, δ_f, Hz								
	-0.06	-0.045	-0.03	-0.015	0.00	0.015	0.03	0.045	0.06
	Polyharmonic signal								
Method AIP	0.057	0.17	0.33	0.28	0.00	-0.05	0.16	0.31	0.26
Method AIP (sample number adjustment)	$3.9\ 10^{-3}$	$1.3\ 10^{-3}$	$5.7\ 10^{-4}$	$2.3\ 10^{-3}$	0.00	$2.3\ 10^{-3}$	$1.8\ 10^{-3}$	$1.2\ 10^{-3}$	$4.5\ 10^{-3}$
Method LPF (MAF filter)	7.3	5.5	3.7	1.9	0.00	1.8	3.5	5.1	6.5
Method LPF (MAF, order adjustment)	0.14	0.25	0.11	0.028	0.00	0.028	0.11	0.25	0.21
Method LPF (Butterworth filter)	$1.6\ 10^{-2}$	$1.5\ 10^{-2}$	$1.4\ 10^{-2}$	$1.3\ 10^{-2}$	$1.2\ 10^{-2}$	$1.1\ 10^{-2}$	$9.8\ 10^{-3}$	$9.2\ 10^{-3}$	$8.6\ 10^{-3}$
Method LPF (Chebyshev I type filter)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Method LPF (Chebyshev II type filter)	0.36	0.38	0.40	0.41	0.43	0.44	0.46	0.47	0.49
Method LPF (elliptic filter)	0.88	0.87	0.86	0.85	0.84	0.83	0.83	0.82	0.81
Method LPF (Kiser filter)	0.91	0.87	0.78	0.70	0.77	0.84	0.88	0.89	0.88
Method LPF (combined filter)	$2.5\ 10^{-3}$	$2.0\ 10^{-3}$	$1.9\ 10^{-3}$	$1.6\ 10^{-3}$	$2.0\ 10^{-3}$	$1.6\ 10^{-3}$	$1.7\ 10^{-3}$	$1.8\ 10^{-3}$	$1.9\ 10^{-3}$
	Sinusoidal signal+noise								
Method AIP (sample number adjustment)	$4.6\ 10^{-4}$	$4.3\ 10^{-2}$	$3.7\ 10^{-2}$	$1.5\ 10^{-2}$	$1.0\ 10^{-2}$	$7.9\ 10^{-3}$	$9.7\ 10^{-3}$	$5.2\ 10^{-2}$	$8.2\ 10^{-2}$
Method LPF (MAF, order adjustment)	0.42	0.70	0.62	0.68	0.55	0.40	0.63	0.72	0.70
Method LPF (Butterworth filter)	0.16	0.34	0.25	0.10	0.64	0.31	0.20	0.18	0.20
Method LPF (combined filter)	0.09	0.22	0.22	0.27	0.59	0.29	0.21	0.16	0.36
	Sinusoidal signal+flicker(sinusoidal)								
Method AIP (sample number adjustment)	$7.5\ 10^{-2}$	$9.0\ 10^{-2}$	$6.2\ 10^{-2}$	$7.8\ 10^{-2}$	$9.5\ 10^{-2}$	$7.0\ 10^{-2}$	$9.1\ 10^{-2}$	$6.3\ 10^{-2}$	$7.6\ 10^{-2}$
Method LPF (MAF, order adjustment)	2.01	2.10	2.01	1.96	1.93	1.94	1.98	2.12	2.08
Method LPF (Butterworth filter)	1.97	1.97	1.97	1.97	1.98	1.98	1.98	1.97	1.97
Method LPF (combined filter)	1.90	1.91	1.91	1.91	1.91	1.91	1.91	1.91	1.91

– for the low-pass filtration method, analytical relationships are obtained that associate the filter parameters and the measurement error of the active power; a ratio is obtained that allows to calculate the minimum stop-band attenuation of the desired filter;

– it is shown that in the case of applying a moving average filter (single-stage), the method based on instantaneous power filtration is equivalent to the method based on averaging of instantaneous power.

REFERENCES

[1] Emanuel A.E. “Power definitions and the physical mechanism of power flow,” Wiley Chichester, 2010.

[2] Dugan R., McGranaghan M., Santoso S., Beaty W. Electrical power systems quality, New-York: McGraw-Hill, 2004, 525p.

[3] A. N. Serov; A. A. Shatokhin “Analysis of truncation error at digital measurement of power and energy”, MPEI Vestnik, No. 5, 2008, pp. 107-112.

[4] Igor V. Levashov; Dmitry A. Chumachenko; Andrey N. Serov “The Ways of Reducing of the Active Power Measurement Error for the Method of Averaging of the Instantaneous Power,” 2020 International Russian Automation Conference (RusAutoCon), (in press).

[5] Andrey N. Serov; Nikolay A. Serov; Alexey A. Lupachev “The Application of Digital Filtration for the Active Power Measurement,” 2018 IEEE International Conference on Electrical Engineering and Photonics (EExPolytech), 2018, pp. 1-4.

[6] Saranovac L.V. “Digital realization of frequency insensitive phase shifter for reactive var-hour meters,” IEEE Transactions on Instrumentation and Measurement, 2000, vol. 49, issue. 4, pp. 802-808.

[7] A. Serov; A. Shatokhin; A. Novitskiy; D. Westermann “Investigation of the method of RMS measuring based on the digital filtration of the square of samples,” 2018 18th International Conference on Harmonics and Quality of Power (ICHQP), pp.1-6.

[8] Ferrero A. “Harmonic power flow analysis for the measurement of the electric power quality,” IEEE transactions on instrumentation and measurement, 1995, vol. 60, issue 3, pp. 683-685.

[9] Ferrero A. “Harmonic power flow analysis for the measurement of the electric power quality,” IEEE transactions on instrumentation and measurement, 1995, vol. 60, issue 3, pp. 683-685.

[10] Mikhalin S.N.; Gevorkyan V.M. “The problems of digital processing of signals in the system of automated power quality control and accounting quantity of electricity (ASQAE)”, MPEI Vestnik, No. 1, pp. 86-92, 2005.

[11] Andrey N. Serov; Alexander Novitskiy; Alexander A. Shatokhin; Steffen Schlegel; Ekaterina A. Dolgacheva; Dirk Westermann “The Influence of Power Frequency Deviation on the Active and Reactive Power Measurement Error with the Application of DFT,” 2019 20th International Symposium on Power Electronics (Ee), 2019, pp. 1-6.

[12] A.N. Serov; N.A. Serov; S.I. Gerasimov "Application of the Quadrature Demodulation for the Measurement of Electric Power Parameters," 2019 18th International Symposium INFOTEH-JAHORINA, pp. 1-6, 2019.

[13] Laddomada M. On the Polyphase “Decomposition for Design of Generalized Comb Decimation Filters,” IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 55, issue 8, pp.: 2287-2299, 2008.

[14] S.S. Tsai; Li Zhang; A.G. Phadke; Yilu Liu; M.R. Ingram; S.C. Bell; I.S. Grant; D.T. Bradshaw; D. Lubkeman; L. Tang “Study of global frequency dynamic behavior of large power systems,” IEEE PES Power Systems Conference and Exposition, 2004.

[15] M. M. Begovic, P. M. Djuric, S. Dunlap and A. G. Phadke, “Frequency tracking in power networks in the presence of harmonics,” IEEE Transactions on Power Delivery, vol. 8, issue 2, 1993, pp. 480-486.

[16] A. Serov “Frequency estimation methods for stationary signals,” 2017 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM), pp. 1-6.

[17] Technical requirements for the connection and operation of customer installations to the medium voltage network (TAR medium voltage) VDE-AR-N 4110:2018-11.

Comparison of Pearson correlation coefficient and distance correlation in Correlation Power Analysis on Digital Multiplier

Jurica Kundrata[1], Daisuke Fujimoto[2], Yuichi Hayashi[2], Adrijan Barić[1]
[1]University of Zagreb, Faculty of Electrical Engineering and Computing
E-mail: jurica.kundrata@fer.hr, adrijan.baric@fer.hr
[2]Nara Institute of Science and Technology
E-mail: fujimoto@is.naist.jp, yu-ichi@is.naist.jp

Abstract— Correlation power analysis (CPA) is a side-channel attack (SCA) which exploits the information leaked through the power supply current and voltage, or the electromagnetic emissions of the attacked digital system. It uses statistical analysis of a large number of power supply measurements to retrieve the secrets of the digital system. Correlation power analysis uses a number of hypothetical secret keys which are correlated to the measurements of the attacked system. Usually correlation power analysis uses the Pearson correlation coefficient, but the intermediary values and the power supply measurements can have a nonlinear relationship. The paper investigates the application of the distance correlation in the correlation power analysis and compares it to the Pearson correlation coefficient. The comparison is based on a side-channel attack on a multiplication operation of an input message and a secret key. The results of the comparison show that the distance correlation achieves a higher prominence of the correct secret key than the Pearson correlation coefficient.

***Keywords*—side-channel attack; correlation power analysis; distance correlation;**

I. INTRODUCTION

Side-channel attack (SCA) is an attack on a computer system which attempts to retrieve its secrets based on the measurements and analysis of a side-channel source of information. The side-channel source of information can be e.g. the power consumption, electromagnetic emissions or execution timings of the computer system. Compared to the traditional forms of attack where weaknesses of the implemented algorithm are exploited, the side-channel attacks exploit the weaknesses of the implementation of the computer system [1], [2], [3].

The secret information of the computer system can be leaked via the power supply and in the case of digital circuits the information is leaked when the circuit is changing its output state. Changing the output state of the digital circuit consists of charging and discharging some parasitic capacitances which draws a current from the power supply. Depending on the impedance of the power supply network the voltage of the power supply is perturbed. These perturbations can then give insight on the operation of the digital circuit. Side-channel attacks via power supply measurements use two main approaches: simple power analysis (SPA) and differential power analysis (DPA) [4]. Simple power analysis is based on observing and analysing a single waveform of power consumption of the computer system [5], [6]. Differential power analysis is an attack which uses statistical analysis of a large number of waveforms and a popular method of such an attack is the correlation power analysis (CPA) [7].

Correlation power analysis uses a number of hypothetical secrets which determine the intermediate values of the attacked computer system. The intermediate values and the power consumption model are used to construct the hypothetical power consumption waveform. The hypothetical waveforms and the power supply measurements of the attacked system are statistically tested for correlation. The most often used statistical test is the Pearson correlation coefficient [8], [9].

The multiplication operation in digital circuits is a complex operation and it is attacked in cryptography as a part of bilinear pairings [10], [11], [12] or in reverse engineering neural networks [13].

The intermediary values and the power consumption measurements can have a nonlinear relationship, while the Pearson correlation coefficient exclusively identifies linear relationships. A correlation measure which identifies both linear and nonlinear relationships in the tested data is the distance correlation [14]. The distance correlation has a zero value if and only if the tested data is independent.

This paper introduces the application of the distance correlation in the correlation power analysis and compares the performance of the Pearson correlation coefficient and the distance correlation on a test digital circuit which implements a multiplication operation.

The digital multiplier circuit which is used to compare the correlation power analysis using the two correlation coefficients is described in Section II. The methodology of the correlation power analysis, as well as the Pearson correlation coefficient and the distance correlation are introduced in Section III. The results of the measurements, the correlation power analyses, and their comparison are presented in Section IV. The measurement and the comparison results are discussed in Section V, which is followed by conclusions presented in Section VI.

II. DIGITAL MULTIPLIER CIRCUIT

A. Design of the test circuit

The comparison of the correlation power analyses based on the Pearson correlation coefficient and the distance correlation is based on a side channel attack on a

This work was in part supported by FFG, Austria through project "Sensor Fusion".

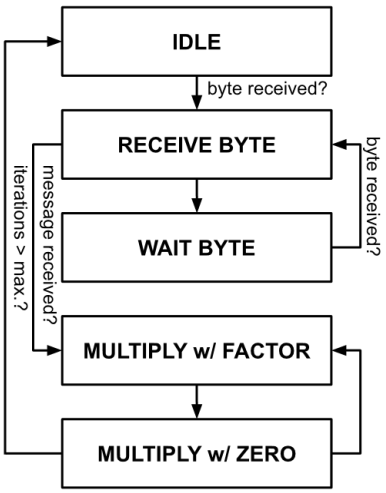


Figure 1. The flowchart of the behaviour of the digital test circuit.

multiplication operation implemented in a digital circuit. The digital circuit is defined in Verilog and then synthesized and implemented in an Artix-7 FPGA, i.e. the development platform Arty A7-35T. The digital design of the test circuit has two main functionalities: receiving the message via the UART interface and then multiplying the message with a secret key. The test circuit implements a 128 x 128-bit multiplication operation and its clock frequency is set to f = 10 MHz.

Figure 1 shows the flowchart of the behaviour of the test circuit. The test circuit is implemented as a finite state machine which first receives the message in bytes via the UART interface and then alternates between multiplying the message and the secret key, and resetting for a defined amount of time. While the state machine alternates between multiplying and resetting, a measurement triggering signal is also generated. Alternating multiplication and resetting ensures that the side-channel information, i.e. power supply measurement is proportional to the Hamming weight of the multiplication operation result.

B. Power supply measurements

Figure 2 illustrates the side channel information leakage through the power supply in a digital circuit. Fig shows an example CMOS circuit in an Artix-7 FPGA which is connected to the power supply voltage $V_{CC,INT}$.

The CMOS circuit is characterized by its parasitic capacitance C_{PAR} which is charged and discharged when the output of the circuit is changing. The charging and discharging of the parasitic capacitance cause perturbations in the power supply voltage and consequently leaks information with respect to the state of the digital circuit.

The power supply of the FPGA core is decoupled by using the capacitor C_{DEC}. In these measurements the decoupling capacitors connected to the power supply $V_{CC,INT}$ are removed in order to distinguish the power supply voltage perturbations which represent the side-channel leakage information.

Additionally, the multiplication and the resetting operations are alternated a large number of times which produces the measured waveform. The waveform is measured using the deep-memory oscilloscope Yokogawa DLM4038 and it is triggered by the signal generated by the test digital circuit.

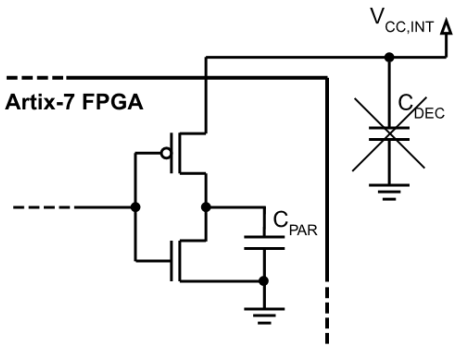


Figure 2. An example CMOS circuit in the FPGA with the power supply.

The waveform of the measured power supply voltage is segmented with respect to the period of the clock signal of the digital circuit. The duration of the measured waveform is set to m = 1000 clock cycles. The waveform segments are then averaged and the peak to peak value of the mean waveform represents the measurement value of the power supply voltage which is used in the correlation power analysis.

III. Correlation Power Analysis

A. Pearson correlation coefficient

The Pearson correlation coefficient $r_{H,V}$ of the Hamming weight of the multiplication result H and the power supply voltage V is estimated using

$$r_{H,V} = \frac{\sum_{i=1}^{n}\left(h_i - \bar{h}\right)\left(v_i - \bar{v}\right)}{\sqrt{\sum_{i=1}^{n}\left(h_i - \bar{h}\right)^2}\sqrt{\sum_{i=1}^{n}\left(v_i - \bar{v}\right)^2}} \tag{1}$$

where n is the number of measurement samples, h_i and v_i are the Hamming weight and the power supply voltage samples, while $\bar{h}$ and $\bar{v}$ are their mean values.

The Pearson correlation coefficient $r_{H,V}$ ranges from $r_{H,V}$ = -1 which indicates a perfect negative linear relationship, to $r_{H,V}$ = +1 which indicates a perfect positive relationship. The Pearson correlation coefficient $r_{H,V}$ = 0 suggests no linear association, but it does not indicate that the two variables are independent.

B. Distance correlation

The distance correlation is a dependency measure between two random vectors which is defined in [14]. Distance correlation of a sample is estimated by using the sample distances a_{ij} and b_{ij} which are defined as

$$a_{ij} = \left|h_i - h_j\right| \text{ and } b_{ij} = \left|v_i - v_j\right| \tag{2}$$

where h_i and h_j are the Hamming weight samples, and v_i and v_j are the power supply voltage samples. The sample distances are then used to define the corresponding double centred matrices $A_{i,j}$ and $B_{i,j}$ by

$$A_{i,j} = a_{ij} - \frac{1}{n}\sum_{l=1}^{n} a_{il} - \frac{1}{n}\sum_{k=1}^{n} a_{kj} + \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl} \tag{3}$$

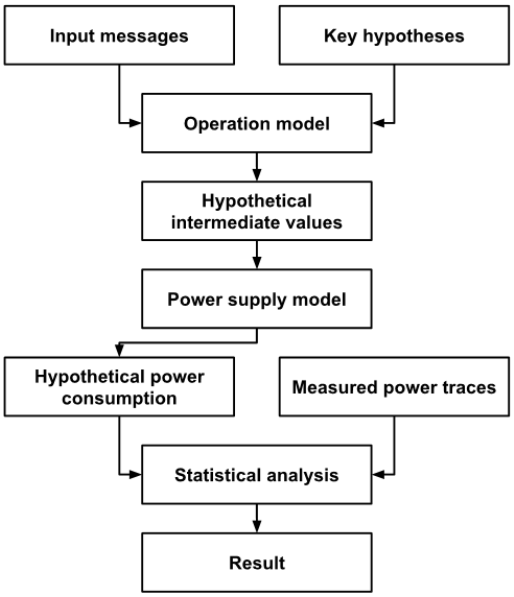


Figure 3. The flowchart of the Correlation Power Analysis (CPA).

$$B_{i,j} = b_{ij} - \frac{1}{n}\sum_{l=1}^{n} b_{il} - \frac{1}{n}\sum_{k=1}^{n} b_{kj} + \frac{1}{n^2}\sum_{k,l=1}^{n} b_{kl} \tag{4}$$

where n is the number of measurement samples. The double centred matrices are used to define the squared distance covariance $N^2(h,v)$ by using

$$\mathrm{N}_n^2(h,v) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{i,j}B_{i,j} \quad . \tag{5}$$

Finally, the squared distance correlation coefficient $d^2(h,v)$ of the Hamming weight of the multiplication result h and the power supply voltage v is estimated by

$$d_n^2(h,v) = \frac{\mathrm{N}_n^2(h,v)}{\mathrm{N}_n^2(h,h)\mathrm{N}_n^2(v,v)} \quad . \tag{6}$$

The distance correlation $d^2(h,v)$ ranges from $d^2(h,v) = 0$ to $d^2(h,v) = 1$, where the correlation $d^2(h,v) = 0$ only if the two variables are independent, i.e. if there is no linear or non-linear relationship between the tested variables.

C. CPA method

Correlation power analysis (CPA) is a statistical method which is uses a large sample of side-channel measurements to determine the secret key. The secret key is determined by observing a statistical measure between the hypothesized value, i.e. output of the digital operation for a given input message and secret key hypothesis, and the side-channel measurement. The most often used statistical measure is the Pearson correlation coefficient, while this paper also considers the distance correlation.

Fig. 3 shows the flowchart of the correlation power analysis. For a given sample of input messages and side-channel, i.e. power supply measurements, the correlation power analysis is done in the following manner. The input messages and the key hypotheses are used to determine the hypothetical intermediate values using the model of the digital operation. The power supply model uses the hypothetical intermediate values to determine the hypothetical power consumption. Finally, the hypothetical power consumption is statistically compared with the measured power traces. The key hypothesis that yields the highest correlation is determined as the correct secret key.

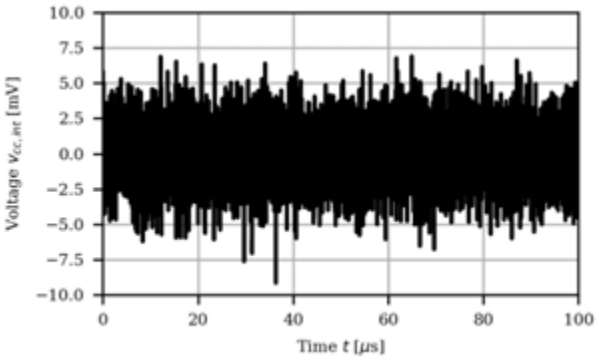


Figure 4. An example measurement of the power supply voltage.

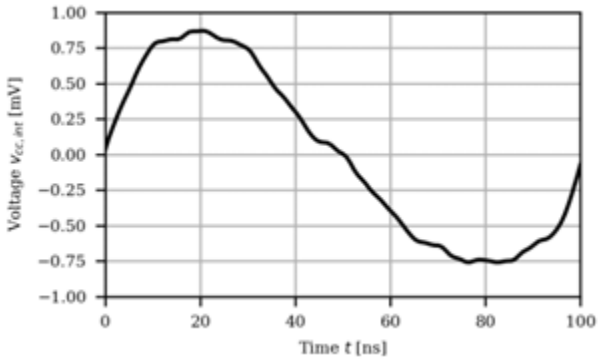


Figure 5. The mean waveform of the segmented measured waveform in Fig. 4.

The model of the digital operation in this paper is straightforward and the intermediary value is defined as a product of the input message and the secret key.

The power supply model relates the intermediary value and the power consumption of the digital circuit, and in this paper the Hamming weight of the intermediary value is directly related to the measurements of the power supply.

Fig. 3 shows a general correlation power analysis algorithm, while a practical attack on multiplier using multi-word operands consists of two steps. First step attacks the individual words of the operands, while the second step uses the results of the first step and attacks the complete multi-word operands [10]. Aim of this paper is to compare the two correlation measures and the example attack is based on attacking a multiplication operand, i.e. the input message consisting of a single word.

IV. Results

A. Power supply voltage measurements

The power supply measurements on the test digital circuit are done $N = 380$ times, and the 128-bit input messages are generated with a uniform distribution of Hamming weights in the range $h \in [1, 128]$. The targeted secret key is set to 186_{10} (10111010_2). The secret key is an 8-bit number which limits the number of key hypotheses to $2^8 = 256$ hypotheses.

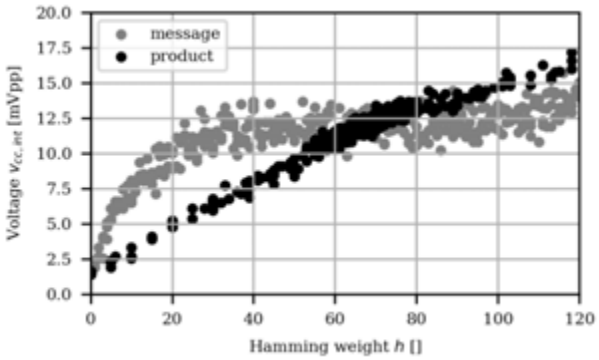


Figure 6. The peak-to-peak value of the averaged power supply waveform with respect to the Hamming weight of multiplication result.

Fig. 4 shows an example of a measured waveform of the power supply voltage $v_{cc,int}$, while Fig. 5 shows the mean value of the segmented, measured waveform shown in Fig. 4. It is observed that the example waveform has a peak-to-peak value $v_{cc,int} \approx 15$ mV$_{PP}$, while the mean waveform has a peak-to-peak value $v_{cc,int} \approx 1.5$ mV$_{PP}$. It is also determined that the RMS value of the example waveform is $v_{cc,int} = 2$ mV$_{RMS}$, while the RMS value of the mean waveform is $v_{cc,int} = 0.6$ mV$_{RMS}$.

Fig. 6 shows the peak-to-peak value of the averaged power supply waveform with respect to the Hamming weight of the input message and the correct multiplication result. Fig. shows that the Hamming weights of the input messages are largely uncorrelated with the power supply measurements, while the weights of the correct multiplication results are strongly correlated with respect to the measurements. These results also show that the relationship between the Hamming weights and the power supply measurements are slightly nonlinear, i.e. the results show a concave relationship.

B. CPA via Pearson correlation coefficient

Fig. 7 shows the Pearson correlation coefficient r with respect to the key hypothesis with the highlighted target key. The Pearson correlation coefficient r is calculated using (1) for the Hamming weight of the multiplication operation result and the peak-to-peak value of the mean power supply voltage. The x-axis of the graph represents the binary value of the key hypothesis and the correct secret key is marked with a star.

It is observed that the highest value of the Pearson correlation coefficient is $r_{max} = 0.9689$, while the lowest value is $r_{min} = 0.7683$ and the median value is $r_{med} = 0.9065$. The highest value of the Pearson correlation coefficient corresponds to two secret key values: 93_{10} (01011101_2) and 186_{10} (10111010_2).

C. CPA via distance correlation

Fig. 8 shows the distance correlation d with respect to the key hypothesis with the highlighted target key. The distance correlation d is calculated using (2)-(6) for the Hamming weight of the multiplication operation result and the peak-to-peak value of the mean power supply voltage.

It is observed that the highest value of the distance correlation is $d_{max} = 0.9671$, while the lowest value is $d_{min} = 0.7488$ and the median value is $d_{med} = 0.8586$. The highest value of the distance correlation corresponds to two secret key values: 93_{10} (01011101_2) and 186_{10} (10111010_2), like the case with the Pearson correlation coefficient.

D. Comparison of correlation power analyses

Fig. 9 presents the comparison between the Pearson correlation coefficient r and the distance correlation d with respect to the number of measurements n. Fig. 9 shows the correlation values for the target key r_{trgt} and d_{trgt}, and the false keys with maximum correlation value $r_{false,max}$ and $d_{false,max}$.

It is observed that the target key correlations in both cases rapidly increase with respect to the number of measurements and with $n = 20$ measurements the correlations exceed r, $d > 0.9800$. Furthermore, at $n \approx 40$ measurements both correlations succeed at identifying the correct secret key, i.e. $r_{trgt} > r_{false,max}$ and $d_{trgt} > d_{false,max}$. As the number of measurements increases, the target key correlations remain largely constant, while the maximum false key correlations steadily fall. It is observed that the decrease in the maximum false key correlation is more pronounced in the case of the distance correlation compared to the Pearson correlation coefficient.

Fig. 10 shows the comparison between the prominence of the Pearson correlation coefficient Δr and the prominence of the distance correlation Δd with respect to the number of measurements n. The prominence is defined as a relative measure between the correlation values of the target key and the false key with maximum correlation, i.e. $\Delta d = 1 - d_{trgt}/d_{false,max}$ and $\Delta r = 1 - r_{trgt}/r_{false,max}$. It is observed that the prominence of both correlations is negative while the number of measurements is $n < 40$, i.e. the Correlation power analysis yields an incorrect secret key solution. In the case where the number of measurements $n > 40$ both correlations succeed at identifying the correct secret key,

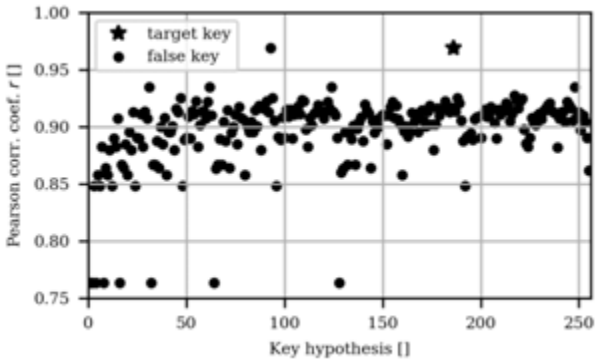


Figure 7. The Pearson correlation coefficient r w.r.t. the key hypotheses.

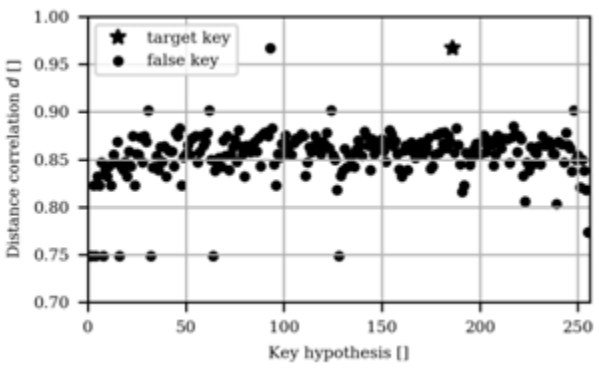


Figure 8. The distance correlation d w.r.t. the key hypotheses.

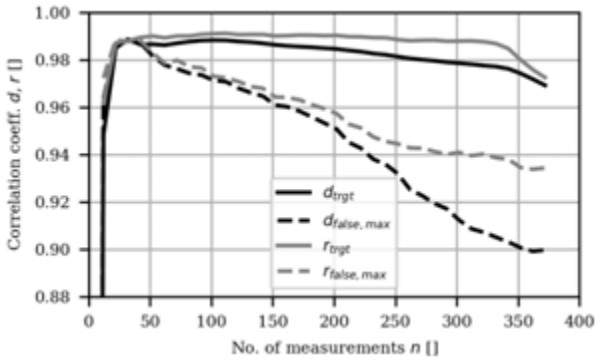


Figure 9. Comparison of the Pearson correlation coefficient r and the distance d w.r.t. number of measurements n.

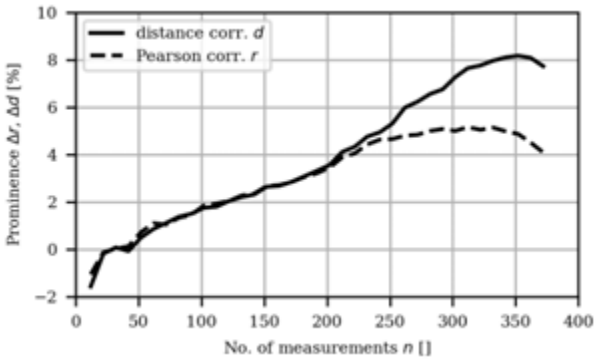


Figure 10. Comparison of the prominence of the Pearson correlation coefficient r and the distance d w.r.t. number of measurements n.

i.e. $\Delta r > 0$ and $\Delta d > 0$. Furthermore, the Pearson correlation coefficient has a maximum prominence $\Delta r \approx 5\%$, while the distance correlation has a maximum prominence $\Delta d \approx 8\%$.

V. Discussion

The test digital circuit implements a 128 x 128-bit multiplication operation, but an 8-bit secret key is used in the measurements to limit the number of key hypotheses and simplify the correlation power analysis and consequently the comparison of the two correlation coefficients. This example can be extended to longer secret keys by extracting the individual bytes of the secret key using the described procedure. In that case a larger number of measurements would be required to complete the analysis and it will depend on the ratio of the number of tested bits and the total number of bits, and the measurement noise level as shown in [7]. Additionally, CPA becomes computationally more prohibitive as it is applied on longer secret keys, e.g. an 8-bit CPA consists of testing 256 key hypotheses, while a 16-bit CPA consists of testing 65536 key hypotheses.

The measurements of the power supply voltage are segmented and averaged in order to separate the voltage perturbation due to the dynamic power consumption and the noise present in the measurements. The results of the waveform measurements and averaging show that this is necessary as the peak-to-peak value of the raw measurement is 10 times greater than the peak-to-peak value of the averaged waveform, while the RMS value of the raw measurement is approx. 3 times greater than the RMS value of the averaged waveform.

The correlation power analysis shows that both the Pearson correlation coefficient and the distance correlation have high values which means that the measurements are highly correlated to the Hamming weights of the multiplication result. This is caused by the fact that the digital circuit only executes one operation, i.e. the multiplication of the input message and the secret key. Also, high correlation values are caused by the fact that the secret key is determined by a single byte. In the case where other operations are executed in parallel to the multiplication or the secret key consists of multiple bytes, the correlation values would be lower.

The correlation power analysis in both cases identifies two secret keys with a maximum correlation. The secret key 93_{10} (01011101_2) is a shifted version of the secret key 186_{10} (10111010_2). The multiplication with the secret key 93_{10} (01011101_2) produces results which have an equal Hamming weight as the results of the multiplication with the secret key 186_{10} (10111010_2).

Although the correlation Power Analyses using both of the correlation coefficients have a similar maximum value, i.e. $r_{max} \approx d_{max} \approx 0.97$, the minimum value and in particular the median value of the analysis using the distance correlation is lower than in the case of the analysis using the Pearson correlation coefficient. This leads to a conclusion that during the correlation power analysis the statistical test that uses the distance correlation better discriminates between the target and the false key hypotheses than the Pearson correlation coefficient.

The relative comparison of the two correlation coefficients confirms the previous conclusion that the difference between the target and the maximum false key correlations is more pronounced in the case of the distance correlation, i.e. the result of the correlation power analysis using the distance correlation yields a more prominent secret key solution, than the correlation power analysis using the Pearson correlation coefficient.

Further research in comparing the two correlation coefficients includes developing a more complex test digital circuit which executes parallel operations along the attacked multiplication. The parallel operations represent an interference during the correlation power analysis and allows for further comparison of the two correlation coefficients. Further research also includes implementing the correlation power analysis in the case of longer secret keys.

VI. Conclusions

Multiplication operation which gives a product of an input message and a secret key is attacked via a side-channel. The multiplication operation is implemented in an FPGA and the side-channel information consists of measuring the perturbations of the power supply voltage. The measurements are averaged and further reduced to their peak-to-peak values. The secret key is extracted by using the correlation power analysis which is a statistical method which postulates the number of key hypotheses and then tests the correlation between the key hypotheses and the side-channel measurements, i.e. the power supply measurements. Two statistical measures of correlation are used in the correlation power analysis: the Pearson correlation coefficient and the distance correlation. The Pearson correlation coefficient is typically used in correlation power analysis and it is sensitive to linear relationships, while the distance correlation is sensitive to

non-linear relationships as well. The two correlation coefficients are compared in the correlation power analysis used in the side-channel attack on the secret key of the multiplication operation. The results show that the measurement is subject to a high level of noise, i.e. the RMS value of the raw measurement is 3 times greater than the RMS value of the averaged waveform and that the measurement averaging is necessary. Measurement results show that the relationship between the measurements and the Hamming weights of the multiplication result is concave. Both correlation coefficients used in the correlation power analysis yield the correct secret key and their values are high as the test digital circuit features only the attacked operation. The analysis also extracts the shifted version of the correct secret key. Analysis results show that the correlation power analysis using the distance correlation yields a more prominent secret key solution when compared to the analysis using the Pearson correlation coefficient. This conclusion is supported by the fact that the distance correlation is sensitive to nonlinear relationships and the measurements show that the power supply is related to the Hamming weight of the multiplication result in a concave manner.

REFERENCES

[1] A.R. Javed et al. "AlphaLogger: detecting motion-based side-channel attack using smartphone keystrokes." *Journal of Ambient Intelligence and Humanized Computing*, 2020, pp. 1-14.

[2] H. Gupta et al., "Impact of Side Channel Attack in Information Security," *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Dubai, United Arab Emirates, 2019, pp. 291-295.

[3] S. Shijie, K. Chen, Y. Zhang. "Overview of Side Channel Cipher Analysis Based on Deep Learning," *Journal of Physics: Conference Series*, vol. 1213, no. 2, IOP Publishing, 2019.

[4] S. Mangard, E. Oswald, T. Popp, "Power analysis attacks: Revealing the secrets of smart cards," *Springer Science & Business Media*, vol. 31, 2008.

[5] I. Kabin et al., "Horizontal Attacks against ECC: from Simulations to ASIC," *Computer Security*, Springer, Cham, 2019, pp. 64-76.

[6] R. Mayer-Sommerm, "Smartly analyzing the simplicity and the power of simple power analysis on smartcards," *International Workshop on Cryptographic Hardware and Embedded Systems*, Springer, Berlin, Heidelberg, 2000.

[7] E. Brier, C. Clavier, F. Olivier, "Correlation power analysis with a leakage model," *International Workshop on Cryptographic Hardware and Embedded Systems*, Springer, Berlin, Heidelberg, 2004.

[8] T. H. Le et al., "A proposition for correlation power analysis enhancement," *International Workshop on Cryptographic Hardware and Embedded Systems*, Springer, Berlin, Heidelberg, 2006.

[9] Y. Ding et al., "Block-oriented correlation power analysis with bitwise linear leakage: An artificial intelligence approach based on genetic algorithms," *Future Generation Computer Systems*, 2020, vol. 106, pp. 34-42.

[10] T. Unterluggauer, E. Wenger, "Practical attack on bilinear pairings to disclose the secrets of embedded devices," *Ninth International Conference on Availability, Reliability and Security*, IEEE, 2014, pp. 69-77.

[11] D. Jauvart et al., "Improving side-channel attacks against pairing-based cryptography," *International Conference on Risks and Security of Internet and Systems*, Springer, Cham, 2016, pp. 199-213.

[12] N. El Mrabet, G. Di Natale, M. L. Flottes, "A practical differential power analysis attack against the miller algorithm," *Ph. D. Research in Microelectronics and Electronics*, IEEE, 2009, pp. 308-311.

[13] L. Batina, S. Bhasin, D. Jap, S. Picek, "{CSI}{NN}: Reverse Engineering of Neural Network Architectures Through Electromagnetic Side Channel," *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019, pp. 515-532.

[14] G. J. Székely, M. L. Rizzo, N. K. Bakirov. "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, 2007, vol. 35.6, pp. 2769-2794.

A Realization of Adaptive Compressive Sensing System

K. Sever*, T. Vlašić** and D. Seršić**
* Ericsson Nikola Tesla d.d. Research and Development Centre, Zagreb, Croatia
** University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
E-mails: karlo.sever@ericsson.com, {tin.vlasic, damir.sersic}@fer.hr

***Abstract* - There are numerous papers analyzing the theoretical background of compressive sensing (CS), but a few practical implementations exist due to the realization complexity. In this paper, a realization of a system for CS of analog one-dimensional signals in a shift-invariant subspace is proposed. We use statistical compressive sensing that allows efficient sampling of signals that follow some statistical distribution. In fact, principal component analysis is used to obtain a small number of principal components in which the majority of signal's energy is compacted. The signal is demodulated using the obtained principal components, which provides optimal measurement procedure in the mean-square error sense. Orthonormality allows for a linear reconstruction which is perfectly suited for a real-time embedded system implementation. Due to hardware realization constraints, the adaptive demodulating signal is quantized. We provide qualitative and quantitative validations based on simulations of the proposed CS system realization. Furthermore, we validate performance of the system when additive white Gaussian noise is present in the measurements.**

Keywords – sparse representation; statistical compressive sensing; inverse problems; signal reconstruction

I. INTRODUCTION

The conventional bridge between continuous and discrete-time signals is the Nyquist-Shannon sampling theorem [1], [2], which states that the sampling frequency must be at least twice the maximum frequency present in a signal. There are several problems associated with the theorem, that provide motivation to seek for a different sampling technique such as the fact that real world signals are seldomly bandlimited and often better represented in other bases than the Fourier basis [3]. Furthermore, the Nyquist-Shannon approach results in a high sampling rate whenever a signal has a wide bandwidth, even if the information rate is low. However, the Nyquist-Shannon theorem can be seen as the projection of the signal onto the bandlimited subspace and only the signals included within the reconstruction space can be reconstructed perfectly [4]. This formulation of the classical theorem leads to more realistic sampling and reconstruction that are obtained by filtering operations in a more general class of shift-invariant (SI) spaces [3], [4]. Alternatively, information content of a signal can be leveraged to create a more efficient sampling and reconstruction technique that overcomes the limitations of the classical sampling theorem. Signal sparsity in a transform domain is a type of information description that inspired the development of compressive sensing (CS) [5], [6], a sampling and reconstruction technique that aims to go beyond the Nyquist limit.

Compressive sensing is a non-adaptive signal acquisition process that aims to completely recover the input signal from an incomplete set of linear measurements. Consider a discrete signal $\boldsymbol{x} \in \mathbb{R}^N$ and a transform basis $\boldsymbol{\Psi}$ in which we can represent the signal as the inner product of the transform basis and a vector $\boldsymbol{s} \in \mathbb{R}^N$:

$$\boldsymbol{x} = \boldsymbol{\Psi s}. \tag{1}$$

Signal $\boldsymbol{x}$ that has a sparse representation in the transform basis $\boldsymbol{\Psi}$ is represented by $\boldsymbol{s}$ that has only K nonzero entries, where $K \ll N$. Signal acquisition in CS is a process of measuring the signal through linear projections by a measurement matrix $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$, where $M < N$. Signal measurements $\boldsymbol{y} \in \mathbb{R}^M$ can be written as:

$$\boldsymbol{y} = \boldsymbol{\Phi x} + \boldsymbol{w} = \boldsymbol{\Phi \Psi s} + \boldsymbol{w}, \tag{2}$$

where $\boldsymbol{w}$ is a disturbance term usually assumed to be small and independent of $\boldsymbol{s}$. Being that the number of measurements M is less than the length of the signal, the problem in (2) is ill-posed and $\boldsymbol{s}$ is recovered by solving an l_1 optimization problem.

Compressive sensing framework has focus primarily set on discrete-time signals. However, there were several frameworks developed that leverage CS to the analog domain. In [7], Tropp *et al.* introduce the random demodulator (RD), an analog-to-information conversion (AIC) system for acquiring of sparse bandlimited signals. The signal is demodulated by a high-rate pseudo-random sequence of ± 1s, known as chipping sequence; integrated, and sampled at a low rate. Another efficient hardware implementation of an AIC system front-end, named modulated wideband converter (MWC), is introduced in [8]. The MWC consists of p channels, where $2K \leq p \ll N$, that mix the signal with a chipping sequence prior to low-pass filtering and following by low-rate sampling. There are a few other realizations of the CS front-end that are basically modifications of the RD and the MWC, and

This research was supported in part by the Croatian Science Foundation under the project IP-2019-04-6703, in part by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS), and in part by Ericsson Nikola Tesla d.d. and University of Zagreb Faculty of Electrical Engineering and Computing under the project ILTERA.

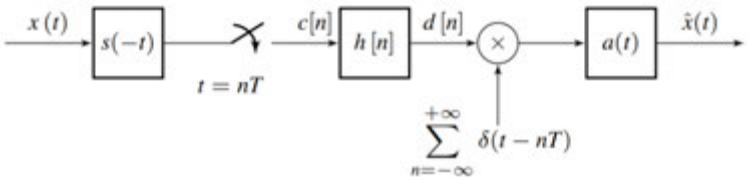


Fig. 1. Sampling and reconstruction of a signal in shift-invariant spaces.

are summarized in [9]. A method for low rate sampling of analog signals in a union of shift-invariant subspaces is proposed in [10]. The sparsity is modeled by treating the case in which only k out of N signal generators are active. The signal is filtered by a bank of only p filters, where $2k \leq p \ll N$, and is uniformly sampled at the sub-Nyquist rate before being recovered. In [11], [12], [13], the authors introduce and develop one-bit CS that extends the scope of sparse recovery by showing that sparse signals can be accurately reconstructed when their linear measurements are subject to the extreme quantization of binary samples (only the sign of each measurement is maintained). In all of the mentioned CS systems the reconstruction is based on solving an l_1 minimization problem.

In this paper, a realization of CS system for acquisition and reconstruction of signals in a shift-invariant subspace that follow some statistical distribution is proposed. The front-end consists of parallel channels that mix, integrate, and sample an input signal. The signal is mixed by functions that lie in an SI subspace characterized by an arbitrary sequence of coefficients. A known statistical distribution and subspace priors of the signal allow for reconstruction from a reduced set of measurements by statistical compressive sensing (SCS) [14], [15]. A small number of principal components obtained by the principal component analysis (PCA) on a training set of signals determine coefficients of the mixing functions. Thus, the measurements obtained by the proposed front-end are optimal in the mean-square error (MSE) sense. Statistical CS allows for a linear recovery of the coefficients, which is equivalent to knowing the orthogonal projection of the signal onto the mixing functions' subspace.

Experimental results of a system realization are presented in the paper. Due to the hardware realization constraints and contrarily to quantization of measurements in one-bit CS, we quantize piecewise constant mixing functions. Qualitative and quantitative validations of the experimental results are provided based on quantization resolution. We, as well, validate the performance of the system when additive white Gaussian noise (AWGN) is present in the measurements.

This paper is organized as follows. In Section II we give a short review of sampling in SI spaces. In Section III the SCS is explained. We propose a SCS measurement model for sampling and reconstruction of signals in B_0-spline space in Section IV. Finally, we provide experimental results in Section V and conclude the paper in section VI.

II. Sampling and Reconstruction in SI Spaces

Let us assume that a signal $x(t)$ lies in an appropriate SI subspace $\mathcal{A}$ of L_2 which can be expressed as a linear combination of the integer shifts of a generator [3]. Any $x(t) \in \mathcal{A}$ is of the form [4]

$$x(t) = \sum_{n \in Z} d[n] a(t - nT), \tag{3}$$

where $a(t)$ is the generator of $\mathcal{A}$. Any signal $x(t)$ is uniquely characterized by a sequence of coefficients $d[n]$ which are not necessarily samples. For bandlimited signals, $a(t)$ corresponds to the *sinc* function. However, one can define a much broader class of generators that are easier to handle.

In order to guarantee a unique signal representation in $\mathcal{A}$ and a stable sampling framework, the generator $a(t)$ is typically chosen to form a Riesz basis or a frame [16]. A countable set of vectors $\{a(t - nT)\}$ is a Riesz basis for L_2 if it is complete and there exist constants $\alpha > 0$ and $\beta < +\infty$ such that [4]:

$$\alpha \|\boldsymbol{d}\|_{l_2}^2 \leq \left\| \sum_{n \in Z} d[n] a(t - nT) \right\|^2 \leq \beta \|\boldsymbol{d}\|_{l_2}^2, \tag{4}$$

where $\|\boldsymbol{d}\|_{l_2}^2 = \sum_n |d[n]|^2$ is the squared l_2 -norm (or energy) of $d[n]$. This ensures that a small modification of coefficients $d[n]$ results in a small distortion of the signal.

Fig. 1. shows a scheme of sampling and reconstruction of signals in an SI subspace. Generally, the signal is prefiltered prior to uniform sampling with a prefilter $s(-t)$ [3], [4], [10]. The samples $c[n]$ can be expressed as

$$c[n] = \int_{-\infty}^{\infty} x(t) s(t - nT) \triangleq \langle x(t), s(t - nT) \rangle, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the L_2 inner product. Let us denote the sampled cross-correlation sequence by

$$r_{sa}[n] = \langle s(t - nT), a(t) \rangle \tag{6}$$

and its discrete time Fourier transform (DTFT) pair by φ_{SA}. The signal $x(t) \in \mathcal{A}$ can be perfectly reconstructed from samples $c[n]$ if [3]

$$\left| \varphi_{SA}(e^{j\omega}) \right| > \alpha > 0, \tag{7}$$

where the DTFT of $r_{sa}[n]$ is given by

$$\varphi_{SA}(e^{j\omega}) = \frac{1}{T} \sum_{k \in Z} S^* \left(\frac{\omega}{T} - \frac{2\pi}{T} k \right) A \left(\frac{\omega}{T} - \frac{2\pi}{T} k \right). \tag{8}$$

Here, $(\cdot)^*$ denotes the complex conjugate, and $A(\omega)$ and $S(\omega)$ are the continuous-time Fourier transform of $a(t)$ and $s(t)$, respectively. Coefficients $d[n]$ are recovered by filtering the samples $c[n]$ using a discrete-time filter [3], [4], [17]:

$$H(e^{j\omega}) = \frac{1}{\varphi_{SA}(e^{j\omega})}, \tag{9}$$

since:

$$C(e^{j\omega}) = D(e^{j\omega})\varphi_{SA}(e^{j\omega}). \tag{10}$$

In the case when basis functions $a(t)$ and $b(t)$ are orthonormal, the cross-correlation sequence $r_{sa}[n] = \delta[n]$. Thus, the filter $H(e^{j\omega}) = 1$ and the samples $c[n]$ do not have to be filtered in the discrete time before the reconstruction. Obtained coefficients $d[n]$ are modulated by an impulse train $\sum_{n\in Z}\delta(t-nT)$ prior to filtering with the corresponding analog filter $a(t)$, for the purpose of reconstructing $x(t)$. Efficient implementation of (9) is particularly possible in spline spaces, when generator $a(t)$ and prefilter $s(t)$ are both *B*-splines [4], [18], [19].

III. Statistical Compressive Sensing

Statistical compressive sensing uses statistical models in order to develop an efficient measurement strategy for signals that follow some statistical distribution [14], [20]. The statistical CS framework introduces an optimal reconstruction algorithm implemented through linear filtering that is computationally less complex than the reconstruction algorithms based on optimization. This paper employs the SCS framework by extracting statistical features from collections of signals in B_0-spline subspace. The coefficients follow a Gaussian distribution, thus enabling development of an optimal measurement technique that leads to linear signal reconstruction.

Signals are divided into K intervals of length N characterized by N expansion coefficients. Each of the signal intervals $\{\boldsymbol{x}_k\}_{k=1}^{K} \in \mathbb{R}^N$ are assumed to follow a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the mean vector and covariance matrix, respectively. In order to induce sparsity, the PCA is applied to the Gaussian model to obtain a small number of principal components in which the majority of signal's energy is compacted. Energy compaction property of the PCA can be useful if observed as a special case of structured sparsity [20]. For a given set of signals, mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ are calculated. Applying the eigenvalue decomposition on the covariance matrix results in obtaining the principal components in $\boldsymbol{V}$ and a diagonal matrix $\boldsymbol{\Lambda}$ whose diagonal elements are the sorted eigenvalues of the covariance matrix:

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T. \tag{11}$$

A measurement matrix $\boldsymbol{\Phi} \in \mathbb{R}^{M\times N}$, where $M < N$, consists of M principal components corresponding to the largest eigenvalues:

$$\boldsymbol{\Phi} = [\mathbf{V}_{N-M+1:N}^T]. \tag{12}$$

Under the assumption that the variance reflects the informational content of the signal, the obtained SCS measurement matrix $\boldsymbol{\Phi}$ optimally describes the observed signal in terms of the MSE [20]. Rows in $\boldsymbol{\Phi}$ can be denoted as a set of vectors $\{\boldsymbol{\phi}_i\}_{i=1}^{M}$. Compressive sensing measurement is a linear process that computes inner products between the signal and a set of measurement vectors, $y_i[k] = \langle \boldsymbol{x}_k, \boldsymbol{\phi}_i \rangle$. In statistical CS, measurement procedure can be seen as the projection of the signal onto the principal components. The signal $\boldsymbol{x}_k$ is reconstructed from the measurements $\boldsymbol{y}[k] = \{y_i[k]\}_{i=1}^{M}$ by a decoder that minimizes the MSE, which is obtained by a linear maximum *a posteriori* (MAP) estimator [14], [20], [21]:

$$\boldsymbol{x}_k = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}[k]. \tag{13}$$

Instead of computationally complex, time consuming solving of the l_1 optimization problem in conventional CS, Gaussian SCS has an optimal decoder (13) calculated via a closed-form linear filtering for any $\boldsymbol{\Phi}$.

IV. A SCS Measurement Model in B_0-Spline Space

A single front-end channel of the proposed system consists of a mixer, an ideal integrator, and an ADC. Typically, the mixing (or demodulating) signal is a piecewise constant function with pseudo-random amplitudes. Such demodulating functions are frequent in the conventional CS frameworks [7], [8], [22]; in which measurement precedes the l_1 optimization. This paper focuses on mixing of the signal by a deterministic piecewise constant function, where both the signal and mixing function lie in a B_0-spline subspace. We define a generator in the B_0-spline subspace as a constant causal signal with a duration T. The mixing function in the i-th channel is characterized by the i-th row of the SCS measurement matrix $\boldsymbol{\Phi}$ in (12) and can be written as:

$$z_i(t) = \sum_{m=1}^{N} \phi_i[m] \sum_{k\in Z} s(t - mT - kNT), \tag{14}$$

where $s(t)$ is a generator of the B_0-spline subspace. The mixing function is periodic with period NT, and values $\phi_i[m]$ are cyclically repeated. The mixed signal passes through an ideal integrator before being sampled with period $\tau = NT$. Note that the sampling rate is N times lower than the rate of the subspace basis functions. Fig. 2. illustrates the acquisition scheme.

A single measurement in the i-th channel and k-th interval is given by:

$$y_i[k] = \int_{k\tau}^{(k+1)\tau} z_i(t)x(t)dt. \tag{15}$$

By applying (3) and (14), (15) can be expanded to:

$$\begin{aligned} y_i[k] &= \sum_{m=1}^{N} \phi_i[m] \sum_{n\in Z} d[n] \\ &\quad \cdot \int_{k\tau}^{(k+1)\tau} s(t - mT - k\tau)a(t - nT)dt \\ &= \sum_{m=1}^{N} \phi_i[m] \sum_{n=1}^{N} d[n + kN] \\ &\quad \cdot \int_{k\tau}^{(k+1)\tau} s(t - mT - k\tau)a(t - nT - k\tau)dt. \end{aligned} \tag{16}$$

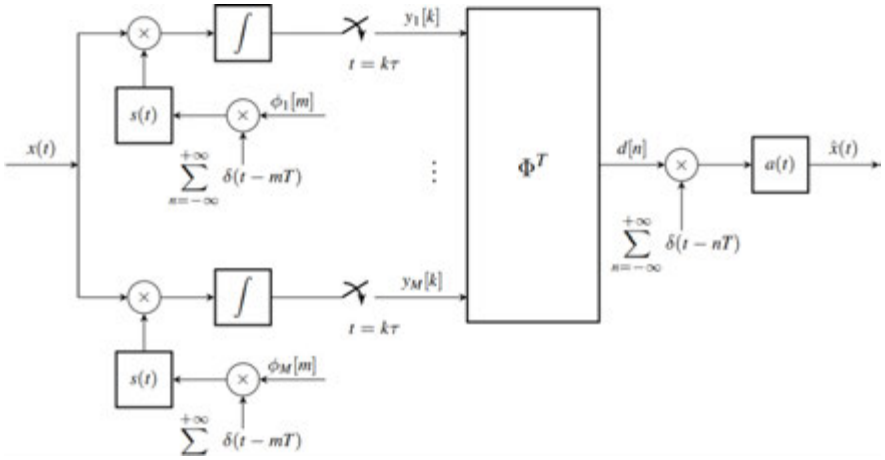


Fig. 2. Statistical compressive sensing of a signal in B_0-spline space. The signal is multiplied by $\boldsymbol{M}$ mixing functions generated by analog filters $\boldsymbol{s(t)}$ and sequences $\{\boldsymbol{\phi_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{M}}$ corresponding to the principal components. The sequences $\{\boldsymbol{\phi_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{M}}$ are modulated by an impulse train before being filtered by $\boldsymbol{s(t)}$.

The integration in (16) can be arranged in a matrix $\boldsymbol{R}$. By introducing $\boldsymbol{R}$, (16) becomes

$$y_i[k] = \sum_{m=1}^{N} \phi_i[m] \sum_{n=1}^{N} d[n+kN] R_{m,n}. \tag{17}$$

Since the basis functions of the B_0-spline subspace are orthonormal, the integral in (16) is equal to $\delta_{m,n}$. It follows that $\boldsymbol{R}$ is an $N \times N$ identity matrix. A single CS measurement in the i-th channel is then given by:

$$y_i[k] = \sum_{m=1}^{N} \phi_i[m] d_k[m], \tag{18}$$

where $d_k[m] = d[m+kN]$.

After finding the expression for a single measurement, we can write the measurement procedure of the whole system in a matrix form:

$$\begin{bmatrix} \phi_1[1] & \cdots & \phi_1[N] \\ \vdots & \vdots & \vdots \\ \phi_i[1] & \cdots & \phi_i[N] \\ \vdots & \vdots & \vdots \\ \phi_M[1] & \cdots & \phi_M[N] \end{bmatrix} \begin{bmatrix} d_k[1] \\ \vdots \\ d_k[N] \end{bmatrix} = \begin{bmatrix} y_1[k] \\ \vdots \\ y_M[k] \end{bmatrix} \tag{19}$$

or

$$\boldsymbol{\Phi d}_k = \boldsymbol{y}[k]. \tag{20}$$

Once we obtained the measurement vector $\boldsymbol{y}[k]$, the optimal linear decoder (13) is applied in order to recover the expansion coefficients $\boldsymbol{d}_k$. The recovered $\boldsymbol{d}_k$ are projections of signal $x(t)$ onto the SI subspace spanned by B_0-spline basis functions $\{s(t-mT)\}$. An optimal measurement matrix for the coefficients $\boldsymbol{d}_k$ is then obtained by applying the PCA to a training set of expansion coefficients which follow a Gaussian distribution. That way the decoder for coefficients $\boldsymbol{d}_k$ is optimal in the MSE sense.

The signal reconstruction procedure is illustrated in Fig. 2. Since B_0-spline basis functions are orthonormal, there is no need for discrete-time filtering when compared to the scheme in Fig. 1., i.e. $h[n] = \delta[n]$. The obtained coefficients $d[n]$ are modulated by an impulse train with period T and filtered by the B_0-spline analog filter $a(t)$. Due to the properties of SCS, the reconstructed signal has the information rate same as the input signal, despite being sampled at a far lower rate.

V. Experimental Results

The proposed CS system was simulated by acquiring and reconstructing signals from the multi-speaker speech Sinhala TTS dataset [23]. We assume that the samples represent coefficients of a signal that lies in the B_0-spline subspace. Next, the coefficients are assumed to follow a Gaussian distribution. In the preprocessing phase, the signal is centered with respect to the mean so that the expansion coefficients $d[n]$ follow Gaussian distribution $\mathcal{N}(0, \boldsymbol{\Sigma})$. The PCA is applied to the training dataset in order to obtain principal components, which construct a measurement matrix as described in Section III. Test signals are divided into intervals of length N, i.e. the sampling rate in a single channel is N times lower than the rate of the B_0-spline basis functions. This leads to a system with the sampling rate of $M/(NT)$, where M represents the number of measurement channels in the front-end of the system. The compressive sensing measurement ratio is defined by $r = M/N$. We simulate an efficient low-cost front-end with a small number of measurement channels, from $M = 2$ to $M = 12$. Such a front-end can be designed as an extended circuit connected to an embedded system development board. Thus, we treat cases where test signals are divided into rather small intervals of length $N = 10$ and $N = 20$.

We quantize values of the measurement matrix entries to simulate a real-world implementation. The measurement matrix is quantized for three different cases where the number of bits used in quantization is $n_q = \{1, 2, 4\}$, in order to provide 2^{n_q} levels of a mixing function. For $n_q = 1$, the mixing function is a chipping sequence of $\pm 1s$, which is particularly practical to implement. Alternatively, when the number of quantization bits are $n_q > 1$, the mixers can be realized by an off-the-shelf multiplying digital-to-analog converter, e.g. [24].

TABLE I. SIGNAL-TO-NOISE RATIOS OF RECONSTRUCTED SIGNALS IN DECIBELS

Measurements done without additive noise						**Measurements done in presence of additive noise**					
$N = 10$	r = 0.2	r = 0.3	r = 0.4	r = 0.5	r = 0.6	$N = 10$	r = 0.2	r = 0.3	r = 0.4	r = 0.5	r = 0.6
$n_q = 1$	0.69	3.94	7.05	7.79	9.82	$n_q = 1$	0.67	3.87	6.87	7.57	9.44
$n_q = 2$	2.61	5.50	7.80	8.79	9.87	$n_q = 2$	2.57	5.36	7.53	8.48	9.42
$n_q = 4$	2.47	7.44	10.47	12.47	14.49	$n_q = 4$	2.43	7.22	10.04	11.81	13.45
ideal	2.48	7.61	10.77	12.93	15.14	ideal	2.43	7.38	10.28	12.15	13.88
$N = 20$	r = 0.2	r = 0.3	r = 0.4	r = 0.5	r = 0.6	$N = 20$	r = 0.2	r = 0.3	r = 0.4	r = 0.5	r = 0.6
$n_q = 1$	3.90	6.98	8.24	9.51	10.36	$n_q = 1$	3.83	6.79	7.98	9.12	9.88
$n_q = 2$	5.47	8.44	7.85	8.65	10.16	$n_q = 2$	5.34	8.15	7.64	8.39	9.78
$n_q = 4$	6.94	11.67	13.83	15.56	17.15	$n_q = 4$	6.75	11.05	12.90	14.23	15.31
ideal	6.98	11.84	14.15	16.03	17.65	ideal	6.79	11.19	13.11	14.52	15.57

Validating the impact of quantization on the success of a reconstruction is done by using four different measurement matrices, specifically the ideal measurement matrix and three quantized matrices for a given number of bits $n_q = \{1, 2, 4\}$. In the simulations, the measurement ratio is set to $r = \{0.2, 0.3, 0.4, 0.5, 0.6\}$. Signal-to-noise ratio (SNR) is used to assess the quality of the linear reconstruction algorithm described in Section IV. The SNR is defined as the ratio of powers of the reconstructed signal and the reconstruction error. Table I. shows reconstruction results for different scenarios based on the measurement ratio and the quantization resolution. The left-hand side of Table I. shows calculated SNRs of reconstructions that include the effects of quantization errors without the presence of noise. Notice that the SNR values for $n_q = 4$ are close to the ideal case. Thus, the proposed SCS system shows robustness to the quantization errors. Furthermore, the resolution dependency effect [25] is visible in the reconstruction results with respect to the number N of coefficients being recovered in a single SCS reconstruction process. The reconstruction quality for the same measurement ratio gets better as N increases. Note that the reconstruction is obtained by multiplying the pseudoinverse of the PCA basis (13) with the measurement vector. Even better results are achievable by applying the l_1 optimization as in the conventional CS setting, but that way the computational complexity is much higher.

In order to simulate the system more realistically, we include additive noise in the simulations. We model the disturbance term in (2) by AWGN and add it to the measurement vector $\boldsymbol{y}[k]$. We add AWGN so that the SNR of the measurement vector is 20 dB. The right-hand side of Table I. shows calculated SNRs of reconstructions that include the effects of the quantization and additive noise. The results show that after introducing additive noise to our CS system, we can still expect reconstructions with similar SNRs. The system remains robust to the effect of quantization even in the presence of additive noise. The resolution dependency effect is visible for reconstructions from noisy measurements, as well. A small segment of the original and reconstructed signal is shown in Fig. 3.

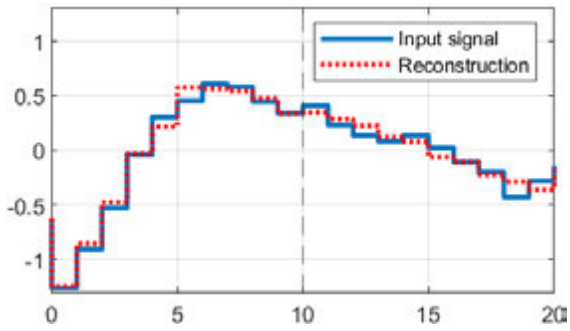


Fig. 3. A segment of a signal from Sinhala TTS dataset represented by B_0-spline basis functions and its reconstruction. A visualization for N=10, M=4, $\boldsymbol{n_q}$=4, with additive noise and SNR=17.08 dB for this particular segment.

VI. CONCLUSION

We proposed a realization of a compressive sensing system for acquisition of signals that follow a Gaussian distribution and lie in a B_0-spline subspace. An input signal is mixed by functions that are characterized by the principal components of a Gaussian distribution, that way minimizing the mean-square error. The signal is reconstructed from measurements that are linear projections of the signal onto the mixing functions by linear filtering. The realization is perfectly suited for an implementation of the reconstruction on an embedded device. We provided extensive qualitative validations based on simulations of a realistic realization of the proposed system. The simulations showed that the proposed system can be efficiently realized with a small number of quantization bits and is robust to the additive noise.

REFERENCES

[1] H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *Transactions of the American Institute of Electrical Engineers,* vol. 47, no. 2, pp. 617-644, 1928.

[2] C. E. Shannon, "Communication in the Presence of Noise," *Proc. IRE,* vol. 37, no. 1, pp. 10-21, 1949.

[3] Y. C. Eldar and T. Michaeli, "Beyond Bandlimited Sampling," *IEEE Signal Process. Mag. ,* vol. 26, no. 3, pp. 48-68, May 2009.

[4] M. Unser, "Sampling-50 years after Shannon," *Proc. IEEE,* vol. 88, no. 4, pp. 569-587, April 2000.

[5] E. J. Candes, J. Romberg and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory,* vol. 52, no. 2, pp. 489-509, Feb. 2006.

[6] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory,* vol. 52, no. 4, pp. 1289-1306, April 2006.

[7] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg and R. G. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Trans. Inf. Theory,* vol. 56, no. 1, pp. 520-544, Jan. 2010.

[8] Y. C. Eldar and M. Mishali, "From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals," *IEEE J. Sel. Topics Signal Process.,* vol. 4, no. 2, pp. 375-391, April 2010.

[9] M. Rani, S. B. Dhok and R. B. Deshmukh, "A Systematic Review of Compressive Sensing: Concepts, Implementations and Applications," *IEEE Access,* vol. 6, pp. 4875-4894, 2018.

[10] Y. C. Eldar, "Compressed Sensing of Analog Signals in Shift-Invariant Spaces," *IEEE Trans. Signal Process.,* vol. 57, no. 8, pp. 2986-2997, Aug. 2009.

[11] P. T. Boufounos and R. G. Baraniuk, "1-Bit compressive sensing," in *Proc. of the 42nd Annual Conference on Information Sciences and Systems (CISS), IEEE*, 2008.

[12] L. Jacques, J. N. Laska, P. T. Boufounos and R. G. Baraniuk, "Robust 1-bit compressive sensing via stable embeddings of sparse vectors," *IEEE Trans. Inform. Theory,* vol. 59, pp. 2082-2102, 2013.

[13] K. Knudson, E. Saab and R. Ward, "One-bit compressive sensing with norm estimation," *IEEE Trans. Inform. Theory,* vol. 62, pp. 2748-2758, 2016.

[14] G. Yu and G. Sapiro, "Statistical Compressed Sensing of Gaussian Mixture Models," *IEEE Trans. Signal Process.,* vol. 59, no. 12, pp. 5842-5858, Dec. 2011.

[15] J. M. Duarte-Carvajalino, G. Yu, L. Carin and G. Sapiro, "Task-Driven Adaptive Statistical Compressive Sensing of Gaussian Mixture Models," *IEEE Trans. Signal Process.,* vol. 61, no. 3, pp. 585-600, Feb. 2013.

[16] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inf. Theory,* vol. 36, pp. 961-1005, Sept. 1990.

[17] M. Unser and A. Aldroubi, "A general sampling theory for nonideal acquisition devices," *IEEE Trans. Signal Process.,* vol. 42, no. 11, pp. 2915-2925, Nov. 1994.

[18] M. Unser, "Splines: A perfect fit for signal and image processing," *IEEE Signal Process. Mag.,* vol. 16, no. 6, pp. 22-38, Nov. 1999.

[19] M. Unser, A. Aldroubi and M. Eden, "B-spline signal processing. I. Theory," *IEEE Trans. Signal Process.,* vol. 41, no. 2, pp. 821-833, Feb. 1993.

[20] I. Ralašić, A. Tafro and D. Seršić, "Statistical compressive sensing for efficient signal reconstruction and classification," in *2018 4th International Conference on Frontiers of Signal Processing (ICFSP)*, Poitiers, 2018.

[21] S. M. Kay, Fundamentals of Statistical Signal Processing, vol. I, Englewood Cliffs, NJ: Prentice-Hall, 1998.

[22] T. Vlašić, J. Ivanković, A. Tafro and D. Seršić, "Spline-Like Chebyshev Polynomial Representation for Compressed Sensing," in *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Dubrovnik, Croatia, 2019.

[23] Google, Inc., "Open Speech and Language Resources," [Online]. Available: http://openslr.org/52/. [Accessed 15 January 2020].

[24] Analog Devices, "AD5424," [Online]. Available: https://www.analog.com/en/products/ad5424. [Accessed 28 January 2020].

[25] B. Roman, A. Hansen and B. Adcock, "On asymptotic structure in compressed sensing," *arXiv:1406.4178,* 2014.

Delta/Alpha Ratio for Diagnosis of Autism Spectrum Disorder

Z. Šverko*, S. Vlahinić *, M. Vrankić* and I. Markovinović*
* University of Rijeka – Faculty of Engineering/Department of Automatics and Electronics, Rijeka, Croatia
zsverko@riteh.hr, sasa.vlahinic@riteh.hr, miroslav.vrankic@riteh.hr, imarkovinovic@riteh.hr

***Abstract* - EEG data signal processing has been performed in order to outline the parameters for the comparison between participants with autism (5 subjects) and typically developing children acting as the control group (14 subjects). Based on the previous studies, the autistic group has a significantly higher delta/alpha ratio than the typical developing children. In this paper, we have tested this hypothesis on new participants. Furthermore, statistical features such as mean value, grand-mean value and their ratios have been calculated. The research was conducted by observing signals on electrodes F3, Fz, F4, C3, Cz, C4, P3, Pz and P4. We observed power delta-alpha, theta-high beta, theta-low beta, theta-alpha ratio, grand mean values of magnitude spectrum and 1st moments of the grand mean values of the spectral distribution. The statistically significant difference between observed groups we found when tested 1st moments of the grand mean values of the spectral distribution.**

Keywords - EEG, signal processing, autism, delta/alpha ratio

I. INTRODUCTION

Defined by the Diagnostics and Statistics Manual (DSM-5) [1], types of disorders named neurodevelopmental disorders are a category of heterogeneous malady classified with a disruption or a delay in skill gaining in different developmental domains. These developmental domains are: social, motor, cognitive and language domain. The neurodevelopmental disorder diagnosis includes autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD), intellectual disability (ID) and global developmental disorder (GDD) [1].

ASD is a category of complex and heterogeneous developmental disorders containing multiple neural system dysfunctions [2]. Currently, diagnosis of autism spectrum disorder is still based on extensive clinical interviews and behavioral reconnaissance. The Childhood Autism Rating Scale (CARS) and the Checklist for Autism in Toddlers [3] are frequently used forms for defining diagnosis, instead of using laboratory tests.

A method which is used for measuring neurophysiological changes linked to postsynaptic activity in the neocortex is named EEG (electroencephalography). From practical and theoretical points of view, the EEG can be used as an optimal biomarker for identifying neurodevelopmental disorders [4]. If we compare functional magnetic resonance imaging (fMRI – low temporal resolution, high spatial resolution, requires extensive funding) with EEG, we can say that EEG has a bigger temporal resolution (used for resolving problems on a millisecond scale). Furthermore, EEG has the advantage that can be used throughout an expanded range of age groups, it also has a higher tolerance for movement, EEG devices are more accessible for clinical and general use. Moreover, it can be used to collect gathering reiterated measurements because it is non-invasive and it studies different developmental disabilities [2]. That the EEG is a powerful tool for research complex neuropsychiatric disorders is shown in [2; 5-8].

The neurophysiological procedure that applies computerized mathematical analysis to transform the raw EEG data recorded from a multi-electrode into different frequency ranges (primarily delta, theta, alpha and beta frequency bands) is named Quantitative Electroencephalography (qEEG) [9]. Delta waves appear as a rule in deep sleep. It is assumed that it is related with tasks for detection of attention and salience [10]. While observing memory processes, theta waves were studied [11]. Alpha waves appear in the relaxed awake state of the subjects; beta waves are related to the active task engagement, motor behavior and alertness [12]. Gamma waves are related to the working memory [13] and early sensory responses [2; 14,15].

Power can be calculated in two ways, as an absolute and a relative power [16]. The quantity of the EEG activity in one band independent of activity in other bands is called absolute power. Respectively, the quantity of EEG activity in an individual band divided by the quantity in all bands is called relative power.

The analysis of absolute power is recommended for improving the understanding of electrophysiological changes caused by ASD and for construing the differences in relative power in ASD group. [17].

Alterations in the EEG related to age was studied [9] and it can be concluded that ASD group (4-7 years) have a decrease of absolute and relative alpha power and an increase of relative theta power in comparison with typically developing children (control group). Opposite to that, while observing older participants, these changes are not noted, but it is the consequences of the fact that there is a decreasing of power on low frequencies and increasing on high frequencies from childhood to adulthood in control group [9]. 'U' shape is proposed [2], with the explanation that ASD participants have an increasing power in low frequencies and high frequencies, and decreasing power in alpha and beta bands in

comparison with the control group. In opposition, the observed resting state EEG of ASD participants [18] showed an increase of alpha power and coherence. Furthermore, the decreasing of power in low-frequency bands (delta, theta) is presented in [19]. Moreover, an increase of power in the left hemisphere in comparison with right hemisphere is studied in [20-22]. Delta/alpha ratio (DAR), relative alpha and relative delta power were studied on five-minute qEEG data in 90 normal control participants and 66 children with ASD [23]. The results of the study were considerably lower relative alpha and higher relative delta and delta–alpha ratio [23].

Based on these previous studies and results, we resolved to do this study by observing power delta-alpha ratio. In desire to expand the previous studies, we also observed theta-high beta (THBR), theta-low beta (TLBR) and theta-alpha ratio (TAR). Furthermore, we observed grand mean values of magnitude and 1st moments of the grand mean values of the spectral distribution. For the purpose of this study, we monitored EEG signals recorded on the F3, Fz, F4, C3, Cz, C4, P3, Pz and P4 electrodes. The placement of these electrodes is shown in Figure 1 encircled in blue.

II. METHODS

A. Experiment description

The experiment was performed on fourteen [24] typically developing participants and five participants with clinically diagnosed autism spectrum disorder. The recording procedure included open eyes, closed eyes, hyperventilation, post-hyperventilation, body moving [24]. Measurements were made without causing any physical and psychological pain to participants.

B. EEG recordings

The EEG recordings were conducted using wet passive electrodes and Nicolet EEG system with 32 channels (sampling rate 512 Hz). EEG electrodes were positioned based on International 10-20 standard for electrode placement. Ground electrode was located on right mastoid and Reference electrode was located on left mastoid.

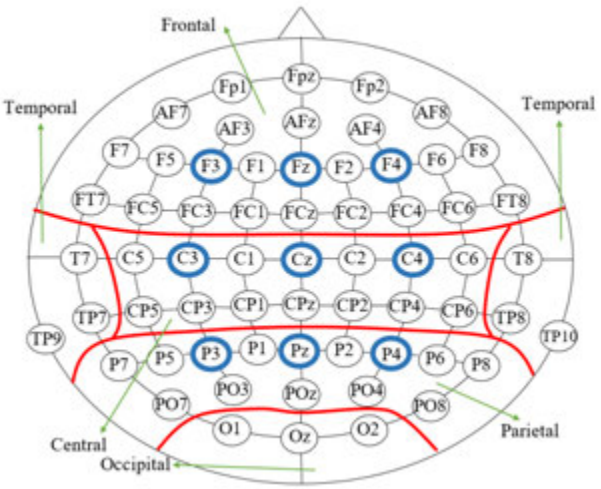


Figure 1. Placement of electrodes [24].

III. ANALYSES OF MEASUREMENTS

The Matlab toolbox EEGLab was used to import recordings. The analysis procedure was done based on the following steps:

- Step 1: Normalization of raw data and removing DC component
- Step 2: Filtration with 32 order bandpass IIR filter; 1 – 47 Hz
- Step 3: Removal of artifacts above/below ± 70 μV thresholds
- Step 4: FFT calculation repeatedly implemented on non-overlapping epochs of 1 s duration, giving 1 Hz resolution
- Step 5: Features extraction.

Furthermore, after FFT calculation, spectrum was divided into six frequency ranges (delta (δ_f) 1 – 4 Hz, theta (θ_f) 4 – 8 Hz, alpha (α_f) 8 – 13 Hz, low beta (β_{Lf}) 13 – 18 Hz, high beta (β_{Hf}) 18-30 Hz, gamma (γ_f) 35 – 45 Hz).

Comparisons were made between two groups of participants (ASD and control group) based on power ratios, mean power value, grand mean power value, standard deviation. The formula for calculating the relative power value of each frequency band is given with (1):

$$P_{\Delta band} = \frac{\sum_{\Delta band} \left(|f_i|\right)^2}{\sum_{all} \left(|f_i|\right)^2}, \quad (1)$$

- f_i – FFT coefficients at i-th spectral component
- $P_{\Delta band}$ – relative power of independent band.

In Fig. 2, grand mean values of magnitude spectral distributions on C3, Cz, C4, F3, Fz, F4, P3, Pz and P4 are given. Increasing values of magnitude spectral distribution are presented while monitoring ASD participants through all the observing electrodes in comparison with control group participants. This fact is stable, observing all subbands except alpha subbands on P3, Pz and P4 electrodes. It is noticeable that while observing control groups on these electrodes we can see the increase magnitude in the alpha subband in comparison with other electrodes. We can conclude that there is a noticeable shape of the magnitude spectral distributions of control group like inverted "U" shape on these electrodes (the lower magnitude values of the theta, the higher of the alpha, the lower of the low beta band). Furthermore, we can confirm [2] that the magnitude spectral distributions of ASD participants have a shape little bit similar to "U" (the higher magnitude values of the delta and theta band, the lower of the alpha and low-beta band, the higher of the high-beta band) on all electrodes.

Statistical analysis is performed for the grand mean values of the spectral distributions and results are presented in numerical format (Table 1). When observing 1st moments of the grand mean values of the spectral distributions, we can conclude that ASD participants have

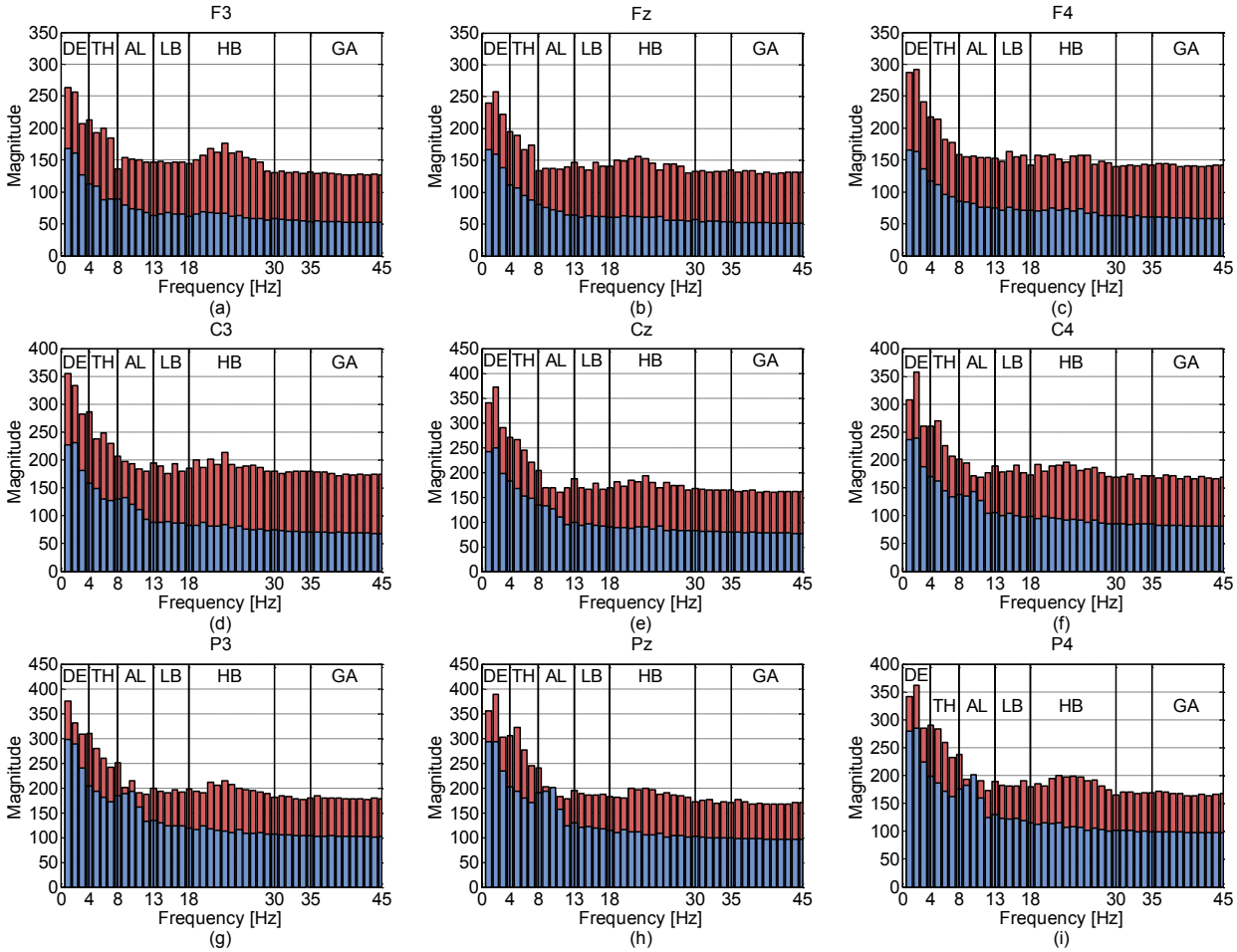


Figure 2. Grand-mean values of magnitude spectral distributions on C3, Cz, C4, F3, Fz, F4, P3, Pz, P4 electrode, for ASD (red) and control group (blue) participants (DE – delta, TH – theta, AL – alpha, LB – low-beta, HB – high-beta, GA – gamma band)

about two times higher values in comparison with control group participants.

Features were statistically analyzed using Excel. Comparison between the observed groups was done using student t-test (two tailed, [9]).

The p-value ≤ 0.05 was considered as a confirmation of significant difference between groups.

Furthermore, because p-values are lower than 0.05 on F3, Fz, F4, C3 and Pz, we can conclude that there is statistically significant difference observing 1st moments of the grand mean values of the spectral distribution between ASD and control group. The highest difference is on F3 electrode.

TABLE I. 1ST MOMENTS OF THE GRAND-MEAN VALUES OF THE SPECTRAL DISTRIBUTIONS

Electrode	1st Moments of the grand mean values of the spectral distributions ± SD, p-value)		
	Control	*Patients*	*p-value*
F3	58.30±40.62	135.68±13.66	0.00000949
Fz	57.10±32.25	136.78±63.76	0.0464
F4	64.52±36.42	147.67±60.64	0.0338
C3	78.01±41.28	182.26±51.40	0.0065
Cz	86.63±40.29	170.26±80.61	0.0799
C4	90.58±44.40	175.97±93.70	0.1113
P3	113.24±57.95	187.95±74.69	0.0902
Pz	108.12±57.62	180.16±55.41	0.0412
P4	108.50±50.04	176.56±82.97	0.1441

Moreover, grand mean values of power ratios are observed. Fig. 3 shows the grand-mean values of power ratios (TAR – theta/alpha ratio, DAR – delta/alpha ratio, TLBR –heta/low beta ratio and THBR – theta/high beta ratio) on all observing electrodes. The increase of TAR and DAR on C3, C4, Cz (central electrodes) and on P3, Pz, P4 (parietal electrodes) can be seen from Fig 3. When observing TLBR and THBR we could not notice any difference between ASD and control group of participants based on visual comparison. Therefore, a statistical analysis is made to research significant statistical difference between the observed groups.

Table 2 and Table 3 show grand mean values and standard deviations of power ratios. P-values are also shown in these tables. Furthermore, because p-values are bigger than 0.05, we can conclude that there is statistically no significant difference between ASD and control group based on participants included in this study on all electrodes.

IV. CONCLUSION

In this paper, a new study was done to confirm the previously studied conclusions and test new hypothesis to define the difference between clinically diagnosed ASD

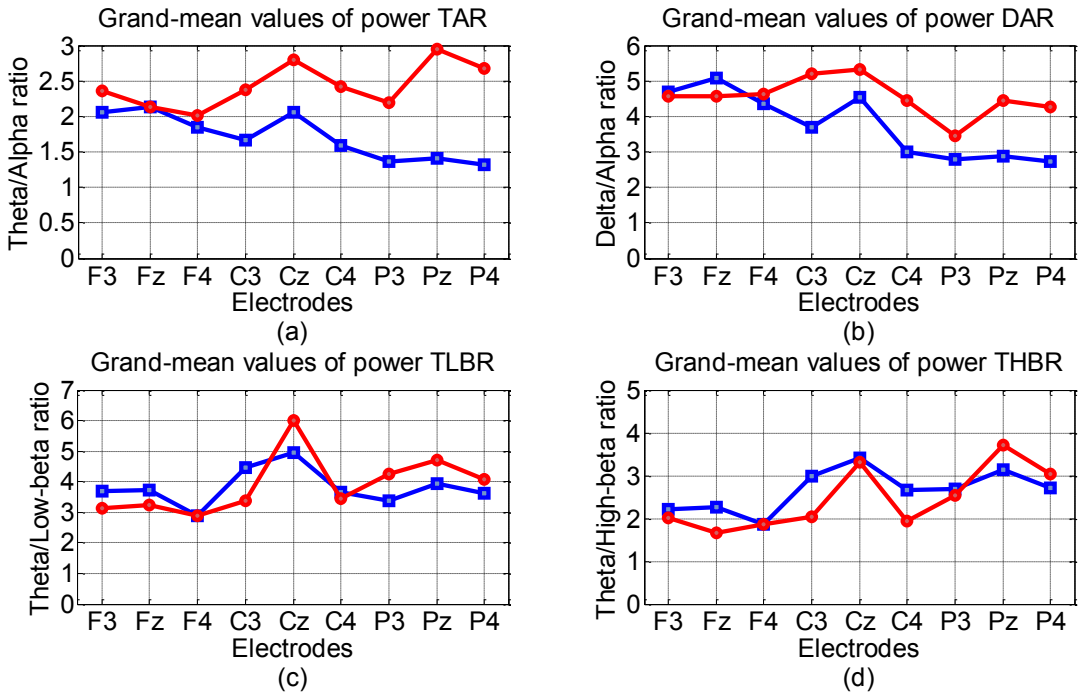


Figure 3. Grand-mean values of power ratios on all observing electrodes (TAR – theta-alpha ratio, DAR – delta-alpha ratio, TLBR – theta-low beta ratio, THBR – theta-high beta ratio) for ASD (red) and control group (blue) participants.

TABLE II. GRAND-MEAN VALUES OF POWER THETA-ALPHA AND DELTA-ALPHA RATIOS

Electrode	Theta-alpha (grand mean ± SD, p-value)			Delta-alpha (grand mean ± SD, p-value)		
	Control	*Patients*	*p-value*	*Control*	*Patients*	*p-value*
F3	2.06±0.85	2.35±0.72	0.473	4.7±2.09	4.58±2.19	0.923
Fz	2.13±0.6	2.14±0.64	0.985	5.07±1.85	4.58±2.04	0.646
F4	1.84±0.36	2±0.6	0.583	4.35±1.77	4.63±2.06	0.796
C3	1.66±0.48	2.38±1.27	0.279	3.69±1.82	5.19±3.19	0.364
Cz	2.05±0.68	2.79±1.65	0.379	4.53±3.53	5.31±3.69	0.694
C4	1.6±0.5	2.42±0.75	0.0678	2.99±1.05	4.44±2.34	0.243
P3	1.36±0.61	2.19±0.89	0.109	2.79±1.26	3.46±1.9	0.495
Pz	1.4±0.63	2.94±1.7	0.112	2.89±2.01	4.43±3.04	0.337
P4	1.32±0.56	2.68±1.4	0.090	2.74±1.78	4.26±3.06	0.343

TABLE III. GRAND-MEAN VALUES OF POWER THETA-LOW BETA AND THETA-HIGH BETA RATIOS

Electrode	Theta-low beta (grand mean ± SD, p-value)			Theta-high beta (grand mean ± SD, p-value)		
	Control	*Patients*	*p-value*	*Control*	*Patients*	*p-value*
F3	3.7±2.25	3.13±0.98	0.449	2.2±1.52	2.02±1.17	0.793
Fz	3.73±1.86	3.24±2.57	0.710	2.25±1.32	1.67±1.24	0.399
F4	4.35±1.77	2.89±1.34	0.087	1.85±1.27	1.86±1.2	0.993
C3	4.46±2.82	3.37±2.05	0.378	2.98±2	2.05±1.55	0.318
Cz	4.94±2.52	5.99±4.7	0.656	3.42±2.39	3.31±2.86	0.941
C4	3.65±1.3	3.45±1.94	0.840	2.67±1.6	1.94±1.33	0.349
P3	3.38±1.25	4.24±2.01	0.411	2.69±1.52	2.54±1.69	0.867
Pz	3.93±2.01	4.7±2.6	0.568	3.13±1.63	3.71±3.04	0.705
P4	3.62±1.39	4.09±2.67	0.721	2.71±1.32	3.03±2.85	0.823

participants and control group. According to the results of this study it can be reported that "U" shape is confirmed, showing a magnitude spectral distributions of ASD participants, which is reported as conclusions in previous studies. Furthermore, statistical analysis was performed calculated 1st moments of the grand mean values of spectral distributions. Observing this, we found statistically significant difference between ASD and control group on the F3, Fz, F4, C3 and Pz electrodes.

Moreover, a hypothesis was investigated, stating that differences in DAR are significant between these groups. Statistical analyses have confirmed that the differences were statistically nonsignificant. Moreover, the difference between TAR, TLBR and THBR was analyzed, but we have concluded that there is statistically no significant differences as well.

For current and future studies in the diagnostic difference between ASD and control group, we suggest a comparison of EEG based on subtypes of ASD (language problems, communication skills, intellectual ability etc.). Also, because of the small number of subjects, the results presented in this paper needs to be verified on larger sample.

REFERENCES

[1] American Psychiatric Association, Diagnostic and statistical manual of mental disorder: DSM-5, 5th ed., Washington, American Psychiatric Association, 2013.

[2] J. Wang, J. Barstein, L. E. Ethridge, M. W. Mosconi, Y. Takarae and J. A. Sweeney, "Resting state EEG abnormalities in autism spectrum disorders", Journal of Neurodevelopmental Disorders 2013, vol. 24, pp. 5-24, September 2013.

[3] E. Schopler, R. J. Reichler, R. F. DeVellis and K. Daly: "Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS)", Journal of Autism and Developmental Disorders, vol. 10 (1), 91-103, March 1980.

[4] S. S. Jeste, J. Fronhlich and S. K. Loo, "Electrophysiological biomarkers of diagnnosis and outcome in neurodevelopmental disorders", Current opinion in Neurology, vol. 28 (2), pp. 110-116, April 2015.

[5] C. A. Mann, J. F. Lubar, A. W. Zimmerman, C.A. Miller and R. A. Muenchen, "Quantitative analysis of EEG in boys with attention-deficit-hyperactivity disorder: controlled study with clinical implications", Pediatric Neurology, vol. 8, pp. 30–36, January-February 1992.

[6] J. Wang, R. Brown, K. R. Dobkins, J. E. McDowell and B. A. Clementz, "Diminished parietal cortex activity associated with poor motion direction discrimination performance in schizophrenia", Cerebral Cortex, vol. 20, pp. 1519–1755, July 2010.

[7] L. E. Ethridge, J. P. Hamm, J. R. Shapiro, A. T. Summerfelt, S. K. Keedy, M. C. Stevens, et al., "Neural activations during auditory oddball processing discriminating schizophrenia and psychotic bipolar disorder", Biological Psychiatry, vol. 72, pp.766–774, November 2012.

[8] R. Coben, "The importance of electroencephalogram assessment for autistic disorders", Biofeedback, vol. 37, pp. 71–80, Summer 2009.

[9] H. El-Habashy, O. Raafat, L. Afifi, H. Raafat and K. Abdullah, "Quantitive EEG Ratios and Power in Children with Autistic Spectrum Disorder", Medical Journal of Cairo University, vol. 84, pp. 1017-1027, September 2016.

[10] G.G. Knyazev, "EEG delta oscillations as a correlate of basic homeostatic and motivational processes", Neuroscience & Biobehavioral Reviews, vol. 36, pp. 677–695, January 2012.

[11] W. Klimesch, "Memory processes, brain oscillations and EEG synchronization", International Journal of Psychophysiology, vol. 24, pp. 61-100, November 1996.

[12] C. Neuper and G. Pfurtscheller, "Event-related dynamics of cortical rhythms: frequency-specific features and functional correlates", International Journal of Psychophysiology, vol. 43, pp. 41-58, December 2001.

[13] C. Tallon-Baudry, "Oscillatory synchrony and human visual cognition", Journal of Physiology-Paris, vol. 97, pp. 355–363, March-May 2003.

[14] J. E. Skinner, M. Molnar and Z. J. Kowalik, "The role of the thalamic reticular neurons in alpha- and gamma-oscillations in neocortex: a mechanism for selective perception and stimulus binding", Acta Neurobiologiae Experimentalis 2000, vol. 60, pp. 123–142, 2000.

[15] W. Singer and C. M. Gray, "Visual feature integration and the temporal correlation hypothesis", Annual Review of Neuroscience, vol. 18, pp. 555–586, March 1995.

[16] L. Billeci, F. Sicca, K. Maharatna, F. Apicella, A. Narzisi, Giulia Campatelli, et al., "On the application of quantitive EEG for characterizing autistic brain: A systematic review", Frontiers in Human Neuroscience, vol. 7, pp. 442-456, August 2013.

[17] R. T. Pivik, R. J. Broughton, R. Coppola, R. J. Davidson, N. Fox and M. R. Nuwer, "Guidelines for the recording and quantitative analysis of electroencephalographic activity in research contexts", Psychophysiology, vol. 30, pp. 547–558, November 1993.

[18] R. Srinivasan, P. L. Nunez and R. B. Silberstein, "Spatial filtering and neocortical dynamics: estimates of EEG coherence", IEEE Transactions on Biomedical Engineering, vol. 45, pp. 814–826, July 1998.

[19] T. J. Whitford, C. J. Rennie, S. M. Grieve, C. R. Clark, E. Gordon and L. M. Williams, "Brain maturation in adolescence: concurrent changes in neuroanatomy and neurophysiology", Human Brain Mapping, vol. 28, pp. 228–237, February 2007.

[20] T. A. Stroganova, G. Nygren, M. M. Tsetlin, I. N. Posikera, C. Gillberg, M. Elam, et al., "Abnormal EEG lateralization in boys with autism", Clinical Neurophysiology, vol. 118, pp. 1842–1854, August 2007.

[21] A. M. Daoust, E. Limoges, C. Bolduc, L. Mottron and R. Godbout, "EEG spectral analysis of wakefulness and REM sleep in high functioning autistic spectrum disorders", Clinical Neurophysiology, vol. 115, pp. 1368–1373, June 2004.

[22] S. K. Sutton, C. P. Burnette, P. C. Mundy, J. Meyer, A. Vaughan, C. Sanders, et al., "Resting cortical brain activity and social behavior in higher functioning children with autism", The Journal of Child Psychology and Psychiatry, vol. 46, pp. 211–222, September 2005.

[23] A. S. Chan, S. L. Sze and M. Cheung, "Quantitative Electroencephalographic Profiles for Children With Autistic Spectrum Disorder", Neuropsychology, American Psychological Association, vol. 21, pp. 74–81, 2007.

[24] Z. Šverko, S. Vlahinić, M. Vrankić and I. Markovinović, "EEG Data Processing in ADHD Diagnosis and Neurofeedback", in press.

Group Delay of Fractional n+α-Order Bessel Filters

Andro Ćoza*, Vesna Županović*, Domagoj Vlah* and Dražen Jurišić*
*University of Zagreb/Faculty of electrical engineering and computing, Unska 3, Zagreb, 10000, Croatia
Phone: (+385 1) 612 99 49; Fax: (+385 1) 612 96 52
e-mail: drazen.jurisic@fer.hr

***Abstract* - In this paper we present the realization of fractional-order analog filter with Bessel approximation, having linear phase response, which is also usually represented by a constant group delay. There are many applications such as analogue video signal processing, radar and sonar receivers, hard disk drive read channel applications, analog front end of biomedical signal processing, where linear phase response is desirable. Optimally designed Bessel filter can provide better transient response in the passband, it reduces overshoot, ringing and provides minimal phase distortion. In this work we research Bessel approximation with the non-integer order n+α; n is integer number and 0<α<1. The influence of non-integer filter order to the value of the constant group delay of the filter is presented. It is shown that the group delay value can be additionally tuned by changing the parameter α.**

Keywords - active filters; analog circuits; circuit synthesis; filtering theory; Bessel approximation; linear-phase response; group delay; fractional-order filters.

I. INTRODUCTION

In recent times fractional-order calculus has gained significant attention by many researchers in modelling biological tissues, describing processes in control theory, materials theory, diffusion theory, robotics, signal processing, oscillators, filters, viscoelasticity, etc. [1][2].

This paper researches advantages of applying fractional order on Bessel approximations (see also [3]). Most of the research on fractional-order filters has been presented for Butterworth and Chebyshev filters [4]–[6]. It is demonstrated that by including of parameter α, the group delay can be fine-tuned.

In definition of fractional derivative, two approaches exist: 1) fractional differentiations and integrations are limits of finite differentiations and integrations, and 2) generalization of convolution type representation of repeated integration [1][2].

In the second approach the most popular is the Caputo fractional derivative, defined by:

$$D^{\alpha} f(t) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dt} \int_a^t \frac{f(\tau)}{(t-\tau)^{\alpha}} d\tau , \qquad (1)$$

where $0<\alpha<1$. Γ (gamma) function is generalized factorial $\Gamma(p+1) = p!$, where p is integer or non-integer number. The well-known Euler definition of Γ is:

$$\Gamma(p) = \int_0^{\infty} e^{-x} t^{p-1} dx; \ \mathrm{Re}[p] > 0 . \qquad (2)$$

Fractional derivative (1) by Caputo is useful in the circuit and filter theory because it can use standard (integer-order) initial conditions.

It possesses a Laplace transform defined by

$$L\left[D^{\alpha} f(t) \right] = s^{\alpha} F(s) - \sum_{k=0}^{n-1} s^k {}_a D_t^{\alpha-k-1} f(0) . \qquad (3)$$

When the initial conditions are zero, this becomes

$$L\left[D^{\alpha} f(t) \right] = s^{\alpha} F(s) , \qquad (4)$$

simplifying substantially analysis in the frequency domain. In the sequel we will use Laplace transformation in the analysis of fractional-order systems.

II. BESSEL FILTERS

In this Section we briefly repeat the *maximally-flat* constant-phase approximation, which leads to Bessel polynomials [7]. We need this, to compare the results with the new fractional-order (FO) approximation of linear phase to maximally-flat integer-order (IO) approximation.

In approximations of ideal frequency characteristic of Butterworth or Chebyshev filters the emphasis was put on the realization of amplitude-frequency (A-F) characteristic. On the shape of phase-frequency characteristic (P-F) was stated no condition. In real world, it is very often needed to realize electrical systems having linear phase in the pass band. Linear phase is equivalent to the constant group delay. Transfer function, which has constant amplitude characteristic and linear phase in the whole frequency range is defined by:

$$H(s) = K \cdot e^{-sT_0} . \qquad (5)$$

This work was fully supported by the Croatian Science Foundation under project (IP-2016-06-1307) "Fractional analog and mixed systems for signal processing".

For amplitude-frequency characteristic we have to include $s=j\omega$ into (5) and after application of natural logarithm we have:

$$\ln H(j\omega) = \ln K - j\omega T_0 = \alpha(\omega) + j\varphi(\omega)\,, \qquad (6)$$

providing linear phase-frequency characteristic as:

$$\varphi(\omega) = -\omega T_0\,. \qquad (7)$$

Group time delay is then defined as:

$$T_g(\omega) = -\frac{d\varphi(\omega)}{d\omega} = T_0\,, \qquad (8)$$

and it is constant. Note that $\alpha(\omega)$ in (6) is in Nepers.

In literature (e.g. [7]) it is well-known that the above characteristics is satisfied by Bessel polynomials. They are given by

$$\begin{aligned} B_0(s) &= 1 \\ B_1(s) &= s+1 \\ B_2(s) &= s^2+3s+3 \\ B_3(s) &= s^3+6s^2+15s+15 \\ B_4(s) &= s^4+10s^3+45s^2+105s+105 \end{aligned} \qquad (9)$$

and so on. They can be calculated from recursive equation (starting from known B_0 and B_1):

$$B_n(s) = (2n-1)B_{n-1}(s) + s^2 B_{n-2}(s)\,. \qquad (10)$$

If the Bessel filter transfer function (5) must have the d.c. gain $H(0)=1$ then, for example, using $B_3(s)$ from (9) for the third-order filter it is given by ($K=15$):

$$H_3(s) = \frac{15}{s^3+6s^2+15s+15}\,. \qquad (11)$$

This transfer function has group delay T_0 equal to unity. If we want to normalize the Bessel filters, that they have magnitudes comparable to those of, for example, Butterworth filters, they must have gain of $1/\sqrt{2}$ at $\omega=1$. For this to be accomplished, the group delays T_0 differ dependent on the filter order n, and are given in Table I.

TABLE I. CONSTANT T_0 VALUES

n	1	2	3	4	5	6	7
$\mathbf{T_0}$	1.0	1.362	1.756	2.114	2.424	2.699	2.947

The denormalization of the transfer function is performed by the frequency transformation $s \to s/\omega_0$, where the denormalization frequency is $\omega_0 = 1/T_0$. If it is applied, for example, on our third-order Bessel characteristic, having $T_{03}=1.756$, its transfer function is given by:

$$\begin{aligned} H_3(s) &= \frac{15}{(T_{03}s)^3 + 6\cdot(T_{03}s)^2 + 15\cdot(T_{03}s) + 15} \\ &= \frac{2.78449}{s^3 + 3.4227s^2 + 4.88212s + 2.78449}\,. \end{aligned} \qquad (12)$$

Phase characteristics is given by

$$\varphi_3(\omega) = -\arctan\frac{4.86636\omega - \omega^3}{2.77179 - 3.41749\omega^2}\,, \qquad (13)$$

while for the group delay we have:

$$\begin{aligned} T_{g3}(\omega) &= 1.756\cdot\frac{7.68284 + 4.73629\omega^2 + 1.924654\omega^4}{7.68284 + 4.73629\omega^2 + 1.924654\omega^4 + \omega^6} \\ &= 1.756\left(1 - \frac{\omega^6}{7.6828 + 4.7363\omega^2 + 1.92465\omega^4 + \omega^6}\right). \end{aligned} \qquad (14)$$

Note that the nth-order Bessel group delay has the form:

$$\begin{aligned} T_{gn}(\omega) &= T_{0n}\cdot\frac{a_0 + a_2\omega^2 + \ldots + a_{2(n-1)}\omega^{2(n-1)}}{a_0 + a_2\omega^2 + \ldots + a_{2(n-1)}\omega^{2(n-1)} + \omega^{2n}} \\ &= T_{0n}\cdot\left(1 - \frac{\omega^{2n}}{a_0 + a_2\omega^2 + \ldots + a_{2(n-1)}\omega^{2(n-1)} + \omega^{2n}}\right). \end{aligned} \qquad (15)$$

By including $\omega=0$ into (15) we obtain a value of the group delay $T_{gn}(0)=T_{0n}$. This special form of (15) is also known as *maximally flat* approximation. Note also, although the phase characteristic in (13) is a transcendent function, the group delay is a rational function of two polynomials [7].

Amplitude-frequency characteristics, phase-frequency characteristics and group delays are shown in Figs. 1(a), (b) and (c), respectively for the Bessel functions up to 4th-order. It can be seen that the magnitude functions have all −3dB cut off frequency equal to 1 rad/s, and posses different group delays. Different group delays are visible in Fig. 1(c), and they are given in Table I.

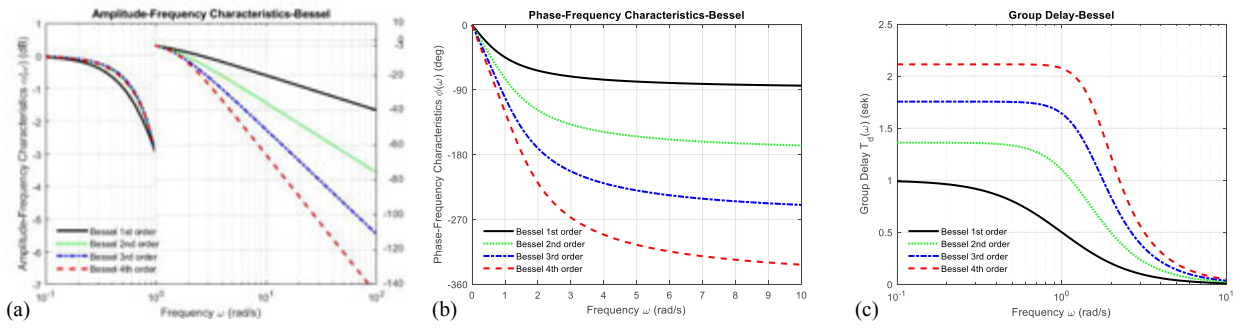


Figure 1. Classical Bessel filters of 1st to 4th-order. (a) Amplitude-frequency characteristics, (b) Phase-frequency characteristics, (c) Group delay.

It is not possible *to continuously adjust the values of the group delay* T_0 when using integer order Bessel filters, because only discrete values are possible. In what follows it will be shown that it is possible to obtain other values of group delays in-between the values given in Table I by using fractional-order Bessel filters, only.

III. FRACTIONAL-ORDER BESSEL FILTERS

The normalized transfer function of the fractional low-pass filter of order $1+\alpha$ is given by

$$H_{FOB}^{1+\alpha}(s)=\frac{k_1}{s^{1+\alpha}+k_3 s^{\alpha}+k_2}. \tag{16}$$

The low-frequency gain is k_1/k_2.

The 3+α-order FO transfer function (which is a fourth-order when α=1) is possible by combining (16), with one biquadratic transfer function, because of the stability issues [4]. The obtained transfer function is given by:

$$\begin{aligned} H_{FOB}^{3+\alpha}(s)&=\frac{N(s)}{D(s)}=\frac{a_4}{s^2+a_6 s+a_5}\cdot\frac{a_1}{s^{1+\alpha}+a_3 s^{\alpha}+a_2}\\ &=\frac{k_1}{s^{3+\alpha}+k_7 s^{2+\alpha}+k_6 s^2+k_5 s^{1+\alpha}+k_4 s+k_3 s^{\alpha}+k_2}. \end{aligned} \tag{17}$$

The values k_1 to k_7 are the positive constants to be obtained numerically to approximate Bessel passband amplitude-frequency characteristic.

We apply the nonlinear least square routine from Matlab called *lsqcurvefit* in the frequency range ω from 0 to 1 to the transfer functions defined by (17) [4][5]. This function uses the trust-region-reflective algorithm with termination tolerances of the function value and the solution set in Matlab by default to 10^{-6}. The routine *lsqcurvefit* is defined in Matlab as:

$$\begin{aligned} &\min_x \left\| \left|H_{FOB}^{j-1+\alpha}(x,\omega)\right|-\left|H_{INT_j}(\omega)\right| \right\|_2^2\\ &=\min_x \sum_{i=1}^{k}\left(\left|H_{FOB}^{j-1+\alpha}(x,\omega_i)\right|-\left|H_{INT_j}(\omega_i)\right|\right)^2, \end{aligned} \tag{18}$$

where j is the order of the integer-order transfer function, k is the total number of frequency points for the comparison, and x is the filter coefficients (k_1 to k_7) vector. The lowest bound for the solutions vector x was set to 0.1. In our example, the order of the fractional transfer function is chosen 3.5 (with n=3 and α=0.5). The obtained coefficient values of 3.5th-order FO transfer functions are given in Table II on the next page. The corresponding normalized amplitude-frequency characteristics, phase-frequency characteristics, and group time delays for 3rd-, 3.5th- and 4th-order filters are shown in Figs. 2(a) to (c), respectively. We usually have to denormalize the transfer function [e.g. (17)] to some cut-off frequency, e.g. to ω_c and the curves in Fig. 2 to the $f_c=\omega_c/(2\pi)$. Then we have to apply the frequency transformation $s\rightarrow s/\omega_c$. Frequency characteristics of fractional-order filters are readily obtainable using FOMCON Matlab toolbox [8].

Besides, A-F, P-F and T_g characteristics of the FO transfer functions, such as (17), can be calculated analytically. We do it by substitution $s=j\omega$ into s^{α}, which is defined by:

$$(j\omega)^{\alpha}=\omega^{\alpha}\cdot e^{j\alpha\frac{\pi}{2}}=\omega^{\alpha}\cdot\left[\cos\left(\alpha\frac{\pi}{2}\right)+j\sin\left(\alpha\frac{\pi}{2}\right)\right]. \tag{19}$$

Using a short notation $H=N/D=H_{FOB}^{3+\alpha}(j\omega)$, where N means numerator and D denominator, we obtain the following expressions for the three characteristics:

$$\left|H(j\omega)\right|=\frac{k_1}{\sqrt{\mathrm{Re}^2[D]+\mathrm{Im}^2[D]}}, \tag{20}$$

$$\varphi(\omega)=\angle H=-\arctan\frac{\mathrm{Im}[D]}{\mathrm{Re}[D]}, \tag{21}$$

$$T_g(\omega)=\frac{\mathrm{Re}[D]\frac{d}{d\omega}\mathrm{Im}[D]-\mathrm{Im}[D]\frac{d}{d\omega}\mathrm{Re}[D]}{\mathrm{Re}^2[D]+\mathrm{Im}^2[D]}. \tag{22}$$

Where, after some calculations, we have:

$$\begin{aligned} \mathrm{Re}[D]&=k_2+k_3\omega^{\alpha}\cos\left(\alpha\frac{\pi}{2}\right)+k_5\omega^{1+\alpha}\cos\left[(1+\alpha)\frac{\pi}{2}\right]\\ &-k_6\omega^2+k_7\omega^{2+\alpha}\cos\left[(2+\alpha)\frac{\pi}{2}\right]+\omega^{3+\alpha}\cos\left[(3+\alpha)\frac{\pi}{2}\right], \end{aligned} \tag{23}$$

$$\begin{aligned} \mathrm{Im}[D]&=k_3\omega^{\alpha}\sin\left(\alpha\frac{\pi}{2}\right)+k_4\omega+k_5\omega^{1+\alpha}\sin\left[(1+\alpha)\frac{\pi}{2}\right]\\ &+k_7\omega^{2+\alpha}\sin\left[(2+\alpha)\frac{\pi}{2}\right]+\omega^{3+\alpha}\sin\left[(3+\alpha)\frac{\pi}{2}\right]. \end{aligned} \tag{24}$$

The derivations of (23) and (24) to ω gives:

$$\begin{aligned} \frac{d}{d\omega}\mathrm{Re}[D]&=\frac{1}{\omega^{1-\alpha}}\left[k_3\alpha-k_7(2+\alpha)\omega^2\right]\cos\left(\alpha\frac{\pi}{2}\right)\\ &-2k_6\omega+\omega^{\alpha}\left[-k_5(1+\alpha)+(3+\alpha)\omega^2\right]\sin\left(\alpha\frac{\pi}{2}\right), \end{aligned} \tag{25}$$

$$\begin{aligned} \frac{d}{d\omega}\mathrm{Im}[D]&=\frac{1}{\omega^{1-\alpha}}\left[k_3\alpha-k_7(2+\alpha)\omega^2\right]\sin\left(\alpha\frac{\pi}{2}\right)\\ &+k_4+\omega^{\alpha}\left[k_5(1+\alpha)-(3+\alpha)\omega^2\right]\cos\left(\alpha\frac{\pi}{2}\right). \end{aligned} \tag{26}$$

For the FO Bessel filter with α=0.5, i.e. for its transfer function of the order 3.5, instead using FOMCON we can plot the FO curves using expressions (20) to (26). Thus, A-F characteristic in (20) is shown in Fig. 2(a) in dB. Also the P-F characteristic in (21) and T_g in (22) of the order 3.5 are in Figs. 2(b) and (c), respectively. All those FO curves are plot as a magenta dotted line.

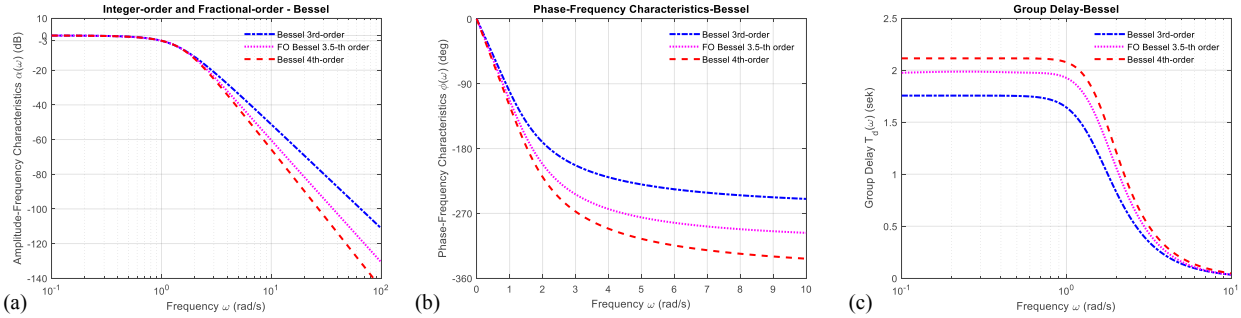


Figure 2. Bessel filters of 3rd, 3.5th and 4th-order. (a) Amplitude-frequency characteristics, (b) Phase-frequency characteristics, (c) Group delay.

To calculate the group delay T_g for the fractional-order Bessel filter depending on the fractional parameter α we can use (22), in which we also have to substitute (23) to (26). Because, some of those expressions have ω in the denominator we cannot include ω=0, as we did above into (15), to obtain group delay $T_g(0)=T_0$. Instead, because all transfer functions are normalized to cut-off frequency equal to 1, we can use a smaller value of ω=0.1 and we will obtain $T_{03.5}$=1.9752 from (22). Note that the obtained value for the order 3.5 is between T_{03}=1.756 and T_{04}=2.114, which can also be seen in Fig. 2(c).

If we want to find the dependence of the group delay $T_{0n+\alpha}$ on the value of the fractional parameter α, and especially if we want to go the other way around, to find the order $n+\alpha$ for given $T_{0n+\alpha}$ the things are going to be unsolvable in an analytical way. We also have to notice that not only because the expressions (22) to (26) are implicit, and cannot be expressed in the inverse form providing α, all the coefficients k_i; i=1,...,7 in (17) that are obtained by optimization are determined by the values of α and n. In our case those were n=3, and α=0.5, and for the new values of $n+\alpha$ we will obtain the new values of the constants k_i. Therefore, we must form the nomograph connecting the values of $T_{0n+\alpha}$ and $n+\alpha$, which goes beyond the scope of this paper.

Furthermore, as mentioned above, when denormalizing a filter to a cut-off frequency f_c, we can calculate the real group delays T_{0c} of the filter circuit using:

$$T_{0c} = \frac{T_0}{\omega_c} = \frac{T_0}{2\pi f_c}. \tag{27}$$

E.g., if f_c=1kHz from above normalized values of T_{03}=1.756, $T_{03.5}$=1.9752 and T_{04}=2.114 using (27) we have

$$T_{03c} = 279.4\mu s;\ T_{03.5c} = 314.3\mu s;\ T_{04c} = 336.4\mu s. \tag{28}$$

Stability check. For the FO system to be stable the angles of all poles in (17) must be greater than the minimum required angle $|\theta_W| > \pi/2m = 9°$ (see also [4]). It is obvious from Table II that the presented filter is stable.

TABLE II. COEFFICIENT VALUES OF 3.5TH FO BESSEL FILTER

Order	k_1	k_2	k_3	k_4
n+α=3.5	2.9219	2.9132	0.1000	5.3243
	k_5	k_6	k_7	$\lvert\theta_W\rvert_{\min}$
	1.2695	2.6296	2.5214	13.2411

IV. FILTER REALIZATION

A. *Active-RC filter circuit*

In this section we will physically realize 4th-order and 3.5th-order transfer function using active-RC filters with operational amplifiers in Pspice. Consider first Tow-Thomas circuit shown in Fig. 3.

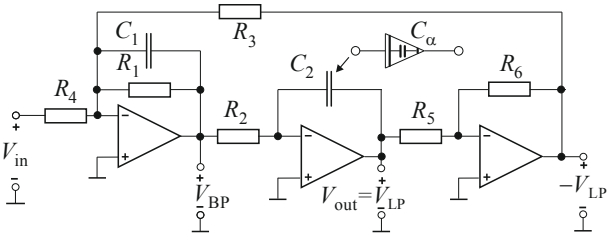


Figure 3. Tow Thomas circuit

If the circuit contains one fractional capacitance C_α, which replaces the capacitor C_2, then the transfer function of the circuit is given by:

$$H(s) = \frac{V_{out}(s)}{V_{in}(s)} = \frac{R_3}{R_4} \cdot \frac{\dfrac{1}{C_1 C_\alpha R_2 R_3}}{s^{1+\alpha} + s^\alpha \dfrac{1}{R_1 C_1} + \dfrac{R_6}{C_1 C_\alpha R_2 R_3 R_5}}. \tag{29}$$

In the case of α=1, we can compare (29) to the low-pass transfer function of the second-order in general form, and choosing R_5=R_6 we can recognize the parameters:

$$\omega_p = \frac{1}{\sqrt{C_1 C_2 R_2 R_3}};\ q_p = R_1\sqrt{\frac{C_1}{C_2 R_2 R_3}};\ k = \frac{R_3}{R_4}. \tag{30}$$

If in the design we choose C_1=C_2=C_0, R_2=R_3=R_0, we will have the following simple design expressions:

$$R_0 = \frac{1}{\omega_p C_0};\ R_1 = q_p R_0;\ R_4 = kR_3 = kR_0. \tag{31}$$

The design parameters of the above fourth-order (integer-order) Bessel filter from [7] are given by:

$$\begin{aligned} &\omega_{pA} = 1.43017; q_{pA} = 0.521935; k_A = 1; \\ &\omega_{pB} = 1.60336; q_{pB} = 0.805538; k_B = 1, \end{aligned} \tag{32}$$

for the two sections as in Fig. 3, denoted by A and B, and connected in cascade. We then denormalize the cut-off frequency of the filter to the frequency $\omega_c=2\pi f_c$; f_c=1kHz. We have chosen C_0=1nF, by that we have obtained the denormalization impedance $R_0=(\omega_c C_0)^{-1}$=159.155kΩ. Using above expressions in the design of the fourth-order Bessel filter, i.e. from (32) and using (31) we have obtained the elements given in Table III. We have chosen arbitrarily $R_5=R_6$=10kΩ.

B. Realization of constant-phase element (CPE)

The impedance of a constant-phase element (CPE) is defined by:

$$Z_{CPE}(s)=\frac{1}{s^{\alpha}C_{\alpha}}, \tag{33}$$

where C_α is the pseudo-capacitance with units [F/sec$^{1-\alpha}$] and $0<\alpha<1$ is the order.

To be able to perform Pspice analysis we must replace CPE by a passive RC network as its approximation. Only a fixed approximation for a specific pseudo-capacitance with value of α valid over a bandwidth (around specified center) with acceptable magnitude and phase error is possible. Known methods for approximating the variable s^α are Carlson's method, Matsuda's method, Oustaloup's method and the continued fraction expansion (CFE) method (see also [9]–[12]). They all form a rational function of "*s*" as a driving point function, or immittance function (impedances and admittances are referred to as immittances). Among approximation methods, one of the most efficient is the CFE method, and will be used below.

A CFE method approximates the impedance of the fractional capacitance $Z_{CPE}(s)$ in (33) in the following general form:

$$Z_{CPE}(s)\cong\frac{1}{C\omega_0}\frac{P_0s^4+P_1\omega_0s^3+P_2\omega_0^2s^2+P_3\omega_0^3s+P_4\omega_0^4}{P_4s^4+P_3\omega_0s^3+P_2\omega_0^2s^2+P_1\omega_0^3s+P_0\omega_0^4}, \tag{34}$$

where from [10] we have:

$$\begin{aligned}
P_0&=\alpha^4+10\alpha^3+35\alpha^2+50\alpha+24,\\
P_1&=-4\alpha^4-20\alpha^3+40\alpha^2+320\alpha+384,\\
P_2&=6\alpha^4-150\alpha^2+864,\\
P_3&=-4\alpha^4+20\alpha^3+40\alpha^2-320\alpha+384,\\
P_4&=\alpha^4-10\alpha^3+35\alpha^2-50\alpha+24.
\end{aligned} \tag{35}$$

Note that the chosen order of the approximation in (34) is four, and ω_0 is the center frequency around which the approximation is performed.

According to classical network theory [7] the rational function of *s* in (34) can be realized by a network shown in Fig. 4, which is also known as Foster I one-port.

TABLE III. ELEMENTS OF 1KHZ FOURTH-ORDER BESSEL FILTER

Section	$C_1=C_2$	R_1	$R_2=R_3=R_4$	$R_5=R_6$
A	1.0nF	58.083kΩ	111.28kΩ	10.0kΩ
B	1.0nF	79.961kΩ	99.264kΩ	10.0kΩ

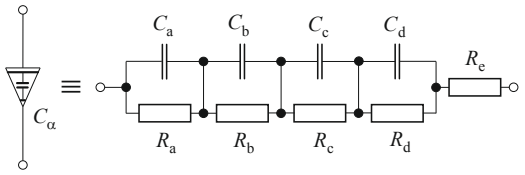


Figure 4. Constant-phase element simulation by Foster I realisation

The expression of the impedance $Z(s)$ of the Foster I network in Fig. 4, is constructed from passive resistors R_i ($i=a, ..., e$) and capacitors C_j ($j=a, ..., d$). They form the following impedance function:

$$Z(s)=\sum_{i=a,b,c,d}\frac{1/C_i}{s+1/(R_iC_i)}+R_e. \tag{36}$$

Furthermore, the rational function of *s* in (34) can be given in the form of partial fraction expansion:

$$Z(s)=\sum_{i=1}^{4}\frac{r_i}{s-p_1}+k, \tag{37}$$

where *k* and r_i are constant terms, p_i are the poles of impedance (34).

The values of passive elements in Fig. 4 readily follow from equating the coefficients of (37) with those derived from the partial fraction expansion of $Z(s)$ in (36). The coefficient *k*, residues r_i and poles p_i in (37) are used to calculate the passive elements using the following expressions:

$$C_i=\frac{1}{r_i},R_i=\frac{1}{C_i|p_i|},R_e=k. \tag{38}$$

Passive element values for approximating a CPE with C=1nF at 1 kHz (i.e. $|Z|$=159.15kΩ), α=0.5, using the Foster I topology in Fig. 4, are given by: R_a=40.063kΩ, C_a=526.27pF, R_b=59.906kΩ, C_b=1.8706nF, R_c=142.89kΩ, C_c=3.3414nF, R_d=1.1718 MΩ, C_d=4.3683nF, R_e=17.684 kΩ. This is a well-know Foster I network synthesis algorithm.

The frequency dependence of the absolute value of the impedance $|Z(j\omega)|$ and the phase angle $\varphi(\omega)=\arg[Z(j\omega)]$ of the CPE are shown in Fig. 5 and 6, respectively.

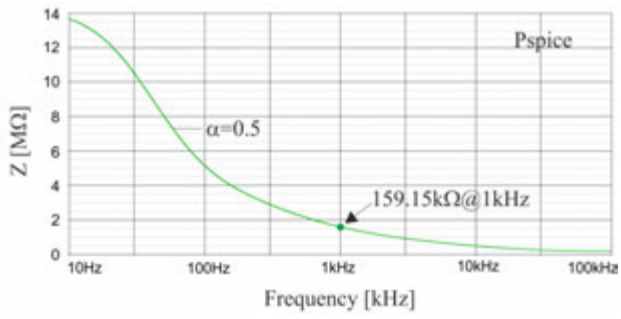


Figure 5. Impedance-frequency characteristic of CPE in Fig. 4

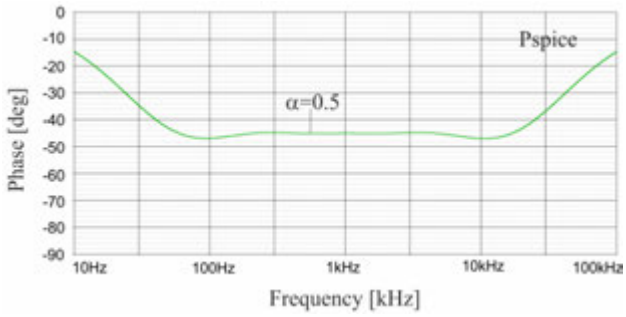


Figure 6. Phase-frequency characteristic of CPE in Fig. 4

C. *Integer-order and fractional-order filter realization and analysis using Pspice*

The integer-order (IO) realization of the fourth-order Bessel filter and fractional-order (FO) Bessel filter of the 3.5th order are analyzed using Pspice.

For the realization of IO Bessel filter the Tow-Thomas section shown in Fig. 3 is used with computed elements in Table II. For the realization of FO Bessel filter Modified-Follow-the-Leader-Feedback (MFLF) realization is used. MFLF is not presented in this paper, because it needs a lot of attention and will be described in details in a follow-up journal publication. A-F characteristics of two filters is shown in Fig. 7. Note that in the case of 4th-order filter, the roll-off at high frequencies is 80dB/decade, while for the 3.5th-order filter, with fractional-step attenuation in the stopband, the roll-off is 70dB/decade. Group delay for two realizations is shown in Fig. 8. It can be noted that the 4th-order filter has T_{0c}=336.6μs, whereas the 3.5th-order filter has T_{0c}=313.7μs. Note that those simulated values are very close to, and confirm, the calculated values (28).

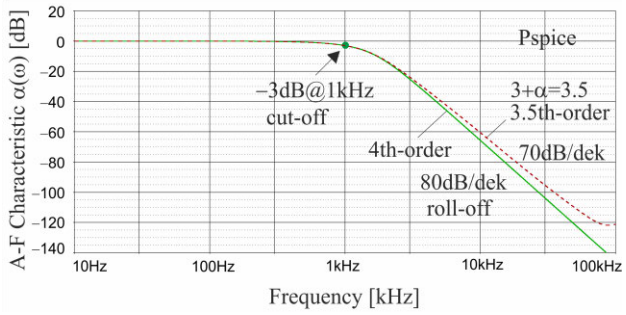


Figure 7. Amplitude-frequency characteristics IO and FO Bessel filters

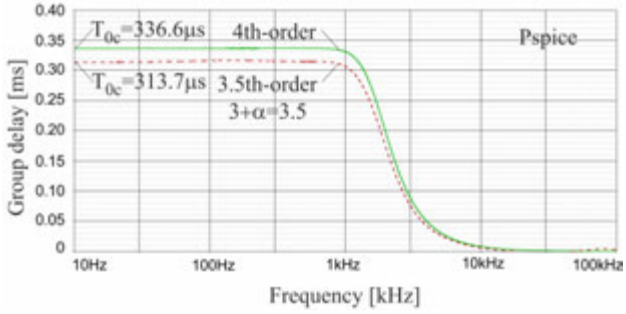


Figure 8. Group delay characteristics IO and FO Bessel filters

We can conclude that the fine adjustment of the group delay T_0 value is impossible with IO Bessel filter design, only. The FO approach provides more degrees of freedom in the design. In doing so besides the fine-tuning of the group delay, we can keep the same amplitude frequency characteristics in the pass-band.

V. CONLUSION

In this paper we present the design of maximally-flat, fractional-order Bessel filter. We have first stated that as a consequence of normalizing the cut-off frequencies of integer-order Bessel filters to 1rad/s (−3dB magnitude), we have only discrete and fixed values of group delays. To obtain more degrees of freedom in the design, we must construct Bessel filters with fractional order. In this way we can fine-tune group delays, and it is possible to obtain more freedom in the design of theirs values, that can be between initial discrete values. To design fractional-order Bessel filters we have used the least square optimization routine in Matlab. With different group delays, we succeed to keep the same amplitude frequency characteristics in the pass-band. The group delay of the fractional-order Bessel filters is increasing with increasing order (i.e. by both increasing IO order *n* and FO order α).

REFERENCES

[1] A. S. Elwakil, "Fractional-order circuits and systems: an emerging interdisciplinary research area," IEEE Circuits and Systems Magazine, vol. 10, pp. 40–50, November 2010.

[2] K. B. Oldham, and J. Spanier. *Fractional calculus: theory and applications, differentiation and integration to arbitrary order*. Academic Press, New York, 1974.

[3] N. Soubhagyaseetha, and D. V. Kamath, "Gm-C fractional Bessel filter of order (1+α)," In Proc. of the 2nd International Conference ICSSIT 2019, Tirunelveli, India, Nov. 27-29 2019, pp. 454–459.

[4] T. J. Freeborn, B. Maundy, and A. S. Elwakil, "Approximated fractional order Chebyshev lowpass filters," Mathematical Problems in Engineering, vol. 2015, Article ID 832468, 2015.

[5] T. J. Freeborn, A. S. Elwakil, and B. Maundy, "Approximated fractional-order inverse Chebyshev lowpass filters," Circuits, Systems, and Signal Processing, vol. 35, no. 6, pp. 1973–1982, 2016.

[6] T. J. Freeborn, B. Maundy, and A. S. Elwakil, "Field programmable an alogue array implementations of fractional step filters," IETDigital Library: IETCircuits, Devices & Systems, vol. 4, no. 6, pp. 514–524, 2010.

[7] L. A. Weinberg. *Network Analysis and Synthesis*. McGraw-Hill, New York, 1962.

[8] A. Tepljakov, E. Petlenkov, and J. Belikov, "FOMCON: Fractional- order modeling and control toolbox for Matlab," in Proc. of the 18th International Conference MIXDES 2011, Gliwice, Poland, June 16-18 2011, pp. 684–689.

[9] A. Oustaloup, F. Levron, B. Mathieu, and F. M. Nanot, "Frequency-band complex noninteger differentiator: characterization and synthesis," IEEE Trans Circuits Syst– I: Fundam Theory App 2000; vol. 47, no. 1, pp. 25–39.

[10] B. T. Krishna, "Studies on fractional-order differentiators and integrators: a survey," Signal Process 2011, vol. 91, no. 3, pp. 386–426.

[11] G. Tsirimokou, "A systematic procedure for deriving RC networks of fractional-order elements emulators using MATLAB," AEU-International Journal of Electronics and Communications, vol. 78, pp. 7–14, August 2017.

[12] G. Tsirimokou, C. Psychalinos, and A. Elwakil. *Design of CMOS analog integrated fractional-order circuits: applications in medicine and biology*. Springer, 2017.

DS-BE

International Conference on

DATA SCIENCE AND BIOMEDICAL ENGINEERING

Steering Committee

Chairs: Karolj Skala, Ruđer Bošković Institute, Zagreb, Croatia
Roman Trobec, Jožef Stefan Institute, Ljubljana, Slovenia
Davor Davidović, Ruđer Bošković Institute, Zagreb, Croatia

Members: Marian Bubak, ACC Cyfronet AGH University of Science and Technology, Krakow, Poland
Tiziana Ferrari, EGI Foundation, Amsterdam, Netherlands
Dieter Kranzlmüller, Ludwig-Maximilians-University, Munchen, Germany
Ludek Matyska, Masaryk University, Brno, Czech Republic
Jesús Carretero Pérez, University Carlos III, Madrid, Spain
Dana Petcu, West University of Timisoara, Romania
Uroš Stanič, Biomedical Research Institute-BRIS, Ljubljana, Slovenia
Tibor Vámos, Hungarian Academy of Science, Budapest, Hungary
Yingwei Wang, University of Prince Edward Island, Charlottetown, Canada
Matjaž Veselko, University Medical Centre Ljubljana, Slovenia

Program Committee

Members: Enis Afgan (Ruđer Bošković Institute, Zagreb, Croatia)
Viktor Avbelj (Jožef Stefan Institute, Ljubljana, Slovenia)
Davor Davidović (Ruđer Bošković Institute, Zagreb, Croatia)
Matjaž Depolli (Jožef Stefan Institute, Ljubljana, Slovenia)
Simeon Grazio (University Clinical Centar Sestre milosrdnice, Zagreb, Croatia)

Marjan Gusev (Ss. Cyril and Methodius University, Skopje, North Macedonia)
Vojko Jazbinšek (University of Ljubljana, Slovenia)
Jurij Matija Kališnik (Klinikum Nürnberg Süd, Germany)
Zalika Klemenc-Ketiš (Angela Boškin Faculty of Health Care, Jesenice, Slovenia)
Gregor Kosec (Jožef Stefan Institute, Ljubljana, Slovenia)
Miklos Kozlovszky (Obuda University, Budapest, Hungary)
Lene Krøl Andersen (Danish e-infrastructure Cooperation Denmark, Kgs. Lyngby, Denmark)
Tomislav Lipić (Ruđer Bošković Institute, Zagreb, Croatia)
Željka Mihajlović (University of Zagreb, Croatia)
Tonka Poplas Susič (Health Centre Ljubljana, Slovenia)
Aleksandra Rashkovska Koceva (Jožef Stefan Institute, Ljubljana, Slovenia)
Karolj Skala (Ruđer Bošković Institute, Zagreb, Croatia)
Viktor Švigelj (University Medical Centre Ljubljana, Slovenia)
Ivan Tomašić (Mälardalen University, Västerås, Sweden)
Roman Trobec (Jožef Stefan Institute, Ljubljana, Slovenia)
Roman Wyrzykowski (Institute of Computer and Information Science, Częstochowa, Poland)

INVITED PAPER

Information Communication Society - a Possible Ecology of the Future

Zorislav Šojat*
* Ruđer Bošković Institute, Centre for Informatics and Computing, Zagreb, Croatia
sojat@irb.hr

***Abstract* - As Henry A. Giroux recently stated, the situation caused by the present pandemic of the Novel Corona Virus (SARS CoV-2) is not only a medical crisis, but even much more a "Socio-Political and Ideological crisis", "crisis of Greed" and "Pedagogical crisis". Consequently it is now absolutely essential to develop our Collective Imagination to produce possible Visions of the Future in which a novel global system of ecological production, distribution and usage of long-life high-quality products will only by necessity support the basic economic local production, distribution and usage of human survival essentials. In this sense an Information Communication Society, heavily leaning on organisational "skills" of computer and networking applications, will have to be envisioned, developed, organised and implemented on the global scale. The Information Communication Technologies and Computer Science developments will have to heavily concentrate primarily on enabling such an Information Communication Society, a future global society of local societies based on Kairology (the science of the right moment) and Synergy (co-operation) and rooted in proper present and future developments of the knowledge of Cybernetics (the science of dynamic systems) and Ecology (the science of the whole). This article tries to give a vision of how we could approach this goal.**

Keywords – COVID-19; Information; Communication; Society; Ecology; Cybernetics; Computer Science; Dew Computing; Fog Computing; Cloud Computing; Rainbow Ecosystem

I. INTRODUCTION

As my late professor László Bulcsú used to stress often when starting a lecture, first we shall analyse and understand the Title, as in the Title most is already said. And often he used to finish his lecture with this analysis. And that is quite a proper method, as once upon a time Confucius said: “If words do not mean what they mean, nothing can be established”. It is quite essential that there is a common denotational and, as much as possible also conotational, understanding of particular terms used.

So what are the important messages we can mine from the title "Information Communication Society" and what is to be understood by "- a Possible Ecology of the Future". Here we have a lot of complex words all put together.

Let us start from the very end, with the last word of the Title: Future.

II. FUTURE

We all like to “look at”, or even “into” the Future, trying to see what is in “front” of us, where are we going, what will happen, or even how to continue...

Ancient Greeks had quite a different view on the notions of Past and Future. When our civilisation talks about the Future, we always imagine it in front of us, the Past is past, somewhere behind our backs. But they, and quite correctly, regarded the Past to be in front, and the Future behind their back. And it is true. We cannot "see" an infinitesimal instant into the Future, but we can quite well "see" all of the Past, as much as our knowledge and memory allow us. The Future we can only wish, try to predict (based on what we "see" in the Past leading to the Now), imagine or visualise. To get a really physical feeling of how we step into the Future, stand up and try to walk backwards, but without ever turning your head.

Speaking in term of physics, we can only see (feel) something that already happened. The visual time distance of anything 1m away is approx. 3.3 nS. That actually means that a person we see is in that moment already 3.3 nS in her or his own future. And in that same moment when we are seen, we are already 3.3nS older then seen[1]. This is a natural law of existence (actually in Computer Science we like to call it Latency). This has the immediate consequence that each and every system with input, on any level of complexity, is always in its own, individual, unique Now, and all inputs reflect the Past of the other system/environment. Each and every one of us is actually living a completely separate individual Now, which is an infinitesimal jot of being. Consequently, it is impossible for any system to react in “real time”, as any input disturbance of its self-organising stable state is already sent from the past of the other system/environment. Therefore there is no theoretical possibility of pre-emptive reaction, except based on the knowledge of previous similar conditions which led to a specific type of signal from the other system/environment.

The unexpected pandemic which was long time expected, the sudden emergence of the Novel Corona Virus (SARS CoV-2), which has shown to be very unpredictable, in its (a-)symptoms, viability, virulence and

[1] If we look at our Milky Way Galaxy Centre, for example, what we see now, happened there around 26 thousand years in the past from our Now, so the Centre is actually, in our present, 26 thousand years older than it is as we see it. This is quite close to the time of a full precessional cycle of our planet Earth.

lethality, attacks not only, as we would expect from a human disease, humans, but even in a much greater measure our general civilisation and its basic "working" principles. As prof. Tomislav Pletenac stated it recently, this Virus is like a robber who stands in front of you with a gun saying "Money or Life?!" [1]. And it gives you only one possible choice, whatever you try – preserve Life, forget Money. From present statistical data it is quite obvious that those countries which preferred Money lose both Life and Money, whereas those which preferred Life lose in the process much less Money. To cite Henry A. Giroux, the present Now is a "Socio-Political and Ideological crisis", "crisis of Greed" and "Pedagogical crisis" [2].

This sudden crisis did not in the first moment seem like something which would completely devastate our past's days "economic" habitus, almost completely stop the global moveability of people (and goods!), clear up the skies, giving Earth a moment to recover, and drastically change the general view on what is actually essential – the basic human needs for food and shelter; showing that local production is essential for the survival, and always newer and fancier products completely unnecessary.

There is an interesting story that happened in Croatia when the pandemic started and the boarders of all nations suddenly closed, and essential goods, like alcohol, where not exported any more. It came out that there is no local (i.e. in Croatia) production of alcohol, necessary for the production of disinfectants, as due to "market economy" reasons the production stopped several years ago, and "cheaper" alcohol was imported. But, fortunately, a sugar company which used to produce it remembered that they did not throw away the necessary machinery, they have stored it away. So they rebuilt in several days the production line, and even pulled back some people who knew how to run it from retirement. The next problem suddenly was that in Croatia there was (almost) no production of yeast. Fortunately, after searching for a while, the company found a yeast producer in Croatia, and started, in extremely short time, alcohol production again [3].

This is a striking example of the necessity to have local economy, while keeping a balance with the global ecology. We will have to organise our civilisation differently, and take into account the "time- / effort- / energy- / availability- distances". The whole global system of delivery of goods, the system of "having something" and "not having something" suddenly looks like travelling through the Galaxy. At once things started being very, very far away.

So here we are, standing with our backs towards the Future, looking at the past and trying to imagine what possible path to take into the Future. And it is very obvious that the only possible paths for our common survival have to be those which take into account and *understand* the Ecology of our individual and collective environment(s).

III. ECOLOGY

So we come to the second to last word of the Title, Ecology. Οἶκος is the ancient Greek word for 'home', 'house', 'estate', 'whole'. Therefore Ecology is the science about the balance of all elements and subsystems of a 'home', 'estate', and the 'whole', i.e. its main area of exploration is an Eco-system. Actually Ecology is per definitionem Holistic, and cybernetically recursive. That means that it recursively regards each and every of the individual ecosystems, down to the smallest (sub-)system, and up to the largest (supra-)system, actually the Universe, as being a "viable system". What is a Viable System? Let me cite the words of Stafford Beer: "a viable system is a system capable of independent existence", but "no living being is absolutely independent" [4]. Or even further, there is nothing in this Universe which could exist independently. So a viable system is a system which can survive in its environment, and which is as self-standing as necessary to survive in that environment. An ecosystem is a system of all its sub-systems and their inter-relationships and inter-behaviour, excluding, naturally, the environment of that ecosystem, but is viable only in the sense of survival in its environment[2].

If we try to look at some possible 'whole', οἶκος, it is something like we can see in Fig. 1. It consists of a huge amount of little "dots", which themselves are self-organising and viable systems. So in this global picture of an ecosystem, regarding Ecology, we have a real problem, as many of the individual viable ecosystems, on each level of recursion, are quite remote from each other. The specific interests, in this global whole, are completely different, and it is necessary to balance all those interests in such a way that they do not destroy the viability of other (sub-)systems and specifically not the viability of the whole!

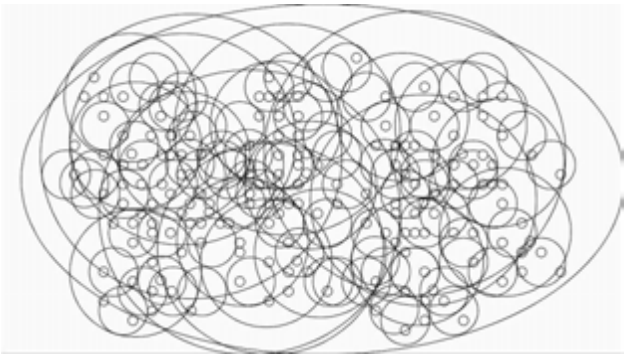


Figure 1: "Ecology": Science of a recursive viable system of viable systems

It is important to stress again that we, as a civilisation, have to think about Ecology, the science of the Whole, in a global way. General Ecology should be global, partial ecology has to be used for individual sub-systems, but only in correspondence with global Ecology, and no partial ecology may be used for global aims, because we have to look at our survival as a civilisation, and even as bare humans, on this planet Earth.

[2] For example: the Earth as an ecosystem is viable only to the point some environmental supra-system, or some sub-system of it, like we humans, changes its orbit. After that new self-organisation of our planet with the surrounding universe would bring it to a new stable state, however the Earth ecosystem would change drastically.

But what is then the Economy? Economy also comes from the οἶκος, from the 'home', 'estate', 'whole', and the other part is from όνομα, 'name'. So actually Economy is the science which tries to see what we have, where we have it, and how to organise that what we have into something which is essential for our survival, and, if survival is not endangered, for the quality of our life. In this sense Economy is primarily local.

It is obvious, looking at Fig. 1, that all of these systems are parts of higher systems, but not all of them are parts of the same higher systems, except being elements of the global supra-system. What normally any system "likes to do" is trying to find a kind of equilibrium, a state-space of relative stability. With such complex interrelationships of different sub-systems, a state of dynamic equilibrium in the supra-system can easily be strongly disturbed by introduction or realignment of its individual elements. The supra-system will certainly "find" its new dynamic equilibrium after a certain time, but that adaptation may not be "inhabitable" by many of its previous sub-systems, depending on their adaptability and severity of the necessary change of their environment. For our civilisation to survive, as it seems, it will be necessary to attain a new equilibrium on our planet. We will have to adopt ways of living which will have to be based on as much as possible local economy, and much, much more global ecology.

In [5] a recursive model of balanced functioning of a person is given, which can also be regarded as a (possible) model of a much wider ecosystem, based on four dimensions: Spiritual harmony, Mental harmony, Strategical harmony and Operational harmony. For the discussion in this Article we are specifically interested in the central point, which represents the balance of all recursive sub-systems, and is actually what Gerd Sommerhoff would call the "focal condition" [6], the point of equilibrium, dynamic stability. (Note that all the recursive sub-systems have their own focal conditions, i.e. dynamic stability points.) This point of balance is actually constantly pulled by each of the dimensional sub-systems, to move that central point of equilibrium towards some direction. Being able to homeostatically keep the focal condition (focal point) of a system will actually make that system viable.

So, cybernetically speaking, what a global ecosystem has to do is to keep all those myriads of elements which are existing around us in a viable balance. And if we intend to survive on Mother Earth, our Ecology has to be developed strong enough at least not to move the Earth's ecosystem into a focal condition which would not include us as members of this new global system dynamic equilibrium.

We are really not at all powerful over the forces of nature. It is interesting, and probably very important, to note, regarding the global ecosystem at the level of our planet, that if Earth would be as tall as we are on average (adult male and female), that is cca. 1.65 m, our size would shrink to some 129 nm (nano-meters)! The present CoViD-19 pandemic is caused by the SARS CoV-2, whose size is measured up to 140 nm [7]. So, looking at our host, the Earth, we are actually just as small, and just as nasty as the Novel Coronavirus is to us!

We can see and we can feel that the system which is presently the οἶκος of our civilisation is not really viable, which this virus has very well additionally exposed.

IV. Possible

The next previous word of the Title gets us to the notion of possibility. However, it is not that we can accept any possible future, as we do not want to destroy the environment to the point human life will lose its dignity, or even to the point that human life will cease to exist. We even do not want to lose many of our civilisation's attainments. But we will have to get rid of many of our civilisation's "attainments"! It is a matter of essential ethical values.

Some possible things are very hard to attain, like balancing on a ball. In this moment of history we are really trying to balance on a jumpy ball, not only because of the pandemic, but primarily because of the whole bunch of global ecosystem equilibrium problems we produced by completely conscienceless exploitation of all kinds of natural viable systems (what we euphemistically like to call "resources"), up to the point that we presently live in the Sixth Mass Extinction [8]. Mass extinctions do happen from time to time. The previous, fifth one, was the Cretaceous–Paleogene extinction event, around 66,000,000 (66 million) years ago. But this one, the sixth, was not caused by some natural processes, but by our own greed, our consumeristic ideology, our constant fight for "growth" and "expansion", our tragically flawed philosophy of "bigger – better – faster", deeply ingrained in the lack of ethical and ecological conscience, education and promotion on all levels.

So much about Possible. Though it is, and must be, possible to develop a new civilisation ecosystem which will enable us to survive, and heed all ethical rights of humans, animals, plants..., and the nature as a whole.

So what can we do? Let us look at the next word: Society.

V. Society

So what is a Society? (See Fig. 2.) A society is an ecological sub-system with a huge amount of different people, which have different ideas, which have different wishes, which have different hopes, different needs, wants, skills, knowledge..., who all try to build one big global system. And this is not possible without proper education and without proper organisation, which should not be a kind of "up → down" organisation, as is much the case in modern world, but a "bottom → up" autopoietic organisation [9]. It is very obvious in the nature and it is very well known in cybernetic systems that all the more complex dynamic systems organise from bottom upwards. Consequently it is extremely important to have the lowest (smallest) system well organised to be able to build up a system which is then also well organised.

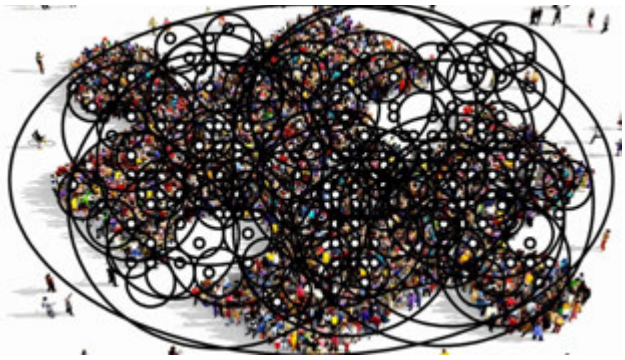


Figure 2: "Society": a recursive system of individuals trying to build a collective

In this sense, what is very important in the society, is Identity. Each of those little circles has a specific identity of a specific social environment in which the specific individua are living, in the same way as each specific individuum has its own ecosystem inside, in which all the elements which make us viable human beings live, all our cells, viruses, bacteria etc. So identity is very, very important for the notion of viability of a system. Now, what is the systemic identity of each of those (sub-)systems? To cite Stafford Beer again: The systemic identity "is a highly complex ethos of self-recognition, deriving from the interplay of many components – and in particular of the observed system with the multiplicity of its environments which, indeed, are major forces in formulating that ethos and forging that identity" [10]. Individual identities are in constant flux, adapting to the immediate and wider environment of that individual system. For us, humans on the planet Earth, it is really essential to start identifying ourselves with the "top" human identity level, the level of realisation that each of us is *just* a human, and nothing more nor less, and that each of us is an unseparable part of the nature and our planet's ecosystem. To be able to attain this urgently needed global identity, without which we can not act as humanity, it is really essential to introduce this identity into education and promotion all around the world population.

It is worth here, again, to remember the Ancient Greeks, and their principle of democratic "election" of those who will temporary be the main "political" force, based largely on sortition, that is being randomly chosen by a "lottery" principle from (electable) population. They even had a machine to help this randomisation, the Kleroterion [11]. Though in Ancient Athens only certain portions of the population were "electable", i.e. had the right (and obligation) to participate in politics, the system itself is very democratic (that's why they call it democracy), as it allows a cybernetically sane diversity (variety) of individua (identities) from all drives of life to actively participate with their knowledge and viewpoints. A random sample of population solving political issues, specifically if re-chosen for each issue to be discussed/solved, will statistically give a much wider spectrum of opinions and solutions existing in the population, as the variety of possible states of the whole society will not be so drastically lost to higher hierarchical solution/decision levels, as when using non-random systems.

As opposed to Athenian Democracy, all of the present-day existing socio-political systems[3] are, whatever their name and ideology, based on either top-down methodology, like dictatorship, monarchy, führer-ship, technocracy, bureaucracy, meritocracy etc., or on Roman republican[4] systems, where an election of "representatives" or "delegates" is in place, choosing particular (known) people into a purely hierarchical political "machine". (Additionally, many of those systems contain non-systemic combinations of different structures and procedures.) Note that both the "representatives" and the "delegates" system have the same systemic problem – individuals becoming politicians. And there is a general (purposeful) lack of *obligation* to be active part of the socio-political life.

William Ross Ashby stated a very important law, the Law of requisite variety: "If a system is to be able to deal successfully with the diversity of challenges that its environment produces, then it needs to have a repertoire of responses which is (at least) as nuanced as the problems thrown up by the environment." [12] In other words, the variety (number of states of a system) needed to properly control another system must be either same or greater [13]. Therefore, it is quite obvious that in a hierarchical system each hierarchically higher level would have to cope with exponentially growing variety of possible regulation responses. Therefore variety attenuation measures are introduced, like statistics, specific measurements tracking etc. This in turn disables the higher system, due to the lack of variety, to control (organise, lead) the lower (wider) system(s). So, for example, the extreme lack of variety in the system of labelling everything with an amount of "money", and the "monetary system" raised from this labelling, actually contort the underlying economic and ecologic system to the point of extremely high stress at the edge of viability.

VI. Communication

Now let us look at what Communication is. In Fig. 3 we have two systems with their internal "elements" represented by small circles. It is obvious that those two systems would like to communicate, but there is nothing they can communicate about, because their internal system organisation is so much different that there is no possible common "understanding".

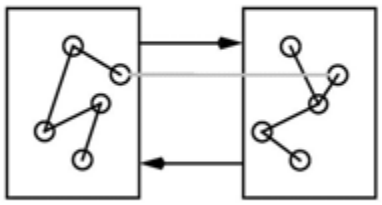


Figure 3: Two systems which can not communicate

In communication that is a major problem. To be able to communicate, there is a requisite of having some basic

[3] Excluding small(ish) self-standing self-sufficient isolated populations, where a number of differing traditional systems are in place.

[4] Lat. res publica, denotationally meaning "public thing".

common “understanding”, or, in other words, to have at least some isomorphic or nearly-isomorphic subsystems, which would be “recognised” as the receptor(s), i.e. context, of the communication transfer.

On the other hand, looking at communication, there is another problem. Let's imagine a Viable system 1 (VS1) and a Viable system 2 (VS2), and they want to communicate. The basic information theory defines that there shall be a bi-directional communication channel between them. But what is actually important here is the Latency of the communication transfer from VS1 to VS2 through the Channel. How long will it take for VS1 and VS2 to mutually exchange enough information to “rectify” the amount of non-isomorphism of their structures, as to finally be able to understand each other?

Normally you have statistics of a state (global systems and their smaller sub-systems). However, the major problem with statistics is that you get it once in a while (like once in a year). So how can you, with such a huge latency, between, for example, a government and an economic process, regulate it? You can not. Because you are very, very late. We can see the same with specific pandemic reactions around different world countries.

How we could have regulated the pandemic spread, and what we had used in Croatia, thanks to the excellent organisation of the Croatian Crisis Headquarters, in regulating the epidemic, is the fact that at one-day intervals all the data, all the information (because we understand what that data means) needed to make very quick adjustments in the response was available.

So one of the important things in the future, and in these efforts Information-Communication Technology and Computer Science can give a huge obol, is actually the shortening of the time necessary to react to a specific emergence. If you know that some of your resources are approaching a low level of availability in time, you can try to replenish them, and you have still enough time to try to do something about it. You are not any more dependent on a last minute resort, or distortion of your system. Unfortunately, in many of the economic crises we saw, we can perceive the same problem: that the data did not transform itself into information to come to some place which could react in time. As a consequence of that, you have a phase shift, a delay between the actuality and the perceived situation. A phase shift between what you know is existing and what actually is existing, at a specific moment. Now reacting on the perceived situation after a delay during which the actuality already changed, gives a wrong control impulse, and consequently generates resonance, when problems start to come in reasonably regular waves [14]. It is known and obvious that long latency regulation is extremely complicated, very much harder than short latency control, like in robotics[5]. The latency of the Novel Coronavirus is one to two weeks (the incubation period), and than it is very hard to control the happening without an a priory experiential model which will allow you to react reasonably pre-emptive.

[5] Imagine moving your hand if you would get the necessary information about its “present” position an hour later than it already was in that position. The same problem applies, for example, also to spaceship remote control.

VII. INFORMATION

And finally the last (actually the first) word of the Title: Information. So first let us look at some data: (Fig 4).

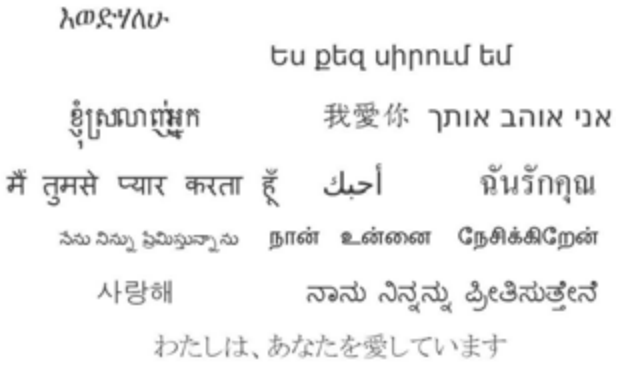


Figure 4: Some data: by themselves meaningless markings

It is important specifically to stress that most modern-day Computer Science is talking about and concentrating on *Data*. But, having data – what do they mean? What do we want to tell with them?

'Data' is the plural of 'datum', Latin word meaning 'a given', so 'data' means 'givens', a set of individual things given to us, with no internal or external relationships. To understand, for example, the data in Figure 4 it is necessary to have additional knowledge, not provided in that Figure: those markings are not random, sequences expand in different directions, the markings represent something related to human endeavour in putting thoughts into specific types and sequences of markings, the markings represent a way of writing sentences, they are very different in style, but individual sections are from the same style, the markings represent a way of writing a sentence in different languages... To fully understand the givens of Fig. 4 it is further necessary to know 13 human languages[6] and the way those markings in that specific language relate to the words and grammar of that language, i.e. we need to know how to read and understand that language. Only with all that knowledge we have proper information: Those givens (data) all represent the same human emotional thought expression.

Namely, to be able to establish communication it is essential to have some common pseudo-isomorphic substructures, where the non-isomorphism of those approximately common (similar) structures of both communicating systems is actually the information to be transferred (in both directions).

Information is indivisible from the common context, and can be defined as sub-structural non-isomorphism of two communicating systems with common pseudo-isomorphism, i.e. partial isomorphism. This general principle is very obvious in any kind of inter-human communication, as much common knowledge / ideology / frame of mind is necessary to properly understand each other.

[6] Amharic, Armenian, Khmer, Chinese, Hebrew, Hindi, Arabic, Thai, Telugu, Tamil, Korean, Kannada, Japanese.

To enable further development, the connotation (common meaning) of Information Communication Technology has to be expanded to the full denotation (explicit meaning) of that term, i.e. the Technology (science of attainment) of communicating information has to be well expanded into different non-machine based communication channels, which implies development of (novel) educational and promotive activities, based on advances in the understanding of societal and interpersonal communication principles (i.e. linguistical skills necessary) and channels, of their proper use, and of their ethical values.

VIII. Dew – Fog – Cloud

In a way we, as global and local societies, had a kind of luck that the CoViD-19 pandemic came during an era of highly developed communication structures and reasonably developed computer science. Much of the everyday life, conversation, exploration, study and work was suddenly transferred into the virtual space of modern information-communication and computing technologies.

However, to actually develop a computing and communication assisted global viable ecosystem, which we, due to our own gradual destruction of our social and natural environments, necessitate quite quickly, it is important to have a Computer Science model of communication and computing hierarchy in which particular recursive elements of the larger viable system can be implemented on the level of machines, taking into account all the necessities of natural and human well-being. In recent years such a hierarchical architecture consisting of three major conceptual paradigmatic (and architectural) levels started being developed: the Dew – Fog – Cloud computing dynamic system [15].

At the lowest level, connected directly with the immediate human and natural environment we recognise the level of Dew Computing. This is the area below the Edge of Internet, consisting of a vast number of self-organizing Dew Droplets, information gathering, processing and distributing elements which directly control our Physical Environment, from energy supply and distribution, heating, cooling, lighting to traffic flow organisation and health monitoring/assistance. The major prerequisite of these systems is that they must be self-sufficient, but can be coordinated, inside fixed limits set by the environmental process they regulate. This level, as said, is not part of the Internet, but has diverse means of coordination-communication, and is able to communicate with higher paradigmatic levels through Internet-included specific Dew Droplets. The Dew devices are not supposed to be user-programmable, but are specific fixed hardware / software entities with inbuilt self-organizing principles allowing them to "behave" in the specific environment for which they were made.

The second level up the paradigmatic computing hierarchy is the level of Fog Computing. This is actually the layer on the "Edge of Internet", the more or less able generic programmable devices primarily oriented towards Human – Computer interaction. This level is a kind of "gateway" between the stubbornness of computers and the wilfulness of humans, and between the basicity of Dew and the vastness of Clouds. Adaptive network reorganisation, local computing and services support, as well as human interaction are essential parts of the Fog. At this level it is extremely important to design proper ergonomic information exchanges between humans and machines, by using knowledge and understanding from psychology of human perception, to be able to present the vast variety of the global ecosystem in a human understandable form, with as much variety preserved as possible.

And finally, on the highest paradigmatic architectural level we have Cloud Computing, where huge amount of information (and still, unfortunately, data) is preserved, and can be used in global self-organisation, and future prediction based on past knowledge. It is at the core of Internet, caring about information, knowledge and intelligence. Of these three layers the Cloud Computing is presently architecturally the most developed, but it will have to be quite drastically adapted to enable seamless integration of information processing as a constituent part of the Dew-Fog-Cloud System.

Development of the Dew-Fog-Cloud system will enable us to solve a lot of problems in viable distribution of goods and resources, starting from local economy, by keeping the distribution necessary as local as possible, and than, over the higher Cloud level, recursively in wider areas, as the need may arise, and finally, as the last resort, up to the global planetary space; therefore keeping the economy local and the ecology global.

IX. Rainbow Ecosystem

From the philosophical and pragmatical points of view it is important to envision a reasonably simple recursive model of a global ecosystem, which would allow proper integration of the extreme diversity of the future Nature-Human-Machine cohabitation and cooperation into a human understandable and manageable viable system, without the necessity to significantly impede (attenuate) the requisite variety. Wise use of information processing equipment as an integral part of the global ecosystem gives us a possibility to surmount the ecological and sociological obstacles we have been building up so fervently.

The use of colours as symbols is ages old. Therefore, as a means of broadly defining major elements of a viable system, the whole is divided into 9 symbolic colours, and each of them is assigned a generic meaning. The Philosophy of Computing view of such a system is given in [16, 17].

Here, the Rainbow Ecosystem is presented in short as a final integratory factor of the elements discussed in this Article, from a viewpoint of general philosophy and pragma regarding the development of a viable nature-human-machine ecosystem. Each of these colours is given a basic meaning and a very broad area of elements to be covered, which are, per se, thoroughly intertwined with all others, both horizontally and recursively.

It should be noted that Green, the Nature, is in the middle, flanked by the "production" side, from bottom (Infrared) up, by Energy and Matter, transformed through

Creativity and, extremely important, the Appropriateness Filter, into Nature, and flanked on the other side by the self-"organisation" path from top (Ultraviolet) down by Visions, Interference, Cooperation and Communication, balancing the whole, and making it a recursive viable system.

The following pseudo-paragraphs only name some of the various notions symbolised by the colours, without entering into their interrelatedness and particular areas:

A. Ultraviolet – Visions

Wisdom, Prudence, Conscience, Responsibility, Holism

B. Violet – Interference

Security, Limits of Expansion, Nature-Human-Machine Interference

C. Indigo – Cooperation

Ethics, Information Use, Redundancy, Knowledge Gathering and Preservation

D. Blue – Communication

Information, Knowledge, Interaction, Languages

E. Green – Nature

Environment, Health, Well-Being, Backup Systems, Global Ecosystem

F. Yellow – Appropriateness Filter

Consequences

G. Orange – Creativity

Stimulation, Ideas, Education

H. Red – Matter

Resources, Products

I. Infrared – Energy

Efficiency, Garbage, Quality

X. CONCLUSION

So, what should be the first steps we ought to take? What is immediately obvious, specifically after the onset of the CoViD-19 pandemic, is that we have to radically (and drastically) change our global civilisation aims and methods. The global and local response on this pandemic has shown a completely uncoordinated and much ideologically influenced, often panicky, reaction, which succeeded to completely destabilise our social and economic habitus. Or, better to say, this pandemic has hard hit an already long time unstable socio-politico-economic (and cybernetically obviously unviable) system. But what it also succeeded to do is to blatantly uncover a huge amount of "rotten" points (sub-systems and their interactions) of the pre-pandemic global and local systems.

In many such historical situations of great destabilisations revolutions would arise, and impose a new ideological system. But as revolutions, in addition to "eating their own children", always like to disregard the knowledge of the past, they blindly stumble backwards into the future. And it is true that already in certain areas of the world we perceive the first signs of revolutions (and the counter-revolutions which are their lasting companion).

Therefore we need to act quickly but prudently, because we need Evolution, and not revolutions. But evolution is a slow viable process, and it is hard to act quickly. However, we already have developed a huge amount of tools which can help us, and we already have the Internet, and we already know much about how to use its services in a vast field of our endeavours.

So the first steps we have to take is to think out a viable cybernetic system which would gradually incorporate many of the existing computer based solutions, primarily in the localisation of economy and drastic shortening of the supply chains, and the "reshuffling" of goods and resources (e.g. food, water, energy...) around the whole planet, to get to a reasonably equilibrical state of the global social aspects of basic human rights (like the right to survive). A strong incentive (primarily educational and promotional) has to be made to enable the world population to understand the need for, and principles of, self-sustainable local economies providing for essential human necessities, and the need for all of us to internalise the global "Earthling" identity[7], together with basic principles of global ecology. The information-communication network has to be available to (but not forced onto) each and every citizen of Earth, and we already produced so many high-tech and quality-of-life enhancing "gadgets" that we all have (can have) everything.

The next step shall be a from of active *democratic* participation of the global Earth population in all matters of common interests recursively up, from the smallest ecosystems of home, neighbourhood, village, to the level of planetary decisions. The present day information-communication and computing technologies open up unprecedented cooperation opportunities, but if we do not actually start to cooperate, our future could be a fall backwards from a cliff.

And finally, we will never succeed in the evolution if we do not think about long term development. And a long term development can be obtained only by proper and all-reaching Education, primarily the education on Ethics, Viability, Ecology, Synergy and Kairology.

[7] Internalising the "Earthling" identity does in no way de-internalise the existing necessary variety of local individual and social identities (cultures, habits, languages, traditions, methods, faiths, nations...). Ethics can be defined as "a set of concepts and principles that guide us in determining what behavior helps or harms sentient creatures" [18]. Therefore, if some of those mentioned "local" identities lack a certain element of Ethics, to sustain global Ecology it will be necessary to try to, again through education and promotion, viably adapt those indentity-related views and standards of conscientious behaviour.

Note: A supplemental video of the on-line ICT-20 @ COVID-19 conference presentation of this Article can be found at [19].

Acknowledgment

The Author wishes to give special thanks, for invaluable scientific input, support and ideas, to Gordana Gredičak Šojat and Karolj Skala, the Croatian MIPRO Society for organising the ICT-20 @ COVID-19 on-line Conference, and also to the myriads of scientists, artists and thinkers who enabled the existence of our present day knowledge.

References

[1] Tomislav Pletenac. “We have to work also on societal immunity” (in Croatian), Interview, https://www.glas-slavonije.hr/430850/11/Moramo-raditi-i-na-drustvenom-imunitetu [8/5/2020].

[2] Henry A. Giroux, “Is impossible to be radical without being optimistic”, Interview, https://www.youtube.com/watch?v=EpRjHfDXMmY [8/5/2020].

[3] Maja Brnić, "U Sladorani sutra kreće proizvodnja alkohola, našli radnike i kvasac", Poslovni dnevnik, 24/3/2020, https://www.poslovni.hr/hrvatska/u-sladorani-sutra-krece-proizvodnja-alkohola-nasli-radnike-i-kvasac-4221244 [15/6/2020, 1:20].

[4] Stafford Beer, “The Intelligent Organization“ https://www.youtube.com/watch?v=7COX-b3HK50 [14/6/2020, 17:19].

[5] Zorislav Šojat, Borna Bebek, "Harmony: Spiritual, Mental, Strategical, Operational", https://www.researchgate.net/publication/328466718_Harmony_Spiritual_Mental_Strategical_Operational [15/6/2020, 2:22]

[6] Gerd Sommerhoff, “Analytical Biology”, Oxford University Press, 1950.

[7] Zsuzsanna Varga, Andreas J Flammer, Peter Steiger, Martina Haberecker, Rea Andermatt, Annelies Zinkernagel, et al., "Electron microscopy of SARS-CoV-2: a challenging task – Authors' reply", The Lancet, 19/5/2020, https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31185-5/fulltext [15/6/2020, 3:17].

[8] Gerardo Ceballos, Paul R. Ehrlich, and Rodolfo Dirzo, "Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines", Proceedings of the National Academy of Sciences of the USA, July 25, 2017 114 (30) E6089-E6096, https://www.pnas.org/content/114/30/E6089 [15/6/2020, 4:19].

[9] Ante Lauc, Josip Miličević, Ivan Šešo, Zorislav Šojat, "Global autopoietic university as fides et ratio”, http://grgur.irb.hr/Library/PDF/Lauc.Milichevicy.Shesho.Shoyat.Global-Autopoietic-University-as-Fides-et-Ratio.pdf [15/6/2020, 4:42]

[10] Stafford Beer, “World in torment – A time whose idea must come”, Kybernetes 22, no. 6 (1993): 15–43, https://ia601606.us.archive.org/11/items/World_in_torment-Stafford_Beer/World_in_torment-Stafford_Beer.pdf [24/5/2020, 21:13]

[11] Allessandro Orlandini,”KLEROTERION, simulation of the allotment of dikastai”, Academia.edu, https://www.academia.edu/36510282/KLEROTERION._simulation_of_the_allotment_of_dikastai [14/6/2020, 18:36].

[12] John Naughton, "Ashby's Law of Requisite Variety", Edge, 2017, https://www.edge.org/response-detail/27150 [15/6/2020, 7:04]

[13] W. Ross Ashby, "An Introduction to Cybernetics", Chapman & Hall, London, 1956, Internet (1999), http://pcp.vub.ac.be/books/IntroCyb.pdf [15/6/2020, 7:01].

[14] Stafford Beer, “Fanfare for effective freedom – Cybernetic Praxis in Government”, Third Richard Goodman Memorial Lecture, Brighton 14/2/1973, https://www.scribd.com/document/2310055/Fanfare-for-Effective-Freedom-Cybernetic-Praxis-in-Government-Beer [24/5/2020, 21:08].

[15] Zorislav Šojat, Karolj Skala, "The dawn of Dew: Dew computing for advanced living environment", MIPRO 2017, https://www.researchgate.net/publication/318690476_The_dawn_of_Dew_Dew_Computing_for_advanced_living_environment [15/6/2020, 8:58].

[16] Zorislav Šojat, Karolj Skala, "The Rainbow: Integrating computing into the global ecosystem”, MIPRO 2019, https://www.researchgate.net/publication/333378758_The_Rainbow_Integrating_Computing_into_the_Global_Ecosystem [15/6/2020, 9:00].

[17] Zorislav Šojat, "From Dew Over Cloud Towards the Rainbow - Ecosystem of the Future: Nature — Human — Machine", Intelligence in Big Data Technologies - Beyond the Hype, Chapter 1, Advs in Intelligent Syst., Computing, Vol. 1167, Springer Nature, 2020, Singapore, ISBN 978-981-15-5284-7.

[18] Richard Paul, Linda Elder, "The miniature guide to understanding the foundations of ethical reasoning, Foundation for Critical Thinking Free Press, 2006, USA, ISBN 978-0-944583-17-3.

[19] Zorislav Shoyat, “Information Communication Society – a possible ecology of the future”, ICT-20 @ COVID-19 on-line conference video recording, May 2020, http://gridvision.irb.hr/mipro/Lectures/Zorislav_Shoyat-Information_Communication_Society.html [14/6/2020, 21:52]

PAPERS

DATA SCIENCE

Comparison of Machine Learning Methods in Classification of Affective Disorders

I. Kinder*, K. Friganovic*, J. Vukojevic**, D. Mulc**, T. Slukan**, D. Vidovic**, P. Brecic** and M. Cifrek*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
** University Psychiatric Hospital Vrapče, University of Zagreb, Zagreb, Croatia
ivana.kinder@fer.hr; kresimir.friganovic@fer.hr

***Abstract* - Depression belongs to a group of psychiatric disorders called affective disorders. In medical practice, patients are diagnosed according to the criteria in standardized diagnostic manuals. The criteria for diagnosing such disorders focus on the symptoms presented by the patient as well as on disqualifying other potential causes of the symptoms. Electroencephalography (EEG) is a non-invasive brain imaging technique that measures the electrical activity of the brain across different sites on the surface of the scalp. In this paper, 15 EEGs of depression patients and 15 EEGs of healthy control subjects are observed. The depressed and healthy subjects are paired according to age and gender to achieve a dataset that is balanced across classes, gender, and age of subjects. 475 different features are extracted from each EEG and used in the evaluation of different binary classification methods. The best F1-score of 0.7586 is achieved with the K-nearest neighbor algorithm. Sequential feature selection is performed, and sequentially selected features are used to evaluate the former binary classification methods. The best F1-score of 0.8750 is achieved with the K-nearest neighbor algorithm. Classification results are compared across different methods, as well as before and after excluding features that were not deemed significant by the sequential selection algorithm.**

Keywords* - *electroencephalography; affective disorders; depression; feature selection; binary classification

I. INTRODUCTION

Affective disorders are psychiatric disorders characterized by problems in mood regulation [1]. They are diagnosed and treated by psychiatrists. The diagnostic process relies on interviewing the patient in order to determine the symptoms which are present, as well as the intensity of those symptoms. A diagnosis is then determined according to a standardized diagnostic manual, e.g., ICD-10 (International Classification of Diseases – Tenth Edition) or DSM-5 (Diagnostic and Statistical Manual of Mental Disorders – Fifth Edition).

Electroencephalography is a method of measuring brain activity. The prefix of the term regards the nature of the activity, which is electrical, as a result of the functioning of the cells of which the central nervous system is comprised. In the process of diagnosing affective disorders, visual inspection of electroencephalograms (EEGs) is done to rule out brain damage or epileptogenic activity. However, EEGs contain more information, which is not accessible by visual inspection of the signal in the time domain but can be extracted with methods of signal processing and analysis.

A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention [2]. It has been hypothesized that the characteristics of EEGs could serve as biomarkers of affective disorders such as major depressive disorder (MDD). The identification of biomarkers for psychiatric disorders then serves as a step toward identifying underlying mechanisms of dysfunction, as an alternative to the current descriptive, diagnostic systems [3].

For example, parameters of EEG recorded during sleep have been found to be of discriminative and predictive value in MDD [4]. Another example are the parameters of the loudness dependency auditory evoked potential as predictors of antidepressant treatment outcome [5]. The absolute and relative power of alpha rhythms have been observed to be higher in MDD patients compared to healthy controls [6], whilst comparatively slower alpha rhythms have been linked to the ineffectiveness of treatment with certain antidepressants [7].

The goal of this paper is to process and analyze resting-state EEGs of individuals with depressive disorders as well as healthy control subjects. EEG recordings are first preprocessed to remove artifacts. Then, feature extraction is performed in order to prepare the input for machine learning algorithms, which classify EEGs into one of two classes: *depression* or *healthy*. Different binary classification methods are compared according to accuracy and F1 scores. Sequential feature selection is used to identify the most significant features, and cross-validation of different models is repeated using said features.

II. METHODS

A. Data acquisition and description

EEG dataset is obtained from the University Psychiatric Hospital Vrapče, Zagreb. Subjects are patients diagnosed with various affective disorders and healthy controls. The EEGs of 149 affective disorder patients (107 female, 42 male, age 53, +/- 12) were procured from the hospital archive. Patient EEGs are recorded as a part of the standardized admission procedure. EEGs of 15 healthy volunteers (8 female, 7 male, age 31 +/- 7) were recorded using the same protocol. The study was carried out in accordance with the guidelines of the Declaration of Helsinki and approved by the ethics committee of the University of Zagreb, University Psychiatric Hospital Vrapče.

EEGs were recorded using a 19-channel EEG amplifier, with standard 10-20 electrode placement, as shown on Fig. 1. Non-reference EEG electrodes are Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1 and O2, with Oz as the reference electrode. The sampling frequency is 256 Hz.

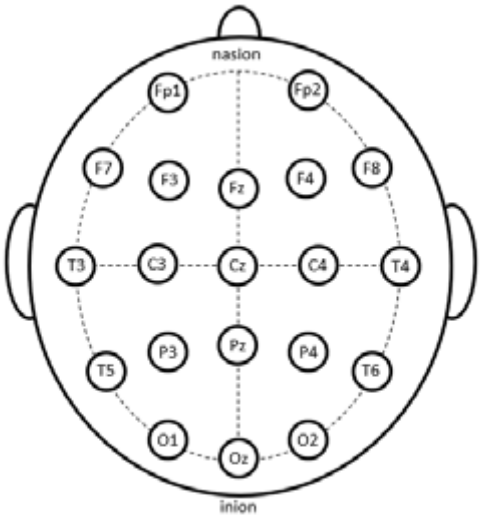


Figure 1. 10 - 20 electrode montage with Oz as the reference electrode

The recording protocol consists of three main parts, as visualized in Fig. 2: the first consisting of resting-state EEG with eyes opened followed by eyes closed, the second being EEG during photo-stimulation, and the third being EEG during and after induced hyperventilation. The subject is in a lying down position, with their body positioned to alleviate muscle tension. The room is kept quiet and peaceful. The protocol starts with the subject being asked to keep their eyes open. After 8 to 10 seconds, the subject is asked to keep their eyes closed until told otherwise. After 45 seconds to a minute, the subject is asked to open their eyes. This is repeated 5 to 7 times. Next, photo-stimulation is performed with 5 different flash frequencies (4 Hz, 8 Hz, 16 Hz, 24 Hz, and 30 Hz), lasting 15 seconds each. Lastly, state hyperventilation is induced; 1 to 2 eyes-opened eyes-closed segments are recorded, after which the subject is instructed to breathe steadily for the rest of the recording. During the recording, the technician marks the onset of each recording event: instructions given to subject as well as non-protocol events that cause EEG artifacts, such as coughing, movement, or speaking.

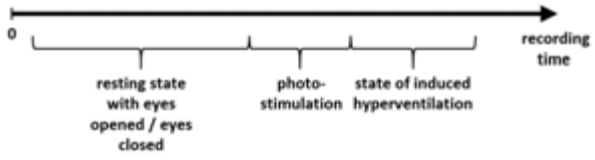


Figure 2. EEG recording protocol

Diagnosing of subjects is carried out by appointed physicians according to the International Classification of Diseases, 10th Edition (ICD-10). Due to the much greater number of subjects diagnosed with affective disorders than that of healthy controls, 15 subjects with severe depression are selected for the depression class, to be compared to the 15 in the healthy control class. This is done to achieve a balanced dataset for the classification problem at hand. Depression subjects are chosen to individually match each healthy subject by gender, as well as to be of similar age. This results in having the gender-balanced classes, as well as an approximately equal age distribution across classes, as shown in Tab. 1.

TABLE 1. AGE AND GENDER OF SUBJECTS IN DEPRESSION AND HEALTHY CLASSES

Class	Female subjects	Male subjects	Mean of age	Standard deviation of age
Depression	7	8	40	10
Healthy	7	8	39	10

Of the 15 subjects in the depression class, 6 were diagnosed with F32.2, 1 with F32.3, 7 with F33.2 and 7 with F33.3, at the time of the recording of their EEGs, as shown in Tab. 2.

TABLE 2. DIAGNOSIS OF SUBJECTS IN DEPRESSION CLASS

ICD-10 diagnosis	Description	Number of subjects
F32.2	Major depressive disorder, single episode, severe without psychotic features	6
F32.3	Major depressive disorder, single episode, severe with psychotic features	1
F33.2	Major depressive disorder, recurrent severe without psychotic features	7
F33.3	Major depressive disorder, recurrent severe with psychotic features	1

B. Preprocessing

Raw EEG data is contaminated with artifacts such as line noise at 50 Hz and its harmonics as well as its aliased higher harmonics, muscle artifacts, blink artifacts, and artifacts from the shifting of the EEG electrodes across the surface of the scalp.

Therefore first, the raw EEG data is filtered with a passband filter from 0.1 to 40 Hz. The lower frequency edge is chosen to eliminate the slow drifts in the signal [8], while the high-frequency edge is chosen to retain the frequency components of interest, whilst removing the 50 Hz line noise.

After passband filtering, channels are re-referenced to the average of all 19 channels.

Finally, independent component analysis (ICA) is performed using the *runica()* function implemented in

eeglab, a framework within *MATLAB*. Components are then labeled using *ICLabel*, a pre-trained independent component classifier. Each component is classified according to its proposed signal source, the sources being either brain activity, muscle activity, cardiac activity, blinking, or other, along with the certainty of said classification expressed as a percentage of certainty. The output of *ICLabel* for one of the EEGs is shown in Fig 3.

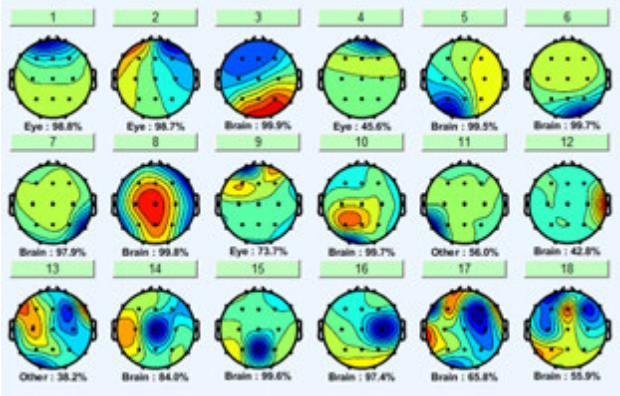


Figure 3. Example of ICA component labeling

Labeled independent components are then manually inspected, and their subsets are then subsequently removed to find a minimal subset that removes the prominent artifacts while retaining the components that make up the useful part of the signal.

C. Feature extraction

There are five main characteristic brain rhythms that are recognizable in EEGs. They can be visually recognized by their characteristic amplitudes and frequencies. They occur and dissipate depending on the state of the subject. These states include eyes opened, eyes closed, concentration, relaxation, and sleep, as well as frustration, intellectual strain, and stress [8].

The main characteristic of each brain rhythm is the frequency band it belongs to. From the lowest to the highest frequencies there is the: delta (δ) band, theta (θ) band, alpha (α) band, beta (β) band, and gamma (γ) band, as shown in Tab. 3.

TABLE 3. CHARACTERISTIC BRAIN RHYTHMS

Rhythm symbol	Rhythm name	Frequency band [Hz]
δ	delta	0.5 – 4
θ	theta	4 – 8
α	alpha	8 – 13
β	beta	13 – 30
γ	gamma	30 - 45

For each of characteristic EEG waves (theta, delta, alpha, beta, and gamma) 5 features were selected and extracted from each of the 19 available EEG channels.

Absolute band power for each characteristic brain rhythm is calculated from power spectral density approximated by expression (1):

$$PSD = \frac{1}{N}|FFT(xcorr(x(n)))| \quad (1)$$

Where $xcorr(x(n))$ is the autocorrelation function of the discrete-time EEG signal, $FFT()$ the Fast Fourier Transform, and N the number of data samples. Absolute band power is then equal to the sum of the PSD values within each frequency band, as shown in expression (2):

$$absPow_{rythm} = \sum_{f \in rhythm\ freq.band} PSD(f) \quad (2)$$

Relative band power for each frequency band of interest is calculated, dividing the absolute power of each frequency band by the total absolute power of the EEG signal, as shown in expression (3):

$$relPow_{rythm} = \frac{absPow_{rythm}}{absPow_{signal}} \quad (3)$$

Spectral centroids describe the shape of EEG spectra in defined frequency bands. Spectral centroids for each of the frequency bands corresponding to their respective characteristic brain rhythm are calculated using expression (4):

$$SC_{rhythm} = \frac{\sum_{k=b_1}^{b_2} f_k s_k}{\sum_{k=b_1}^{b_2} s_k} \quad (4)$$

where f_k is the frequency in Hz corresponding to frequency bin k of the discrete Fourier transform (DFT) of a discrete signal $x(n)$, s_k the spectral value of bin k and b_1 and b_2 the bins corresponding to the band edges of the characteristic brain rhythms.

The final two features are calculated based on the detail coefficients of signal components obtained by decomposing the signals using the one-dimensional wavelet transform implemented in *MATLAB* function *wavedec()*, resulting in details and coefficients corresponding to frequency bands which approximate that of the characteristic brain rhythms, as shown in Tab. 4.

TABLE 4. WAVELET DECOMPOSITION

Coefficients	Frequency band [Hz]
d1	6 - 128
d2	32 – 64
d2	16 – 32
d4	8 – 16
d5	4 – 8
d6	2 – 4
d7	1 – 2
d8	0.5 – 1
a8	0 – 0.5
a8	0 – 0.5

Wavelet energy is defined by expression (5):

$$E_n = \sum_j |d_{n,j}|^2 \ ; \ n = 1, \ldots, N \quad (5)$$

where $d_{n,j}$ is the ith detail coefficient at level n of the discrete wavelet decomposition of the signal.

Total wavelet energy is the sum of the wavelet energy across all levels of decomposition and is calculated using expression (6):

$$E_{total} = \sum_{n=1}^{N} E_n \;\; ; \; n = 1, \dots, N \qquad (6)$$

where E_n is the wavelet energy at decomposition level n [9].

Relative wavelet energy (RWE) is calculated using (7):

$$\rho_n = \frac{E_n}{E_{total}} \; ; n = 1, \dots, N \qquad (7)$$

where E_n is the wavelet energy at decomposition level n and E_{total} the total wavelet energy across all decomposition levels [9].

Wavelet entropy is calculated using expression (8):

$$WE_n = -\sum_{n=1}^{N} \rho_n \, ln(\rho_n) \qquad (8)$$

where ρ_n is the RWE at decomposition level n [9].

Features are extracted only for the signal samples of the first part of the recording protocol, the resting state with eyes opened and eyes closed. For 19 channels, 5 characteristic brain rhythms, and 5 different feature types, the total number of features per subject amounts to a total of 475 features.

D. Classification methods and feature selection

For evaluating different classification methods, 5-fold cross-validation is done using different classification methods. Cross-validation is performed using all 475 features to rank the various methods according to accuracy and F1-score.

Sequential feature selection was performed using *MATLAB* function *sequetialfs()* to identify the most significant subset of features for separating the two classes. Cross-validation is then done again, and classification methods are ranked by accuracy and F1-score.

III. RESULTS

5-fold cross-validation results for binary classification of classes *depression* and *healthy* are shown in Tab. 5. All 475 available features are used. The support vector machine model shows the highest classification accuracy, followed by the K-nearest neighbor algorithm.

Sequential feature selection results in features *Fp1_RelPow_Beta* (relative power of beta rhythm on channel location Fp1), *Fp2_SpecCent_Alpha* (spectral centroid of alpha rhythm on channel location Fp2), *Cz_SpecCent_Delta* (spectral centroid of delta rhythm on channel location Cz) *and Pz_SpecCent_Alpha* (spectral centroid of alpha rhythm on channel location Pz) chosen as most significant.

TABLE 5. RESULTS FOR ALL 475 FEATURES

Method	Acc.	F1 score
SVM	**0.7667**	**0.7586**
KNN	**0.7667**	**0.7586**
Naïve Bayes	0.7000	0.7097
Linear Discriminant	0.7000	0.6400
Tree	0.5333	0.5625
Logistic Regression	0.5333	0.5333

Fig. 4 shows the scatter plot of points defined by feature *Cz_SpecCent_Delta* on the x-axis and *Pz_SpecCent_Alpha* on the y-axis.

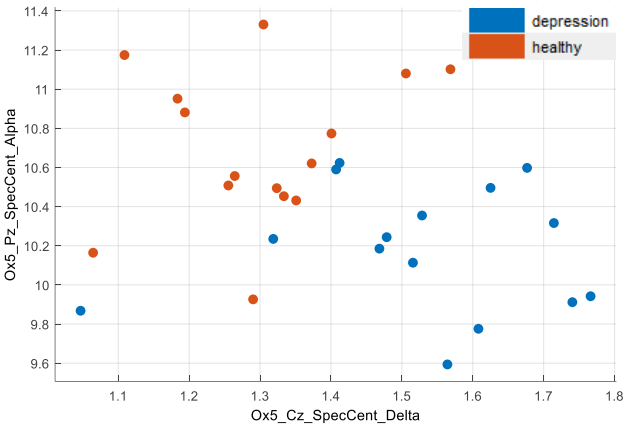


Figure 4. Scatter plot of features *Cz_SpecCent_Delta* and *Pz_SpecCent_Alpha*

Fig. 5 shows the scatter plot of points defined by feature *F1_RelPow_Beta* on the x-axis and *Fp2_SpecCent_Alpha* on the y-axis.

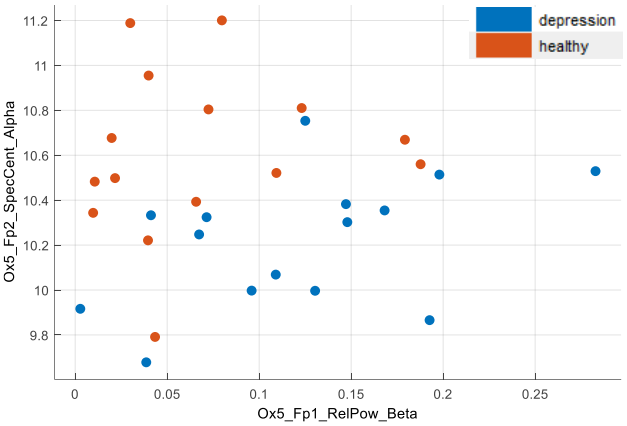


Figure 5. Scatter plot of features *F1_RelPow_Beta* and *Fp2_SpecCent_Alpha*

5-fold cross-validation results for binary classification of classes *depression* and *healthy* are shown in Tab. 6. Of the 475 available features, only 4 are used:

- *Fp1_RelPow_Beta*
- *Fp2_SpecCent_Alpha*
- *Cz_SpecCent_Delta*
- *Pz_SpecCent_Alpha*

TABLE 6. RESULTS FOR ALL SEQUENTIALLY SELECTED FEATURES

Method	Acc.	F1 score
KNN	**0.8667**	**0.8750**
SVM	**0.8667**	0.8571
Naïve Bayes	**0.8667**	0.8571
Tree	0.8333	0.8276
Linear Discriminant	0.8000	0.7857
Logistic Regression	0.7667	0.7407

The K-Nearest Neighbor (KNN) algorithm shows the highest classification accuracy and F1 score, followed by the Support Vector Machine (SVM) and the Naive Bayes.

IV. CONCLUSION AND FUTURE WORK

EEGs of 15 depression patients and 15 healthy controls are filtered, re-referenced to average reference, decomposed using independent component analysis, and then cleaned of blink, muscle, and cardiac artifacts by removing artifact components. Resting-state EEG segments are extracted from the individual EEG recordings, and feature extraction is performed on said segments.

For 5-fold cross-validation for binary classification of complete feature sets (475) labeled with labels *depression* or *healthy* SVM and KNN show the highest accuracy and F1 scores amongst the evaluated methods.

Sequential feature selection identifies the relative power of beta rhythm on Fp1, the spectral centroid of the alpha rhythm on Fp2, the spectral centroid of the delta rhythm on Cz, and the spectral centroid of the alpha rhythm on channel location Pz as the most significant features.

5-fold cross-validation with sequentially selected features shows higher accuracy and F1 scores across all evaluated methods, with the Linear Discriminant Analysis and Logistic Regression showing the biggest rise in accuracy and F1 score.

To determine whether these findings can be generalized to fit EEGs outside the observed dataset, trained models should be tested with feature – label pairs belonging to new EEGs. Moreover, the models should be trained with a bigger dataset to see whether an improvement in evaluation metrics can be achieved. Feature selection should be repeated on a larger dataset. The inclusion of other features should be examined as well as the use of other classification methods.

REFERENCES

[1] N. Sartorius *et al.*, "The ICD-10 Classification of Mental and Behavioural Disorders," *IACAPAP E-textb. child Adolesc. Ment. Heal.*, vol. 55, no. 1993, pp. 135–139, 2013.

[2] V. G. De Gruttola *et al.*, "Considerations in the evaluation of surrogate endpoints in clinical trials: Summary of a National Institutes of Health Workshop," *Control. Clin. Trials*, vol. 22, no. 5, pp. 485–502, 2001.

[3] T. Insel *et al.*, "Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders," *Am. J. Psychiatry Online*, no. July, pp. 748–751, 2010.

[4] S. Olbrich, R. Van Dinteren, and M. Arns, "Personalized Medicine: Review and Perspectives of Promising Baseline EEG Biomarkers in Major Depressive Disorder and Attention Deficit Hyperactivity Disorder," *Neuropsychobiology*, vol. 72, no. 3–4, pp. 229–240, 2016.

[5] D. Spronk, M. Arns, K. J. Barnett, N. J. Cooper, and E. Gordon, "An investigation of EEG, genetic and cognitive markers of treatment response to antidepressant medication in patients with major depressive disorder: A pilot study," *J. Affect. Disord.*, vol. 128, no. 1–2, pp. 41–48, 2011.

[6] N. Jaworska, P. Blier, W. Fusee, and V. Knott, "Alpha power, alpha asymmetry and anterior cingulate cortex activity in depressed males and females," *J. Psychiatr. Res.*, vol. 46, no. 11, pp. 1483–1491, 2012.

[7] G. Ulrich, E. Renfordt, G. Zeller, and K. Frick, "Interrelation between changes in the EEG and psychopathology under pharmacotherapy for endogenous depression. A contribution to the predictor question," *Pharmacopsychiatry*, vol. 17, no. 6, pp. 178–183, 1984.

[8] S. Sanei and J. A. Chambers, *EEG Signal Processing*. 2013.

[9] S. D. Puthankattil and P. K. Joseph, "Analysis of EEG Signals Using Wavelet Entropy and Approximate Entropy : A Case Study on Depression Patients," *Int. J. Medical, Heal. Biomed. Bioeng. Pharm. Eng.*, vol. 8, no. 7, pp. 420–424, 2014.

Evaluating Microsoft Face API in the context of student classroom attendance

M. Marjanović*, T. Kramberger *, R. Kramberger* and I. Cesar*
* Zagreb University of Applied Sciences/Department of computer science, Zagreb, Croatia
icesar@tvz.hr

***Abstract* - Face recognition technology is having a tremendous impact on the vast majority of today's human activities – mostly for leisure and security purposes. In this paper, we evaluate Microsoft Face API technology in the context of estimating classroom student attendance. A short overview of face recognition technology is presented, alongside common algorithms used in today's state-of-the-art face recognition systems. Finally, Microsoft Face API is evaluated with respect to image quality, blur, occlusion, and in-the-wild with an attempt to evaluate the efficiency of the recognition system in the context of student classroom attendance.**

Keywords* - *Microsft Face API, face detection, face recognition

I. INTRODUCTION

Vision, intelligence, and learning capabilities enable humans to identify distinct elements from their surroundings. Humans can easily distinguish between different objects or recognize faces from different persons. Computer scientists are intensively working to apply similar concepts to machines and computer vision. An extremely active area of research is recognizing and identifying persons based on their facial characteristics. Systems like these find their application in areas of security, marketing, social networks, and many other areas. Although current face recognition systems proved to be successful, experts are still working on the development of more accurate and more performant algorithms, aiming to increase detection accuracy, while reducing inference speed.

II. RELATED WORK

Research on face recognition algorithms begins in 1960. with Woody Bledsoe, Helen Chan Wolf, and Charles Bisson [1] with the semi-automated process of determining face identity based on distances and relations between eyes, ears, nose, and mouth. Due to limited computing power, this just marked the beginning of extensive research in years to come. Goldstein, Harmon, and Lesk increased the system accuracy by including hair color and lips dimensions, which summed up to a total of 21 face markers, which were still inserted manually [2]. As computing power increased, 1988. linear algebra was introduced in the process, known as the Eigenface approach [3], and extended in 1991. by Matthew Turk and Alex Pentland who developed the algorithm of detecting a human face from image automatically and pioneered first cases of automatic face recognition [4]. Face Recognition Technology, known as FERET, was one of the key projects which impacted the area of face recognition research in 1993. aiming to improve face recognition technology by creating a shared face dataset [5]. Initially, 5.000 images were collected, and in later stages ended up with 14.126 images of 1.199 different individuals [6]. One of the milestones in face recognition technology was the Face Recognition Vendor Test (FRVT) [7], whose primary goal was measuring face recognition algorithms efficiency on data previously hidden from researchers and programmers. The Face Recognition Grand Challenge (FRGC) yielded some algorithms which were able to surpass humans in face recognition [7]. To utilize face recognition, face detection algorithm needs to be as accurate as possible. Most algorithms focus on locating eyes in the image [7], after which other face parts such as nose, mouth, and eyebrows are searched for. Recognition systems try to locate the most similar face from the database of known faces within the system [7].

Face recognition algorithms can be split into those which recognize a face based on 2D representation (image) and those which use 3D representation. Principal Component Analysis (PCA) is one of the successful algorithms used in 2D face recognition, a statistical method which reduces the input dimension by applying the orthogonal transformation to input entities in an attempt to produce linearly uncorrelated variables referred to as principal components, hence simplifying the search space with a small trade-off in overall accuracy. Therefore, PCA is one of the popular dimensionality-reduction and face feature selection technologies [11] which, however, suffers from sensibility regarding dataset size [8]. In practice, due to its limitations, it is not often used as sole face recognition technology, but rather represents the first step in more complex algorithms. Support Vector Machines, a binary linear classifier, proved to be successful face recognition technology [8, 10, 12]. SVM models attempt to increase the margin between two classification groups. Independent Component Analysis (ICA) is one of the techniques of finding principal elements from multi-dimensional data [10, 13]. Algorithms based on the Gabor wavelet technique proved to be effective in face recognition [8, 14]. Linear discriminant analysis (LDA) is also a method of dimensionality-reduction similar to PCA, but instead of finding the axis with maximum variance on the entire dataset, it attempts to find axes which ensure maximum class separability [8, 15, 16]. Lately, deep learning plays

an important role in modern face recognition technology, primarily using convolutional neural networks for feature extraction. Recent work was focused mainly on discovering a preferable loss function for maximum accuracy [22, 23, 24].

III. THE EXPERIMENT

We used Microsoft Azure Face API as one of the commercially available face recognition systems and evaluated its performance against several scenarios of both artificially altered images and images taken in a classroom with 10-15 students at regular intervals throughout approximately 2 months. Unfortunately, the Azure Face API does not specify the exact algorithms used for its face recognition capabilities, so we were unable to verify its expected performance on some of the standard face recognition tests.

The system is based on supplying one or more faces into the database of „known faces“, which can later be used for identification. In the first experiment, a group of 5 distinct famous person photos is created, with one image for each person. The system was tested with 5 different images of the same person alone in the photo, as shown in **Figure 1**.



Figure 1. Testing with five images of Brad Pitt, with only one image initially in the group

Next, sensitivity to image size is tested. The system successfully recognized the actor in the image for sizes 1200x1200, 600x600 and 100x100 px (Figure 2). It is required that face size in the image is at least 36px on each dimension [25], otherwise, the face will not be detected by the system, and therefore recognition will not even take place. The system proved robust when evaluating with images with multiple different persons in the same image, but also several occurrences of the same person in one photo (Figure 3).



Figure 2. Testing sensitivity to image size, from left to right: 1200x1200px, 600x600px, 100x100px.

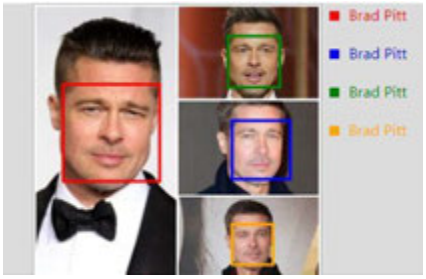


Figure 3. Multiple occurrences of the same person in a photo

System robustness with regard to age, fashion items, face expressions, and facial hair is presented in Series of actor Johnny Depp's photos were evaluated in the system due to a large number of available different poses and expressions, fashion items, facial hair differences and age difference of available photos. Series of tests with brightness alteration and blur were performed, with modifications shown in Figure 4.

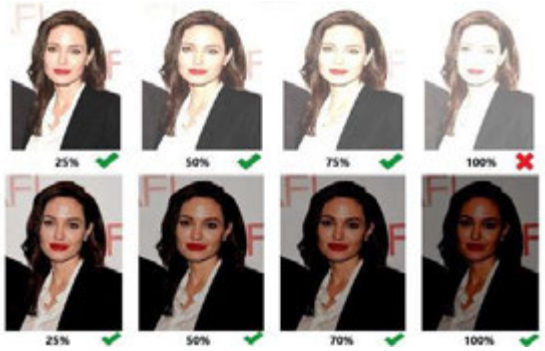


Figure 4. Tests with brightness alteration. Top row: increasing brightness from 25% to 100%. Bottom row: decreasing brightness by 25% in each step.

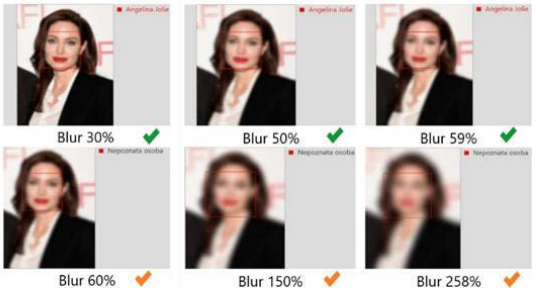


Figure 5. Blur transformations: 30%, 50%, 59%, 60%, 150%, 250%.

The system was further tested on a real-world example with a database of 21 students, initialized with only one photo that is at least 2, and at most 5 years old, from the student identification document upon registering at the university, as shown in Figure 6. We tested the system by trying to detect and recognize students in the classroom regularly, throughout two months, yielding different hair styles, slightly different occlusion ratio and different poses and positions. All photos were taken with a 12MP mobile camera. An example of one such photo is shown in Figure 7. We analyzed the data collected throughout 60 days and recorded if the student is present in the photo, if the student's face was detected and if the student was correctly identified. To simulate testing system efficiency in classrooms of larger size, we conducted a simple

experiment of taking the photo of a person at 10m, 20m and 30m distance from the camera. An example of photo taken from 30m distance is shown in Figure 8. Also, we supplied the original 12MP photo to the system, and reduced the photo in each experiment to 3MP and 1MP.



Figure 6. Student images database used to initialize the system, blurred to obfuscate identities.

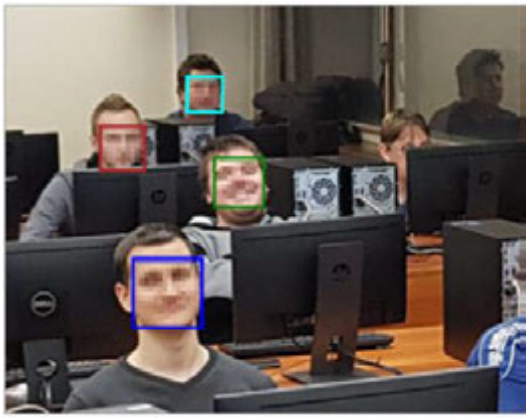


Figure 7. Example crops from the classroom. Faces were pixelated to obfuscate identity.



Figure 8. Experiment with face distance from camera at 10m, 20m and 30m. Face pixelated to obfuscate identity.

IV. RESULTS

The initial part of the experiment was conducted with storing one image of the person in the group, with two images and finally with five images in the group in the database. Makeup seems to have some impact on successful recognition, especially when eyebrows were altered. Also, younger photos proved to be difficult for recognition (TABLE I.). Adding or reducing brightness does not have a significant impact on person recognition (Figure 4), but applying too much blur does impact recognition accuracy. With blur up to 60%, a person is successfully recognized, but more blur affects recognition accuracy, and blur of 250% for that particular image also affects face detection accuracy (Figure 4 and Figure 5). Additionally, several other image filters were applied and tested, and the system proved robust to mosaic filter, tanning, grayscale transformation, poster, even fringe. However, negative transformation yielded unsuccessful face detection.

TABLE I. SYSTEM ROBUSTNESS REGARDING AGE, OCCLUSION AND FACE EXPRESSIONS

Parameters tested		**Number of images in the group**		
		1	***2***	***5***
Age	Younger	-	-	-/+
	Older	+	+	+
Face occlusion	Facial hair	+	+	+
	Glasses	+	+	+
	Glasses + facial hair	+	+	+
	Makeup	-/+	-/+	-/+
	Hair (covering part of face)	+	+	+
	Contact lenses	+	+	+
Face expressions and poses	Neutral	+	+	+
	Smiling	-/+	+	+
	Eyes closed	+	+	+
	Angry	+	+	+
	Profile shot	+	+	+

The second part of the experiment was conducted over a period of one whole semester, with images taken in the real-world environment within the classroom of 15-20 attendees. The results are presented in Figure 9. This experiment showed a system efficiency of 80% successful identifications out of 564 faces in 40 images. There were 14% successful detections, but without successful identification, and 6% of faces were not detected at all.

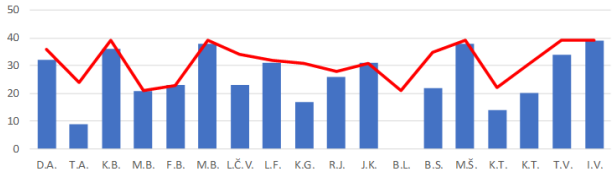


Figure 9. Indicating the total number of successful identifications and occurrences for each student. Student names and surnames are represented with initials. The red line shows the recorded number of

attendances, while the blue columns represent successful identifications for each student.

We further analyzed the factors which made the system fail to recognize or detect faces. Occlusion proved to be the crucial factor for face detection or recognition failure, with 54% misidentifications or misdetections. Face expression caused 20% of errors, head pose 17% and blur caused 9% of misdetections.

The last part of the experiment was testing the detection robustness based on the distance from the camera. For any larger classroom, minimal requirement proved to be a 16MP camera, as it is shown in TABLE II.

TABLE II. EXPERIMENT RESULTS REGARDING FACE DISTANCE FROM CAMERA AND IMAGE QUALITY

Photo resolution	**Distance from camera**		
	10m	***20m***	***30m***
16MP	OK	OK	OK
4MP	OK	NOT OK	NOT OK
1MP	NOT OK	NOT OK	NOT OK

V. CONCLUSION

Face detection and identification, even with a series of negative implications and concerns lately raised remains an active area of research. We gave a historical overview of face recognition technologies and presented popular algorithms for face identification. The main contribution of this paper is a detailed analysis of the Microsoft Azure Face API – system for face detection and recognition, with intent to use in student attendance statistics in the classroom. In our experiment, we conclude that the high number of misidentifications and misdetections was due to face occlusion, which does not leave a lot of space for improvements, but precautions can be taken to reduce the occlusion in real-world usage. In future work, we will experiment in a larger classroom with 100+ students to evaluate system efficiency in such conditions.

REFERENCES

[1] Ballantyne, M; Boyer, R S; Hines L: "Woody Bledsoe: His life and legacy", AI Magazine, Vol. 17. No. 1, pp. 7–20, Spring 1996

[2] Goldstein AJ; Harmon LD; Lesk AB: "Identification of Human Faces", Proc. IEEE 1971; Vol. 59, No. 5, 748-760.

[3] Kirby M; Sirovich L: "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces", IEEE Transactions on Pattern Analysis and Machine Intelligence 1990; vol. 12, No. 1, pp. 103-108

[4] Turk M; Pentland A: "Face recognition using eigenfaces" IEEE Conference on Computer Vision and Pattern Recognition 1991.

[5] Phillips P J; Wechsler H; Rauss P: "The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing" 1998; 16(5):295–306.

[6] Phillips P J; Rauss P; Der S: "FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results", U.S. Army Research Laboratory – via NIST, Arlington: DARPA, 1996.

[7] Zhao W. et al: "Face Recognition: A Literature Survey", ACM Computing Surveys, 2003.

[8] Bhaskar, N; Das, K M: "A Survey on Face Recognition Techniques", IJTIMES

[9] Lal M; Kumar K: "Study of Face Recognition Techniques: A Survey", IJTIMES, 2018.

[10] Admane, A. et al (2019). "A Review on Different Face Recognition Techniques", IJTIMES

[11] Turk M; Pentland A: "Eigenfaces for recognition. Journal of cognitive neuroscience", vol.3, pp.71-86, 1991.

[12] Phillips PJ: "Support vector machines applied to face recognition.", Adv. Neural Inform. Process. Syst. 11, 803–809, 1998.

[13] Bartlett MS; Movellan JR; Sejnowski TJ: "Face recognition by independent component analysis", IEEE Trans Neural Networks 13, 1450–1464, 2002.

[14] Jin Y; Ruan QQ: "Face recognition using gabor-based improved supervised locality preservingprojections." Computing and Informatics 28.1, 81-95, 2012

[15] Lu,Kostantinos J; Plataniotis N; Venet sanopoulos AN: "Face Recognition Using LDA- Based Algorithms" IEEE Transactions in Neural Network,Vol.14 No.1, 2003.

[16] Yu H; Yang J: "A direct LDA algorithm for high-dimensional data—with application to face recognition,Pattern recognition,Pergamon", vol.34, pp.2067-2070, 2001.

[17] Simonyan K; Zisserman A: "Very deep convolutional networks for large-scale image recognition", ICLR, 2015

[18] Hoo-Chnag S et al: "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning", IEEE Trans Med Imaging, 2016

[19] Stanford, "CS231n Convolutional Neural Networks for Visual Recognition", http://cs231n.github.io/convolutional-networks/#overview, 20.06.2019

[20] Buduma N; Lacascio N: "Fundamentals of Deep Learning", USA: O'REILLY, 85-115, 2017

[21] Krizhevsky A; Sutskever I; Hinton G.E: "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, pp. 1097–1105, 2012

[22] Srivastava Y; Murali V; Dubey S.R: "A Performance Comparison of Loss Functions for Deep Face Recognition", https://arxiv.org/pdf/1901.05903.pdf, 21.06.2019

[23] Deng J; Guo J; Xue N: "ArcFace: Additive Angular Margin Loss for Deep Face Recognition ", https://arxiv.org/pdf/1801.07698.pdf, 20.06.2019

[24] Deng J., Zhou Y., Zafeiriou S.,"Marginal Loss for Deep Face Recognition ", CVPR

[25] Microsoft Azure Face API Documentation, https://docs.microsoft.com/en-us/azure/cognitive-services/face/ (01.05.2019)

A Scalable K-Nearest Neighbor Algorithm for Recommendation System Problems

A. Sagdic*, C. Tekinbas*, E. Arslan*, and T. Kucukyilmaz*
* TED University, Computer Engineering Department, Ankara, Turkey
ahmet.sagdic@tedu.edu.tr, cihad.tekinbas@tedu.edu.tr,
enes.arslan@tedu.edu.tr, tayfun.kucukyilmaz@tedu.edu.tr

Abstract - **Memory-based classification techniques are commonly used for modeling recommendation problems. They rely on the intuition that similar users and/or items behave similarly, facilitating user-to-item, item-to-item, or user-to-user proximities. A significant drawback of memory-based classification techniques is that they perform poorly with large scale data. Thus, using the off-the-shelf classification techniques for recommendation problems generally lead to impractical computational costs.**

In this study, we propose a recommendation problem-specific enhancement for a widely known memory-based classification algorithm, *K*-Nearest Neighbor. For this purpose, the movie recommendation problem is selected, and the scalability of the proposed enhancement is evaluated on three publicly available datasets. In the proposed enhancement, user- and item-proximities are pre-calculated during the first offline recommendation, while an auxiliary data structure is constructed for keeping user-to-user proximities. The stored neighborhood information is then facilitated in order to speed up later recommendations. The experiments show that the proposed algorithm has performed superior to both the classical classification technique and the state-of-the-art off-the-shelf toolkits.

Keywords - **Recommendation Systems, Collaborative Filtering, Memory Based Classification, Recommendation Performance.**

I. INTRODUCTION

Due to the widespread use of the Internet and increasing popularity of using Web technologies in commerce and advertisement, recommender systems become an integral part of many businesses in various domains such as video streaming, gaming, e-commerce, and social network services. In general terms, recommender systems provide users with item suggestions where, an item can be anything that is related with the provided service. The basic idea of recommender systems is to utilize (item or user) feedback in order to infer user interests. With the growing size of the online data, providing high-quality recommendations in a fast and scalable way became an important objective of research in the area. Various techniques and algorithms are proposed for the recommendation problem in the literature. Among them, collaborative filtering and content-based filtering are two of the widely used and commonly known techniques for recommendation. Collaborative filtering is a technique that infers differences/similarities between users or users' item preferences while content-based filtering is based on inferring the same relations based on the similarities between the items that users has interacted within the past.

Due to the obvious similarities between the recommendation and the classification problem, utilization of the supervised classification techniques for solving recommendation problems is a common practice in many studies [1][2]. Although such a formulation is straightforward, using off-the-shelf classification tools and techniques for solving recommendation problems has serious drawbacks: First, unlike the classification problems, the distinction between an attribute and an objective is not clear for a recommendation problem. Attributes (past item or user interactions) that define the similarities in order to infer a recommendation for a user could become the inferred recommendation for some other user [3]. Consequently, attributes defining the similarities between two instances may have missing attribute values for classification due to the nature of the problem and may require additional forecasting or attribute completion techniques, ultimately reducing the quality of the results. Second, recommendation problem requires application of forecasting in an incremental or iterative fashion while, supervised classification techniques are based on a pre-trained model. Hence, many of the off-the-shelf classification techniques are not fully optimized for recommendation problems. As an illustrative example, memory-based classification algorithms calculate item and user similarities in a per-instance basis while, it is possible for a recommender system to utilize past-computations for future predictions, significantly reducing the computation time.

In this paper, we focus on the latter of the drawbacks of using supervised classification techniques for the recommendation problem: the computational cost. We propose a recommendation system-specific enhancement on a widely used memory-based classification algorithm, *K*-Nearest Neighbor (*K*-NN) in order to improve its

computational performance. To this end, we perform extensive experiments on the scalability of the algorithm and report our results.

This study is organized as follows. In Section 2, the related work is presented. The proposed enhancements are described in detail in Section 3. Experimental framework, the datasets used in this study, and the results of the experiments are presented in Section 4. Section 5 concludes with a short discussion on the findings.

II. Related Work

Collaborative filtering is a popular and widely used technique in recommender systems. There are two commonly used models for collaborative filtering: memory- and model-based techniques.

Memory-based collaborative filtering techniques are also known as neighborhood-based collaborative filtering techniques due to the fact that the similarities or dissimilarities between users and/or items determine the outcome of the recommendations. Memory-based techniques can also be divided into two categories: user-to-user filtering where, similarities between users are facilitated in order to infer recommendations. The motivation behind user-to-user filtering is that the recommendations should be based on users' taste, and similar users would prefer similar items. In user-to--item filtering, similarities between items are used in order to suggest new items to a user. The motivation of user-to-item filtering is that the recommendations should be performed using the similarities between a particular item and a previously interacted set of items of a user. Popularity[4] and *K*-NN[5] are widely used examples of memory-based collaborative filtering techniques.

In model-based techniques, machine learning and data mining techniques are used in the context of generating predictive models [5]. The models are developed by evaluating the past ratings of the users. Positive Matrix Factorization (PMF)[6] and Bayesian methods[7] are some examples of model-based techniques.

Memory-based methods have the advantage of the ease of implementation, simplicity and interpretability. However, they often encounter scalability challenges. In general, the complexity of the memory-based methods is $O(fm^2)$ where, f is the number of attributes and m is the number of users. Poor scalability makes such methods impractical for many real life applications. Another important disadvantage of such methods is their limited coverage due to sparsity[8]. Since memory-based methods have low variance, they perform poorly for unobserved data. As an illustrative example, in a movie recommendation problem, if a movie has not been observed by the users in the training data, memory-based methods cannot infer similarities in order to provide sensible predictions.

Among recommendation problems, the movie recommendation problem is considered a staple benchmark for many studies due to large dataset size, number of involved users and the amount of variety in the problem space. A large number of studies has preferred the movie recommendation problem as a benchmark in order to present their findings, using both collaborative filtering, content-based filtering, and hybrid filtering methods [9][10].

III. Modified K-NN Framework

K-NN algorithm is a widely known memory-based classification algorithm. It is a supervised learning algorithm that trains a classification model via memorizing distance between instances in the training and the testing sets. The algorithm then performs predictions via a majority voting between the *K* closest neighbors for each tested instance. For calculating the distances between training and testing instances, it is possible to use several distance or similarity measures. In this study, we have adopted cosine similarity measure as the distance metric. Additionally, each recommendation in our version of the problem is considered as binary. Thus, selecting an odd *K* value guarantees that a decision is reached at the end of each majority voting.

Algorithm 1 provides a pseudocode for solving the recommendation problem using *K*-NN under the supervised classification setting. The inputs of Algorithm 1 are: a set of users denoted by the vector U, a set of movies denoted by the vector M, and a binary matrix R representing the user interactions with the movies. The algorithm proceeds as follows: First, as denoted at lines 1-3, the users are partitioned into two: some users are selected as the training data, denoted by U_{tr} and the rest of the users are selected as the testing data, denoted as U_{te}. Conforming to the above partition, matrix R is also

Algorithm 1 Pseudocode for the movie recommendation problem using *K*-NN

Input: U: users $U = \{u_1, u_2 ..., u_r\}$.
M: movies $M = \{m_1, m_2 ..., m_s\}$.
R: $R \in R_{r\,x\,s}$ where, $R_{i\,j} = 1$ if interaction is observed, $R_{i\,j} = 0$ otherwise.

Output: Calculated recommendations $R' \in R$

1: Training user samples U_{tr}: $U_{tr} \subset U$.
2: Testing user samples U_{te}, $U_{te} \subset U$ and $U_{tr} \cap U_{te} = \varnothing$.
3: Pending recommendations $R' \subset R$.
4: **for** r $\in R'$ **do**
5: let D be a set of neighbors where, $|D| = K$, $D = \varnothing$;
6: **for** $t_{te} \in U_{te}$ **do**
7: **for** $t_{tr} \in U_{tr}$ **do**
8: distance = *dist.* (t_{te}, t_{tr})
9: **If** distance < *max* ($D_i \in D$)
10: **If** $|D|$==K then $D = D$ - {*max* ($D_i \in D$)
11: $D = D \cup t_{tr}$;
12: **end for**
13: R'_{te} = *argmax*($|d \in D|$: $R_{d,r} = 1$, $|d \in D|$: $Rd_{,r} = 0$)
14: **end for**
15: **end for**

Algorithm 2 Pseudocode for movie recommendation problem using modified *K*-NN

Input: U: users $U = \{u_1, u_2 ..., u_r\}$.
M: movies $M = \{m_1, m_2 ..., m_s\}$.
R: $R \in R^{r \times s}$ where, $R_{i\,j} = 1$ if interaction is observed, $R_{i\,j} = 0$ otherwise.

Output: Calculated recommendations $R' \subset R$

1: Training user samples U_{tr}: $U_{tr} \subset U$.
2: Testing user samples U_{te}, $U_{te} \subset U$ and $U_{tr} \cap U_{te} = \varnothing$.
3: Pending recommendations $R' \subset R$.
4: **for** $t_{te} \in U_{te}$ **do**
5: let D be a set of neighbors where, $max(|D|) = K$, $D = \varnothing$;
6: **for** $t_{tr} \in U_{tr}$ **do**
7: distance = *dist.* (t_{te}, t_{tr}). // calculating distance
8: **If** distance < $max\ (D_i \in D)$
9: **If** $|D| == K$ **then** $D = D - \{max\ (D_i \in D)\}$.
10: $D = D \cup t_{tr}$.
11: **end for**
12: **for** $r \in R'$ **do**
13: $R'_{te} = argmax(\ |d \in D|: R_{d,r} = 1\ ,\ |d \in D|: R_{d,r} = 0)$
14: **end for**
15: **end for**

partitioned where, a set of interactions for the testing users are selected as pending recommendations (i.e., the objective values) denoted as R'.

For each pending recommendation in R', the algorithm proceeds as follows: at lines 6-12, the distance between each training and testing user is calculated via a distance function, modeled using the values in R, representing the attributes of an instance. Among all calculated distances, the K closest training instances for each testing instance is recorded. At line 13, a majority voting is performed in order to give a recommendation decision.

Note that the computational bottleneck of the algorithm is the distance calculation. For each pending recommendation, a distance calculation between all training-testing user pairs is performed. Consequently, the complexity of the *K*-NN algorithm for the supervised classification setting is $O(|U_{te}| \times |U_{tr}| \times |R'|)$.

An important observation on Algorithm 1 is that the training-testing user distance calculation at line 8 is repeated for each pending recommendation in R'. Attributes representing a training and a testing user does not change during each iteration of the distance calculation. Consequently, it would be preferable to calculate the distances once in order to determine the K nearest training users for each testing user, store them in an auxiliary data structure, and check only these users' involvements in order to give a recommendation decision. The modified *K*-NN algorithm is presented in Algorithm 2.

In the algorithm, the order of the loops is changed in such a way that the distance calculations are only calculated once for each testing user. Additionally, the loop at lines 6-11 is unrolled so that, the distances between each training-testing pair is immediately used to infer all pending recommendations after determining the K closest training users for each testing user. Due to the above modifications, the complexity of the *K*-NN algorithm tailored for the recommendation problem is $O(|U_{te}| \times |U_{tr}| + |R'|)$ where, the first component of the notation denotes the distance calculation while, the second component denotes majority voting for each pending recommendation.

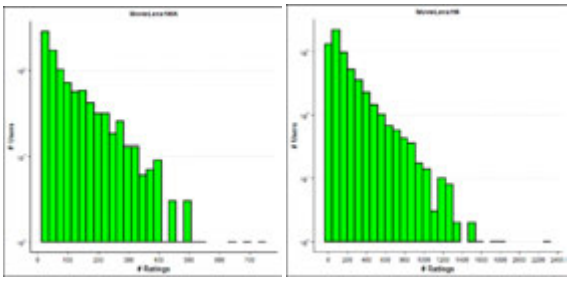
IV. Empirical Analysis

In this section, the datasets that are used in this study are presented along with the results of several experiments on the scalability of the *K*-NN algorithm. Specifically, we have compared the performances of both the *K*-NN supervised classification technique (KNN), a modern implementation of the *K*-NN algorithm in Phyton SkLearn library facilitating fast distance calculations (KNN-sklearn), and the proposed *K*-NN algorithm (KNN-m) on three publicly available datasets. To this end, four parameters are evaluated. These parameters are: size of the training set in terms of instances, size of the training set in terms of attributes, size of the testing set in terms of instances (i.e., the number of pending recommendations), and size of the testing set in terms of attributes.

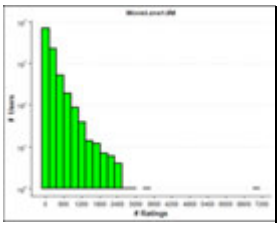
Three datasets with varying number of users, movies, and user ratings are selected for the experiments, which are collected by the GroupLens Research Project at the University of Minnesota [11]. The datasets contain a set of users, a set of movies, and rating values of users for these movies where, the rating scores are ranked between 1 and 5. For the purpose of this study, out of the three datasets two of them, namely MovieLens100k containing 100,000 user reviews and MovieLens1M containing roughly 1,000,000 user reviews are directly adopted for the experiments. We have randomly selected 10,000 users due to exceedingly large execution times from the MovieLens10M dataset containing roughly 1,800,000 reviews in order to form the third experimental dataset (denoted as MovieLens1.8M). Table 1 summarizes the contents of these datasets. Additionally, Figure 1 represents the distribution of ratings among users within the datasets. As evident from the figure, in terms of movie ratings, all datasets are quite sparse.

TABLE I. MovieLens Datasets Details

Name	# Users	# Movies	# Ratings
MovieLens100K	943	1,682	100,000
MovieLens1M	6,040	3,706	1,000,209
MovieLens1.8M	10,000	10,263	1,823,642



A. MOVIELENS100K B. MOVIELENS1M



C.MOVIELENS1.8M

Figure 1. Distributional characteristics of the MovieLens Datasets.

As a preliminary step, a user-movie matrix *R* is created for our experiments. In this matrix, rows represent the users and columns represent the movies. For the purpose of our experiments, it is assumed that it is possible to fit matrix *R* into memory; hence the following discussions do not involve disk access times. Each movie rating is then converted into a binary recommendation value such that the user interest (i.e. user providing a rating for the movie) is converted to a 1, and if the user has no interaction with the movie, the recommendation value is set to 0. Next, the dataset is divided into a training and a testing set for simulating the recommendation problem. To this end, 80% of the user recommendations in *R* are selected as the training set and the remaining recommendations are used as the testing set for the purpose of the experiments in a random fashion.

A. Experimental results

To demonstrate the scalability of the proposed enhancements, four experiments are conducted. As for the value of *K*, we have adopted 5 in the experiments. The computational performance of the modified *K*-NN is then compared with the other algorithms.

TABLE II. COMPUTATION TIMES FOR VARYING NUMBER OF TRAINING INSTANCES

Dataset	# Training Instance	KNN (sec.)	Knn-sklearn (sec.)	KNN-m (sec.)
MovieLens100K	754	134.42	72.75	1.06
	377	66.09	44.10	0.42
	188	32.61	27.55	0.31
	94	16.56	8.13	0.28
MovieLens1M	4832	27702.95	6401.47	37.57
	2416	13622.27	3113.18	20.04
	1208	6764.18	1467.38	11.00
	604	3383.45	656.49	6.98
MovieLens1.8M	8000	555353.51	63993.50	240.88
	4000	255673.41	32417.94	126.70
	2000	161513.56	14864.81	69.51
	1000	96784.45	8797.75	40.78

TABLE III. COMPUTATION TIMES FOR VARYING NUMBER OF TRAINING ATTRIBUTES

Dataset	# Training Attribute	KNN (sec.)	Knn-sklearn (sec.)	KNN-m (sec.)
MovieLens100K	1346	154.55	78.60	1.01
	673	101.34	72.75	0.52
	336	78.54	71.56	0.45
	168	69.45	69.32	0.42
MovieLens1M	2965	27702.95	6401.47	37.59
	1482	16016.06	5688.29	24.11
	741	10804.85	5537.47	16.86
	371	8156.76	5421.70	13.52
MovieLens1.8M	8210	555353.51	63993.50	239.83
	4105	327899.19	56162.83	133.97
	2052	160666.51	55163.38	83.96
	1026	138313.31	55153.98	57.90

In the first experiment, we evaluate scalability with respect to the training dataset size. For this purpose, a set tests are conducted for each dataset using varying number of users in the training set. For each dataset 100%, 50%, 25%, and 12.5% of the users in the training set are selected in a random fashion for training the recommendation model. The results of the experiment is provided in Table 2.

The table is divided into two partitions representing the results for each dataset with varying number of training users. The rows of the table represent the computation time for each setting (in seconds). The results indicate that, as the number of users in the training set increase (along with the number of user-movie pairs in *R*), the computation time also increases linearly for the *K*-NN classification algorithm. The same linear scalability can also be observed for the KNN-sklearn algorithm. Due to fast calculation of the distance metric, KNN-sklearn performed better than the original K-NN algorithm.

Similar relation between the size of the training dataset and computation time can also be observed in the modified *K*-NN algorithm. However, as described at lines 6-11 of Algorithm 2, due to the pre-computation and re-use of the neighborhood information across multiple recommendations, the computation time for the modified algorithm is 1.4% to 0.5% of the original algorithm.

In the second experiment, we conducted our experiments for evaluating the scalability in terms of the number of attributes in the training dataset. Similar to the first experiment, 100%, 50%, 25%, and 12.5% of the available attributes (i.e., movies) are randomly selected for training the recommendation model. The results are presented in Table 3.

The effect of varying attribute size is more subtle for both algorithms. The reason behind this phenomenon is

TABLE IV. COMPUTATION TIMES FOR VARYING NUMBER OF TESTING INSTANCES

Dataset	# Testing Instance	KNN (sec.)	Knn-sklearn (sec.)	KNN-m (sec.)
MovieLens100K	189	143.90	78.68	0.73
	94	72.07	49.55	0.34
	47	35.80	36.66	0.18
	24	18.26	22.51	0.11
MovieLens1M	1208	27702.95	6401.47	38.05
	604	14181.07	3226.14	18.88
	302	7084.42	1706.85	9.76
	151	3548.03	994.83	4.54
MovieLens1.8M	2000	555353.51	63993.50	243.66
	1000	304628.79	34614.60	121.71
	500	163354.66	17489.91	61.04
	250	95486.71	8447.86	30.67

that, as illustrated in Fig. 1, the number of positive recommendations is quite sparse for both datasets. Both algorithms incur a computation only if a training-testing user pair has a common attribute interaction (i.e., both training and testing users has interacted with the particular movie) over the same movie. Otherwise, the attribute does not incur an additional distance cost and can be trivially omitted. Hence, due to the sparsity of the movie interactions, doubling the attribute size has a relatively smaller effect on the computation time. In both cases however, similar to the first experiment, the modified *K*-NN algorithm performed significantly better than other algorithms with respect to computational time.

In the third experiment, we evaluate the testing time scalability of the proposed method. In this setting, a set of experiments are conducted with varying number of testing users. For this purpose, 100%, 50%, 25%, and 12.5% of the testing users are randomly selected for evaluating the testing performance of the algorithm. The results are presented in Table 4.

As both line 7 of Algorithm 1 and Line 6 of Algorithm 2 suggests, both versions of the *K*-NN algorithm performs distance calculations for each of the testing users. Hence, the computation time in both versions scale with the number of testing users linearly. The proposed enhancements does not change scalability with respect to the testing size. For the KNN-sklearn algorithm, same phenomenon can also be observed.

TABLE V. COMPUTATION TIMES FOR VARYING NUMBER OF TESTING ATTRIBUTES

Dataset	# Testing Attribute	KNN (sec.)	Knn-sklearn (sec.)	KNN-m (sec.)
MovieLens100K	336	145.20	77.96	0.79
	168	72.30	37.51	0.58
	84	36.04	19.40	0.53
	42	18.00	9.75	0.51
MovieLens1M	741	27702.95	6401.47	38.22
	370	13030.50	2966.44	36.27
	185	6469.29	1876.78	35.59
	93	3255.17	913.83	35.09
MovieLens1.8M	2053	555353.51	63993.50	244.18
	1026	301070.10	32232.79	235.01
	513	121511.66	18144.05	231.99
	257	59206.34	10394.11	230.43

The last experiment involves scalability with respect to the number of pending recommendations calculated by each algorithm. Similar to the previous experiments, in this setting we have tested involvement of varying number of movies (i.e., attributes) in the testing set. That is, for each user, 100%, 50%, 25%, and 12.5% of the movies are selected for pending recommendations in the testing set. Table 5 summarize the results of the experiments.

Unlike experiment 2, the variation in the number of pending recommendations linearly affect the *K*-NN computational performance. The result is more visible in the original *K*-NN: The computation time roughly doubles with the number of pending recommendations. Same can also be observed for KNN-sklearn. In the proposed *K*-NN algorithm, as stated in Algorithm 2 at lines 6-10, for each pending recommendation, an auxiliary data structure is set up, which is later used for all upcoming recommendations for the same user. This operation adds an extra initial cost to the algorithm, which is then facilitated in the forthcoming iterations. Although the rest of the computation again scales linearly with the number of pending recommendations, the overall computational cost is minimized due to the use of the auxiliary data structure during computation.

V. CONCLUSION

This study is motivated by the fact that despite the similarities, the classification and recommendation problems have inherent differences. Although it is possible to convert a recommendation problem into a classification setting, and there are already many toolkits and applications in the market for solving supervised classification efficiently, the algorithms such packages provide are not inherently optimized for solving the recommendation problem. In this work, we examined a particular memory-based algorithm, *K*-NN and propose improvements in order to enhance its computation time specifically for the recommendation problems. Both analysis and experimental results show that the proposed algorithm has better performance than both classical and off-the-shelf *K*-NN algorithms, and its computation time scales linearly with the number of training instances, number of testing instances, the number of attributes, and the number of pending recommendations.

REFERENCES

[1] G. Jawaheer, P. Weller, and P. Kostkova, "Modeling User Preferences In Recommender Systems: A classification framework for explicit and implicit user feedback", ACM Transactions on Interactive and Intelligent Systems 4, 2, Article 8 (May 2014), 26 pages, 2014.

[2] V. Pronk, W. Verhaegh, A. Proidl, and M. Tiemann, "Incorporating user control into recommender systems based on

Naïve Bayesian classification", in Proceedings of the 2007 ACM Conference on Recommender Systems. pp. 73-80, 2007.

[3] C.C. Aggarwal, "Basic models of recommender systems" in Recommender Systems, Springer, pp. 8-10, 2016.

[4] X. Yu, X. Ren, Y. Sun, B. Sturt, U. Khandelwal, and Q. Gu, "Recommendation in heterogeneous information networks with implicit user feedback", in Proceedings of the 7th ACM conference on Recommender Systems. pp. 347-350, 2013.

[5] D.A. Adeniyi, Z. Wei, and Y.Yongquan, "Automated Web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", Applied Computing and Informatics. vol. 12, pp. 90-108, 2014.

[6] L. Hu, Y. Wang, Z. Xie, and F. Wang, "Semantic preference-based personalized recommendation on heterogeneous information network", in IEEE Access. vol. 5, pp. 19773 – 19781, 2017.

[7] P. Valdiviezo-Diaz, F. Ortega, E. Cobos, and R. Lara-Cabrera, "A collaborative filtering approach based on Naïve Bayes classifier", In IEEE Access. vol. 7, pp. 108581 – 108592, 2019.

[8] C.C. Aggarwal, "Neighborhood-based collaborative filtering" in Recommender Systems, Springer, pp. 44 – 45, 2016.

[9] Y. Wang, M. Wang, and W. Xu, "A sentiment-enhanced hybrid recommender system for movie recommendation: a big data analytics framework", Wireless Communications & Mobile Computing, pp. 33-42, 2018.

[10] C. Christakou, S. Vrettos, and A. Stafylopatis, "A hybrid movie recommender system based on neural networks", in *5th International Conference on Intelligent Systems Design and Applications,* pp. 500-505, 2005.

[11] F. Maxwell Harper and Joseph A. Konstan, "The MovieLens datasets: history and context." ACM Transactions on Interactive Intelligent Systems (TiiS) vol. 5, No 4, Article 19 (December 2015), 19 pages, 2016.

Retinal OCT Image Segmentation: How Well do Algorithms Generalize or How Transferable are the Data?

M. Melinščak[1,2] and S. Lončarić [2]
[1] Karlovac University of Applied Sciences, Karlovac, Croatia
[2] Faculty of Electrical Engineering and Computing, Zagreb, Croatia
martina.melinscak@gmail.com

***Abstract* - The success of deep learning depends, among other things, on a large amount of labeled data. However, in medical applications, large labeled datasets are the exception, rather than the rule. Manual image labeling is time-consuming and is generally performed only with the purpose of developing algorithms, and not as a part of standard clinical practice. The goal of this study is twofold. Since there is always a trade-off between the ability to collect data and achieve the best possible performance, we wanted to explore how performance depends on the amount of data. For this purpose, a database of manually annotated OCT images was collected. Also, we wanted to see how much transfer learning can help. Retinal OCT images vary depending on the type of device, therefore developed methods should be as robust as possible. Transfer learning was performed so that the model was trained with similar OCT images and then fine-tuned with images from the collected database. It has been shown that transfer learning helps in terms of generalization and better prediction in case the source database is similar to the target database. We can also assume that further improvement can probably be achieved by adding images from another distribution (medical or nonmedical).**

Keywords* – deep learning*; retinal OCT images; image segmentation;

I. INTRODUCTION

Age-related macular degeneration (AMD) is one of the most common diseases leading to vision loss in the elderly population. Currently, 170 million people worldwide are suffering from it and, by some estimates, 288 million will be affected by 2040 [1]. Due to the disease, pathological biomarkers such as intraretinal fluids appear and there is a significant disturbance in the distribution and thickness of the retinal layers [2]. Since the introduction of optical coherence tomography (OCT), imaging has been done using OCT devices. Diagnostics and treatment in clinical practice rely only on the qualitative analysis and assessment of the ophthalmologist. For more successful diagnostics and therapy, computer segmentation is therefore necessary in quantitative analysis. It should also contribute to a better understanding of the disease that has not yet been fully explored.

Recently, the segmentation of medical and thus ophthalmic images is performed almost exclusively using deep learning methods. One of the preconditions for the success of deep learning algorithms is a large amount of data available. But in medicine, this is rarely the case. In the case of supervised learning, images manually annotated by an expert are required, which is very time-consuming. Even more, image labeling is not part of clinical practice, and labeling is done solely to develop machine learning methods. As images from different OCT devices and different device generations vary, it is required that the segmentation algorithms be as robust as possible. This means that in case of migration to a new device it is not necessary to re-annotate a larger number of images, and existing algorithms will be upgraded with a smaller number of new images.

Although transfer learning is often applied in medicine, by using the known architectures of neural networks (Xception, Inception V3, ResNet50, VGG16, VGG19, MobileNet) pretrained on the ImageNet dataset, in ophthalmology this is rarely the case [3]. Several papers [4]–[6] examine the adequacy of such a method since ImageNet dataset is a database of natural images that bears no resemblance to medical ones. In a recent paper [7], a hybrid architecture was proposed in which only the first few layers of networks such as the ResNet architecture (pretrained on ImageNet) would be used and then some lighter versions of CNN would be applied (and fine-tuned) in the end.

The aim of this study was twofold. First, to see how the performance of the algorithm depends on the amount of data. To begin with, the network was trained from scratch with a different amount of data to see how performance varies. Segmentation was done using the U-net architecture [8], which is almost standard in the segmentation of medical images. In order to perform the segmentation of retinal structures and pathological biomarkers, a database of images with manually labeled features was collected. And second, we explored how many additional labeled images are needed and how much transfer learning can help in the case the network was trained on one dataset (source data) and then migrated to another (target data). As a source dataset, the image database from the RETOUCH challenge (The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge) was used [9]. Transfer learning was then performed so that the network was trained with part of the

images from RETOUCH challenge and then fine-tuned with the images of the target database: again, with a different number of images.

II. MATERIAL AND METHODS

A. Database

For the research, a database of manually annotated images was collected. For convenience, let it be called the target dataset. In collaboration with Sisters of Charity Hospital (KBC Sestre milosrdnice), Zagreb, images were annotated for 25 patients with neovascular age-related macular degeneration (nAMD). Macular SD-OCT volumes were recorded with the Zeiss Cirrus HD OCT 4000 device. Each OCT volume consisted of 128 B-scans with a resolution of 1024 x 512 pixels (pixel size 1.96 x 11.74 μm). Retinal fluids were annotated for 1224 B-scans. Of the intraretinal fluids, the following were annotated: pigment epithelial detachment (PED), subretinal fluid (SRF), and intraretinal fluid (IRF) (Fig. 1).

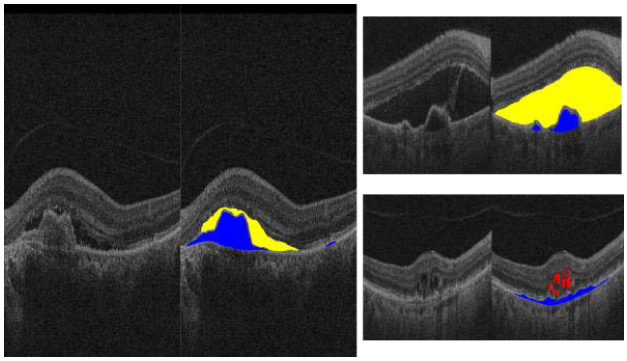


Figure 1. Examples of B-scans from a dataset with manual annotations of retinal fluids. PED is colored with blue, SRF with yellow, and IRF with red. On the left there is an example of an entire B-scan (1024x512 pixels). It is visible that there is a large amount of background and thus a large class imbalance. Examples cropped to the region of interest (ROI) are on the right.

Data collection adhered to the tenets of the Declaration of Helsinki and the standards of Good Scientific Practice of Sisters of Charity Hospital. The presented study was approved by the Ethics Committee of the Sisters of Charity Hospital and Faculty of Electrical Engineering and Computing, Zagreb, Croatia.

To estimate human error, for 75 B-scans, the same expert made a re-annotation with a time delay (and without insight into former annotations). Also, annotations were made by another expert. An intra-observer and an inter-observer error were calculated. Dice scores for inter-observer and intra-observer errors are 0.822 and 0.889, respectively. Fig. 2 shows two examples. It became evident that there are some even major differences in the opinions of experts. The pathology of the disease is not sufficiently known, and it is often difficult to decide what type of fluid was present, especially since in addition to intraretinal fluids there are other pathological phenomena (hyperreflective foci, druse, cysts, etc.). Due to the poor quality of images (a large amount of speckle-noise), and the necessary prior knowledge of retinal anatomy and pathological changes, it is difficult to expect great accuracy (especially in the regime of small data) as it is a demanding task even for humans.

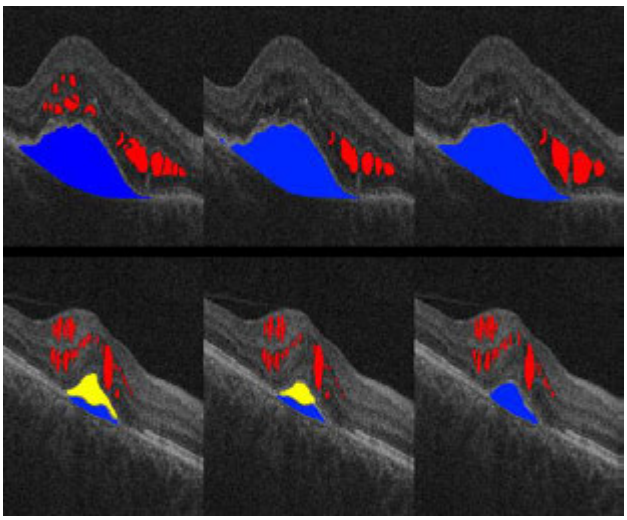


Figure 2. Two examples of B-scans from a dataset with manual annotations (from left to right): annotation from the 1st expert (considered as ground truth); annotation form the 1st expert with time delay and with no insight into previous annotation (used to calculate intra-observer error) and annotation from the 2nd expert (used to calculate inter-observer error). All images are cropped to show just ROI.

The other dataset we used is from the RETOUCH challenge [9] – let it be called source dataset. The database consists of images from three types of devices (Topcon, Spectralis, and Cirrus). For each type of device, all B-scans for 24 patients were labeled. We used only part of the images from the Cirrus device (1000 randomly picked images) to train the network. Unlike the target dataset, where all B-scans contained intraretinal fluids, many B-scans did not contain pathological changes (whole B-scans belonged to the background class). It resulted in the even greater disproportion of background pixels and pixels belonging to fluids. The database included images of patients with AMD, retinal vein occlusion (RVO), and diabetic macular edema (DME). More details can be found on the website [9].

B. Segmentation approach

For segmentation we used the U-net architecture [8]. The U-net architecture, its various modifications [10], stacking of multiple U-net architectures, and combining the U-net architecture with other machine learning methods [11], [12] have become standard methods for segmenting OCT images in recent years. Eight teams participated in the RETOUCH challenge in 2017, and the results and algorithms of all participants were presented in [13].

To make it easier to see the dependence of the algorithm's performance on the amount of data, the problem was reduced to binary segmentation (all three fluids belonged to the same class). Data augmentation, which is usually used in the case of a small amount of data, hadn't been performed, in order to better isolate the dependence of network performance on the amount of data.

In the first case, the target dataset was divided into a training set (1000 images) and a test set (224 images). Furthermore, the training set was divided into three groups with a different number of images: 256, 512, and 1000. The network was trained from scratch. Dice loss was used to train the model, since the dice score is usually used as a metric for evaluation, and the goal was to maximize it. We tested the model training with binary cross-entropy loss, but worse results were obtained. The batch size was set to 4. Adam optimizer with learning rate 1e-5 was used. The model had 7759521 trainable parameters. Training was performed during 70 epochs, without early stopping. Dropout was used to prevent overfitting. The validation dataset was 10% of the training dataset.

In the second case, the network training was performed with the source dataset. The same hyperparameters were used. The network had only been trained for 20 epochs, as the purpose was only to have a pretrained network. Fine-tuning of all layers with the same three data groups from the target dataset (256, 512, and 1000) was then performed. Training was executed during 50 epochs with all the same parameters as in the first case.

The model was trained on Google Colab [14] with a GPU. The code was written in Keras with the TensorFlow backend. Code is available on GitHub: https://github.com/mmelinscak/MIPRO_2020

III. RESULTS

In the first case, when the model is trained from scratch, as expected, the performance is better due to the larger number of images (it is well known that deep learning algorithms are "data-hungry"). Obtained dice scores (mean ± SEM) on a test dataset (224 images) for 256, 512, and 1000 images are: 0.579±0.272, 0.632±0.279, and 0.731±0.268, respectively.

Fig. 3 shows examples for one raw image and ground truth mask, followed by model predictions on the test set in case the model is trained with 256, 512, and 1000 training images. As it is as expected, segmentations are better with a larger amount of data. With an experiment like this, we can get a better insight into how much data we need, since there is always a trade-off between the amount of manually annotated data we can/want to collect and the performance of algorithms. The second row shows segmentation predictions for the second case (with the model pretrained on source dataset and then fine-tuned on target dataset). In this case, obtained dice scores (mean ± SEM) on the test dataset were 0.612±0.284, 0.677±0.272, and 0.826±0.270 respectively. It is evident that predictions after transfer learning have more "serrated" shapes while predictions from a model trained from scratch are more oval. As shape types are one of the key features in distinguishing different kinds of intraretinal fluids, this is a promising result.

Fig. 4 shows a graph with mean dice scores for all cases. It is visible that the mean value of the dice score is higher in the second case (with transfer learning).

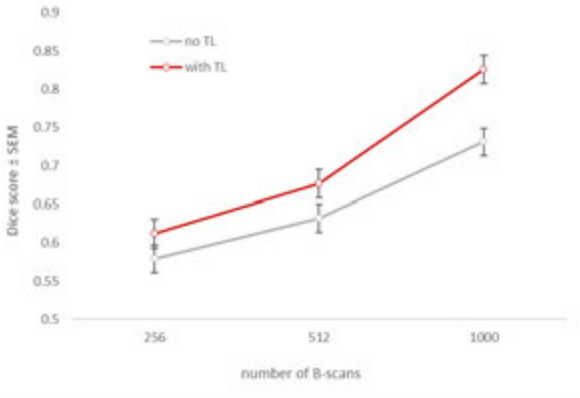


Figure 4. Mean dice score with error bars (standard error of the mean – SEM) for training model from scratch (no TL) and in case of fine-tuning the pretrained model (with TL).

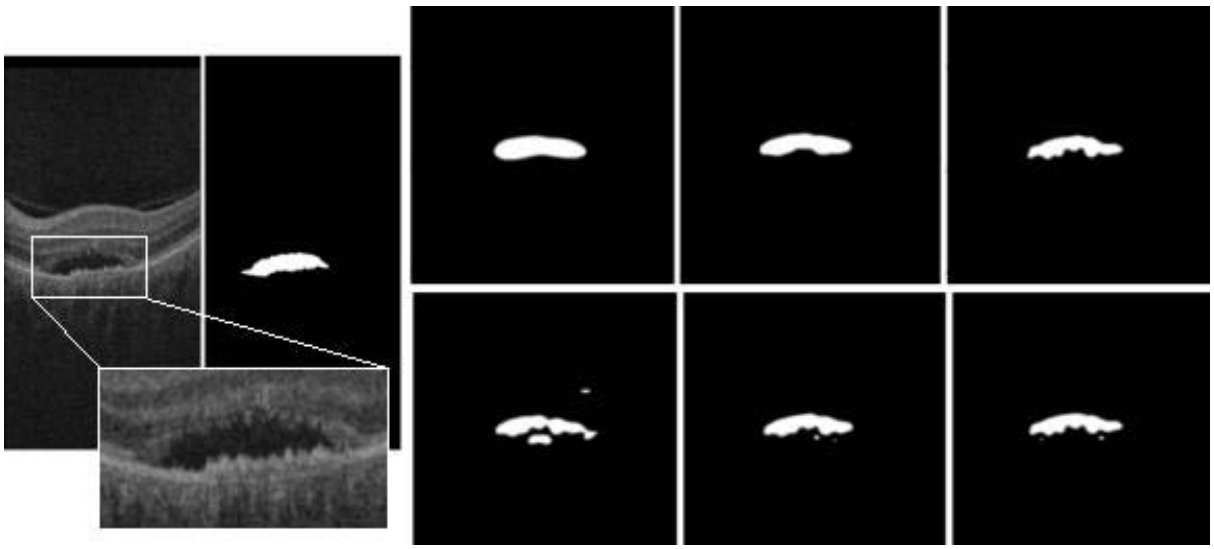


Figure 3. First row (from left to right): raw image from test dataset, ground truth (manual annotation), prediction from a network trained with 256 images, 512 images, and 1000 images. Second row (from left to right): predictions after training the network with 1000 images from source dataset and fine-tuning pretrained network with 256 images, 512 images, 1000 images (from left to right) from target data.

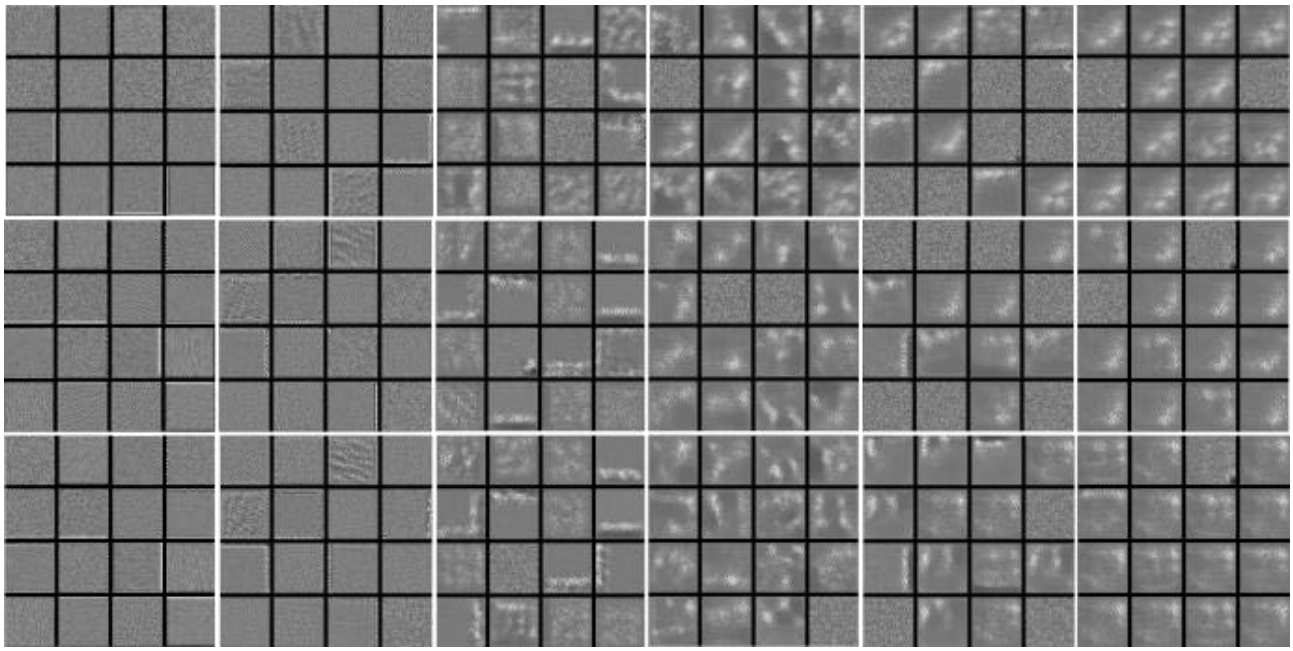


Figure 5. Visualization of convolution layers 2, 4, 7, 10, 13, 16 (from left to right). Layers 2, 4, and 7 belong to the encoder part of U-net architecture, and 10, 13, and 16 to the decoder part. First row: network trained from a scratch with 256 images (target data). Second row: network trained with 1000 images (source data). Third row: network fine-tuned with target data (256 images) after learning with source data (1000 images).

Although the model trained with 1000 images from the source dataset achieved a high value for dice score on training test (0.717), the value of the dice score is very low for the test set (0.348). The reason is probably that both datasets cannot be considered to fit the same distributions despite the same type of OCT device being used in both. One of the reasons are the differences in annotations (we have seen that there is quite a significant inconsistency between expert opinions). Another reason may be that the source dataset contains images of patients with three types of disease, and the target dataset only contains patients with AMD. Therefore, we have a case of a model trained on data from one distribution and tested on a different image distribution.

Fig. 5 shows a visualization of convolution layers in case of a training model from scratch with 256 images (first row). Then for the pretrained model (second row), and after fine-tuning the pretrained model with 256 images (third row). First three layers belong to the encoder and the last three layers to the decoder. It is visible that in the third row, convolution layers appear to be a combination from convolution layers shown in the first two rows, as might be expected. It is evident that convolution layers are clearer and more completed after fine-tuning with additional 256 target images then after only training with 1000 source images. It can be concluded that using images from different distributions leads to better results than increasing the number of images from the same distribution. However, the generality of this finding requires further investigation. Also, it is visible that the first layers of the encoder which should look like Gabor filters are not very clear, and maybe it would help to initialize the first few convolution layers with Gabor filters, and then to see if freezing or fine-tuning them leads to better results.

IV. CONLUSION

In this study, we wanted to examine how the performance of deep learning segmentation approaches depends on the amount of data. Since medical labeled images are difficult to collect, we found it useful to get an insight into the performance dependence of the algorithm on the number of images. Furthermore, we investigated how transfer learning affects the robustness of the algorithm and how much new data needs to be collected if the model is not trained from scratch but fine-tuned after it was pretrained on existing data.

Segmentation was performed with the basic U-net architecture. No data augmentation or other common ways of increasing performance were used to get a better result, since in the case of more complex architecture and changing multiple parameters it would be difficult to assess the dependence only on the amount of data.

It was demonstrated that training the model on very similar OCT images had a low dice score on the test set despite the relatively high dice score on the training dataset. This means that in the case of migration to a new OCT device, it would still be necessary to additionally train the model, but with a smaller number of images than in case of training the model from scratch.

Transfer learning is often applied in medicine, by using the known architectures of neural networks pretrained on the ImageNet dataset. ImageNet is a database with over 14 million natural images. Intuitively, it is not clear how would transfer learning help in case of medical images which are often grayscale and dissimilar to natural images. Optimizing transfer learning (with model pretrained on ImageNet) is still an active area of research in the case of medical images.

By visualizing convolutional filters, it becomes visible what the network is actually learning. In the study, it has been demonstrated that the initial layers, that are usually similar to Gabor filters, are not clear even after transfer learning when the source database is similar. Our assumption is that better generalization and prediction is possible by adding images from another distribution (medical and/or non-medical images). In addition to adding images from another distribution, our assumption is that incorporating some prior knowledge would help. Mapping from human to computer methods is not always completely identical but can help and serve as a useful analogy. Ophthalmologists use some general features of vision developed during life, some specific features learned about ophthalmic images, and use knowledge of retinal anatomy and pathological changes. We surmise that optimal results might be achieved by similar combinations: learning on a large number of images such as ImageNet, learning on the ophthalmic images themselves, and finally by adding some prior knowledge. Further research will test these assumptions.

ACKNOWLEDGMENT

We would like to thank the Sisters of Charity Hospital, the Clinic for eye diseases (Zagreb, Croatia). Special thanks to Prof. Zoran Vatavuk, MD, and to Marin Radmilović, MD who did image labeling.

REFERENCES

[1] U. Schmidt-Erfurth, A. Sadeghipour, B. S. Gerendas, S. M. Waldstein, and H. Bogunović, "Artificial intelligence in retina," Prog. Retin. Eye Res., Aug. 2018, doi: 10.1016/j.preteyeres.2018.07.004.

[2] P. Mitchell, G. Liew, B. Gopinath, and T. Y. Wong, "Age-related macular degeneration," *The Lancet*, vol. 392, no. 10153, pp. 1147–1159, Sep. 2018, doi: 10.1016/S0140-6736(18)31550-2.

[3] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[4] J. Lee, P. Sattigeri, and G. Wornell, "Learning New Tricks From Old Dogs: Multi-Source Transfer Learning From Pre-Trained Networks," p. 11.

[5] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *ArXiv14111792 Cs*, Nov. 2014, Accessed: Jan. 07, 2020. [Online]. Available: http://arxiv.org/abs/1411.1792.

[6] K. He, R. Girshick, and P. Dollár, "Rethinking ImageNet Pre-training," *ArXiv181108883 Cs*, Nov. 2018, Accessed: Jan. 07, 2020. [Online]. Available: http://arxiv.org/abs/1811.08883.

[7] M. Raghu, J. Kleinberg, C. Zhang, and S. Bengio, "Transfusion: Understanding Transfer Learning for Medical Imaging," p. 11.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *ArXiv150504597 Cs*, May 2015, [Online]. Available: http://arxiv.org/abs/1505.04597.

[9] "RETOUCH - MICCAI 2017 Workshop," Dec. 13, 2017. https://retouch.grand-challenge.org/workshop/ (accessed Dec. 13, 2017).

[10] A. G. Roy *et al.*, "ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomed. Opt. Express*, vol. 8, no. 8, p. 3627, Aug. 2017, doi: 10.1364/BOE.8.003627.

[11] D. Lu, M. Heisler, S. Lee, G. Ding, M. V. Sarunic, and M. F. Beg, "Retinal Fluid Segmentation and Detection in Optical Coherence Tomography Images using Fully Convolutional Neural Network," *ArXiv171004778 Cs*, Oct. 2017, [Online]. Available: http://arxiv.org/abs/1710.04778.

[12] S. Apostolopoulos, S. De Zanet, C. Ciller, S. Wolf, and R. Sznitman, "Pathological OCT Retinal Layer Segmentation Using Branch Residual U-Shape Networks," in Medical Image Computing and Computer-Assisted Intervention − MICCAI 2017, vol. 10435, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham: Springer International Publishing, 2017, pp. 294–301.

[13] H. Bogunovic *et al.*, "RETOUCH -The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge," *IEEE Trans. Med. Imaging*, pp. 1–1, 2019, doi: 10.1109/TMI.2019.2901398.

[14] "Google Colaboratory." https://colab.research.google.com (accessed May 17, 2020).

Smartphone Based Range of Motion Measurement in Physiotherapy

M. Njirjak*, E. Otović*, M. Budimir**, H. Vlahović***, M. Tomić*, V. Marijančić***
* University of Rijeka, Faculty of Engineering/Department of Computer Engineering, Rijeka, Croatia
**Thalassotherapia Opatija, Physical medicine and rheumatology, Opatija, Croatia
*** University of Rijeka, Faculty of Health Studies/Department of Physiotherapy, Rijeka, Croatia
marko.njirjak@riteh.hr, erik.otovic@riteh.hr, marko.budimir2@gmail.com, hrvoje.vlahovic@uniri.hr, mladen.tomic@riteh.hr, verner.marijancic@uniri.hr

***Abstract* – Range of motion measurement is often performed in physiotherapy. It is important for both initial patient's joint health status evaluation, and later on for tracking progress of a rehabilitation therapy. Standard goniometers are often cumbersome to use and many therapists resort to approximative by eye angle assessment. In this paper, we investigate a modern smartphone application goniometer, based on an inertial measurement unit (IMU). We analyze IMU properties and their influence on angle measurement accuracy. The application is tested in a controlled environment and later on in real-life situations. The same set of measurements is performed by experienced physiotherapist using both the smartphone application and traditional goniometers. Shoulder abduction and elbow pronation and supination were examined. The results obtained clearly show that the smartphone application is comparable in accuracy to traditional ways of measurement. There was no statistically significant difference between the measurements made with traditional instruments and the mobile application. It was shown that for some ROM measurements it was efficient and easy to use and can be considered to be a viable and a better alternative to traditional goniometers.**

Keywords – physiotherapy, smartphone, range of motion, shoulder, elbow, IMU

I. INTRODUCTION

Physiotherapists use holistic approach in conducting physiotherapy assessments. They evaluate measurable parts of human bodies to detect dysfunctions or anomalies. Common procedure in physiotherapy assessments is the range of motion (ROM) measurement. The measurement quantifies mobility of a part of a body involved in the formation of joints. The ROM is expressed in angular degrees and the measurement is made using either a two-axis or gravity goniometers [1].

A two-axis goniometer is a device made of plastic or metal. It features two hinged "arms", of which one is held stationary during a measurement while the other one is unrestricted and follows a body part motion. The arms are positioned on specific points of a body with the center of a goniometer being aligned to the joint of interest. Hash marks on the hinge allow a therapist to precisely measure ROM in degrees. The other type of goniometer is a gravity goniometer. It is, usually, a plastic device with angular degree scale and fluid in a half-filled round tube, which then points to a correct ROM angle readout. Both, types of goniometers can also feature a digital readout of a measured angle. A ROM measurement can also be performed indirectly, using X-ray images, which is a more accurate method, but rarely available [2]. To perform a ROM measurement, a two-axis goniometer needs to be aligned with both parts of a body, adjoining the joint of interest, while gravity goniometers are placed only on the moving part of a body. Due to a human body configuration, not all ROM measurements can be made using either device. Measurement accuracy is highly dependent on both, a therapist's experience and on a patient's ability to closely follow therapist's instructions. For that reason, successive measurements, generally, lead to a slight readout variation. Often, 3 measurements are taken and their average considered to be the actual ROM. Depending on the joint of interest, measurements can be cumbersome and require special positioning of a patient. A large number of different goniometers is a clear indication of application insufficiencies of their construction, which mostly arise from inability of achieving a firm fixed positioning of the devices on a human body. Today, a lot of effort is being put in digital goniometer development, with the goal of making the measuring procedures faster, more accurate and more convenient. Out of many approaches, smartphone applications are becoming increasingly popular. Due to advances of smartphone hardware components, most of the phones currently shipped are already equipped with inertial measurement units (IMU), which can enable ROM angle measurement. Considering the general availability of the smartphones, it is possible to perform ROM measurements at any place or time, without requiring specialized instruments – goniometers. However, it is not easy for such applications to match the accuracy of a traditional measuring devices and procedures.

A number of authors examined accuracy of some of the available smartphone applications for specific joint's ROM, with varying degree of success. Ulluci et el. [3] focused on the upper cervical range of motion. Their measuring setup required a camera mount to achieve satisfying phone alignment and they found the smartphone to be reliable and accurate enough in that setup. Vauclair et al. [4] examined elbow joint measurements. They found the application to be acceptable for pronation and extension. However, elbow flexion was overestimated by a mean of 6.4°, while for elbow supination, the angle was overestimated by a mean of 5.9°. When inspecting upper limb range of motion measurement, Lin et al. [5] found the standard error of measurement of the smartphone application to be 8°-10°,

which is too large and is outside the clinically acceptable range. Santos et al. [6] focused on wrist pronation and supination. Generally, they found the application to be accurate enough, however, it was conditional to precise identification of bony landmarks, needed for proper phone alignment. Mehta et al. [7] examined a smartphone application accuracy on a knee joint. They found that the application had superior reliability indices and smaller measurement error in assessing knee ROM over a universal goniometer in patients with knee osteoarthritis or patients who had a total knee replacement surgery. On the other hand, the application overestimated knee flexion by a mean of 5°. The authors conclude that precise positioning is crucial for measurement accuracy. Hancock et al. [8] compared the range of knee motion measurements using a classic goniometer, digital inclinometer, and a smartphone application that uses an accelerometer. The worst results were obtained using a smartphone application. Finally, Keogh et al. [9] performed a systematic review of literature which confirmed general findings of above authors. They conclude that there is relatively strong evidence regarding validity of smartphones and applications on joint ROM assessment, which tend to be observed across multiple joints, joint actions, populations, smartphones and applications. However, they also note that applications show poorer reliability in axial rotations, compared to flexions or extensions. Generally, there are three common reasons for any inaccuracies:

- axis mis-alignment that occurs when a sensor's coordinate axes are not aligned with anatomically meaningful axes,
- secondary movements that naturally occur in body parts and are often unavoidable,
- inability to keep proper smartphone alignment and positioning during a joint action.

In this paper, we examine an Android application GoniometerPro (available on Google Play Store), proposed by authors Njirjak, Otovic and Tomic, that tries to alleviate the three problems that lead to inaccuracies. We first test the application in a lab environment and then in a clinical practice. Shoulder abduction and elbow pronation and supination are considered.

II. Inertial Measurement Units

Smartphones use inertial measurement units (IMU) to detect their orientation (and sometimes, position). IMU is an electronic device that typically contains three microelectromechanical system (MEMS) sensors. These are accelerometer, gyroscope and a magnetometer. IMU devices containing all three components are said to have 9 degrees of freedom, because they can detect acceleration in three directions, rotation about all three axes, and magnetic field strength in all three directions. Many applications have emerged that utilize IMUs. They found great use in robotics, drone control, entertaining, augmented and virtual reality, etc.

A. Accelerometer

Accelerometer is an IMU component that senses acceleration force [18]. It is motion dependent and

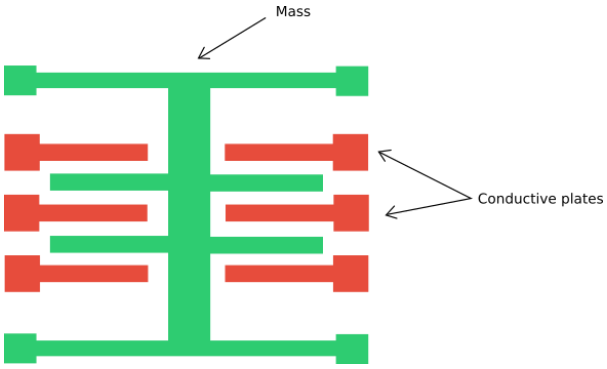


Figure 1. Internal structure of MEMS accelerometer

measures the forces along its three axes, x, y and z. When stationary, accelerometers display the direction of Earth's gravity vector and can be used to determine roll and pitch components of attitude. Accelerometers measure absolute state – resultant force on the sensor, not the relative change. MEMS accelerometers work by measuring capacitance between fixed plates. Between the plates, there is a mass attached to springs. Depending on a sensor movement, the mass shifts and the capacitance between the plates changes (Fig. 1). The circuit inside MEMS accelerometer uses that information to calculate and output acceleration data, as:

$$V_x = V_0 \frac{C_2 - C_1}{C_2 + C_1}, \tag{1}$$

where C_1 and C_2 are capacitances between the two different plates, V_0 is a voltage generated by an oscillator and V_x is a voltage that is generated due to a mass shift.

Prevailing accelerometer errors are bias and scale error. Because of the imperfections in the IMU, A-D conversion and other factors, accelerometers experience bias so the calculated values have offset from real values [19]. While bias can be compensated for, scale factor cannot be corrected. True acceleration is given by:

$$a = \nabla * (A - b), \tag{2}$$

where ∇ is axis scale factor, A is the acceleration measured by the sensor and b is bias. Accelerometer errors are shown in Fig. 2.

B. Gyroscope

Gyroscopes are sensors that measure angular velocity. Like accelerometers, they also suffer from bias and scale error. When determining attitude and heading using gyroscope, angle velocity needs to be integrated with respect to time to obtain angle estimations. This results in drift after a certain time period. The scale of the drift largely depends on gyroscope properties. High-precision gyroscopes tend to drift less.

Similar to MEMS accelerometers, MEMS gyroscopes measure capacitance between plates when a displacement of mass occurs in a certain plane [16, 17]. MEMS gyroscopes are based on a Coriolis effect. Mass inside the gyroscope vibrates in one plane while no external force is applied. When the object rotates, the displacement of mass occurs in another plane, between capacitive plates, and

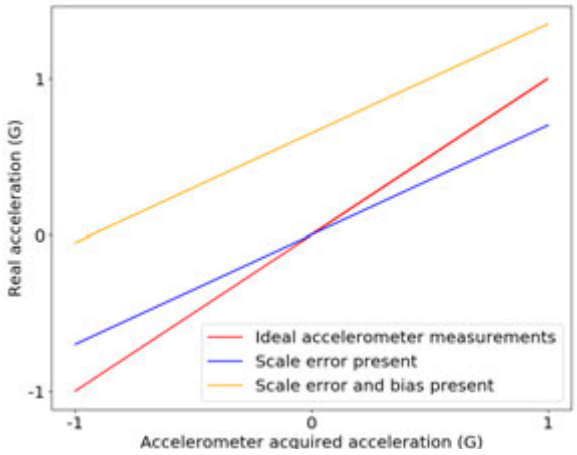


Figure 2. Accelerometer errors

angular rate can be calculated. A mass m moving with velocity v, in a frame of reference rotating at angular velocity ω, undergoes a force:

$$\vec{F}_c = -2m(\vec{\omega} \times \vec{v}) \quad (3)$$

C. Magnetometer

Magnetometer is a sensor that senses Earth's magnetic field and can be used to calculate geographic north direction. Magnetometers are absolute sensors and are highly sensitive to other magnetic materials in their environment. A so-called soft-iron and hard-iron effects are present. Hard-iron distortions are caused by materials that exhibit additive field to the Earth's magnetic field. Hard-iron distortions cause a permanent bias in magnetometer measurements. Soft-iron distortions are caused by materials that skew the Earth's magnetic field, but don't necessarily generate one. Usually ferromagnetic materials. While hard-iron distortions are constant and independent of sensor's orientation, soft-iron distortions are not and require more sophisticated procedure to be compensated for. Magnetometer measurement distortions by hard-iron and soft-iron error are depicted in Fig. 3. The measurements were taken during a magnetometer rotation in a 2D plane, around z-axis.

To compensate for errors, sensor calibration is needed [20], however, if the magnetic configuration of a space changes (or a magnetometer is taken to another location), a recalibration is needed. This is a tedious process and represents a challenge in indoor applications. Outdoor spaces are less affected by such problems and one calibration is usually enough.

The majority of MEMS magnetometers utilize Hall effect to detect and measure Earth's magnetic field. The Hall effect is a phenomenon in which the electrons travelling through electric conductor are deflected due to an external magnetic field. This causes electric potential difference in the conductor which can be measured, and magnetic field strength and direction can be extracted (Fig. 4).

D. Additional Errors

Prevailing errors have been explained for each sensor, however, they are not the only sources of error. All

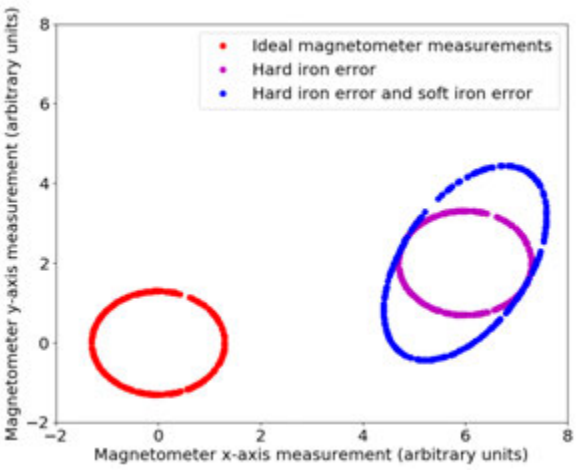


Figure 3. Magnetometer errors

mentioned sensors also suffer from temperature drift. Temperature drift manifests itself as an inaccuracy in measurements due to an increase or decrease in ambient temperatures. There are also other error sources that contribute to the overall IMU accuracy and ought to be compensated for [10, 11].

III. Sensor Fusion

Fusion algorithms are employed to combine measurements of all three sensors into a single rotation quaternion. Accelerometers provide direction of gravity and can be used to determine roll and pitch components. Magnetometers, on the other hand, are used to calculate Earth's north vector and, consequently, determine heading. Heading is attitude's yaw component. Gyroscopes supplement former sensors by providing more accurate attitude and heading estimates during abrupt rotations. Three dominant algorithms that are used today are: Kalman filter, Madgwick and Mahony.

Kalman filter [12] was devised by Rudolf E. Kalman. It became a de facto standard in system state estimation. It is a linear quadratic estimation algorithm that estimates joint probability distribution over variables for each timeframe. By doing so, it yields more accurate results. Due to a fact that the Kalman filter is used for linear filtering, modifications of the algorithm have been developed, most popular being extended Kalman filter and unscented Kalman filter

Mahony [13] algorithm is an enhanced complementary filter. Its main advantage over complex filters, such as Kalman, is much shorter computational time and lower

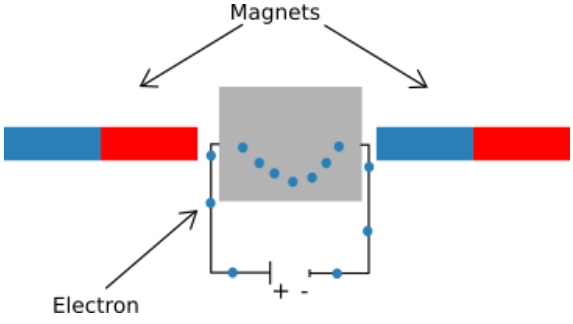


Figure 4. Hall effect

complexity. That is why it is an ideal algorithm to be used in small embedded systems, which usually have no memory or CPU power to spare.

As a part of his Ph.D. research in 2009., Sebastian Madgwick developed a novel approach to IMU data filtering. The Madgwick filter [14] is based on a Newton optimization using an analytic formulation of a gradient that is derived from a quaternion representation of motions. It is similar to Mahony algorithm. It also has short computational time, low complexity and can be used to process data at a much higher rate than Kalman filter, which can be useful in applications such as human motion tracking.

Ludwing and Burnham conducted a study in which they compared the three sensor fusion algorithms, based on a quadcopter flight data [15]. The results showed that the three approaches achieved relatively similar results, while Mahony had the fastest execution time.

IV. Angle Estimation

In the proposed application, quaternion number system is used for angle estimation. After a smartphone is placed in a required position and measurement is started, initial quaternion is obtained from the Android system. Sensors are then polled periodically in order to calculate the delta quaternions:

$$\Delta Q_{gyro} = \left(\hat{v}_{gyro}, |\vec{v}_{gyro}| * \Delta t\right), \quad (4)$$

$$\Delta Q_{accel} = \left(\hat{v}_{\perp accel}, \angle\left(\vec{v}_{accel,t-1}, \vec{v}_{accel,t}\right)\right), \quad (5)$$

where $\hat{v}_{\text{gyro}}$ represents axis of rotation based on gyroscope data and $\hat{v}_{\perp\text{accel}}$ represents axis of rotation based on accelerometer data:

$$\hat{v}_{gyro} = \frac{\vec{v}_{gyro}}{|\vec{v}_{gyro}|}, \quad (6)$$

$$\hat{v}_{\perp accel} = \frac{\vec{v}_{accel,t-1} \times \vec{v}_{accel,t}}{|\vec{v}_{accel,t-1} \times \vec{v}_{accel,t}|}. \quad (7)$$

Since accelerometers measure acceleration forces, their measurements combine gravity acceleration with external motion. The component of interest for angle estimation is gravity. Although it is a reasonable assumption that there should be no sudden and sharp movements during the measuring period, to improve gravity isolation, an exponential moving average low pass filter is applied to accelerometer measurements:

$$A_{accel,\gamma,t} = \alpha * A_{accel,\gamma,t-1} + (1-\alpha) * A_{accel,\gamma,t}, \quad (8)$$

where γ symbolizes three accelerometer axes – *x*, *y* and *z*, A represents accelerometer measurement and α is a constant defining exponential moving average decay. Equation (8) is applied to each axis individually. Finally, by multiplying the current quaternion with a delta quaternion [21], a new state is deduced:

$$Q_{c,t} = Q_{c,t-1} * \Delta Q \quad (9)$$



Figure 5. GoniometerPro measurement screen

V. GoniometerPro

The application we propose is tailored to physiotherapists. It increases accuracy, when compared to common Android inclinometer or ROM applications, but, even more important, the usability. The usability is often disregarded, which can lead to an accurate application, with no clinical significance or usage. The improvements are achieved by multiple features.

GoniometerPro allows for measurement in any plane, instead of exclusively the vertical plane, that is mostly used in current applications. This is significant as it makes it much easier for physiotherapists to make measurements, as exact alignment of a smartphone with anatomically meaningful axis is no longer necessary. This also speeds up positioning and fixating (if necessary) of a phone to body parts and eliminates inaccuracies caused by missing exact vertical alignment. Another benefit is the ability to perform measurements on joint actions that cannot be made in vertical plane. It also allows for more comfortable postures for patients during measurements as well as measurements on immobile patients.

To further improve measurement accuracy, a correction algorithm is built into the application. In clinical practice, it is not always possible to achieve strict single plane movement of body parts. The correction algorithm tries to detect and exclude of such secondary movements.

To increase usability, except for the real-time angle display during measurement and a manual measurement stop, a therapist can also choose the auto-stop option, which stops the measurement after there is no angle change for a configurable amount of time. Also, during a measurement, a maximum angle value is continuously tracked and displayed, Fig. 5. This enables a therapist to fully focus on a patient and on leading him in performing a joint action correctly, which, in effect, improves reliability of measurements, as well as a user experience of therapists.



Figure 6. Testing platform

Before commencing a measurement, a joint and joint action can be selected. That way, a therapist can get a prompt warning from the application if a measured angle falls outside the normal range of motion for chosen joint action.

The print function allows for simple report printing. The report is primarily intended for patients to encourage them in making additional efforts in their recovery, as well as tracking progress, however it can also be used for therapists' documentation. Patient data and history of measurements are stored in a local database.

VI. Testing Methodology

For testing, a mid-range smartphone, Samsung J6+, was used. Initial tests were performed in a lab environment on a specialized testing platform, Fig. 6. In the first set of tests, the smartphone was exactly aligned with one of the three planes: *x*, *y* or *z*. Testing platform was rotated by multiple angles in all 3 planes (separately) and the exact rotation angle measured with a calibrated protractor.

In the second set of tests, the smartphone was positioned in arbitrary planes in a 3D space, and the platform rotated the same way as in the first set.

A third set of tests were performed to investigate influence of total measurement time on angle estimation accuracy. Total measurement time was increased to 5 minutes, which is longer than any clinical practice measurement time.

In a clinical trial, the range of motion was measured on 10 healthy subjects between the ages of 18-20, and 1 subject with the plexus brachialis paresis condition. The range of motion was measured using the Neutral zero method. For each subject, shoulder abduction ROM was measured with two-axis goniometer, gravity goniometer and the proposed mobile application, while for elbow pronation and supination, only two-axis goniometer and the mobile application were used. Three measurements were taken for each joint action and mean value was calculated and recorded. All measurements were performed by an experienced physical therapist.

VII. Results and Discussion

The laboratory testing showed very high accuracy of the proposed application. The maximum angle estimation error was 0.5°, which is much better than needed for ROM measurements. It was also shown that long measurement times did not negatively impact estimation accuracy.

In a clinical trial, we wanted to determine if there was a statistically significant difference in measurement results when using standard measuring instruments or the mobile application.

The mean ROM value for shoulder abduction was 92.3° for a two-axis goniometer, 91.6° for a gravity goniometer and 91° for the application. The obtained results show that all results are within the normal range of motion in the shoulder joint (90-95°) and that there is no statistically significant difference between the measurements taken with different measuring instruments, p = .751 (Fig. 7). Mean ROM value of pronation movement in the elbow joint, measured with the classical goniometer was 89.6°, and with the application it was 91.3° (Fig. 8). Regarding the supination movement in the elbow, mean value with a two-axis goniometer was 81.6°, while the application showed 83° (Fig. 9). The difference is also not statistically significant – for elbow pronation p = .109, while for elbow supination p = .671.

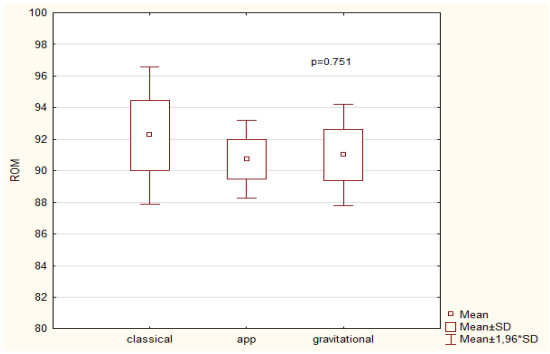


Figure 7. Friedman ANOVA test results - shoulder abduction

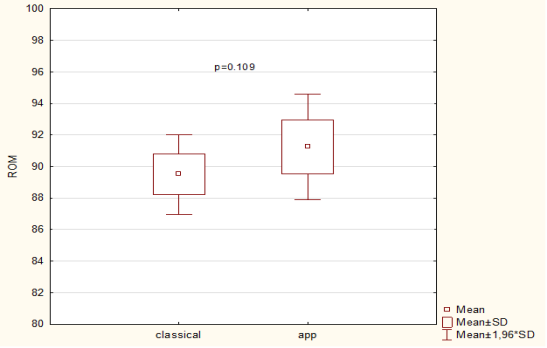


Figure 8. Wilcoxon Matched Pair test results for elbow pronation

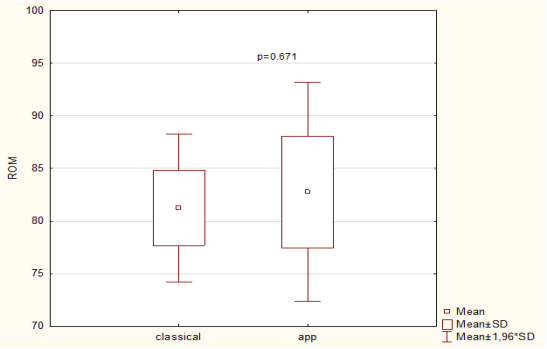


Figure 9. Wilcoxon Matched Pair test results for elbow supination

VIII. CONCLUSION

Joint range of motion measurement (ROM) is a common procedure in physiotherapy. However, many therapists often avoid performing it because traditional goniometers can be cumbersome to use. As a possible alternative, smartphone applications are often considered. Their common shortcomings are restriction of measurement plane to exclusively vertical plane and requirement for a near-perfect alignment of a phone axes to anatomical axes. The conditions limit the joint actions that can be measured, and failure to set near-perfect alignment lead to inaccurate measurements. They also negatively impact therapists' user experience during the procedure.

The proposed application, GoniometerPro, tries to improve the shortcomings by removing or relaxing the two conditions. Instead of only measuring in the vertical plane, the application can perform measurements in any arbitrary plane. This has several benefits. It enables the application to be used to measure any joint ROM. It enables a therapist to measure ROM by setting a patient in a more relaxed and convenient posture. It eliminates inaccuracies caused by imperfect alignments. Finally, it improves therapists' user experience as less time is needed to achieve proper measuring setup. The application also has several other features, such as auto-stop and maximum angle tracking that allow a therapist to fully focus on a patient.

The test results showed excellent accuracy of the application in a lab environment, with the maximum error being only 0.5°, regardless of a measuring plane. The accuracy was also confirmed in clinical testing, in which there was no statistically significant differences between the application and traditional goniometer measurements. Also, the flexibility of smartphone alignment was very appreciated by a therapist. As there are many possible ways to perform measurements, a therapist is free to choose and modify the procedure (placement, fixing, holding, etc.) to best match his preference and the range of motion being examined.

The application proved to be very accurate and for many joints we see it as a better alternative that can fully replace traditional goniometers. It was quick and easy to use and with little practice can be utilized for most joint actions. However, due to a smartphone size, the advantages could be less obvious for some of the smaller body joints.

REFERENCES

[1] Zulle, M., Fužinac-Smojver, F., Lulić Drenjak, J. Mjerenje opsega pokreta i antopometrijsko mjerenje, Medicinski fakultet, Rijeka, 2012., ISBN: 978-953-6384-95-2

[2] Clarkson, M. H. Valutazione cinesiologica Esame della mobilita articolare e della forza muscolare, Edi-ermes, Milano, 2006., ISBN: 978-887-0512-42-7

[3] Ullucci, Paul & Tudini, Frank & Moran, Matthew. (2018). Reliability of Smartphone Inclinometry to Measure Upper Cervical Range of Motion. Journal of Sport Rehabilitation. 28. 1-12. 10.1123/jsr.2018-0048.

[4] Vauclair, F., Aljurayyan, A., Abduljabbar, F.H. et al. The smartphone inclinometer: A new tool to determine elbow range of motion?. Eur J Orthop Surg Traumatol 28, 415–421 (2018). https://doi.org/10.1007/s00590-017-2058-x

[5] Lin, Natalie & Hayward, Kathryn & D'Cruz, Kate & Thompson, Eloise & Li, Xia & Lannin, Natasha. (2019). Validity and reliability of a smartphone inclinometer app for measuring passive upper limb range of motion in a stroke population*. Disability and Rehabilitation. 1-7. 10.1080/09638288.2019.1585972.

[6] Santos, Claude & Pauchard, Nicolas & Guilloteau, A.. (2017). Reliability assessment of measuring active wrist pronation and supination range of motion with a smartphone. Hand Surgery and Rehabilitation. 36. 10.1016/j.hansur.2017.06.005.

[7] Mehta, S.P., Barker, K., Bowman, B.S., Galloway, H., Oliashirazi, N., & Oliashirazi, A. (2017). Reliability, Concurrent Validity, and Minimal Detectable Change for iPhone Goniometer App in Assessing Knee Range of Motion. Journal of Knee Surgery, 30, 577–584.

[8] Hancock, G. E., Hepworth, T., & Wembridge, K. (2018). Accuracy and reliability of knee goniometry methods. Journal of experimental orthopaedics, 5(1), 46. doi:10.1186/s40634-018-0161-5

[9] Keogh JWL, Cox A, Anderson S, et al. Reliability and validity of clinically accessible smartphone applications to measure joint range of motion: A systematic review. PLoS One. 2019;14(5):e0215806. Published 2019 May 8. doi:10.1371/journal.pone.0215806

[10] Ma, L., Chen, W., Li, B., You, Z., & Chen, Z. (2014). Fast field calibration of MIMU based on the Powell algorithm. Sensors (Basel, Switzerland), 14(9), 16062–16081.

[11] Renaudin, Valérie & Afzal, Muhammad & Lachapelle, Gérard. (2010). Complete triaxis magnetometer calibration in the magnetic domain. Hindawi Publishing Corporation J. Sens.. 1-10.

[12] Kalman, R. E. (March 1, 1960). "A New Approach to Linear Filtering and Prediction Problems." ASME. J. Basic Eng. March 1960; 82(1): 35–45.

[13] R. Mahony, Tarek Hamel, Jean-Michel Pflimlin. Nonlinear Complementary Filters on the SpecialOrthogonal Group. IEEE Transactions on Automatic Control, Institute of Electrical and ElectronicsEngineers, 2008, 53 (5), pp.1203-1217.

[14] Madgwick, S.O. (2010). An efficient orientation filter for inertial and inertial / magnetic sensor arrays.

[15] S. A. Ludwig and K. D. Burnham, "Comparison of Euler Estimate using Extended Kalman Filter, Madgwick and Mahony on Quadcopter Flight Data," 2018 International Conference on Unmanned Aircraft Systems (ICUAS), Dallas, TX, 2018, pp. 1236-1241.

[16] Passaro, V., Cuccovillo, A., Vaiani, L., Carlo, M., & Campanella, C. E. (2017). Gyroscope Technology and Applications: A Review in the Industrial Perspective. Sensors (Basel, Switzerland), 17(10), 2284.

[17] Xie, Huikai & Fedder, Gary. (2003). Integrated Microelectromechanical Gyroscopes. Journal of Aerospace Engineering. 16. 10.1061/(ASCE)0893-1321(2003)16:2(65)

[18] Mohammed, Z., Elfadel, I., & Rasras, M. (2018). Monolithic Multi Degree of Freedom (MDoF) Capacitive MEMS Accelerometers. Micromachines, 9(11), 602.

[19] Albarbar, A. & BadriÂ¹, A & Sinha, Jyoti & Starr, Andrew. (2009). Performance evaluation of MEMS accelerometers. Measurement. 42. 10.1016/j.measurement.2008.12.002.

[20] A. Hosseinzadeh, A. Khayatian, P. Karimagahee, O. Daneshmandi and B. Raeesi, "Three Axis Fluxgate Magnetometer Sensor Calibration," 2019 27th Iranian Conference on Electrical Engineering (ICEE), Yazd, Iran, 2019, pp. 431-436.

[21] Jia, Yan-Bin, Quaternion and Rotation, Com S 477/577 Notes Sep 5, 2017.

Comparative Analysis of Network Embeddings for Functional Annotation in Protein Interaction Networks

I. Ivanoska, E. Petreska, B. R. Stojkoska, S. Kalajdziski and K. Trivodaliev
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, Macedonia
ilinka.ivanoska@finki.ukim.mk, eva.petreska@students.finki.ukim.mk, biljana.stojkoska@finki.ukim.mk, slobodan.kalajdziski@finki.ukim.mk, kire.trivodaliev@finki.ukim.mk

***Abstract* - One of the major problems in bioinformatics is the computational prediction of functions for the large number of sequenced proteins which will facilitate the expensive and long process of wet lab verification. Protein-protein interaction networks (PINs) are considered as one of the richest sources of information for solving this problem. PINs can be represented as graphs, where the nodes are the proteins with their functions as node labels and the edges are their physical interactions. In this paper embedding vectors are created to represent the nodes of the graph which are later used as the input data for a classification model. This is a graph node classification problem and because of the property of proteins to have multiple functions, it is also a multi-label problem. The classification model used is linear SVM, while the embeddings are built with 4 algorithms, HOPE, SDNE, GF and node2vec and then a comparative analysis is done on the results. Hamming loss is used as an evaluation metrics, because of the multi-label problem. Based on the comparative evaluation, recommendation for using a specific network embedding in specific scenarios is given.**

Keywords – graph embeddings; data mining; protein-protein interaction network

I. INTRODUCTION

All living organisms are consisted of cells that use proteins, the most important biochemical molecules, to perform activities required for that organism. Proteins play a key role in almost all biological activities. This has led to the emergence of a revolutionary number of protein sequencing techniques in recent decades. Although knowledge of protein sequences is useful in many fields, such as phylogenetics and evolutionary biology, an understanding of protein functions is still needed. It is useful for discovering how biological processes work at the molecular level, developing personalized medicine, more effective therapeutic treatments, understanding the emergence of diseases, understanding biological entities as complete systems, etc.

Discovering protein functions experimentally is done through "in vitro" or "in vivo" experiments, which are very expensive and time-consuming. As the number of proteins without known functions becomes too large to be classified in this way, there is need for computational methods for protein functions classification. While more methods of this type are emerging, the number of sequenced proteins without functional annotations is expected to continue to grow, making computational function prediction one of the fundamental problems in bioinformatics today [1].

Ontologies are often used to describe protein functions, such as the Gene Ontology (GO) project, which is currently the largest bioinformatics initiative designed to unify genes and genes products representations of all species. GO divides functions into three different ontologies: cellular component (CC), molecular function (MF), and biological process (BP) [2]. Any function is in exactly one of these three ontologies, and any protein can have multiple functions. Functions are structured as a hierarchy where the root of all three hierarchies is a function that denotes the ontology itself.

There are several approaches to solving the problem of determining protein functions, and they can use protein structure, protein interactions with other proteins, genetic interactions, evolutionary interactions, or a combination of them. There are several major challenges to methods for predicting protein functions. One of them is the complex relationship between protein structure, sequence, and function. Another challenge is the size and complexity of the output space required for any classification method. Additionally, one has to pay attention to the structure in which these functions are located, the hierarchy to which they belong, leading to emergence of proteins with different number of functions depending on the level of abstraction. This means that predicting protein functions is a multi-label, multi-class problem. A final challenge is that proteins are not isolated from other proteins and determining the function of one protein requires knowledge of the functions of other proteins with which it interacts [3].

Methods for protein function prediction using Protein Interaction Network (PIN) data type assume that proteins with similar functions are topologically close in the network, which represents interactions between proteins. This means that it is necessary to perform a graph/network analysis to solve this problem. Main graph analysis tasks are divided into four categories: node classification, edge prediction, clustering, and visualization. Protein functions appear as node labels in protein interaction networks; therefore, the classification of these functions can be

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss.Cyril and Methodius University"

identified as node classification. Node classification attempts to determine node labels based on other labeled nodes and network topology [4,5].

As graphs are used to represent different types of information, their analysis has attracted increasing attention in recent times, not only in the field of biology, but also in social sciences, linguistics, etc. Typically, methods defined to solve graph-based problems, either work with the original graph adjacency list or work with derived vector graph representations. Vector graph representations (embeddings) are becoming increasingly used because they eliminate the need for complex classification models to be applied directly to the graph. Instead, these network embeddings are used as input to a classification model [6,7].

This paper aims at facilitating future research in computational protein function prediction by comparative analysis of the performance of different PIN embeddings in a baseline classification of the protein function. The network embeddings are obtained with four different algorithms on a PIN derived from HIPPIE, the Human Integrated Protein-Protein Interaction Reference Database [12]. The produced embeddings are used as input to a multi-label classifier that uses the one-vs-rest method, i.e. it generates SVM classifiers for each GO label and makes a binary classification for it. Experiments are performed on different embedding dimensions considering GO labels (with and without including GO hierarchy) that are frequent, i.e. well represented.

The rest of the paper is structured as follows: in Section II the data and algorithms used in the research is presented, as well as the data preprocessing and analysis that leads to the specific experimental setups. Section III presents the different experiments performed, the obtained results and discussion about the potential recommendations. Finally, Section IV concludes the paper.

II. MATHERIALS AND METHODS

In this section the data and algorithms used in the research is presented, as well as the data preprocessing and analysis that leads to the specific experimental setups.

A. Dataset

1) Gene Ontology (GO)

Due to the incredible complexity of biological systems and vast databases, biomedical research and development is increasingly dependent on knowledge organized in a unified form that can be analyzed. That is a problem that projects like the Gene Ontology (GO) solves. GO is currently the golden standard for organized knowledge about gene functions and gene products. The GO database is a comprehensive model for biological systems consisting of two main components: 1) Gene ontology (GO), which is the logical structure of biological functions and their relationships to one another, in the form of a directed acyclic graph; 2) the corpus of GO annotations, which are annotations derived from evidence linking a particular gene product to a particular ontological term.

Functional annotations can be found in one of the three sub-ontologies, depending on what they describe: cellular component (CC), molecular function (MF), or biological process (BP). Each of these three sub-ontologies is represented as a directed acyclic graph, with one root representing the ontology and the relationships between functions. The nodes in that graph are GO function annotations, and the edges are the relationships between functions.

2) Protein interaction network

The protein interaction network used in this paper is taken from HIPPIE (the Human Integrated Protein Protein Interaction Reference) [12]. HIPPIE integrates data from 10 databases and 11 studies, such as BioGRID, DIP, MINT, HPRD, IntAct databases, etc. This network has a large number of interactions that are unique that cannot be found in any other database. An important part of this network is the assessment of confidentiality that exists between each of the interactions involved. Evaluation of this is made using the weighted sum of the number of studies in which that interaction was detected, the number and quality of the experimental technique used to measure the interaction, and the number of organisms in which the interaction was reproduced, other than human organisms. The scheme to determine the evaluation grade is optimized by a group of experts and a computer algorithm.

B. Graph embeddings methods

Graph embedding (i.e. network embedding) is a representation of a graph in a vector space, where relevant graph features are preserved. A graph is consisted of nodes and edges, therefore, the biggest challenge in learning a graph embedding is the choice of metrics, properties and features to be preserved in the vector representation, since vectors are more easily handled in several domains of machine learning. Graph embeddings methods as vector representations of networks are developed as part of dimensionality reduction techniques. A lot of graph embeddings methods have been proposed and are categorized into three broad categories: (1) matrix factorization based, (2) random walk based, and (3) neural networks (or deep learning) based. This research uses Graph Factorization (GF), High-Order Proximity Preserved Embedding (HOPE), Structural Deep Network Embedding (SDNE) and node2vec embedding methods for protein interaction network vector representation incorporated in a web tool for graph embeddings previously developed in the group [13].

1) Graph Factorization (GF)

Graph Factorization [8] is one of the matrix-factorization based algorithms. The possible matrices to use are the graph adjacency matrix, Laplacian matrix, Katz similarity matrix, etc. GF allows to identify unnoticed factors associated with nodes that can be used for further graph conclusions. The algorithm uses regularized Gaussian matrix factorization on the graph adjacency matrix and minimizes the loss function. This is done on observed edges as opposed to all possible edges, which is in the interest of scalability, but raises a noisy solution possibility.

2) HOPE

High-Order Proximity Preserved Embedding (HOPE) [9] is another matrix factorization-based algorithm. It maintains high-order closeness by minimizing a similarity matrix. Several similarity metrics, such as the Katz Index, Rooted Page Rank, Common Neighbors, and Adamic-Adar score can be used. HOPE uses Singular Value Decomposition (SVD) to obtain vector representations efficiently.

3) SDNE

Structural Deep Network Embedding (SDNE) [10] is a method that belongs to the group of deep learning based methods. Deep autoencoders have been used to reduce dimensionality, leading to SDNE's use of this feature to generate a model that can maintain graph non-linearity. SDNE, uses deep autoencoders to maintain first and second-order network proximities. This is done by jointly optimizing both proximities. Nonlinear functions are used to obtain the vector representation and the model has a supervised and unsupervised part. The unsupervised part tries to find a vector for a node that can reconstruct its neighborhood, while the supervised part is based on Laplacian Eigenmaps (LE) algorithm which punishes when similar nodes are mapped away from each other in the vector space.

4) Node2vec

Node2vec [11] is one of the graph embeddings algorithms that uses random walks. These algorithms are useful when the graph is too large to be analysed fully or when it is possible to observe only part of the graph. Node2vec maintains high-order nodes proximity by maximizing the probability of successive nodes with fixed length random walks. It uses biased-random walks that provide a balance between breath-first search (BFS) and depth-first search (DFS) and thus produce high quality vector representations.

C. Dataset preprocessing and analysis

In this paper functional annotations were analyzed in terms of their hierarchy. Using the GO relational graph, that is, the three graphs for each of the three sub-ontologies, CC, MF and BP, three hierarchical tree-like structures with multiple inheritance possibilities have been constructed. Only "is a" or direct sub-type relationships are used. Functions up to level 5 depth of these trees are analyzed, where the root is at level 0. In lower levels, functions are more specific and occur in a lower number of proteins. In GO a transitive closure can be made i.e. to each protein having a function, all ancestors of that function can be added. In addition, duplicate functions are removed and only those with the highest evaluation grade from the HIPPIE database are retained (higher than 0.6). An undirected weighted graph is constructed using HIPPIE, from which the largest connected component is taken out for further analysis. Proteins that do not have functions taken from GO are filtered and GO functions that do not occur in any protein are also removed. A multi-label classification problem is created in this way. Since the number of all functions cannot be directly classified, calculations have been made of the number of functions appearing at levels 1 to 5 that can be seen in Table 1.

After level 2 the number of functions grows to a size where one-vs-rest strategy of classification is infeasible. To obtain a relevant comparison between the different embeddings a representative set of functions (labels) is derived for each level of the hierarchy. Frequent itemsets are constructed up to level 5 for each GO sub-ontology, with at least 500, 1000, 2000 and 4000 proteins as support. The results of the analysis of frequent itemsets can be seen in Tables 2-4. For CC and BP sub-ontologies, functions and sets of at least 2000 proteins are presented, and for MF sub-ontologies are presented those of at least 1000 proteins. Exceptions are level 5 in CC, where the most common function occurs in only 597 proteins, which is why this level is rejected for further analysis and level 5 in MF, where there are no functions in more than 1000 proteins, resulting in the minimum support to be reduced to 500 proteins. These frequent itemsets are used as the target labels in the experiments.

TABLE I. NUMBER OF FUNCTIONS BY LEVEL FOR SUB-ONTOLOGIES CC, MF AND BP

Sub-ontol.\level	**1**	**2**	**3**	**4**	**5**
CC	16	163	609	584	273
MF	13	108	506	996	1274
BP	25	244	1026	2161	3355

TABLE II. NUMBER OF FREQUENT ITEMSETS IN AT LEAST 2000 PROTEINS BY LEVEL 1-5 IN CC SUB-ONTOLOGY

	Level 1	Level 2	Level 3	Level 4	level 5
Fr. Itemset	/	18	10	3	1
Functions	16	6	5	3	1

TABLE III. NUMBER OF FREQUENT ITEMSETS IN AT LEAST 1000 PROTEINS BY LEVEL 1-5 IN MF SUB-ONTOLOGY (*FOR LEVEL 5 AT LEAST 500 PROTEINS)

	Level 1	Level 2	Level 3	Level 4	Level 5
Fr. Itemset	/	17	6	2	9
Functions	13	8	6	2	5

TABLE IV. NUMBER OF FREQUENT ITEMSETS IN AT LEAST 2000 PROTEINS BY LEVEL 1-5 IN BP SUB-ONTOLOGY

	Level 1	Level 2	Level 3	Level 4	Level 5
Fr. Itemset	/	38	73	106	40
Functions	25	9	16	14	10

III. EXPERIMENTS, RESULTS AND DISCUSSION

Three scenarios are employed to produce a more comprehensive comparison of the network embeddings. In the first scenario protein embeddings are classified using the frequent functions itemsets obtained as described in the previous section. The second scenario extends the first by including the hierarchy in the data i.e. the set of each protein's functions is augmented using transitive closure for the GO. In both scenarios all 4 network embedding algorithms, with corresponding vector dimensions ranging from 32 to 256, are used. The last scenario involves grid search for fine tuning of the SVM algorithm. This is done

only with last level functions expanded with their ancestors and with embeddings of dimension 32.

A. *Comparison with frequent functions per level*

Figures 1-3 show classification results for each graph embedding by sub-ontology level using Hamming loss metric. This evaluation method is most commonly used for multi-label classification since it calculates the percentage of misclassified labels (in range 0-1, optimal value 0). Best results are obtained with node2vec with greater dimensions, since, node2vec for all levels gives a Hamming loss value that varies between 0.15 and 0.32, whereas, the other three graph embeddings algorithms vary, but are very similar to one another. SDNE yields worst results, especially for greater dimension, where for MF level 4 and CC level 4 hamming loss goes up to 0.46 and 0.48, respectively. HOPE and GF are similar, with HOPE giving better results in a number of levels.

B. *Comparison with function hierarchy included*

Previous experiments only include labels that are at a certain level, but do not utilize sub-ontologies hierarchy. It is not taken into account that any protein that has a specific function additionally has all the functions that are ancestors of that given function. Therefore, at any level, its ancestors are added, and due to multiple inheritance, for functions with more than one parent, it is determined whether all their parents satisfy the rule of belonging to a certain number of minimum proteins. That number for BP and CC sub-ontology functions is a minimum of 2000 proteins, and for MF is 1000 (just 500 for the last level). Finally, those ancestors that satisfy the rule are added.

The number of labels obtained is much lower than expected, due to functions inherited from the same parents and functions that were removed and reasonably sized to allow multi-label classification (Table V).

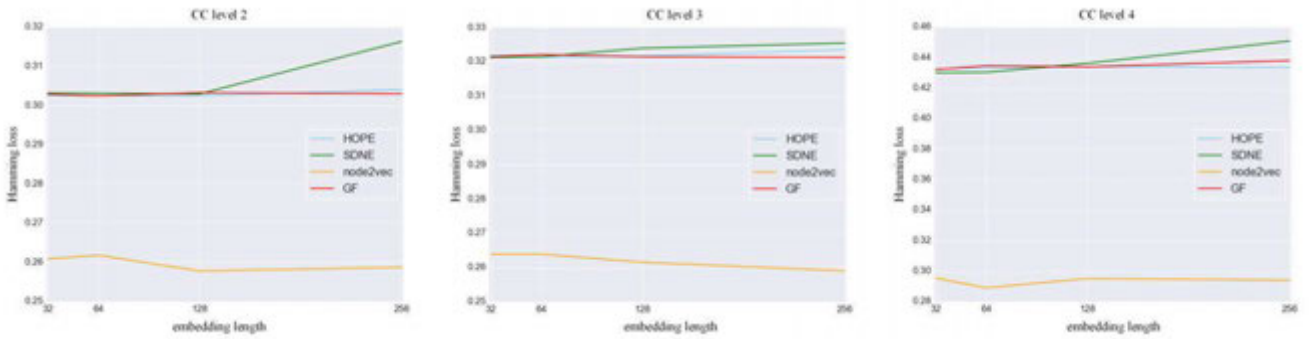


Figure 1. Results by level 2 (left), 3 (center), 4 (right) in CC sub-ontology

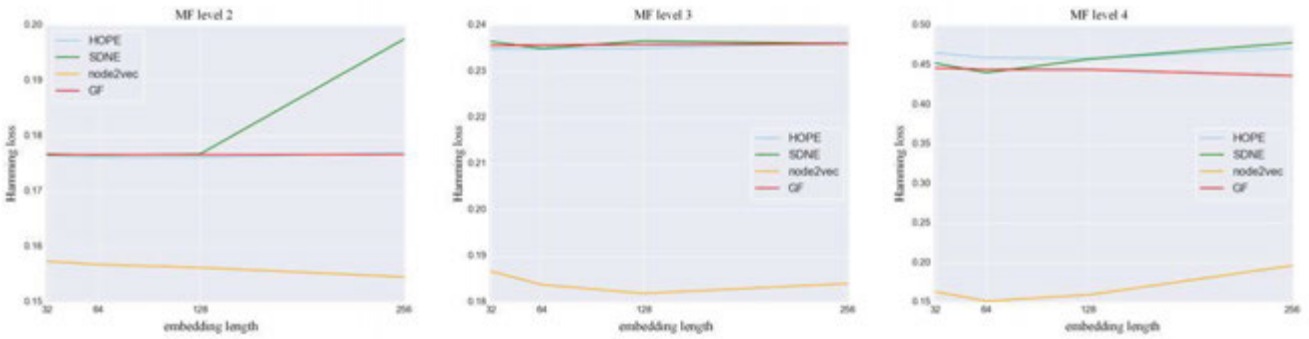


Figure 2. Results by level 2 (left), 3 (center), 4 (right) in MF sub-ontology

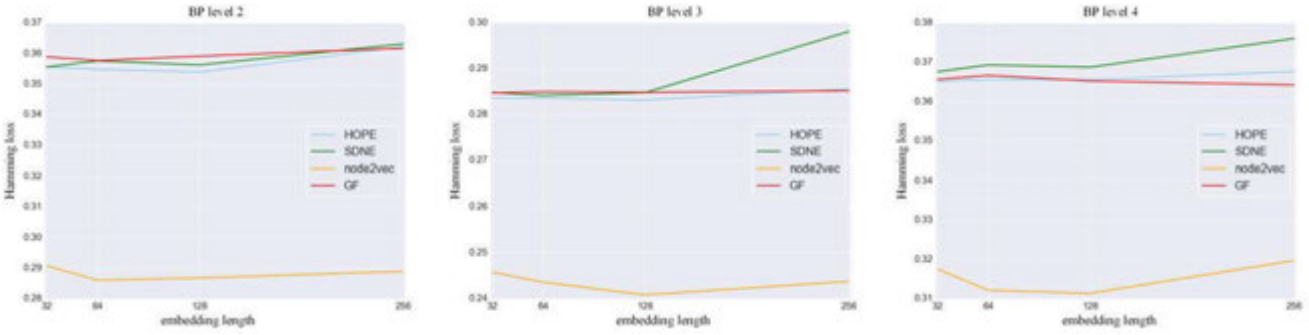


Figure 3. Results by level 2 (left), 3 (center), 4 (right) in BP sub-ontology

TABLE V. NUMBER OF FUNCTIONS BY LEVEL FOR SUB-ONTOLOGIES CC, MF AND BP

Sub-ontology \ Level	**2**	**3**	**4**	**5**
CC	11	12	13	/
MF	11	13	6	17
BP	14	26	32	35

Figures 4-6 show the classification results using the sub-ontologies hierarchy. Best results for Hamming loss are obtained with node2vec, which does not exceed 0.3. The impact of the graph embedding dimension is minimal for node2vec, where the worst result is using lowest dimension. HOPE and GF are again similar and ranked second in results, although using the hierarchy the differences between them are reduced. SDNE shows worst results, but only for higher graph embedding dimension. For the other dimensions, the differences between SDNE, HOPE and GF are insignificant. In addition, using the first level hierarchy gives similar results as in the first experimental scenario (without using GO hierarchy). However, higher levels show classification improvement by using function ancestors in all sub-ontologies and graph embeddings. The greatest improvement is seen in MF, where at level 3 there is a Hamming loss reduction of almost 0.1 for SDNE, GF and HOPE, at level 4 there is an improvement of 0.3, additionally, at level 5 where there is an improvement of 0.2.

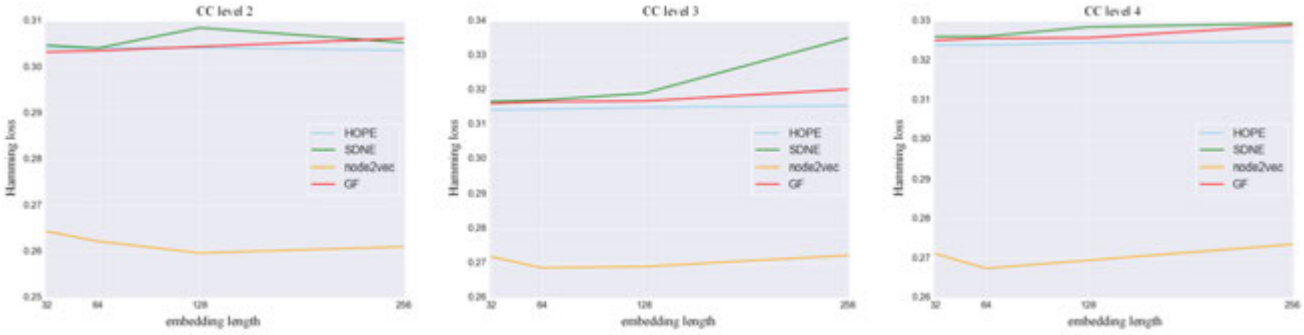


Figure 4. Results by level 2 (left), 3 (center), 4 (right) in CC sub-ontology with function hierarchy included

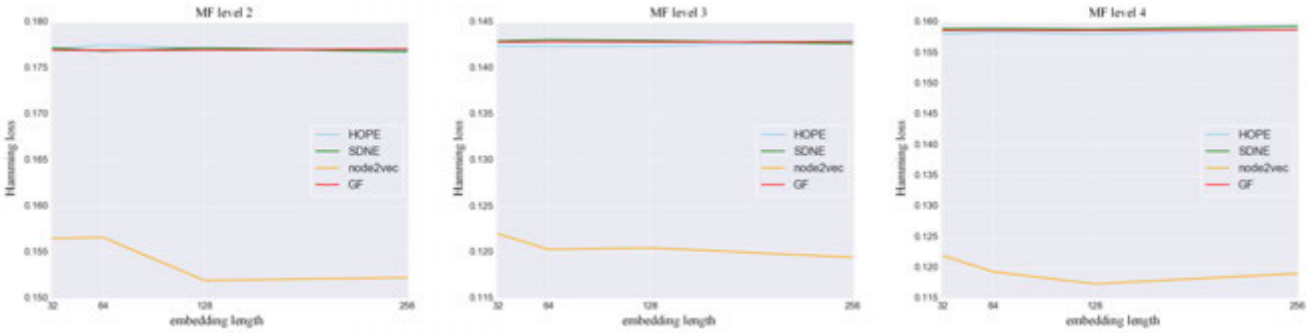


Figure 5. Results by level 2 (left), 3 (center), 4 (right) in MF sub-ontology with function hierarchy included

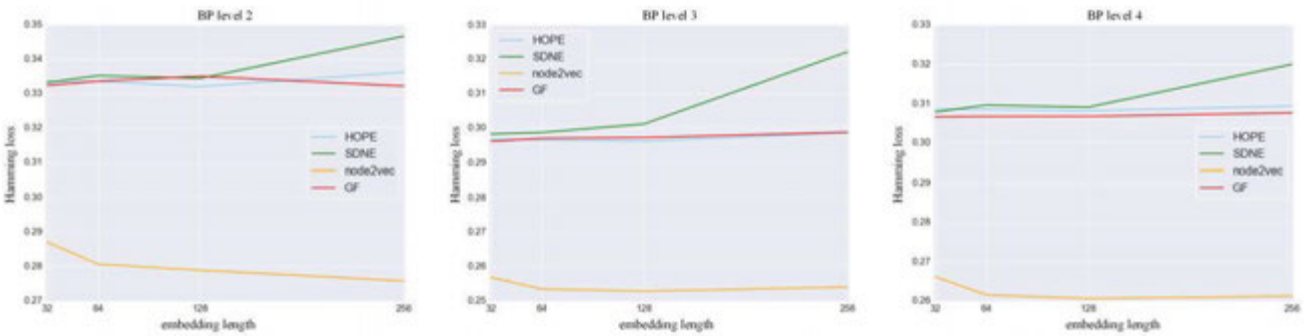


Figure 6. Results by level 2 (left), 3 (center), 4 (right) in BP sub-ontology with function hierarchy included

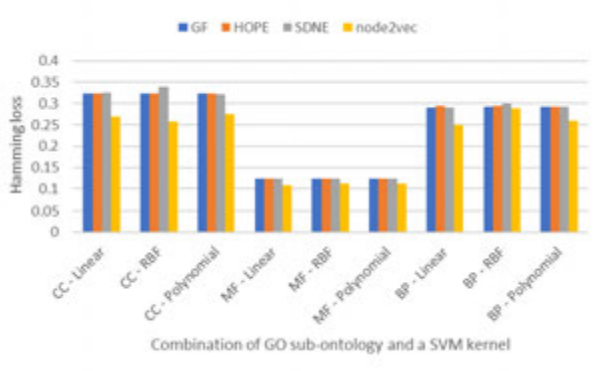


Figure 7. Results for different kernel functions in GO sub-ontologies

C. *Comparison with SVM optimization*

SVM optimization is performed with a grid search method and Hamming loss optimization function. This is done using protein functions of the last levels of the three sub-ontologies expanded with their ancestors (level 5 for MF and BP and level 4 for CC), as used in the second experimental scenario. Three versions of kernel functions are tested: linear, polynomial and RBF and various parameters are tested depending on the kernel function. Figure 7 shows the classification performance with hamming loss for different kernel SVM functions with grid search optimization for each graph embeddings algorithm and sub-ontologies. The results for the linear function and the polynomial function are almost identical. For MF sub-ontology kernel function makes no difference and the results are equal in all three cases, similarly, with CC sub-ontology, with an improvement in node2vec when using RBF kernel, and an improvement in SDNE when using linear function. There is a difference in BP sub-ontology, where both node2vec and SDNE show best results when using a linear kernel. The kernel function choice does not affect the results using HOPE and GF.

IV. CONCLUSION

In this paper a comparative analysis of the performance of different protein interaction network embeddings in a baseline classification of the protein function is performed to facilitate future research in computational protein function prediction. Four algorithms: HOPE, SDNE, node2vec and GF, are used to obtain the protein interaction network's nodes' vector representations (embeddings). The resulting vectors are used as input to a multi-label classifier that uses the one-vs-rest method. Three different experimental scenarios are employed to produce a more comprehensive comparison of the network embeddings, with and without considering the inherent hierarchy of the functions, with and without optimization of the classifier. All classifications are performed on a well represented (frequent) set of protein functions at different levels of the GO. The effects of dimensionality of the embeddings are also explored.

The results show that the most successful algorithm for extracting vector representations (embeddings) from protein interaction networks is node2vec. The difference between node2vec and other graph embeddings algorithms is from 0.15 to 0.3 using the hamming loss metris, however, the rest of the algorithms are similar in performance with negligible differences. Furthermore, node2vec is again the best choice when adding a hierarchy to the classification, in which case the margin between this algorithm and the others is greatly reduced and is maximum 0.05. The vector representations dimension has a minimal impact, with the exclusion of SDNE which shows performance degradation with dimensionality increase. The kernel function choice shows that the linear kernel improves the results of node2vec and SDNE, and gives best performance. These results and comparative analysis can be the foundation of further studies that can move the computational protein function prediction closer to a more conventional straight-forward machine learning problem, but also studies that will be able to explain the embedding process in respect to the biology it embeds.

REFERENCES

[1] N. Zhou et al., "The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens," Genome Biology, vol. 20, no. 1, p. 244, 2019.

[2] G. O. Consortium, "The gene ontology resource: 20 years and still GOing strong," Nucleic acids research, vol. 47, no. D1, pp. D330-D338, 2019.

[3] Z. Lv, C. Ao, and Q. Zou, "Protein function prediction: from traditional classifier to deep learning," Proteomics, vol. 19, no. 14, p. 1900119, 2019.

[4] K. Trivodaliev, A. Bogojeska, and L. Kocarev, "Exploring function prediction in protein interaction networks via clustering methods," PloS one, vol. 9, no. 6, 2014.

[5] K. Trivodaliev, S. Kalajdziski, I. Ivanoska, B. R. Stojkoska, and L. Kocarev, "SHOPIN: semantic homogeneity optimization in protein interaction networks," in Advances in protein chemistry and structural biology, vol. 101: Elsevier, 2015, pp. 323-349.

[6] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," Knowledge-Based Systems, vol. 151, pp. 78-94, 2018.

[7] K. Trivodaliev, M. Josifoski, and S. Kalajdziski, "Deep Learning the Protein Function in Protein Interaction Networks," in International Conference on Telecommunications, 2018, pp. 185-197: Springer.

[8] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola, "Distributed large-scale natural graph factorization," in Proceedings of the 22nd international conference on World Wide Web, 2013, pp. 37-48.

[9] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu, "Asymmetric transitivity preserving graph embedding," in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 1105-1114.

[10] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 1225-1234.

[11] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 855-864.

[12] G. Alanis-Lobato, M. A. Andrade-Navarro, and M. H. Schaefer, "HIPPIE v2. 0: enhancing meaningfulness and reliability of protein–protein interaction networks," Nucleic acids research, p. gkw985, 2016..

[13] I. Ivanoska, M. Milenkoski, S. Kalajdziski, and K. Trivodaliev, "Web tool for graph embeddings representation techniques evaluation," in 2019 42nd MIPRO, 2019, pp. 983-988: IEEE.

Modernization of the PIC codes for exascale plasma simulation

I. Vasileska [*] , P. Tomšič[**] and Leon Kos[***]
Faculty of Mechanical Engineering, University of Ljubljana, Ljubljana, Slovenia
[*] ivona.vasileska@lecad.fs.uni-lj.si
[**] pavel.tomsic@lecad.fs.uni-lj.si
[***] leon.kos@lecad.fs.uni-lj.si

***Abstract* - Particle simulation in the field of nuclear fusion is a well-established technique which has spawned dozens of codes around the world through years (e.g. BIT1, VPIC, VSIM, OSIRIS, REMP, EPOCH, SMILEI, FBPIC, GENE, WARP, PEPC) with varying degrees of specialization for different physic areas and accessibility. By leveraging advanced algorithmic features we propose to implement key for genuine pre-exascale capability. While more expensive than conventional mesh-based particle simulation, proposed approach eliminates artificial grid-generated noise by design and offers a more direct method for treating complex simulation case. A combined library would make coupling of mesh-based and mesh-free methods a lot easier. Strategically this work is expected to benefit not only by future users, but will also provide a fully alternative approach for high-fidelity modelling.**

Keywords – HPC, GPUs, PIC codes, pre-exascale computing.

I. INTRODUCTION

Numerical modeling in the field of fusion physics often demands a kinetic approach to handle extreme nonlinearities methods in the codes. Currently the model of choice in this work are the particle-in-cell (PIC) codes. The PIC codes are used to describe the time evolution of the distribution functions of different particle species (ions, electrons, impurities) in different kind of studies, for example: in gyrokinetic modelling for fusion devices or in laser-based particle beam and etc.

State of the art three dimensional PIC simulations involve up to 1012 particles on 106 cores. The complexity of the problem requires calculation of more than 10^6 equations with many variables. This possesses huge computational demand, as well as computing time. With computer architecture advancement, computational demand can be sufficiently handled, but time efficiency (needed CPU hours) remains a challenge. During the last decade technological advancement has made GPU calculation less time consuming than CPU calculation. For that reason it is beneficiary to make transition of computation from CPUs to GPUs.

It would clearly be advantageous to offer generic algorithmic components for kinetic plasma simulation on a single platform or within an easily portable library which are capable of exploiting modern supercomputing architectures. The general aim here is to transfer HPC best practices and software technology known to work well in algorithmically related, but physically much simpler applications, in order to enable PIC codes for Exascale-class simulations.

The paper is organized in six sections. After introduction, problem description and raising computational demand of PIC codes is briefly presented. Transition from traditional usage of CPUs towards modern day implementation of GPUs and its benefits is described. In the fourth section different techniques of transition of current CPU orientated towards GPU orientated codes is described. In the section Results and benefits the current activities and related challenges are described. In the closing section Conclusion, summary assessments and a view towards future challenges is presented.

II. PROBLEM DESCRIPTION

Even when performed on modern supercomputers, PIC simulation still has its limitations: apart from the computational expense of solving the particle and field equations for the full 3D electromagnetic system, the necessity of transferring information to and from the spatial grid makes it inherently noisy, collisional regimes are only accessible with the help of ad-hoc extensions, and some form of adaptive mesh refinement is often required to handle geometrically complex problems. Various simplifications are possible by exploiting geometrical symmetry, enabling 1D and 2D codes to be used for preliminary studies or parameter scans.

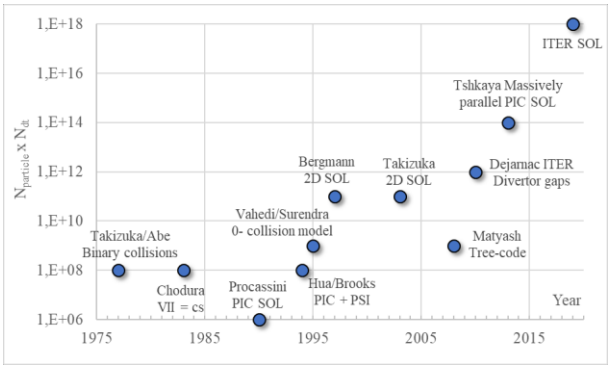


Figure 1. History of PIC simulations of fusion edge plasmas

The need for computational effort is ever growing as the PIC simulation in fusion become more complex,

shown in Figure 1. In the last 20 years the complexity of the calculations have led to increase of numerical points of factor more than 10E+6. If the phenomena of interest in PIC simulations require decreasing the time intervals, code fragmentation using implicit algorithms can be used to overcome timestep limitations of explicit codes.

III. FROM CPU TOWARDS GPU

Advancement in computer architecture design has led to an explosion in amount of data that needs to be evaluated. Processing the ever-growing data in a reasonable amount of time has been one of the key aspects of computing development.

In recent years Graphics Processing Units (GPUs) have been developed rapidly and can easily outperform Central Processing Units (CPUs) in processing large-scale data parallel workloads, but are considered weak in processing serialized tasks and communication with other devices. Current trends lead towards the CPU-GPU heterogeneous computing era [1]. For the purpose of high-performance computing and embedded computing GPUs have been widely adopted thanks to their massively-parallel architecture [2], high memory throughput and energy efficiency.

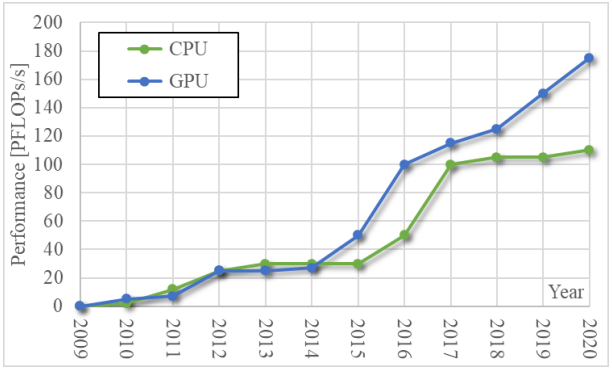


Figure 2. Top HPC systems CPU versus GPU performance growth [3]

The CPU is composed of a few cores with lots of cache memory, in contrast to a GPU that is composed of hundreds of cores that can handle thousands of threads simultaneously. The GPU makes the parallel computing more easer. This is the reason why nowadays the PIC codes, which are massively parallel, are being rewritten and are used with GPU [4]. As we compare Fig. 1 and Fig.2 we can see that during the years, as the complexity of the PIC codes has increased, also the GPUs computing performance has been on the rise. Hardware vendors have supported the move of a wider range of applications to a GPU with constantly adding new features to the accelerator so that they no longer need to work under strict close control of a CPU device. Nowadays GPUs can assume more responsibilities that would be typically carried out by the CPUs. Therefore there is a need of adopting current computer HPC codes for GPUs.

The EU has recognized the need for HPC to develop top-of-the-range supercomputers for processing big data. EuroHPC is a Joint Undertaking of the EC and 29 European countries with an aim to fund world-class integrated European HPC and data infrastructure and support a highly competitive and innovative HPC and Big Data ecosystem.

IV. METHODS

The main approach in this work is to utilize and exploit techniques which are known to work successfully in related particle-kinetic codes, adapting and transferring them. Nevertheless, we anticipate that considerable innovation will still be necessary to address the moving goal-posts of modern HPC architecture. Interestingly, some of the key components of the PIC can be characterized as:

- Generic algorithmic components of particle codes as library (field solvers, domain decomposition, particle pusher);
- Application of new developments in I/O, task-based parallelism, vectorization and memory management, paying due attention to available prototype hardware and likely adoption in future Exascale systems.
- Benchmarking against competitor PIC codes such as OSIRIS, SMILEI, PSC, PIConGPU, in approach, scalability and flexibility.

For each of the target codes, there are specific priorities to be addressed to overcome recognized bottlenecks.

For a better fit load-balancing, vectorization, tasks and I/O, the data-structures have to be evaluated and refactored. In general this comprises field and particle structures, possibly grouped for locality. Here it is important to keep localized, regular (streamed) access to particle/field data in mind in order to use GPUs/MICs effectively.

Consider a new domain-decomposition in form of an over-decomposition to enable better load-balancing over MPI ranks by using space-filling curves. This has been demonstrated previously in the PSC code [5]. As additional benefit, removing the link between the number of domains and available cores might help with approaches to fault tolerance (failing nodes).

Implementation of a task-based programming model to enable more asynchronous execution of code, improving scalability by reducing synchronization points between nodes, possibly reducing the effect of load-imbalances between nodes. This also benefits load-balancing within nodes (hybrid parallelisation) when tasks on different data can overlap.

The main challenge is on reorganization of the PIC code structure to allow more asynchronous communication in order to reduce the time consuming simulations. The much higher scaling up to 100k cores can be achieved with task-based parallelism, for which a number of options are available. This means that the main part of the code needs to be rewritten, for example as in a recent implementation of the cosmology code SWIFT [6]. Several task-based implementations exist, and differ mainly in how tasks and dependencies are created/specified, for example:

- Task spawning: any function can spawn a task, i.e. call a function that will be executed as a task. Dependencies are implicitly given by the order in which tasks are spawned, e.g. Cilk [7], OpenMP 4.0 [8].
- Dependency deduction: tasks and the data they operate on are specified explicitly, dependencies are deduced from the data and the order in which the tasks are created, e.g. QUARK, OmpSs [9], StarPU [10].
- Explicit task graph construction: tasks and dependencies are explicitly specified by the user, e.g. Intel TBB [11], QuickSched [12].

In order to prepare for the technical and scientific challenges in Exascale computing, we propose to improve the scalability and adaptability of the PIC prototype codes. This work consists of two main tasks:

A. refactoring of the codes to make them more accessible and adaptable,

B. implementing task-based parallelisation to achieve performance portability and scalability.

A. Refactoring

Code refactoring is a process of restructuring an existing code, a process which requires braking the complex system into smaller parts, which are computational easier to handle. Thus the refactoring also improves the code structure clarity. The benefits of refactoring are maintainability (it is easier to fix bugs because the source code is clearer), structural improvement (enables faster handling of the code) and speeding up development or/and analyzing.

During refactoring the codes some risks might occur. Inside the code bugs could be introduced due to incorrect utilization and using inadequate mathematical techniques.

The PIC codes are complex structured because they consist of a lot of subcodes, which describe the particles motions, fields (electric and magnetic), particles collisions and cross sections. With their connection, the PIC codes need a lot of CPU time consuming for simulations. For that reason with correct refactoring and transferring from CPU to GPU will speed up the simulations and will decrease the time consuming.

The refactoring of the PIC codes demands different types of restructuring, thus increasing the pool of new code parts for transversal activities and assures its reusability.

Refactoring of the existing codes to have a clearer separation of concern is the key to achieve the adaptability and portability needed to fully utilize extreme-scale HPC systems. Code modernization and modularization will allow for more accessible and adaptable codes with a clear separation between different levels of the code. This separation will make it easier to use external libraries and to fine-tune computational kernels to different computer architectures in order to achieve maximal performance.

In light of the current trend of increasing diversity of architectures within a single high-end HPC system (CPUs with large vector registers, GPUs connected via PCIe or NVLINK, FPGAs, vector processors as accelerators etc.), it is clear that a modular design with clear separations is needed more than ever. In order to cater for specific programming models, special treatment of data or other architecture-dependent tweaks that are required to fully utilize the hardware, it is crucial to either limit the code changes needed to a small part of the code or to offload the burden to an external library.

B. Task-based parallelisation

Task-based parallelization is a process running many different tasks at the same time of the same data. This form of parallelization uses multiple processors in parallel computing environments. One of the most important benefits using task-based parallelization is to increase the potential computing power, which the main key in CPU to GPU transfer. The risks that can occur using the parallelization are the lack of flexibility, scaling, requiring a large amount of arrays to be effective and difficulties regarding getting a good performance.

This task offers a methodological approach towards parallel computing, both in terms of algorithms and computing tasks and also in terms of diverse hardware environments.

By splitting the computation into reasonably small tasks and by defining the dependencies among the tasks, it is possible to efficiently utilize the full computing power of a single computational node, even in a heterogeneous architecture that may contain e.g. GPUs or other accelerators. Task-based parallelisation in combination with message-passing communication (MPI) for multi-node parallelisation is well suited to achieve good performance and scalability on extreme-scale computing resources containing a diversity of architectures.

For the exploration phase of this parallelization scheme, we propose to implement a simpler but (from the numerical point of view) similar grid-based code with StarPU using task-based parallelism. The main purpose of this approach is to reduce the development risk and time of the implementation of task-based parallelism in the target codes, by anticipating upcoming pitfalls, mistakes and problems. All insights will be valuable for finally introducing the task-based parallelism into the target codes. The preparation phase will develop ideas, workflows, algorithms that might be used in all of the three main codes. It will also serve as a testbed for new ideas concerning the task-based parallelism.

Because of the structure complexity of PIC codes, it is not trivial to do task-based parallelisation. For that reason in this work we will use our own prototype code, which includes the same scheme and algorithm as other PIC codes. First step is refactoring the code, following by task-based parallelisation.

V. Results and benefits

Every PIC code has the same basic algorithm, as presented in Fig. 3. The PIC code simulates the motion of each plasma particle (described with the particle pusher)

and calculates all macro-quantities (like density, current density etc.) from the position and velocity of each of these particles. The macro-force acting on the particles is calculated from the field equations i.e. set of Maxwell's equations (presented by block Field solver). If the PIC code solves the whole set of Maxwell's equations it is called electromagnetic solver.

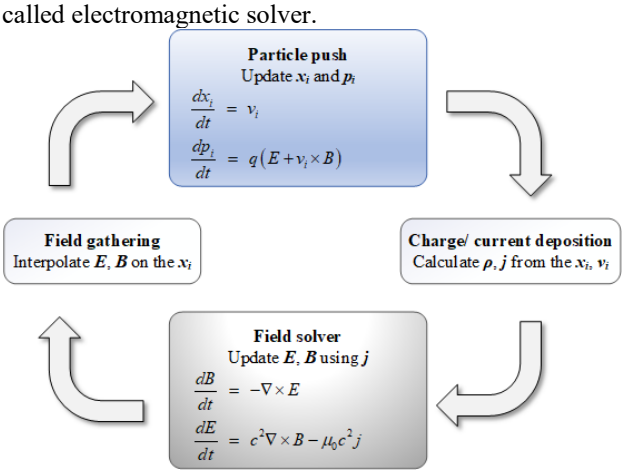


Figure 3. Shematics of PIC code algorithm [13].

The researched PIC code is presented as a prototype code with simple simulation geometry for the ease of evaluation. Such approach simplifies providing an algorithm, which goal is to adapt the code designed for running on CPUs into GPUs with potential of later to be used in other high level codes with greater complexity.

As a runtime task scheduler, which function is to transfer the simulation run tasks from CPUs into GPUs, StarPU (A Unified Platform for Task Scheduling on Heterogeneous Multicore Architecture) [14] is being deployed.

Task based parallelization approach offers the advantages of heterogeneous parallel computing in form of providing codelets, (i.e. performing arithmetic operation, such as vector multiplication on GPU) in to the prototype simple PIC as described in Fig. 4. The combination of StarPU together with the memory management and scheduling engine enables the transfer from the CPU to GPU.

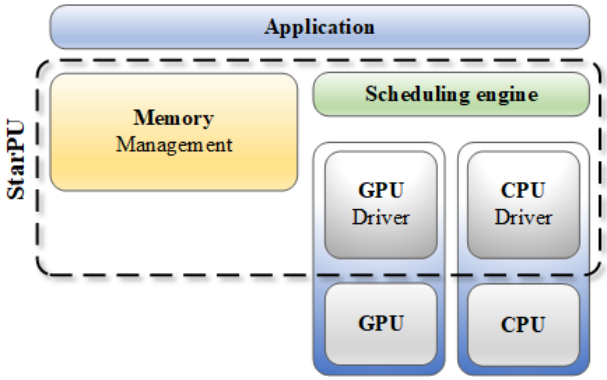


Figure 4. Function sheme of the StarPU run system.

The PIC code, used as our prototype code is a simple one-dimensional electrostatic code, meaning:

$$\nabla \times B = \partial B / \partial \tau \approx 0 \qquad (1)$$

In equation the $E = -\nabla\phi$ present self and applied electric fields E directed along coordinate x. There are no variations in y or z directions. The plane parallel problem consists of single (electron) species. Input file is provided in presented form:

```
NT=200              # of time steps

PPC=100             # of particles per cell

CPP=64              # of cell per process

DTFACTOR=0.001 #defines a fraction of dt vs.
time steps from plasma frequency;

                    #must be positive

NPROC=4             # of processors

LHSV=25000          #applied voltage on left-
hand side; RHS is grounded;
```

Transferring the current run code from CPU to GPU enables computing on a GPU node cluster. This is achieved through building the StarPU. First refactoring of the code was done, during which every solver was divided into more simple parts using extract method. After that parallelization of the code was done.

The output of the simple test PIC is presented in the form of physical particle values, available in the format of four .dat files, containing data:

- e.dat (electric field), file contains dependences of electrical field and coordinate,
- phi.dat (potential), file contains information on how the potential is changed at the coordinate,
- vxx.dat (phase space), file contains the values of the velocity,
- nt.dat; record of time evolution of number of particles per processor

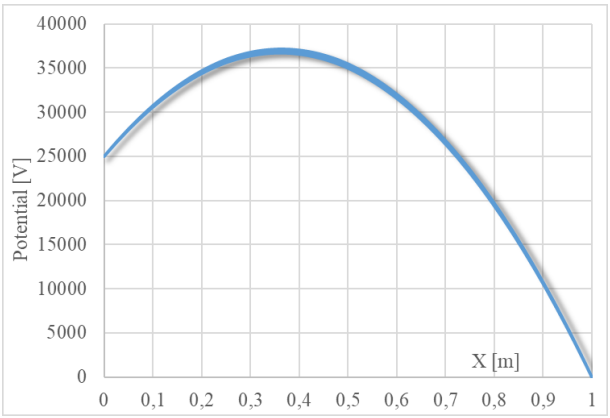


Figure 5. Plot of potential change.

Graphical illustration of a result of a test simulation is presented in Fig. 5. The presented results from the transfer from CPU to GPU show adequacy in the approach as the output values are aligned with original CPU calculations. What are the advantages of such approach on a larger scale is yet to be researched, as it is needed to further

adapt the procedure and codes to be used in cases with higher code complexity. The benefits of this work are:

1. Refactoring of the codes to achieve a clear separation between different levels in the algorithms will not only improve the portability to new architectures but also make it easier to use external libraries when and where they provide optimized solutions. The portability of the simple PIC prototype code enables its universality and therefor is not dependent from the computer architecture. By testing this prototype code, which is a simple one-dimensional electrostatic code with only one species (electrons), it will be easier to refactor the other complex PIC codes, because the workflow scheme is similar. This will also make it easier to follow the latest developments in algorithms and will encourage code reuse in other projects in the community. By making the codes more accessible and maintainable, the benefits from this effort will be far-reaching and long-lasting.
2. Task-based parallelization is expected to directly improve the performance and scalability of the codes allowing for new science. In addition, the restructuring of the codes needed to accommodate for task-based parallelism will be beneficial for the adaptability to new architectures in heterogeneous environments. Using an external runtime task scheduler should also offer an even greater level of independence from the underlying hardware.

The whole transfer from CPU to GPU for every PIC codes is a complex procedure. For that reasons we have split the work in different independent parts that are solved separately. In this work we solve just the beginning part, where we check the compatibility of this transfer using the prototype simple PIC code. For future work we will continuing for solving the other problems to get whole CPU to GPU transfer for other complex PIC codes.

VI. CONLUSION

Transfer of existing computer codes, written for numerical calculation on CPUs towards modern day usage on GPU orientated HPCs is a challenging task. In the work our goal was to test an approach on PIC code for particle simulation in the field of nuclear fusion.

We used the StarPU to transfer the simple PIC code from calculating on CPUs into GPUs. The results obtained show that the method used is adequate. It presents a base for further work. The main issue was compiling the correct StarPU, used for our simple code transfer.

The discussed approach for code adaptation from traditional CPU to GPU calculation is promising and further scaling of the idea for more complex codes seems to be a way forward for the future.

ACKNOWLEDGMENT

This work was supported in part by the grants P2-0073 and P2-0256 of the Slovenian Research Agency (ARRS). This work has been carried out within the framework of the EU's Horizon 2020 Research and Innovation Programme (2014-2020) under grant agreement 823767 (PRACE-6IP WP8). The views and opinions expressed herein do not necessarily reflect those of the European Commission.

REFERENCES

[1] R. Ubal, B. Jang, P. Mistry, D. Schaa and D. Kaeli, "Multi2Sim: A simulation framework for CPU-GPU computing," 2012 21st International Conference on Parallel Architectures and Compilation Techniques (PACT), Minneapolis, MN, 2012, pp. 335-344. https://ieeexplore.ieee.org/abstract/document/7842946

[2] Y. Sun et al., "Hetero-mark, a benchmark suite for CPU-GPU collaborative computing," 2016 IEEE International Symposium on Workload Characterization (IISWC), Providence, RI, 2016, pp. 1-10. https://ieeexplore.ieee.org/abstract/document/7581262

[3] The TOP500 project supercomputer statistics, https://www.Top500.org/ [30.06.2020]

[4] A. Binotto: A Dynamic Scheduling Runtime and Tuning System for Heterogeneous Multi- and Many-Core Desktop Platforms (2011) DOI: 10.13140/RG.2.1.3995.2405

[5] M. Poljak, M. Glavan and S. Kuzmić, "Accelerating simulation of nanodevices based on 2D materials by hybrid CPU-GPU parallel computing," 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 47-52. https://ieeexplore.ieee.org/document/8756964

[6] P. Gonnet, Task-based parallelism and why it is awesome, Tutorial presented at the Summer School on Computational Solar and Astrophysical Modelling, Juelich (2015) DOI

[7] K. Germaschewski, W. Fox, S. Abbott, N. Ahmadi, K. Mayna, L. Wang, H. Ruhl, A. Bhattacharjee, The Plasma Simulation Code: A modern particle-in-cell code with patch-based load-balancing, J. Comp. Phys. 318, 305-326 (2016)

[8] Cilk, https://www.cilkplus.org/ [30.06.2020]

[9] OpenMP,https://www.openmp.org/wp-content/uploads/OpenMP4.0.0.pdf [30.06.2020]

[10] OmpSs, https://pm.bsc.es/ompss [30.06.2020]

[11] StarPU, http://starpu.gforge.inria.fr/ [30.06.2020]

[12] IntelTBBhttps://software.intel.com/content/www/us/en/develop/tools/threading-building-blocks.html [30.06.2020]

[13] Pedro Gonnet, Aidan B. G. Chalk, Matthieu Schaller QuickSched: Task-based parallelism with dependencies and conflicts, Computer Science, (2016) URL

[14] D. Tskhakaya, K. Matyash, R. Schneider, F. Taccogna The Particle In Cell Method, CPP,. 47, 563-594 (2007) DOI

[15] Cédric Augonnet, Samuel Thibault, Samuel Thibault, Raymond Namyst, Raymond Namyst, Pierre, Pierre-André Wacrenier, StarPU: a unified platform for task scheduling on heterogeneous multicore architectures, Concurrency and Computation: Practice & Experience, 23, (2011), DOI

GPU accelerated RBF-FD solution of Poisson's equation

Mitja Jančič[1,2], Jure Slak[1,3], Gregor Kosec[1]
[1] "Jožef Stefan" Institute, Parallel and Distributed Systems Laboratory, Ljubljana, Slovenia
[2] "Jožef Stefan" International Postgraduate School, Ljubljana, Slovenia
[3] Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
mitja.jancic@ijs.si, jure.slak@ijs.si, gregor.kosec@ijs.si

Abstract – The Radial Basis Function-generated finite differences became a popular variant of local meshless strong form methods due to its robustness regarding the position of nodes and its controllable order of accuracy. In this paper, we present a GPU accelerated numerical solution of Poisson's equation on scattered nodes in 2D for orders from 2 up to 6. We specifically study the effect of using different orders on GPU acceleration efficiency.

I. Introduction

In contrast to the traditional numerical methods for solving partial differential equations (PDE) that require connectivity between discretization nodes, meshless methods can operate on scattered nodes. Although seemingly minimal difference, this feature made meshless methods popular [1] in various fields of science and engineering ranging from computational fluid dynamics [2, 3, 4] to option pricing [5]. The key to becoming so popular is that discretization of arbitrary domain with scattered nodes is considered much easier problem in comparison with meshing. To some degree it can also be automated in dimensionless sense [6].

From a historical point of view, meshless methods were introduced in 1990s. Since then, many meshless methods have been developed, e.g. the Element Free Galerkin Method [7], the Diffuse Element Method [8], the Partition of Unity Method [9], the Local Petrov-Galerkin Methods [10], the h-p Cloud Methods [11], etc.

The radial basis function-generated finite differences (RBF-FD) was first mentioned in [12] as a local strong form meshless method for solving PDEs. The method approximates differential operators only using scattered nodes. However, often used RBFs, e.g. Gaussians, include a shape parameter that can be crucial to the overall method stability [13]. Stability issue has been recently addressed by using Polyharmonic splines (PHS) and additionally augmenting them with polynomials [14].

The generality of the meshless methods comes with the price, namely higher complexity. First, the shape functions of stencil weights have to be computed every time from scratch. Second, the support sizes are typically much bigger than in mesh-based methods, especially in high order RBF-FD methods. The natural way to accelerate most computationally demanding parts of the solution procedure is to employ parallel computing. There have been several reports on parallel meshless solution procedures in the past, including distributed computing as well as shared memory approaches [4, 15, 16, 17, 18]. In this paper we analyze the graphics processing unit (GPU) acceleration of RBF-FD explicit solution of Poisson problem with Dirichlet boundary condition. We are especially interested in the effect of using different polynomial orders.

The rest of the paper is organized as follows: in section II a short presentation of local meshless methods is given, in section III we present our workflow and how GPU was included in our computations, in section IV problem is explained, in section V results are presented, and in section VI final conclusions are given.

II. Local strong form meshless methods

In general, the idea of meshless methods is the use of local discretization points to construct an approximation of the considered field and later use it for manipulation with differential operators. Discretization points are referred to as computational nodes or just *nodes*. The nodes are placed within the domain and on its boundary, and their distribution can be scattered, as presented in Fig. 1.

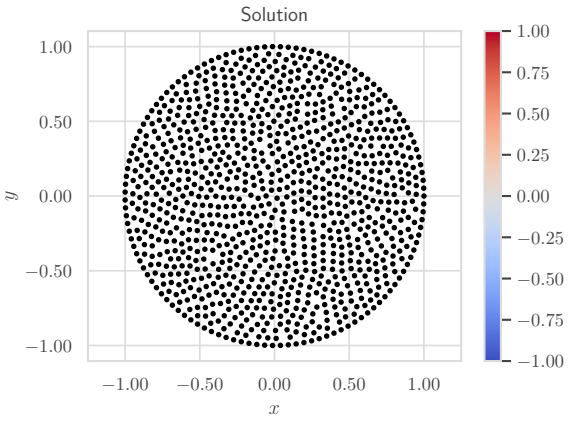


Figure 1. Solution of Poisson's problem with Dirichlet boundary conditions on scattered 2D nodes. Chosen highest polynomial degree $m = 2$, support size $n = 15$ and number of nodes $N = 1027$.

Local strong form meshless methods approximate derivatives in the form

$$(\mathcal{L}u)(x) \approx \sum_{p \in N(x)} w_i^{\mathcal{L}}(x) u(p), \qquad (1)$$

where $N(x)$ is a set of neighboring nodes of x and $\mathcal{L}$ is a differential operator. Different ways of computing the weights w exists. We will use the RBF-FD method [19].

We have implemented the RBF-FD based solution procedure using object oriented approach and C++'s strong template system. Node positioning, support selection, differential operator approximation and PDE discretization and other modules are all available in one package, the Medusa library [20]. Please refer to our open source Medusa library for more features and examples.

III. GPU IMPLEMENTATION

Compared to a CPU-based system, the memory bandwidth in GPU based systems is often an order of magnitude higher, however at the price of higher memory latency. A single access to on board GPU memory takes approximately 400-600 cycles compared to a floating point operation which takes approximately 1-2 cycles. Current GPUs support large numbers of processing elements and must be programmed in a high data-parallel way in order to observe the advantages of GPU programming and to help hide the memory latency. Large data is also needed to keep the processing units busy and good knowledge of GPU systems is required to reduce the memory communication and optimize the kernel functions [17].

In this work only part of the problem was dumped to the GPU. The illustration of the GPU implementation is in Fig. 2. Green boxes are specific to GPU programming while the yellow box presents the actual calculations done on the GPU. Shapes, node positions and support nodes were still tasks executed by the CPU, only the explicit time loop (yellow box in Fig. 2) was ported to the GPU. Time loop execution was also timed using high resolution timer, enabling us to estimate the speedups and thus evaluate the performance of CPU and GPU.

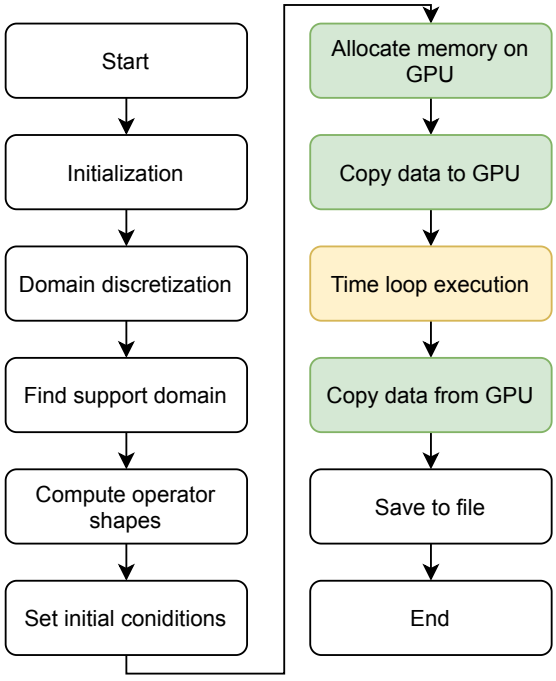


Figure 2. The implementation scheme.

A. Time loop implementation

Once the shapes, node positions and support nodes computation was carried out on the host (CPU), memory was allocated on the device (GPU) using `cudaMalloc` memory allocation function. As necessary data structures for numerically feasible problem sizes on the onboard GPU memory, we simply copied relevant data to the host using the `cudaMemcpy` function with `cudaMemcpyHostToDevice` argument and left the data there during time loop execution.

Time loop computation on the GPU was executed by introducing a `solveOnGpu` kernel function as shown in code listing 1.

```
int tpb = 32;
int bpg = (N + tpb − 1) / tpb;

for (int step = 0; step < timeSteps; ++step) {
  solveOnGpu <<<bpg, tpb>>> (d_u1, d_u2, ...);
  cudaMemcpy(d_u1, d_u2, sizeof(double) ∗ N,
                  cudaMemcpyDeviceToDevice);
}
```

Listing 1: Sample code of time loop execution on GPU.

In Compute Unified Device Architecture (CUDA) the actual use of multi-processors and processing units is controlled by block size [21]. This reflects in the number of threads which act on each processor. In order to fully exploit the benefits of a GPU, proper definition of block size is of great importance. When chosen to small, processing units tend to stall, if chosen to large other effects (e.g. register spilling) might interfere with top performance [17].

Several tests of block size of multiples of 32 were made. Best time performance showed at $N_{tpb} = 32$ threads per block and $N_{bpg} = (N+N_{tpb}-1)/N_{tpb}$ blocks per grid with N nodes in the domain.

Once the time loop execution on the GPU is finished, `cudaDeviceSynchronize` function is used for synchronization and results are copied from the device to host using `cudaMemcpy` function with the `cudaMemcpyDeviceToHost` argument. The results are saved to a file and post-processing is done with python 3. To prevent memory leakage, `cudaFree(varibaleName)` was used before the shutdown.

IV. PROBLEM DESCRIPTION

With the aim to analyze the performance of CPU and GPU in terms of support size and highest augmented polynomial degree, proposed solution procedure and its implementation is studied on a Poisson problem with Dirichlet boundary conditions:

$$\frac{\partial}{\partial t}u(\boldsymbol{p}) - \nabla^2 u(\boldsymbol{p}) = f(\boldsymbol{p}) \qquad \text{in } \Omega, \qquad (2)$$

$$u(\boldsymbol{p}) = \prod_{i=1}^{d} \sin(\pi_i) \qquad \text{on } \partial\Omega, \qquad (3)$$

where $f(\boldsymbol{p}) = d\pi^2 \prod_{i=1}^{d} \sin(\pi p_i)$ and domain space Ω is a $d = 2$ dimensional unit disk with boundary $\partial\Omega$.

The problem is discretized in both time and spatial domain space

$$u_2(\boldsymbol{p}) = u_1(\boldsymbol{p}) + \mathrm{d}t(f(\boldsymbol{p}) + \nabla^2 u_1(\boldsymbol{p})), \qquad (4)$$

where $\mathrm{d}t$ is an infinitesimal time step and ∇^2 is approximated as described in section II.

The closed form solution of the above problem is $u(\boldsymbol{p}) = \prod_{i=1}^{d} \sin(\pi p_i)$ allowing us to validate the numerically obtained solution.

V. Results

Numerical results are computed using RBF-FD with PHS radial basis functions $\Phi(r) = r^3$ and monomial augmentation. Radial function was kept the same for all cases. The convergence order is directly controlled by the highest augmented polynomial degree $m \in \{2, 4, 6\}$, however the larger the polynomial degree, the larger the recommended support size $n = \binom{m+d}{m}$. Recommended support size in terms of domain space dimensionality and highest polynomial degree was first given by Bayona [14], however larger supports can also be used [22]. In our test $n \in \{12, 15, 20, 30, 45, 60\}$.

All computations were performed on a single core of a computer with `Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz` processor and 64 GB of DDR4 memory and graphical processing unit `NVIDIA GeForce RTX 2080 Ti, 11GB GDDR6` with `4352 CUDA cores`. Code was compiled using `g++ (GCC) 8.1.0` on Linux with `-O3 -DNDEBUG -std=c++11` flags while GPU code was compiled using `CUDA release V9.1.85`.

An example of numerical solution is shown in Fig. 1.

A. Execution times

The discretized Poisson problem (4) was solved for 10^5 time steps with time step $\mathrm{d}t = 10^{-6}$ seconds, resulting in total simulation time $t = 0.1$ seconds.

The polynomial degree takes an important role in the shape computation stage. After shapes are stored the differential operator only depends on the support. As illustrated on implementation scheme from Fig. 2, the shapes and supports are in our case calculated on the CPU. Different time loop execution times for various polynomial degrees m are therefore not expected neither are they observed.

In Fig. 3 we observe how the performance from CPU-only code is directly proportional to number of nodes, while multiple regimes can be seen in the single-GPU implementation.

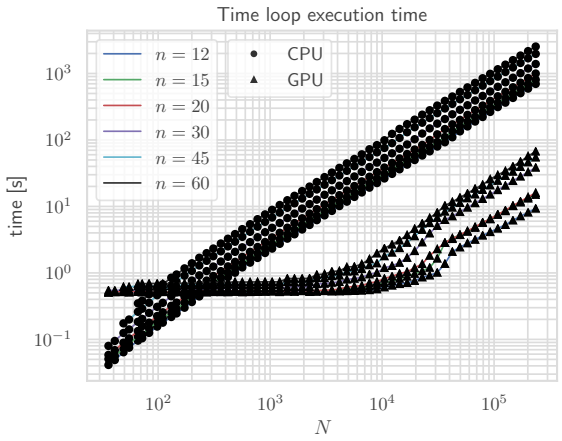


Figure 3. Time loop execution times on single CPU and GPU. Chosen highest polynomial degree $m = 2$.

In the first regime it makes no significant difference what portion of the available L2 cache is occupied by the data set. There is just not enough data dumped to the GPU to fully exploit the advantages of parallelization as most of the time is spent on communication and data distribution rather than on computation itself. For larger data sets ($N > 10^3$), the computation part prevails and faster computation times are observed. Increasing data set size even further again leads to computation times proportional to the number of nodes - similar to the CPU. As expected generally higher execution times are observed for larger support sizes.

Speedup is defined as ratio between the execution of the time loop on a single CPU t_{CPU} and single GPU t_{GPU}. Speedups are shown on Fig. 4. Similar regimes as in Fig. 3 can be seen starting with low or zero benefit from GPU implementation. Increasing the data exploits the parallelization advantages which peak at certain data set size dumped to the GPU.

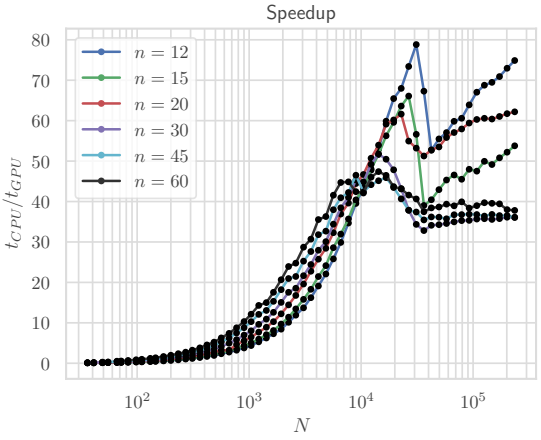


Figure 4. Observed speedups.

To explain the notable drop of speedup in Fig. 4, we have to understand hardware structure in more detail. The L2 cache size of the `NVIDIA GeForce RTX 2080 Ti` graphics processing unit is 5.5 MB [23]. Therefore increasing the data set size exploits the advantages of GPU implementation but only until the L2 cache of the GPU is

almost, but not yet, full. Once the L2 cache is filled, more memory communication with higher latency is needed.

A rough estimation of how much data is copied to the GPU can be made. In our implementation we used doubles and integers, $N(4+n)$ double-precision and N_i+Nn integer numbers, where N_i is number of nodes in interior. Knowing that size of double and integer is 8 and 4 bytes respectively, we can estimate when the L2 cache is full and approximately estimate the speedup peak position. The calculations are gathered in table 1. Note that the estimated $N_{5.5\,\mathrm{MB}}$ is calculated by enforcing $N_i = N$. The larger the number of nodes, the more acceptable this enforcement is. Estimated data size copied on GPU is also shown in Fig. 5.

Support size	Estimated $N_{5.5\,\mathrm{MB}}$	N speedup peak
12	30556	31085
15	25463	26601
20	19928	19527
30	13889	14313
45	9549	8991
60	7275	7707

TABLE 1. ESTIMATED AND OBSERVED PEAK PERFORMANCE AS FUNCTION OF NUMBER OF NODES N.

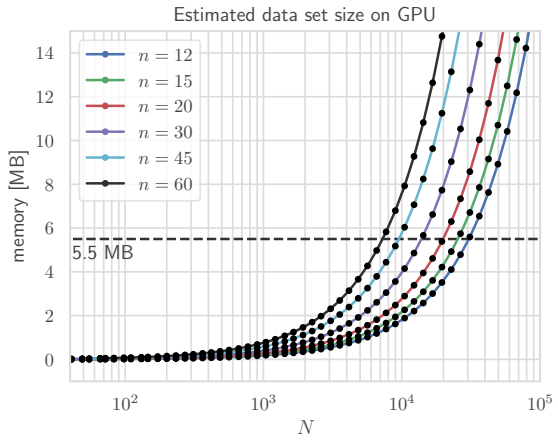


Figure 5. Estimated data set size copied to GPU.

Table 1 shows how observed speedup peaks are close to the estimated calculations. We therefore conclude that the drop of speedup performance is closely related to the size of L2 cache.

VI. CONCLUSIONS

The execution performance of solution of a Poisson problem on 2D scattered nodes with Dirichlet boundary conditions was analyzed in this paper. We measured how changing support sizes, a crucial parameter in using high order RBF-FD, and total number of discretization nodes affects the parallel execution performance. We observed the speedup peaks, which can be explained by measuring the data set size dumped to the GPU and comparing it with the available L2 cache memory size.

In this paper only the execution of time loop was considered as performance important, however some researchers already reported on using GPUs to calculate the shapes. The next step could therefore be to dump even more computation to the GPU. Additional extension could also be by solving the Poisson problem implicitly or by comparing the computational performance using mulitple CPU and GPU devices.

ACKNOWLEDGMENTS

The authors would like to acknowledge the financial support of the ARRS research core funding No. P2-0095 and the Young Researcher program PR-08346.

REFERENCES

[1] Bengt Fornberg and Natasha Flyer. Solving pdes with radial basis functions. *Acta Numerica*, 24:215–258, 2015.

[2] Gregor Kosec. A local numerical solution of a fluid-flow problem on an irregular domain. *Adv. Eng. Software*, 120:36–44, 2018.

[3] Eko Prasetya Budiana et al. Meshless numerical model based on radial basis function (rbf) method to simulate the rayleigh–taylor instability (rti). *Computers & Fluids*, page 104472, 2020.

[4] Jia-Le Zhang, Zhi-Hua Ma, Hong-Quan Chen, and Cheng Cao. A gpu-accelerated implicit meshless method for compressible flows. *J. Comput. Phys.*, 360:39–56, 2018.

[5] Jamal Amani Rad, Kourosh Parand, and Saeid Abbasbandy. Pricing european and american options using a very fast and accurate scheme: The meshless local petrov–galerkin method. *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences*, 85(3):337–351, 2015.

[6] Jure Slak and Gregor Kosec. On generation of node distributions for meshless PDE discretizations. *SIAM Journal on Scientific Computing*, 41(5):A3202–A3229, jan 2019.

[7] T. Belytschko, Y. Y. Lu, and L. Gu. Element-free galerkin methods. *Int. J. Numer. Methods Eng.*, 37(2):229–256, 1994.

[8] B. Nayroles, G. Touzot, and P. Villon. Generalizing the finite element method: Diffuse approximation and diffuse elements. *Comput. Mech.*, 10(5):307–318, Sep 1992.

[9] J. M. Melenk and I. Babuška. The partition of unity finite element method: Basic theory and applications. *Comput. Methods Appl. Mech. Eng.*, 139(1):289 – 314, 1996.

[10] Satya N. Atluri and Tulong Zhu. A new meshless local petrov-galerkin (mlpg) approach in computational mechanics. *Comput. Mech.*, 22(2):117–127, 1998.

[11] C. Armando Duarte and J. Tinsley Oden. H-p clouds—an h-p meshless method. *Numerical Methods for Partial Differential Equations: An International Journal*, 12(6):673–705, 1996.

[12] A. I. Tolstykh and D. A. Shirobokov. On using radial basis functions in a "finite difference mode" with applications to elasticity problems. *Comput. Mech.*, 33(1):68–79, 2003.

[13] Natasha Flyer, Bengt Fornberg, Victor Bayona, and Gregory A. Barnett. On the role of polynomials in RBF-FD approximations: I. Interpolation and accuracy. *J. Comput. Phys.*, 321:21–38, 2016.

[14] Victor Bayona, Natasha Flyer, Bengt Fornberg, and Gregory A. Barnett. On the role of polynomials in RBF-FD approximations: II. Numerical solution of elliptic PDEs. *J. Comput. Phys.*, 332:257–273, 2017.

[15] Roman Trobec and Gregor Kosec. *Parallel scientific computing: theory, algorithms, and applications of mesh based and meshless methods*. Springer, 2015.

[16] Jesse M. Kelly, Eduardo A. Divo, and Alain J. Kassab. Numerical solution of the two-phase incompressible navier–stokes equations using a gpu-accelerated meshless method. *Eng. Anal. Boundary Elem.*, 40:36–49, 2014.

[17] G. Kosec and P. Zinterhof. Local strong form meshless method on multiple Graphics Processing Units. *CMES: Computer Modeling in Engineering and Sciences*, 91(5):377–396, 2013.

[18] Evan F. Bollig, Natasha Flyer, and Gordon Erlebacher. Solution to pdes using radial basis function finite-differences (rbf-fd) on multiple gpus. *J. Comput. Phys.*, 231(21):7133–7151, 2012.

[19] Jure Slak, Blaž Stojanovič, and Gregor Kosec. High order RBF-FD approximations with application to a scattering problem. In *2019 4th International Conference on Smart and Sustainable Technologies (SpliTech)*, pages 1–5. IEEE, 2019.

[20] Jure Slak and Gregor Kosec. Medusa: A C++ library for solving pdes using strong form mesh-free methods, 2019.

[21] John Nickolls, Ian Buck, Michael Garland, and Kevin Skadron. Scalable parallel programming with cuda. *Queue*, 6(2):40–53, 2008.

[22] Mitja Jančič, Jure Slak, and Gregor Kosec. Analysis of high order dimension independent RBF-FD solution of Poisson's equation, 2019.

[23] William Harmon. Nvidia geforce rtx 2080 ti graphics card review, 2019.

Parallel point sampling for 3D bodies

Urban Duh[1,2], Matjaž Depolli[1], Jure Slak[1,2], Gregor Kosec[1]
[1] "Jožef Stefan" Institute, Parallel and Distributed Systems Laboratory, Ljubljana, Slovenia
[2] Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
urban.duh@student.fmf.uni-lj.si, matjaz.depolli@ijs.si, jure.slak@ijs.si, gregor.kosec@ijs.si

Abstract – Point sampling in interiors and on surfaces of 3D bodies has important applications in computer graphics, visualization and physical simulations. Some of those application require certain statistical properties of the generated sample, but many don't and it is often only desired that the final points are distributed quasi regularly, with approximately equal distances to the neighboring points. We present a parallel version of an algorithm for point sampling that supports generation of locally regular point sets, but with globally variable point spacing. The behavior of the algorithm is show-cased on interiors and surfaces of 2D and 3D models. The speedups of the parallel execution are also reported.

I. Introduction

Point sampling in interiors and on surfaces of 3D bodies has important applications. In computer graphics it is used for rendering, imaging, and geometry processing [1]. In statistics, point sampling with desired distributions can be used to generate samples [2] and help with Monte-Carlo simulations. Such point sets often posses certain probabilistic properties. However, these properties are often not needed, such as in physical simulations, where point sets can be used as a base for mesh generation or to generate point sets for meshless methods [3, 4]. In many cases, it is also common to generate nodes on the boundary first, so that they can be stored and treated separately and even used as initial data for interior discretization.

Many approaches for generating locally uniform point distributions exist. A common technique is to distribute the points in such a way that they minimize some potential, analogous to simulation of free charged particles or charged particles bound to surfaces [5, 6]. This approach is often iterative, starting from an initial distribution and improving it. Other iterative approaches, such as bubble simulation and Voronoi relaxation are also used [7]. These approaches can always be applied to an already generated point set to improve its quality.

Methods for generating node sets use different techniques. Among more common are the advancing front techniques [8] and circle/sphere packing techniques [9]. Poisson disk sampling is popular for generating node sets with excellent spatial and spectral properties, which are very useful in computer graphics. Efficiently parallelized Poisson sampling has been well researched [10, 11]. It usually builds on top of dart throwing algorithm, which produces results with excellent spectral properties, which are required in computer graphics. But compared to iterative techniques it is generally slower, does not naturally support variable nodal density, and is difficult to expand past its native to 2 dimensions.

In this paper, we present a point sampling algorithm for interiors and surfaces of 3D bodies, which can be used when only a quasi regular distribution is required without strict spectral properties. The method is very efficient in its sequential form, but moreover, it also exhibits great parallel scalability.

The rest of the paper is organized as follows: in section II an algorithm for point generation in interiors of d-dimensional bodies is described. The algorithm is generalized to sampling on boundaries in section III. Section IV describes the parallel implementation and shows the execution times and speedups. Finally, conclusions are presented in section V.

II. Point sampling in interiors

A. Description of the algorithm

Let $\Omega \subset \mathbb{R}^d$ be a body of arbitrary dimension d we wish to sample and let $h\colon \Omega \to \mathbb{R}$ be a function mapping points from Ω to desired distances to neighboring points, called a spacing function. The interior sampling algorithm optionally accepts a set of seed points $\chi \subset \Omega$. Often, seed points can be generated by sampling the boundary of the domain, which is described in Section III. If the set of seed points is not given, one seed point is chosen in Ω randomly. Initially, all the seed points are enqueued and then expanded until the queue is empty.

The algorithm progresses by popping elements from the queue sequentially. In each iteration, a point $\boldsymbol{p}_i$ is dequeued and a set of n candidates $C_i = \{\boldsymbol{c}_{i,j}; j = 1, 2, \dots n\}$ is generated on a unit sphere with radius $r_i = h(\boldsymbol{p}_i)$ around it. The generation of points on a unit sphere can be done in many ways, in this paper we shall use a uniformly discretised sphere rotated by a random angle around its axis, which was found to be desirable in a previous paper [4].

After that, the set C_i is iterated and for every candidate $\boldsymbol{c}_{i,j}$ it is checked, if it is at least r_i away from all the preexisting points. If the check is successful, candidate $\boldsymbol{c}_{i,j}$ is added to the list of accepted points and enqueued for expansion. Calculating the distances to nearest neighbors requires an efficient spatial search structure, in this paper we shall be using k-d trees [12] provided by the `nanoflann` library [13].

Once the queue is empty, all the points have been expanded and the set of accepted points is complete. The number of generated candidates at each expansion (n) is a free parameter of this interior sampling algorithm. Generally, the optimal value for n depends on the dimension d and regularity of the domain. In this paper we shall be

using $n = 15$.

A more detailed presentation of the algorithm and its use in physical simulations using meshless methods is given in [4].

B. Time complexity

We establish output sensitive theoretical time complexity of the algorithm. Let N be the total number of generated points. Every point is inserted once and for every generated point, n candidates are considered when expanding. Therefore, the total time complexity of interior sampling algorithm is

$$T_{\text{interior}} = T_{\text{init}} + \sum_{i=1}^{N}(T_{\text{insert}} + \sum_{j=1}^{n} T_{\text{query}}), \quad (1)$$

where T_{init} is the time complexity of initialization of the spatial search structure and T_{query}, T_{insert} are time complexities for querying and inserting into the spatial search structure. For k-d trees, we know that $T_{\text{query}} = T_{\text{insert}} = O(\log N)$ and $T_{\text{init}} = O(N_s \log N_s)$, where $N_s = |\chi|$ [13]. By assuming $N_s \ll N$ we arrive at the total time complexity of

$$T_{\text{interior}} = O(nN \log N). \quad (2)$$

C. Examples

Although the interior sampling algorithm can be used in arbitrary dimensions, we will focus on 2D and 3D domains in this paper. In Figure 1 a simple and a more complex 2D domain were sampled with variable and constant spacing function h.

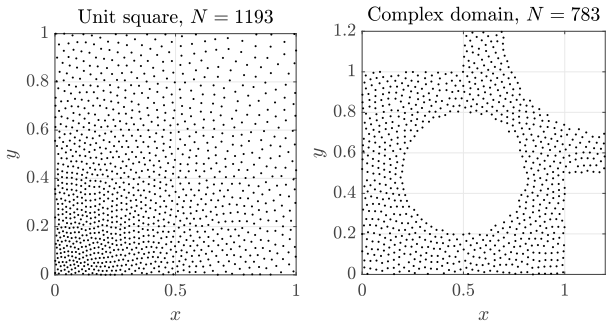


Figure 1. Unit square sampled with $h(x, y) = 0.02(x^{1.5} + y^{1.5}) + 0.015$ and a more complex 2D domain sampled with $h = 0.03$ by our interior sampling algorithm.

Similar results in 3D are presented in Figure 2. It can be observed that the algorithm adapts local spacing variably, and does so in a smooth fashion, to not cause too large discrepancies in distances to neighboring points.

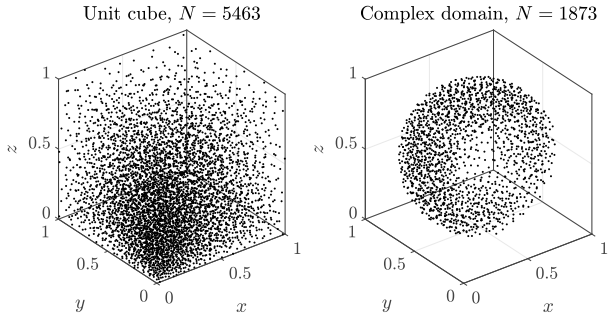


Figure 2. Unit cube sampled with $h(x, y, z) = 0.033(x^{1.5} + y^{1.5} + z^{1.5}) + 0.025$ and a complex 3D domain sampled with $h = 0.05$ by our interior sampling algorithm.

III. Point sampling on parametric surfaces

A. Brief description

Interior sampling algorithm described in Section II can be adapted to work on domain boundaries, for example curves in 2D and surfaces in 3D. Let $\partial\Omega$ be a domain boundary parametrized with a regular parametrization $\boldsymbol{r} : \Lambda \subset \mathbb{R}^{d-1} \to \partial\Omega \subset \mathbb{R}^d$ and let $h\colon \partial\Omega \to \mathbb{R}$ be the desired spacing function. The set Λ will be called the parametric space and its elements are called parameters.

We may consider the given problem as the problem of sampling the interior of the parametric space Λ in a way, that when mapped to $\partial\Omega$ by $\boldsymbol{r}$ we get a boundary sampled with spacing h. Therefore, the general logic of enqueueing points awaiting expansion and iteratively expanding them can stay the same, we only need to generate candidates for expansion in a different way. Let $\boldsymbol{\lambda}_i \in \Lambda$ be a parameter in the parametric domain we wish to expand. What we actually want is to generate a set of candidate parameters $H_i \subseteq \Lambda$ so that for each $\boldsymbol{\eta}_{i,j} \in H_i, j = 1, 2, \ldots n$

$$\|\boldsymbol{r}(\boldsymbol{\eta}_{i,j}) - \boldsymbol{r}(\boldsymbol{\lambda}_i)\| = h(\boldsymbol{r}(\boldsymbol{\lambda}_i)), \quad (3)$$

where $\|\cdot\|$ is the Euclidean norm. Let

$$\boldsymbol{\eta}_{i,j} = \boldsymbol{\lambda}_i + \alpha_{i,j}\vec{s}_{i,j}, \quad (4)$$

where $\vec{s}_{i,j}$ is a unit vector and $\alpha_{i,j} > 0$. We can approximate

$$\boldsymbol{r}(\boldsymbol{\eta}_{i,j}) = \boldsymbol{r}(\boldsymbol{\lambda}_i + \alpha_{i,j}\vec{s}_{i,j}) \approx \boldsymbol{r}(\boldsymbol{\lambda}_i) + \alpha_{i,j}\nabla\boldsymbol{r}(\boldsymbol{\lambda}_i)\vec{s}_{i,j} \quad (5)$$

using first order Taylor's expansion. By substituting (5) into (3) we get

$$h(\boldsymbol{r}(\boldsymbol{\lambda}_i)) \approx \|\boldsymbol{r}(\boldsymbol{\lambda}_i) + \alpha_{i,j}\nabla\boldsymbol{r}(\boldsymbol{\lambda}_i)\vec{s}_{i,j} - \boldsymbol{r}(\boldsymbol{\lambda}_i)\| = \\ = \alpha_{i,j}\|\nabla\boldsymbol{r}(\boldsymbol{\lambda}_i)\vec{s}_{i,j}\|, \quad (6)$$

from where we can express

$$\alpha_{i,j} \approx \frac{h(\boldsymbol{r}(\boldsymbol{\lambda}_i))}{\|\nabla\boldsymbol{r}(\boldsymbol{\lambda}_i)\vec{s}_{i,j}\|}. \quad (7)$$

Thus, the set of candidates for expansion of $\boldsymbol{\lambda}_i$ can be written as

$$H_i = \left\{\boldsymbol{\lambda}_i + \frac{h(\boldsymbol{r}(\boldsymbol{\lambda}_i))}{\|\nabla\boldsymbol{r}(\boldsymbol{\lambda}_i)\vec{s}_{i,j}\|}\vec{s}_{i,j};\ \vec{s}_{i,j} \in S_i\right\}, \quad (8)$$

where S_i is a random discretization of a unit sphere.

With the above definition of H_i, Generated candidates $\boldsymbol{r}(H_i) \subset \partial\Omega$ in each iteration are approximately $r_i = h(\boldsymbol{r}(\boldsymbol{\lambda}_i))$ away from $\boldsymbol{r}(\boldsymbol{\lambda}_i)$, the point that is currently being expanded. We can now accept candidates that are at least r_i away from already existing points using a spatial search structure built on $\partial\Omega$ and enqueue them for expansion. When all the points are expanded, surface sampling algorithm terminates and returns the set of accepted points.

This algorithm also has a free parameter n, which also depends on the dimension d and regularity of the given domain. We will be using $n = 2$ for curves in 2D and $n = 15$ for surfaces in 3D.

The time complexity of surface sampling algorithm is also $O(nN \log N)$, assuming the usage of k-d trees as spatial search structures, since only the candidate generation process is prolonged by a constant factor and the core logic remains the same. This assumes that evaluation of $\boldsymbol{r}$ and its derivatives takes constant time.

B. Examples

Parametric surface sampling algorithm works on hypersurfaces of arbitrary dimensions, but we will once again focus on curves in 2D and surfaces in 3D. In Figure 3 a circle parametrized with

$$\boldsymbol{r}(\varphi) = 0.5(\cos\varphi + 0.5, \sin\varphi + 0.5) \tag{9}$$

with $\varphi \in \Lambda = [0, 2\pi)$ and a more complex curve in 2D with parametrization

$$\boldsymbol{r}(\varphi) = ((3+\cos(3\varphi))\cos\varphi, (3+\cos(3\varphi))\sin\varphi), \tag{10}$$

$\varphi \in \Lambda = [0, 2\pi)$ were sampled. From the results on the more complex curve in 2D we can see that the results of the surface sampling algorithm are only approximate. However, the used Taylor approximation (5) gets better with smaller spacing functions h. If the results are not good enough, higher order approximations can also be used. Further analysis of the algorithm and the error is left for future work.

Similar results on a sphere parametrized with

$$\boldsymbol{r}(u, v) = 0.5(\cos u \sin v, \sin u \sin v, \cos v) + 0.5, \tag{11}$$

$(u, v) \in \Lambda = [0, 2\pi) \times [0, \pi)$ and a more complex surface in 3D with parametrization

$$\boldsymbol{r}(u, v) = (\sqrt{0.25 + u^2}\cos v, \sqrt{0.25 + u^2}\sin v, u), \tag{12}$$

$(u, v) \in \Lambda = [-1, 1] \times [0, 2\pi)$ are shown in Figure 4.

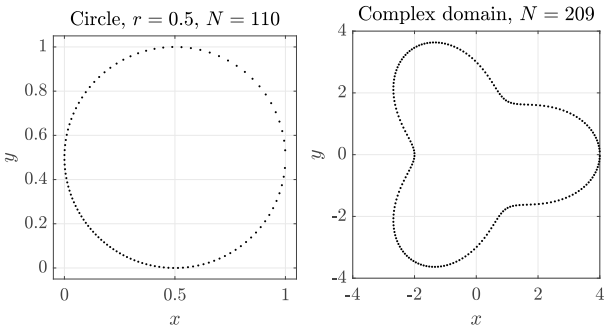


Figure 3. Circle from (9) sampled with $h(x, y) = 0.02(x^{1.5} + y^{1.5}) + 0.015$ and a complex curve in 2D from (10) sampled with $h = 0.09$.

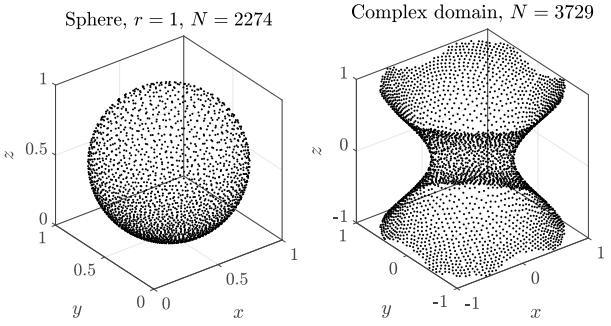


Figure 4. Sphere from (11) sampled with $h(x, y, z) = 0.033(x^{1.5} + y^{1.5} + z^{1.5}) + 0.025$ and a complex surface in 3D from (12) sampled with $h = 0.06$.

Once again we see that the algorithm produces locally regularly spaced points with smooth transitions in density.

IV. Parallel implementation

A. The idea

Point sampling algorithm, as described in Section II, lends itself nicely to shared-memory parallelization. Our basic idea of parallelization was to let multiple threads sample in the domain with candidate points, starting from multiple spatially distant starting points. This idea requires no communication between the threads but requires that two global data structures be shared by all the threads - a list of existing points and a spatial search structure that indexes the list. Since there is no explicit communication between the threads, this idea seems to allow for very good parallel efficiency as long as the common structures can also be shared with minimal synchronization overhead.

To maximize the efficiency, shared access to the spacial search structure has to be efficiently managed. While reads from structure can be processed concurrently, writes require exclusive access. The point sampling algorithm creates a lot more candidate points than final points that are inserted into the domain. Thus, many more reads are required, which can be concurrent, than writes, which have to be synchronized.

There are many options to creating parallel spatial search structure with efficient shared-memory access. We settle for a search structure, which is essentially a k-d tree of k-d trees. The first level k-d tree is initialized in the

first step and is immutable afterwards. The second level k-d trees are rooted to leaves of the first level k-d tree and are grown in real-time by the worker threads. This structure can also be viewed as a single k-d tree, with an immutable part, which is inherently thread-safe, and many mutable sub-trees, which have to be individually protected for concurrent use by multiple threads. Such a spatial search structure promises near zero parallelization overhead if the threads can be bound to work on separate sub-trees.

B. Description of the algorithm

Within the first step, which is done sequentially, the first level k-d tree is generated. The first level requires the generation of seed points that are evenly distributed across the target domain. Seed point density can be orders of magnitude lower than the final point density, since only as many seeds as there are worker threads is required. Therefore, sequential point sampling as described in Section II can be used to perform this step. With the low target density, its execution time will be orders of magnitude shorter tha final point sampling execution time would be. Spacing function for this step is created by using the scaling up the final spacing function by a constant factor. For now, the exact factor value can be thought of as a user parameter of the parallel algorithm.

The second step is done in parallel by N_t threads, where $N_t \leq N_s$. Each thread gets a set of N_s/N_t seed points, rounded up or down as necessary to distribute all the seed points among threads. Then all the threads start the same algorithm that was described in Section II but using the parallel spatial search structure described in previous Subsection. When performing the search for the nearest neighbour, the thread must perform the following:

1. Descend the first level k-d tree until the nearest seed point to the search query is found. Since this part of the spatial search structure is immutable, no synchronization is required between threads.
2. Lock the second level k-d tree, which is rooted in the found seed point, for reading. This lock allows other threads to start concurrent reads in the same tree but puts all attempts for modifications on wait.
3. Descend the second level of the spatial search structure to obtain query results.
4. Remove the read-lock.

Insertion of candidate points to the tree is similar but uses write lock instead, which blocks all other threads from accessing the same tree. Since threads start sampling the domain from spatially distinct starting points, they tend to modify distinct second level trees and thus work in relative isolation, with minimal parallel overheads. Only when the areas sampled by different threads start to converge, the parallel overhead starts to rise somewhat. Even then it is usually limited to only two threads competing for the same resource, and is not capable of severely compromising parallel scalability of the algorithm.

Each thread finishes when its queue of candidate points is exhausted, and the parallel algorithm finishes when all the threads finish.

The distribution of points that was obtained with the parallel algorithm will differ from that of the sequential algorithm. In our preliminary experiments with the generated points, we could successfully use them as point sets for meshless methods when performing physical simulations. As was expected, we experienced no problems, since there are no special requirements for the probabilistic properties of the nodal distribution in this use case.

C. Experimental results

To estimate the efficiency and scalability of the represented parallel point sampling algorithm, a set of experiments is performed. The complex 3D domain introduced in Figure 2 is selected as the target domain, while the number of worker threads and the density of sampling are varied. Then execution time and speedup as functions of the varied parameters are calculated.

The platform where experiments were performed is a 16-core, 32-thread AMD TR 2950X processor with 32 GB RAM, running Ubuntu 19.04. CPU implements a two way simultaneous multithreading (SMT), which enables it to execute 2 threads on a single physical core simultaneously, albeit at lower efficiency as if the same two threads were executed on 2 physical cores. Threads are mapped to cores in such a way that the first 16 threads are mapped to separate cores, and threads from 17 to 32 are mapped as SMT threads. Therefore, if using up to 16 threads, SMT will not be utilized at all.

The results of experiments are summarized in Figure 5. Main finding is that the algorithm scales linearly on the experimental computer as long as only physical cores are utilized. The SMT is in general beneficial, and enables the speedups to continue rising sub-linearly, up to the maximum speedup of 22 on 32 threads. For large problem sizes and on number of threads roughly between 8 and 16, speedups are visibly super-linear, which means that the parallel efficiency is above 1. This is likely due to the program having access to increasingly more memory caches, when increasingly many CPU cores are utilized. Also worth noting is that the execution times are very low, e.g. below 100 seconds for the maximal problem size of 10 million points on 8 threads.

Speedups also depend on the computational complexity of the problem instance and fall roughly in two classes. In terms of the execution time, the divide between classes is at around 0.1 second. Adding more resources to a program that finishes under a tenth of a second produces modest gains at best, but can also prolong execution time in worst case. But if the program is taking longer than 0.1 s, adding more threads in general helps shorten execution time. If problem size is expressed as the number of points, then the class divide lies between 20000 and 50000 points. For 20000 points or less, the speedups are limited, and using SMT is contra-productive. For 50000 points or more, the behavior is not monotonous. First there is a dip in speedup when the first few SMT threads are added (just after 16 threads), but after that, the speedup goes up again. Maximal speedup is then achieved at 32 threads. Therefore, if SMT is fully utilized and the problem size is sufficient, then SMT is beneficial in the presented example.

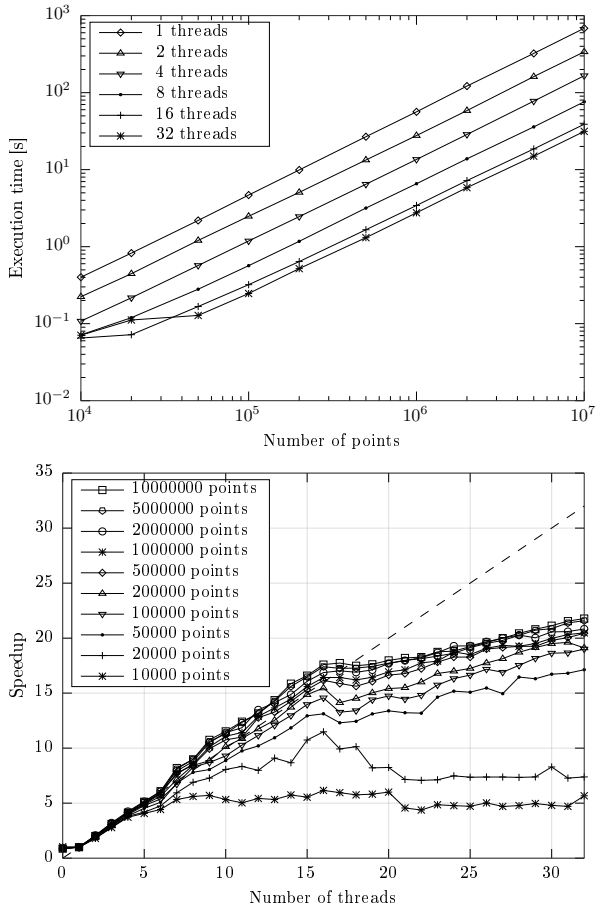


Figure 5. Results of the parallel efficiency tests; raw execution times on top and speedups below.

V. Conclusion

An efficient parallel algorithm for generating point sets on boundaries and in interiors of 3D bodies was presented. The theoretical complexity and execution time of the algorithm were demonstrated, as well as the speedups from parallel execution.

Future work will focus on further developing and analyzing the surface sampling algorithm. Additionally, the parallel algorithm needs to be investigated further in terms of impact on quality of the node distribution.

The algorithm can also be generalized to support surfaces based on splines and, in turn, general inputs from computer-aided design (CAD) programs.

Acknowledgments

The authors would like to acknowledge the financial support of the ARRS research core funding No. P2-0095 and the Young Researcher program PR-08346.

References

[1] Robert L. Cook. Stochastic sampling in computer graphics. *ACM Transactions on Graphics (TOG)*, 5(1):51–72, 1986.

[2] M. Alexa, S. Rusinkiewicz, Diego Nehab, and Philip Shilane. Stratified point sampling of 3D models. In *Proc. Eurographics Symp. on Point-Based Graphics*, pages 49–56, 2004.

[3] Varun Shankar, Robert M. Kirby, and Aaron L. Fogelson. Robust node generation for mesh-free discretizations on irregular domains and surfaces. *SIAM Journal on Scientific Computing*, 40(4):A2584–A2608, 2018.

[4] Jure Slak and Gregor Kosec. On generation of node distributions for meshless PDE discretizations. *SIAM Journal on Scientific Computing*, 41(5):A3202–A3229, October 2019.

[5] D. P. Hardin and E. B. Saff. Discretizing manifolds via minimum energy points. *Notices of the AMS*, 51(10):1186–1194, 2004.

[6] Douglas P. Hardin, Edward B. Saff, and J. Tyler Whitehouse. Quasi-uniformity of minimal weighted energy points on compact metric spaces. *Journal of Complexity*, 28(2):177–191, 2012.

[7] Y. Choi and S. Kim. Node generation scheme for meshfree method by Voronoi diagram and weighted bubble packing. In *Fifth us national congress on computational mechanics, Boulder, CO*, 1999.

[8] Rainald Löhner and Eugenio Oñate. A general advancing front technique for filling space with arbitrary objects. *Int. J. Numer. Methods Eng.*, 61(12):1977–1991, 2004.

[9] Xiang-Yang Li, Shang-Hua Teng, and Alper Ungor. Point placement for meshless methods using sphere packing and advancing front methods. In *ICCES'00, Los Angeles, CA*. Citeseer, 2000.

[10] Li-Yi Wei. Parallel Poisson disk sampling. *ACM Transactions on Graphics (TOG)*, 27(3):1–9, 2008.

[11] Xiang Ying, Shi-Qing Xin, Qian Sun, and Ying He. An intrinsic algorithm for parallel poisson disk sampling on arbitrary surfaces. *IEEE Transactions on Visualization and Computer Graphics*, 19:1425–1437, 2013.

[12] Andrew W. Moore. An introductory tutorial on kd-trees. phdthesis 209, Computer Laboratory, University of Cambridge, 1991.

[13] Jose Luis Blanco and Pranjal Kumar Rai. nanoflann: a C++ header-only fork of FLANN, a library for nearest neighbor (NN) with KD-trees, 2014.

Sarajevo War Tunnel - Mobile Virtual Cultural Heritage Application

I. Prazina*, I. Ivkovic-Kihic*, T. A. Chahin*, A. Jajcanin*, S. Rizvic* and V. Okanovic*
*Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina
iprazina1@etf.unsa.ba, iivkovic2@etf.unsa.ba, tchahin1@etf.unsa.ba, ajajcanin1@etf.unsa.ba,
srizvic@etf.unsa.ba and vokanovic@etf.unsa.ba

***Abstract*—This paper explores the new way of presenting one existing VR application, which was described in our previous work - Virtual Reality Experience of Sarajevo War Heritage. The goal of the application was to introduce more people with the Sarajevo siege and allow them to experience the Tunnel crossing at that time. Before this application, we made two versions, the first one for VR setup and the second for the web. In this paper, we introduce a mobile version with the same content. The challenge was to optimize the content for the mobile experience. The assets were optimized so a wider number of mobile phones with different hardware capabilities can run the application. The advantages and disadvantages of this approach are pointed out, and the limitations of the mobile application are emphasized. The memory usage and frame rate are measured for different Android devices with different operating system versions and hardware capabilities. The results show the optimized application can be run on different Android mobile devices. Nevertheless, for better user experience a higher number of frames per second is needed, which may include reducing the quality of the assets.**

***Keywords*—mobile application; cultural heritage; virtual reality**

I. Introduction

Digital technologies offer the users to travel virtually, as well as to be present and interact in a different space and time. In that sense, they can be very useful for the presentation of cultural heritage that cannot be experienced, visited, or explored anymore in reality. One of those is the Sarajevo War Tunnel which saved the city of Sarajevo during the siege. It was crucial for the defense of the city during the war in the 1990s. The citizens were without water, electricity, heating and the Tunnel was the only way through which food, medical supplies, and weapons were delivered. Today the Tunnel is closed and only 20m of this object is open for public within the Tunnel museum.

This paper presents a mobile application with interactive digital storytelling about the historical content of Sarajevo siege and details on the Tunnel construction. The application enables the users to experience the Tunnel crossing at that time. The application teaches the users about historical events in an attractive way and the conditioning the access to the mobile simulation by passing the quiz. The quiz evaluates the knowledge gained through stories, thus the application has an educational aspect.

The application was primarily made for the VR setup. In the paper [1], the VR solution was introduced first, and then the process of developing the application for the Android platform was described in detail. In the following, the paper also presents measured average FPS and average memory usage for different Android devices.

The motivation behind the creation of the mobile application was to make the content available to a wider number of users. VR technology is expensive and less practical for usage, while smartphones are used by big variety of users. Since the purpose of the application is primary education and promotion of cultural heritage, it is of great importance to make it available and affordable, even though some of the features are lost this way.

The paper is structured as follows. In Chapter II related work is presented. Two solutions, the VR solution, and the mobile application are described in Chapter III. The development process and performance evaluation are given in Chapter III. The conclusion is presented in Chapter IV.

II. Related works

Using virtual reality (VR) in museum exhibitions is lately very common thing. Authors of paper [2] describe the effects of VR on one London exhibition visitors. Authors notice that visitors are more engaged and active when the VR setup is used. Participants, in this case, felt more involved in the exhibition. Nevertheless, some participants felt pressure from the fact that other visitors are waiting for them to finish.

In the paper [3] authors describe the technical solution of one virtual system where multiple users are present. In the solution, users control objects and environment and see their virtual avatar.

Authors in paper [4] present their solution which uses 3D game concepts in a virtual museum. One of the conclusions from this paper is that the user's immersion and satisfaction depends on the PC setup hardware capabilities.

In the paper [5] authors describe different aspects of serious games in cultural heritage. They state that serious games provide a high level of empathy with a game character which can enhance learning and understanding of historical events, characters and scenes.

Different factors of success for serious games are described in the paper [6]. Some of the factors are story quality and balance between game task difficulty and the educational goals of these tasks.

Some of the applications use anthropomorphic virtual guides in virtual historic scenes, as in paper [7]. The authors

present one of the advantages of their application is availability. The application is available even when visitors leave the museum because the application is downloadable from the web. VR applications have a limitation that they are usually only available in museums and require VR equipment.

In the paper [8] authors present a mobile application that, as they state, improves interaction and cooperation between users in storytelling and placemaking of cultural heritage.

Mobile applications in cultural heritage are often used as trail narrating and guiding as in papers [9], [10] and [11]. We have not found any paper describing the process of porting the application from the VR to the mobile experience.

III. VR TO MOBILE

A. The VR solution

In [1] authors described the Sarajevo War Tunnel VR project, a virtual cultural heritage application combining VR storytelling about the Sarajevo siege with the VR simulation of the tunnel crossing. It was concluded that VR technology can be useful in preserving intangible cultural heritage. The immersion that can be gained through such applications is beneficial for user experience. Even though it is said that VR is an attractive and efficient way of presenting cultural heritage. It has limitations in the term of availability and cost since the VR equipment is expensive and not widely available.

B. The mobile app

Using a mobile application in presenting intangible cultural heritage can be beneficial in overcoming previously mentioned limitations since mobile phones are commonly used amongst people of all ages. This comes with the cost of losing immersion and ease of use. Also, mobile as a platform is a challenge for deploying these kinds of applications in the means of efficiency. In this paper, the mobile application with the same content is presented. Users first have to watch to the 360° videos containing stories about the siege, shown in figure 1a, than they are taking the quiz meant to prove that they learned the basics about the topic, and then as a reward, the user gains access to the experience of tunnel crossing, shown in figures 1b and 1c.

C. The development process

All the models have been created manually in Blender. Because of the large number of logs used in building the model and the original Tunnel we had to do something to reduce the size and at the same time improve the performance of the simulation. To achieve that we reduced the number of vertices in the Blender model using the decimate modifier and increasing the number of un-subdivide iterations before importing it into Unity. Also wherever it was possible, instead of creating details on the models by hand which would increase the number of vertices, we used PBR (Physically Based Rendering) textures that gave the model also a more realistic look.

[1]The profiler did not display memory usage, this value is from the application settings

The major difference when changing the platform from a VR based simulation to an Android one is the movement. How to give the user a more natural movement feel throughout the app. In the VR based platform the user has a headset to move the camera around, and joysticks in both hands with triggers, but on an Android platform it's all about fingers. So we enabled the user to move the camera around by swiping his fingers around the phone screen. Also, we added a joystick and a jump button to give as much of a closer feeling in using the app on the phone, to the one on the VR platform as possible. That involved writing new movement scripts.

One of the biggest goals in our app is for the users to learn about history, and that wouldn't be possible if there were no videos included. But having videos in the project brings huge trouble, especially since our videos are 360° So to reduce the size of the project we decided to stream the videos, which reduced the size to almost 15% of the initial size and for what we wrote a new script that handles the streaming. The script creates a Video Player component and fetches a few frames from the provided URL now and then and loads the component with these frames. The component starts playing as soon as the first frames are loaded.

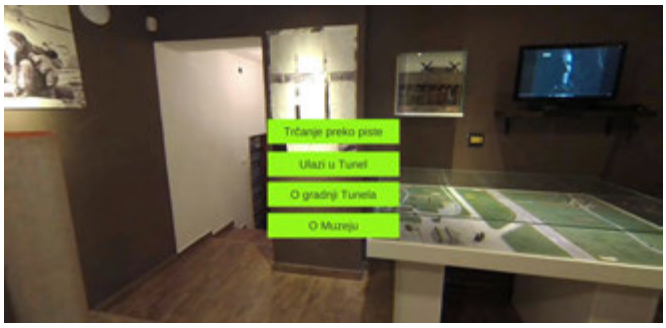
The final step in building the app was choosing the correct build settings since the options that you choose in this step could also improve the overall app performance. Since our app has a lot of graphics involved, the key setting we focused on was the graphics rendering API. There is a variety of APIs like Vulkan, OpenGL, DirectX, Metal, etc. But for our app, the best fit was the OpenGL ES 2.0 since it is a cross-platform software and is known to be working on most of the devices around.

D. The evaluation on different mobile devices

When developing this app one of the goals was to make it available for more people. We reduced its size and used commonly used libraries so it can work on different Android devices. To ensure that after the optimizations it can be used on mobile devices we profiled the application. The application is profiled in Snapdragon and Adreno profiler on Android devices with different operating system versions, and different hardware capabilities. The most challenging part of the application is the virtual walk through the tunnel. This part can overwhelm the devices GPU and use most of the RAM. It is mainly because of the high number of objects and vertices.

In the Snapdragon profiling tool (and Adreno for older devices) we measured average FPS and average memory usage. The average FPS is calculated only for the virtual walk part, for other parts of application FPS is above 30 even for older devices. For an older device with less than 2GB of available RAM, the average FPS was around 3. For newer devices with more than 2GB of available RAM, the average FPS was between 8 and 12 (Table I).

The application is also run on Android phones with Mali GPU. Snapdragon profiling tools can not measure FPS on those devices. Even though we could not profile applications on those devices perceived performance was similar.



(a) Screenshot I



(b) Screenshot II



(c) Screenshot III

Fig. 1. The mobile application interface

TABLE I
THE EVALUATION OF FPS AND MEMORY USAGE

Device	FPS	Memory usage	OS	Total memory	GPU
Samsung A6+	8.7	2.26GB	Android 9	3GB	Adreno 506
Motorola Moto X Pure	11	2.3GB	Android 7.2	3GB	Adreno 418
Samsung J5 (2016)	3	600MB[1]	Android 7.1.1	2GB	Adreno 306
Redmi Note 5 AI	11.28	2.8GB	Android 8.0	3GB	Adreno 509

IV. CONCLUSION

The VR application presented earlier in this paper was made for users to perceive and understand the feelings and difficulties of people who used the war tunnel. Even though this application can achieve these goals it can only be used on specialized equipment that is not available for most users. To solve this problem a mobile application with similar capabilities was made. The development of the mobile application was not trivial. There were problems with 3D models optimization, 360 video streaming, and libraries' support on mobile devices. The approach we used for the transition of VR application to mobile was replacing VR related assets with the ones supported by mobile phones. Also, the tunnel model and animations had to be optimized. The main problems are solved, the application is available for more users. Even though it can be installed on most android devices, there are still open problems as low FPS and nonintuitive controls. User evaluation is not performed at this time although it is planed when open problems are solved.

REFERENCES

[1] S. Rizvic, D. Boskovic, V. Okanovic, I. Ivkovic-Kihic, S. Sljivo, "Virtual Reality Experience of Sarajevo War Heritage", EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2019).

[2] E. Parker and M. Saker, "Art museums and the incorporation of virtual reality: Examining the impact of VR on spatial and social norms," Convergence: The International Journal of Research into New Media Technologies, p. 1354856519897251, Jan. 2020.

[3] L. Nisiotis, L. Alboul, and M. Beer, "Virtual Museums as a New Type of Cyber-Physical-Social System," in Lecture Notes in Computer Science, Springer International Publishing, 2019, pp. 256–263.

[4] G. Lepouras and C. Vassilakis, "Virtual museums for all: employing game technology for edutainment," Virtual Reality, vol. 8, no. 2, pp. 96–106, Jun. 2004, doi: 10.1007/s10055-004-0141-1.

[5] M. Mortara, C. E. Catalano, F. Bellotti, G. Fiucci, M. Houry-Panchetti, and P. Petridis, "Learning cultural heritage by serious games," Journal of Cultural Heritage, vol. 15, no. 3, pp. 318–325, May 2014.

[6] W. S. Ravyse, A. Seugnet Blignaut, V. Leendertz, and A. Woolner, "Success factors for serious games to enhance learning: a systematic review," Virtual Reality, vol. 21, no. 1, pp. 31–58, Sep. 2016.

[7] V. Lombardo and R. Damiano, "Storytelling on mobile devices for cultural heritage," New Review of Hypermedia and Multimedia, vol. 18, no. 1–2, pp. 11–35, Mar. 2012.

[8] K. Han, P. C. Shih, M. B. Rosson, and J. M. Carroll, "Enhancing community awareness of and participation in local heritage with a mobile

application," in Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14, 2014.
[9] M. Roussou and A. Katifori, "Flow, Staging, Wayfinding, Personalization: Evaluating User Experience with Mobile Museum Narratives," Multimodal Technologies and Interaction, vol. 2, no. 2, p. 32, Jun. 2018, doi: 10.3390/mti2020032.
[10] N. Basaraba, O. Conlan, J. Edmond, and P. Arnds, "Digital narrative conventions in heritage trail mobile apps," New Review of Hypermedia and Multimedia, vol. 25, no. 1–2, pp. 1–30, Apr. 2019, doi: 10.1080/13614568.2019.1642963.
[11] A. Rolando A. Scandiffio "Mobile Applications as Tool for Exploiting Cultural Heritage in the Region of Turin and Milan" vol. XL-5/W2 pp. 525-529 2013.

Modeling Agricultural Production Activities Using Weather and Soil Parameters

Tomislav Kovačević*, Lovre Mrčela†, Andro Merćep‡ and Zvonko Kostanjčar§
Department of Electronic Systems and Information Processing
University of Zagreb
Email: *tomislav.kovacevic@fer.hr, †lovre.mrcela@fer.hr, ‡andro.mercep@fer.hr, §zvonko.kostanjcar@fer.hr

***Abstract*—Activities concerning agricultural production processes are highly influenced by weather and soil parameters. Such parameters determine when certain activities in the agricultural process should be carried out. For example, it is known that fertilization should be applied when humidity of the soil is appropriate for fertilizer to get worked into it, as too dry soil causes fertilizer to be washed away. Furthermore, pest and disease control is mostly affected by wind direction and speed, which may cause pesticide to be overly spread or concentrated in some spot on the field. There is well known expert knowledge about connection between those parameters and scheduling production process activities. However, expert knowledge can be enriched and automatized with data-driven models, which are relatively unknown due to lack of research in this field.**

We propose a Bayesian logistic regression model in order to predict probability of conducting field activities, based on weather and soil parameters. We have trained and validated our model on historical data which contains activities during potato production in Croatia and weather station measurements gathered from several fields over a period of one season.

***Index Terms*—Bayesian logistic regression, weather, soil, production process, pesticide, fertilization, potato**

I. Introduction

Pressure on the agricultural system is growing as the human population and food consumption continue to expand [1]. The decline in land fertility prompts the need for increased crop yield and rational usage of chemical fertilizers, pesticides and herbicides due to high production costs and huge environmental impact. Integration of new technologies with expert knowledge leads to precision agriculture, which is also known as site specific crop management. It allows us to optimize production process by applying the right amount of fertilizers and pesticides at the right time, according to weather and soil parameters. Recent development of Internet of Things drastically increased the amount of available data, which is suitable for modeling field activities. Climate and soil parameters strongly influence activities related to agricultural production processes. For example, proper amount of rainfall can be beneficial to dissolve dry fertilizer and move nutrients into the soil rooting zone, but excessive rain can increase runoff potential. Also, there is a short window of opportunity to safely and efficiently control pests and diseases in agricultural crops. That timeframe may become even shorter under unsuitable weather conditions. In this paper, we propose a Bayesian logistic regression model which predicts optimal point in time for such activities. We have also compared Bayesian logistic regression model with standard logistic regression model.

II. Importance of weather and soil parameters

Warm temperatures cause rapid plant growth above and below ground, leading to root expansion which improves nutrient uptake. It also highly affects soil temperature, which has already been shown in several studies [2]–[4]. Soil temperature is one of the crucial parameters that affect chemical and biological processes within the soil. It affects nitrogen release from organic N sources [5], [6]. As it has been shown in [5], increasing temperature enhanced N release from all N sources that were considered in study, but magnitude of the response to temperature varied between sources, as well as most appropriate temperature interval. It has been discussed in [7] that soil temperature also has great effect on pesticide degradation. They showed that the decrease in soil temperature from 24 °C to 10 °C lead to increase of 2,4-D pesticide half-lifes in both experimental soil types. Due to a wide spreading area of fertilization spreaders and pesticide sprayers, wind can cause deviation in anticipated particle trajectories, resulting in over-application and under-application of nutrient, thereby leading to undesired financial and ecological effect [8]. Modeling ballistic flight of fertilizer particles showed that wind speed of only 3 m/s can already affect the landing position of fertilizer particles [9], [10]. Although it is stated in [6] that soil temperature does not effect nitrogen release from artificial NPK fertilizers, it is important to note that precipitation has a greater impact on chemical fertilizers. The same goes for pesticide usage, huge amounts of rain may cause drift of the particles of both chemical fertilizer as well as pesticide into ground waters, causing huge environmental impact and losing the key nutrients for the crop. Researches in [11] suggested that replacing chemical fertilizers with organic or slow-release ones may drastically reduce N and P losses and runoff which decreases ecological impact on ground waters.

III. Similar work

In [12] authors developed a model that predicts suitable days for conducting field additives from long-term weather and soil records for different locations using the Integrated Farm System Model. This model uses simulations that integrate many biological and physical processes on crops. Suitable day component is predicted on daily time step based on soil

moisture. Although it was pointed out that model is usable for different locations and different soil types, authors did not mention whether and how they validated the model. In the [13] authors researched impact of planting date on maize and winter wheat yield, and benchmarked methods for determining the planting date. They proposed novel models that considered not only average daily temperature, but also precipitation amount, soil temperature, and also soil water content with historical data from 5 to 10 days in order to determine maize and winter wheat planting dates. The results indicated that those parameters, as well as climate change, had a great impact on planting date and that it should be specific to crop and site. Simulations showed that variable planting dates results in greater yield, when compared with fixed planting dates. A new method for predicting the readiness of a soil and field for tillage was proposed in [14]. Field readiness was defined as combination of field workability and field trafficability, where workability is positive if stress propagation withing the soil at 50 cm soil depth is below 50 kPa, false otherwise, while trafficability is positive if the soil water content is between some predefined upper and lower bounds. If both workability and traficability were positive, field is ready for tillage. The authors compared newly proposed method to similar ones on conceptual basis, without any reported error rate pointed out. The decision support tool is made to improve informed choices. Data driven decision support system was proposed in [15]. The model predicts days suitable for field activities, particularly days suitable for post-planting nitrogen applications. Authors obtained root mean squared error of 1 day per week using weather parameters, soil parameters, and crop growth stage. Similar regression models for different crops and activities were developed in [16], [17]. Another research [18] was conducted in order to quantify machinery and production costs considering local weather and soil parameters for predicting probability of conducting field activities. As it turned out, harvesting dates of miscanthus should be optimized and corrected, but investments in simulated production costs are also increasing by 5% for the machinery. The most similar study was conducted in [19], where authors also used logistic regression in order to predict probability of conducting field activities. Rather than evaluating their error in predicting activities, the authors monitored soil moisture threshold for the specific location and its impact on planting date. It was shown that soil moisture has a great impact on planting date and that planting dates should be reevaluated for specific sites and climates.

IV. Methodology

Similarly to [19], we modeled the probability of fertilizing and protection (not distinguishing between pesticide and herbicide usage) activities which depend on weather and soil parameters. As we did not research impact of the right conditions of fertilizer and protection on final yield, an expert chose farmers with good practice during potato crop season. Chosen farmers were supplied with weather stations that measure weather and soil parameters on their fields, and their activities are logged in the Agrivi agricultural management software. This allowed us to gather a training dataset that contains weather and soil parameters as input features $\boldsymbol{x}$, and activity that the farmer did (or did not) conduct as the target variable y for each input vector $\boldsymbol{x}$. Because of different input features, two different models were proposed – one for the fertilization activity and the other one for protection activity. The probability of conducting activity was modeled using both standard logistic regression as well as Bayesian logistic regression. A brief introduction of those models is given below.

A. Logistic regression

Logistic regression models the posterior probability of a certain class or, in our case event. From a machine learning standpoint, logistic regression is used for binary classification problems. Given an example $\boldsymbol{x} \in \mathbb{R}^n$ with features x_i, logistic regression models the posterior probability of positive class $\mathcal{C}_1$ as

$$p\left(\mathcal{C}_1|\boldsymbol{x}\right) = h\left(\boldsymbol{x}|\beta_0, \boldsymbol{\beta}\right) = sigm\left(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}\right), \tag{1}$$

where $\boldsymbol{\beta}$ and β_0 represent weights and bias term (i.e. model parameters), and $sigm(z)$ denotes the sigmoid or logistic function defined as

$$sigm\left(z\right) = \frac{1}{1+e^{-z}}. \tag{2}$$

Assuming that the example vector $\boldsymbol{x}$ contains a fixed dummy feature x_0 = 1, (1) simplfies to $p(\mathcal{C}_1|\boldsymbol{x}) = sigm\left(\boldsymbol{\beta}^\top \boldsymbol{x}\right)$, where $\boldsymbol{\beta}$ is a vector of parameters $\boldsymbol{\beta} = \left[\beta_0\ \beta_1\ ...\ \beta_n\right]^\top$. For a dataset $\mathbb{D} = \left\{\boldsymbol{x^{(i)}}, y^{(i)}\right\}_{i=1}^{N}$, where $y_i \in \{0, 1\}$, the likelihood function can be written as

$$\boldsymbol{\mathcal{L}}\left(\boldsymbol{\beta}\right) = \prod_{i=1}^{N} h\left(\boldsymbol{x^{(i)}}\middle|\boldsymbol{\beta}\right)^{y^{(i)}} \cdot \left[1 - h\left(\boldsymbol{x^{(i)}}\middle|\boldsymbol{\beta}\right)\right]^{1-y^{(i)}}. \tag{3}$$

The error function can be defined as a negative logarithm of the likelihood, which results in cross-entropy function defined as

$$\begin{aligned} -\log \boldsymbol{\mathcal{L}}\left(\boldsymbol{\beta}\right) = -\sum_{i=1}^{N} \Big\{ & y^{(i)} \ln h\left(\boldsymbol{x^{(i)}}\middle|\boldsymbol{\beta}\right) \\ & + \left(1 - y^{(i)}\right) \ln\left(1 - h\left(\boldsymbol{x^{(i)}}\middle|\boldsymbol{\beta}\right)\right)\Big\}. \end{aligned} \tag{4}$$

Optimal parameters $\boldsymbol{\beta}$ are obtained by minimizing the error function (5) using iterative optimization methods, such gradient descent [20].

B. Bayesian logistic regression

The main difference between standard and Bayesian logistic regression models is that we assume prior distributions of $\boldsymbol{\beta}$ in the Bayesian model. By doing inference, we can calculate posterior distribution of parameters. This allows us to sample from parameters posterior distribution in order to obtain uncertainty of our model output. The posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{x})$ is defined as

$$p\left(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{y}\right) = p\left(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\beta}\right) \cdot p\left(\boldsymbol{\beta}\right), \tag{5}$$

where $p(\boldsymbol{x}, y|\boldsymbol{\beta})$ denotes the likelihood, and $p(\boldsymbol{\beta})$ is the prior distribution. There are two types of prior distributions for the parameters $\boldsymbol{\beta}$ – informative and non-informative. Informative ones, when something is known about the values of the parameters and non-informative ones, when there is no knowledge about values of the parameters or if we'd like to ensure that prior distribution doesn't play role in the analysis. For this study, we assumed a multivariate non-informative Gaussian prior for the parameters $\boldsymbol{\beta}$:

$$\beta_j \sim \mathcal{N}(\mu_j, \sigma_j). \tag{6}$$

For the non-informative Gaussian prior, μ_j is usually zero, while σ_j is set to 1000. Likelihood is same as (3), so we model it as

$$p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{N} \left[sigm(\boldsymbol{\beta})^{y^{(i)}} \cdot (1 - sigm(\boldsymbol{\beta}))^{1-y^{(i)}} \right] \tag{7}$$

Posterior distribution can be then derived as

$$p(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{N} \left[sigm(\boldsymbol{\beta})^{y^{(i)}} \cdot (1 - sigm(\boldsymbol{\beta}))^{1-y^{(i)}} \right] \cdot \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_j}} \exp -\frac{1}{2} \left(\frac{\beta_j - \mu_j}{\sigma_j} \right)^2 \tag{8}$$

Marginal posterior distributions of parameters $\boldsymbol{\beta}$ are hard to solve, therefore we use random sampling methods such as Metropolis sampler in order to solve it, or to get an estimate of the marginal posterior distribution. For more information on the sampling method and the marginal posterior distribution, see [21].

In order to make a prediction for a new input example, we can sample from parameters posterior distribution. Thus, for each output y^{new}, we have distribution for the output. Using that, we can model conditional probability of the output and also quantify the risk of model output.

C. Model selection

The data was split in train and test subsets. Train set contained 75% of overall examples, while the rest was used for testing. Since the dataset was highly imbalanced, we used stratified train-test splitting meaning that approximately the same percentage of samples of each target class was in both subsets. For the standard logistic regression we performed 5-fold cross-validation method on train set in order to obtain best hyper parameters.

V. The data

A. Acquired data

On-site measurements were conducted using Pessl Instruments weather stations during period from February 2019 to June 2019 in Bjelovarsko-Bilogorska county of Croatia. Potato crop production activities were recorded on 24 different fields in the same broader area between 45°40'17.23" N, 17°5'51.50" E and 45°32'51.23" N, 16°55'33.40" E. Although weather stations provide more parameters, we used only those that were analyzed for their impact on fertilization and pesticide application (see Section II). Those parameters are also available from online weather application programming interfaces. That way, models can be used for different locations from those that were trained on. The obtained data are summarised in Table I. The data scale is the daily average of hourly measurements. Overall number of data points was 2640.

B. Preprocessing

The main problem we had to deal with was imbalanced data, since the number of fertilizing field activities during previously mentioned period was 64 which is less than 3% of all dataset. Fertilizing activity was hardly separable in feature space, so we defined a new feature that generally discards the possibility of conducting fertilizing activity if it was done in recent time. For the pesticide application, number of times activity was conducted was 660 or 25% of all dataset, which makes it more suitable for model training. Data normalization was done using standard scaling by subtracting feature mean value for each example and dividing the difference with feature standard deviation. For Bayesian logistic regression we additionally under-sampled the majority class in training set using cluster based under-sampling in order to obtain balanced dataset, as for us it showed better results on test set. More about cluster based under-sampling can be found in [22].

TABLE I
STATISTICAL DATA OF WEATHER AND SOIL PARAMETERS THAT WERE USED FOR MODEL TRAINING AND EVALUATION.

	Air temperature [°C]	Soil temperature [°C]	Wind speed [m/s]	Precipitation [mm]
Mean	12.63	11.34	2.74	2.59
St. dev.	5.14	2.36	1.20	5.01
Minimum	-0.44	9.47	0.00	0.00
Maximum	26.01	19.99	8.39	49.32

VI. Results and discussion

Scores obtained for the fertilization activity are reported in Table II. If we compare F1 measure and AUC (Area Under Receiver Operating Curve) for both standard and Bayesian logistic regression, we can conclude that there exists a decision threshold that could yield a higher F1 score, rather than the standard one of 0.5. Since all metrics are relatively high on the test set, we can conclude that the model is well suited for the problem and has good generalization properties for fertilization activity. If we compare standard and Bayesian logistic regressions in expectation, they are very similar. Results for the protection activity are shown in Table III. Comparing these two models, the one for fertilization and the one for protection, we can conclude that adding the feature that removes the possibility of repeating the activity within the 10 days from the last occurrence had a significant impact on the overall performance. In order to improve results, more informative priors for Bayesian logistic regression could be considered in modeling the problem. As for the Bayesian logistic regression performance and uncertainty modeling, results on test field

TABLE II
SCORES ON TEST DATA FOR FERTILIZATION ACTIVITY.

	Logistic regression	Bayesian logistic regression
F1 score	0.77	0.78
AUC	0.81	0.80
Precision	0.73	0.79

TABLE III
SCORES ON TEST DATA FOR PESTICIDE APPLICATION ACTIVITY.

	Logistic regression	Bayesian logistic regression
F1 score	0.67	0.67
AUC	0.74	0.70
Precision	0.65	0.66

are displayed on Figures 1 and 3. We explained in Section IV-B that Bayesian logistic model parameters β are probability distributions. Marginal posterior distributions of parameters for both fertilization and protection model are displayed in Figures 2 and 4. Sampling from those distributions, we can obtain uncertainty for model output. Interval of one standard deviation from model output is displayed together with model output means. We can see that in different time steps model output has different standard deviation, which means that the model is more certain of its prediction in case of lower values of standard deviation.

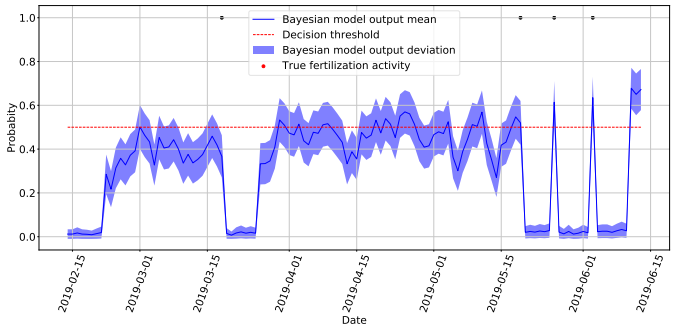


Fig. 1. Probability of conducting fertilization activity during potato production process on test site.

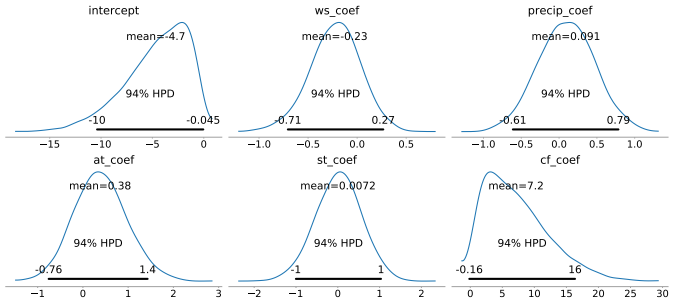


Fig. 2. Marginal posterior distribution of Bayesian logistic regression parameters β for fertilization activity.

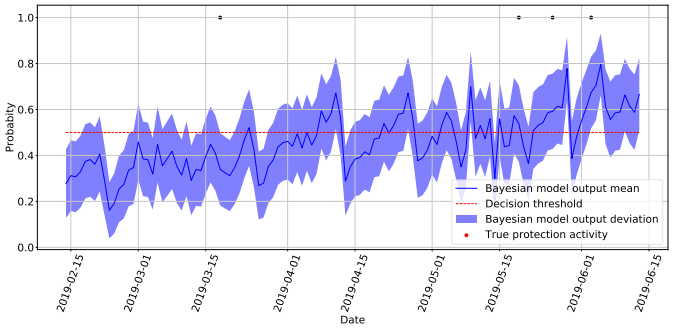


Fig. 3. Probability of conducting protection activity during potato production process on test site.

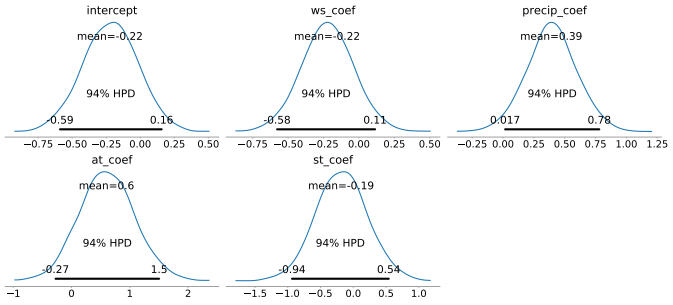


Fig. 4. Marginal posterior distribution of Bayesian logistic regression parameters β for protection activity.

VII. CONCLUSION

Weather and soil parameters highly impact agricultural production process activities. It is possible to model field activities conditioned on those parameters. In this paper we modeled probability of conducting field activity using standard and Bayesian logistic regression. Both standard and Bayesian logistic regression proved to have suitable capacity for modeling agricultural process filed activities. Features that discard the possibility of conducting some of the field activities can have significant impact in model performance. We compared scores from standard and Bayesian logistic regression; the obtained scores were similar between the models. Modeling agricultural production process with Bayesian logistic regression enabled us to model production process uncertainty. If we would use Bayesian logistic regression model in decision support system, we could weigh our decision by considering uncertainty, rather than a single number as an output of the model. Finally, uncertainty of model output can be used for quantification of production risk, which can in turn be used for either financial planning or modeling farmers' production process quality, i.e. for agricultural insurance.

ACKNOWLEDGMENT

This research has been fully supported by Operational programme Competitiveness and Cohesion 2014-2020 under the grant KK.01.2.1.01.0091. The authors would like to thank the reviewers for their comments which helped improve the quality of the paper.

REFERENCES

[1] K. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine Learning in Agriculture: A Review," *Sensors*, vol. 18, no. 8, p. 2674, aug 2018.

[2] V. R. Hasfurther, R. D. Burman, and . Nunn, John R. (John Robert), "Model for predicting soil temperature from air temperature, A," 1972.

[3] L. M. DWYER, H. N. HAYHOE, and J. L. B. CULLEY, "PREDICTION OF SOIL TEMPERATURE FROM AIR TEMPERATURE FOR ESTIMATING CORN EMERGENCE," *Canadian Journal of Plant Science*, vol. 70, no. 3, pp. 619–628, jul 1990.

[4] N. Persaud and A. C. Chang, "Estimating soil temperature by linear filtering of measured air temperature." *Soil Science Society of America Journal*, vol. 47, no. 5, pp. 841–847, 1983.

[5] S. Agehara and D. D. Warncke, "Soil moisture and temperature effects on nitrogen release from organic nitrogen sources," *Soil Science Society of America Journal*, vol. 69, no. 6, pp. 1844–1855, nov 2005.

[6] M. E. Engelsjord, O. Fostad, and B. R. Singh, "Effects of temperature on nutrient release from slow-release fertilizers - I. Commerical and Experimental products," *Nutrient Cycling in Agroecosystems*, vol. 46, no. 3, pp. 179–187, mar 1996.

[7] R. H. Veeh, W. P. Inskeep, and A. K. Camper, "Soil depth and temperature effects on microbial degradation of 2,4-D," pp. 5–12, 1996.

[8] S. R. Cool, J. G. Pieters, J. Van Acker, J. Van Den Bulcke, K. C. Mertens, D. R. Nuyttens, T. C. Van De Gucht, and J. Vangeyte, "Determining the effect of wind on the ballistic flight of fertiliser particles," *Biosystems Engineering*, vol. 151, pp. 425–434, nov 2016.

[9] H. T. Søgaard and P. Kierkegaard, "Yield Reduction Resulting from Uneven Fertilizer Distribution," *Transactions of the ASAE*, vol. 37, no. 6, pp. 1749–1752, 1994.

[10] D. A. Desmarteau, A. M. Ritter, P. Hendley, and M. W. Guevara, "Impact of Wind Speed and Direction and Key Meteorological Parameters on Potential Pesticide Drift Mass Loadings from Sequential Aerial Applications," *Integrated Environmental Assessment and Management*, p. ieam.4221, dec 2019.

[11] Z. Liu, J. Yang, Z. Yang, and J. Zou, "Effects of rainfall and fertilizer types on nitrogen and phosphorus concentrations in surface runoff from subtropical tea fields in Zhejiang, China," *Nutrient Cycling in Agroecosystems*, vol. 93, no. 3, pp. 297–307, jul 2012.

[12] C. A. Rotz and T. M. Harrigan, "PREDICTING SUITABLE DAYS FOR FIELD MACHINERY OPERATIONS IN A WHOLE FARM SIMULATION," *Applied Engineering in Agriculture*, vol. 21, no. 4, pp. 563–571, 2005.

[13] L. Dobor, Z. Barcza, T. Hlásny, T. Árendás, T. Spitkó, and N. Fodor, "Crop planting date matters: Estimation methods and effect on future yields," *Agricultural and Forest Meteorology*, vol. 223, pp. 103–115, jun 2016.

[14] G. Edwards, D. R. White, L. J. Munkholm, C. G. Sørensen, and M. Lamandé, "Modelling the readiness of soil for different methods of tillage," *Soil and Tillage Research*, vol. 155, pp. 339–350, jan 2016.

[15] B. M. Gramig, R. Massey, and S. D. Yun, "Nitrogen application decision-making under climate risk in the U.S. Corn Belt," *Climate Risk Management*, vol. 15, pp. 82–89, jan 2017.

[16] J. Aurbacher, P. S. Parker, G. A. Calberto Sánchez, J. Steinbach, E. Reinmuth, J. Ingwersen, and S. Dabbert, "Influence of climate change on short term management of field crops - A modelling approach," *Agricultural Systems*, vol. 119, pp. 44–57, jul 2013.

[17] L. H. d. S. Vieira, P. C. Sentelhas, and A. B. Pereira, "Agrometeorological and Soil Criteria for Defining Workable Days for Rational Mechanized Sugarcane Harvest in Southern Brazil," *Agricultural Sciences*, vol. 10, no. 05, pp. 597–621, may 2019.

[18] Y. N. Shastri, A. C. Hansen, L. F. Rodríguez, and K. Ting, "Impact of probability of working day on planning and operation of biomass feedstock production systems," *Biofuels, Bioproducts and Biorefining*, vol. 6, no. 3, pp. 281–291, may 2012.

[19] B. J. Tomasek, M. M. Williams, and A. S. Davis, "Optimization of agricultural field workability predictions for improved risk management," *Agronomy Journal*, vol. 107, no. 2, pp. 627–633, mar 2015.

[20] C. M. Bishop and C. M., *Pattern recognition and machine learning*. Springer, 2006.

[21] D. Koller and N. Friedman, *Probabilistic graphical models : principles and techniques*.

[22] S. J. Yen and Y. S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications*, vol. 36, no. 3 PART 1, pp. 5718–5727, 2009.

Improving 3D printing of garments by using HPC Cloud

D. Tomić*, D. Davidović*, T. Šubić**, J. Mesarić***, K. Skala*
* Ruđer Bošković Institute/ Center for Informatics and Computing, Zagreb, Croatia
** Arctur d.o.o/Nova Gorica, Slovenia
***Faculty of Electrical Engineering and Computing, Zagreb, Croatia

***Abstract* - 3D garment printing is a technology that is experiencing rapid development, and more and more garments are being printed on 3D printers. Advanced processes such as Selective Laser Sintering and PolyJet 3D printing are used. However, 3D printing of clothes still experiences some disadvantages, of which the biggest is the slow 3D printing speed and usually a small printing area of 3D printers. The novel, large multi-head printers can significantly increase the printing speed, but such printers require an additional step to prepare given 3D models that increase. However, pre-processing time grows exponentially with the number of printer heads, therefore computer resources that exceed the capabilities of a single workstation are required to prepare a printout. This implies the use of HPC resources, and because of its flexibility, lower cost and ease of use, HPC Cloud is optimal platform for such jobs. We have shown that the structure of these jobs fits perfectly in the HPC Cloud environment.**

Keywords - 3D garment printing, HPC Cloud, multi-head printers, PolyJet 3D printing, Selective Laser Sintering

I. INTRODUCTION

The connection between fashion [1] and 3D printing [2] is great and more and more projects are emerging. New additives [3] are coming, printing is becoming more precise, and new applications [4] are appearing. 3D printing offers huge opportunities for the fashion industry, from 3D printed clothing to 3D printed footwear and accessories, the possibilities are endless. Primarily, 3D printing was used more for an artistic touch, but the fashion industry is now moving towards the development of 3D printed wearable garments. The production of additives is interesting for fashion because it makes it easier to work on fashion designs and create amazing things for the fashion industry like clothes, jewelry and nets. This technology gives designers a lot of freedom in terms of geometry. For example, it is possible to create intricate designs for various projects within the fashion industry. From shoes and accessories to 3D printed dresses, the fashion industry is starting to embrace the full potential of 3D printing and develop interesting items.

With the current developments in 3D technology it becomes cheaper, more affordable and easier to use. The new opportunities in the fashion industry are arising and the reasons why designers are choosing 3D printing is changing: 3D printing turns out to be more sustainable and eco-friendlier. For instance, the production of textile generates the waste problem around the world, that is why a lot of elements of the manufacturing process need to be rethought in order to be eco-friendlier. Using 3D printing allows you to reduce waste, for example the 3D printer uses the exact amount of material that is needed to create a product. Experimenting with and creating new materials used in 3D printing in a more responsible way is the other reason for using 3D printing for fashion companies. The use of 3D printing to make clothes is constantly evolving. Extremely complex and impressive pieces were created first. The use of 3D printing in this sector is at a turn point to push the boundaries of the fashion world to create visually impressive creations. Nowadays, some designers are more interested in using the 3D printing technology to make their collections of regular clothing, using additive manufacturing, rather than using standards method of making. They use 3D printing not only for design benefits, but to create customizable and comfortable clothing. The introduction of advanced additives is a good way for all industries to improve their production process. It allows designers to work on their prototypes with faster and cheaper methods, and to produce customizable and cheaper products. 3D printing is a perfect technique for mass customization, which can otherwise be very expensive. Relatively fast tailoring is possible thanks to 3D printing. 3D printing could also fully create clothing that adapts to outdoor conditions. 3D printing thus transcends the boundaries of high fashion. There are already several examples on the market. The Spider dress from Anouk Wipprecht [5] has mechanical arms that extend and pull in response to external stimuli when people approach. The owner's breath helps to occupy the defensive position of the robotic arms. The dress is completely printed with Selective Laser Sintering (SLS) technology [6]. This technology prints the object parts additively by sintering tiny particles of polymer powder to bond them. The part is made layer by layer, according to the 3D model. During 3D printing, the SLS printer preheats the powder material just below its melting point, to allow the laser to melt faster. The roller applies a layer of polymer powder, then the laser sinters the powder according to the 3D file and the fabrication platform is lowered before applying a new layer of powder. The

procedure is repeated until the desired part is created. Julia Daviy [7] uses 3D printing to create biodegradable fashion. Her collection includes 3D printed dresses. Clothes created with 3D printing are usually not easy to wear. But that's not the case with Julie Daviy's clothes. Another example is the American designer Travis Fitch [8]. He prints fashion pieces with colorful multiple materials. The material is flexible to respond to body movements. The use of colors in this project is quite unique. This technique offers new possibilities because multicolor 3D printing is not often used in fashion projects. The Ministry of Supply [9] has developed a 3D printed knitted blazer. The process of 3D printing and the work on the 3D design they create allows the piece to be adapted to the body and its movements. Their goal is to develop an agile supply chain and create products that are more comfortable and durable. Fashion designer Danit Peleg [10] decided to 3D print her entire fashion collection. It took about 2,000 hours to print it completely. Her goal is to create comfortable clothes that she could wear herself. Alexis Walsh [11] created an impressive 3D dress printed using Selective Laser Sintering. The dress has 400 tiles assembled by hand. It took the designer six months to develop this garment. Designer dresses are not always comfortable because that is not their main goal. However, by 3D scanning the model, Jessica Rosenkrantz [12] created a dress that perfectly fits her body. Another great example is Viptie 3D [13]. This company focuses more on the aspect of mass use of 3D printing. They create ties and strands for their customers. PolyJet 3D Printing [14] works similarly to inkjet printing, but instead of jetting drops of ink onto paper, PolyJet 3D Printers jet layers of curable liquid photopolymer onto a build tray. During the pre-processing, the software automatically calculates the placement of photopolymers and support material from a 3D CAD file. During the printing, the resin 3D printer jets and instantly UV-cures tiny droplets of liquid photopolymer. Fine layers accumulate on the build tray to create several highly-detailed 3D printed products. Where overhangs or complex shapes require support, the Polyjet 3D printer jets a removable support material, named FullCore 750. The process of fine layer polymerization is repeated until the object is finished. After the photo polymerization is completed, the batch is placed under a pressurized water jet. This allows any excess liquid or supports to be removed with a few human interactions. The final product has a naturally smooth surface which can be polished further to create near full transparency for the clear resin. 3D printed models and parts are ready to handle and use right out after the cleaning process, there is no need for post-curing. There is a growth regarding the use of different 3D printing materials and techniques. In the future, metal 3D printing could be implemented to create some ornaments, but also laser cutting techniques.

II. HPC CLOUD FOR 3D PRINTING

Despite the great success of 3D printing in the garments production and design, one of the main drawbacks of this technology is the printing speed. As an example, Danit Peleg spent about 2,000 hours printing her collection, and Julia Daviy 6 months. This implies a higher price of such clothing, often unacceptable to the average customer. A multi-head printer with heads able to print a single object at the same time can significantly speed up printing [15], but then the time required to prepare the given 3D object (model) to the printer increases exponentially with the number of its heads. One solution to this problem is to leave the preparation of such files to the HPC system. However, traditional HPC resources are often difficult to access, and the cost of using them is relatively high, so it is good to consider using such resources on a Cloud system. In general, it is shown that a Cloud system is a suitable for HPC jobs in which no or minor inter-communication [16] is required. The calculation process is performed in two steps. In the first step, a code is generated from a CAD file (e.g. STL file) that describes the garments

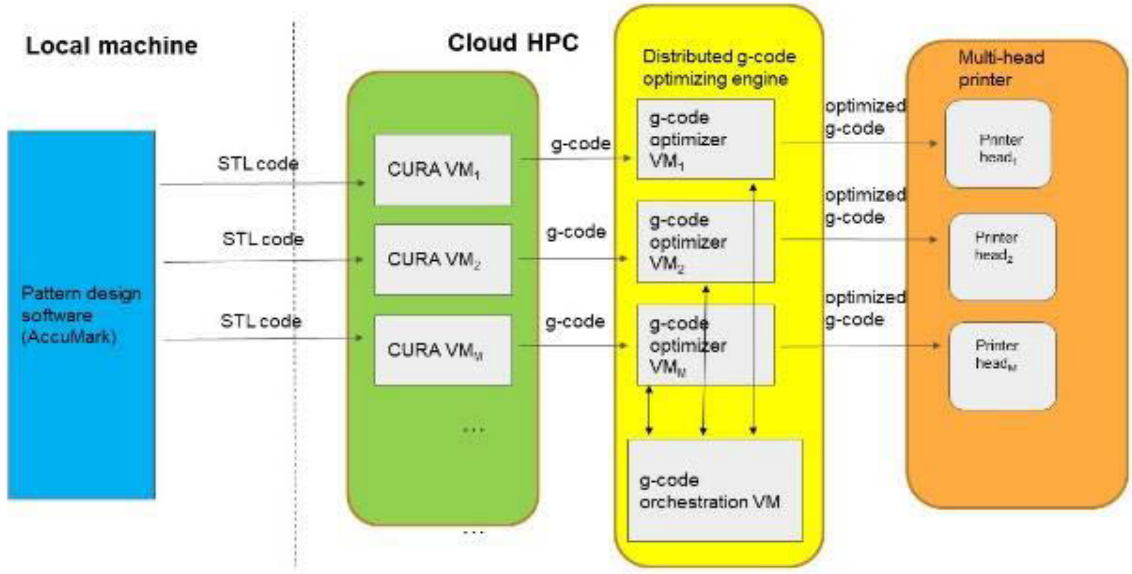


Fig 1: A schematic presentation of the 3D printing process in HPC Cloud. STL code fragments are generated on the local machine and sent to CURA virtual machines. These machines generate g-code and send it to the distributed g-code optimizing engine. Finally, an optimized g-code is generated and sent to the multi-headed printer.

using a G-code generator, such as the CURA Ultimaker software [17]. G-code or RS274 is the most commonly used computer numerical control programming language, used to control automated machine tools, such as 3D printer. The G-code translates the 3D-generated model to a set of machine instructions. In the case of the standard, off-the-shelf 3D printers, no additional pre-processing is required. The obtained G-codes are sent to the 3D printer(s) for printing. However, in the case of a specialized multi-head and/or multi-bridge 3D printers, the CURA generated G-code needs to be further optimized for that particular 3D printer. The whole process is schematically presented in Fig1. The goal of the optimization step is to divide the G-code, describing a single object that needs to be printed, into a set of G-codes, each tailored for one printing head. Then, each printing head can print its own part of the object without collision with other heads.

Let's the number of print heads be M, and the number of layer Ln represented with matrix A(i,j), with N matrix elements of value 1 for points i,j with print material. Then, the movement of mth head between the certain points of value 1 can be represented as a function of A:

$$hm = fm(A); m = 1, \dots, M \quad (1)$$

In addition, let each mth head print Nm points. Then, the total distance Hm that m^{th} head needs to traverse is:

$$Hm = \sum_{i=1}^{Nm} hm \quad (2)$$

The total number of printed points by all heads is:

$$N = N1 + N2 + \cdots + Nm \quad (3)$$

As the load should be equally distributed across all head, the following optimization problem can be posed:

$$\text{minimize } (Hm) \text{ w.r.t. } N1 = N2 = \cdots = Nm \quad (4)$$

Apparently, (4) can be split to M independent parallel tasks, with the communication occurring between them when they need to interchange the optimization results. Minimizing Hm is similar to the well-known travelling salesman problem [18] and the number of computational steps (NCS) required is exponential to the number of points traversed:

$$NCS = Kmexp^{Nm} \quad (5)$$

Value Km is dependent of the shape of the printing segment m. On the other side, the number of communications (NC) between tasks is:

$$NC = M! \quad (6)$$

The maximum number of printing heads M within one 3D printer on today's market is 6. Therefore, even for the objects with many segments the communication is neglectable to the processing:

$$NCS << NC \quad (7)$$

Inequality (7) suggests that the optimization of G-Code can be effectively performed on Cloud.

III. CONCLUSION

3D printing of clothing is increasingly coming to the fore today. However, this is a slow process, which increases the cost of its production. Multi-headed printers are able to print faster, but the pre-processing time of the required files grows proportionally with the number of printer heads. Fortunately, inequality (7) and the results from [16] suggests that this time can be significantly shortened by parallel processing these files on high-performance computing (HPC) systems. Further analysis of this problem shows that due to its specific structure, it can be efficiently solved on the Cloud. This increases the flexibility of the solution and reduces production costs.

ACKNOWLEDGMENT

This research has been supported by the European Commission H2020 project CloudiFacturing under grant agreement no. 768892, by the European Regional Development Fund under the grant KK.01.1.1.01.0009 – DATACROSS, and by EOSC Hub Horizon 2020.

REFERENCES

[1] Laver J: The Concise History of Costume and Fashion, H. N. Abrams, New York, 1979.

[2] Ngo TD, Kashani A, Imbalzano G, Nguyen TK, Hui D. Additive manufacturing (3D printing): A review of materials, methods, applications and challenges. Composites Part B: Engineering, Volume 143, 15 June 2018, Pages 172-196

[3] Vaezi M, Chianrabutra S, Mellor B, Yang S. Multiple material additive manufacturing – Part 1: a review, Virtual and Physical Prototyping, 8:1, 19-50, 2013. DOI: 10.1080/17452759.2013.778175

[4] Shahrubudin N, Lee T.C., Ramlan R. An Overview on 3D Printing Technology: Technological, Materials, and Applications. Procedia Manufacturing, Volume 35, 2019, Pages 1286-1296. https://doi.org/10.1016/j.promfg.2019.06.089

[5] https://www.arch2o.com/the-spider-dress-anouk-wipprecht/

[6] Kruth J.P., Mercelis P, Vaerenbergh V, Froyen L, Rombouts M, Binding mechanisms in selective laser sintering and selective laser melting, Rapid Prototyping Journal, vol 11, issue 1, Feb. 2015.

[7] https://www.3dnatives.com/en/julia-daviy-280520194/

[8] https://www.futurecurrent.net/travis-fitch

[9] https://www.ministryofsupply.com/

[10] https://en.wikipedia.org/wiki/Danit_Peleg

[11] https://www.interlaced.co/article/keeping-up-with-alexis-walsh-1529511915

[12] https://n-e-r-v-o-u-s.com/about_us.php

[13] https://www.viptie3d.com/

[14] Meisel N.A., Elliott A.M., Williams C.B. A procedure for creating actuatedjoints via embedding shape memoryalloys in PolyJet 3D

printing. Journal of Intelligent Material Systemsand Structures2015, Vol. 26(12) 1498–1512

[15] http://www.goya3d.com/#xl_xr_page_index

[16] D. Tomić, Z. Car and D. Ogrizović, "Running HPC applications on many million cores Cloud," *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, 2017, pp. 209-214, doi: 10.23919/MIPRO.2017.7973420.

[17] https://ultimaker.com/software/ultimaker-cura

[18] Lawler E. L. (1985). *The Travelling Salesman Problem: A Guided Tour of Combinatorial Optimization* (Repr. with corrections. ed.). John Wiley & sons. ISBN 978-0471904137.

Global repeat map algorithm (GRM) reveals differences in alpha satellite number of tandem and higher order repeats (HORs) in human, Neanderthal and chimpanzee genomes – novel tandem repeat database

I.Vlahović[*, **], M.Glunčić[*], K.Dekanić[*], L.Mršić[**], H.Jerković[**], I.Martinjak[****], V.Paar[*, ***]
[*] Department of Physics, Faculty of Science, University of Zagreb, Zagreb, Croatia
[**] Algebra University College, Zagreb, Croatia
[***] Croatian Academy of Sciences and Arts, Zagreb, Croatia
[****] Faculty of Science, University of Zagreb, Zagreb, Croatia

ines.vlahovic@algebra.hr, matko@phy.hr, kdekanic@phy.hr, leo.mrsic@algebra.hr, hrvoje.jerkovic@algebra.hr, imartinjak@phy.hr, paar@hazu.hr

***Abstract* - Development of novel sequencing methods (NGS) enabled investigation of DNA sequence in centromere region that, as is shown in previous studies, are replete with alpha satellite sequences, 171bp long monomer units. Those units (monomers) are found in each of human chromosomes as tandem repeats or as higher order repeat (HOR) structures. Except for their role in chromosome segregation and microtubule attachment to kinetochores, recent studies show that alpha satellite copy number variations are associated with cancer outcomes, aneuploidy, aging, differences in gene expression and evolutionary development. Using Global Repeat Map algorithm (GRM) and its extended ALPHAsub tool for finding all alpha monomers in genomic sequence, we created tandem repeat database for human, Neanderthal and chimpanzee genomes. This database has almost 3000 records (only main data table is shown on-line, http://genom.hazu.hr) with information about build of inserted genomes, start position in chromosome, sequence length of array, monomer length and number of monomers for specific array. Analysis of large set of data based on alpha satellites will help us to understand their role in centromeric organization and function and enable us to make conclusions about evolution between closely related species even though technologic limitation of assembly of centromere region are still present.**

Keywords – human, chimpanzee and Neanderthal genomes; DNA sequence; tandem repeats; Alphasub; Global Repeat Map algorithm, alpha satellites; database

I. INTRODUCTION

The emergence of new methods and technologies for DNA sequencing (such as NGS, nanopore sequencing, single molecule real time sequencing, optical mapping…) has boosted development of new databases for storing and analyzing DNA or protein sequences. Those databases are core part for scientific researches in life science and for understanding the roles and functions of specific DNA or protein sequence segments in diseases (e.g. in cancer research), evolutionary studies of closely related species, binding motifs and drug discovery. With growing collected amount of data from DNA or protein sequencing, new specialized databases become more sophisticated by offering more complex analysis tools for their analysis. Recently, the paper with an overview of currently available databases was published [1] – that describes NAR online Molecular Biology Database Collection. That paper shows that there are ~1600 databases categorized in 15 categories and 41 subcategories. Some of the most relevant databases for life science on that list are GWAS catalog, EWAS catalog, KEGG, Cancer database COSMIC, NCBI, Ensemble, UCSC genomic browser, ClinVar, AWESOME, TRDB…

Although there are vast number of databases, only few are oriented toward collecting and analysis of repetitive elements such as tandem repeats (TRDB, STRBase, Plant Repeat Database, RepBase).

Repetitive elements in genomes can occupy more than 50% of DNA sequence (e.g. in human genome [2]) and they are divided to different types: tandem repeats, dispersed repeats, segmental duplications, complex repeats. Tandem repeats usually have primary repeat unit, called monomer, whose copies are arranged in head-to-tail fashion in DNA sequence and by which they can be further classified according to its length of base pair to microsatellites (between 1 bp to ~6 bp), minisatellites (between ~ 6 bp and ~ 10 bp), satellites (between ~100 bp and ~2 kbp) and macrosatellites (>~2 kbp).

Studies have shown that tandem repeats have roles as markers for genotyping a number of pathogens whose flexibility gives opportunity for fast adaptation [3] and that copy number variation (duplications or deletions of monomers created with processes such as unequal crossing-over, replication slippage) in them can modulate molecular paths influencing individuals in population that might lead to changes of phenotype and adaptation [4] so they can be responsible for rapid changes and variation in structure and function as well as development of some disease (Huntington, schizophrenia, cancer…) [5,6].

One of the most researched tandem repeats are alpha satellites in human and primates whit primary unit ~171 bp long. They are found in each human chromosome in centromere region in long arrays that are usually organized in form of higher order repeat structures (HORs). Centromeric alpha satellites are important for segregation of chromosomes in cell division and changes in that area of chromosome might cause missegregation, aneuploidy and cancers [6].

Some algorithms have been developed for studying tandem repeats such as TRF, Sputnik, Mreps, RepeatMasker, SRF and STAR [7] that are based or on signal processing or on probabilistic methods. Comparison of those tools gave differences in number of monomer units in repeats, their sizes and length distributions [7, 8] showing their limitations and advantages that comes from their input parameters.

In this paper we show distribution for human, Neanderthal and chimpanzee alpha satellites in their genomes and differences, analyzed with Global Repeat Map algorithm and ALPHAsub algorithm. All data about positions, length, and monomer number of alpha satellite arrays in chromosomes are available at http://genom.hazu.hr/.

II. METHODS

A. Data

For human genome analysis we used hg38 sequence assembly from NCBI as well as for chimpanzee genome Pan_tro_3.0 assembly. DNA sequence for Neanderthal genome was provided by Janet Kelso group from Max Planck institute, Department of Evolutionary Genetics in Leipzig.

B. Global Repeat Map Method (GRM)

The main feature of GRM algorithm is its ability to identify a broad variety of repeats of unbounded length that can be arbitrarily distant in sequences as large as human chromosomes. The efficacy is due to the use of complete set of a *K*-string ensemble which enables a new method of direct mapping of symbolic DNA sequence into frequency domain, with straightforward identification of repeats as peaks in GRM diagram. GRM uses the frequency domain, but avoiding mapping of symbolic DNA sequence into numerical sequence, and uses *K*-string matching, but avoiding statistical methods of locally optimizing individual *K*-strings mapping. (Fig. 1) [2].

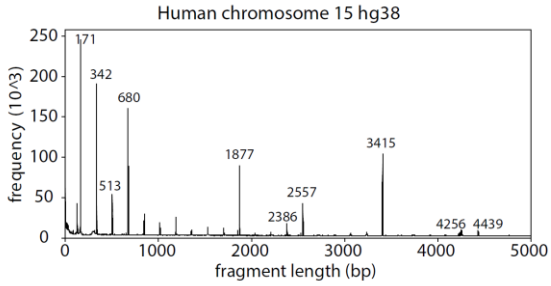


Figure 1. GRM diagram for Human chromosome 15. On figure main peaks are labeled that indicates appearance of monomer unit of alpha satellites ~171 bp and its multiples 2x171=342 bp, 3x171=513 bp ie. indicates existence of HOR structure.

In this way, the level of background noise is drastically reduced with respect to other available algorithms, enabling identification of very complex higher order repeats. For repeats which can be submerged in the background of noise we can use GRM as computational "magnifying glasses" for their identification and analysis, sizably increasing the resolution power. In the GRM diagram each significant peak ("fragment length") presents the length of repeat unit, which we in analysis use for determination of its position and sequence using K-sting matching with only one input parameter of k-mer length (usually 8 bp long). Further analysis determines repeat consensus sequence and divergences between monomers and consensus sequence of repeat unit in one tandem array as well as between all monomer units in that array using Needleman-Wunsh algorithm for string matching calculations [9]. The last step is crucial for determination of HOR existence and its structure in DNA sequence since HOR structures have smaller divergence in respect to divergence between monomers in array as Fig. 2 shows (see detailed description in [2]).

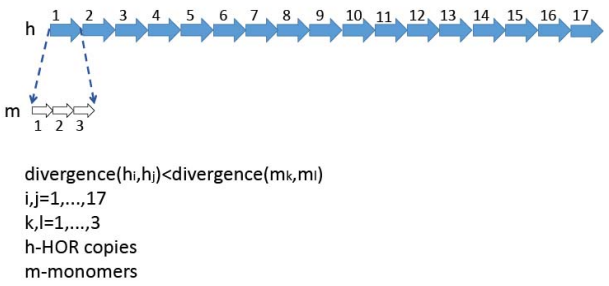


Figure 2. Higher order repeat structure definition. Figure shows 3mer HOR structure that repeat itself in tandem of 17 copies. Each HOR copy in tandem repeat array is composed from 3 repeat units – monomers. Divergence between HOR copies is much less then divergence between monomers in array (~5% vs. ~50% for alpha satellites and for example from this scheme, that represents NBPF family gene HOR, those divergences are ~1% vs. ~20% [12]). Figure is taken from [13].

C. ALPHAsub algorithm

For automatization of alpha satellite location identification in chromosomes we use 28 bp long sequence motif (in direct and its reverse complement version): TGAGAAACTGCTTTGTGATGTGTGCATT as "ideal key". For position determination of alpha satellite array for this "ideal key" in chromosomes we allow at most 9 bp mismatch using the Levenshtein distance algorithm. Based on those matching positions in DNA sequence we calculate distances between neighboring position hits and their multiples (1x~171 bp, 2x~171…) that are approximately equal to 171 bp. Details of method are described in [10].

III. RESULTS

DNA sequence analyzed using GRM and ALPHAsub algorithm show that in these three species chromosome alpha satellites are clustered in a vast range of tandem arrays, from 1 to 139 per chromosome. We clustered them in three ranges for purpose of their distribution analysis. For alpha satellite sequences in the centromeric/pericentromeric region our code identifies a novel substructure: selected subarrays (referred to as

TABLE I. NUMBER OF MONOMERS IN ALPHA SATELLITE ARRAYS IN CHIMPANZEE, NEANDERTHAL AND HUMAN GENOMES.

Chromosome	Species number of monomers		
	Chimpanzee	***Neanderthal***	***Human***
1	78	745	19884
2/(2A in chimpanzee)	-	534	11998
2B	784	-	-
3	2243	940	18432
4	5836	373	1632
5	3446	3158	24858
6	423	934	9119
7	3481	6185	24484
8	4674	2938	14604
9	922	1487	15936
10	77	762	14699
11	3237	6386	25604
12	4642	3538	18980
13	16	17	12039
14	6	332	12673
15	-	223	16221
16	1512	2055	14202
17	185	751	23749
18	-	210	26353
19	2871	3310	17141
20	-	510	18123
21	-	545	14409
22	50	31	14413
X	1408	3560	25774
Y	2910	792	2397

alpha satellite arrays) consisting of tandem of alpha satellite monomers approximately characterized by the 28 bp "alpha recognition nucleotide subset", as defined in this work. According to the number of alpha satellite subarrays we classify chromosomes in three classes: (1) with less than 50, (2) with 50-100, and (3) with more than 100 alpha satellite subarrays.

We show that in Neanderthal genome 19 chromosomes have alpha satellite arrays clustered mostly in first range group (<50 tandem array per chromosome), and that in this range group in human and chimpanzee genomes are 12 chromosomes and 10 chromosomes respectively. Also, in 9 human chromosomes alpha satellites are clustered in second range group (with between 50 and 100 alpha satellite tandem arrays per chromosome) and in third range group (with more than 100 tandem arrays) are 3 chromosomes.

In second range group of alpha satellite arrays per chromosome for Neanderthal and chimpanzee genomes are 4 and 8 chromosomes. For the third range group of number of alpha satellite arrays for Neanderthal and chimpanzee we show that they are concentrated in 1 and 2 chromosomes. This distribution of number of alfa satellites tandem arrays in their chromosome are shown in Fig. 3. Overall, the highest number of alpha satellite arrays has human (1249), then chimpanzee (979) and Neanderthal genome (712).

In table 1 we show number of monomers in alpha satellite arrays in each chromosome for these three species. Results in table 1 shows that, as expected, number of monomers is largest for human genome (overall 397724). In chimpanzee there are 38801 monomers and in Neanderthal genome 40316 monomers distributed in alpha

TABLE II. NUMBER OF NUCLEOTIDE BASES IN ALPHA SATELLITE ARRAYS/ AND % OF ALPHA SATELLITES IN CHROMOSOMES IN CHIMPANZEE, NEANDERTHAL AND HUMAN GENOMES.

Chromosome	Species-number of bases in alpha satellite arrays/ % of alpha satellites in chromosomes		
	Chimpanzee	Neanderthal	Human
1	13238/(0.006%)	125992/(0.051%)	3377695/(1.357%)
2/(2A in chimpanzee)	-	86234/(0.035%)	2074000/(0.856%)
2B	133329/(0.1%)	-	-
3	382224/(0.189%)	159448/(0.081%)	3107741/(1.567%)
4	992205/(0.51%)	63470/(0.033%)	277751/(0.146%)
5	588454/(0.323%)	541200/(0.299%)	4236103/(2.333%)
6	72301/(0.041%)	158847/(0.093%)	1549137/(0.907%)
7	602625/(0.363%)	1054718/(0.662%)	4173185/(2.619%)
8	785517/(0.531%)	502009/(0.343%)	2487664/(1.714%)
9	158180/(0.135%)	255594/(0.181%)	2712061/(1.96%)
10	13117/(0.01%)	130654/(0.096%)	2550572/(1.906%)
11	553998/(0.408%)	1093466/(0.81%)	4352845/(3.222%)
12	792398/(0.578%)	605501/(0.452%)	3227251/(2.421%)
13	2665/(0.003%)	2838/(0.002%)	2051463/(1.794%)
14	1020/(0.001%)	56672/(0.053%)	2159477/(2.017%)
15	-	38630/(0.038%)	2766845/(2.713%)
16	58050/(0.071%)	350305/(0.388%)	2422346/(2.681%)
17	31452/(0.038%)	128158/(0.158%)	4025311/(4.835%)
18	-	35657/(0.046%)	4472621/(5.565%)
19	488714/(0.797%)	564976/(0.955%)	2924481/(4.989%)
20	-	87011/(0.138%)	3084429/(4.786%)
21	-	93518/(0.194%)	2455375/(5.257%)
22	8424/(0.022%)	5399/(0.011%)	2455649/(4.832%)
X	242122/(0.156%)	608419/(0.392%)	4416919/(2.831%)
Y	496190/(1.883%)	134881/(0.228%)	407986/(0.713%)
Genome	6416223/(0.216%)	6883597/(0.222%)	67768907/(2.194%)

satellite arrays. Accordingly, number of bases of alpha satellite arrays in human is also highest as is shown in table 2. In table 2 we show also the percentage that alpha satellites occupy in the total number of bases of each chromosome. In human, the highest percentage of alpha satellites in DNA sequence are in chromosomes 17 – 22

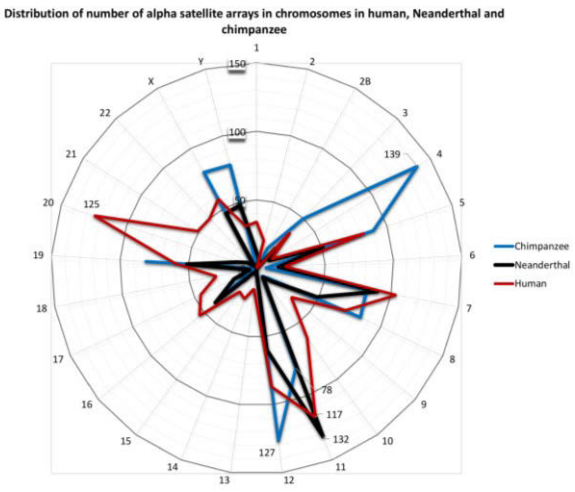


Figure 3. Distribution of human, Neanderthal and chimpanzee alfa satellites number of arrays in chromosomes.

(~5%), while in Neanderthal and chimpanzee genomes this percentage is much smaller (table 2). One of the chromosomes with this high percentage is chromosome 21, for which we have shown in one of our previous paper using our methods for studying alpha satellites, that it has an eleven *nmer* HOR arrays with n ⩾8 [10].

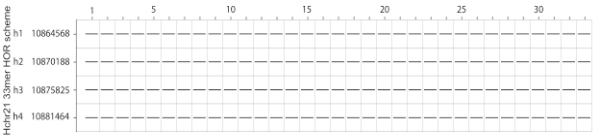


Figure 4. 33mer HOR structure in human chromosome 21. Numbers 1 to 33 at top represent monomer unit of alpha satellite ~171 bp. 33mer HOR copies consist of four HOR copies in array (denoted hi, i = 1, 2, 3, 4). Monomers of the same type (from consensus HOR) in different HOR copies are presented by a bar in the same column. Image has been downloaded and revised from reference [10].

This chromosome is responsible for occurrence of Down syndrome (DS) and some studies show that alpha satellite arrays in it can be used as potential molecular diagnostic markers of DS [11] because of its polymorphisms.

One of those HORs is 33mer HOR, based on alpha satellite monomer unit that is the largest alpha satellite HOR among all human somatic chromosomes for which we have shown its structure (Fig. 4.). Across the human, Neanderthal and chimpanzee genomes alpha satellites occupy only, in the order indicated, 2.19%, 0.22% and 0.22% of the assembled DNA sequence. As mentioned in methods, one of the steps of our analysis is the determination of divergences among monomer units in tandem array. We calculated all those divergences in every alpha satellite array in each chromosome for human, Neanderthal and chimpanzee genome. Average divergences of monomers in arrays for whole genome of human, Neanderthal and chimpanzee, in that order, are 23.27%, 26.04% and 25.24%. If we look for every chromosome individually for each species, those divergences per chromosome per each alpha satellite array reveal vast divergence ranges, mostly from ~9% to ~45% except for outliners as Fig.5. shows.

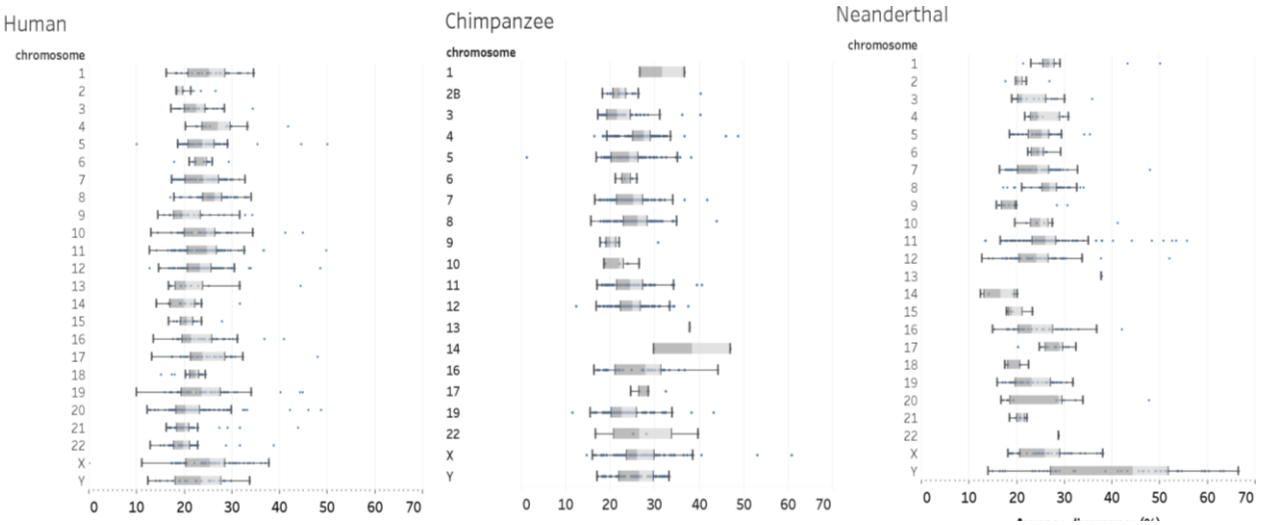


Figure 5. Average divergence of alpha satellite monomers in every array in every chromsome in human, chimpanzee and Neanderthal genome. Inside of boxes, line between dark grey and light grey box represents median of divergences for all alpha satellite array in each chromosome.

Our results show that alpha satellite distribution in chromosomes for human, Neanderthal and chimpanzee genomes analyzed with Global Repeat Map algorithm and ALPHAsub algorithm vary greatly in number of alpha satellite tandem arrays. Also, results show that in this build versions of genomes for human, Neanderthal and chimpanzee number of alpha satellite monomers in tandem is highest for human. The number of tandem arrays in which alpha satellites concentrates is much less in chimpanzee and Neanderthal regards to human genome but in some chromosomes this number is higher for chimpanzee and Neanderthal. In chimpanzee chromosome 4 there are 5836 monomers in 139 alpha satellite arrays, while in human chromosome 4 there are 1632 alpha satellite monomers and they are concentrated in only 14 tandem arrays. In Neanderthal chromosome 4, 373 monomers are arranged in 11 arrays. Chimpanzee have in few more chromosomes higher number of arrays than human. Those are chromosome 3 with 50 arrays (2243 monomers) vs. human 36 arrays (18432 monomers), chromosome 5 with 90 arrays (3446 monomers) vs. human 83 arrays (24858 monomers), chromosome 8 with 84 arrays (4674 monomers) vs. human 72 arrays (14604 monomers), chromosome 12 with 127 arrays (4642 monomers) vs. human 87 arrays (18980), chromosome 19 with 81 array (2871 monomers) vs. human 60 arrays (17141 monomers), chromosome X with 80 arrays (1408 arrays) vs. human 58 arrays (25774 monomers) and chromosome Y who have both more arrays and monomers than human - 78 arrays (with 2910 monomers) vs. human 32 arrays (2397 monomers).

Neanderthal chromosomes have mostly smaller number of tandem arrays than human and chimpanzee in this sequence version except for chromosome 7 that have 90 arrays (6185 monomers) regard to chimpanzee that

have alpha satellites in 82 arrays (3481 monomers) and similarly, in chromosome 16 Neanderthal have 39 arrays with alpha satellites with 2055 monomers while chimpanzee have 34 arrays with 1512 monomers. Neanderthal chromosome 11 has greater number of alpha satellite subarrays (132) than chromosome 11 in human (117) and chimpanzee (78), but the summary number of alpha satellite monomers is higher in human (25604) and smaller in chimpanzee (3237).

Mostly, tandem repeats in genomes are analyzed with other tools, such as the most popular Tandem Repeat Finder, for storing in databases, but those tools have problems with long repeating units and with low power of predictions of divergent monomers and divergence in monomer unit lengths because their input parameters [7]. Our GRM algorithm is very robust for copy number variation and length of monomer unit deviations as we

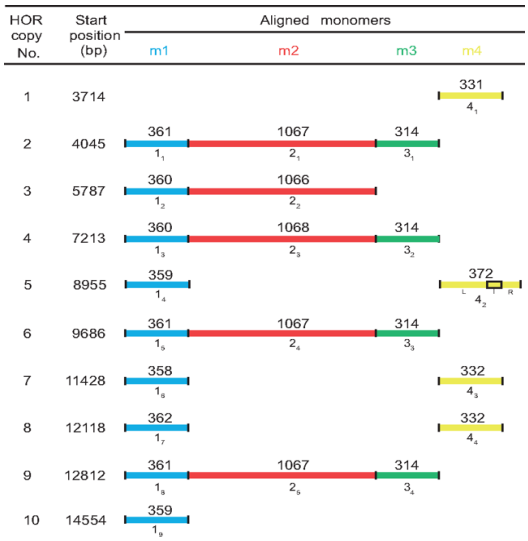


Figure 6. Aligned monomer structure of 4mer HOR in DS497953.1 is composed from three Tcast-360 monomer types (m1, m3, and m4) and one ~1067 bp monomer type (m2) in insect *T.castanum*. Monomer types constituting HOR are denoted by m1, m2, m3, and m4 (shown at top of columns). In complex repeats can have indels like the monomer 4_2 (of length 372 bp) from HOR copy No. 5 with a 48 bp extraneous insertion (shown as short rectangle along a part of horizontal bar representing this monomer). Image has been downloaded and revised from reference [13].

have previously shown for insect *Tribolium castaneum* satellites that vary in monomer unit lengths from ~331 bp to 369 bp [13]. Those variation in monomer unit length also can give us insight in complex repeats and their structure [13] as Fig.6 show.

IV. Conlusion

This analysis of alpha satellite arrays of closely related species, human, Neanderthal and chimpanzee give us indication about their dynamic and evolution formed with different processes of these elements in DNA sequence and because their mutational rate are much higher than in other parts of genomes. Although roles of tandem repeats in DNA sequence are not still resolved, more and more studies are showing their importance such as studies on plants that show that genes with tandem repeats in side of them in introns are connected with transcription and translation processes in cell [14] and that they can influence the increase of protein variations and modifications [15] and possible roles of alpha satellite arrays in cancer [16]. Our previous work showed that differences in DNA sequence of closely related species are in large tandem arrays and their HORs [17, 10]. Repetitive DNA sequences are of increasing importance because of their possible regulatory role and of important factors for evolution. A convincing example of crucial evolutionary role of HORs is a 3mer HOR pattern, based on 1.6 kb NBPF monomers related to brain function. This HOR represents one of the outstanding human-chimpanzee differences: human chromosome 1 contains 47 tandem NBPF HORs and chimpanzee none (discovered by our GRM algorithm), which can be one of the key elements of human cognitive abilities developed during last six million years of evolution since the human - chimpanzee split.

Identification of HOR structures based on alpha satellite arrays are one of our main goal for further work as well as explanation of their structures. By now our database show data for alpha satellites, but in further work we will expand it with other tandem repeats for those species (as well as newly added investigated species) for which we will provide detailed data (sequence composition of arrays in terms of GC content, consensus sequence of repeat based on monomer in tandem arrays, complete array sequence that could be used for comparison and phylogenetic analysis of similar sequence in closely related species as well as for uses in other bioinformatic tools such as Blast). Also, for identified HORs we will provide data about their structures and divergences. Also, we will make platform with online GRM and ALPHAsub tools for identification of repeats in species submitted by other researchers.

Main flaw for our study of repeats in different species genomes are errors that are occurring during assembly of DNA sequences [17] especially for centromere region. Solution of this problem would be novel, more precise and cheaper sequencing methods or combination of existing ones for genome sequencing and assembly. Still, investigation of repeats in genomes can give us much novel information that could help researchers to reveal their functions in cell.

Acknowledgment

We acknowledge support from the QuantiXLie Centre of Excellence, a project cofinanced by the Croatian Government and European Union through the European Regional Development Fund - the Competitiveness and Cohesion Operational Programme (Grant KK.01.1.1.01.0004), and the grant IP-2014-09-3626 from Croatian Science Foundation.

References

[1] D. J. Rigden and X. M. Fernández, "The 26th annual Nucleic Acid Research database issue and Molecular Biology database collection", *Nucleic Acids Research, Vol. 47,* pp. *D1–D7, January 2019.*

[2] M. Glunčić and V. Paar. “Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm.” Nucleic acids research, Vol. 41, 1 (2013): e17. doi:10.1093/nar/gks721.

[3] P.Le Flèche. “A tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis”. MC Microbiol, 1, 2 (2001) doi:10.1186/1471-2180-1-2.

[4] V. Romero, K. Hosomichi, H. Nakaoka, H. Shibata and I. Inoue. “Structure and evolution of the filaggrin gene repeated region in primates.” BMC Evol Biol 17, 10 (2017) doi:10.1186/s12862-016-0851-5.

[5] J.W. Fondon and H.R. Garner. “Molecular origins of rapid and continuous morphological evolution.” PNAS. 2004;101:18058–63.

[6] M. Glunčić, I. Vlahović and V. Paar. “Discovery of 33mer in chromosome 21 – the largest alpha satellite higher order repeat unit among all human somatic chromosomes.” Sci Rep 9, 12629 (2019) doi:10.1038/s41598-019-49022-2.

[7] S. Leclercq, E. Rivals and P. Jarne. “Detecting microsatellites within genomes: significant variation among algorithms. BMC bioinformatics”, 2007. Vol 8, 125. doi:10.1186/1471-2105-8-125.

[8] A. Merkel and N. Gemmell. “Detecting short tandem repeats from genome data: opening the software black box”, Briefings in Bioinformatics, Volume 9, Issue 5, September 2008, Pages 355–366.

[9] S. B. Needleman and C. D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". Journal of Molecular Biology. 1970. 48 (3): 443–53. doi:10.1016/0022-2836(70)90057-4. PMID 5420325.

[10] M. Glunčić, I. Vlahović and V. Paar. Discovery of 33mer in chromosome 21 – the largest alpha satellite higher order repeat unit among all human somatic chromosomes. Sci Rep 9, 12629 (2019) doi:10.1038/s41598-019-49022-2.

[11] Q. Chen, B. Tan, J. He, X. Liu, X. Chen, R. Gao, J. Zhu, Y. Wang, H. Q."Mutational spectrum of CENP-B box and α-satellite DNA on chromosome 21 in Down syndrome children". Molecular Medicine Reports 15.4 (2017): 2313-2317.

[12] V. Paar, M. Glunčić, M. Rosandić, I. Basar, I. Vlahović. “Intragene Higher Order Repeats in Neuroblastoma BreakPoint Family Genes Distinguish Humans from Chimpanzees”. Molecular Biology and Evolution, Volume 28, Issue 6, June 2011, Pages 1877–1892.

[13] I. Vlahović, M. Glunčić, M. Rosandić, Đ. Ugarković, V. Paar. “Regular Higher Order Repeat Structures in Beetle *Tribolium castaneum* Genome”. *Genome Biology and Evolution*, Volume 9, Issue 10, October 2017, Pages 2668–2680, https://doi.org/10.1093/gbe/evw174.

[14] Z.Zhao, C.Guo, S. Sutharzan, P. Li, C. S. Echt, J. Zhang and C. Liang. “Genome-wide analysis of tandem repeats in plants and green algae.” G3 (Bethesda). 2014;4(1):67–78.

[15] O. K Tørresen, B. Star, P. Mier, M.A Andrade-Navarro, A. Bateman, P. Jarnot, A. Gruca, M. Grynberg, A. V Kajava, V. J Promponas, M. Anisimova, K. S Jakobsen, D. Linke. “Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases”. Nucleic Acids Research, Volume 47, Issue 21, 02 December 2019, Pages 10994–11006, https://doi.org/10.1093/nar/gkz841

[16] M. Levy-Sakin et al. „Genome maps across 26 human populations reveal population-specific patterns of structural variation”. Nature Communications volume 10, Article number: 1025 (2019).

[17] V. Paar, M.Gluncic, I. Basar, M. Rosandic, P. Paar and M. Cvitković. „Large Tandem, Higher Order Repeats and Regularly Dispersed Repeat Units Contribute Substantially to Divergence Between Human and Chimpanzee Y Chromosomes.“ Journal of molecular evolution. 2011.72. 34-55. 10.1007/s00239-010-9401-8

Air Traffic Control Competencies Measurements based on Functional Near-Infrared Spectroscopy

Ivan Kesedžić*, Jelena Božek*, Milan Radoš**, Siniša Popović*, Krešimir Ćosić*
* Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
** Croatian Institute for Brain Research, School of Medicine, University of Zagreb, Zagreb, Croatia
ivan.kesedzic@fer.hr

***Abstract* – Air traffic control officers (ATCOs) are rigorously selected from the vast number of applicants and extensively trained, in order to have the required psychological and behavioural characteristics for this stressful occupation, as well as specialised performance competencies while controlling air traffic, such as situational awareness, workload management, problem-solving and decision-making. Recent studies have demonstrated the possible improvements of ATCO selection and training processes by using physiological, speech and oculometric features related to stress resilience, cognitive load, visual attention and fatigue. The functional near-infrared spectroscopy (fNIRS) is a brain imaging technique which provides a good insight into brain activation by measuring the haemoglobin concentration in the prefrontal cortex, which is one of the key brain areas responsible for complex cognitive behaviour, decision-making, working memory, and stress regulation. This paper focuses on the ATC performance competencies that could be measured with the fNIRS, demonstrates computational processing of signals obtained with this technique, and discusses the applicability of fNIRS measurements in the current ATC selection and training procedures. The experimental paradigm and the results obtained from the paradigm demonstrate that fNIRS shows differences in brain activation among individuals during cognitive tasks, thus allowing the more objective performance measurement.**

Keywords – air traffic control competencies, performance, fNIRS, signal processing

I. INTRODUCTION

Air traffic control (ATC) is today characterised as one of the most stressful professions which requires a specific set of cognitive and psychological skills. ATC officers (ATCOs) regulate air traffic flow using international regulations while keeping track of flights on radars and computer systems. They need to be able to deal with unexpected events, such as changes in weather, unscheduled traffic, and emergency situations. ATCOs can work at the airports as part of the tower control which controls landings and departures, but also in the approach centre where they stream aircraft at lower altitudes, or in the area centres, where they control the airspace at higher altitudes. The complexity of ATCOs' job depends on the type of their job (tower, approach or area controller) but also on the number of aircraft arriving to and departing from the airport for tower and approach controllers, and the number of aircraft they navigate in their designated airspace for area controllers.

The path to the ATCO licensing is both cost- and time-consuming. The selection process is rigorous since only a small percentage of applied candidates fulfil the selection criteria, e.g. only 6 % of the applied candidates fulfil these criteria in the Maastricht area [1]. During the selection process, candidates' psychological, behavioural, health, and ATC-related set of skills are being rated. Some of the skills evaluated in the selection process are logical reasoning, teamwork, communication, motivation, situational awareness, visual perception etc. Lately, new tools and methods for enhancement of ATCO selection processes have been proposed which enable the measurements of the candidates' psychophysiological features [2, 3]. Although there is a significant amount of ATC-related research, the lack of application of these new tools in the selection processes is still present.

The ATCO selection process is divided into several phases, where each phase evaluates candidates according to different criteria required for ATCOs. During the preselection phase, the candidates are rejected if they do not fulfil basic requirements from the public call for the ATCO candidates, such as the appropriate age and the level of education. The first phase usually includes an evaluation of the candidates' English language, mental arithmetic, working memory, visual perception, spatial orientation, and attention. The candidates' psychological testing is usually performed in this phase. The second phase assesses the application of these skills in the working conditions and measures the multi-tasking ability, trainability, collaboration, and stress management, mostly on the simulator tasks. A series of interviews and other simulator tasks with varying fidelities complete the second phase of the selection process. The last phase of the selection process is the medical examination. The duration of the selection process depends on the number of applicants and the process usually lasts several months. Selection is followed by the extensive education and training of the selected candidates, over the period of 2-3 years [4]. The attrition rates during the ATCO training vary between countries and are between 25 and 50 % [4, 5, 6].

One of the essential skills required in the ATC profession is stress management, which is usually

IK has been supported through the funding from the Croatian Science Foundation. JB has been supported through the funding from the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS).

assessed in the later phases of the selection process through behavioural observation during simulated exercises. During the recent selection process in the Croatia Control, the novel multimodal psychophysiological measurements of ATCO candidates were obtained, for physiological assessment of candidates' stress resilience, oculometric assessment of visual attention and fatigue, and speech-based assessment of cognitive load [2, 3]. This multimodal approach to the assessment of various ATC-relevant mental capabilities, states, and traits could be further augmented to include the portable brain imaging measurements, like the functional near-infrared spectroscopy (fNIRS).

II. NEUROPSYCHOPHYSIOLOGY IN THE ATC-RELATED RESEARCH

The complexity of ATCOs' profession offers research opportunities for the analysis of the complex human behaviour, such as decision-making, problem-solving, stress management, and many more. These research efforts combine neuropsychophysiological modalities with standardised tasks (e.g. basic arithmetic task, Stroop task, etc.) or ATC-related simulator tasks, in order to gain additional knowledge regarding the cognitive challenges and required abilities within this profession. The participation of various ATC institutions in these studies, such as Eurocontrol, Federal Aviation Administration, Ecole Nationale de l'Aviation Civile, Società Nazionale per l'Assistenza al Volo, Croatia Control, etc., indicate the importance of research in this field [2, 3, 7, 8, 9, 10, 11, 12].

Some of the neuropsychophysiological modalities in the ATC-related research field include brain activity analysis, such as fNIRS, and electroencephalography (EEG) recordings; peripheral physiology analysis, including electrocardiography (ECG), electromyography (EMG), respiration measures, electrodermal activity (EDA), and skin temperature measures; acoustic characteristics of the human speech; oculometric analysis, including the eye gaze and pupil analysis; facial analysis etc. Although the ATC-related neuropsychophysiological research is present for decades, its application in the ATCO selection and training procedures is currently not common. However, the recent publications started to demonstrate how the selection and training processes could benefit from the objective multimodal performance measurements [2, 3], which could be extended to include neuropsychophysiological modalities.

Studies in the field involve current and veteran ATCOs, as well as ATCO trainees and ATCO candidates, while the paradigms include simulator tasks, generic tasks or a combination of both. The most researched constructs of interest are workload, performance, stress, situational awareness, and fatigue. Both single modality, as well as the combination of modalities, are used within these studies. Situational awareness is mostly assessed using the oculometric measurements [13], while the mental workload is mostly assessed using the human voice analysis and brain imaging measurements [7, 8, 9, 10, 11, 12].

The simplest approaches in evaluating the mental workload are based on subjective self-reports, which can be reported both retrospectively, i.e. after the task is finished, or while the cognitive task is performed. Beside the biases that can be present in the self-reporting questionnaires, the perception of the task difficulty or mental workload can be affected by the time interval between the task and the report, while the self-reporting during the task can even increase the mental workload since it may be regarded as another task to be performed within the experiment [14].

The mental workload can be observed in psychophysiological responses on various tasks. In this context, one of the most researched psychophysiological measurements is related to the cardiac activity (e.g. heart rate variability) which changes during the mental workload [9]. The galvanic skin response and skin temperature were correlated with the mental workload [9], as well as features extracted from voice [15]. Eye-tracking presents an easy technique for mental workload measurements, with the pupil dilation being one of the most researched features for mental workload extracted from the oculometry [13].

The ATCOs' mental workload can be examined both objectively and subjectively. The fNIRS could provide performance metrics directly from brain-based measures to assess safety and performance of operators in high-risk fields and, in some specific tasks, their difficulties can be classified using only fNIRS-based brain measurements [16]. Besides the fNIRS, other brain imaging techniques, such as EEG, are used in ATC-related research. In a study by Di Flumeri [11], professional ATCOs' EEG was recorded while performing an air traffic management task. The workload was measured subjectively and objectively, showing a high positive and significant correlation between these two measures, as well as the discriminability of the task difficulty levels from the EEG recordings. A high significant correlation between the neurophysiological signals and the subjective workload questionnaire was also shown in [7] where the professional ATCOs performed realistic air traffic management scenarios.

A recent study by Arico et al. [8] demonstrated that neurophysiologically-based adaptive automation could have its utilization in the ATC. The ATCO students performed the air traffic management scenario with and without the support of the adaptive automation solutions. Based on the EEG recordings, the adaptive automation was enabled mostly during the high-demanding conditions in order to reduce the mental workload under which the ATCOs were operating. The longstanding goal for the future of the ATC is the ability of technology to reliably identify people that are at risk of becoming overloaded and automatically adjust their task demand accordingly [17]. However, there is much research to be done in this field, since the accuracy of these systems may have to approach 95% to be acceptable in the ATC workplace [18].

III. ATC COMPETENCIES AND FNIRS-BASED MEASUREMENTS

During the ATCO selection and training processes, various skills are measured on a simulator and other tasks, but they are also graded by the ATCO instructors during candidates' selection and training phase. The competencies evaluated by the ATCO instructors are listed and described in the ATC Refresher Training Manual [19] published by the Eurocontrol. The competency is defined as a combination of skills, knowledge and attitudes required to perform a task to the prescribed standard [20], i.e. to provide safe and efficient air traffic services. The competencies defined in [19] and observed within the ATCO selection and training processes are situational awareness, traffic and capacity management, separation and conflict resolution, communication, coordination, management of non-routine situations, problem-solving and decision-making, self-management and continuous development, workload management, and teamwork.

Each of these competencies is described with a set of performance criteria. The performance criteria are simple, evaluative statements on the required outcome of the competency element and a description of the criteria used to judge whether the required level of performance has been achieved [20]. The performance criteria within the competencies are measured by the ATCO instructors, which may result in the subjective performance evaluation. Introduction of new objective neuropsychophysiological tools and methods in the competencies and performance criteria measurements could enrich the current evaluation procedures both during the selection as well as during the training processes. An example of the objective psychophysiological measurements used in the ATCO selection process, described in [3], resulted with a multimodal performance assessment radar chart illustrating the candidates' variability within the multimodal and multidimensional performance space, where the axes of the space are related to stress resilience, visual attention and fatigue, and cognitive load. Such multimodal performance space representation could be further enhanced with additional modalities and features and could be used as criteria not only during the ATCO selection but also during the ATCO training processes.

Using the multimodal system for psychophysiological measurements described in [3], various performance criteria could be more objectively and comprehensively assessed, than in relying exclusively on ATCO instructors' evaluations of the candidates. For example, performance criteria under competency situational awareness could be estimated using the oculometric features calculated from the data collected during the specifically designed computer-based visual task for the situational awareness measurement. Furthermore, communication, as one of the described competencies, could be more objectively measured using the speech analysis. However, the emphasis of this work are the ATC competencies that could be objectively measured using the fNIRS. Examples of such competencies are problem-solving and decision-making as well as workload management. As illustrated in the following paragraphs, these are relatively common constructs of interest within studies using the fNIRS, not only in the ATC-related research but also in cognitive neuroscience in general, since they are highly related to the brain regions specifically targeted by fNIRS measurements, i.e. the prefrontal cortex (PFC). The fNIRS is an emerging simple, non-invasive and portable neuroimaging technique that provides a good insight into the PFC activation during task-related stimuli.

The PFC is part of the frontal cortex that is located at the very front of the human brain. It is critical in situations when the mapping between the sensory inputs, thoughts, and actions is needed [21]. The PFC plays a crucial role in the most complex aspects of human thought, such as reasoning and planning [22]. The dorsolateral part of the PFC is related to working memory [23] which enables mental manipulation of information over a short period of time [24]. The working memory is a key executive function for many professions, including the ATC and flying an aircraft. This function is particularly critical when pilots have to recall a series of ATCO's instructions because working memory limitations may jeopardize flight safety [25]. Following ATCO's instructions is known as one of the hardest tasks for working memory as it requires memorisation of critical flight information (e.g. heading, altitude, speed) and the input of these parameters into the flight deck [25]. The participants with better working memory performance experienced higher levels of PFC activation measured with the fNIRS [26].

Relevant cognitive and sensory attributes required for high-performance levels at ATC radar workstations are spatial scanning, movement detection, image and pattern recognition, prioritizing, visual and verbal filtering, coding and decoding, inductive and deductive reasoning, short- and long-term memory, and mathematic and probabilistic reasoning [27]. Most of these attributes are controlled by the PFC and their measurements with brain imaging technique could provide more objective insight into ATCOs' performance.

Besides studies conducted on the ATCO professionals, veterans, trainees, and candidates, fNIRS is an emerging tool in the research of cognitive neuroscience in other professions, such as pilots [28]. The fNIRS measurements were sensitive to variations in task difficulty in both flight simulator and laboratory settings, showing increased activation in the PFC as the tasks became more complex. The paradigms for the fNIRS measurements are often based on arithmetic tasks, making it the perfect tool in assessing individuals' problem-solving and decision-making strategies. Besides the arithmetic tasks, other simple tasks, such as the Stroop task, make the scenario for decision-making strategies. The workload is already broadly researched within the ATC field using both fNIRS and other brain imaging techniques [7, 8, 9, 10, 11, 16]. Altogether, the literature reviewed in this section demonstrates the potential of the fNIRS brain-imaging-based performance metrics in the ATC, which calls for their further research and development.

IV. METHODS AND PARTICIPANTS

Thirty-three students from the University of Zagreb (mean age: 24.23, SD: 1.03) signed informed consent statements and participated in this study. The aim of the study was to assess participants' performance using the fNIRS. In the Stroop test [29] subjects either read words or name the colour in which the words are written, with three possible variations of the task – neutral, congruent, and incongruent. Neutral Stoop task is the task in which only the text of the colour is displayed, usually in black or white colour. The congruent task is the task in which the meaning of the word and the colour of the word are the same (e.g. word "red" written in red), while the incongruent task is the task in which the colour of the word and the meaning of the word differ. Stroop test measures inhibition, attention, and processing speed. The test is challenging when naming the colour of a conflicting stimulus, i.e. the incongruent Stroop task (e.g. the word "green" displayed in red) because there is a strong prepotent tendency to read the word ("green"), which competes with the response to the colour ("red"). This illustrates one of the most fundamental aspects of cognitive control and goal-directed behaviour: the ability to select a weaker, task-relevant response (or source of information) in the face of competition from an otherwise stronger, but task-irrelevant one [21].

The Stroop task conducted in this experiment was designed as a block paradigm, consisting of two neutral blocks, two congruent and incongruent blocks, and five resting blocks. Each block lasted for 30 seconds, with 20 tasks in each of the task blocks, which made this experimental paradigm 4.5 minutes long. The sequence of the blocks is shown in Figure 1. During the neutral Stroop task, all the words were presented in white colour on a black screen, and the participants had to match the meaning of the word with one of the two words presented at the bottom of the screen by clicking the left or right arrow on the keyboard. During the incongruent task, the participants had to match the colour of the word with one of the two colours presented at the bottom of the screen. Figure 2 shows the illustration of these tasks. Colours used in this task were red, green, blue, yellow, and white, all presented on a black background.

The fNIRS Imager 1100 with 16 optodes and COBIStudio software [30] were used for data acquisition. The system consists of four LED diodes emitting light at two different wavelengths (730 and 850 nm), and ten light detectors. The imager was placed over the participants' PFC with light-source separation of 2.5 cm. Oxygenated and deoxygenated haemoglobin concentration rates were

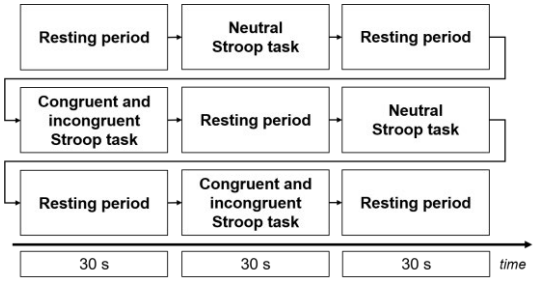


Figure 1 Sequence of the blocks during the experimental paradigm



Figure 2 Neutral (left) and incongruent task (right) examples (written in Croatian: "plava"="blue", "žuta"="yellow")

measured with a sampling rate of 2 Hz. During the experiment, the participants were seated facing a screen showing the tasks and a keyboard for their response. The 10 seconds baseline preceded the recording phase. The participants were instructed to avoid movements of their head and body. They were also instructed that their reaction time and accuracy will be measured during the Stroop task. The participants were shown examples of the tasks prior to the scanning.

The collected raw data were transformed into oxygenated (HbO) and deoxygenated (HbR) haemoglobin levels using the modified Beer-Lambert law [31]. Visual inspection of the collected signals was required since some of the signals or signal segments ended up rejected due to poor signal quality, which is usually a result of bad fNIRS headband placement. Haemoglobin concentrations were then filtered in order to remove physiological artefacts such as heartbeats and other motion-related artefacts. Motion artefacts were removed using the sliding-window motion artefact rejection algorithm [32]. These cleaned signals were used in the subsequent analysis. The signals were normalised and averaged across 16 optodes and participants. Mean values of the signal across all the segments were extracted from the haemoglobin concentration. Since the signals HbO and HbR were strongly negatively correlated, in this paper only the HbO is analysed.

V. RESULTS AND DISCUSSION

The mean activation in the PFC was higher in blocks of congruent and incongruent Stroop tasks compared with the activation in the resting periods and neutral Stroop tasks. The averaged signal of the PFC activation across all the participants is shown in Figure 3. The congruent and incongruent task segments present more complex decision-making and problem-solving tasks, compared with the neutral Stroop task and resting periods, which is shown in the increase of the PFC activation.

Besides the calculated mean brain activation, other features can be computed from the oxygenated and deoxygenated haemoglobin levels. An example of such features is given in the [33]. This set of features was used in the classification of task segments, i.e. classification between the resting periods and task segments. The signal slope is a feature that might significantly contribute to the task classification accuracy, since the signal slope is mostly negative during the resting blocks after the congruent and incongruent tasks, and positive in the congruent and incongruent task blocks.

Using the Stroop test in ATC-related research is not uncommon [3, 34]. In a study by Arbula et al. [34] a group of ATCO trainees was compared with a matched group of university students and longitudinally tested measuring the psychometric responses in Stroop paradigms. The

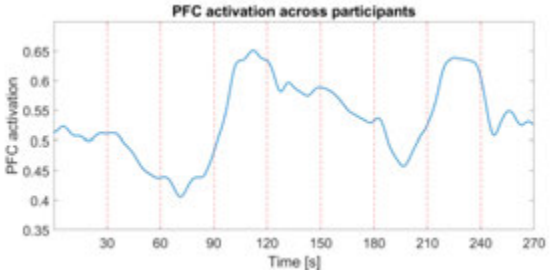


Figure 3 Mean PFC activation across all the participants

results showed that ATCO trainees had substantially reduced Stroop interference effect, both prior to and during their ATCO training phase, indicating that ATCO selection and training resulted in changes in their performance with respect to the control group.

ATCO candidates' speech features from the paradigm that consisted of three variations of Stroop tasks obtained during their ATCO selection process were shown in [3]. An example of these features is the speech loudness, calculated as the difference between the average value of loudness across the incongruent Stroop condition (higher cognitive load) and the average value of loudness across the congruent Stroop condition (lower cognitive load). Besides this feature, the psychometric Stroop test performance-based error rates were also calculated. While the psychometric performance-based features can be intuitively interpreted, i.e. lower error rates mark better performance, the lower speech loudness indicates either lower sympathetic responses, or higher parasympathetic responses, or better balance of autonomic nervous system, making those candidates better suited for highly stressful professions.

In this study, the averaged and normalised PFC activation shows higher brain activation in more complex decision-making and problem-solving tasks, despite the variability between participants shown in Figure 4. This figure shows two participants' PFC activation during the experimental paradigm, one with the highest and the other with the lowest psychometric performance accuracy score in the tasks. The features extracted from their brain signals significantly differ since the participant with the lowest accuracy score has a constant decrease of the PFC activation, while the participant with the highest accuracy score has an increased PFC activation response in task segments where the cognitive demand was the highest during the experimental paradigm. Those brain activation variabilities present valuable information about the participants' strategies and performance, since the overall psychometric performance, i.e. task accuracy, is in the

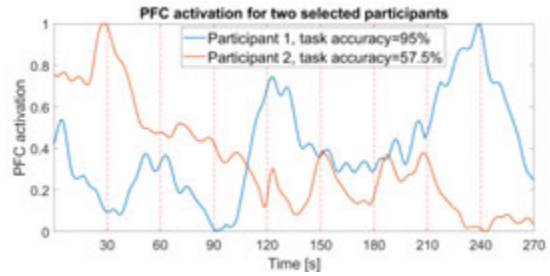


Figure 4 PFC activation for two selected participants with the best and the worst accuracy

literature already correlated with the brain activation [35]. Besides that, brain activation patterns might indicate a withdrawal in the performed task, visible in the decrease of the brain activation during the task.

The fNIRS, as a technique for brain imaging and performance evaluation, demonstrated the possibility in understanding the ATCO's decision-making and problem-solving tasks overall performance, since both attributes are a construct of interest in the Stroop test and already reported in literature [36]. Besides the decision-making and problem-solving, the effect of workload can be evaluated on other simple paradigms. Previous literature demonstrated the possibilities in the objectivization of the workload measurement using the fNIRS [9, 10, 16]. The results obtained from the brain imaging activation indicate that fNIRS might be a useful tool in the ATCO selection and training procedures in the future, since the specific brain activation patterns may indicate better performance on the Stroop task, according to the Figure 4 and [35], as well as testing the performance on the Stroop task in the ability to select a weaker task-relevant response, which is extremely important in the ATC profession [34].

VI. CONCLUSION

The human performance assessment in the aviation industry using brain imaging techniques is a common research topic [16]. Various cognitive and sensory attributes could be more objectively measured using brain imaging and other multimodal neuropsychophysiological measurements [3]. Direct measurements of ATCOs' performance based on brain imaging techniques, such as fNIRS, and other emerging wearable technologies can improve current ATCO selection and training processes. Previous research has emphasised the need to combine different multimodal features as a logical step forward in the enhancement of the ATCO selection protocols, thus minimising the relatively high attrition rates in the ATCO training [3].

This paper describes the ATC competencies that are today evaluated by ATCO instructors and for that reason might be affected with various biases, suggesting the possible neuropsychophysiological methods for improving their objectivization. The example of the experimental paradigm measuring the brain activation using the fNIRS device is also shown in this paper.

Monitoring the cognitive workload in ATCOs, as well as in other highly stressful professions, such as pilots, military personnel, etc., may enable an increase in safety by ensuring the optimal cognitive arousal in these high-risk fields. The future of the ATC technology development may result with the assistance system that would be able to objectively evaluate ATCOs' performance with higher accuracies based on their multimodal responses. The measuring system based on fNIRS allows continuous and objective monitoring and may be a valuable tool in such assistance system. The simplicity of the described measuring system may allow its possible future implementation even in real-time as part of the ATCO assistance system, which would minimise the errors that may occur in this complex environment.

LITERATURE

[1] Eurocontrol. Frequently Asked Questions: How many candidates pass the tests? Accessed: Jan. 13, 2020. [Online]. Available: https://atco.eurocontrol.int/.

[2] K. Ćosić, M. Šarlija, V. Ivkovic, Q. Zhang, G. Strangman, and S. Popović, "Stress resilience assessment based on physiological features in selection of air traffic controllers," IEEE access, 7, 41989-42005., 2019.

[3] K. Ćosić et al., "New Tools and Methods in Selection of Air Traffic Controllers Based on Multimodal Psychophysiological Measurements," IEEE Access, 7, 174873-174888., 2019.

[4] D. Broach, "Attrition in US Air Traffic Control Specialist (ATCS) Training: A Review of 50 Years of Data." In 19th International Symposium on Aviation Psychology (p. 542), 2017.

[5] T.R. Carretta and R.E. King, "Improved military air traffic controller selection methods as measured by subsequent training performance," Aviation, space, and environmental medicine, 79(1), 36-43, 2008.

[6] N. De Koninck, "Reaching out to the next generation: Student air traffic controller selection & training process," World ATM Congress Madrid, 7 March 2017.

[7] P. Aricò, G. Borghini, G. Di Flumeri, A. Colosimo, S. Pozzi, and F. Babiloni, "A passive brain–computer interface application for the mental workload assessment on professional air traffic controllers during realistic air traffic control tasks," In Progress in brain research (Vol. 228, pp. 295-328), Elsevier, 2016.

[8] P. Aricò et al., "Adaptive automation triggered by EEG-based mental workload index: a passive brain-computer interface application in realistic air traffic control environment," Frontiers in human neuroscience, 10, 539, 2016.

[9] P. Aricò et al., "Human factors and neurophysiological metrics in air traffic control: a critical review," IEEE reviews in biomedical engineering, 10, 250-263, 2017.

[10] H. Ayaz et al. "Cognitive workload assessment of air traffic controllers using optical brain imaging sensors," Advances in understanding human performance: Neuroergonomics, human factors design, and special populations, 21-31, 2010.

[11] G. Di Flumeri et al., "On the use of cognitive neurometric indexes in aeronautic and air traffic management environments," In International Workshop on Symbiotic Interaction (pp. 45-56). Springer, Cham, October 2015.

[12] M. Hagmüller, G. Kubin, and E. Rank, "Evaluation of the human voice for indications of workload induced stress in the aviation environment," Eurrocontrol Experimental Centre, EEC Note No. 18/06, 2005.

[13] K. Moore and L. Gugerty, "Development of a novel measure of situation awareness: The case for eye movement analysis," In Proceedings of the human factors and ergonomics society annual meeting (Vol. 54, No. 19, pp. 1650-1654), Sage CA: Los Angeles, CA: SAGE Publications, September 2010.

[14] A.J. Tattersall and P.S. Foord, "An experimental evaluation of instantaneous self-assessment as a measure of workload," Ergonomics, 39(5), 740-748, 1996.

[15] I. Mijić, M. Šarlija, and D. Petrinović, "MMOD-COG: A database for multimodal cognitive load classification," In 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA) (pp. 15-20), IEEE, September 2019.

[16] J. Harrison et al., "Cognitive workload and learning assessment during the implementation of a next-generation air traffic control technology using functional near-infrared spectroscopy," IEEE Transactions on Human-Machine Systems, 44(4), 429-440, 2014.

[17] J. Benerradi, H.A. Maior, A. Marinescu, J. Clos, and M.L. Wilson, "Exploring Machine Learning Approaches for Classifying Mental Workload using fNIRS Data from HCI Tasks," In Proceedings of the Halfway to the Future Symposium 2019 (pp. 1-11), November 2019.

[18] G.F. Wilson and C.A. Russell, "Real-time assessment of mental workload using psychophysiological measures and artificial neural networks," Human factors, 45(4), 635-644, 2003.

[19] Eurocontrol, "ATC Refresher Training Manual", Edition 1.0, Edition date: 06/03/2015.

[20] ICAO, "NGAP Implementation Workshop Introduction to PANS-TRG Document" (ICAO DOC 9868).

[21] E.K. Miller and J.D. Cohen, "An integrative theory of prefrontal cortex function," Annual review of neuroscience, 24(1), 167-202, 2001.

[22] R. Levy and P.S. Goldman-Rakic, "Segregation of working memory functions within the dorsolateral prefrontal cortex," In Executive control and the frontal lobe: Current issues (pp. 23-32), Springer, Berlin, Heidelberg, 2000.

[23] H. Ayaz, P.A. Shewokis, S. Bunce, K. Izzetoglu, B. Willems, and B. Onaral, "Optical brain monitoring for operator training and mental workload assessment," Neuroimage, 59(1), 36-47, 2012.

[24] F. Al-Shargie, M. Kiguchi, N. Badruddin, S.C. Dass, A.F.M. Hani, and T.B. Tang, "Mental stress assessment using simultaneous measurement of EEG and fNIRS," Biomedical optics express, 7(10), 3882-3898, 2016.

[25] T. Gateau, G. Durantin, F. Lancelot, S. Scannella, and F. Dehais, "Real-time state estimation in a flight simulator using fNIRS," PloS one, 10(3), 2015.

[26] Y. Ogawa, K. Kotani, and Y. Jimbo, "Relationship between working memory performance and neural activation measured using near-infrared spectroscopy," Brain and behavior, 4(4), 544-551, 2014.

[27] G. Costa, "Occupational stress and stress prevention in air traffic control," Geneva: International Labour Office, 1996.

[28] M. Causse, Z. Chua, V. Peysakhovich, N. Del Campo, and N. Matton, "Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS," Scientific reports, 7(1), 1-15, 2017.

[29] J.R. Stroop, "Studies of interference in serial verbal reactions," Journal of experimental psychology, 18(6), 643, 1935.

[30] H. Ayaz, P.A. Shewokis, S.C. Bunce, and B. Onaral, "U.S. Patent No. 9,946,344," Washington, DC: U.S. Patent and Trademark Office, 2018.

[31] W.B. Baker, A.B. Parthasarathy, D.R. Busch, R.C. Mesquita, J.H. Greenberg, and A.G. Yodh, "Modified Beer-Lambert law for blood flow," Biomedical optics express, 5(11), 4053-4075, 2014.

[32] H. Ayaz, M. Izzetoglu, P.A. Shewokis, and B. Onaral, "Sliding-window motion artifact rejection for functional near-infrared spectroscopy," In 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology (pp. 6567-6570), IEEE, August 2010.

[33] N. Naseer, F.M. Noori, N.K. Qureshi, and K.S. Hong, "Determining optimal feature-combination for LDA classification of functional near-infrared spectroscopy signals in brain-computer interface application," Frontiers in human neuroscience, 10, 237, 2016.

[34] S. Arbula, M. Capizzi, N. Lombardo, and A. Vallesi, "How life experience shapes cognitive control strategies: the case of air traffic control training," PloS one, 11(6), 2016.

[35] R.H. Grabner, D. Ansari, G. Reishofer, E. Stern, F. Ebner, and C. Neuper, "Individual differences in mathematical competence predict parietal brain activation during mental calculation," Neuroimage, 38(2), 346-356, 2007.

[36] V. Tidikis and I.K. Ash, "Greater Stroop Effect Predicts Better Performance on Creative Insight Problems, But Not on Divergent Thinking Tasks," The international journal of creativity & problem solving, 28(2), 27-37, 2018.

Public transport analyzing using bee colony

Matej Babič*, Božo Soldo**, Janez Povh***,****
*Faculty of Information studies, Novo mesto,Slovenia
** University of North, Croatia
***Faculty of Mechanical Engineering, University of Ljubljana, Slovenia
****Development center, Novo mesto, Slovenia
babicster@gmail.com

Abstract - **Urban public transport is crucial to the functionality of every urban city. Good urban transport affects the functionality of the city itself and of course reduces the number of cars on the roads. The wider Novo mesto region, as a key industrial center, is heavily dependent on urban transport. Unfortunately, urban transport in Novo mesto still has a relatively weak effect on the traffic connections of the wider area. In fact, its functionality has never been systematically scientifically researched.**

Keywords – bee colony, public transport, bicycle

I. INTRODUCTION

Mass transport more effectively solves the problem of increasing the carrying capacity of the street than a personal car. Using public transport as the main way to move around the city can reduce the number of traffic jams. Cars take up too much space in the city. An integrated approach is needed to develop public transport. One of the measures to develop public transport is to change the transport behavior of citizens. Why should it be changed? Because today in our view, a personal car is associated with status and financial well-being, and public transport causes negative associations. For this, coercive measures are applied that would encourage citizens to refuse personal transport in favor of public transport - paid parking (the softest measure, by the way), paid entry, additional taxes and other restrictions. Perhaps the master in this matter is Singapore. In addition, citizens need to form positive associations with public transport. Various holidays, contests, excursions, advertising posters - all this is aimed at creating a positive image of public transport.

Support for public transport and cycling in day-to-day mobility is a priority in large as well as small cities around the world for an effective means of reducing air pollution, congestion and carbon emissions. Common bicycle systems have proven effective in increasing cycling in many urban areas, especially in combination with public transport. Public bicycles are innovative rental or free bicycle schemes in urban areas that can be used for day-to-day mobility as one-way use is possible and can be considered as part of a public transport system. Different from traditional, mostly leisure bike rental services, they provide fast and easy access and have a variety of organizational layout, business models and useful technology for "smart bikes" (smart card or mobile phone rental).

Aims of the research the analysis of the urban transport network in Novo mesto is:

• to gain as accurate an insight into the current state and general functionality of Novo mesto urban transport by means of a survey data collection on the use of meat bus lines,

• make a detailed analysis of the current situation using the approaches and methodology described above,

• attempt to analyze the impact of economic factors on potential changes to urban transport,

• to draw conclusions from the whole in the form of proposals or. "recipes", how could the quality of the service offered by Novo mesto city transport be improved. Of course, these proposals will depend, among other things, on the available transport (or bus) resources. We will try to formulate our results in the most general way possible so that they are useful regardless of actual resources.

II. METHDLOGY

GoNM is an automatic bicycle rental system in the City of Novo mesto. The system makes it easy to rent and ride bikes currently available at 14 stations. To rent a bike, you need to have a proper card and be registered. The card then picks up the bike at one of the stops and places it at one or the other stop. We will introduce modeling methods that describe the behavior of a bike path or the destination of a bike trip. This urban transport planning process often treats bicycles and pedestrians as a system of passenger network. It is therefore necessary to develop a methodology for estimating and analyzing bicycle rental demand.

Colonies of honey bees have been of interest to mankind for many years, these creatures are recognized as one of

the most amazing creations of wildlife. A clear organization of the life of bees in the colonies, the correct geometry of honeycombs, the cultivation of millions of offspring and hundreds of other factors have led to the beginning of modern scientific research that gives us an understanding of one of the most unusual structures. Studying the processes of the brain and the visual functions of these insects makes it possible to create new technical devices and introduce modern technologies, such as Face-ID, using the amazing properties of bees in their modeling, which is certainly relevant for modern society. The main goal of the research work was to identify the psychological characteristics of bees, manifested in the behavior of each individual.

The main steps of the algorithm are given below.

- Initial food sources are produced for all employed bees
- REPEAT
    - Each employed bee goes to a food source in her memory and determines a closest source, then evaluates its nectar amount and dances in the hive
    - Each onlooker watches the dance of employed bees and chooses one of their sources depending on the dances, and then goes to that source. After choosing a neighbour around that, she evaluates its nectar amount.
    - Abandoned food sources are determined and are replaced with the new food sources discovered by scouts.
    - The best food source found so far is registered.
- UNTIL (requirements are met)

III. RESULTS AND DISCUSSION

In Fig. 1 is presented 14 stations in Novo mesto, Slovenia, where it is possible to rent a bike with network. So, Fig. present Bicycle Rental and Return Network by 14 different GoNM stations for all days. Bicycle network has 14 nodes and 28 edges between stations. So, we have problem how to visit all nodes in this network one time. Fig. 2 Optimal way for transport and pick up bicycles across 14 stations but using bee colony.

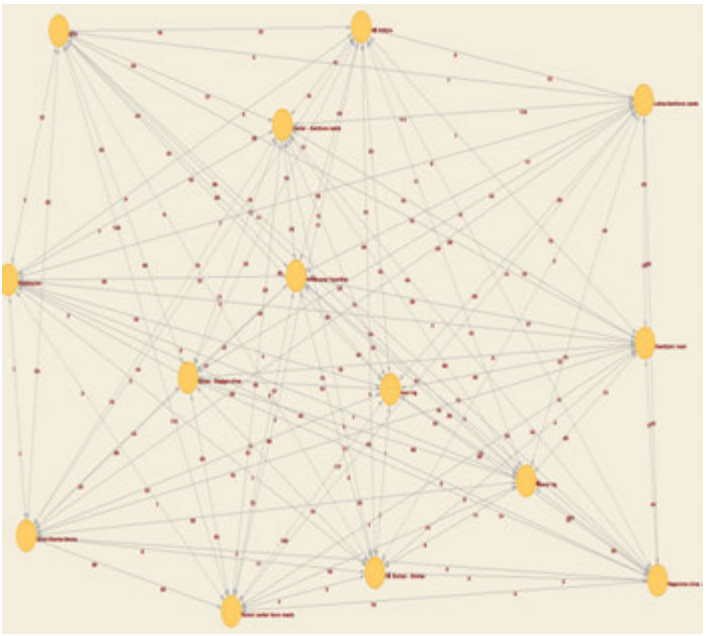


Figure 1. Bicycle Rental and Return Network by 14 different GoNM stations for all days

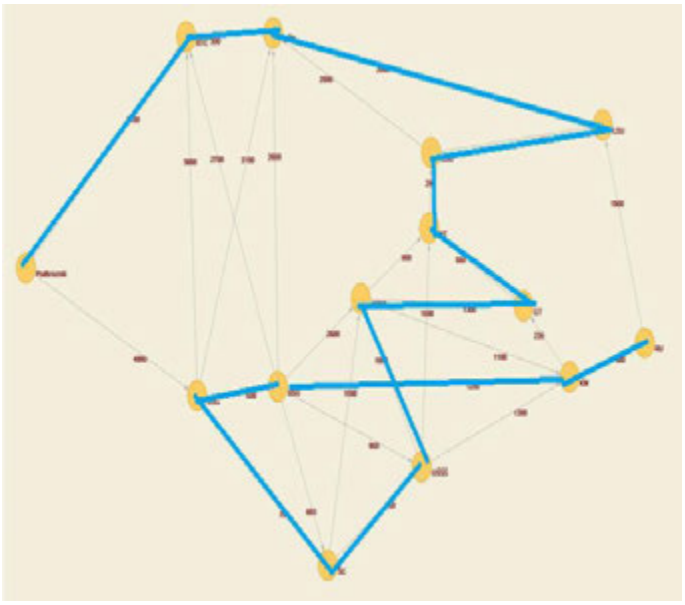


Figure 2. Optimal way for transport and pick up bicycles across 14 stations but using bee colony

We use bee colony for optimizing path in city Novo mesto, Slovenia to transport and pick up bicycles across 14 stations.

IV. CONCLUSSION

In this paper, we use intelligent public transport using bee colony to transport and pick up bicycles across 14 stations. Locations where public bikes can be found in the city center should be easy for the user to find. The

location of public bicycles must be well planned according to the expected demand. Existing examples show that renting bicycles to users during the day often leads to an uneven distribution of bicycles throughout the city. In this case, it is necessary to redistribute the bikes to ensure the availability of the bikes, as this avoids frustration for users who cannot find any bikes or cannot leave the bikes at full racks. Public bicycles must be available on the website of the important public transport station. This allows the combined use of bicycles and public transport services and increases the attractiveness of the system.

ACKNOWLEDGMENT

The investment is co-financed by the Republic of Slovenia and the European Union from the European Regional Development Fund.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] Buehler, R. & Pucher, J. (2009) Sustainable Transport Freiburg: Lessons from Germany’s Environmental Capital, International Journal of Sustainable Transportation, Vol 5, pp. 42-70.

[3] Buehler, R. & Pucher, J. (2011) Sustainable Transport in Freiburg: Lessons from Germany’s Environmental Capital. International Journal of Sustainable Transportation, Vol. 5, pp.43–70.

[4] Anthill Art (12 March 2015). "Largest Aluminum Fire Ant Colony Cast So Far (Cast #072)".

[5] Foitzik, S.; DeHeer, C. J.; Hunjan, D. N.; Herbers, J. M. (7 June 2001). "Coevolution in host-parasite systems: Behavioural strategies of slave-making ants and their hosts". Proceedings of the Royal Society B: Biological Sciences. 268 (1472): 1139–1146.

[6] Carse, A., Goodman, A., Mackett, R.L., Panter, J. & Ogilvie, D. (2013) The factors influencing car use in a cycle-friendly city: the case of Cambridge, Journal of Transport Geography, Vol. 28, pp. 67–74.

[7] Delmelle, E. M. & Cahill Delmelle, E. (2012) Exploring spatio-temporal commuting patterns in a university environment, Transport Policy, Vol. 21, pp. 1–9.

[8] Emanuel, M. (2012) Trafikslag på undantag – Cykeltrafiken i Stockholm 1930-1980.

[9] Heinen, E., van Wee, B. & Maat, K. (2009) Commuting by Bicycle: An Overview of the Literature, Transport Reviews, Vol. 30(1), pp. 59-96.

[10] Heinen, E., Maat, K. & van Wee, B. (2013) The effect of work-related factors on the bicycle commute mode choice in the Netherlands, Transportation, Vol. 40, pp. 23–43.

[11] Koglin, T. & Rye, T. (2014) The marginalisation of bicycling in Modernist urban transport planning. Journal of Transport and Health, Vol. 1, pp. 214-222.

[12] Krenn, P., Oja, P. & Titze, S. (2015) Development of a Bikeability Index to Assess the Bicycle-Friendliness of Urban Environments, Open Journal of Civil Engineering, Vol. 5, pp. 451-459.

[13] Marsden, G., Frick, K. T., May, A. D. & Deakin, E. (2011) How do cities approach policy innovation and policy learning? A study of 30 policies in Northern Europe and North America, Transport Policy, Vol. 18, pp. 501–512.

[14] Pucher, J. & Buehler, R. (2007) At the Frontiers of Cycling: Policy Innovations in the Netherlands, Denmark and Germany, World Transport Policy & Practice, Vol. 3, pp. 8-56.

Measure for Traffic Anomaly Detection on the Urban Roads Using Speed Transition Matrices

L. Tišljarić*, Ž. Majstorović*, T. Erdelić*, T. Carić *
* Faculty of Transport and Traffic Sciences, University of Zagreb, Vukelićeva 4, 10000 Zagreb, Croatia
ltisljaric@fpz.unizg.hr

***Abstract* – Road traffic anomaly detection is an essential research topic within the Intelligent Transport System (ITS) context. Urban road anomaly detection systems are a crucial part of the ITS regarding the trip planning, road security, and congestion estimation applications. In this paper, the method for traffic anomaly detection using Speed Transition Matrices (STM) is presented. The paper's main goal is to present the novel method for measuring the distance between two STM, as standard distance measures are inapplicable for the anomaly detection and road traffic analysis interpretation The method is based on the Euclidean distance measure between STM's Center of Mass (COM), and the average STM that represents normal traffic conditions. The Global Navigation Satellite System (GNSS) data on the road-network of City of Zagreb were used as a case study, as it is, the capital and the largest city in Croatia, suitable for the application of the proposed methodology. The anomaly detection method resulted in 73 anomalous points which are presented on the digital map. The proposed method is compared to the other distance metrics used in the literature, and advantages over each of the metrics are highlighted.**

Keywords – Anomaly detection; Speed transition matrix; Quantifying anomaly; Intelligent transport systems; Traffic state estimation; Urban road network

I. INTRODUCTION

As a process of detecting potentially dangerous events, anomaly detection is a crucial part of the ITS [1]. The importance of anomaly detection lies in the potential useful actionable information for traffic management systems. Anomalous traffic in an urban road network could indicate a severe traffic accident, traffic congestion, or a violation of the regulations.

The focus of this research is on detecting the recurrent anomalies on the urban road networks. In [2], authors report that recurrent anomalous traffic is covering around 85% of all congestion on the urban road networks. Although, the algorithms for anomaly detection in other research areas have reached high efficiency and performance, the development of traffic anomaly detection methods in urban areas is still in the early stages. The authors in [3] present the review on the traffic anomaly detection and report that this is an early-stage research area, and further research is required in every direction.

In this paper, a method for anomaly detection based on the STM is proposed. The methodology consists of the three mains steps: (i) data preprocessing and STM generation, (ii) COM estimation, and (iii) measure of distance between STMs. For the STM generation, a large real-life GNSS data are used. Anomaly detection is based on the measure of distance between the COMs of the observed STMs and the median STM, which represents normal traffic conditions.

Contributions of this paper are as follows: (i) novel methodology for the anomaly detection based on the COM measurement and (ii) results of the anomaly detection are analyzed on the urban road network in the City of Zagreb, Croatia.

The rest of the paper is organized as follows. Section 2 presents the review on the literature regarding the anomaly detection method, and data used to model the traffic parameters. Section 3 presents the methodology for the anomaly detection based on the measure of distance between STMs' COMs. Section 4 presents the case study results on the urban road network in the City of Zagreb. Alongside the anomaly detection results, a comparison between standard and most used distance metrics is conducted. In section 5, the future work, and the conclusion is given.

II. RELATED WORK

In [4], the authors developed a framework for detecting multiple public events which were differentiated from regular crowd moving patterns. The method is based on the tensor factorization where crowd movement data, collected from the bike trips in a bike-sharing system, is augmented with semantic social activity data. In [5], the authors relied on multiple data sources to demonstrate the advantages of the proposed method to detect anomalies with unknown external influences, and to detect an anomaly before it reaches its peak. The tensor decomposition methods [6] are used by authors in [7] and [8] for traffic data modelling, which contained spatial and temporal data, and the Local Outlier Factor algorithm, to successfully detect the anomalies, and to determine the time interval of the anomaly for the certain area.

The authors in [9] based their research on a data collected from the traffic loop sensors and a dictionary-based compression theory for analyzing the traffic data to identify the features of the spatial and temporal patterns. The proposed method resulted with clear geographically distributed features of spatial and temporal traffic flow pattern, represented as a heatmap, where anomalies can be easily identified. Taxi vehicles are also a valuable source of the GPS data considering the nature of their activities. In [10] authors used Taxi GPS data as traffic flow data to build traffic flow matrix and proposed anomaly detection method based on the wavelet and principal component analysis for

detecting traffic anomalies in urban subregions. The proposed method proved to be effective in discovering anomalies in the two adjacent regions as well as identifying traffic flows that cause the traffic anomalies. The authors in [11] also used GPS data collected from taxi vehicles and proposed deviation-based outlier detection method, which focuses on the road segments instead of paths. The proposed method resulted in detection of a large number of events (anomalies), where each event was associated with the road segment and time interval when the event happened.

Simple features such as speed or distance can be used as the anomaly indicator as well. Based on the physical distance between the origin and the destination, the authors in [12] developed a method which long taxi route marks as an anomalous event. The authors in [13] used the vehicle speed recorded by the mobile device sensors to detect speed bumps and road damages, which turn out to represent anomalies in urban roads. In [14], the authors detected detect traffic jams based on the estimated free-flow speed and actual speed for each road segment.

III. METHODOLOGY

Given the set of precomputed STMs, this paper aims to obtain a method for traffic anomaly detection on the urban road network. The proposed methodology is presented in Figure 1. This section briefly describes the three main steps of the method that includes data preprocessing, STMs generation, and anomaly detection based on the COM estimation.

A. Speed Transition Matrix

In most of the literature, the traffic data are represented by using a time series vector [15]. In some recent papers, the traffic data are represented as images by using a two-dimensional matrix that represents the spatial and temporal characteristics of the traffic flow [16], [17].

In this paper, the STM is used for the representation of traffic data. The STM is a matrix that represents a probability of changing the speed value when a vehicle travels between two consecutive road segments. Similar data modeling was used in [18] for predicting the traffic density by using the GNSS data. The change of speed between two consecutive road network segments is called a transition. The transition can be defined as a change in vehicle trajectory when traveling between two consecutive road segments r_i and r_j in the time interval t. Then, the average speed on the input road segment r_i is labeled as source speed v_o, and the average speed on the output segment r_j is labeled as destination speed v_d.

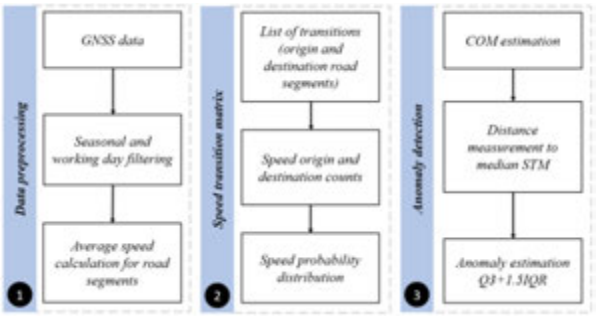
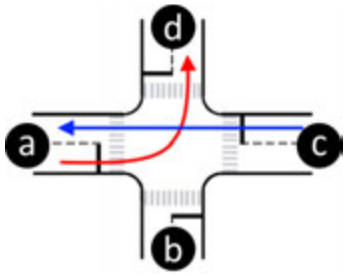
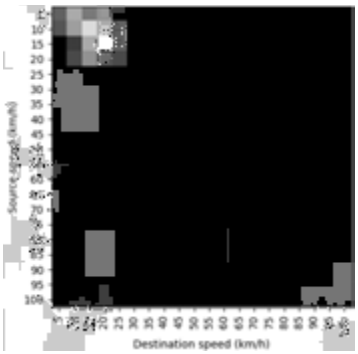


Figure 1. Methodology for the anomaly detection using STMs

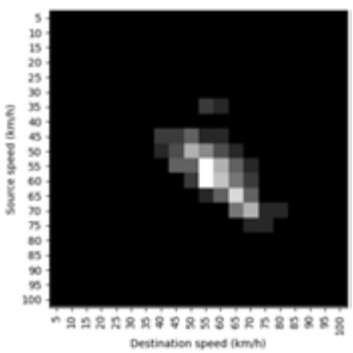
Two examples of the transition are visually presented in Figure 2a). The first transition can be presented with the vehicle traveling between road segments a and d (red), and the second transition with the vehicle traveling between the road segments c and a (blue). Source speeds v_o are average speeds on the edges a and c, and destination speeds v_d are average speeds on the road segments d and a for the corresponding transitions. The STM can be constructed by counting the v_o and v_d for the time interval under observation. Then, the speed counts are transformed into the speed transition probability distribution to get the probabilities for every transition. Figure 2b) and c) show the examples of the STMs that represent anomalous and normal traffic conditions. On the x-axis are the destination speed, and on the y-axis is the source speed, with the discretization period of 5 km/h. Based on the speed count, the probability matrix $\boldsymbol{X}(t)$ is computed. The dimensions of the matrix depend on the chosen resolutions (sensitivity)



a) Transitions on the simple road network



b) STM containing the anomaly



c) STM representing the normal traffic flow

Figure 2. Examples for two types of transition of the STMs containing the anomaly and the normal traffic flow

of the speed change and the maximal speed that can be captured:

$$\boldsymbol{X}(t) = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ p_{m1} & \cdots & \cdots & p_{mn} \end{pmatrix} \quad (1)$$

where p_{ij} presents the probability of the observed transition that vehicle had source speed v_o and destination speed v_d in the time interval t. For this paper, 5 km/h is the chosen discretization period and 100 km/h for the maximal possible speed, which resulted in matrix dimensions 20×20. The specific maximal speed value is chosen because case study is conducted on the road segments with a speed limit between 50 and 80 km/h.

B. Quantifying anomaly

Determining the COM is a key factor for detecting anomalous road segments. It can be observed that the position of the grouped data in the STMs indicates the traffic conditions on the observed transitions. If grouped data are positioned further from the center of the matrix, an anomalous event can be detected. For example, if the data are grouped in the far upper left corner, the STM shows that the observed speed transitions are extremely low, which can indicate an anomaly (Figure 2b). Similar conclusions can be given if the data are grouped in the other corners of the STM.

For COM estimation, a method based on the determination of the expected values is selected [19]. First, the marginal distributions for the x and y coordinates of the STM are calculated. Then, the coordinates of the COM are calculated as follows:

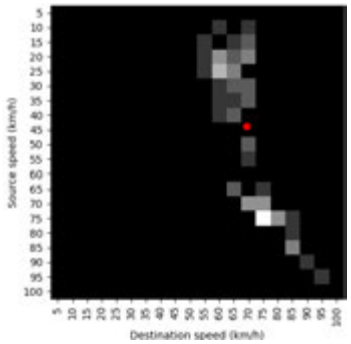
$$c_x = \sum_{i=1}^{20} p_x(x_i) \cdot i \quad (2)$$

$$c_y = \sum_{j=1}^{20} p_y(y_j) \cdot j \quad (3)$$

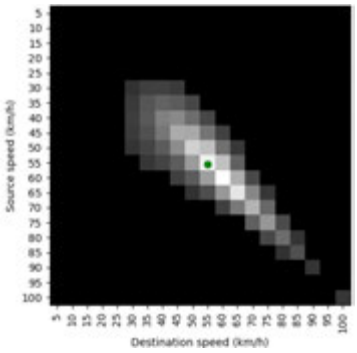
where $p_x(x_i)$ represents the value of the marginal distribution of the x_i coordinate, i represents the index of the x coordinate, $p_y(y_j)$ represents the value of the marginal distribution of the y_i coordinate, j represents the index of the y coordinate. Applying the equations (2) and (3) on the STMs gives the result as a coordinate of COM denoted as c_x and c_y. The examples of the COM estimation are shown in Figure 3 represented as red and green dots.

Figure 3a) shows an example of STM with determined COM, and figure 3b) shows the results of the COM estimation on the median STM. The median STM is used as the representation of the normal traffic on the observed transitions. The median matrix is characterized by the speeds grouped diagonally from the center of the matrix to the lower right corner, which indicates high-speed values close to the speed limit.

With known coordinates of COM, Euclidean distance can be calculated between the COM coordinates of the STM under observation and the COM coordinates of the median STM:



a) STM with estimated COM (red dot)



b) Median STM with estimated COM (green dot)

Figure 3. Examples of the STMs with estimated COM

$$d\left(c_{STM(i)}, c_M\right) = \sqrt{(c_{x2} - c_{x1})^2 + \left(c_{y2} - c_{y1}\right)^2} \quad (4)$$

where c_{x1} and c_{y1}represent the COM coordinates of the STM under observation denoted as $c_{STM(i)}$ and c_{x2} and c_{y2} represents the COM coordinates of the median STM denoted as c_M. To get more interpretable measure of the anomaly level, Euclidean distance is expressed as the relative distance to the longest distance in a matrix (diagonal):

$$\mathrm{d}_{rel} = \frac{d\left(c_{STM(i)}, c_M\right)}{20\sqrt{2}} \quad (5)$$

where d_{rel} is the relative Euclidean distance. The Euclidean distance is divided with $20\sqrt{2}$ because STM has dimensions 20×20, and its diagonal can be calculated by Pythagorean theorem. This method gives the normalized results in interval $[0, 1]$, and the value quantifies the traffic anomaly. If the result is closer to 1, the speed data are grouped on the edges of the STM, which indicates a significant anomaly. This can indicate significantly the lower or higher speeds on the source and destination road segments represented by the STM in the traffic context. The result closer to zero indicates that data are grouped closer to the COM of the median STM, which represents the normal traffic conditions in the source and destination road segments. This indicates that the traffic flow has the speed closer to the speed limit.

Given the relative Euclidean distance values $\mathrm{d}_{\mathrm{rel}}$, we consider the standard statistical classification of the distance as an outlier [20]:

$$d_{rel} > Q_3 + 1.5 \cdot \mathrm{IQ}R \quad (6)$$

where Q3 represents the value of the d_{rel} that is on the 75 % of the sorted data, and the IQR represents the interquartile range of the data

IV. RESULTS

A. Data

The GNSS data used for the case study are acquired from the vehicles equipped with the tracking devices. The data are provided by the company Mireo Inc. as a part of the SORDITO project [15], [21]. Geographical longitude and latitude, speed, timestamp, and heading are included in every record. Most of the data are sampled with the sampling rate of 100 m for vehicles in driving mode and every 5 min for turned off vehicles. GNSS records are map-matched to the road segments in a digital map of Croatia. The data were recorded in five-year period between August 2009 and October 2014 by approximately 4200 vehicles, which were versatile, mostly consisting of taxi cars and delivery vehicles (vans and caddies). The total number of records was around 6,55 billion. The data were analyzed, and anomalies were detected using the proposed method for a large road segments in the City of Zagreb. The seasonality of the traffic flow is considered to lower the deviation. Summer months, July and August, are excluded from the case study. They significantly influence the results due to the different, and lower traffic flows caused by vacations [22]. Further filtering is done by segmenting the data on working and weekend days. Traffic conditions during the working days are different from the weekend traffic conditions mostly due to the daily commuters. The data used for the case study include only working days with summer months excluded to extract only the most relevant traffic conditions.

B. Anomaly detection

By applying the proposed method, the 73 anomalous matrices out of total 25,188 STMs were detected. Matrices that were declared anomalous, with the relative Euclidean distance closer to 1 (e.g. $d_{rel} = 0.41$) will have data grouped on the edges of STM visible in Figure 4a). On the contrary, matrices with relative Euclidean distance closer to 0 (e.g. $d_{rel} = 0.03$) will have the data grouped around the center of matrix as shown in Figure 4b).

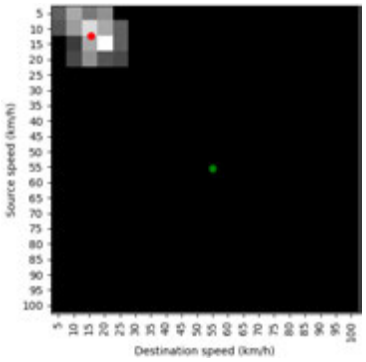
The visualization of detected anomalies on the map of Zagreb is shown in Figure 5. It is noticeable that the anomalies (the red lines) are detected in different parts of the city. Results can be segmented in two types of anomalies: the first type are anomalies detected on the roads such as city's access roads, positioned on the edges of the city, and the second type are the anomalies detected on the city streets closer to the city center. The anomalies detected on the city access roads are the result of people's daily migrations from surrounding cities, and those anomalies occur in peak hours during the morning and the evening rush hour. The second type of anomalies that were detected in the city center is the result of the migration of people within the city. Those anomalies occur not only during the morning and evening rush hour, but also in the other parts of the day. Such an outcome is the result of a high number of people working in the city, a lot of tourists, and availability of wide a variety of entertainment and nightlife.

C. Comparison

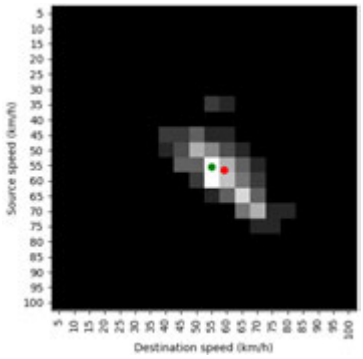
In this section, the proposed method for measuring the distance between STMs is compared to other distance metrics. Table 1. presents the results for the anomaly detection using different distance metrics. The proposed relative Euclidean distance between COMs is compared to the Manhattan, Cosine, and Jaccard distances, which are one of the most used distance metrics for the comparison of the matrices. For the comparison, $Q1$, $Q3$, IQR, $Q3 + 1.5 \cdot IQR$ values, and anomaly count are used.

1) Metrics

Manhattan distance is also known as Taxicab geometry. It is called Manhattan because it is the distance that a car or a taxi would drive in a city such as Manhattan, which has straight streets with intersections at right angles and buildings in square blocks formation. If two points A and B are at the intersections on the same street, the distance between them is measured as a straight line (Euclidean distance), and the measure for the distance is expressed by counting the number of blocks. In another case, if two points A and B are not on the same street, then the distance between two points is the number of blocks taxicab must travel to get from A to B using the shortest possible route. In contrary to the Euclidean distance, where the shortest distance is unique, when using the Manhattan



a) STM with d_{rel} value closer to 1



b) STM with d_{rel} value closer to 0

Figure 4. Examples of the STMs showing the importance of the relative Euclidean distance between COMs in anomaly detection



Figure 5. Visualization of detected anomalies in the City of Zagreb

distance, it is possible to get the shortest distance with multiple paths through the traffic network. Manhattan distance can be calculated as [23]:

$$d_M(STM, M) = \sum_{i=1}^{n} |STM(i) - M(i)| \qquad (7)$$

where d_M is Manhattan distance, $STM(i)$ represents i-th value of the vectorized STM, $M(i)$ represents i-th value of the vectorized values of median STM, and n is the number of samples.

Cosine distance is a distance metric for comparing two datasets where both datasets are represented as a vector with its own direction and magnitude. If two datasets are represented as vectors, the similarity between those two datasets corresponds to an angle between those two vectors, which is quantified as the cosine of the angle between vectors. The Cosine similarity between the two datasets is defined as [24]:

$$d_C(STM, M) = \frac{STM \cdot M}{\|STM\| \times \|M\|} \qquad (8)$$

where Cosine distance is denoted as d_C, and M represents flattened median STM matrix. Result values of Cosine distance are in $[-1, 1]$ range, where value closer to 1 means higher similarity between the compared datasets (vectors with same orientation) and the result with value closer to -1 indicates that compared matrices are opposite (vectors are the same, but with different orientation). Result with value 0 means that the compared matrices are unrelated (orthogonal vectors).

Jaccard distance is the distance metric that measures the distance as similarity of two matrices. It is calculated as a ratio between the number of elements of intersected data between two matrices, and a number of union elements of those two matrices.

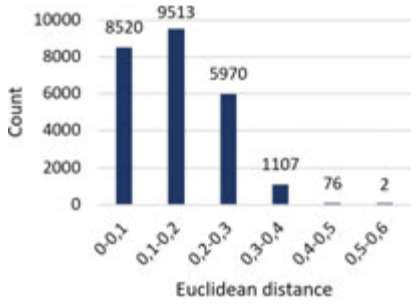
Table 1 - Comparison of the results for the anomaly detection using different distance measures

Measures	d_{rel}	d_M	d_C	d_J
Q1	**0,08**	6363	0,341	0,9822
Q3	**0,21**	6390	0,666	0,9873
IQR	**0,13**	27	0,325	0,0051
Q3+1.5*IQR	**0,405**	6430,5	1,1535	0,99495
Anomaly count	**73**	1339	0	69

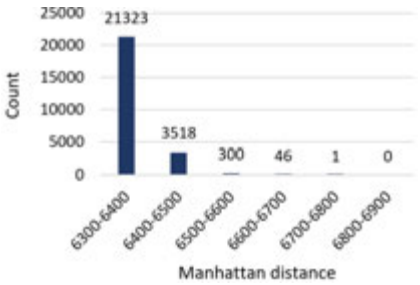
Jaccard distance can be calculated as follows [25]:

$$d_J(STM, M) = \frac{|STM \cap M|}{|STM \cup M|} \qquad (9)$$

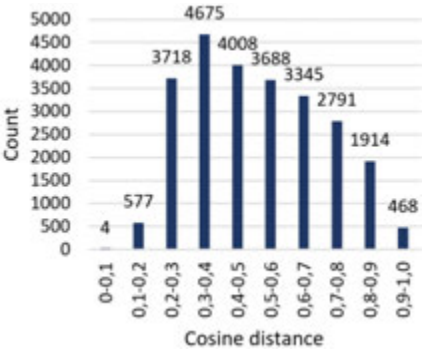
where Jaccard distance is denoted as d_J with the results in range $[0, 1]$. The result closer to 1 indicates that the compared matrices are more similar, and the result closer to 0 indicates that compared matrices are more different.



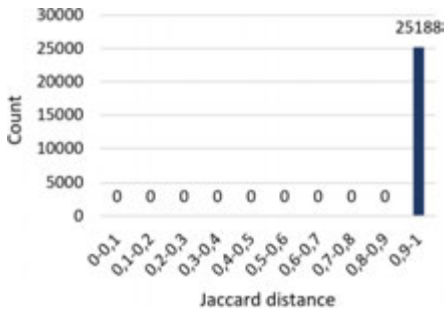
a) Relative Euclidean distance distribution



b) Manhattan distance distribution



c) Cosine distance distribution



d) Jaccard distance distribution

Figure 6. Distributions of the distance measures used for the anomaly detection

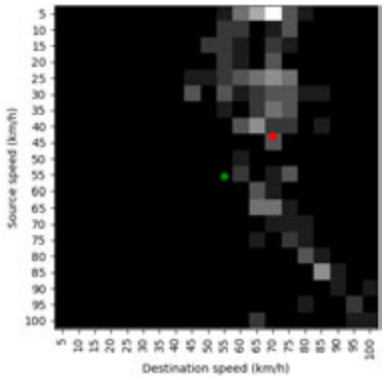
D. Results of the anomaly detection

Alongside the mentioned measures, the comparison is conducted by examining the distributions of distance values and it is presented in Figure 6. By examining the distribution of relative Euclidean distances, it can be noticed that most of the results are grouped in the first half the distributions, which is not the case with other distance measuring methods. The other measuring methods have distributions that are dispersed and/or compressed in certain parts.

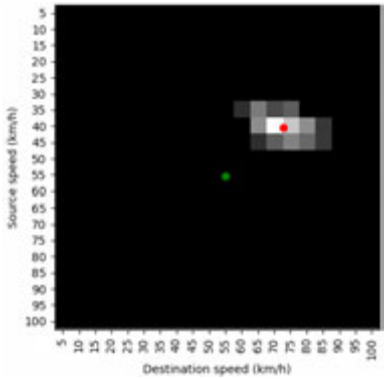
Manhattan distance: Out of 25,188 matrices Manhattan distance detected 1,339 anomalous matrices, which is significantly more than by using relative Euclidean distance method. Considering the distribution of the distance values (Figure 6) it can be concluded that the data are grouped mostly at the beginning of the distribution, which indicates that this measure is not well suited for the anomaly detection based on the STMs. This claim is justified by examining the results in Figure 7. The Manhattan distance for the examples a), and b) are very high, and the anomaly is detected. This is not the correct assumption because on the example a) the speeds are scattered, and on the example b) the destination speeds are high and cannot be classified as the anomaly. On the other hand, relative Euclidean distance for the presented examples is very low, and the anomaly is not detected.

Cosine distance: By using this metric, anomalous STMs cannot be detected with proposed approach because the measure $Q3 + 1.5 \cdot IQR$ resulted in $d_c > 1.15$. This value is not suited for anomaly detection because it exceeds the range of $[-1, 1]$. High values for $Q1$, $Q3$ and high IQR value indicate that the result values are dispersed through entire distribution in range $[0, 1]$ which is visible in the distribution of Cosine distances shown in Figure 6.

Jaccard distance: detected 69 anomalous matrices, which is the result similar to the relative Euclidean distance. Considering the results of statistical processing, $Q1$ and $Q3$ have the values closer to 1 and are very closely spaced with low IQR value. It can be concluded that all the values are grouped in the upper part of the distribution, which is visible in Figure 6. Such grouped result suggests that this method is not appropriate for anomaly detection based on the STMs. The claim is justified by examples of detected anomalous STMs by the Jaccard distance, shown in Figure 8. The Jaccard distances for the examples a) and b) are very high, and the anomaly is detected. Since the speeds are high and scattered on both examples a) and b), they cannot be classified as an anomaly. On the contrary, relative Euclidean distances for the mentioned examples are very low, and anomalies were not detected.



a) STM with scattered speeds and high d_M



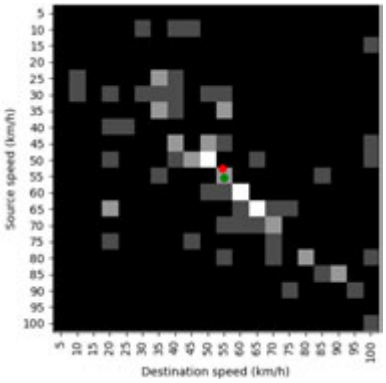
b) STM with high speed and high d_M

Figure 7. Example of anomalous STM according to the Manhattan distance

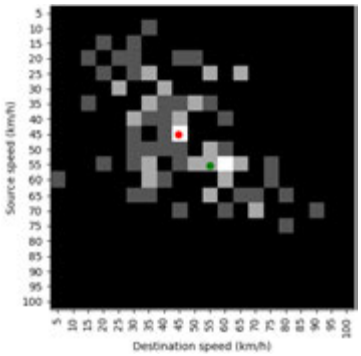
V. CONCLUSION

In this paper, the method for anomaly detection based on the STMs is presented. The method consists of the three mains steps: data preprocessing and STM generation, COM estimation, and distance measurement. As a distance metric, relative Euclidean distance between the COMs of the observed STM and the median STM, which represents the normal traffic conditions, is proposed. The results show that the proposed metric is better suited for the anomaly detection problem based on the STMs because it can use the valuable traffic information, which depends on the position of the grouped data. The method is validated by comparison to other distance metrics from the literature. It showed the best results for the anomaly detection because other distance metrics are not able to use the information about the speeds position in the STM.

Future work will be related to extending the analysis with the temporal components of the anomalies by utilizing the tensor decomposition methods. This type of anomaly detection algorithm will provide more detailed spatiotemporal traffic anomaly information and can be useful for the potentially actionable traffic information, routing applications, or to the road network maintenance authorities.



a) STM with d_J value close to 1 and d_{rel} value close to 0



b) STM with $d_J$value close to 1

Figure 8. Example of detected anomalous STMs using the Jaccard distance

ACKNOWLEDGMENTS

This research has been supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS). Data used for this research is collected during the SORDITO project, European Regional Development Fund under contract RC.2.2.08-0022. Authors are also very grateful to industrial partner MIREO Inc. and the Faculty of Transport and Traffic Sciences.

REFERENCES

[1] P. Škorput, S. Mandžuka, and N. Jelušić, "Real-Time Detection of Road Traffic Incidents," *Promet – Traffic&Transportation*, vol. 22, no. 4, pp. 273–283, 2010.

[2] A. Chow, A. Santacreu, I. Tsapakis, G. Tanasaranond, and T. Cheng, "Empirical assessment of urban traffic congestion," *J. Adv. Transp.*, vol. 48, no. 8, pp. 1000–1016, 2014.

[3] Y. Djenouri, A. Belhadi, J. C. Lin, D. Djenouri, and A. Cano, "A Survey on Urban Traffic Anomalies Detection Algorithms," *IEEE Access*, vol. 7, pp. 12192–12205, 2019.

[4] L. Chen, J. Jakubowicz, D. Yang, D. Zhang, and G. Pan, "Fine-Grained Urban Event Detection and Characterization Based on Tensor Cofactorization," *IEEE Trans. Human-Machine Syst.*, vol. 47, no. 3, 2017.

[5] H. Zhang, Y. Zheng, and Y. Yu, "Detecting Urban Anomalies Using Multiple Spatio-Temporal Data Sources," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 1, 2018.

[6] S. F. B and H. Fanaee-t, "Evolving Social Networks Analysisvia Tensor Decompositions : From Global Event Detection Towards Local Pattern," vol. 1, pp. 385–395, 2019.

[7] C. Lin, Q. Zhu, S. Guo, Z. Jin, Y. R. Lin, and N. Cao, "Anomaly detection in spatiotemporal data via regularized non-negative tensor analysis," *Data Min. Knowl. Discov.*, vol. 32, no. 4, pp. 1056–1073, Jul. 2018.

[8] M. Zhang, T. Li, H. Shi, Y. Li, and P. Hui, "A decomposition approach for urban anomaly detection across spatiotemporal data," in *IJCAI International Joint Conference on Artificial Intelligence*, 2019, vol. 2019-August.

[9] Z. Zhang, Q. He, H. Tong, J. Gou, and X. Li, "Spatial-temporal traffic flow pattern identification and anomaly detection with dictionary-based compression theory in a large-scale urban network," *Transp. Res. Part C Emerg. Technol.*, vol. 71, pp. 284–302, Oct. 2016.

[10] W. Kuang, S. An, and H. Jiang, "Detecting Traffic Anomalies in Urban Areas Using Taxi GPS Data," *Math. Probl. Eng.*, vol. 2015, 2015.

[11] C. Huang and X. Wu, "Discovering road segment-based outliers in urban traffic network," in *2013 IEEE Globecom Workshops, GC Wkshps 2013*, 2013.

[12] Y. Ge, H. Xiong, C. Liu, and Z. H. Zhou, "A taxi driving fraud detection system," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2011.

[13] N. Silva, V. Shah, J. Soares, and H. Rodrigues, "Road anomalies detection system evaluation," *Sensors (Switzerland)*, vol. 18, no. 7, Jul. 2018.

[14] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. Van De Wetering, "Visual traffic jam analysis based on trajectory data," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, 2013.

[15] T. Erdelić, M. Ravlić, and T. Carić, "Travel time prediction using speed profiles for road network of Croatia," in *2016 International Symposium ELMAR*, 2016, pp. 97–100.

[16] X. Liu, X. Liu, Y. Wang, J. Pu, and X. Zhang, "Detecting Anomaly in Traffic Flow from Road Similarity Analysis," *Lect. Notes Comput. Sci.*, pp. 92–104, 2016.

[17] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction," *Sensors (Switzerland)*, vol. 17, no. 4, pp. 1–16, 2017.

[18] P. S. Castro, D. Zhang, and S. Li, "Urban traffic modelling and prediction using large scale taxi GPS traces," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7319 LNCS, pp. 57–72, 2012.

[19] I. J. Jordaan, *Decisions under uncertainty : probabilistic analysis for engineering decisions*. Cambridge University Press, 2005.

[20] R. Dawson, "How significant is a boxplot outlier?," *J. Stat. Educ.*, vol. 19, no. 2, 2011.

[21] T. Carić and J. Fosin, “Using Congestion Zones for Solving the Time Dependent Vehicle Routing Problem,” *Promet-Traffic;Transportation*, vol. 32, no. 1, pp. 25–38, 2020.

[22] R. Zochowska and K. Grzegorz, “ITS services packages as a tool for managing traffic congestion in cities,” in *Intelligent Transportation Systems - Problems and Perspectives*, A. Sladkowski and W. Pamula, Eds. 2015, pp. 81–103.

[23] T. Ada and A. Kurtulus, “Project-Based Learning to Explore Taxicab Geometry,” *PRIMUS*, vol. 22, no. 2, pp. 108–133, 2012.

[24] A. Huang, “Similarity measures for text document clustering,” in *New Zealand Computer Science Research Student Conference, NZCSRSC 2008 - Proceedings*, 2008.

[25] J. M. Hancock, “Jaccard Distance (Jaccard Index, Jaccard Similarity Coefficient),” in *Dictionary of Bioinformatics and Computational Biology*, American Cancer Society, 2014.

Calculating the Fire Weather Index for US wildfires using a GPU

Jana Kuzmanova*, Marjan Gusev** and Vladimir Zdraveski**
* University Ss Cyril and Methodius, Faculty of Computer Science and Engineering
Skopje, North Macedonia
jana.kuzmanova@students.finki.ukim.mk
** University Ss Cyril and Methodius, Faculty of Computer Science and Engineering
Skopje, North Macedonia
{marjan.gushev,vladimir.zdraveski}@finki.ukim.mk

***Abstract*—The Forest Fire Weather Index allows the assessment of fire danger using weather variables in order to increase preparedness to prevent or halt the spread of wildfires. It often needs to be computed over large areas, taking weather data from hundreds of thousands of stations. CUDA parallel programming can be used to do this more efficiently. This paper presents a CPU and a GPU version as a solution to this problem, using historic datasets of wildfires and weather in the US to measure performance.**

I. Introduction

One of the main areas of application of computation, especially big data, in the research of climate and climate change is the assessment of risk and damages of natural disasters, as well as finding links between climate and natural disasters. One such threat are wildfires. They can have adverse effects on the vegetation and species distribution in forests, which can lead to depletion of the global carbon budget. They can also be a danger to human and animal life, as well as cause economic damages due to larger suppression costs and damage to property. In order to minimize the consequences and make suppression efforts effective, it's important to correctly predict the occurrence and spread of wildfires. While human-caused fires can't be predicted, there are multiple models which aim to asses wildfire risk. Most of them rely on weather factors, such as humidity, temperature, and wind speed for their calculations.

Previous research has shown that when fuel is available, the most important factors that determine whether a fire will occur are weather and vegetation. Some indexes currently used to predict wildfires are the Forest Fire Weather Index (FWI), McArthur Forest Fire Danger Index (FFDI) and National Fire Danger Rating System (NFDRS). FWI and FFDI both use temperature, humidity, precipitation and wind speed to calculate their results. They have been developed using empirical data from Canada and Australia, respectively, and can both be adapted to different types of vegetation and climate. The NFDRS was developed using empirical data from the US and includes not only weather factors, but also the topography of the area, fuel type and moistrue and the state of the vegetations. Di Giuseppe [1] shows that FWI and FFDI are better at predicting fires when they exist, while NFDRS results in fewer false positives. The indexes have different performance when applied to different parts of the world, however, the results showed that all three indexes are better than random prediction in a very large majority of countries. Forest fires are most often caused by lightning. Studies [2], [3], [4] have shown that the length of the fire season and intensity of fires in North America is on the rise, caused by both anthropogenic climate change, as well as natural variability.

The problem of computing these indexes, especially over longer periods of time or over large areas, is a good fit for parallel computation. The index assessing the danger for each location is independent of the index at any different location. Some of these indexes can choose to take the index of the previous days as input, but since the parallel threads only need to read the same data in order to redundantly compute the indexes for the predetermined number of days, the computation over a longer consecutive period is also possible without worrying about race conditions.

This paper shows this by calculating the Canadian Forest Fire Weather Index (FWI) for wildfires that occurred in the US between 1992 and 2015. The FWI uses fire weather observations for temperature, relative humidity, wind speed and rain to calculate three fuel moisture codes and three fire behavior indexes, including the final fire weather index. Taking into consideration that the formulas for the calculation of this index were calibrated on data from Canadian forest fires, there is a possibility the FWI isn't the most appropriate predictor of fire behavior for the US. However, it has been applied successfully to assess fire danger in other countries such as France, Australia and Croatia, which indicates that it can be a useful tool to quantify wildfire risk in countries other than the one it was initially developed for.

The rest of this paper has the following structure. Section II gives a more thorough explanation of the FWI and its applications, followed by an elaboration of a serial and parallel CUDA implementation of a program that computes it in Section III. The performance results of all algorithms are presented in Section IV. The paper concludes with a discussion of these results and the algorithms that produced them.

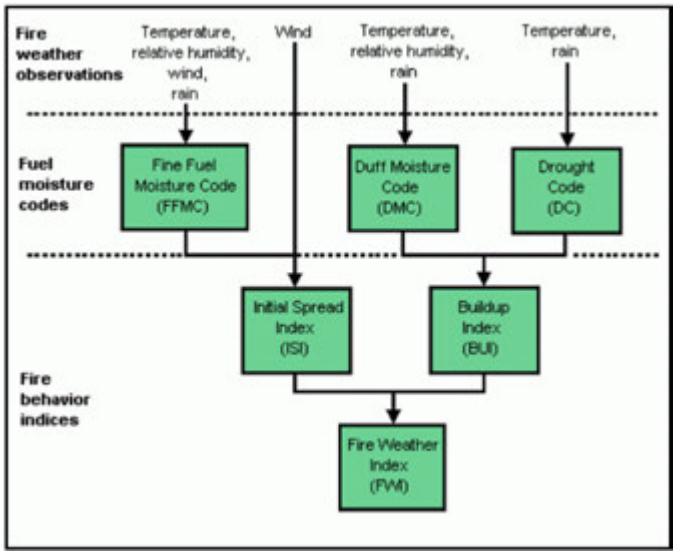


Fig. 1. FWI model

II. Model Description

The Fire Weather Index is calculated by a series of equations that use weather information to compute numeric ratings of fire danger in wildland areas. The final index combines five components: the Fine Fuel Moisture Code, Duff Moisture Code, Drought Code, Initial Spread Index and Buildup Index. The weather data needed to compute these indices includes the temperature, relative humidity, wind speed and rainfall. Furthermore, since day length is necessary for the calculation of the Duff Moisture Code and Drought Code, the month and latitude at which the fire occurred also need to be passed to the function that calculates the FWI [5].

Figure 1 illustrates the components of the FWI, their input values and how they combine to compute the final index. The moisture codes are used to determine the water content and dryness of various types of fuel. The FFMC quantifies the moisture content of light, surface fuels, the DMC is used to compute the moisture in light, decomposing organic matter, while the DC is applied to heavier and compact organic matter. Since the latter two types of fuel dry slowly, their moisture codes don't only depend on the current weather conditions, but also the available daylight. For this reason, the systems needs functions for the calculation of the day length at the given latitude in the month the fire occurred.

The next two indices, the ISI and BUI help describe how fast and far the fire is able to spread. Specifically, the ISI represents the rate of spread, while the BUI shows how much fuel is available to the given fire.

One possible application for FWI computation on a large scale is to calculate the FWI over a year or multiple years in order to identify periods that require special attention. One example of what the result could look like is shown in Figure 2.

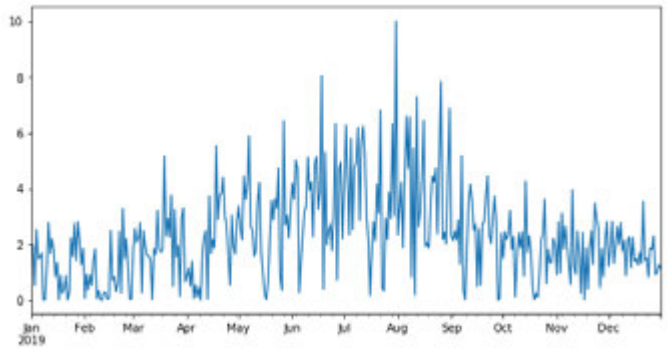


Fig. 2. FWIs for one location over a year calculated using randomly generated data

III. Implementation

The temperature and rainfall data used in this paper was collected from the GHCN Daily [6], while the relative humidity and wind speed data came from the North American Regional Reanalysis [7]. The wildfire events were taken from [8].

The code used to obtain the results presented in this paper was written in Python. The CUDA version was programmed and ran using the JIT compiler numba and its CUDA module, whose functions map directly to the CUDA C language. Numba supports all CUDA features needed for the implementation presented in this paper, including device functions, atomic operations, memory management for global, shared and local memory and thread synchronization. The CPU version was compiled using native Python code.

The code for the calculation of the FWI was adapted from [9]. It sequentially executes helper functions that compute the components of the index. Both the serial and parallel implementation take the 6 elements of weather data needed for the FWI calculation on the day of the fire for each of the n fires flattened into a one-dimensional numpy array. The output of both versions is a one-dimensional array containing the Fire Weather Indices of each of the fires. The serial implementation iterates over all elements of the input array and calculates the FWI. In the parallel version, each thread works on the weather data for one fire.

All functions described loop through all input elements and perform a constant number of operations on each of them, so they are all in O(n).

A. CPU Implementation

The serial implementation of the function that computes the FWI is straightforward. It iterates over the input array 6 elements at a time and passes them to the functions containing the series of equations that compute each of the components. Since the FFMC, DMC and DC require previous-day values, these are taken to be 85, 6 and 15, respectively, as per the original research. Once the result has been computed, it is appended to the output array.

The computed FWIs can then be used as input to the histogram function. In this case the histogram consist of 7

bins. The start of the first interval is 0, and each interval after that has a starting point that is higher by 5 than the one before, that is, the first interval counts the number of fires with a FWI value from 0 to 5, the second from 5 to 10 and so on. The final bin represents the number of fires with a FWI above 30. The function to compute this starts by initializing a 7-element array where every element is 0. It then iterates over all elements of the input array, and increments the element corresponding to the interval the value belongs to by one.

B. CUDA Implementation

This is a problem that can be easily parallellized. The weather data relating to each fire is independent which means that their FWIs can all be calculated concurrently. Another possible way to add parallelism to this problem is to exploit the inherent parallelism of the model. The FFMC, DMC and DC doesn't depend on each other and can be computed at the same time. The ISI depends on the FFMC and the BUI function takes the DMC and DC values as input, so they have to wait for the previous results, but the two indices themselves are independent and can also be computed in parallel. Finally, the FWI can be computed once the ISI and BUI functions have returned their results.

In the parallel CUDA version, each thread works on the weather data that corresponds to one fire. The kernel function makes use of six device functions, one for each component of the FWI. There are two additional device functions, called by the functions that calculate the DMC and DC, used to compute the day length during the month in which the fire occurred. This implementation only uses global and shared memory, as the threads don't need to communicate with each other. The input data is read from the global memory and directly passed to the functions for calculation. The device functions declare the numerous variables needed to get their results locally, and their return values are also stored in local variables. These local variables are passed to the next functions and again stored locally until the final function returns its result. This result is then written to the global memory.

Algorithm 1: FWI calculation 1

```
load data from global memory
ffmcVal = ffmc(temp, rh, wind, rain, 85)
dmcVal = dmc(temp, rh, wind, rain, 6, month, lat)
dcVal = dc(temp, rain, 15, month, lat)
isiVal = isi(wind, ffmcVal)
buiVal = bui(dmcVal, dcVal)
fwiVal = fwi(isiVal,buiVal)
store fwiVal in global memory
```

The second proposed parallelization version uses three threads per input data for one wildfire. It parallelizes not only the computation of the FWI for each fire, but also the computation of the individual components of the FWI. The blocks are initialized as a 2-dimensional grid of type $(n, 3)$ where n is the number of FWIs being worked on in parallel per block. The thread index in the y-dimension represents the separate components of the final index that the threads works on. When considering the computation of one FWI, each of the three threads working on it calls one of the device functions for the FFMC, DMC and DC. The intermediate results are written to the shared memory. The shared memory array contains $3n$ elements, equal to the number of threads in the block. Each thread writes to exactly one element of the shared memory array, however some of them read values from more than one element. This necessitates the addition of a barrier between each step of the FWI calculation. The threads computing the FFMC, DMC and DC all have to finish their work and write their results to the shared memory before the first two threads continue to calculate the ISI and BUI, the latter of which also uses the result of the third thread, and write them to the shared memory, overwriting the previous values that resided there. Another barrier is placed between this step and the final one, where the results are again passed to the function that calculates the FWI and its result directly written to the global memory.

Algorithm 2: FWI calculation 2

```
tx = threadIdx.y + 3*threadIdx.x
allocate shared memory
load data from global to local memory
ty = tx % 3
if ty == 0 then
  | sh[tx] = ffmc(temp, rh, wind1, rain, 85)
else if ty == 1 then
  | sh[tx] = dmc(temp, rh, wind, rain, 6, month, lat)
else if ty == 2 then
  | sh[tx] = dc(temp, rain, 15, month, lat)
if ty == 0 then
  | sh[tx] = isi(wind,sh[tx])
else if ty == 1 then
  | sh[tx] = bui(sh[tx],sh[tx+1])
sync threads
if ty == 0 then
  | store result in global memory
```

The problem of computing a histogram is another one that allows for easy parallelization. The histogram kernel takes the FWIs as input and computes the frequency of FWIs belonging to the ranges as described in the description of the CPU algorithm. In the parallel version, an array of shared memory with 7 elements is allocated for each block. Every thread works on one element of the input array and atomically increments the corresponding element of the local histogram in the shared memory. After all the threads have finished this task, the first 7 threads of each block atomically add the element of the shared array at their threads index to the element with the same index in the global memory array, which, when all threads have finished their computation, contains the final result.

IV. Results

The performance was measured on a desktop computer with an Intel Celeron G1620 2.70 GHz CPU and an Nvidia GeForce

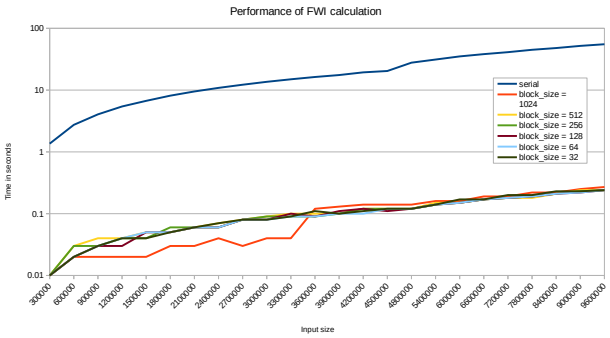


Fig. 3. FWI calculation performance of first parallel version. The x-axis shows the total number of input values passed to the function. This is the number of values needed to calculate the FWI of 1.6 million wildfires. The y-axis shows the time taken by the function in seconds.

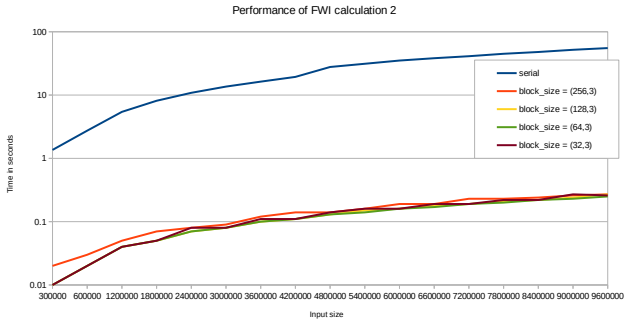


Fig. 4. FWI calculation performance of second parallel version. The x-axis shows the total number of input values passed to the function. This is the number of values needed to calculate the FWI of 1.6 million wildfires. The y-axis shows the time taken by the function in seconds.

GT 730 graphics card. The graphics card has a compute capability of 3.5.

The computed FWIs won't be compared to historical results because the US uses the NFDRS to asses fire danger.

Figure 3 shows the performance of the function for the computation of the FWI. The length of the input array is represented on the x-axis, while the y-axis represents the execution time in seconds. The tests of the CUDA parallel implementation were ran for block sizes from 2^5 to 2^{10}, represented by the different colored lines on the graph, with an additional line for the performance of the serial implementation. Since each thread works on multiple elements of the input array, and writes exactly one value to the output array, the number of blocks is

$$\left\lceil \frac{\text{lenth of output array}}{\text{blockSize}} \right\rceil. \tag{1}$$

The performance was measured for input arrays of sizes between 300K and 9.6M, with resulting output arrays that contain between 50K and 1.6M elements.

We can see that even when the size of the input array is small, the serial version is significantly slower than the parallel one. In the serial case, the program needs over a second for an input array containing the weather data for 50K wildfires or 300K input elements, while the CUDA version takes less than half a second even for an input array containing data for over 1.6M fires. The gap in the performance of the two implementations widens as the number of input elements increases. The serial version becomes prohibitively slow for large input arrays, needing over 50 seconds for the largest number of elements the performance was measured for. This kind of input size or larger can be common when calculating the fire danger over large areas, with data coming from hundreds of thousands of stations, or when doing historical analysis, like in this case.

The results of the performance tests of the second parallel version which also parallelizes the component computation are illustrated in Fig. 4. In this case, blocks are two-dimensional and the number of blocks is defined as

$$\left\lceil \frac{\text{length of output array}}{\text{blockSize in x-dimension}} \right\rceil \tag{2}$$

This parallel version still significantly outperforms the serial version, however, it's slower by a few milliseconds than the other proposed parallel version in most cases.

The performance of the histogram function is shown by Fig. 5. It was tested for input sizes between 50K and 1.6M, with the same block sizes as in the previous tests. Here, the number of blocks is a function of the input size, since each threads works on one elements of the input array, and is defined by

$$\left\lceil \frac{\text{lenth of input array}}{\text{blockSize}} \right\rceil \tag{3}$$

Again, the x-axis shows the input size, while the y-axis shows the time needed to compute the result in seconds.

While the differences in the performance of the histogram aren't as large as in the previous case, especially for small input sizes, the increase of time needed to compute the result with increasing input size is much smaller for the parallel than the serial version. For the smallest input sizes, the difference between the two implementation is measured in milliseconds, but it quickly increases with the input size. At 800K input elements, the gap is still under a second, but nonetheless growing rapidly, and it reaches 1.5 seconds for an input size of 1.6M elements, which can become a bigger problem at larger input sizes.

V. Discussion

The performance results presented in the previous section show that using a GPU to solve this problem leads to a significantly improved performance compared to a CPU. As shown in Figure 6, the speedup of the more efficient problem solution presented, the histogram, is 5 times faster when executed on a GPU even for small input sizes. Figure 7 shows that the algorithm for the calculation of the FWI is over 100 times faster on the GPU than on the CPU, regardless of input size.

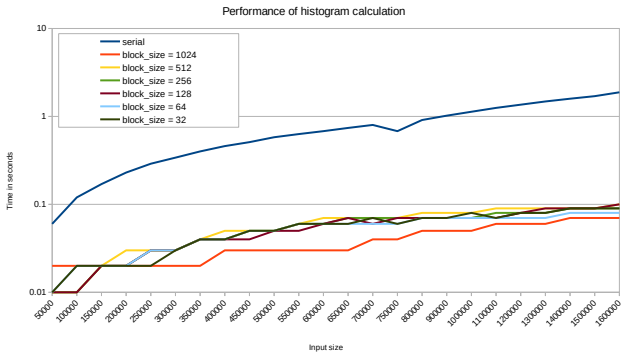


Fig. 5. Histogram calculation performance. The x-axis shows the total number of input values, in this case, FWIs, passed to the function. The y-axis shows the time taken by the function in seconds.

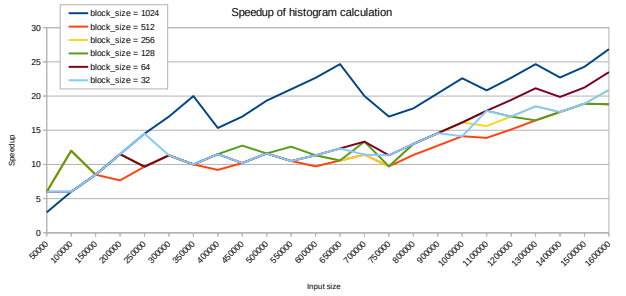


Fig. 6. Speedup of histogram calculation

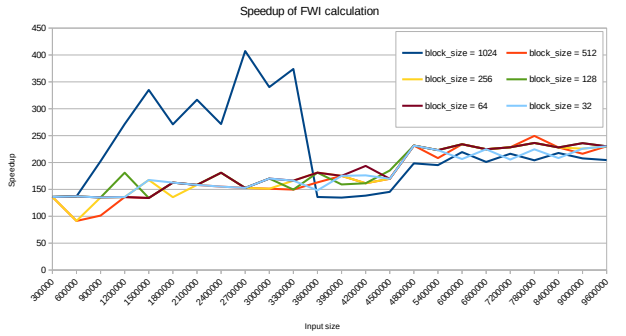


Fig. 7. Speedup of FWI calculation

The speedup of this function sharply falls for input array length of 3600000 elements, or when 600000 threads have to be launched and executed. This happens because the number of blocks that can be executed at the same time is limited by the number of streaming multiprocessors, and the number of blocks that can run on one streaming multiprocessors is limited by the number of registers and shared memory used per thread. While a larger block size needs fewer blocks to compute the result for the same number of elements, it's also limited to a smaller number of blocks per streaming multiprocessor, while smaller block sizes allow a larger number of blocks per SM, which means that more blocks, and for certain block sizes, more threads, can be executed simultaneously. In the case of the FWI calculation described here, block sizes of 256 and 128 threads allow a higher number of active threads per SM than a block size of 1024. This doesn't make much of a difference when the number of threads is smaller, but becomes obvious as the number of threads increases.

The two proposed parallel implementations are both an improvement over the serial version of the algorithm, however, the version that calls all device functions sequentially is consistently faster by a few milliseconds when compared to the version that parallelizes these calls. One possible reason for this is the fact that the latter version uses additional memory accesses and keeps the intermediate results in the shared instead of the local memory, as well as the addition of barriers, which could offset the sequential computations of these values. Another disadvantage of this implementation is that for part of the calculation, 1/3 of the threads are idle, and in the final step, 2/3 of all threads are idle.

VI. Conclusion

The computation of the FWI is not very computationally intensive when considering the data for one place and one day. However, it often needs to be calculated over large areas to assess fire danger, and its accuracy increases when it considers the weather conditions of multiple days instead of just the current, which leads to an even higher number of calculations that need to be done. The speedup provided by the CUDA implementation presented in this paper could be useful in those cases.

References

[1] F. Di Giuseppe, F. Pappenberger, F. Wetterhall, B. Krzeminski, A. Camia, G. Libertá, and J. San Miguel, "The potential predictability of fire danger provided by numerical weather prediction," *Journal of Applied Meteorology and Climatology*, vol. 55, no. 11, pp. 2469–2491, 2016.

[2] J. S. Fried, M. S. Torn, and E. Mills, "The impact of climate change on wildfire severity: A regional forecast for northern california," *Climatic Change*, vol. 64, no. 1-2, pp. 169–191, 2004.

[3] M. M. Fauria and E. Johnson, "Climate and wildfires in the north american boreal forest," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1501, p. 23152327, 2008.

[4] J. T. Abatzoglou and A. P. Williams, "Impact of anthropogenic climate change on wildfire across western us forests," *Proceedings of the National Academy of Sciences*, vol. 113, no. 42, p. 1177011775, 2016.

[5] C. Van Wagner, P. Forest *et al.*, "Development and structure of the canadian forest fireweather index system," in *Can. For. Serv., Forestry Tech. Rep*. Citeseer, 1987.

[6] M. J. Menne, I. Durre, B. Korzeniewski, S. McNeal, K. Thomas, X. Yin, S. Anthony, R. Ray, R. S. Vose, B. E. Gleason, and T. G. Houston, "Global historical climatology network - daily (ghcn - daily), version 3.26," http://doi.org/10.7289/V5D21VHZ, 2012.

[7] National Oceanic and Atmospheric Administration, Department of Commerce, "Ncep North American regional reanalysis (NARR), for 1979 to present," https://catalog.data.gov/dataset/ncep-north-american-regional-reanalysis-narr-for-1979-to-present, 2019.

[8] R. Tatman, "1.88 million us wildfires, 24 years of geo-referenced wildfire records," https://www.kaggle.com/rtatman/188-million-us-wildfires, 2018.

[9] V. Wagner, "Vba module containing functions to calculate the components of the canadian fire weather index system, development and structure of the canadian forest fire weather index system," http://www.columbia.edu/~rf2426/index_files/FWI.vba, p. 37, 1987.

Iris Identification Using Wavelet Decomposition and Gabor Filter

Hannes Nitz Petterson*, Jonas Rehnholm*, Samuel Vikström*, Martin Åslund*,
Elaine Åstrand**, Ivan Tomasic**
*Mälardalen University, Master's Programme in Robotics, Västerås, Sweden
**Mälardalen University, Division of Intelligent Future Technologies, Västerås, Sweden
{hpn15004, jrm16003, svm16001, mad10004}@student.mdh.se, {elaine.astrand, ivan.tomasic}@mdh.se

***Abstract*—Biometric authentication has seen a widespread increase in popularity as supporting technology has become common in mass produced consumer electronics. Like fingerprints, each individual has unique patterns in the iris, which makes it a common approach for implementing visual biometric authentication. The paper describes a novel system for extracting the iris pattern and using it for identification of people. The system uses Haar wavelet decomposition and 2D Gabor filtering to extract the pattern data. The pattern data is then used with bitwise XOR comparison for final identification matching. Instead of manually selecting parameters for the Gabor filter, a machine learning method called Particle Swarm Optimization was used. The parameters that gave the best matching result were then implemented in the filter design. The implemented system was evaluated on images obtained from 6 individuals in different settings. The evaluation showed that matching identification could be achieved for the database used. The prepossessing of images with Independent Component Analysis was also used to remove the reflections on the images but that did not improve the classification significantly. Still we were able to perfectly distinguish between the individuals. Further preprocessing and a larger training database would be required to get more general and robust results.**

***Index Terms*—Iris identification, wavelet decomposition, 2-D Gabor filter, PSO**

I. Introduction

For a long time, the only way to unlock personal devices such as phones or laptops has been by using passwords or pin codes. Recently however biometric authentication such as fingerprints, face recognition and iris identification has become available in such devices. Today almost all smartphones have some sort of biometric authentication.

Biometric authentication has many advantages over regular passwords and pin codes. Perhaps the biggest advantage is the fact that it is very fast and convenient for the user but still provides the security level needed for personal devices. Among all the biometric authentication methods, iris biometrics is the most efficient technique [1].

Investigated in this paper is iris identification for biometric authentication. Just like each individual has a unique fingerprint, textures in the irises are unique. Therefore it is possible to identify people from an image of the eye by extracting and comparing the textures of the iris.

Even though iris image-based biometric systems have already been commercialised, there is still room for improvement [2]. In the previous work various methods have been used in all the standard steps in iris recognition: image acquisition, pre-processing, iris segmentation, feature-extraction, classification (recognition) [2, 3]. Methods used in this work for extracting the texture information are:

- Daugmans' rubbersheet model, which is used because it can transform any eye image to a specific size regardless of variance in pupil dilatation.
- Haar wavelet decomposition, because it can compress the image while keeping most of the information.
- Gabor filtering which can effectively extract any information encoded in the patters of the iris that is unique to each individual.

Gabor filter is used for extracting information in a lot of similar works [4, 5, 6]. Gabor filter parameters in those papers are selected by manually testing different combinations of parameters, leaving a lot of room for improvement, given that a human can not go through and test all the combination of parameters. Instead this paper proposes the use of particle swarm optimization (PSO) to find the optimal parameters. By using PSO a computer can test far more combination than a human could feasibly manually select. This should in theory lead to a better quality filter which will in turn result in a better classification.

II. Approach and method

A. Iris and pupil location

Hough circular transform is used to find locations of outer and inner perimeters of irises.

Every circle can be described with a center point (x,y) and a radius (r). The Hough transform tries to find good values for these three parameters in an image. The Hough transform places circles with the same radius at every edge pixel. If enough of these placed circles touch at the same point, that point will be classified as the center of the circle with the same radius as the placed circles. The Hough transform iterates through different values for the radius to see which one gets the best results (see Fig. 1). The solid circle is the circle that is in the image and the dotted circles are those that the Hough transform creates.

B. Rubbersheet

There are some prepossessing problems with iris identification which mostly can be solved by using Daugmans

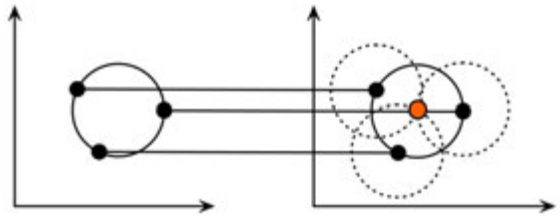


Fig. 1: Hough transform principle

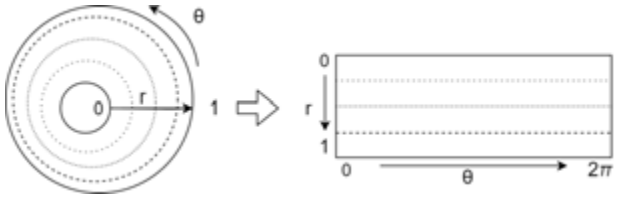


Fig. 2: Rubbersheet model

rubbersheet model [7]. Some of the problems are that the pupil and the iris usually are not concentric and that the pupil dilates depending on light stimulants.

Daugmans rubbersheet model takes the annulus shaped iris and unwraps it out to a rectangular form, see Fig. 2, 3. The rectangle gets a height of r which is the distance from the inner perimeter to the outer perimeter of the iris with an interval of [0 1] and the width is θ with an interval of [0 2π].

The remapping of the iris can be described as taking a pixel at I(x,y) on the iris and transferring it to I(r,θ) on the rectangle. This can be written as the following equations:

$$I(X(r,\theta),Y(r,\theta)) \rightarrow I(r,\theta) \tag{1}$$

$$X(r,\theta) = (1-r)X_p(\theta) + rX_a(\theta) \tag{2}$$

$$Y(r,\theta) = (1-r)Y_p(\theta) + rY_a(\theta) \tag{3}$$

where (X_p,Y_p) are the points along inner perimeter of the iris and (X_a,Y_a) are the points along outer perimeter of the iris.

Because the size of the rectangle depends on the interval of r and θ which is manually selected, it will have a constant size which does not depend on the starting image size. The size of the rectangle that we used is 64x512.

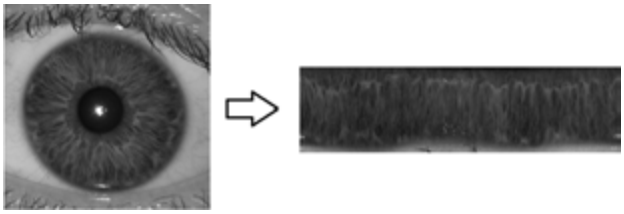


Fig. 3: Rubbersheet model example on a real eye

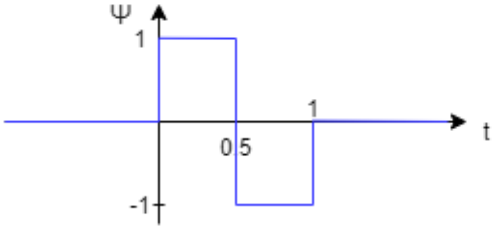


Fig. 4: Haar wavelet used for wavelet decomposition

C. Wavelet decomposition

Wavelet decomposition (also called wavelet transform) has both the advantages of a sparse image representation and a complete representation. The sparse representation means that a reduction in the size of the image is achieved. Due to the wavelet decomposition's ability to have a complete representation, the potential texture information in the image is conserved. The smaller image size also helps speed up the rest of the processing of the image. Different wavelets can be used for wavelet decomposition, but in this work the Haar wavelet was used and can be seen in Fig. 4. The Haar wavelet transform is the simplest of the wavelet transforms and has been found effective for image compression due to its simple and effective approach for analyzing local aspects of a signal [8, 9].

The wavelet transform runs a high pass and an low pass filter on rows of the original image separately, yielding two new images: one low pass image (L) and one high pass image (H). The two new images then get their columns downsampled by only keeping the even-indexed columns. Then both of the downsampled images are run through a high pass and a low pass filter on the columns, yielding four new images. One image is low pass filtered in both rows and columns (LL), one is filtered with a low pass on the rows and a high pass on the columns (LH), the other two are (HL) and (HH). These four images then get their rows downsampled by only keeping the even-indexed rows. The LL image is an approximation of the original image. The LH image shows high frequency

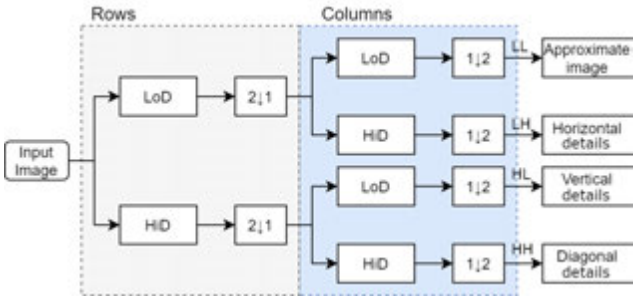


Fig. 5: Wavelet decomposition, LoD is the Low pass decomposition filter on rows or columns, HiD is the High pass decomposition filter on rows or columns. 2↓1: Downsample columns: keep the even-indexed columns. 1↓2: Downsample rows: keep the even-indexed rows

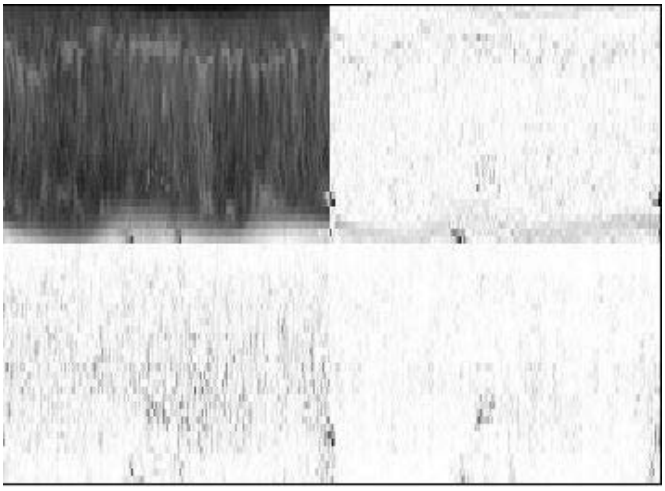


Fig. 6: The results of the Wavelet decomposition for level 1. Top left, approximate image. Top right vertical details. Bottom left, horizontal details. Bottom right, diagonal details

horizontal details, HL shows high frequency vertical details and HH shows high frequency diagonal details.

The decomposition steps can be seen in Fig. 5 and the resulting images can be seen in Fig. 6. This is the output for level 1, however wavelet decomposition can be done in multiple levels, to compress the image further wavelet decomposition is done but with the approximate image. In this work, the approximate image from level 2 is used.

D. The Gabor filter

Gabor filtering is a common method used in various image processing and computer vision applications. Using 2D Gabor filters has been shown to be a useful tool for extracting textural information from images. This has proven to be very useful in texture classification [10], as well as for identification of specific patterns like palm prints or finger prints [5],[11]. It has also been suggested that Gabor filters are useful in modeling of the V1 neurons in the visual cortex [12],[13]. The filter is applied by the way of 2D convolution. The method for 1D convolution is simply extended to work along the two spatial axes of a 2D image. The image is passed through a number of filters with different orientations, where each orientated filter will preserve information in the location where pattern changes occurs in the orientation matching the filter orientation.

The Gabor wavelet is composed of a 2D Gaussian function, as seen in Fig. 7. which is the envelope of the wavelet. It has a complex sinusoidal plane wave as the carrier function, Fig. 8. The sinusoid has a particular frequency and orientation in the 2D plane. The Gabor wavelet is then the product of the envelope and carrier functions as shown in Fig. 9.

We used the Gabor filter which is defined in equation (4) as a complex function with six arguments. The x and y are the spatial coordinates of the 2D filter. $\mathcal{F}$ is the central frequency of the sinusoidal wave. σ is the standard deviation of the Gaussian function that defines the size of the envelope. θ is

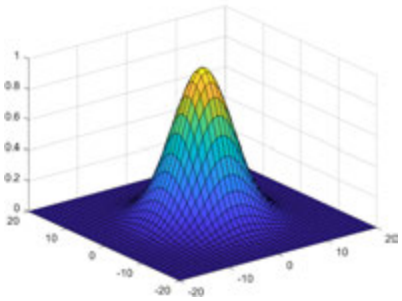


Fig. 7: 2D Gaussian function

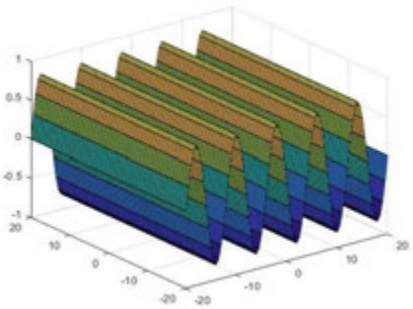


Fig. 8: 2D sinusoidal wave function

the orientation in spatial domain. As suggested in [14], larger number of oriented filters has little effect on the ability of the filter to extract relevant information. Therefore, we used a fixed set of orientations for θ defined as $N_i \cdot pi/4$ where $N = \{0, 1, 2, 3\}$. Lastly, Φ is the phase of the sinusoidal function.

$$\Psi(x, y, \mathcal{F}, \sigma, \theta, \Phi) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(\frac{-(x^2 + y^2)}{2\sigma^2}) \cdot exp(\mathcal{F} \cdot 2\pi i(x \cdot \cos(\theta) + y \cdot \sin(\theta) + \Phi)) \quad (4)$$

E. Gabor filter parameter selection

The decomposed image generated from the wavelet decomposition is passed through the Gabor filter. The filter parameters that can be tuned are the spatial size of the filter,

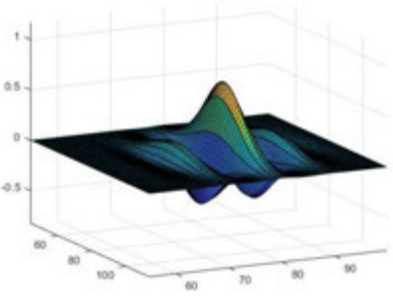


Fig. 9: Gabor filter in spatial space

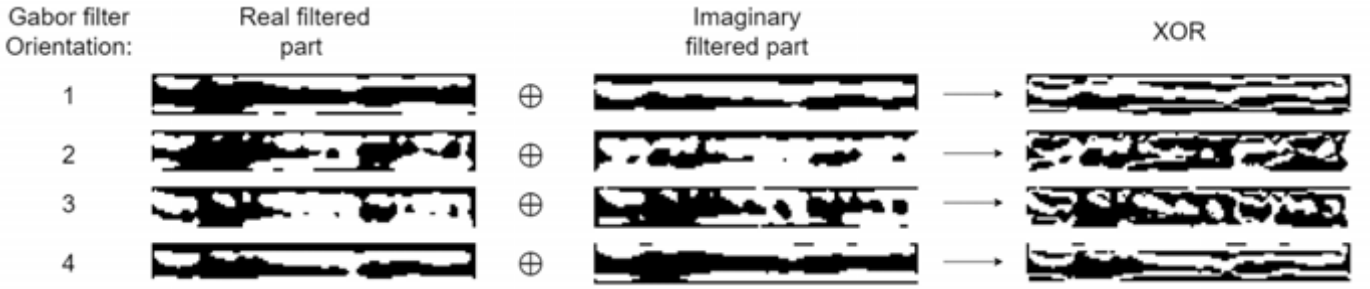


Fig. 10: Real and imaginary parts from applying a 2-D Gabor filter and the results from using the XOR operator on them

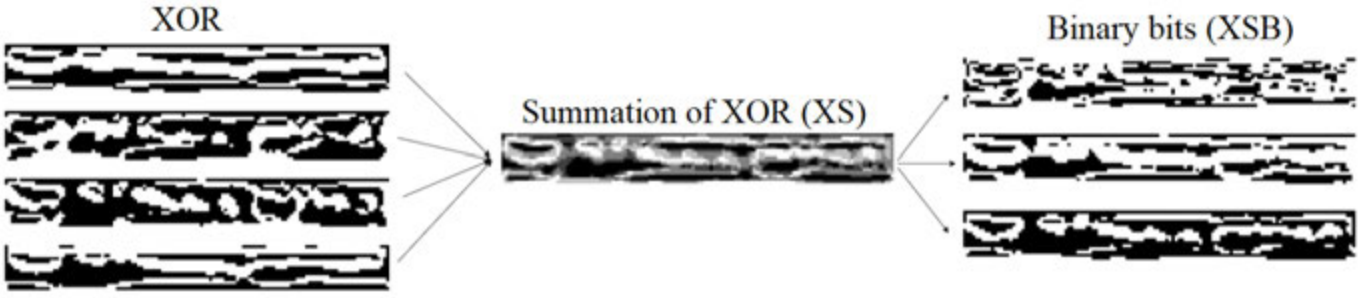


Fig. 11: Binary images get summed together and then coded back into binary images

the central frequency $\mathcal{F}$ and phase Φ of the sinusoidal part of the function, and the standard deviation σ of the Gaussian envelope function.

To find optimal values for these parameters we used the optimization algorithm PSO (particle swarm optimization). PSO starts by initializing a cluster of particles in a n:th order dimension where n is the number of parameter that should be optimized. In this work the particles are initialized in a 4:th dimensional space. After initialization the particles move according to a velocity vector which is based on the particle's personal best position, the best position of any particle and the previous velocity vector of the particle. After some number of iterations the particles will start to swarm around a local minimum point.

F. XOR-sum

When applying the Gabor filter to the iris image it will result in a real part and an imaginary part for every N different orientations of the filter. Both parts contain equal amounts of potential information. To conserve the information in both parts they first are binarized and then fused together using the XOR operator for each corresponding orientation, see Fig. 10.

All of the results from the XOR operations gets summed up into one image XS(x,y). Thereafter the image gets coded into (N+1)/2, rounded up, binary images XSB(x,y,b) see Fig. 11.

The XS(x,y) gets coded into binary as follows:

$$XSB(x,y,b) = \\ = \begin{Bmatrix} 1 & if & b \leq XS(x,y) < b + \frac{N+1}{2} \\ 0 & otherwise & \end{Bmatrix} \quad (5)$$

TABLE I: Three bit encoding of XORSUM

	b = 1	b = 2	b = 3
XS(x,y) = 0	0	0	0
XS(x,y) = 1	0	0	1
XS(x,y) = 2	0	1	1
XS(x,y) = 3	1	1	1
XS(x,y) = 4	1	1	0

where b = 1,2,...(N+1)/2. Table I shows the coding of three bit images from four orientations of the Gabor filter.

The final identification is done with these three binary images. A matching score is calculated by taking two sets of these binary images and comparing them with a bit-wise XOR operator. If the two iris images are similar the received score will be lower. By setting a threshold, it is possible to determine if the images are of the same eye or not.

G. ICA based specular reflection minimization

Most images contain some reflections. These can be divided into specular and diffuse reflections. An attempt was made to

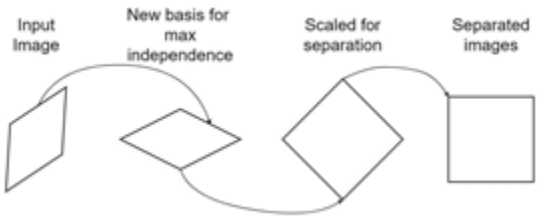


Fig. 12: Step by step illustration of ICA based separation

remove/minimize the specular reflection which can interfere with feature detection in the iris. The chosen method was Independent Component Analysis (ICA) [15, 16, 17] (see Fig. 12), because its purpose is to separate independent signals, which in our case are the reflection, and the image of the iris. ICA results in two images, first ideally representing the reflections, and second containing the eye without reflections. In order to focus the separation on the reflections and the eye, the iris and pupil area was extracted and used for image separation. This method require multiple images captured through different polarizing filters. We used two images captured with approximately 90° rotation of the polarizing filter.

III. Results and Discussion

First the PSO was performed and resulted with the set of optimal parameters presented in Table II.

The developed iris identification program was tested on eye images collected from three males and three females. For every person, six images have been taken close to the eye, each with slightly different angle and lighting.

Then the algorithm was applied on each image and the results were split in to a reference list (3 images) and a identification list (3 images). Doing a XOR-sum comparison between all images in the reference lists and one image from individual 2 identification list gives the results that are given in Table III.

Lower score indicates a better matching between the identification image and a reference image. A score of 0 would mean a perfect match and would happen only if the identification image is also present in the reference list. As seen in the table, there is some variation in score both between individuals and within one individual. Still, for the example shown in Table III, a clear distinction in score can be seen for the correct matching individual, i.e. individual 2.

In Table IV the average matching score for all individuals is shown. Each row represents average matching score between the 3 images in the identification list for that person, and the 3 images in the reference lists of each other person. For example, cell (INDV 1, INDV 2) represents the average matching score between the 3 images from individual 1 identification list and the 3 images in individual 2 reference list. It is evident that the scores within individual (diagonal scores) are significantly different than the scores between individuals (off-diagonal elements).

Since the used database consisted of eyes from only six people with light coloured eyes, the database is obviously too narrow for the presented approach to be considered as proven to be accurate. While the tested database would imply that identification is possible, by applying the algorithm on more images, a more reliable conclusion about the accuracy of our algorithm could be made. We plan to do that in the future work.

A larger database of eye images would also enable tuning the Gabor filter parameters to more eye colours. We did tests also with images of brown eyes, but due to the iris and the pupil having similar colours, the Hough circular transform was not consistently able to find the eyes. Here, other eye detection algorithms could be explored.

An image enhancement technique used in [18] was also implemented and tested with the aim to enhance the image by calculating a uniform illumination of the image. The method has shown to be inefficient because it caused the already small image to become more pixelated. Instead some preprocessing on the images like CLAHE (contrast limited adaptive histogram equalization) might be used to enhance the contrast in the images, making details easier to detect in darker images.

The results from our implementation of removing specular reflections with ICA has not proved to be useful, since it reduced the clarity of the iris image and made the image features less apparent (Fig. 13).

This could be caused by our data collection method: taking photographs with a handheld smartphone with only a relatively weak integrated polarizing filter. It might also be because our

TABLE II: Optimal 2-D Gabor filter parameters from PSO

Size	$\mathcal{F}$	σ	Φ
15	0,9828	1.0066	3,6159

TABLE III: Reference matching score between an image of individual 2 and 3 reference images of each individual.

INDV 1	INDV 2	INDV 3	INDV 4	INDV 5	INDV 6
2681	896	1809	2285	2028	1617
2385	835	2401	2277	2251	1655
1882	941	2360	2014	2048	1643

TABLE IV: Average reference matching score for each individual

		Reference INDV					
		INDV 1	INDV 2	INDV 3	INDV 4	INDV 5	INDV 6
identification INDV	INDV 1	1560	2388	2734	2810	2998	2908
	INDV 2	2342	865	2245	2232	2086	1606
	INDV 3	2697	2414	1464	2441	2406	2690
	INDV 4	2926	2323	2331	1441	1789	2129
	INDV 5	2940	2195	2090	1869	921	1821
	INDV 6	2754	1829	2471	2124	1905	853

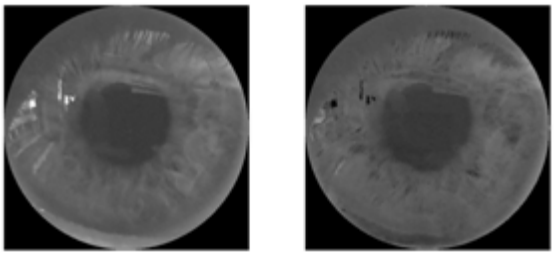


Fig. 13: Left: One original image in grayscale; Right: Resulting image with ICA reduced specular reflections

images were too Gaussian for the ICA to be effective. There are other possible methods to tackle the problem of reflections which we plan to explore in our future work.

IV. Conclusion

An algorithm for extracting pattern information from the iris for the purpose of identification was developed and shown to be effective on six persons. We offered a deeper look into the application of wavelets in processing and filtering of two dimensional data. Machine learning optimisation method called particle swarm optimization was used for the design of the Gabor filter which is applied on the results of wavelet filtering.

In order to improve the capability of the system, a more general database of eye images should be used. For dealing with poor lighting condition, especially with darker coloured eyes, some preprocessing method could be implemented.

V. Acknowledgment

I. Tomasic acknowledges the financial support by the Swedish Knowledge Foundation (KKS) through the project Embedded sensor systems for health plus (ESS-H+), project number 20180158.

References

[1] M. Lavanya and V. Kavitha. "A survey on iris biometric and recognition". In: *Journal of Advanced Research in Dynamical and Control Systems* 11.11 (2019), pp. 331–337. DOI: 10.5373/JARDCS/V11SP11/20193039.

[2] J. Jenkin Winston and D. Jude Hemanth. "A comprehensive review on iris image-based biometric system". In: *Soft Computing* 23.19 (2019), pp. 9361–9384. ISSN: 14337479. DOI: 10.1007/s00500-018-3497-y.

[3] Puja S. Prasad and D. Baswaraj. "Iris Recognition Systems: A Review". In: *ICCCE 2018*. Ed. by Amit Kumar and Stefan Mozar. Singapore: Springer Singapore, 2019, pp. 521–527. ISBN: 978-981-13-0212-1.

[4] Wai Kin Kong, David Zhang, and Wenxin Li. "Palmprint feature extraction using 2-D Gabor filters". In: *Pattern Recognition* 36.10 (2003), pp. 2339–2347. ISSN: 0031-3203. DOI: https://doi.org/10.1016/S0031-3203(03)00121-3.

[5] Deepti Tamrakar and Pritee Khanna. "Palmprint verification with XOR-SUM Code". In: *Signal, Image and Video Processing* 9.3 (2015), pp. 535–542. ISSN: 1863-1711. DOI: 10.1007/s11760-013-0475-9. URL: https://doi.org/10.1007/s11760-013-0475-9.

[6] Ritesh Vyas, Tirupathiraju Kanumuri, and Gyanendra Sheoran. "Iris recognition using 2-D Gabor filter and XOR-SUM code". In: Aug. 2016, pp. 1–5. DOI: 10.1109/IICIP.2016.7975369.

[7] J. G. Daugman. "High confidence visual recognition of persons by a test of statistical independence". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.11 (Nov. 1993), pp. 1148–1161. ISSN: 1939-3539. DOI: 10.1109/34.244676.

[8] Sulaiman Khan et al. "An efficient JPEG image compression based on Haar wavelet transform, discrete cosine transform, and run length encoding techniques for advanced manufacturing processes". In: *Measurement and Control* 52.9-10 (2019), pp. 1532–1544.

[9] Kamrul Hasan Talukder and Koichi Harada. "Haar Wavelet Based Approach for Image Compression and Quality Assessment of Compressed Image". In: *CoRR* abs/1010.4084 (2010). arXiv: 1010.4084.

[10] Mahamadou Idrissa and Marc Acheroy. "Texture classification using Gabor filters". In: *Pattern Recognition Letters* 23.9 (2002), pp. 1095–1102. ISSN: 0167-8655. DOI: https://doi.org/10.1016/S0167-8655(02)00056-9.

[11] Chih-Jen Lee and Sheng-De Wang. "Fingerprint feature extraction using Gabor filters". In: *Electronics Letters* 35.4 (Feb. 1999), pp. 288–290. ISSN: 0013-5194. DOI: 10.1049/el:19990213.

[12] John G. Daugman. "Two-dimensional spectral analysis of cortical receptive field profiles". In: *Vision Research* 20.10 (1980), pp. 847–856. ISSN: 0042-6989. DOI: https://doi.org/10.1016/0042-6989(80)90065-6.

[13] John G. Daugman. "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters". In: *J. Opt. Soc. Am. A* 2.7 (July 1985), pp. 1160–1169. DOI: 10.1364/JOSAA.2.001160.

[14] Francesco Bianconi and Antonio Fernández. "Evaluation of the effects of Gabor filter parameters on texture classification". In: *Pattern Recognition* 40.12 (2007), pp. 3325–3335. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2007.04.023.

[15] M. Yamazaki, Yen-Wei Chen, and Gang Xu. "Separating Reflections from Images Using Kernel Independent Component Analysis". In: *18th International Conference on Pattern Recognition (ICPR'06)*. Vol. 3. Aug. 2006, pp. 194–197. DOI: 10.1109/ICPR.2006.1045.

[16] M. M. Bronstein et al. "Separation of reflections via sparse ICA". In: *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*. Vol. 1. Sept. 2003, pp. I–313. DOI: 10.1109/ICIP.2003.1246961.

[17] H. Farid and E. H. Adelson. "Separating reflections and lighting using independent components analysis". In: *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. Vol. 1. June 1999, 262–267 Vol. 1. DOI: 10.1109/CVPR.1999.786949.

[18] Aditya Nigam et al. "Iris Recognition Using Discrete Cosine Transform and Relational Measures". In: *Computer Analysis of Images and Patterns*. Ed. by George Azzopardi and Nicolai Petkov. Cham: Springer International Publishing, 2015, pp. 506–517. ISBN: 978-3-319-23117-4.

An Analytics Toolbox for Cyber-Physical Systems Data Analysis: Requirements and Challenges

M. Zanin*, E. Menasalvas*, †, A. Rodriguez González*, † and P. Smrz‡
* Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Madrid, Spain
† Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, Madrid, Spain
‡ Brno University of Technology, Brno, Czech Republic
[massimiliano.zanin, ernestina.menasalvas, alejandro.rg]@upm.es, smrz@fit.vutbr.cz

Abstract - **The fast improvement in telecommunication technologies that has characterised the last decade is enabling a revolution centred on Cyber-Physical Systems (CPSs). Elements inside cities, from vehicles to cars, can now be connected and share data, describing both our environment and our behaviours. These data can also be used in an active way, by becoming the tenet of innovative services and products, i.e. of Cyber-Physical Products (CPPs). Still, having data is not tantamount to having knowledge, and an important overlooked topic is how should them be analysed. In this contribution we tackle the issue of the development of an analytics toolbox for processing CPS data. Specifically, we review and quantify the main requirements that should be fulfilled, both functional (e.g. flexibility or dependability) and technical (e.g. scalability, response time, etc.). We further propose an initial set of analysis that should in it be included. We finally review some challenges and open issues, including how security and privacy could be tackled by emerging new technologies.**

Keywords – Cyber-Physical Systems; Cyber-Physical Products; data analytics; information security; information privacy

I. INTRODUCTION

The widespread availability of Cyber-Physical Systems (CPSs) [1, 2], combined with the deployment of telecommunication technologies allowing the transmission of large volumes of data with negligible costs, are triggering a revolution in our societies. Specifically, new business models are being developed, in which novel services (also called Cyber-Physical Products, CPPs) are created by leveraging on the data provided by CPSs. As a familiar example, one may think on the use of geolocalised information from mobile phones to detect traffic jams, and the resulting service suggesting optimised routes based on real-time data [3]. Note that the innovation behind CPPs resides in the data they use, as they would not be feasible without the collaboration and sharing made possible by many distributed CPSs.

In spite of many advancements and research works, one important aspect has mostly been neglected: data availability does not equate knowledge availability. In other words, having access to data does not directly imply gaining new knowledge, as data have to be processed. If data analytics is in general a mature field, the same may not be true for data analytics in a CPP environment, due to the specific requirements of the latter – e.g. large data volumes, restricted bandwidths, or privacy of the information. In other words, it may be naïve to think that *one-size-fits-all*, and that existing analytics solutions may seamlessly be reused in this context.

In this work we focus on the data processing requirements and challenges that will soon have to be tackled, if CPPs have to become a reality. We start by defining a prototypical CPP scenario, developed within the European project Cross-CPP [4], involving a set of cars and building sharing data about the environment they perceive (see Section II). We then tackle two fundamental questions. Firstly, what are the requirements associated to a data analytics toolbox for CPP data? We discuss them in Section III, and observe how existing analytics solutions do not by and large comply with them (Section IV). Secondly, what do these requirements imply, in terms of research challenges and future research works (Section V)? Finally, Section VI draws some conclusions.

II. A CPP SCENARIO

The scenario here envisioned stems from the existence of two complementary paths for value generation.

CPSs producing large volumes of data create new value in themselves, for instance by supporting the improvement of existing services. At the same time, such novel data may be the fuel of new business opportunities, not only for the CPS manufacturer, but especially for cross-sectorial industries and applications. There are therefore two perspectives that have to be taken into account:

- The Data Providers, who manufacture or own the CPSs. These should receive an incentive to share their data, i.e. they should receive a benefit beyond what already guaranteed by the in-house consumption of the data. In addition, they are the party most concerned about privacy and security, as CPS data can

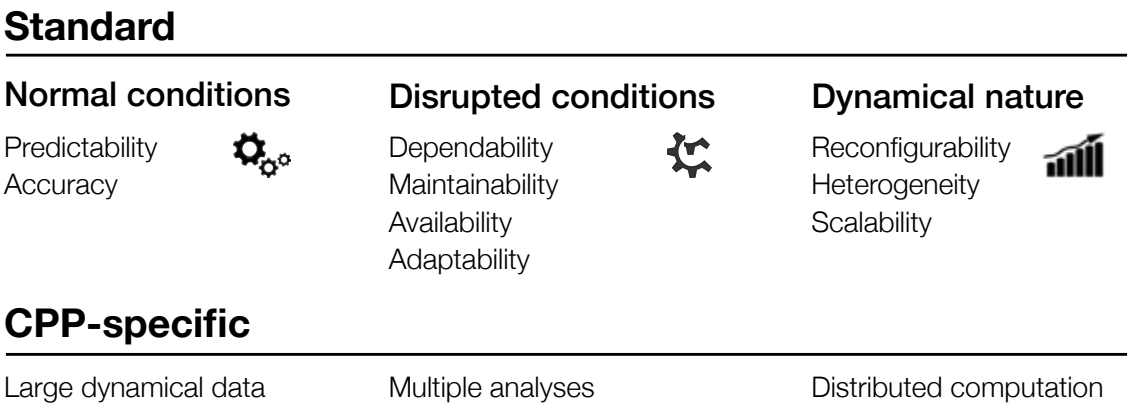


Figure 1. List and organisation of requirements – see Section III for details.

be used to infer activity patterns of the owners and business strategies of the manufacturers.

- The Data Customers, that is, the providers of the new services constructed around the CPS data. They benefit from an easy and homogeneous access to the data, i.e. not depending on multiple protocols and restrictions.

We suppose that data are provided by cars and buildings in cities. This is not arbitrary, but responds to several rationales. Firstly, modern cars are already generating thousands of signals, used to monitor and optimise the functioning of the vehicle; and tests have already been performed on transmitting these data to both local networks [5] and central repositories [6, 7]. While less evident, the same applies to intelligent buildings, in which parameters like temperature and occupancy are constantly monitored [8]. Secondly, cars and buildings can be used to describe the status of a city with high resolution, as they cover most of our daily activities – note that cars can here be intended *lato sensu*, and include e.g. public transportation vehicles. Thirdly, it is easy to envision new services based on these data. To illustrate, weather-related data (including temperature, humidity, or presence of rain) can be used to create detailed weather maps, and hence high resolution weather forecast models; vehicle data can be used to monitor mobility, to detect traffic jams and suggest alternative re-routings in real-time; and, when combined with energy production and consumption trends, to optimise electric vehicle recharging. These three examples share two important features: the user of the CPS (i.e. the car driver or the building inhabitant) gets something in return from sharing the data, e.g. avoiding unexpected traffic jams; and these services would not be feasible without a large scale sharing of data.

In synthesis, the CPP scenario here considered has several characteristics that we expect to be common to most applications, namely:

- Large volumes of data, generated in a continuous (streaming) fashion.
- Redundancy of data, with the same (or similar) values generated by multiple CPSs.
- Potentially low input data reliability, but a required high output reliability.
- Confidentiality of data.

III. Requirements for a Data Analysis Toolbox

The four characteristics listed above have important implications in terms of the requirements associated to a data analysis toolbox, which are discussed in what follows. For the sake of clarity, these have been organised in two groups: standard and CPP-specific requirements – see Fig. 1.

The review of existing literature, and especially of references [9, 10], has led us to the identification of several fundamental requirements for data handling and processing. We term these requirements as "standard", in that they are not specific to CPPs, but are instead common to many data-handling systems. They are organised in three global types, depending on the operative condition of the system.

First of all, under normal conditions, the system will be expected to provide:

- Predictability: ability of foreseeing the system's behaviour or functionality in the future, under standard operational condition. Such prediction could have a qualitative or, ideally, a quantitative nature.
- Accuracy: how close is the system's observed behaviour with the one forecasted or expected from it.

From the point of view of the functioning of the system under disrupted conditions, the system will require:

- Dependability: the user can expect to be able to access the main system's functionalities, even under disrupted operations, without a significant degradation.
- Maintainability: easiness of repairing the system in case of failure, and in general to maintain its optimal operational conditions. This includes both the resources needed, and the resulting downtime.

- Availability: capability of being ready for access even when disruptions occur.
- Adaptability: capability of the system to adapt its internal state and operation, in response to a change in the environment or in the users' requirements, without losing its main characteristics (especially performance-wise).

Finally, when considering the dynamical nature of the system, and thus the possibility of expanding its scope in real-time, the following characteristics will be required:

- Reconfigurability: capability of the system to be easily changed to perform new functions or operations, following an external request. Note that this is conceptually different from adaptability, as in the latter case it is the system itself that decides to change its functioning, as opposed to in response to an external directive.
- Heterogeneity: the system should be able to internally incorporate heterogeneous components, as for instance different computational modules, in order to be able to cope with heterogeneous external inputs (as, for instance, the existence of heterogeneous data sources).
- Scalability: adaptability of the system to the specific problem of an increase (or decrease) in the workload. The system should thus be able to respond to this type of changes, and be able to dynamically use more or less resources as needed. To illustrate, at the beginning of operations, few users may join a new service; but this number may increase fast in case of commercial success. Note that, while scalability is usually understood in the sense of an increment, the opposite is also relevant, as it allows a saving of resources – for instance when a service is not expecting many users at night.

The second group includes requirements that are specific to the characteristics of CPS and CPP data, and that mostly derive from the description of the scenario presented in Section II:

- Capability of handling large quantities of dynamical data, sent by the different sensors composing the system in near real-time. In order to avoid excessive communication overheads, especially when leveraging on mobile communication networks, it may be necessary to send these data in packets. The system should then be able to analyse the incoming information both in streaming and in batches, and eventually to dynamically change between both operational modes.
- Capability of performing several types of analysis, including (but not limited to): supervised classification of different scenarios [11]; unsupervised identification of common patterns in the data [12]; time series forecast [13]; detection of anomalies and transitions in the dynamics of one or more systems [14].
- Possibility of distributing the computation, such that for instance a sensor may decide to transmit information only when the perceived environment changes, or it presents an unexpected behaviour. Besides reducing the workload of the central system, such distributed approach would also help tackling the problem of data confidentiality, as some personal information may be masked at the source.

IV. LIMITATION OF EXISTING SOLUTIONS

With the widespread adoption of data-based solutions in many real-world scenarios, it is not surprising to find a large number of analytic solutions, spanning from cloud pipelines to commercial and freeware software, and both stemming from research activities or having a commercial nature. It is thus only natural to explore the following question: should an analytics toolbox for CPP be made from scratch, or would it be more efficient to use one of these solutions? For the sake of completeness, the most important of them are here reviewed, with a special attention to their shortcomings with respect to the requirements of a CPP application. For the sake of simplicity, those solutions are grouped according to the place where the computation is performed – i.e. local vs. on the cloud.

On one hand, one can find both commercial and freeware software tools for data analysis that are designed to work on a local environment. In this category one can include, for instance, KNIME [15], SPSS Modeller [16], RapidMiner [17] or Alteryx. These software usually have a broad focus, allowing to process any (or most) kind of data. While *prima facie* an advantage, such generality can actually become a drawback. Specifically, most of the functions provided by such tools may not be relevant in a CPP context, and, conversely, some essential functions or algorithms may not be included. A clear example of the latter case are algorithms to detect changes in time series [18], which are not available in the mentioned packages. These are essential to detect when changes in the dynamics of the system occur, which in turn can be used to raise alerts or start computation tasks [19]. Additionally, while it is in principle possible to deploy such software tools in a cloud computing environment, the process is not straightforward, thus hindering the scalability of the system.

On the other hand, it is worth discussing a new approach being introduced by Google, Microsoft or Amazon, among others. This is based on full cloud environments, and in the creation of web-based pipelines in which data are fed, processed, and returned to the user in a completely automatic way. This approach presents two advantages: a complete scalability, and a simplified user experience. At the same time, it usually provides a limited spectrum of possible analysis - for instance, Google ML Engine completely relies on Tensor Flow algorithms [20]. In addition, data have to be transferred to the cloud platform, implying a loss of privacy and control over the data. In some cases the location of the data and of the processing may even not be transparent to the user, something in conflict with recent regulations on personal data protection [21, 22].

Finally, a third category of data analytics solutions is worth mentioning, which positions itself in between the two families previously described. While they are designed for a cloud deployment, they can easily be complemented by a local infrastructure; and they shift the focus towards an intuitive representation of the results and simplified user experience. Among others, these include Sisense, Looker, Zoho Analytics, or Tableau. They usually allow to summarise data on high-level dashboards, with specific applications including business analytics [23] or website usage tracking. They nevertheless do not provide the analytical flexibility required by a CPP application.

The previously mentioned toolboxes have not been developed with CPS applications in mind; their limits with respect to the requirements listed in Section III are therefore not surprising. Still, it is interesting to note that many CPS and CPP applications have been proposed in the literature (the reader may refer to [24, 25] for more exhaustive reviews), which may be used as a starting point for the present development. Nevertheless, in most cases these just included a proof of concept of CPP and CPS scenarios; they thus did not dig into the specificities of data analysis algorithms. In spite of this, a brief overview of three relevant projects is reported below, selected for presenting some commonalities with the CPP scenario here considered.

- Carweb [26]. This project proposed the integration of information coming from cars and mobile phones, focusing on the position and movements of users. The information was then integrated at a higher level, and presented in an interactive way to the user, through a webpage, for the identification of traffic jams in a city.
- SIPRESK [27]. Similarly to the previous project, here authors propose the integration of data coming from distributed sensors to dynamically detect congestion points in a city. With respect to the previous case, here more information sources are considered, including social networks and road sensors.
- CBMANET [28]. Integration and analysis of sensor data gathered from a battlefield. As opposed to the previous cases, here information is generated by fixed stations, and not by the vehicles themselves.

In synthesis, even though many data analytics solutions have been proposed during the last decade, as a result of the increasing relevance of data in business and scientific fields, these do not easily fit the requirements of CPP applications. At the same time, it is not trivial to learn from previous CPS research projects, as most of them did not go beyond the proof-of-concept phase.

V. Challenges and Open Issues

While many requirements listed in Section III are not new, and therefore techniques and technologies have been developed to deal with them, some are pushing the boundaries of standard solutions. In this last part we discuss some challenges and open issues, which may become the foundation of novel lines of research.

A. Information volume, communication bandwidth and data pre-processing

CPP systems are going to generate extremely large volumes of data, both due to the capacity of individual sensors to record measurements at very high frequencies, and to the high number of elements participating in the scenario. While current technology allows to process such massive volumes of data in economical ways, for instance through cloud computing, CPPs present an additional challenge: the bottleneck of transmitting the information from the individual sensors to the data processing system. In other words, the communication bandwidth may be limited, especially if costs have to be kept under control, and such limit will be tighter than those of any subsequent computational step. While communication limitations were part of previous studies of similar scope [5, 6, 7], none of them has considered future scenarios with data volumes of the order of GB per hour.

There are *prima facie* solutions to deal with this, and specifically to reduce the quantity of information to be transmitted by every sensor. Data can be pre-processed locally by each CPS, and transformed through sampling, windowing or dimensionality reduction. Several frameworks, integrating some of these techniques, are available in the literature [29]. Yet, the application of these techniques is far from trivial, especially because they usually require some *a priori* or external knowledge about the system. To illustrate, let us consider the case of a data sampling, in which only one every n measurements is transmitted. Choosing the value of n requires balancing the cost of transmitting potentially redundant information, with the necessity of analysing data with a high-enough time resolution. On top of that, the optimal n may not be constant, but may instead vary according to both the target applications and the dynamics of the system; and multiple elements may share the same infrastructure, such that each element's n must be chosen to optimise a global (as opposed to a local) cost function.

The ideal solution will then require a pre-processing of the data agnostic to the final application, but still optimal in some sense; for instance, a reduction of the data volume that does not reduce the actual quantity of encoded information. While some mathematical instruments are available for this, e.g. entropies [30, 31], their applicability in real-world scenarios has not hitherto been tested.

B. Security: avoiding malicious uses

As any other large-scale infrastructure, CPP systems are going to be vulnerable to external attacks. Besides external attackers (or hackers) that may try to illicitly access the system, there is also the threat of internal elements or users trying to disrupt its functioning. We will here consider the simplest situation of a sensor being programmed to transmit false information - note that this may also be due to a hardware malfunction, and not necessary to a malicious intention.

The solution to this scenario may come from the data integration and fusion step. Once multiple sensors have sent their corresponding data, these should be integrated, in order to recreate a coherent picture of the system. The system may then realise that a sensor is unreliable, or that has come under external attack [32, 33], and thus exclude it from the collective [34, 35, 36]. The simplest solution involves selecting between conflicting options through, for instance, a majority vote. It has nevertheless to be noted that this does not protect against coordinated attacks, in which multiple sensors are programmed to yield erroneous yet coherent information. Additionally, while not essential in the scenario discussed in Section II, security may become critical in the future, especially if CPS data will be used to support safety-critical applications.

C. *Privacy of information*

A final mayor challenge related to CPP systems is the confidential nature of the information by them handled. As previously mentioned, information can be private due to what can be inferred about the behaviour of users, but also about the internal functioning of CPSs.

The standard approach to confidentiality involves, on one hand, ensuring that no private information leaves the system, being thus related to the previous security challenge. On the other hand, it also requires trust, as for instance the service providers have to use data only for the analysis tasks they are authorised to do. This second point may be problematic in a CPP environment, as users have no way of checking the service providers' activity – in other words, they have to blindly trust them.

An alternative solution may be offered by the rising field of *secure multiparty computation* (SMC) [37, 38]. SMC is based on a set of cryptographic protocols that allows a group of parties (or participants) to perform a computation over a data set without actually accessing the data themselves. In other words, the computation is performed over encrypted data, such that the only information gained by parties is the final result.

The use of SMC would yield an important advantage: as CPSs would only share encrypted information, service providers cannot gain any knowledge, and therefore do not have to be trusted by the users. In other words, and to be more precise, the problem of trust is shifted from the service providers to the SMC protocols themselves. On the other hand, SMC also presents some important challenges. First of all, no practical general protocol (what is usually called *fully homomorphic encryption*) has hitherto been proposed [39]; instead, specific protocols exist only for specific applications. Secondly, the computational cost of SMC is usually significant [40], especially because part of the computation has to be performed locally at each CPS. Thirdly, different SMC protocols guarantee different levels of security, e.g. against just one, a few or a majority of malicious participants. In synthesis, while SMC can be a promising future solution, it still does not provide a universal and well-defined answer to CPP privacy concerns.

VI. CONCLUSIONS

The data availability offered by the widespread adoption of CPSs will soon trigger a revolution, in which new services based on them will become part of our daily lives. This will nevertheless not come for free: data have to be analysed, and existing analytics solutions are not designed to deal with the idiosyncrasies of CPPs. As discussed in this contribution, aspects like data pre-processing and compression, security, and privacy, are not easily dealt with by current technologies, and will require future research work.

ACKNOWLEDGMENT

This paper presents work developed in the scope of the project Cross-CPP. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 780167. The content of this paper does not reflect the official opinion of the European Union. Responsibility for the information and views expressed in this paper lies entirely with the authors.

REFERENCES

[1] W. Wolf, "Cyber-physical systems," Computer, vol. 3, pp. 88-89, 2009.

[2] R. Rajkumar, I. Lee, L. Sha, and J. Stankovic, "Cyber-physical systems: the next computing revolution," In Design Automation Conference, pp. 731-736, IEEE, 2010.

[3] T. Jeske, "Floating car data from smartphones: What google and waze know about you and how hackers can control traffic," Proc. of the BlackHat Europe, pp. 1-12, 2013.

[4] A. Correia, C. Wolff, M. Zanin, E. Menasalvas, E. A. Herrmann, V. Corral, M. Kachelmann, R. DeLong, and P. Smrz, "Cross-cpp - an ecosystem for provisioning, consolidating, and analysing big data from cyber-physical products," in Proceedings of the 1st Workshop on Cyber-Physical Social Systems co-located with the 9th International Conference on the Internet of Things (IoT 2019), 2019.

[5] H. Huang, M. Hassan, G. Geers, and L. Libman, "CarAgent: Collecting and disseminating floating car data in vehicular networks," Vehicular Communications, vol. 13, pp. 114-127, 2018.

[6] G. Leduc, "Road traffic data: Collection methods and applications," Working Papers on Energy, Transport and Climate Change, vol. 1, pp. 1-55, 2008.

[7] K. Rehrl, R. Brunauer, and S. Gröchenig, "Collecting floating car data with smartphones: results from a field trial in Austria," Journal of Location Based Services, vol. 10, pp. 16-30, 2016.

[8] Z. Han, R. X. Gao, and Z. Fan, "Occupancy and indoor environment quality sensing for smart buildings," in 2012 IEEE international instrumentation and measurement technology conference proceedings, pp. 882-887, IEEE, 2012.

[9] E. Simmon, K. S. Kim, E. Subrahmanian, R. Lee, F. De Vaulx, Y. Murakami, K. Zettsu, and R. D. Sriram, "A vision of cyber-physical cloud computing for smart networked systems," Gaithersburg, Maryland: US Department of Commerce, National Institute of Standards and Technology, 2013.

[10] V. Gunes, S. Peter, T. Givargis, and F. Vahid, "A survey on concepts, applications, and challenges in cyber-physical systems," KSII Transactions on Internet & Information Systems, vol. 8, 2014.

[11] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer Science & Business Media, 2009.

[12] H. B. Barlow, "Unsupervised learning," Neural computation, vol. 1, pp. 295-311, 1989.

[13] J. D. Hamilton, Time series analysis, New Jersey: Princeton, 1994.

[14] S. Agrawal, and J. Agrawal, "Survey on anomaly detection using data mining techniques," Procedia Computer Science, vol. 60, pp. 708-713, 2015.

[15] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel, "KNIME-the Konstanz information miner: version 2.0 and beyond," AcM SIGKDD explorations Newsletter, vol. 11, pp. 26-31, 2009.

[16] T. Wendler, and S. Gröttrup, Data mining with SPSS modeler: theory, exercises and solutions, Springer, 2016.

[17] M. Hofmann, and R. Klinkenberg, RapidMiner: Data mining use cases and business analytics applications, CRC Press, 2016.

[18] S. Aminikhanghahi, and D. J. Cook, "A survey of methods for time series change point detection," Knowledge and information systems, vol. 51, pp. 339-367, 2017.

[19] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "Segmenting time series: A survey and novel approach," in Data mining in time series databases, pp. 1-21, 2004.

[20] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "Tensorflow: A system for large-scale machine learning," in 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), pp. 265-283, 2016.

[21] M. Goddard, "The EU General Data Protection Regulation (GDPR): European regulation that has a global impact," International Journal of Market Research, vol. 59, pp. 703-705, 2017.

[22] S. Bu-Pasha, "Cross-border issues under EU data protection law with regards to personal data protection," Information & Communications Technology Law, vol. 26, pp. 213-228, 2017.

[23] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," MIS quarterly, pp. 1165-1188, 2012.

[24] F. J. Wu, Y. F. Kao, and Y. C. Tseng, "From wireless sensor networks towards cyber physical systems," Pervasive and Mobile computing, vol. 7, pp. 397-413, 2011.

[25] V. Gunes, S. Peter, T. Givargis, and F. Vahid, "A survey on concepts, applications, and challenges in cyber-physical systems," KSII Transactions on Internet & Information Systems, vol. 8, 2014.

[26] C. H. Lo, W. C. Peng, C. W. Chen, T. Y. Lin, and C. S. Lin, "Carweb: A traffic data collection platform," in The Ninth International Conference on Mobile Data Management (mdm 2008), pp. 221-222, IEEE, 2008

[27] H. Khazaei, S. Zareian, R. Veleda, and M. Litoiu, "Sipresk: A big data analytic platform for smart transportation," In Smart City 360°, pp. 419-430, Springer, 2016.

[28] L. A. Tang, J. Han, and G. Jiang, "Mining sensor data in cyber-physical systems," Tsinghua Science and Technology, vol. 19, pp. 225-234, 2014.

[29] G. D. F. Morales, and A. Bifet, "SAMOA: scalable advanced massive online analysis," Journal of Machine Learning Research, vol. 16, pp. 149-153, 2015.

[30] M. Gell-Mann, and S. Lloyd, "Information measures, effective complexity, and total information," Complexity, vol. 2, pp. 44-52, 1996.

[31] M. Zanin, L. Zunino, O. A. Rosso, and D. Papo, "Permutation entropy and its main biomedical and econophysics applications: a review," Entropy, vol. 14, pp. 1553-1577, 2012.

[32] J. Newsome, E. Shi, D. Song, and A. Perrig, "The sybil attack in sensor networks: analysis & defenses," in Third international symposium on information processing in sensor networks, 2004. IPSN 2004, pp. 259-268, IEEE, 2004.

[33] J. Park, R. Ivanov, J. Weimer, M. Pajic, and I. Lee, "Sensor attack detection in the presence of transient faults," in Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems, pp. 1-10, 2015.

[34] J. Chen, S. Kher, and A. Somani, "Distributed fault detection of wireless sensor networks," in Proceedings of the 2006 workshop on Dependability issues in wireless ad hoc networks and sensor networks, pp. 65-72, 2006.

[35] E. M. Clarke, and P. Zuliani, "Statistical model checking for cyber-physical systems," in International Symposium on Automated Technology for Verification and Analysis, pp. 1-12, 2011.

[36] L. A. Tang, X. Yu, S. Kim, Q. Gu, J. Han, A. Leung, and T. La Porta, "Trustworthiness analysis of sensor data in cyber-physical systems," Journal of Computer and System Sciences, vol. 79, pp. 383-401, 2013.

[37] Y. Lindell, "Secure multiparty computation for privacy preserving data mining," in Encyclopedia of Data Warehousing and Mining, pp. 1005-1009, IGI Global, 2005.

[38] R. Cramer, I. B. Damgård, and J. B. Nielsen, Secure multiparty computation, Cambridge University Press, 2015.

[39] M. Naehrig, K. Lauter, and V. Vaikuntanathan, "Can homomorphic encryption be practical?," in Proceedings of the 3rd ACM workshop on Cloud computing security workshop, pp. 113-124, 2011.

[40] M. Zanin, T. T. Delibasi, J. C. Triana, V. Mirchandani, E. Á. Pereira, A. Enrich, D. Perez, C. Paşaoğlu, M. Fidanoglu, E. Koyuncu, and G. Guner, "Towards a secure trading of aviation CO2 allowance," Journal of Air Transport Management, vol. 56, pp. 3-11, 2016.

Assisting Motorists Using Parking Prediction through a Car App

C. Attard*, A. Naudi ***, S. Mallia**, D.Gauci* and R. Farrugia**
* University of Malta/ Department of Computer Information Systems – Faculty of ICT, Msida, Malta
** University of Malta/ Department of Communications and Computer Engineering Faculty of ICT, Msida, Malta
*** University College London (UCL)/Department of Computer Science London, England
conrad.attard@um.edu.mt

Abstract*—More persons depend on private cars, particularly when alternative transport such as public transport is not as efficient as required. The majority of motorists get caught in queues moving slowly through large cities. Parking becomes more of a challenge in areas where existing car parks provide limited parking spaces. The model for the study was created following an observational study. This required a drone taking top down images for building a dataset, which in turn was used to flag available parking slots consulting historic patterns. The dataset is currently available for research purposes. The vehicle-detection tool developed for this study was used to evaluate the manual logs of the dataset and obtained generally satisfactory results, albeit presenting some limitations. Different regression algorithms were tested on the dataset and the best one overall was selected for making predictions. After considering various techniques, a car app using web technologies and a Node.js framework was built. Through this solution, predictions made using the dataset have been stored in a MongoDB database, and passed on to a motorist through the app. A total of 18 motorists took part in a controlled experiment designed to enable the functional and usability testing of the app.

Keywords—parking dataset, usability, app development, vehicle detection

I. Introduction

The number of cars and new licences released per day is increasing. This surge is not only affecting traffic congestion, but also the time taken to find a free parking space. A study carried out by the Institute of Climate Change and Sustainable Development at the University of Malta revealed that commuters were wasting more than an hour every week [1]. This tends to cause frustration amongst motorists, forcing many to set off at much earlier times in order to reach their respective destinations on time.

The work presented in this paper discusses a dataset of images and manual logging of the activities within different car parks at the University of Malta (UM) chosen for a case study. This data was adopted to train machine-learning models that would be used to predict the number of free parking slots at a particular time. The mobile app in question was evaluated by 18 participants within a real-world, controlled environment.

The following objectives were established towards addressing the existing problem:

- Creating a suitable model for the case study being investigated – a dataset was created for developing and testing the proposed solution. The chosen case study was designed to represent car parks having similar configurations enabling the use drones for capturing images in the top-down approach.
- Developing a vehicle-detection tool for evaluating the manual logs of the dataset. Different algorithms were tested on the dataset.

Finally, this paper presents the design of a controlled experiment that employed the car app designed for this research in a real scenario, whereby the recruited participants were instructed to find a vacant parking space. A usability questionnaire was distributed among participants for the purpose of re-evaluating the said design. This was undertaken with a view to establish the effectiveness of the solution being put forward.

II. Review of Existing Work

A. Parking Spaces Dataset Review

Different research teams have created their respective types of datasets, each one addressing a different aspect of parking-related issues [2] - [5]. This is mainly because of the number of practical applications that could be used when adopting different techniques. An example is the utilisation of CCTV cameras [6]. In fact, CCTV cameras already present in a parking space were exploited using video-processing techniques adopted for analysing the footage captured by these cameras in order to determine the availability of a parking space.

The system is largely dependent on the existing infrastructure. If a CCTV does not cover an entire parking area, the free parking spaces cannot be detected.

Another approach taken was that of using a dash-mounted mobile device in combination with computer-vision techniques to seek unoccupied parking spaces [7]. While this approach addressed the issues resulting from the limitations of fixed cameras by using a mobile camera, this solution would only succeed in identifying an empty parking space when this is already visible to the driver. Nonetheless, it might still be useful in a very busy car park, where a driver might not be able to see an empty parking space easily.

Mathew [8] focused on collecting a large dataset consisting of the number of cars leaving and entering a specific parking facility. Such data was then used to calculate

statistics aimed at making the parking area more efficient in terms of available space. One such statistic is the 'Parking Accumulation', which refers to the number of cars parked at a given point in time. The study then describes an in-out survey. In such a study, the occupancy count in the selected parking area would be taken at the beginning. This requires counting the number of vehicles entering and leaving the parking area during a particular time interval. Accumulation is calculated by adding the number of parked vehicles at the outset to the number of arrivals, minus the number of departures. Occupancy is calculated by dividing the accumulation by the total number of spaces and multiplying the result by 100. This value represents the average occupancy of the specific time interval. All vehicles would be counted at the beginning of the survey. Then, after a fixed time interval that may vary between 15 minutes and 1 hour, the count is taken again. One disadvantage of such an exercise is the possibility of missing the number of vehicles parked for a short duration. Since parking duration and turnover could not be obtained, this method could not be used alone to calculate an ideal parking fare.

B. Computer Vision

Traditionally, object detection could be categorised into two main classes. The first is the sliding-window approach, whereby each image is scanned in raster-scan order, multiple times for different scales and aspect ratios [9]. The second is the region-proposal approach, which uses algorithms to propose different locations, scales and aspect ratios in the image to be passed on to the relatively more expensive feature extraction and classification algorithms. The most recent advances in object detection come from research that strongly relies on convolutional neural networks (CNNs). A faster region CNN (RCNN) [10] uses a region-proposal network that locates many objects such as regions in the image, which are then used in another neural network that classifies the object and adjusts the bounding coordinates. In contrast, [11] adopt a different deep-learning architecture – the YOLOv2 – to the whole image. Each grid is responsible for detecting the bounding box coordinates, classification and confidence of multiple objects. The authors in [12] use a hybrid CNN with multiple convolutional layers connected to a common preceding convolutional layer. In doing so, the model was found to be better at extracting features at multiple scales. In [13], the faster R-CNN [10] model's region-proposal network (RPN) was modified to detect smaller objects more accurately. These deep-learning methods were found to achieve state-of-the-art performance, significantly outperforming classical-based methods.

III. DESCRIPTION OF STAGES

The data was collected through a pre and post study refer to Fig 1 where twelve stages where conducted. To complete the experiment design dataset creation stages 1 to 7 where required. Design of car app was conducted at stage 8 where various prototypes where created and evaluated to achieve the final version. The actual study was conducted from stage 9 to 11 and evaluation of result at stage 12.

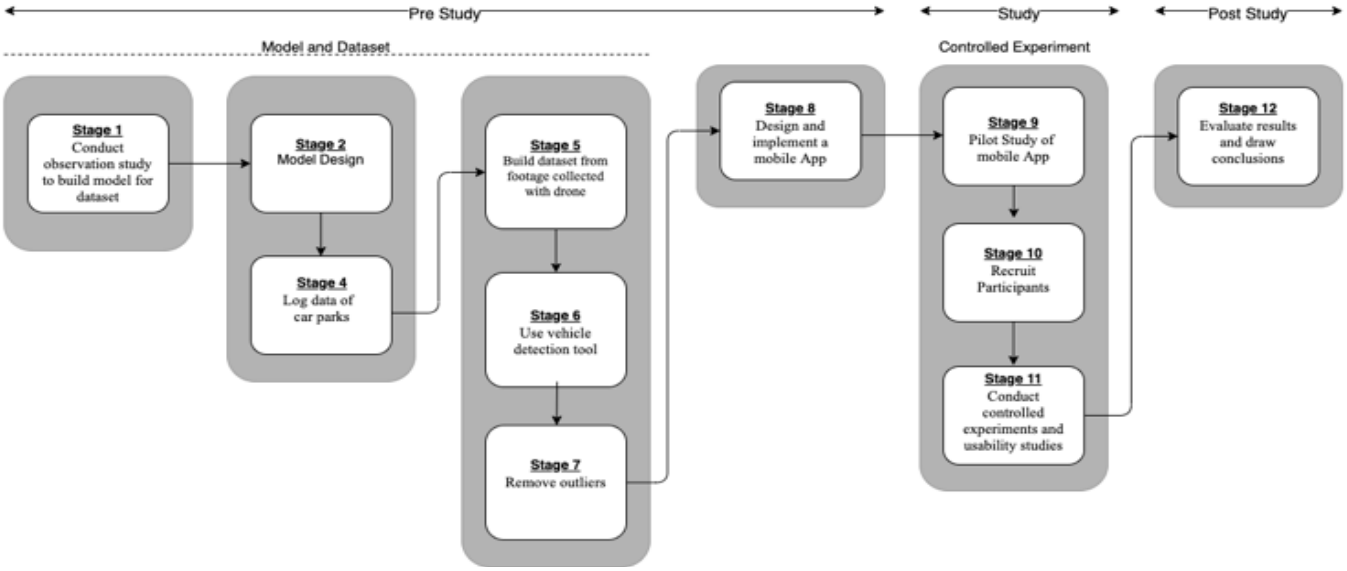


Figure 1. Description of stages

IV. DATASET

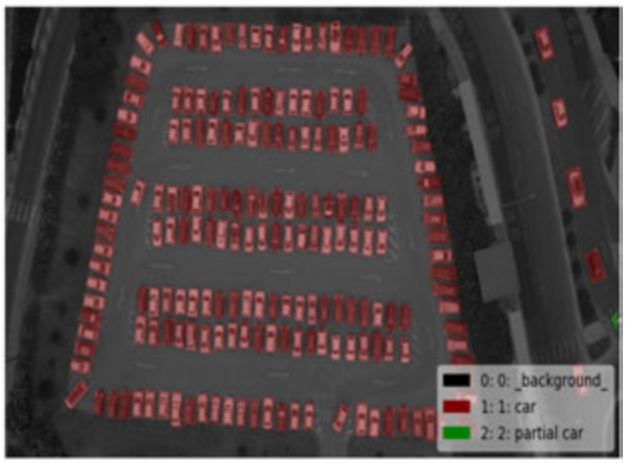


Figure 2. Labelled car park dataset.

University of Malta (UM) has been chosen as a case study and a number of datasets were subsequently created for this purpose. (refer to Stages 1 to 5 in Fig. 1.). The dataset contains a video and images of parking areas. This study was conducted at a specific car park within the said university, over which a drone was used for capturing top-down imagery such as Fig. 2. (Labelled car park dataset). This setup enabled the research team to compile a total of 137 images (101 in November 2018 and 36 in March 2019) over a total of 34 full flights and 1 short flight. – Fig. 1, Stage 5. These images were annotated using the LabelMe image annotation software[1], whereby all the cars within the parking were labelled by a bounding box, as seen in Fig. 2. The number of vacant parking slots at a specific time was also logged and included as metadata, as indicated in Fig. 1. (Stage 4). The dataset can be accessed at this link shorturl.at/abnX2.

V. PROPOSED METHOD

A. Prediction Model

TABLE I . COMPARISON OF ALGORITHMS ON THE MARCH 19 - DATASET

Model	Value	LR	DT	KNN	SVR	Ideal Algorithm
Initial Monday	R^2	0.3593	0.5625	0.344	0.3593	LR/DT
	RMSE	0.9242	1.0271	1.0969	1.2334	
Departures Monday	R^2	0	0.6019	0.2239	0	DT
	RMSE	1.2135	1.0282	1.2381	1.2247	
Arrivals Monday	R^2	0	-0.3808	0.2239	0	DT
	RMSE	0.8683	1.1997	1.2381	0.866	
Accumulation Monday	R^2	0.6434	0.6099	0.6099	0.6434	LR/DT/KNN
	RMSE	1.073	1.0969	1.0969	1.013	
Initial Wednesday	R^2	0	-0.28	0.0475	0	DT
	RMSE	1.781	2.3783	2.2638	1.3338	
Departures Wednesday	R^2	0	0.3109	-0.1362	0	DT
	RMSE	4.062	3.9987	4.1079	4.4549	
Arrivals Wednesday	R^2	-0.3398	-0.6949	-0.6422	-0.3398	DT
	RMSE	2.1933	2.2879	2.3333	2.1517	
Accumulation Wednesday	R^2	0	-0.247	-0.0455	-0.2718	DT/SVR
	RMSE	4.3211	5.3169	4.7177	4.8636	
Initial Friday	R^2	0.8275	0.8214	0.8721	0.8275	KNN
	RMSE	5.094	5.3268	4.5586	5.0013	
Departures Friday	R^2	0	-0.4418	-0.6149	0.4778	SVR
	RMSE	1.002	2.0991	2.243	0.9225	
Arrivals Friday	R^2	0.9686	0.795	0.5577	0.9686	LR
	RMSE	0.7343	1.1859	1.2374	0.8377	
Accumulation Friday	R^2	0.9029	0.8792	0.9129	0.9029	LR/KNN/ SVR
	RMSE	4.747	6.1498	4.5377	4.7489	

The prediction model was trained using the logged information that included the day, time and number of available parking slots. This data was split into two: 66% used for training and 34% used for testing. For each algorithm, the R2 and RMSE values were recorded. Different regression models were trained on this data including linear regression (LR), decision tree (DT), *k*-nearest neighbours (*k*-NN) and support-vector regression (SVR). The results displayed in *TABLE* I. summarise the performance of these regression models. The best performance was achieved through DT, which will be therefore used in the mobile app to predict the number of available parking slots at a specific time.

B. Vehicle Detection from Drone

Due to the limited amount of labelled data needed to train the YOLOv2 architecture [12], the CNN model was trained on the CARPK dataset [14] and then evaluated on the UM Dataset. The optimal network architecture was identified through Bayesian optimization, using a method similar to the one presented in [15]. A Nvidia Titan X graphics card with 11GB of memory was used to train each architecture individually. These were each trained for 400 epochs – that is, the training algorithm goes through all the training data 400 times, where each train run lasted for approximately 12 hours. The table II below outlines the best-performing CNN architecture.

TABLE II. BEST-PERFORMING YOLOV2 CONFIGURATION

Type	Filters	Size/Stride	Output
Convolutional	32	3x3	544x544
Max-pool		2x2/2	272x272
Convolutional	64	3x3	272x272
Max-pool		2x2/2	136x136
Convolutional	128	3x3	136x136
Convolutional	128	3x3	136x136
Convolutional	128	3x3	136x136
Convolutional	128	3x3	136x136
Convolutional	128	3x3	136x136
Convolutional	64	3x3	136x136
Max-pool		2x2/2	68x68
Convolutional	1024	3x3	68x68
Convolutional	1024	3x3	68x68
Convolutional	128	3x3	68x68
Convolutional	18	1x1	68x68
Avg-pool		Global	
Softmax			

Average Precision: 0.878
Average Precision on test (non-validation) data: 0.881

C. Mobile App

MeteorJS framework was used to develop the mobile app. The main advantage of MeteorJS is its extreme reactivity: in this framework the server does not send HTML but data that reaches the client without the need to refresh the page. Thus, the interface always reflects the state of the database with minimal effort needed, allowing the app to behave in a real-time manner. This characteristic was used to update parking information in a real-time manner depending on the time of day. A Web API was used to implement voice commands to enable hands-free use. The Google Distance Matrix API was used to obtain the driving distance to the parking area. The main features of the app are:

[1] http://labelme.csail.mit.edu/Release3.0/

- Provides parking predictions of various car parks at the current time (refer to Fig.3.)
- Integrates with Google Maps to give directions to the selected car park
- Provides a quick way of selecting the ideal car park at the current time
- Displays a schedule of future predictions.



Figure 3. Main screen of car app. (Ref to table III)

TABLE III. FUNCTIONALITY OF CAR APP

Element	Functionality	Voice Command
1	These buttons are used to show the data of the next or previous car park in the list.	"Next"/ "Next parking"/ "Previous"/ "Previous parking"
2	This button is used to launch Google Maps and get directions to the chosen car park.	"Get directions"/ "Directions"
3	This button is used to open the screen shown in Figure 9.	"Open menu"/ "Menu"
4	This button is used to switch to the suggested (ideal) car park at the current time. Yellow colouring indicates that driver is on the suggested car park.	"Show suggested" / "Suggested"
5	This label shows the name of the chosen car park and the driving distance from the user's location.	N/A
6	This label shows the time interval which the data on the dashboards relates to.	N/A
7	This dashboard shows the number of predicted arrivals for the time interval. On click, the screen in Figure 10 is opened.	"Show arrivals"/ "Arrivals"
8	This dashboard shows the number of predicted empty spaces at the end of the time interval. On click, the screen in Figure 10 is opened.	"Show vacancy"/ "Vacancy"

VI. CASE STUDY

This study revolves around an experiment consisting in monitoring parking facilities that are densely populated on a daily basis, meaning that the number of motorists seeking a parking space is equal to, or higher than, the number of the available parking spaces.

The experiment was conducted by eighteen motorists between 01 April 2019 and 08 May 2019 – three motorists for each case study in the model. Moreover, since the experiments were conducted in April and May, the March dataset was used to make predictions. A Samsung Galaxy S4 phone was provided throughout the experiment. No conditions were imposed on participants, except they had to be at least 18 years old and in possession of a valid driving licence. The experiment involved motorists using the car app to find a vacant parking space on the UM campus. Meanwhile, the investigator accompanied the participants, observing the situation, taking notes and building a log of all that was happening. Safety measures were taken towards ensuring that both driver and investigator were safe at all times.

Since only data was only collected for Car Park 6, a similar dataset for other car parks around the university campus was built.

VII. RESULTS

A. Controlled Experiment

The controlled experiment was a very important component of the study, as it allowed the functional and usability testing of the application. Furthermore, it made it possible to test the prediction module in a real-world environment. Fig. 4. shows the number of motorists who found a vacant spot immediately upon arriving at Car Park 6, those who had to wait for another person to leave, and those who left Car Park 6 in order to park elsewhere.

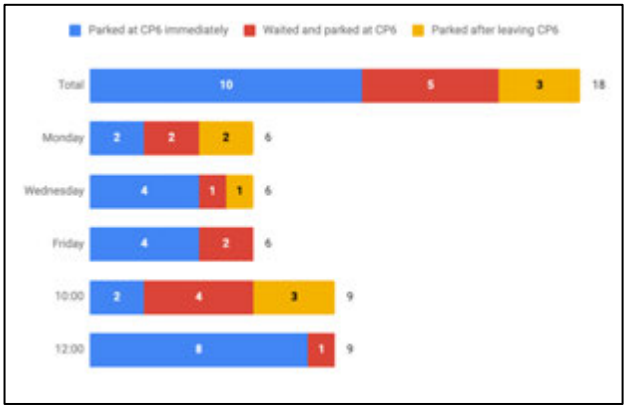


Figure 4. Stacked bar chart showing the outcomes of the controlled experiments.

The different bars indicate that it was most difficult to find a parking spot on a Monday, whereas Friday was the best day. Furthermore, the bar chart shows that 12:00 is a much better time to park than 10:00, with eight motorists out of nine finding a slot upon arrival. In total, fifteen of the participants managed to park in Car Park 6. The other three left the car park after waiting for some time, but eventually succeeded in parking elsewhere. Two out of these three cases fell into the Monday 10:00 slot. This slot was the busiest of the six chosen slots and the prediction module predicted a full car park, with a very low number of departures.

B. Accuracy of Vehicle Detection (VD)

This section discusses the results obtained after using the vehicle-detection tool developed for this study. As discussed in the previous section ('Proposed Method'), Java was used to develop a tool to remove any outliers from the total number of cars found (refer to Stage 7 in Fig. 1.). Furthermore, the number of unparked and illegally parked cars was also removed from the count. After this, the percentage error

between the actual accumulation and VD accumulation was calculated for each row in the dataset. Table IV. below shows the mean error percentages for the November - December dataset, the March dataset and the full dataset. In the case of negative numbers, the modulus was taken.

TABLE IV. MEAN PERCENTAGE ERRORS OF VEHICLE-DETECTION TOOL

Dataset	Mean Percentage Error
November – December	8.7%
March	15.3%
Full	10.4%

The cumulative distribution functions (CDF) shown in Fig. 5. provide the probability (on the y-axis) that the error percentage between the actual accumulation column and the VD accumulation will be less than. Or equal to, particular value a value on the x-axis. These graphs make it possible to compare the probability of getting a specific percentage error for the two datasets.

As shown in Fig. 5., the VD accumulation graph for the November - December dataset was more accurate than the graph concerning the March dataset. Fig. 5. show a 90% probability of the occurrence of an error in the November - December dataset being less than 15%. Meanwhile, this statistic also indicates that for the March dataset there is approximately only a 50% chance of the occurrence of an error less than 15%. For further evidence of the difference between the two datasets, one could also refer to Table IV., which shows a difference of 6.6% between the mean percentage errors of the datasets. Possible reasons for these discrepancies include:

- Difference in lighting
- Varying levels of traffic congestion
- Position and angle of the drone when taking footage.

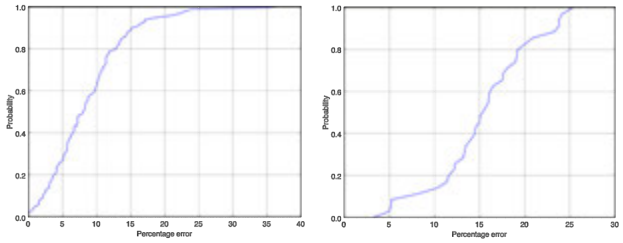


Figure 5. CDFs of percentage error for November - December images and for March images.

C. Interface of the App

Table V lists the different statements concerned with the usability of the interface. The mean rating score (Mean) for each statement is presented together with the standard deviation (Std Dev.), the minimum (Min) choice and the maximum (Max) choice. The mean rating scores allocated to these five statements all exceed 3, indicating that the participants tended more to agree more than to disagree. The *p*-value (0.058) is higher than the 0.05 level of significance, meaning that there is no significant difference between the MRSs of these statements.

TABLE V. STATEMENTS CONCERNING THE INTERFACE OF THE CAR APP

Statement	Mean	Std. Dev.	Min	Max
The experience was satisfactory and smooth rather than frustrating and distracting.	3.78	0.428	3	4
The application worked without errors and delays.	3.72	0.461	3	4
Information on the display was easy to understand and memorise without requiring too much concentration.	3.83	0.383	3	4
The interface was consistent throughout, well laid out and visually appealing.	3.44	0.511	3	4
Using the application was straightforward.	3.72	0.461	3	4

$X^2(4) = 9.125$, $p = 0.058$

D. Usability

The usability questionnaire was answered by the motorists participating in the controlled experiment. The participants were asked to indicate whether they agreed or otherwise with different statements, which mainly related to:

- their overall experience using the car app;
- the appearance and performance of the interface;
- the various screens and features of the app;
- their safety whilst using the app.

The Friedman Test was used to compare MRSs between a number of related statements. These MRSs range from 0 to 4, where 0 corresponds to 'Strongly Disagree' and 4 to 'Strongly Agree'. The higher the MRS, the stronger the concurrence. The null hypothesis specifies that the MRSs allocated to the statements are similar, and the result is accepted if the *p*-value exceeds the 0.05 level of significance. The alternative hypothesis specifies that the MRSs vary significantly between the statements and the result is accepted if the *p*-value is less than the 0.05 criterion.

The MRSs allocated to all the statements concerned with the usability of the interface all exceed 3, indicating that the participants tended to strongly agree. The *p*-value (0.058) is higher than the 0.05 level of significance, meaning that there is no significant discrepancy among the MRSs of these statements. The same could be said for statements concerned with the safety of the motorists, where the *p*-value is 0.727. The participants tended to strongly agree that they felt safe and could focus on the road. When asked about the different methods of input, the participants tended to agree that both the control buttons and the voice commands were simple and effective. Although there is no significant difference between

the MRSs (p-value of 0.119), the participants nevertheless preferred the control buttons to the voice commands. This is further corroborated by the fact that most of the participants actively preferred the control buttons rather than the voice commands when asked to choose between the two.

All participants stated that they would recommend the app to fellow motorists. When asked to rate the app from 0 to 4 (4 being the highest), the mean of all the answers was a highly favourable 3.94.

VIII Discussion

This paper seeks to propose a solution that depends on the quality of the dataset created and the tool used to process the images collected. While being aware that the vehicle-detection tool has its pitfalls, the research team has established that one way of improving the accuracy of the VD accumulation would be to improve the tool itself. The model tends to fail when given different vehicle scales. However, zooming in or out of images could be a solution. Another proposed solution would be to perform augmentation on the data, focusing on scale augmentation. This could be achieved by augmenting the training data with different rescaling's and then optimising the YOLOv2 architecture configuration on the augmented data. Another problem encountered is vehicles hidden by the shade. The solution proposed is to train the model on data extracted at varying degrees of shade.

The VD accumulation column is to improve the process of removing outliers from the total count. Although the outlying areas in the implemented solution were calculated mathematically for each different picture, having the same areas every time would help significantly.

To evaluate the dataset created for this investigation, the research team conducted a study that validated and tested the prediction module in a real-world environment. Success of parking recorded in the study is informed by the statistics gathered from 18 participants.

The accurate dataset has been achieved and could be used for the model designed for this research. The success of the images created for this study have a reliable outcome, and could therefore be used in the future without the need for manual logging. A high success rate that was matched to the logged data was achieved.

A common challenge that researchers in this field very often encounter is the lack of a representative dataset to perform their experiments. Through this study, the research team managed to create a dataset that allowed the team to conduct the experiment and test it in a real-world case study.

The participating motorists managed to use the mobile app successfully, meeting the minimum criteria. Further development of voice-over and information about car park patterns – including an innovative and safe way to guide cars to specific parking spaces – will significantly improve the user experience of driving with the help of technology.

More work could be done to adapt the model for a broader time frame. This would accommodate various combinations, including time of day and different parking combinations that are not limited to the time allocated for this research.

Acknowledgment

This research is part of SmartAkyAuomator project which was partially funded by the University of Malta. For the initial stage, Mr Daniel Cauchi assembled the drone that was used for this research. Transport Malta provided us with all information and support including the license to fly the drone following the safety measures required by law.

References

[1] "33 new cars per day as Malta records highest CO2 emissions increase in EU," *The Malta Independent*, 12-May-2018.

[2] L. Zhang, J. Huang, X. Li and L. Xiong, "Vision-Based Parking-Slot Detection: A DCNN-Based Approach and a Large-Scale Benchmark Dataset," in *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5350-5364, Nov. 2018. doi: 10.1109/TIP.2018.2857407

[3] Example of intelligent parking systems S. U. Raj, M. Veera Manikanta, P. S. Sai Harsitha and M. Judith Leo, "Vacant Parking Lot Detection System Using Random Forest Classification," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2019, pp. 454-458.doi: 10.1109/ICCMC.2019.8819689

[4] S. Di Martino, V. N. Vitale and F. Bock, "Comparing Different On-Street Parking Information for Parking Guidance and Information Systems," *2019 IEEE Intelligent Vehicles Symposium (IV)*, Paris, France, 2019, pp. 1093-1098. doi: 10.1109/IVS.2019.8813883

[5] Almeida, Paulo & Soares de Oliveira, Luiz & Jr, Alceu & Jr, Eunelson & Koerich, Alessandro. (2015). PKLot - A Robust Dataset for Parking Lot Classification. Expert Systems with Applications. 42. 10.1016/j.eswa.2015.02.009.

[6] M. Y. I. Idris, E. M. Tamil, Z. Razak, N. M. Noor, and L. W. Kin, "Smart Parking System using Image Processing Techniques in Wireless Sensor Network Environment," Inf. Technol. J., vol. 8, no. 2, pp. 114–127, Feb. 2009.

[7] S. Patricio Figueroa Sanz, J. Tran, and C. Yu Wang, "Street Parking Detection."

[8] T. V. Mathew, "Parking Studies," in Engineering, Transportation Systems, 2014, pp. 1–11.

[9] P. Felzenszwalb, D. McAllester, and D. Ramanan, A discriminatively trained, multiscale, deformable part model," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8, June 2008.

[10] S. Ren, K. He, R. B. Girshick, and J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks," CoRR, vol. abs/1506.01497, 2015.

[11] J. Redmon and A. Farhadi. YOLO9000: Better, faster, stronger. arXiv preprint arXiv:1612.08242, 2016.

[12] X. Chen, S. Xiang, C. L. Liu, and C. H. Pan, Vehicle detection in satellite images by hybrid deep convolutional neural networks," IEEE Geoscience and Remote Sensing Letters, vol. 11, pp. 1797-1801, Oct 2014.

[13] W. Sakla, G. Konjevod, and T. N. Mundhenk, Deep multi-modal vehicle detection in aerial isr imagery," in 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 916{923, March 2017.

[14] K. Kandasamy, W. Neiswanger, J. Schneider, B. Poczos, and E. Xing. Neural Architecture Search with Bayesian Optimisation and Optimal Transport. ArXiv e-prints, Feb. 2018.

[15] M. Hsieh, Y. Lin, and W. H. Hsu. Drone-based object counting by spatially regularized regional proposal network. CoRR, abs/1707.05972, 2017.

BIOMEDICAL ENGINEERING

Automatic Brain Extraction in Nissl-Stained Histological Sections

T. Smolčić*, T. Petković ** and S. Škokić *
* University of Zagreb / School of Medicine, Zagreb, Croatia
** University of Zagreb / Faculty of Electrical Engineering and Computing, Zagreb, Croatia
tomislav.smolcic@mef.hr

***Abstract* - A simple and robust algorithm for automatic segmentation of Nissl-stained histological brain images is presented. The algorithm preprocesses the input image with anisotropic diffusion filter prior to applying standard binary image manipulations to isolate the brain from the background. Shape smoothing filter is applied to mimic the common practice of human evaluators. The developed code is tested on histological sections of mouse brains and shows good agreement with manual segmentation results.**

Keywords – brain segmentation; anisotropic diffusion; shape smoothing; histology

I. INTRODUCTION

Segmentation of brain in different medical images is an important element of tracking morphological changes in longitudinal observation of stroke, tumors and other brain pathologies. The changes in pathology volume are usually measured and normalized against ipsilateral (affected by pathology) and contralateral (unaffected by pathology) hemisphere volume, or against total brain volume.

During the course of a typical pre-clinical experiment, rodent brain and pathology volume will first be measured on images obtained with *in vivo* imaging methods such as MRI or micro-CT. These findings will subsequently be compared to histological analysis performed on the brain *ex vivo*. In this context, the histological analysis is commonly regarded as the "ground truth," since it allows visualization of biological processes on a cellular level, with specific biological markers and at micrometer resolutions.

Histological staining is a method used to examine structural layout in tissue. Tissue of interest is preserved using chemical fixatives and cut into 10-20 μm thick slices. One of the most frequently used staining techniques is Cresyl violet or so-called Nissl staining [1]. The Nissl substance binds to the acidic components of the neuronal cytoplasm, such as the RNA and gives a violet appearance to regions where it accumulates [2], as shown in Fig. 1.a. After cellular damage is caused by pathology (i.e. stroke), the cells degenerate and the binding weakens, causing the affected area to appear pale in comparison to healthy tissue, as illustrated in Fig. 1.b.

For scanning, histological sections are mounted on a glass slide and transferred to a high-resolution optical scanner. Due to the nature of the process, a number of artefacts can appear: tearing and folding caused by slicing (Fig. 1.c), uneven coloration or various other artefacts such as dust and speckles. These artefacts can manifest as noise in the background or as more prominent image features in the foreground, sometimes even on top of the brain itself (Fig. 1.d).

A necessary prerequisite for segmentation of histological brain images is successful separation of brain tissue from the background. While much simpler than co-registration of different brain regions to histological or MRI brain atlases [3] [4] [5], it is often reported to be done manually or at best semi-automatically. For example, Shiffman et al. [6] manually removed the background, orientation markers and scanning artefacts, and extracted individual brain sections via trial and error intensity thresholding, reporting on average 5 minutes of user time per section. It is also worth noting that manual segmentation, besides making the analysis very time-consuming, is also known to introduce inter- and intra-observer variability [1] [7].

In this paper, we propose a simple and robust algorithm for automatic brain extraction in histological images. The proposed algorithm is based on anisotropic diffusion [8] and shape smoothing [9]. Preliminary results are shown and compared to manual segmentation done by two independent expert evaluators.

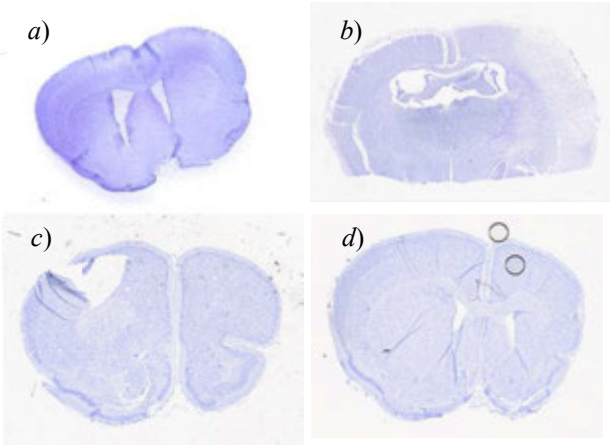


Figure 1. Histological sections. a) Evenly stained healthy brain without artefacts; b) Brain section with ischemic stroke in the right hemisphere; c) Brain with torn tissue and folding artefacts; d) Artefacts appearing both in the background and over the brain tissue.

II. PROPOSED METHOD

The outline of the proposed algorithm is shown in Fig. 2. In addition to the abovementioned anisotropic diffusion and shape smoothing, the procedure involves several common binary operations and filters, as well as a function for centering and rectifying the image after background subtraction, all in order to provide the user with a manageable object for manual inspection and corrections.

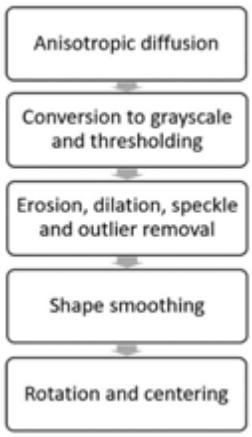


Figure 2. Flowchart of the proposed algorithm. Binary filters are applied before shape smoothing filter in order to remove small artefacts caused by dust and debris.

A. Anisotropic diffusion

Anisotropic diffusion (AD) is a well-known technique aiming to reduce noise in an image without affecting important image features, such as edges. It was proposed in 1990 by Perona and Malik [8] and revised by Tschumperle and Deriche [10]. It combines linear filtering and edge identification into one nonlinear diffusion equation:

$$\frac{\partial I_i}{\partial t} = \sum_{j=1}^{n} \mathrm{trace}(\mathbf{A}^{ij}\mathbf{H}_i) \qquad (1)$$

where I_i is the i-th channel of the RGB image $\mathbf{I}$ ($n = 3$), t is time, $\mathbf{A}^{ij} = \delta_{ij}\mathbf{A}$ is a family of 2x2 symmetric matrices of diffusion tensor $\mathbf{A}$ and $\mathbf{H}_i$ is the Hessian matrix of I_i. The diffusivity is controlled by the diffusion tensor $\mathbf{A}^{ij}$, which varies according to the features of the image. In homogeneous regions, isotropic smoothing is preferred, while around the edges the tensor $\mathbf{A}^{ij}$ becomes anisotropic in order to preserve the orientation of the edge. The effect of the filter constructed this way on a histological brain scan is demonstrated in Figs. 3.a and 3.b.

A welcome side-effect of anisotropic diffusion is the fact that it effectively removes most of the background noise caused by fine dust speckles or introduced by the scanner. Any remaining noise in the background is further suppressed with a median filter.

Table I presents the input parameters of the anisotropic diffusion filter used in this study. The key parameters controlling the shape of the smoothing mask are diffusion limiters A1 and A2, while edge threshold height defines the minimum edge "strength" that will be preserved by the filter [11].

TABLE I. ANISOTROPIC DIFFUSION PARAMETERS

Parameter	*Value*
Number of iterations	20
Smoothings per iteration	10
A1 - Diffusion limiter along minimal variations	0.5
A2 - Diffusion limiter along maximal variations	1.5
dt – Time (iteration) step	20
Edge threshold height	5

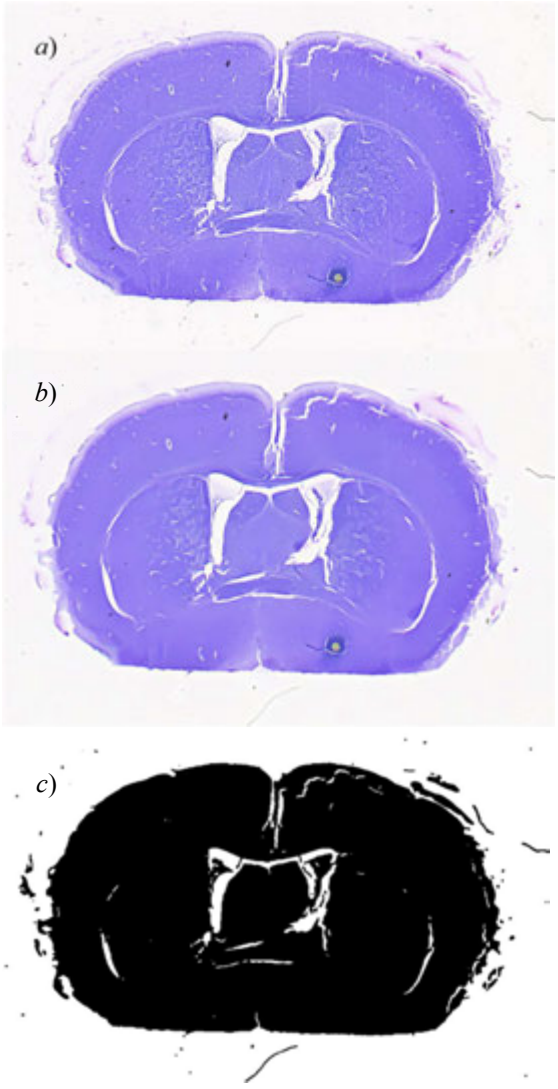


Figure 3. *a*) Histological brain section before AD; *b*) Histological brain section after AD: the main features of brain morphology are preserved while noise and speckles have been suppressed; *c*) binary mask of the AD-filtered brain image with gaps around ventricles and hippocampus

B. Creating a binary mask of the whole brain

Having removed the noise from the background, simple thresholding using the median grey value of the image as a threshold converts the image into a binary mask.

In order to further suppress medium-size artefacts and smoothen the mask edges, a series of binary dilations and erosions are applied. This smooths small wedges, closes

holes in the binary mask and eliminates tiny isolated dots. It should be noted, however, that ventricles and other larger bright structures are not removed and will remain as gaps inside the mask, as shown in Fig. 3.c. These gaps are added to the mask applying the Fill Holes algorithm [12] [13]. The algorithm flood-fills the background of the image starting from the image border, and any holes within the foreground shape cannot be reached. These non-flooded pixels are then included in the binary mask.

The brain is selected as the single largest connected foreground region in the image. However, at this point, besides the mask representing the brain, various other smaller shapes may still be present in the image, whether representing larger pieces of dust, hair or alignment markings, or caused by scanning artefacts. It is also possible that the section has suffered heavy tearing and is split into multiple pieces. To remedy for that, all objects smaller than a set threshold, relative to the largest object in the image, are removed.

C. *Shape smoothing*

During the process of preparing histological sections, all parts of the brain tissue are not evenly colorized. This is especially true for brain edges, where small-scale abrupt indentations in the binary mask can appear as a result of weak staining. A similar effect can also appear due to tissue tearing.

Expert evaluators usually ignore such small-scale discontinuities under the assumption that their contribution to total measured volume is negligible and that the brain surface before slicing was smooth and without discontinuities. Another justification for this practice is time management, since manual tracing of all fine details significantly increases the segmentation time.

To mimic this common practice, the shape smoothing filter is applied [9]. The filter extracts contours of all objects in the image, applies the discrete Fourier transformation to the contours, and truncates the obtained Fourier series to a user-defined number of harmonics. The result is a smooth approximation of the original contour, as shown in Fig. 4.

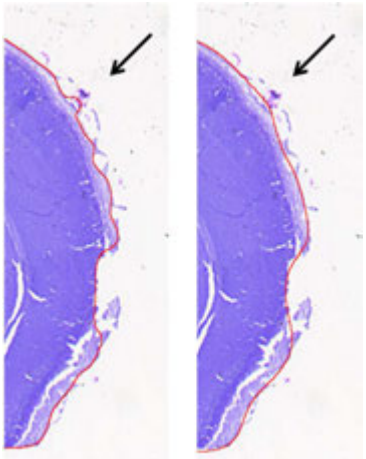


Figure 4. Segmented edge before (left) and after shape smoothing (right). Bumps and indentations in the original contour are straightened out after processing.

When analyzing multiple slices, which is common in pre-clinical practice, it is practical to arrange the analyzed histological slices in a stack. In the proposed algorithm, this takes place after aligning the processed images. Aligning is achieved by fitting an ellipse to the isolated brain and measuring its tilt angle and distance to the frame center. After that, the brain is rotated by the negative value of the measured angle and translated by the measured distance towards the center of the frame.

III. RESULTS AND DISCUSSION

The proposed method was implemented in FIJI [14], a distribution of ImageJ [15]. ImageJ is an open source Java-based image processing platform which allows users to combine built-in image processing tools and develop own image analysis programs in Java, Python, Ruby and other languages.

Our algorithm was implemented using freely available FIJI plugins for anisotropic diffusion [11] and shape smoothing [9], which are combined with built-in binary tools (erosion, dilation, hole filling) and filters (despeckle and outlier removal) [16]. Aside from the optimal parameters for anisotropic diffusion and the truncation parameter of Shape Smoothing filter, all other parameters are automatically set based on the properties of the analyzed image.

The performance of the algorithm was tested on histological scans of mouse brains obtained courtesy of projects RepairStroke and Elettra. Ten glass (?) slides, each containing on average 12 histological slices, as shown in Fig. 5, were used as the test dataset. The output was compared to manual measurements done by two independent expert evaluators, which we consider the ground truth.

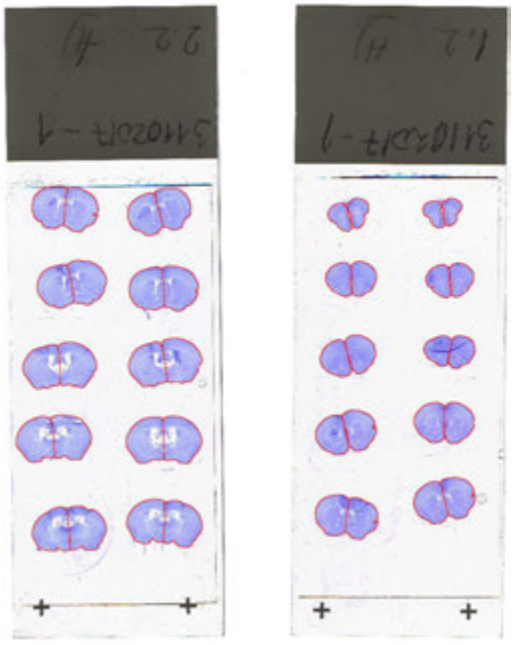


Figure 5. Microscopy glass slides with Nissl-stained histological mouse brain sections, used for testing the developed algorithm. Red outline denotes segmentation by an expert evaluator.

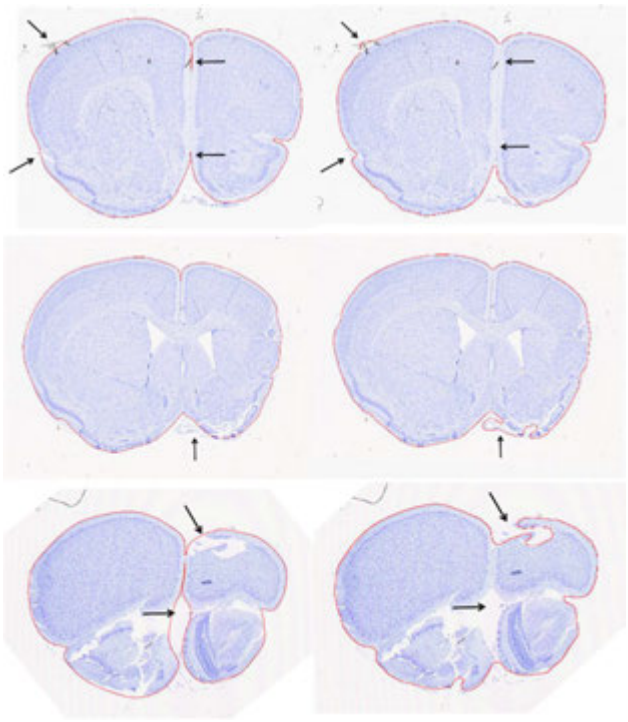


Figure 6. Manual segmentation by an expert evaluator (left column) vs. automatic segmentation via the proposed algorithm (right column).

Fig. 6 shows several typical examples depicting the strengths and limitations of the proposed method. Manually segmented sections are shown in the left column, and sections segmented using our method in the right column.

In the first example, the automatically detected edge fits to the brain well, but is unable to distinguish dust from brain tissue in the top left corner. One can also notice that the manually delineated edge penetrates deeper into the brain at the midline, compared to the edge produced by our algorithm, which is a side-effect of shape smoothing. The expert evaluator, however, also smoothened the edge by ignoring the dust in the top left corner and the gap in the bottom left corner of the brain. This helps justify the use of the shape smoothing in the proposed method. Furthermore, bearing in mind the properties of the Fourier transformation, one can expect that the contributions to the total area of deviations of the Fourier-fitted smooth contour from the initial contour will have a negligible effect on the total delineated area.

The second example demonstrates the ability of the algorithm to follow finer details along the brain edge, despite keeping only a small fraction of Fourier harmonics (4 %). However, it cannot distinguish between brain tissue and any other neighboring stains, unlike the expert evaluator.

The last example is an anterior section of the mouse brain, showing significant tearing along the edge as well as hemisphere separation in the middle. In this part of the brain, cerebrum ends and the olfactory bulb begins, and oftentimes these separate bodies will split during the cutting of histological slices. Our algorithm creates a unified edge around the whole brain, closing the gap in the middle as if it were an inherent part of the cerebrum, similar to the treatment of ventricles in Fig. 3. The expert evaluator, on the other hand, correctly interpreted and excluded this gap. It is interesting to note how the expert evaluator included the empty space around the top midline area in the delineated brain area, effectively reconstructing the missing brain part. This is a common – albeit somewhat scientifically dubious – practice founded in the assumption that the brain was intact prior to the cutting. While taking into account that some sections may suffer tearing or folding, the algorithm presented herein does not attempt to reconstruct empty sections and only attributes real, non-empty pixels to the brain.

Despite all the individual delineation differences, the overall agreement between automatically analyzed scans and manually analyzed scans (ground truth) is very good. The algorithm was tested against two expert evaluators on 10 microscopy glass slides, each containing on average 12 histological slices, as shown in Fig. 5. Preliminary results show an average difference in total measured volume of 1.98 %, while inter-evaluator difference on the same dataset was 0.84 %.

However, the difference in analysis time is much more significant. On average, manual segmentation can take up to 2 minutes per slice depending on the complexity. Our program completes the same task in about 20 s on an average office PC (Intel i3, 16 GB RAM, integrated GPU). Still, bearing in mind that an average brain will yield 50-60 slices, delineating the brain automatically will save on average 1-2 man-hours per animal, or up to 50 hours or even more per experiment. Finally, most segmentation inaccuracies can still be manually corrected afterwards in a fraction of the time, owing to the application of shape smoothing, which reduces the number of anchor points in the detected edge.

IV. CONLUSION

A novel approach to segmentation of Nissl-stained histological brain images has been presented. The proposed algorithm applies anisotropic diffusion filtering to reduce noise and artefacts while preserving key morphological features of the brain, allowing the use of standard binary image manipulations to separate the brain from the background. Shape smoothing filter reduces the number of anchor points in the detected edge, mimicking to some extent the common practice of human evaluators.

By automatically extracting the entire brain from the background, subsequent volumetric assessment of brain hemispheres is significantly simplified, underpinning the practical value of the proposed algorithm either as a tool for assisted segmentation or a starting point for more advanced segmentation techniques based on active contours.

ACKNOWLEDGMENT

This work was supported by the following grants: Croatian Science Foundation project RepairStroke (IP-06-2016-1892), Synchrotron Light Source 'Elettra' grant n.20170140 and European Regional Development Fund, Operational Programme Competitiveness and Cohesion, grant agreement No.KK.01.1.1.01.0007, CoRE – Neuro. We thank our colleagues Anja Baric and Helena Justic for supplying manual segmentation results.

REFERENCES

[1] M. Zille, T.D. Farr, I. Przesdzing, J Muller, C. Sommer, U. Dirnagl and Andreas Wunder, "Visualizing cell death in experimental focal cerebral ischemia: promises, problems, and perspectives," Journal of Cerebral Blood Flow & Metabolism, vol. 32, pp. 213-231, 2012.

[2] K Türeyen, R. Vemuganti, K. A. Sailor and Robert J. Dempsey, " Infarct volume quantification in mouse focal cerebral ischemia: a comparison of triphenyltetrazolium chloride and cresyl violet staining techniques," Journal of NeuroScience Methods, vol. 139, pp. 203 – 207, 2004.

[3] Z. Yang, K. Richards, N. D. Kurniawan, S. Petrou, and D. C. Reutens, „MRI-guided volume reconstruction of mouse brain from histological sections" Journal of neuroscience methods, 211(2), 210–217. 2012.

[4] J. Nie, D. Shen, „Automated segmentation of mouse brain images using multi-atlas multi-ROI deformation and label fusion," Neuroinformatics. 2013;11(1):35-45.

[5] P. A. Yushkevich *et al.*, „3D Mouse Brain Reconstruction from Histology Using a Coarse-to-Fine Approach," WBIR 2006: Biomedical Image Registration, pp. 230-237, 2006.

[6] S. Shiffman , S. Basak, C. Kozlowski, R. N. Fuji, „An Automated Mapping Method for Nissl-stained Mouse Brain Histologic Sections", J. Neurosci Methods, vol 308, pp. 219-227, 2018.

[7] L. Rekik, S. Allassonnière, T. K. Carpenter, and J. M. Wardlaw, "Medical image analysis methods in MR/CT-imaged acute-subacute ischemic stroke lesion: Segmentation, prediction and insights into dynamic evolution simulation models. A critical appraisal," NeuroImage Clin. 1, 164–178. 2012.

[8] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," IEEE Trans. Pattern Anal. Machine Intelligence, 12 (1990), 161-192.

[9] U. Erdenetsogt, T. Wagner, "Shape Smoothing", October 6, 2016. Accessed on June 10, 2020. [Online]. Available: https://imagej.net/Shape_Smoothing

[10] D. Tschumperle and R. Deriche, "Vector-valued image regularization with PDEs: a common framework for different applications," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 4, pp. 506-517, April 2005.

[11] V. Pilny and J. Janacek, "Anisotropic Diffusion 2D", February 14, 2006. Accessed on June 10, 2020. [Online]. Available: https://imagej.nih.gov/ij/plugins/anisotropic-diffusion-2d.html

[12] G. Landini, "Fill Holes". Accessed on July 7, 2020. [Online]. Available: https://imagej.nih.gov/ij/docs/guide/146-References.html#biblio-34

[13] G. Landini, "Novel context-based segmentation algorithms for intelligent microscopy", accessed on July 7, 2020. [Online]. Available: https://blog.bham.ac.uk/intellimic/g-landini-software/

[14] J. Schindelin et al. (2012), "Fiji: an open-source platform for biological-image analysis", Nature methods 9(7): 676-682, 2012.

[15] C. T. Rueden, et al., "ImageJ2: ImageJ for the next generation of scientific image data", BMC Bioinformatics 18:529, 2017

[16] T. Ferreira, W. Rasband, "ImageJ User Guide". September/October 2012. Accessed on July 7, 2020. [Online]. Available: https://imagej.nih.gov/ij/docs/guide/index.html

Laser Speckle Stereo System in Biomedical Imaging

F. Salgueiro[1, 3], P. A. Assunção[1, 3], R. Fonseca-Pinto[1, 2, 3]
[1]Instituto de Telecomunicações, Multimedia Signal Processing Group, Leiria, Portugal
[2] ciTechCare – Center for Innovative Care and Health Technology – Polytechnic of Leiria, Portugal
[3] Polytechnic of Leiria, Leiria, Portugal
rui.pinto@ipleiria.pt

***Abstract* - Speckle is an optical interference phenomenon generated by the reflection of a highly coherent beam of light (e.g., laser) on a rough surface. This phenomenon has been applied in several research areas, aiming at the characterization of static surfaces or the extraction of properties in dynamic processes. In particular, laser speckle has shown huge potential in the identification of high varying perfusion regions in medical images of skin and brain. The acquisition setup in laser speckle is usually composed by a laser source, a magnifying lens, and one video camera. From the obtained sequence of video frames several image descriptors are calculated to characterize the surface under analysis, using information obtained from the scattered patterns. In recent years a new acquisition perspective in laser speckle was proposed by assembling an acquisition setup using two video cameras, thus producing the so called stereo laser speckle. This new approach is challenging regarding acquisition and imaging processing, introducing several variables to be controlled (distance from the camera lens to the surface, direction and width of the reflected beam of light). In this paper, a revision of recent results in stereo laser speckle is presented, highlighting the challenges in acquisition and image processing. Moreover, a new setup is proposed to test the influence of the identified critical variables in stereo laser speckle to be applied in medical imaging contexts.**

Keywords – Laser Speckle Imaging, Stereo Imaging, Image and Video Processing.

I. INTRODUCTION

Despite being a phenomenon that has been theoretically described since the 19th century, the applications of the laser speckle were only possible due to technological advances related to the acquisition systems, sensors and optical components [1], [2].

Speckle is an interference phenomenon resulting from the scattering effect of the wave fronts that are emitted from a rough reflecting surface, when illuminated by a coherent light source. When the coherent light source is the laser, the physical phenomenon is termed as laser speckle [3].

The scattering surface can either be static or presenting micromovements due to some dynamic process able to modify the texture of the reflection surface, which can be characterized by temporal descriptors. The removal of the speckle effect, regarded as noise, is a common process to improve the accuracy of the obtained information in research fields as synthetic aperture radars [4], biomedical imaging ultrasounds [5] and optical coherence imaging [6], in which speckle interference of the reflected wave fronts is filtered. However, the speckle signal can be used to obtain information about the pattern of the reflected front waves.

More recently, in particular regarding the applications in Biomedical Engineering context, perfusion studies related to cortical ischemia [7], blood flow speed [8] and the assessment of burn recovery [9], are pointing laser speckle imagery as a noninvasive, low cost and versatile imaging modality. Moreover, the integration of the dynamic laser speckle signal with the stereo acquisition is seen as a promising technique to advance for a new fusion imaging (by joining the function obtained by speckle and the texture of the surface using the 3D information obtained by stereo imaging) [10].

The possibility of obtaining fusion images by combining the speckle laser technique with stereo acquisition, poses some challenges regarding optimization in the acquisition variables and in hardware assembling [10], [11].

In this work a versatile stereo acquisition system for laser speckle is presented, designed to conduct optimization experiments regarding the central variables for stereo laser speckle imaging (distance between cameras, pitch, roll and yaw in each camera, distance to the Region of Interest).

II. METHODS

The implemented acquisition block system it is represented in Fig.1. This system is composed by two GigE ac1920-25gm cameras, one Gigabit-Ethernet (GS110TP PoE) switch, a hardware trigger generator and a host computer. The stereo acquisition system is divided in two sectors: the trigger generator and the two acquisition cameras.



Figure 1. Setup acquisition.

A. Camera Features

In Table I the camera features are presented. The exposure time is a configuration parameter of central importance in speckle, not only because it must be adjusted to match the short duration of the visible elements of speckle, but also due to its interdependence with the acquisition frame rate. Another central parameter directly related with acquisition frame rate is the electronic shutter operation. In the cameras used in this work only the rolling shutter mode is available, which includes two sub-types (the rolling shutter mode and global reset release shutter mode). These are different reading modes of the camera sensors, which provide a further level of flexibility in the setup parametrization.

TABLE I. CAMERA PARAMETERS (ADAPTED FROM [12])

Parameters	acA1920-25gm
Maximum possible resolution	1920 x 1080
Color	Mono
Possible Synchronization Modes	Via Hardware trigger Via Software trigger Via Free run
Exposure time control	Via hardware trigger Programmable via the camera API
Camera Power Requirements	Via Ethernet connector - Power over Ethernet (PoE) 802.3af Via I/O connector - 12 VDC
I/O Lines	1 optocoupled input line 1 optocoupled output line

B. System Configuration

To achieve a frame rate in the order of 50 frames per second (FPS) (to accomplish the same frame rate as the reference papers for this work and our previous experience in speckle acquisition using one camera) a region-of-interest (ROI) of 960 x 540 resolution was defined to be used in the experimental setup during acquisition. As it can be observed in Table 1, this is not the maximum resolution of these cameras, but this reduction in resolution allows to obtain the maximum frame rate available for a 960 x 540 ROI, thus setting the frame rate close to 50 FPS.

In order to guarantee stereo synchronization of the left and right images, a hardware generated trigger is used to define the instant of acquisition of a frame. In other words, a frame start trigger is generated with a hardware source. The frequency of such signal corresponds to the acquisition frame rate. The actual acquisition of the visual data begins after a predefined delay established by the trigger signal rise, as represented in Fig. 2. This delay is recommended when a hardware trigger is used for image acquisition. As the defined electronic shutter operation mode is the rolling shutter mode, the minimum exposure time (MET) can be calculated using (1).

$$MET > tRow * (AOI_Height - 1) \quad (1)$$

where:

AOI_Height = height of the ROI defined in the frame

tRow = temporal offset from one line to the next

This minimum exposure time value determines when the sensor reading will be taken on all lines at the same time. This value is important to determine the flash window width, that it is represented by the difference between the defined exposure time and the minimum exposure time calculated by (1). The use of a flash window width is necessary to generate a signal to activate a flash light for better illumination of moving objects, whenever necessary. Since the laser speckle is a phenomenon that occurs with a high variation in changing patterns, the electronic shutter operation modes and the exposure time must be tightly controlled, as they can influence the pattern of the laser speckle captured by the cameras. As a security measure, the exposure mode has been set to discard possible triggers while an exposure is in progress.

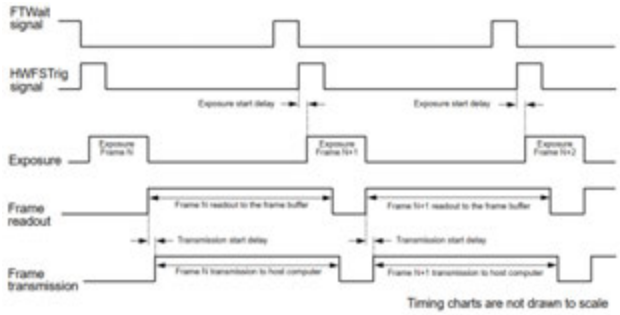


Figure 2. Timing chart of an Exposure Start Controlled with a HWFSTrig Signal, from [12].

Since the objective is to obtain a synchronous stereo acquisition, the process of exposing and reading the image sensors must be performed at exactly the same the frame rate and acquisition period. As it can be observed in Fig. 2, the sequence of processes was defined so that the sum of the total start delay (TSD), defined in (2), with the frame readout time, must belong to the HWFSTrig period. The frame readout time value can be obtained by reading the *Readout Time Abs* parameter.

$$TSD = ESD + ILRT + DS + FSTD \quad (2)$$

Where:

TSD = Total Start Delay

ESD = Exposure Start Delay

ILRT = Input Line Response time

DS = Debouncer Setting

FSTD = Frame Start Trigger Delay

The *PacketTimeout* parameter defines how long the filter driver waits for the next expected packet before initiating a resend request [12]. Other important parameter to consider is the *FrameRetention,* which defines the timeout frame retention timer (in milliseconds). The frame retention timer starts when a leader for a frame is detected. After all frame packets are received the timer resets and the timer will timeout. In case timeout occurs before the

last packet is received, the buffer for the frame is released and an indication of unsuccessful grab is provided [12].

Regarding the hardware trigger, the frame starts trigger is defined by configuring the time when the signal is *high* by setting an Arduino Mega timer in PWM mode, where it is configured for the frequency of acquisition.

C. *Acquisition Process*

The flowcharts of the implemented acquisition system or the synchronous stereo images is illustrated in Fig. 3. The acquisition methodology is initiated by the above mentioned process of setting the parameters. After this first step, the command for the initialization of the acquisition process during a predefined period is executed. Afterwards, the signal generation begins by sending the order to grab the frame. The control of the acquisition period is carried out through a timer, that is started right after the hardware trigger start order. To guarantee the frame synchronization between the two cameras, a validation function checks each "successful acquisition" to confirm whether the acquisition of camera 1 occurred at the same time of the camera 2. After this verification the stereo image data is saved.

As soon as the acquisition process is finished, the conversion of the acquired frames begins, so that this information can be stored in a *.tiff* file.

The synchronous stereo acquisition process is represented in Fig. 3.

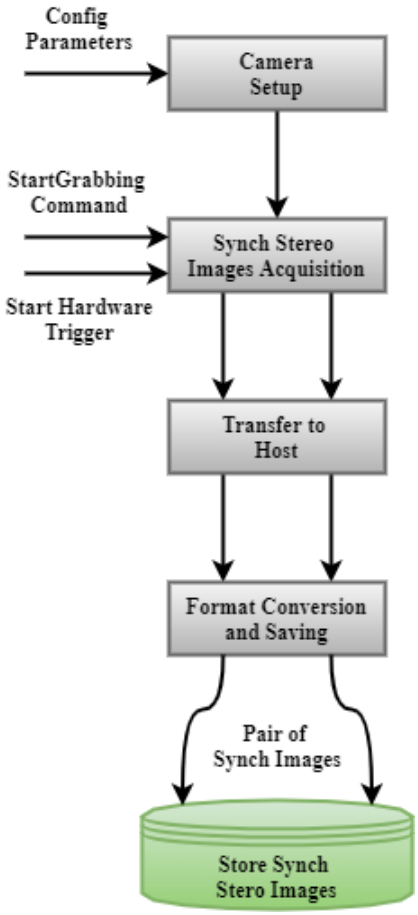


Figure 3. Flow diagram of acquisition process.

III. RESULTS

The main result in the scope of this work is the system itself, from whom the generic configuration is presented in Fig. 1. The results intend to validate the acquisition setup and to confirm that it is fully operational to capture a laser speckle phenomenon.

The setup was developed in order to test distinct acquisition parameters (e.g. distance between cameras, pitch, roll and yaw in each camera, distance to the ROI) and to optimize stereo laser speckle acquisitions.

TABLE II. ACQUIISIITION TIMES

#Frame	Left camera - Acquisition time [s]	Right camera - Acquisition time [s]
1	0,0000	0,0000
100	3,2987	3,2987
200	6,6307	6,6307
300	9,9627	9,9627
400	13,2947	13,2947
500	16,6267	16,6267
600	19,9587	19,9587
700	23,2907	23,2907
800	26,6227	26,6227
900	29,9547	29,9547

In Table II one can observe the acquisition times of the left and right cameras during one speckle acquisition experiment.

In this experiment, one test of isquemia-perfusion in the arm (as described in [3]) was conducted in order to test the system by using an experimental condition similar to the ones we intend to apply in next works with stereo speckle for skin imaging.

As observed from the results shown in Table II, the systems allows perfect synchronization between cameras, which is crucial in the process of stereo laser speckle process. From these results, taking into account the number of frames and time interval, it can be confirmed that the frame period is constant and equal to 33ms.

As a future work, several experimental conditions during speckle experiments will be tested to define the optimal prototype of the system, considering the level of characterization and discrimination of the speckle images, as a quality factor.

IV. CONLUSION

The speckle laser imaging technique has been shown to be very versatile, and has great advantages in the area of medical imaging since it applies non-ionizing radiation and is non-invasive.

More recently, the use of two video cameras to carry out stereo acquisition has shown to be very promising, but at the same time it raises many technical questions, for example related to the timing of the acquisitions, and the definition of parameters of the stereo acquisition setup that ideally has to be designed for the specific context type and optimized under cost functions that include the output of the image processing algorithms and quality of relevant features.

This work presents an acquisition setup for stereo laser speckle, and the synchronization of video acquisitions for the proposed setup is tested. The presented results will allow to use the setup to test other variables whose variation may influence the image processing, and, ultimately, the consistency of the obtained results.

ACKNOWLEDGMENT

This work is funded by Portuguese national funds provided by Fundação para a Ciência e Tecnologia (FCT) through national funds and when applicable co-funded EU funds under the projects UIDB/EEA/50008/2020, UIDP/EEA/50008/2020 and UIDB/05704/2020 and UIDP/05704/2020.

REFERENCES

[1] L. Bento, L. Tavera, P. Assuncao, S. Faria, and R. Fonseca-Pinto, "Evaluation of cutaneous microcirculation patterns by laser speckle imaging," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings*, 2018, pp. 290–293, doi: 10.23919/MIPRO.2018.8400055.

[2] D. R. Tomassi, L. Aronson, C. E. Martínez, D. H. Milone, M. E. Torres, and H. L. Rufiner, "Evaluación de técnicas clásicas de reducción de ruido en señales de voz," *Rev. la Soc. Argentina Bioingenierìa*, vol. 11, no. 2, pp. 19–28, Mar. 2005.

[3] L. Bento, L. Távora, P. A. A. Assunção, S. Faria, and R. Fonseca-Pinto, "Dynamic laser speckle in medical imaging : on the quantification of skin patterns," pp. 36–36, 2017.

[4] M. Kanevsky, *Radar Imaging of the Ocean Waves*. Elsevier, 2009.

[5] M. Forouzanfar, H. Abrishami Moghaddam, and M. Gity, "A new multiscale Bayesian algorithm for speckle reduction in medical ultrasound images," *Signal, Image Video Process.*, vol. 4, no. 3, pp. 359–375, Aug. 2010, doi: 10.1007/s11760-009-0126-3.

[6] H. Lv, S. Fu, C. Zhang, and L. Zhai, "Speckle noise reduction for optical coherence tomography based on adaptive 2D dictionary," *Laser Phys. Lett.*, vol. 15, no. 5, p. 055401, Apr. 2018, doi: 10.1088/1612-202X/AAAEB0.

[7] A. B. Parthasarathy, S. M. S. Kazmi, and A. K. Dunn, "Quantitative imaging of ischemic stroke through thinned skull in mice with Multi Exposure Speckle Imaging," *Biomed. Opt. Express*, vol. 1, no. 1, p. 246, Aug. 2010, doi: 10.1364/boe.1.000246.

[8] Z. Luo, Z. Yuan, M. Tully, Y. Pan, and C. Du, "Quantification of cocaine-induced cortical blood flow changes using laser speckle contrast imaging and Doppler optical coherence tomography," *Appl. Opt.*, vol. 48, no. 10, pp. D247–D255, Apr. 2009, doi: 10.1364/AO.48.00D247.

[9] C. J. Stewart, R. Frank, K. R. Forrester, J. Tulip, R. Lindsay, and R. C. Bray, "A comparison of two laser-based methods for determination of burn scar perfusion: Laser Doppler versus laser speckle imaging," *Burns*, vol. 31, no. 6, pp. 744–752, Sep. 2005, doi: 10.1016/j.burns.2005.04.004.

[10] F. Cunha, L. Távora, P. Assunção, S. Faria, and R. FonsecaPinto, "Stereo laser speckle dissimilarity analysis using self-organizing maps," in *IFMBE Proceedings*, 2020, vol. 73, pp. 51–60, doi: 10.1007/978-3-030-17971-7_8.

[11] R. Fonseca-Pinto and M. Machado, "A textured scale-based approach to melanocytic skin lesions in dermoscopy," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2017 - Proceedings*, 2017, pp. 279–282, doi: 10.23919/MIPRO.2017.7973434.

[12] Basler AG, "Basler ace USER'S MANUAL FOR GigE CAMERAS," 2016.

Lung cancer typology classification based on biochemical markers using machine learning techniques

Jasna Nuhić
Department of Electrical and Electronics Engineering
International Burch University
Sarajevo, Bosnia and Herzegovina
jasna.nuhic@stu.ibu.edu.ba

Jasmin Kevrić
Department of Electrical and Electronics Engineering
International Burch University
Sarajevo, Bosnia and Herzegovina
jasmin.kevric@ibu.edu.ba

***Abstract*—Clinical mistreatment and mismanagement are big issues caused by detection of too many false negative patients. Therefore, lung cancer diagnostic inaccuracy and methods to surpass it in a minimally invasive way is often the subject of research, as it is case of this study. This study focuses on the use of machine learning algorithms as a noninvasive tool to differentiate malignant pleural effusions from benign effusions. It provides performance comparisons between Adaptive neuro-fuzzy inference system (ANFIS), Support vector machine (SVM), RUS Boosted Tree (RUSBoost) and K-Nearest-Neighbor (K-NN) techniques for lung cancer detection. The proposed algorithms were chosen based on the current state of the art in the field of pulmonary diagnostics. The novelty of this work is the application of machine learning models for classification of lung cancer based on expression of tumor markers obtained from serum and pleural fluids. The performance of all models is compared and validated on data samples of 168 patients. Three classification model, SVM, RUSBoost and K-NN performed equally well, whereas underperforming model was ANFIS.**

***Index Terms*—machine learning, fuzzy logic, pleural effusion, biochemical tumor marker, lung cancer detection**

I. Introduction

Lung cancer's first presenting sign, responsible for approximately 30% of malignant cases, is the presence of malignant pleural effusions (MPEs) [1]. MPEs are metastatic manifestations of malignant tumors so tumor marker assessment is consequently possible in case of their presence.

Pleural effusions can be benign or malignant and their differentiation is of the utmost importance due to the link between the presence of malignancy and the median patient survival period of 4 months [2]. A common method for differentiation among them is pleural fluid cytology, with a sensitivity of up to 70%. This relatively low sensitivity is dependent on different factors, such as amount of fluid, the cytologist's experience level, and the tumor load [3]. Therefore, pleural fluid cytology should be methodologically complemented in order to avoid diagnostic inaccuracy and consequent clinical mistreatment and mismanagement of patients.

Considered as a valid and less invasive method for MPE identification is analysis of tumor markers. More accurate diagnosis can be achieved by analyzing tumor markers in serum and pleural fluid. Use of tumor marker analysis reduces the number of false negative results and therefore helps increase diagnostic sensitivity [4].

Even though minimally invasive diagnostic procedures with high specificity levels do exist, namely thoracoscopic and video-assisted pleural biopsy, they are subject to limitations in terms of patient intolerability and unavailability in many hospitals [5]. Therefore, there is an evident need to develop novel noninvasive methods in pleural diffusions' differentiation.

In this study, a group of lung cancer patients affected by cytologically negative effusions with a control group of patients with non-malignant disease will be compared. The comparison will actually test the ability of CEA and CYFRA21-1, identified as optimal set of biochemical tumor markers in [4], to differentiate benign from malignant effusions. A machine learning (ML) approach will be used and different classification models will be developed to achieve accurate differentiation.

II. Previous Research in Lung Cancer Detection

Several researchers have proposed and implemented detection of lung cancer using different machine learning approaches [6]–[15]. This section presents an overview of articles that implemented machine learning techniques for lung cancer detection. Twenty papers were analyzed for this purpose and a brief overview of ten of them is given in the rest of this section.

Sun W. et al tested the feasibility of using deep learning algorithms for lung cancer diagnosis with the cases from the Lung Image Database Consortium database. Implemented deep learning algorithms that were Convolutional Neural Network (CNN), Deep Belief Networks (DBNs) and Stacked Denoising Autoencoder (SDAE). Performance of deep learning algorithms was compared with traditional computer aided diagnosis (CADx) system. The accuracies of CNN, DBNs, and SDAE are 79.76%, 81.19%, and 79.29%, respectively; the accuracy of the CADx is 79.40%, which is slightly lower than CNN and DBNs [6].

In [7], the aim of the study was the development of an efficient lung nodule detection scheme by performing nodule segmentation through fuzzy based clustering models and classification using a Support Vector Machine (SVM) classifier. Three different types of kernels were tested, where the RBF kernel outperforms others with the obtained accuracy of 80.36%.

An early lung cancer detection methodology using nucleus based features was proposed by K. Kancherla et al. in [8]. The performance of well know machine learning techniques in the context of lung cancer detection on the Biomoda dataset was studied. A maximum accuracy of 81.00% using 71 features related to shape, intensity and color was achieved, while addition of the nucleus segmented features improved the accuracy to 87.00% [8].

The usage of SVM classification for lung cancer was investigated in [9]. Its performance was compared with Boosting, Decision trees, k-nearest neighbor, LASSO regressions, neural networks and random forests trained on a large database developed from 259 patients. SVM, having an accuracy of 94.00%, significantly increased classification performance compared to other algorithms [9].

H. Wang et al. compared the deep learning method with four classical machine learning methods to classify mediastinal lymph node metastasis of lung cancer from 18F-FDG PET/CT images [10]. The PET/CT images were studied from 168 patients and from them 1397 lymph nodes were collected. There was no significant difference between the results of convolutional neural network (CNN) and the best classical method, all of them achieving accuracy above 80.00% [10].

An automatic system, based on genetic algorithm (GA) and extreme learning machines (ELM), was proposed by M. Daliri in [11], for the diagnosis of lung cancer. A genetic algorithm was used for feature selection, which were then fed into a fuzzy inference system (FIS). The obtained results indicate that the trained system could be useful in lung cancer diagnosis and could be used in clinical practice [11].

The entropy degradation method (EDM) was proposed as a novel neural network-based algorithm for detection of lung cancer and its early stage prediction. The data used were twelve high-resolution computational tomography (CT) scans obtained from the National Cancer Institute. The trained algorithm finally reached an accuracy of 77.8%, with an evident limitation of low amount of data [12].

Another system for detection of cancerous nodules was proposed by S. Makaju et al. [13]. The scan images were used to classify nodules as benign or malignant. First, they were watershed segmented and then nodules were classified. The SVM classifier was used and obtained an accuracy of 86.60% with a possibility to further increase accuracy by applying data preprocessing techniques and eliminating false objects. Additionally, the possibility of predicting cancer stage using same methodology was observed [13].

Sangamithraa et al., used the K mean unsupervised learning algorithm for clustering the pixel dataset according to certain features. The back propagation network was employed for classification purposes. Features, including entropy, correlation, homogeneity, PSNR, SSIM were extracted using gray-level co-occurrence matrix (GLCM) method [14].

In [15], the lung cancer detection system was developed using watershed segmentation. The accuracy of the system was compared using neural fuzzy model and region growing method. Accuracy the system obtained was 90.10%, which was higher than the other two methods [15].

The Table I below shows the various lung cancer detection techniques and the detection rate for given technique. While most of them exclusively focused on CT image scan analysis (nine out of ten papers analyzed), only few of them managed to obtain detection accuracy that is higher than 90.00%. The highest detection accuracy (98.85%) was obtained in the study that used combined clinical data and X-ray data which implies that relying on image analysis only is not sufficient. Therefore, this paper will focus on using of biochemical tumor markers as a potential tool to obtain highly accurate lung cancer detection using different machine learning approaches.

III. Methods

This study compared supervised learning classification techniques, where two of them were classical machine learning classifying techniques, one ensemble classification technique, one FUZZY technique, and one deep learning technique. The techniques were implemented in MATLAB R2019a. Before classification deployment, the data was standardized. In order to obtain an accurate estimate of the error, validation of classification techniques using 5-fold cross-validation was utilized and optimal parameters were determined.

TABLE I
LUNG CANCER DETECTION TECHNIQUES

Study	Technique performance		
	Technique	Number of *subjects*	*Accuracy (%)*
[6][a]	CNN	1018	79.76
	DBNs		81.19
	SDAE		79.29
	CADx		79.40
[7][a]	SVM linear	1010	71.43
	SVM polynomial		66.07
	SVM RBF		80.36
[8][a]	SVM	28	79.43
	RBF Network		80.48
	Naïve Bayes		13.80
	Multinomial Logistic Model		70.74
	Sequential Minimal Optimization		75.00
	Ada-Boost RBF Network		83.10
	Bagging RBF Network		87.80
	Multilayer Perceptron		70.80
	Random Forest		80.48
	Linear Logistic Model		78.04
	Ada-Boost Linear Logistic Model		75.60
	Bagging Linear Logistic Model		85.36
[9][a]	SVM	259	94.00
	Boosting		86.00
	Decision trees		82.00
	k-nearest neighbor		75.00
	LASSO		93.00
	Neural networks		89.00
	Random Forests		86.00
[10][a]	Random forest	168	85.08
	SVM		82.73
	AdaBoost		85.05
	Back-propagation ANN		80.51
	CNN		85.64
[11]	GA + FIS	32	98.85
[12][a]	EDM	12	77.80
[13][a]	SVM	1018	86.60
[14][a]	Back propagation network	-	90.87
[15][a]	Neural Fuzzy Model	-	89.30
	Region Growing Method		88.50
	CT scan image processing		90.10

[a]Image analysis was used.

Performance evaluation was based on values of sensitivity, specificity, detection accuracy, and the area under the receiver operating characteristic (ROC) curve (AUC). ROC is a probability curve and AUC is a measure of separability which evaluates how much a model is capable of distinguishing between classes. The higher the AUC, the better the model is at distinguishing between subjects with disease and without disease. Sensitivity, also known as a true positive (TP) rate, is a measure of the proportion of positives that are correctly identified. Specificity, also known as a true negative (TN) rate, is a measure of the proportion of negatives that are correctly identified. The false positive (FP) rate is measure of the proportion of negatives that are incorrectly identified as positives and the false negative (FN) rate is ameasure of the proportion of positives that are incorrectly identified as negatives. Detection accuracy is ability to differentiate among cases correctly and is mathematically given by the ratio of true negative and true positive rates of all evaluated cases. Comparison between classification techniques was based on these four measures.

TABLE II
DATASET DESCRIPTION

Patient Data			
Selected patients	*All patients*	*Malignant pleural effusions*	*Benign pleural effusions*
Total number	168	124	44
Female	70	44	26
Male	98	80	18
Age range (years)	25-89	37-80	25-89

A. Data set

This study aims to detect and differentiate between types of effusions in patients pathologically diagnosed either with lung cancer or some type of inflammatory pulmonary disease. In total 168 patients were included in the study. Fluid and serum samples were collected from patients under "Tumor marker in pleural effusion and biopsies" study approved by ethical committee of and carried out at Tor Vegeta University Hospital in Rome, Italy [4].

Data set input features consist of antigen (CEA) and Cytokeratin fragment 21-1 (CYFRA 21.1) diagnostic values taken from both serum and pleural fluids assessed using chemiluminescence and immuno-fluorescence methods, respectively [4].

As described in Table II, samples were collected from 98 male and 70 female patients with median age of 72 for male and 67 for female patients. Out of 124 patients with MPE 64.5% were male and 35.5% female, while 40.9% of 44 patients with benign effusions were male and 59.1% were female.

About 75% (124) of patients were diagnosed with some type of lung cancer. Only six lung cancer patients were diagnosed with small-cell lung cancer (SCLC), while the remaining 118 patients were diagnosed with non-small-cell lung cancer (NSCLC), of which 62 were diagnosed with adenocarcinoma and 56 with squamous-cell carcinoma. The control group consisted of 44 patients diagnosed with non-malignant respiratory disease, where 23 of them were suffering from pneumonia, 18 from congestive heart failure, 2 from tuberculosis and 1 from collagen disease. More male patients were diagnosed with malignant tumor, while more female patients are diagnosed with benign pleural effusions.

B. Selected Machine Learning Techniques

a) Adaptive-Network-based Fuzzy Inference System: ANFIS, as a computational model, has been deployed in medicine for many purposes where brain, breast, lung and skin cancer detection are some of them [16]–[19].

ANFIS serves as a basis for constructing a set of fuzzy if-then rules with appropriate membership functions to generate the stipulated input-output pairs in order to minimize the output error measure or maximize performance index. It is a hybrid learning system based on neural networks and fuzzy inference system that identifies premise and consequent membership function parameters of single output. The network configuration should be of a feed-forward type [20].

Architectural system of ANFIS consists of five layers: fuzzification process, Fuzzy and execution, normalizing membership functions, execution of consequent part of the fuzzy rules and output computation [20].

b) K-Nearest Neighbors: K-NN, as a well performing classifier is often used for cancer detection purposes and has been utilized and proven to be effective in detection of not only lung cancer, but in breast and prostate cancer detection as well [21]–[26]. K-NN is instance based or, in other words, lazy type of learning. The working principle is based on distance function and class of element's k nearest neighbors, where k represents the number of neighbors selected to assign the class to the new element [21].

Although distance may be defined in many ways, in this study the most universal and generally accepted distance, the Euclidean distance, is used [21]. Euclidean distance or 2-distance between two vectors x_{ir} and x_{jr} is defined by the following formula:

$$d(x_{\mathrm{i}}, x_{\mathrm{j}}) = \sqrt{\sum_{r=1}^{n} (x_{\mathrm{ir}} - x_{\mathrm{jr}})^2} \tag{1}$$

c) Random Under-sampling Boosting: RUSBoost has been utilized in different medical diagnostic classification problems and has been proven to be an outperforming, working solution for neurodegenerative diseases detection [27]. It is an ensemble classifier used to handle class imbalance, meaning that number of subjects corresponding to one class is much smaller than the other class which is the case with the data used in this study. For that purpose, the RUSBoost algorithm combined the random under-sampling technique with the standard boosting procedure [27].

RUSBoost working principle is based on the application of multiple classifiers and adjusting samples' weights in each iteration. Learning is of the decision trees type, where the iterative run of distributed training data is applied on weak learners. Decision trees proved to be effective in cancer detection [28]. Furthermore, they are combined into one classifier with higher detection accuracy than accuracy of an individual tree [29].

d) Support Vector Machine: SVM, an algorithm that was first developed in 1963, is recognized as an outperforming classification technique for cancer detection [30]–[34]. The working. principle of the SVM classifier is based on construction of a hyperplane between classes. SVM can be linear or non-linear. Linear SVM, which is deployed in this study, is simpler one and is given by the equation below.

$$f(x) = w^{\mathrm{T}}x + b \tag{2}$$

Equation (1) means that classes are being separated by the hyperplane $f(x) = w^{T}x + b = 0$, where w represents weight vector normal to hyperplane, b represents bias and x represents the data point(s). There may be many hyperplanes that divide the classes, but the SVM is based on the hyperplane that maximizes the separating margins of the classes. Boundary hyperplanes contain elements of the data-set which are called support vectors [30].

Machine learning algorithms were implemented using Matlab® R2019a software and the Classification Learner application of the corresponding MATLAB edition.

IV. Results and Discussion

The data was standardized prior to model training. Prediction models were adjusted to the application and their parameters optimized for the purpose of lung cancer detection. Five nearest neighbors with the Euclidean distance were selected for K-NN model training. RUSBoosted Trees algorithm was trained with 20 as a maximum number of splits and 30 learners. Linear SVM was trained with the automatic Kernal scale and the One-vs-One multiclass method utilized.

Table III reports the performance values of the machine learning techniques used for lung cancer detection. All the performance values are listed as results of 5-fold cross validation procedure. The models cross validated to ensure that problems like overfitting and selection bias are avoided. All the performance values are listed as accuracy, AUC, sensitivity and specificity values. Prediction models were successfully established using the four features as input data.

It can be seen that SVM is the technique with the highest accuracy, AUC and specificity. In terms of sensitivity, both K-NN and RUSBoost with equal sensitivity are better than SVM. ANFIS is the technique of the lowest AUC and it is worse than any other technique in terms of detection accuracy. Other than ANFIS, the other three classification techniques perform well, not significantly different among each other, with an AUC greater than or equal to 0.95.

There is no agreement on which model has a better performance in the analysis of MPEs in serum and pleural fluid and different researchers choose different classifiers for their studies. In this study, it was observed that the SVM classifier insignificantly outperforms other classifiers, suggesting that SVMs may be a more appropriate prediction model in the analysis of MPEs for lung cancer detection

TABLE III
CLASSIFICATION RESULTS

Technique	Accuracy (%)	TP (%)	TN (%)	AUC
ANFIS	86.83	89.50	79.50	0.90
K-NN	94.00	95.00	93.00	0.97
RUSBoost	92.30	95.00	84.00	0.95
SVM	94.60	94.00	95.00	0.98

TABLE IV
SVM PERFORMANCE EVALUATION

Observation	SVM (nr)	SVM (%)
TP	117	94.00
TN	42	95.00
FP	2	6.00
FN	7	5.00
Accuracy (%)	94.60	

purposes. It can be observed and expected that K-NN and RUSBoost classifiers would perform equally well when deployed for the same purposes.

In Table IV, the results of the best performing classifier are presented. As can be seen from Table III, the best performing classifier is SVM, which obtained 94.60% detection accuracy. SVM algorithm makes only two false positive predictions, meaning than two patients are labeled to have MPE by mistake. Similarly, SVM misses seven cases when the patients actually have MPE. This shows that there is a slight room for improvement, which might be obtained by integrating SVM with an ensemble classifier to obtain better detection accuracy.

V. CONCLUSION

In this study, the application of machine learning techniques for lung cancer detection by biochemical marker classification is explored. The experimental results yield promising outcomes for some of tested classifiers.

Detection accuracy values obtained in this study exceed those in previous research studies which exclusively focused on X-ray image classification for lung cancer detection. This phenomena shows great potential use of pleural effusions as complementary diagnostic tool.

The future scope of this project would be to test models trained on new data and validate it for clinical use. Furthermore, investigation of classification performance using combination of features obtained from X-ray images and pleural effusions should be considered for potential detection accuracy enhancement.

REFERENCES

[1] F. Malcolm M. DeCamp, Jr., AID, FCCP; Ster;en j. Mentzer, MD, FCCP; Scott j. S[C(I/lS01" MD; and David j. Sngarbaker, MD, "Malignant Effusive Disease of the Pleura and Pericardium," *Chest*, vol. 112, no. 4 SUPPL., 1997.

[2] D. T. Arnold, D. D. Fonseka, S. Perry, A. Morley, J. E. Harvey, A. Medford, M. Brett, and N. A. Maskell, "Investigating unilateral pleural effusions: The role of cytology," *European Respiratory Journal*, vol. 52, no. 5, 2018.

[3] K. B. Sriram, V. Relan, B. E. Clarke, E. E. Duhig, I. A. Yang, R. V. Bowman, Y. C. Lee, and K. M. Fong, "Diagnostic molecular biomarkers for malignant pleural effusions," *Future Oncology*, vol. 7, no. 6, pp. 737–752, 2011.

[4] S. Elia, G. D'Angelo, F. Palmieri, R. Sorge, R. Massoud, C. Cortese, G. Hardavella, and A. De Stefano, "A machine learning evolutionary algorithm-based formula to assess tumor markers and predict lung cancer in cytologically negative pleural effusions," *Soft Computing*, no. September, 2019.

[5] A. C. Bibby and N. A. Maskell, "Pleural biopsies in undiagnosed pleural effusions; Abrams vs image-guided vs thoracoscopic biopsies," *Current Opinion in Pulmonary Medicine*, vol. 22, no. 4, pp. 392–398, 2016.

[6] W. Sun, B. Zheng, and W. Qian, "Computer aided lung cancer diagnosis with deep learning algorithms," *Medical Imaging 2016: Computer-Aided Diagnosis*, vol. 9785, p. 97850Z, 2016.

[7] S. Sivakumar and C. Chandrasekar, "Lung nodule detection using fuzzy clustering and support vector machines," *International Journal of Engineering and Technology*, vol. 5, no. 1, pp. 179–185, 2013.

[8] K. Kancherla and S. Mukkamala, "Early lung cancer detection using nucleus segementation based features," *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2013 - 2013 IEEE Symposium Series on Computational Intelligence, SSCI 2013*, pp. 91–95, 2013.

[9] T. Sun, J. Wang, X. Li, P. Lv, F. Liu, Y. Luo, Q. Gao, H. Zhu, and X. Guo, "Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 2, pp. 519–524, 2013.

[10] H. Wang, Z. Zhou, Y. Li, Z. Chen, P. Lu, W. Wang, W. Liu, and L. Yu, "Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images," *EJNMMI Research*, vol. 7, no. 1, 2017.

[11] M. R. Daliri, "A hybrid automatic system for the diagnosis of lung cancer based on genetic algorithm and fuzzy extreme learning machines," *Journal of Medical Systems*, vol. 36, no. 2, pp. 1001–1005, 2012.

[12] Q. Wu and W. Zhao, "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm," *Proceedings - 2017 International Symposium on Computer Science and Intelligent Controls, ISCSIC 2017*, vol. 2018-Febru, pp. 88–91, 2018.

[13] S. Makaju, P. W. Prasad, A. Alsadoon, A. K. Singh, and A. Elchouemi, "Lung Cancer Detection using CT Scan Images," *Procedia Computer Science*, vol. 125, no. 2009, pp. 107–114, 2018.

[14] P. B. Sangamithraa and S. Govindaraju, "Lung tumour detection and classification using EK-Mean clustering," *Proceedings of the 2016 IEEE International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2016*, pp. 2201–2206, 2016.

[15] S. Ignatious and R. Joseph, "Computer aided lung cancer detection system," *Global Conference on Communication Technologies, GCCT 2015*, no. Gcct, pp. 555–558, 2015.

[16] R. J. Deshmukh and R. S. Khule, "Brain Tumor Detection Using Artificial Neural Network Fuzzy Inference System (ANFIS)," *International Journal of Computer Applications Technology and Research*, vol. 3, no. 3, pp. 150–154, 2014.

[17] J. Kuruvilla and G. K., "Lung cancer classification using fuzzy logic for CT images," *International Journal of Medical Engineering and Informatics*, vol. 7, no. 3, pp. 233–249, 2015.

[18] M.-l. Huang, Y.-h. Hung, W.-m. Lee, R. K. Li, and T.-h. Wang, "Usage of Case-Based Reasoning , Neural Network and Adaptive Neuro-Fuzzy Inference System Classification Techniques in Breast Cancer Dataset Classification Diagnosis," *Journal of Medical Systems*, vol. 36, no. 2, pp. 407–414, 2012.

[19] S. M. Odeh, "Using an Adaptive Neuro-Fuzzy Inference System (AnFis) Algorithm for Automatic Diagnosis of Skin Cancer," *Jurnal of Communication and Cmputer*, vol. 8, pp. 751–755, 2011.

[20] J. S. R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference System," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 23, no. 3, pp. 665–685, 1993.

[21] S. AhmedMedjahed, T. Ait Saadi, and A. Benyettou, "Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and

Classification Rules," *International Journal of Computer Applications*, vol. 62, no. 1, pp. 1–5, 2013.
[22] A. Lorenz, M. Bluem, H. Ermert, and T. Senge, "Comparison of different neuro-fuzzy classification systems for the detection of prostate cancer in ultrasonic images," *Proceedings of the IEEE Ultrasonics Symposium*, vol. 2, pp. 1201–1204, 1997.
[23] M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm," *Proceedings - 2016 9th International Conference on Developments in eSystems Engineering, DeSE 2016*, pp. 35–39, 2017.
[24] P. Giri and K. Saravanakumar, "Oriental Journal of Breast Cancer Detection using Image Processing Techniques," *Oriental Journal of Computer Science & Technology*, vol. 10, no. 2, pp. 391–399, 2017.
[25] J. Nuhić and J. Kevrić, "Prostate Cancer Detection Using Different Classification Techniques," in *International Conference on Medical and Biological Engineering*, vol. 73. Springer, 2019, pp. 67–73.
[26] J. Nuhić, L. Spahić, S. Ćordić, and J. Kevrić, "Comparative Study on Different Classification Techniques for Ovarian Cancer Detection," in *IFMBE Proceedings - CMBEBIH*, vol. 73, 2019, pp. 159–163.
[27] T. Tong, C. Ledig, R. Guerrero, A. Schuh, J. Koikkalainen, A. Tolonen, H. Rhodius, F. Barkhof, B. Tijms, A. W. Lemstra, H. Soininen, A. M. Remes, G. Waldemar, S. Hasselbalch, P. Mecocci, M. Baroni, J. Lötjönen, W. van der Flier, and D. Rueckert, "Five-class differential diagnostics of neurodegenerative diseases using random undersampling boosting," *NeuroImage: Clinical*, vol. 15, no. March, pp. 613–624, 2017.
[28] A. Osmanović, L. Abdel-ilah, A. Hodžić, J. Kevrić, and A. Fojnica, "Ovary Cancer Detection using Decision Tree Classifiers based on Historical Data of Ovary Cancer Patients," in *Cmbebih 2017*, vol. 62, 2017, pp. 503–510.
[29] V. Murugananthan and U. K. Durairaj, "RUS boost tree ensemble classifiers for occupancy detection," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 272–277, 2019.
[30] K. Machhale, H. B. Nandpuru, V. Kapur, and L. Kosta, "MRI brain cancer classification using hybrid classifier (SVM-KNN)," *2015 International Conference on Industrial Instrumentation and Control, ICIC 2015*, no. Icic, pp. 60–65, 2015.
[31] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," *Digital Signal Processing: A Review Journal*, vol. 17, no. 4, pp. 694–701, 2007.
[32] S. K. Majumder, N. Ghosh, and P. K. Gupta, "Support vector machine for optical diagnosis of cancer," *Journal of Biomedical Optics*, vol. 10, no. 2, p. 024034, 2005.
[33] Y. I. A. Rejani and S. T. Selvi, "Early Detection of Breast Cancer using SVM Classifier Technique," vol. 1, no. 3, pp. 127–130, 2009.
[34] D. Kečo, A. Subasi, and J. Kevric, "Cloud computing-based parallel genetic algorithm for gene selection in cancer classification," *Neural Computing and Applications*, vol. 30, no. 5, pp. 1601–1610, 2018.

Self-help for Elderly Individuals with Morning gymnastics “1000 movements” and “method 23+5”

B. Širaiy*, U.J. Stanič*, Z. Katkič* and N. Grishin*
* School of Health Society, Slamnikarska cesta 18, Domžale, Slovenia
boris.siraiy@hotmail.com

Abstract - **Medical doctors have many patients with chronic health problems such as headache, balance issues, high blood pressure, leg, shoulder and back pain, etc. In such cases doctors are often in a quandary, should they start with medicaments therapy, send the patient to a specialist or wait for the pain to pass by itself. The duration of diagnosis should not exceed 1 month, but if a doctor decides to send a patient to the specialist it can take quite a long time to have a final diagnosis, and that could be very frustrating for the patient. In many cases, self-help gymnastics “1000 movements”, and “method 23+5”, can result in significant improvement of the health status. These methods have been applied to many patients with chronic health problems. The methods are based on the exercises which are understandable, simple and reasonable for everyone. The effect of exercises is multifaceted: improved blood flow to all parts of the body, relaxed musculoskeletal system, improved functioning of the organs, decreased chronic pain, and prevention of many diseases. In utilising one of the methods presented in this paper the patients becomes an active participant in self-healing, starts to believe in themselves, and last but not least they save a lot of time and money which would be spent on medical treatments.**

Key words: self-help; “method 23+5”; chronic health problems; gymnatics “1000 movements”.

I. INTRODUCTION

Regular physical activity has positive effects on morbidity, mortality and general cardiovascular health [1,2]. Benefits of regular physical activity on physical status are mostly studied in young and middle-age individuals, while in elderly population such studies are not as represented, although it was already established in the early nineties, that elderly people can also improve their functional capacities, prevent disability and loss of independence by regular exercising [3,4,5]. Also, it was shown that regular physical activity prevents falls and osteoporosis among elderly subjects, with increasing bone mass and muscle strength [6,7]. However, regardless of the proven positive effects, physical inactivity remains one of the world’s biggest problems, because one in four adults is not sufficiently physically active [8]. Due to this fact the Centres for Disease Control and Prevention (CDC), and the American College of Sports Medicine (ACSM) in 1995 published a preventive recommendation that “Every US adult should accumulate 30 minutes or more of moderate-intensity physical activity on most, preferably all, days of the week” [9].

In 2013, the World Health Organisation (WHO) reported that around 3.2 million deaths each year are caused by physical inactivity [10]. Because of this fact many governments around the world recognise the importance and the large impact of physical activity on health status of their citizens, which has led to the drafting of global and national guidelines for physical activity [11,12,13,14]. Most governments are also faced with continuously growing costs of health care. Recently they invested a lot of money to develop new medical devices based on the ICT and sensor technology to de-accelerate the health systems costs, but with relative lean results. The key problem is the accessibility of the devices and willingness of the citizens to do something actively for their health. In different countries the attitudes towards the self-help approach to the prevention of health problems is very diverse and culturally related and it is conditioned on the notion of wellness and happiness in one's life. The present method that has been developed in Slovenia during the last decade could be recognised as a good practice and could be applicable to other regions in Europe and elsewhere. The key ingredients are voluntary organisation, education of citizens in regards to the health problems and implementation of the Gymnastics and developing skills for self-therapeutic treatment by the participants themselves [15].

Slovenia with population of 2 million has a big propotion of physical active population [16]. Also, analyses show that elderly population is above average physically active [17]. Certainly, morning gymnastics in the open air, entitled “1000 movements” which involved more than 4000 elderly participants [18], has a positive effect on the physical activity of elderly population. The exercises are organized by the School of Health Society (SOH). SOH is a nongovernmental organisation, well connected to the local communities. There are more than 200 SOH groups covering the majority of the regions of Slovenia. First group started exercising in 2006 in Celje, and in 2007 and 2008 new groups were established in the Primorska region. Today, the growth rate is approximately

20% per year. Each SOH group has at least 3 instructors among their members, which assures the continuous presence of at least one instructor each morning from 7:30 to 8:00 a.m., 6 days per week (except Sundays and holidays), during the whole year. The SOH management provides a yearly plan of educational seminars for the instructors which are chosen from the most motivated and capable members. The changing weather conditions present no obstacle for the exercise; only in rain the groups find some sheltered place, also outside [15].

For participating in morning gymnastics "1000 movements" no special sports dress is required. The training encompasses several dimensions of exercises: aerobic exercise training (AET), resistance exercise training (RET), flexibility exercise to preserve or extend range of motion (ROM) around joints, and balance training designed to increase lower body strength and reduce the likelihood of falling. The program consists of 50 different drills gathered in subgroups focused on the following body parts or goals: arms and shoulders, neck, trunk, balance and legs. Morning gymnastics has the following positive aspects for participants: improvement of physical health status, positive social aspect, and maintenance or improvement of mental and cognitive health with theoretical training and development of skills.

Structured exercise programs are a type of physical activity which has been beneficial in older people [19]. The benefits of exercising depend on continued participation; however, a change in lifestyle and inclusion of regular exercising in daily routine is very difficult for many people, regardless of age. In order to develop effective strategies for increasing participation in structured exercise programs by older adults, it is necessary to understand the individual, social, and demographic factors associated with adherence to this health-promoting behaviour [20,21].



Figure 1: Gymnastics "1000 movements" (foto: Dunja Wedam).

Many of the abovementioned people or, moreover especially them, are the one with health issues who need self-help and exercising. In such cases "method 23+5" can help them to improve their physical and health status.

"Method 23+5" is a simple method which includes 23 exercises for different parts of the body and at the end participants made circles with 5 different parts of the body. Exercises included in this method are simple and effective. All exercises could be finished in about 20 minutes, and all information and description of the exercises is available online for free [22]. To participate in "method 23+5" no special equipment or sports dress is required, and everyone can do the exercises at their home.

II. METHOD

This paper will present the effects of group morning gymnastics "1000 movements" and individual self-help "method 23+5" from different points of view. Effects of morning gymnastics will be objectively evaluated by ECG measurements, senior fitness tests (SFT), and questionnaire. Effect of "method 23+5" will be evaluated by medical personnel and statistics of changes in health status of active participants.

A. ECG measurements

For the purpose of this study 19 participants, younger than 70 years-old, who participated in morning gymnastics at least one year were monitored with a telemetric ECG device, and measurements lasts 24-hours [23]. The ECG sensor is small, flexible, and unobtrusive for users, with more than seven-day autonomy and simple for use. It is appropriate for continuous measurements during physical activities [15,24,25,26,27]. ECG measurements acquisition run continuously in the background on the phone, and do not interfere with the usual mobile phone functions. However, short interruptions in measurement can always be present, for example, due to lost radio connection [28].

B. Senior Fitness Tests

In presented paper previous studies including participants who participate in morning gymnastics will be analysed [15,29]. In the above mentioned studies six SFT tests were used: chair 30-s sit-to-stand; chair sit-and-reach; back scratch; 6 min walk; biceps curl; 8-ft up-and-go.

C. Questionnaire

To evaluate objective opinion on the impact of morning gymnastics, a questionnaire was prepared for new participants in 24 new established groups in several Slovenian regions, and it was used in years 2018/19.

Questionnaire contained the following questions:

1) Gender/ sex
2) Do you participate in any physical activity?

3) What is your general health?

4) Are you active in some other activities?

5) Observations related to changes in your mobility and quality of life after 5 months or more of continuous exercise (6 times weekly); Improvements, worsening, unchanged

6) What does the "1000 Movements" exercise mean to you?

D. *Evaluation of the effects of the "method 23+5"*

Effects of the "method 23+5" were evaluated by a medical doctor, who analysed the improvement of health status of participants.

III. RESULTS

A. *ECG measurements*

ECG measurements were attended by 19 participants, younger than 70 years-old, who participated in morning gymnastics "1000 movements" for at least one year. Monitoring lasted 24-hours.

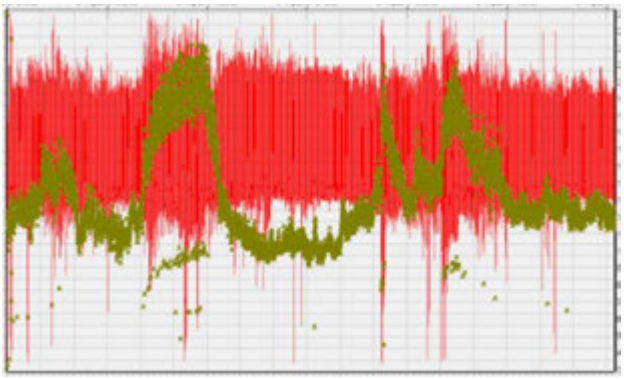


Figure 2: Example of ECG monitoring.

Analyses of the results show that participants in morning gymnastics have a good cardiac status without any detected cardiovascular abnormalities. It was found that heart rate (HR) reaches normal values, (the same as before activity), soon after finishing exercising. Figure 2 showed that HR starts to rise sharply after the beginning of activity, and when cardiovascular system adopts to activity HR falls down to the level before activity. Also, in the second part of the graph we can see that regular daily activities like walking, housekeeping or gardening does not raise HR as much as participating in physical activity. All this indicates good cardiovascular status of measured participants.

B. *Senior Fitness Tests*

Review of previous studies which analysed impact of morning gymnastics "1000 movements" on improvement of physical status by SFT showed the following results [15,29]. After one year of participating in morning gymnastics results were significantly improved ($p<0.05$) in four SFT tests:

- 30-s sit-to-stand,
- chair sit-and-reach,
- back scratch,
- 8-ft up-and-go.

There was no progress in two SFT tests ($p>0.05$):

- 6 min walk,
- biceps curl.

These results indicated that regular participating in morning gymnastics has a positive effect on a physical status of participants, however significant progress has been recorded in tests where strength and resistance of lower extremities are predominant, and in test of flexibility. After a year of exercising strength of upper extremities and endurance shown no improvements.

C. *Questionnaire*

Questionnaire was sent to 333 participants in 24 groups, and 251 completed forms were received. That represents 75.38% response. Results of questionnaire show the following:

1) Gender: 4% males, 96% females.

2) Before participating in morning gymnastics "1000 movements", 40.64% of participants reported a daily walking activity, and 49.80% occasionally swam.

3) In general, participants at the beginning of participating reported problems with squat (39%), running (37%), and being overweight (33%).

4) Beside participating in morning gymnastics "1000 movements", 59.70% of participants reported that they walk more, and 69.65% reported they have better social life and found more friends.

5) Participants reported more than 52% improvement in balance, and in seat-and-reach test; 46.77% improvement in squat; and 41.29% in shoulder mobility.

6) 48.26% of participants reported that morning gymnastics represent recreation and better physical status, and 44.28% reported that it presents them with a high social factor, because they meet new people and establish new friendships.

D. Evaluation of the effects of the "method 23+5"

Medical doctor reported that active participation in the "method 23+5", at least 3 times per week, improved the health status of participants. With this self-help method participants recognised significant improvement in some chronic health issues like stiffness, when body parts become more flexible and relaxed, helping to improve headaches, back pain, leg and arms pain, and it also helps to improve some inner organs issues.

Exercises are prescribed to participants by medical personnel, and they are used two-fold i) preventive exercises, and ii) exercises for improvement of some health issue or bad habit, for example poor posture.

IV. DISCUSSION

Analysis of the self-help method "1000 movements" and "method 23+5" show their benefits for general health, and health improvement of participants who are included in one of the methods. While method "1000 movements" was already analysed in previous studies [15,29], and benefits of the method are already well known, "method 23+5" is new and not so well known until now.

In a previous study, analysis of the method "1000 movements" shows its positive effect on flexibility, increased strength, better cardiovascular fitness and better general well-being [15]. However, the tested sample in that study was too small to make some general conclusions. In the presented paper, the sample was much higher and presented results give the same conclusion: *regular participating in morning gymnastics "1000 movements" has a positive effect on general cardiovascular status of participants.* Review of senior fitness tests show there is a significant improvement in tests: 30-s sit-to-stand, chair sit-and-reach, back scratch, 8-ft up-and-go, which measure strength of lower extremities, and flexibility of the body and upper extremities. In tests: 6 min walk, and biceps curl, which measure endurance and strength of upper extremities no improvement was detected. This indicates that more attention should be paid to the exercises for endurance and strength of upper extremities.

Analysis of the questionnaire shows slightly different results in comparison to the previously observed group [15]. In presented paper approximately the same sample was used, and the structure of participants is 96% female, and 4% male. That is slightly different than in previous analysis when 93% of participants were female, and 7% were males. Participants reported that almost 60% of them now walk more, and almost 70% have a better social life, while in previous study this percentages were much lower, only 46% of them walked more, and 40% reported better social life. In presented questionnaire participants reported slightly lower improvement in some physical tests like squat (46% vs 51%), shoulder mobility (41% vs 44%), etc.

General conclusion of both studies is very similar, for participants morning gymnastics bring general well-being, healthy life and better social life. Many of them start participating in morning gymnastics because they feel good because of social moments, and later they also recognise health benefits. For participants who already participate in morning gymnastics, "method 23+5" can present the upgrade and additional benefit of exercising, while they can practice this self-help method at home after finishing morning gymnastics or in cases when they cannot participate in morning gymnastics for some objective reason. Also, this method could be very beneficial for people who do not attend morning gymnastics or any other exercising.

V. CONCLUSION

This paper showed the benefits of two different self-help methods. While method "1000 movements" were objectively measured with different tests, and its benefits were shown by analysis of the results, "method 23+5" needs some objective measurements to show the results in an objective way. Future studies should include objective tests of participants of the method which will show or deny positive effects of the method on participant health. The more detailed pilot study project is strongly recommended and needed before final impact of the presented self-help approach could be established and highlighted.

ACKNOWLEDGEMENT

Special gratitude to Maja Nagode, Andraž Purger and Neda Galijaš, who generously shared the results of their work and thus contributed to the successful creation of our paper.

REFERENCES

[1] P.O. Astrand, "Why exercise?" Med Sci Sports Exerc, vol 24. 1992, pp. 153-162.

[2] T.L. Bush, S.R. Miller, M.H. Criqui and E. Barrett-Connor, "Risk factors for morbidity and mortality in older populations: an epidemiologic approach," 2nd ed., Principles of Geriatric Medicine and Gerontology. Hazzard WR, Andres R, Bierman EL, Blass JP, editors. McGraw Hill, New York, 1990; pp.125-136.

[3] E.L. Smith, R.P. Di Fabio and C. Gilligan, "Exercise intervention and physiologic function in the elderly," Top Geriatr Rehabil, vol. 6, 1990, pp. 57-68.

[4] J.L. Fleg and A.P. Goldberg "Exercise in older people: cardiovascular and metabolic adaptations," 2nd ed., Principles of Geriatric Medicine and Gerontology. Hazzard WR, Andres R, Bierman EL, Blass JP, editors. McGraw-Hill, New York, 1990; pp.85-100.

[5] J.D. Posner, K.M. Gorman, L.M. Gitlin, L.P. Sands, M. Kleban, L. Windsor and C. Shaw, "Effects of exercise training in the elderly on the occurrence and time to onset of cardiovascular diagnosis," J Am Geriatr Soc, vol. 38, 1990, pp. 205-210.

[6] P. Kannus, "Preventing osteoporosis, falls, and fractures among elderly people: promotion of lifelong physical activity is essential," BMJ, 1999, pp. 205-206.

[7] P. Kannus, S. Niemi, M. Palvanen and J. Parkkari, "Fall-induced injuries among elderly people," Lancet, vol. 350, 1997, pp. 1174.

[8] World Health Organization, *(2018)*. Physical Activity Key Facts [online]. Available from: http://www.who.int/mediacentre/factsheets/fs385/en/ (Updated February 2018).

[9] R. Pate., M. Pratt, S. Blair, W.L. Haskell, C.A. Macera, C. Bouchard, et al., "Physical activity and public health. A

recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine," JAMA, vol. 273, 1995, pp. 402–407.

[10] World Health Organization. Diet and Physical Activity Factsheet. Secondary Diet and Physical Activity Factsheet. 2013. http://www.who.int/dietphysicalactivity/factsheet_inactivity/en/index.html

[11] A. Bauman and C. Craig, "The place of physical activity in the WHO Global Strategy on Diet and physical activity, " Int J Behav Nutr Phys Act, vol. 2, 2005, pp. 10.

[12] U.S. Department of Health and Human Services, "Physical activity guidelines for Americans," U.S. Department of Health and Human Services, Division of Nutrition, Physical Activity and Obesity. Atlanta: National Center for Chronic Disease Prevention and Health Promotion, 2008.

[13] N. Townsend, P. Bhatnagar, K. Wickramasinghe, P. Scarborough, C. Foster and M. Rayner, "Physical activity statistics 2012," London: British Heart Foundation; 2012. *Google Scholar*.

[14] Ministry of Health. Guidelines on physical activity for older people (aged 65 years and over). Wellington: Ministry of Health, 2013.

[15] B. Širaiy, J.U. Stanič,, A.S. Poplas and Z. Katkič, "Impact assessment of the morning gymnastics "1000 movements" via ECG and sport test," 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, Opatija, Croatia, 2018, pp. 284-288.

[16] Eurobarometer, 2014: https://ec.europa.eu/commission/presscorner/detail/sl/IP_14_300. Last visited 16.01.2020

[17] R. Slabe Erker and T. Bartolj, "Telesna (ne)aktivnost starejših v Sloveniji, " Inštitut za ekonomska raziskovanja, 2011.

[18] N. Grishin, "Textbook for the trainers of the Šola Zdravja Soceity "1000 movement," ed. The School of Health Society, 2018.

[19] A.M.A. Picorelli, L.S.M. Pereira, D.S. Pereira,, D. Felício and C. Sherrington, "Adherence to exercise programs for older people is influenced by program characteristics and personal factors: a systematic review," Journal of Physiotherapy, vol. 60(3), 2014, pp. 151-156.

[20] M.A. Dolansky, B. Stepanczuk, J.M. Charvat and S.M., Moore, "Women's and Men's Exercise Adherence after a Cardiac Event: Does Age Make a Difference?" Res Gerontol Nurs., vol. 3(1), 2010, pp. 30–38.

[21] R.A. Seguin, C.D. Economos, R. Palombo, R. Hyatt, J. Kuder and M.E. Nelson, "Strength training and older women: a cross-sectional study examining factors related to exercise adherence," J Aging Phys Activ, vol. 18, 2010, pp. 201–218.

[22] Studio orhideja: https://www.studio-orhideja.com/metoda-samopomoci-23-5/ Last visited 18.01.2022

[23] R. Trobec and I. Tomašić, "Synthesis of the 12-lead electrocardiogram from differential leads," IEEE Trans Inf Tech Biomed, vol. 15, 2011, pp. 615–621.

[24] B. Širaiy, R. Trobec and V. Ilić, "Quality of one-channel telemetric ECG sensors signal in maximum exercise stress tests," Meas Sci Rev, vol. 19(3), 2019, pp. 79-85.

[25] B. Širaiy, V. Ilić and R. Trobec, "Evaluating telemetric ECG body sensor signal in exercise stress test: Pilot study," 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2018, pp. 265-270. IEEE.

[26] B. Širaiy and R. Trobec, "Telemetric ECG monitoring during physical activity in field tests," 21st International Multiconference Information Society 2018, Vol. I, pp. 50-52.

[27] B.J. Širaiy, "Evaluacija jednokanalnog telemetrijskog EKG uređaja u različitim submaksimalnim i maksimalnim laboratorijskim i terenskim testovima," Doctoral dissertation, Univerzitet u Beogradu-Fakultet sporta i fizičkog vaspitanja, 2019.

[28] M. Jan and R. Trobec, "Long-term follow-up case study of atrial fibrillation after treatment", 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2017, pp. 297- 302.

[29] B. Širaiy, J.U. Stanič, A. Purger, M. Nagode, and Z. Katkič, "Effect of "School of Health Society" exercise program on a group of elderly subjects" In 12. International Scientific Conference, Bled, Slovenia, 2019, pp. 495-501.

Design of a Non-invasive ECG-based Glucose Measurement System

M. Gusev*, E. Guseva** and L. Poposka***,
*Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering, Skopje, North Macedonia
**Innovation Dooel, Skopje, North Macedonia
***Ss. Cyril and Methodius University, University Clinic of Cardiology, Skopje, North Macedonia
E-mail: marjan.gushev@finki.ukim.mk, ema.guseva@innovation.com.mk, lidijapoposka@gmail.com

***Abstract*—Diabetic patients have to pay for each glucose reading with a blood drop and a small fortune. In addition, routine finger pricking is troublesome for diabetic patients because it can lead to scarring. It is no surprise then that the idea that glucose measurement can be done cheaply and in a non-invasive way surpasses the wildest dreams of diabetic patients. The goal of this paper is to present the design of a new technology solution for non-invasive glucose measurement based on processing the electrocardiogram obtained via a light easy-to-wear ECG monitor. We present details on how to develop a service that tracks glucose levels based on real-time ECG monitoring, and using sophisticated machine learning and related technologies. Our initial analysis shows that no similar solution is present on the market today, although several research initiatives are ongoing.**

Index Terms*—*ECG*; *HRV*; *diabetes*; *autonomic neuropathy*; *Edge computing

I. Introduction

In the last few years we have spent thousands of hours processing ECG data and developing diagnostic algorithms while working on our product ECGalert [1]. During this time we verified initial exciting new scientific research results [2], [3], [4], [5], which proposed that analysis of ECG signal can determine glucose levels in the body and its ability to manage glucose. This is really exciting because the development of this technology would allow patients to measure both heart activity and glucose levels at the same time with one non-invasive wearable device. Needless to say, this realization motivated us to brainstorm ideas of how to build this new feature into our existing product ECGalert and improve healthcare services.

Since diabetes is a chronic disease, which needs to be managed regularly, patients need to measure their glucose levels multiples times per day. And this is a pain point for them because frequent finger pricking, the measurement technique most commonly used, causes problems such as intolerance, annoying pain, development of massive scarring/callous formation and loss of sensibility/perception hindrance [6].

Glucose measurement methods can be categorized into three categories: invasive, minimally invasive and non-invasive. The least intrusive of these, non-invasive glucose measurement, is based on measuring glucose concentrations from its chemical, electrical, optical, thermal, acoustic property or any combination of these [7]. It includes devices that do not invade the human body and can be based on various methods [8], including 1) spectrometry, 2) analysis of other parameters correlated to the glucose level, such as analysis of electrocardiogram (ECG) or heart rate variability (HRV) parameters, or 3) analysis of other physiological fluids [9], such as urine, sweat, saliva, etc.

The goal of this paper is to present the design details of our system, which is a non-invasive glucose measurement system based on HRV parameters. It analyzes the functionality of the autonomous nerve system by processing an electrocardiogram (ECG) and various heart rate variability (HRV) parameters. An ECG is collected by a wearable sensor and transferred to a nearby smartphone to enable remote monitoring and processing of heart rate variability [8] with sophisticated machine learning algorithms, which inform on how well the human body is managing its glucose level.

The rest of the paper is organized as follows. Related work on non-invasive methods based on HRV is presented in Section II. Section III specifies user and computing requirements, and presents a new architecture that matches these requirements to process the relevant information. Identified problems and open issues are discussed in Section IV, along with an analysis of benefits and impact of such an approach. Finally, Section V gives relevant conclusions and directions on future work.

II. Related work

Villena Gonzales et al. [10] analyze minimally invasive and non-invasive glucose measurement trends based on electric, thermal, and optical methods, including also, nanotechnology approaches. Garcia-Guzman et al. [7] conclude that most suitable non-invasive glucose measurement include chemical, electrical, optical, thermal, acoustic properties or their combination.

Transdermal and optical methods are most popular as non-invasive techniques [11], and those used for wearable and non-invasive devices are based on monitoring interstitial fluid, detection of sweat (eyeglass, flexible wristband, etc.), breath analysis, saliva analysis (tattoo printed on a tooth, etc.) or ocular fluid (smart contact lens).

But there is a new kind of non-invasive glucose measurement on the horizon. We recently reported a comprehensive analysis of noninvasive glucose measurement with special emphasis on using Machine Learning and Neural Network Methods and the possibility of using HRV parameters for

detection [8]. The rest of this section will review findings in this area and argue why it is good to measure HRV and glucose levels together.

The analysis of which HRV parameters are better at differentiating between diabetics and non-diabetics has been addressed by many researchers, including tests of time, frequency and other domain HRV parameters [4], or long versus short term measurements [5], [12], [13]. Typically, published results [14], [4] confirm better sensitivity of long versus short term HRV analysis.

Specifically, detecting decreased heart rate variability can differentiate diabetic patients from non-diabetics [15]. In fact, lower values of both time and frequency domain HRV parameters are detected in diabetic patients [2]. Ewing et al. [16] have conducted many experiments to find which specific tests can differentiate diabetic patients from non-diabetics, including short-term R-R alterations and heart rate response to respiration; standing, Valsalva maneuver, orthostasis, and isometric exercise.

Diabetes and cardiovascular disease often go together, so there is a significant added benefit to develop a technology that measure both. For example, autonomic dysfunction develops when the Autonomic Nervous System (ANS) can not perform its essential tasks, leading to autonomic neuropathy. The causes of autonomic neuropathy may be attacks by the immune system and most commonly by diabetes, especially in cases of poor glucose control, in which case it is called diabetic autonomic neuropathy (DAN) [17]. DAN is a serious and common complication of diabetes [18] with an increased risk of cardiovascular mortality. Resting tachycardia and exercise intolerance are among the most recognized signs and symptoms of autonomic dysfunction. Not surprisingly, Ballinger et al. [19] have reported a correlation of 0.85 between HRV parameters and glucose levels, indicating that diabetes caused progressive autonomic dysfunction. Their study included deep learning methods in the analysis.

Furthermore, one of the most overlooked of all serious complications of diabetes is cardiovascular autonomic neuropathy (CAN)[17], which encompasses damage to autonomic nerve fibers that innervate heart and blood vessels, resulting in abnormalities in heart rate control and vascular dynamics [20]. Slower heart rates and reduction in variability of heart rate is the earliest indicator of diabetes [14], [21].

III. System Description

New sophisticated algorithms from artificial intelligence, machine learning and parallel processing can enable simultaneous processing of multiple data streams of high volume and speed. In addition, our ECGalert monitoring technology is consistent with the modern trend of cloud computing, so that access to measurement results is instant and available from anywhere.

To start off the description of our design, a patient is continuously wearing a light ECG monitor, as is presented in the left part of Figure 1. The wearable ECG monitor can be worn for up to 7 days without a need to recharge and even



Fig. 1. ECG sensor and proposed mockup solution. The application ECGalert (available for Apple IOS) is currently providing a real-time ECG monitoring system, and the glucose level is add-on to this application as a mockup demonstration of the prototype.

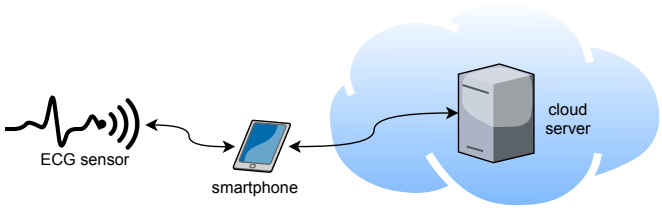


Fig. 2. System architecture of the proposed solution. The sensor sends data to smartphone, as an intermediate dew computing device that communicates with the cloud server.

in the unlikeliest of places, like in the shower [22], [23]. The monitor wirelessly connects to the patient's smartphone, which displays the glucose level, heart rate and ECG, as presented in the right part of Figure 1.

Given that the patient's smartphone is connected to the Internet, the smartphone application will upload the measured ECG data on the cloud where it can be accessible by both the patient, her doctor and a caregiver, anytime and from any computing device.

For this cloud computing purpose, we use our monitoring technology based on a Software as a service provision, which was developed within the ECGalert project [1], [24]. The system concept is presented in Figure 2.

The ECG sensed analog signal is being converted to a digital sampled form by the ECG sensor. In order to save the energy on the ECG sensor, it does not store data samples internally, and therefore all data samples are transferred to the smartphone using a low energy Bluetooth technology [24]. This enables the small battery on the wearable sensor to be recharged on 7 days and enable long-term monitoring.

Therefore, the embedded system in the ECG sensor runs a simple application that initiates an AD conversion on regular time intervals determined by the sampling frequency, and generates data samples with predefined bit resolution and zero offset value. These data samples are stored in a small buffer

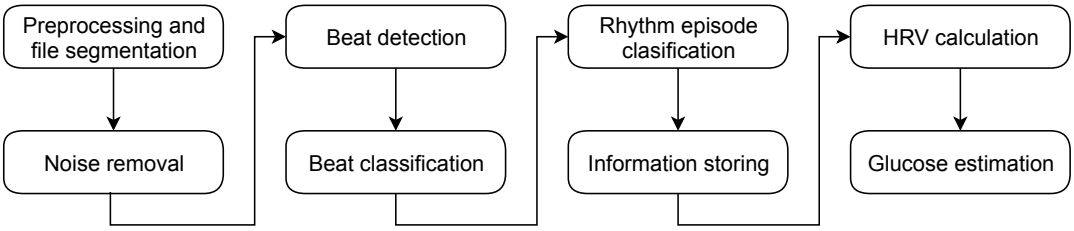


Fig. 3. Software architecture of the smartphone application, presenting the data processing and storing modules.

and when the buffer fills it initiates a data chunk transfer via the Bluetooth device. This simple processing logic ensures that only small energy is required for the ECG sensor performance.

The ECG analog signal is converted to a digital sampled form by the ECG sensor. For battery conservation purposes, the ECG sensor does not store data samples internally, but rather it transfers all data to a smartphone via low-energy Bluetooth technology [1]. This is what allows for the small battery on the wearable sensor to only be recharged every 7 days and thus enables long-term monitoring.

The smartphone captures the ECG data samples and then processes them, temporarily stores them before finally transferring them to the cloud server. The smartphone plays the role of an edge computing device, or a dew computing device found on of the edge of the Internet network. The software on the smartphone acts as a dew server and transfers all data to the server and communicates with the results from high demanding big data and ML algorithm executions on the server in case of Internet connectivity, and serves the entire processing in case of no Internet connectivity.

Figure 3 presents the inner functional parts of the software application that run on the smartphone. First, the ECG signal is pre-processed and noise is removed efficiently by the sophisticated signal processing and filtering algorithms. Next, beat detection and beat classification are performed according to our algorithm [25] with high sensitivity and a positive predictive rate. The beat detection module detects the location (time moment represented by a sample id) of the heartbeats (QRS complexes), while the beat classification, the beat type (normal, ventricular or atrial). They proceed annotations to the rhythm episode module to determine the associated rhythm and updates the annotations with the corresponding information about start and end locations of the rhythm episodes. Finally, the annotations and ECG files are stored in by the information storing module that is also responsible to communicate with the cloud to exchange information.

The presented ECGalert technology is then upgraded by a HRV calculation module and the results are used by glucose estimation algorithms based on sophisticated ML algorithms. Currently we are developing ML-based algorithm using SDNN and RMSSD long-term HRV parameters to detect the ability to regulate glucose level by predicting the HbA1C parameter (usually interpreted as an average level of blood sugar over the past 2 to 3 months.

The cloud server runs a sophisticated SaaS solution [26] which enables medical experts to remotely monitor the patient and analyze the patient's heart condition. The ECGalert technology is then upgraded to analyze HRV parameters considering not just time domain parameters, but also frequency domain and other domains. The obtained values are matched to predefined thresholds and the combination of various HRV parameters set a trigger to classify the patient's ability to control her glucose level using sophisticated ML algorithms.

IV. Discussion

A. Identified problems and open issues

We faced many challenges during our research and development activities and tried to solve them in the most appropriate way. In this section we elaborate on the limitations of obtaining relevant results using the proposed solution:

- *Proper measurement*. Measuring ECG is very sensitive to the body, and plenty of factors may create a bad ECG signal abundant with noise. For example, hair may corrupt good placement of ECG electrodes and prevent sufficient quality contact. In addition, it is extremely important to have good location detection to achieve good indication of the R wave in the QRS complex and enable good interpretation even in case of abundant noise.
  In light of this and with the aim to eliminate the possibility of bad ECG measurements, we conducted clinical research and evaluation on hospital premises. Lab measurements of blood samples and ECG measurements were provided by medical experts. The patients were hospitalized for at least one day, and they were asked not to move during the measurement in order to obtain better results. All patients signed a corresponding consent form and agreed to participate in clinical research and evaluation.
- *Muscle noise*. The biggest enemy to achieving a reliable result is muscle noise since even small negligible movements such as breathing may corrupt the ECG signal. The real problem is that the frequency of energy level of muscle noise is often found to be overlapping with those of the QRS complexes. This prevents simple filtering techniques from producing results with sufficient quality, and can generate a lot of artifacts, misleading the correct beat detection, and correspondingly beat classification.

Although, our algorithm for beat detection and classification [25] has proven extremely high sensitivity and positive predictive rate, we conclude that, in cases of a lot of muscle noise, even an expert's human can not detect the location of a QRS. For this purpose, we have built specific algorithms to detect energy levels and their frequency to mark un-interpretable regions abundant noise.

In addition, we have built advanced monitoring and marking tools for ECG signal processing. These tools were used by medical experts to correct any wrong detections by noise-corrupted signals and clarify whether an artifact is mixed with a QRS wave, or whether a wrong beat was classified. Of course, this activity has been a time-consuming task, especially because we had three independent medical experts mark the beats and noisy uninterpretable intervals. Then, we used the majority rule to establish the correctness of results.

- *Proper HRV calculation*. In order to calculate HRV values one needs to understand the essence of calculating HRV values and sensitivity to small variations. Note that we have concluded that a lot of authors have published various solutions and methods of calculating the HRV values and we found that most of them are not calculating the correct value. The reasons for detected discrepancies include at least the following:
  - not paying attention to the beat type,
  - not making a distinction between RR and NN intervals,
  - the implemented algorithm does not realize the elaborated method,
  - the selection of sampling frequency plays an important role in the calculation of HRV values, and the mistakes due to the sampling frequency can go up to 5%, and
  - the calculation of frequency domain HRV parameters require huge processing and often simpler methods are used which do not calculate the exact frequency domain results.

  In light of this, we implemented our own system to calculate correct and reliable HRV values.
- *Frequent noise intervals*. Even when the intervals with a lot of noise and uninterpretable ECG signal are properly marked, they become an obstacle for long term measurements, especially for:
  - NN analysis of HRV parameters, and
  - RR analysis of HRV parameters.

  This means that frequent noise intervals disable the production of data input for estimating the required HRV parameters.
- *Relevant results*. Our research and development has found a correlation between the level of glucose and specific HRV parameters, which confirmed previous related work results, and additionally, determined correlation factors with higher precision. In regards to this, we conclude with the following limitations which can lead to relevant results:
  - Only long-term HRV measurements are relevant for proper judgement about autonomic dysfunction, specified by a strong Pearson and Spearman correlation of SDNN and RMSSD to HbA1C [27], and
  - The system currently gives information on the average ability to control glucose (2 or 3 months average), and the ongoing research shows to be able to reveal a real-time measurement of glucose levels based on the last one-hour continuous ECG measurements (note that also the convenient invasive methods for blood glucose level gives an estimation of the glucose concentration in the last half an hour, due to the propagation time and the ANS performance).

  As an overall conclusion, better results are obtained when the patient sleeps, so that the information when the patient wakes in the morning is a good starting point to discern glucose levels. This reliability when patients make measurement in the morning prior to breakfast, after at least 8-hour fasting while sleeping, is also present in conventional invasive methods based on blood drops.
- *ML algorithms*. The development of ML algorithms was a time consuming process. We have performed a clinical evaluation with 200 patients classified in various diabetes categories, such as no diabetes, diabetes type 1 or 2, or those with undefined status. The collected data contains many lab results for blood analysis, and includes 1 to 3 days of ECG monitoring of hospitalized patients. Correspondingly, the calculated HRV data contains values starting from 30 sec (very short-term) HRV parameters, which span over the whole measurement period, through 1, 2, 5 and 10 minutes of HRV calculations over the whole measurement period, to long-term HRV calculations of 2, 4, 6, 8, 10, 12, 16, 20 and 24 hours.

  We have used two methods to measure HRV parameters. The first one uses an uninterruptible signal, and the second one uses RR and NN intervals of signal interrupted by noise marked intervals. The whole process generated thousands of data per patient to be correlated, and it took a lot of processing requirements to run the ML methods, including Deep Learning, Random Forest, Logistic Regression and others being currently under development to improve the perfromance [27]. The prototype has shown sufficient results that will be published soon.

B. Analyses of benefits

Table I compares the methodological approaches of invasive, minimally invasive and non-invasive glucose measurement methods.

The following benefits are expected from introduction of a new non-invasive solution

1) Consumer benefit: Consumers stand to benefit from a significant reduction in costs for healthcare and the opportunity to practice preventative medicine and prolong their lifespan. Great things happen when technologies that were once stuck

TABLE I
COMPARISON OF DIFFERENT TECHNOLOGIES AND METHODS

Class	Method	Technology characteristics
Invasive	Finger pricking	Patients pay with a blood drop to make a measurement
Invasive	Implanted sensors	Implanted under skin transmits data each minute to a specialized device
Minimally invasive	An array of small needles	Activated on demand by inserting an array of small needles that collect small blood samples from capillary under the skin
Non-invasive	Infrared, ultrasound or radio waves	Special devices stick to ears, fingers or other body parts
Non-invasive	Measuring other physiological parameters	A special device is analysing collected measurements of blood pressure wrist monitor, heart rate watch etc.
Non-invasive	ECG-based solution	Wearable device for remote real-time ECG and glucose monitoring accessible via Internet

TABLE II
ANALYSES OF BENEFITS

Healthcare problem	This solution
Invasive glucose monitoring methods pose a threat	Non-invasive glucose level monitoring
On-demand approach to monitoring fails to give a holistic image of the patient's health	Continuous monitoring solution that provides rich data for informed treatment decisions
Cardiac disease is closely linked to diabetes, but is often not diagnosed until it is too late	Long-term ECG monitoring makes it possible to practice preventative cardiovascular medicine
Economic problem	**This solution**
Self-monitoring is expensive for diabetic patients	Affordable solution that is more than half the price of existing continuous monitoring methods
Technological problem	**This solution**
Continuous glucose monitoring devices can only last up to 2 weeks	The ECG monitor can be worn up to a week without needing to recharge and lasts more than 2 years
Patients often have to manually record measurement values	Automatic transfer of measurements on the cloud accessible anytime with any computing device

in the lab of a scientist become commercially available consumer products, and when once uninformed patients have a continuous stream of data on their health.

2) Social impact: When living with diabetes is no longer a burden, societies will flourish with increasingly productive individuals in good health. The technological innovation contained in this project will make the healthcare industry more efficient and effective, hence more attractive. So, more time can be spent enjoying a greater quality of life for everyone affected by this innovation.

3) Technological impact: Not only have we have found that ECG can reveal information about glucose levels, but also blood pressure and related issues, which are considered as trending topics for future research and development.

Problems were mainly classified as a specific healthcare, economic and technological problem. Table II presents how this solution solves the identified problem.

Finally, the main benefit of this approach is that patients can monitor both their hearts and glucose levels, all with one light non-invasive sensor.

V. CONCLUSION

The possibility of measuring glucose non-invasively and getting real-time results as opposed to existing invasive expensive and delayed results is of immense value for customers. In this paper, we have presented a new design of a solution for non-invasive glucose measurement system based on a wearable ECG sensor. The presented design aims at developing deeper analysis, research and development of a commercial product to deliver service based on an automated monitoring software agent, which will send estimated glucose level results to a cloud server and enable real-time continuous monitoring, and notification of extreme glucose levels.

Advances beyond the State-of-the-Art created by the approach in this solution are:

- Comfortable easy to wear non-invasive ECG monitor,
- No need to clip on ear or another body part that collects interstitial body fluids,
- Combined measurement and diagnostics of heart function and glucose levels,
- Continuous real-time measurement instead of on-demand activation
- No limited measurement duration
- Cloud-computing data storage solution

The presented design includes a specification of requirements for the design of a smartphone application and SaaS cloud- provided application for the end-user, which permits sharing of measurement data and results between the patient, caregiver and medical experts. Future work includes further research and development of algorithms to find a specific calculation of HRV parameters and their exploitation in ML-based algorithms to detect glucose levels.

REFERENCES

[1] M. Gusev, A. Stojmenski, and A. Guseva, "ECGalert: A heart attack alerting system," in *International Conference on ICT Innovations*. Springer, 2017, pp. 27–36.

[2] H. Kudat, V. Akkaya, A. Sozen, S. Salman, S. Demirel, M. Ozcan, D. Atilgan, M. Yilmaz, and O. Guven, "Heart rate variability in diabetes patients," *Journal of international medical research*, vol. 34, no. 3, pp. 291–296, 2006.

[3] F. Bellavere, "Heart rate variability in patients with diabetes and other noncardiological diseases," *Heart rate variability*, pp. 507–516, 1995.

[4] M. Malik, A. J. Camm, J. T. Bigger, G. Breithardt, S. Cerutti, R. J. Cohen, P. Coumel, E. L. Fallen, H. L. Kennedy, and R. Kleiger, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: Task force of the european society of cardiology and the north american society for pacing and electrophysiology," *Annals of Noninvasive Electrocardiology*, vol. 1, no. 2, pp. 151–181, 1996.

[5] F. Shaffer and J. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, vol. 5, p. 258, 2017.

[6] L. Heinemann, "Finger pricking and pain: a never ending story," *Journal of diabetes science and technology*, vol. 2, no. 5, pp. 919–921, 2008.

[7] J. García-Guzmán, N. González-Viveros, and H. H. Cerecedo-Núnez, "Comparative analysis of optoelectronic properties of glucose for non-invasive monitoring," in *Emerging Challenges for Experimental Mechanics in Energy and Environmental Applications, Proceedings of the 5th International Symposium on Experimental Mechanics and 9th Symposium on Optics in Industry (ISEM-SOI), 2015*. Springer, 2017, pp. 55–63.

[8] M. Gusev, L. Poposka, G. Spasevski, M. Kostoska, B. Koteska, M. Simjanoska, N. Ackovska, A. Stojmenski, J. Tasic, and J. Trontelj, "Noninvasive glucose measurement using machine learning and neural network methods and correlation with heart rate variability," *Journal of Sensors*, vol. 2020, no. article ID 9628281, p. 13 pages, 2020. [Online]. Available: https://doi.org/10.1155/2020/9628281

[9] D. Bruen, C. Delaney, L. Florea, and D. Diamond, "Glucose sensing for diabetes monitoring: recent developments," *Sensors*, vol. 17, no. 8, p. 1866, 2017.

[10] W. Villena Gonzales, A. T. Mobashsher, and A. Abbosh, "The progress of glucose monitoring - a review of invasive to minimally and non-invasive techniques, devices and sensors," *Sensors*, vol. 19, no. 4, p. 800, 2019.

[11] A. Nawaz, P. Øhlckers, S. Sælid, M. Jacobsen, and M. N. Akram, "Non-invasive continuous blood glucose measurement techniques," *Journal of Bioinformatics and Diabetes*, vol. 1, no. 3, p. 01, 2016.

[12] H. J. Baek, C.-H. Cho, J. Cho, and J.-M. Woo, "Reliability of ultra-short-term analysis as a surrogate of standard 5-min analysis of heart rate variability," *Telemedicine and e-Health*, vol. 21, no. 5, pp. 404–414, 2015.

[13] T. Kuusela, "Methodological aspects of heart rate variability analysis," *Heart rate variability (HRV) signal analysis: Clinical applications*, pp. 10–42, 2013.

[14] R. E. Maser and M. J. Lenhard, "Cardiovascular autonomic neuropathy due to diabetes mellitus: clinical manifestations, consequences, and treatment," *The Journal of Clinical Endocrinology & Metabolism*, vol. 90, no. 10, pp. 5896–5903, 2005.

[15] C. Meyer, F. Milat, B. P. McGrath, J. Cameron, D. Kotsopoulos, and H. J. Teede, "Vascular dysfunction and autonomic neuropathy in type 2 diabetes," *Diabetic medicine*, vol. 21, no. 7, pp. 746–751, 2004.

[16] D. Ewing, J. Neilson, C. Shapiro, J. Stewart, and W. Reid, "Twenty four hour heart rate variability: effects of posture, sleep, and time of day in healthy controls and comparison with bedside tests of autonomic function in diabetic patients." *Heart*, vol. 65, no. 5, pp. 239–244, 1991.

[17] M. P. Schumer, S. A. Joyner, and M. A. Pfeifer, "Cardiovascular autonomic neuropathy testing in patients with diabetes," *Diabetes spectrum*, vol. 11, no. 4, p. 227, 1998.

[18] A. I. Vinik, R. E. Maser, B. D. Mitchell, and R. Freeman, "Diabetic autonomic neuropathy," *Diabetes care*, vol. 26, no. 5, pp. 1553–1579, 2003.

[19] B. Ballinger, J. Hsieh, A. Singh, N. Sohoni, J. Wang, G. H. Tison, G. M. Marcus, J. M. Sanchez, C. Maguire, J. E. Olgin *et al.*, "Deepheart: semi-supervised sequence learning for cardiovascular risk prediction," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[20] A. I. Vinik and D. Ziegler, "Diabetic cardiovascular autonomic neuropathy," *Circulation*, vol. 115, no. 3, pp. 387–397, 2007.

[21] S. C. Malpas and T. J. Maling, "Heart-rate variability and cardiac autonomic function in diabetes," *Diabetes*, vol. 39, no. 10, pp. 1177–1181, 1990.

[22] R. Trobec, I. Tomašić, A. Rashkovska, M. Depolli, and V. Avbelj, "Commercial ECG systems," in *Body Sensors and Electrocardiography*. Springer, 2018, pp. 101–114.

[23] A. Rashkovska, M. Depolli, I. Tomašić, V. Avbelj, and R. Trobec, "Medical-grade ECG sensor for long-term monitoring," *Sensors*, vol. 20, no. 6, p. 1695, 2020.

[24] M. Gusev, A. Stojmenski, and I. Chorbev, "Challenges for development of an ECG m-health solution," *Journal of Emerging Research and Solutions in ICT*, vol. 1, no. 2, pp. 25–38, 2016.

[25] E. Domazet and M. Gusev, "Improving the QRS detection for one-channel ECG sensor," *Technology and Health Care*, vol. 27, no. 6, pp. 623–642, 2019.

[26] A. Ristovski and M. Gusev, "SaaS solution for ECG monitoring expert system," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2017, pp. 308–313.

[27] E. Shaqiri, M. Gusev, L. Poposka, and M. Vavlukis, "Correlating HRV to glucose levels," University of Sts Cyril and Methodius, Skopje, Macedonia, Tech. Rep. FCSE-2020-MG07, 2020.

Analysis of Diabetes Dataset

L. Beqiri *, A. Velinov *, B. Fetaji **, L. Loku **, A. Buçuku *, Z. Zdravev *
* University Goce Delcev, Faculty of Computer Science, Stip, Macedonia
** Mother Teresa University, Skopje, Macedonia
e-mail address: lavdim.beqiri@gmail.com, aleksandar.velinov@ugd.edu.mk, bekim.fetaji@unt.edu.mk, lindita.loku@unt.edu.mk, agon.bucuku@gmail.com, zoran.zdravev@ugd.edu.mk

Abstract - **The focus of the research study was analysis of diabetes dataset and how it will perform if we try to do a prediction of diabetes with different machine learning algorithms. We used the original dataset from the National Institute of Diabetes, and Digestive and Kidney Diseases. The dataset can be used to predict whether or not a patient has diabetes, based on certain diagnostics. For analysis we used Amazon Web Services. We used AWS S3 service to store our dataset, and Amazon Sagemaker to perform an analysis. For the given dataset we applied three classification models: Logistic Regression Model, K-nearest Neighbors and Support Vector Machines. For each of the models we also performed a performance measurement. We also compared all the results we got and according to the results, Support Vector Machines has the best performance. Insights and recommendations are provided.**

Keywords – Diabetes Dataset, Machine Learning Algorithms, Amazon Sagemaker, Logistic Regression Model, K-nearest Neighbors, Support Vector Machines.

I. INTRODUCTION

The exponential growth of data has created a new area of interest in technology and business called “Big Data". To be classified as Big Data, a data set or business problem must have data that is so vast, fast or complex that it becomes impossible to store, process, and analyze using traditional data storage and analytics applications [1]. Volume is only one of the criteria because the need for real time processing of the data (also called data in motion) or the need for integrating structured and unstructured data may qualify the problem as Big Data problem [2]. The three characteristics of Big Data are: data itself, analytics and analytics result. Big Data technologies as a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis. Referring to the book by Kalyvas and Overly, the “Big Data” refers to datasets whose size is beyond the ability of typical database software tool to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big dataset needs to be in order to be considered Big Data, don’t define Big Data in terms of being larger than a certain number of terabytes [3]. Looking at the trends of technological growth and data growth, he thought that terminology would change in terms of the number of sizes that determine what Big Data is. Big Data is all about analytics on a broader spectrum of data, and therefore represents an opportunity to create even more differentiation among industry peers. Real value can only emerge from a consumable analytics platform that saves you from having to build applications from scratch [4]. In Big Data, we can use the modeling process of large datasets, which uses statistical and artificial intelligence methods. Data mining makes the information structure readable for use in the future. This is done through the process of analyzing the content of the model, and relationships between variables, validating the findings and applying patterns. The use of classification is to extract the model and to describe the classes. For better understanding, the data at large is used for analysis help. Classification predicts categorical, labels, models continuous valued functions. Classification organize and categorize data in distinct classes.

The focus of this paper is to apply different classification algorithms. For the given dataset we applied three classification models: Logistic Regression Model, K-nearest Neighbors and Support Vector Machines.

Using Big Data should allow us to understand the cause of the disease and to help develop the latest therapies. Personalized medicine has great potential to make treatments more effective by reducing side effects, but it requires access for researchers to Big Data in order for it to progress.

According to the World Health Organization (WHO), diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. [5]. There are three main types of diabetes mellitus. Type 1 and type 2 diabetes both occur when the body cannot properly store and use glucose, which is primary energy source [6]. Sugar, or glucose, collects in the blood and does not reach the cells that need it, which can lead to serious complications. Type 1 diabetes usually begins in childhood or early adulthood; but its onset can occur in adults. In type 1 diabetes, a person with a genetic predisposition who is exposed to a precipitating event, such as a viral infection, experiences autoimmune destruction of the beta cells. Type 2 diabetes is more likely to appear as people age, but many children are now starting to develop it. In this type, the pancreas produces insulin, but the body cannot use it effectively. There has been a substantial increase in the number of cases of type 2 diabetes diagnosed in young children. [7].

II. RELATED WORK

Analysis of dataset is widely applied and used in various fields, such as weather forecasting, GPS, target marketing, engineering, and medical diagnosis. By analyzing a given dataset, we can extract knowledge from it. In this section, we will present some related work which used medical datasets.

An analysis of diabetes dataset was conducted by Alahmar, Mohammed and Benlamri [8]. The dataset they used includes total 9 features which have numerical and nominal attributes. They targeted to make a classifier system by using several ML techniques. In that analysis, they found 79.13% accuracy using Support Vector Machine (SVM).

Similarly, Mir and Dhage applied ML techniques to classify diabetes dataset [9]. They applied a few machine learning (ML) algorithms, in particular: Naive Bayes, SVM, Random Forest (RF) and Simple CART. Among these algorithms, they found 76.5% accuracy using RF in terms of classification of diabetes dataset.

Verma and Mishra conducted a study to identify diabetes cases by using several ML algorithms: Naive Bayes, J48, Sequential Minimal Optimization (SMO), MLP and Reduced Error Pruning Tree (REP-tree) algorithms [10]. In their study, they found the best accuracy of 76.80% using SMO algorithm.

III. CLASSIFICATION

Classification is a technique for identifying and grouping data in such a way that based on a value of the target attribute, the entire dataset can be qualified to belong to a class. This is one of the techniques used in data mining to identify the data behavior patterns [11].

Classification is two-step process: I) Learning or training step where data is analyzed by a classification algorithm. II) Testing step where data is used for classification and to estimate the accuracy of the classification [12].

A. Model 1: Logistic Regression Model

Logistic Regression is the most common method used for binary classification problems (problems with a 1 or 0 outcome) [13].

The function used in Logistic regression is the logistic function. The logistic function is an S-shaped curve that can take any number and map it into a value between 0 and 1. The equation of the Logistic Function is as follows:

$$\frac{1}{(1 + e^{-value})}$$

Where e is the base of the natural logarithms (Euler's number or the exp() function) and value is the actual numerical value that you want to transform. In Figure 1 we can see a plot of numbers on the x-axis transformed into the range 0 and 1 on the y-axis using the logistic function.

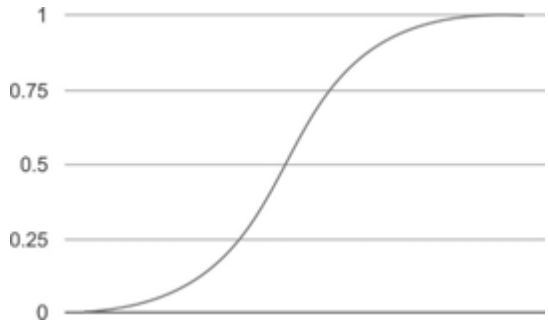


Figure 1. Logistic Function

How is the logistic function used in logistic regression? Input values are combined linearly using coefficient values to predict an output value. How this differs from linear regression is that it is used to model a binary value (0 or 1) instead of a numeric value.

Below is the logistic regression equation:

$$probability = \frac{1}{(1 + e^{-(b_0 + b_1(x))})}$$

b_0 is an intercept term and b1 is a coefficient for the input value (x). Each variable in our input data has a b coefficient that is calculated using the training dataset.

As an example, if we are modelling the probability of a patient having diabetes from a single factor (say blood-glucose level). A logistic regression model could be written as the probability of having diabetes given a person's height blood-glucose level (x).

The above equation can be re-written as follows:

$$ln\left(\frac{probability}{1 - probability}\right) = b_0 + b_1 x$$

The calculation on the right is a linear equation (like linear regression), and the input on the left is a log of a ratio including the probability.

This ratio $\frac{probability}{1-probability}$ is referred to as "**the odds**" of the default class. We could write the above equation as follows:

$$ln(odds) = b_0 + b_1 x$$

usually referred to as ***log odds***

Because we are interested in using the logistic function for a classification problem (deciding if something is 1 or 0), we don't use the probabilities directly. We can convert the probabilities to a binary class value, for example:

- If probability < 0.5 assign: 0
- If probability >= 0.5 assign: 1

B. *Model 2: K-nearest Neighbors Algorithm*

In order to understand the algorithm, it is useful to take a graphical example. Say we have a distribution of 1's and 0's on a plane as shown below. Now say we have a red square on the same plane and don't know whether it belongs to class 1 or 0.



Figure 2. K nearest neighbours

In the K nearest Neighbors algorithm (Figure 2) the "K" is the amount of near-by points we want to check [14]. Let's take an example of K=3, we would make a circle with the red square at the center and increase the size of the circle until it surrounds 3 data points (Figure 3).

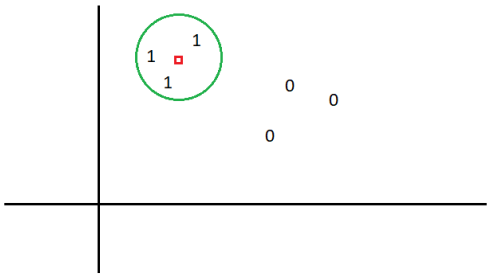


Figure 3. K nearest neighbours (3 points enclosed of class 1)

As we can see, all 3 points which have been enclosed are of class 1. Intuitively we can deduce that our unknown point (the red square) is of class 1.

C. *Model 3: Support Vector Machine Algorithm*

A Support Vector Machine is another algorithm that can be used in classification problems.

SVMs use the concept of finding a dividing boundary that splits the dataset into two seperate classes. This dividing boundary is referred to as a hyperplane. This is shown graphically in Figure 4.

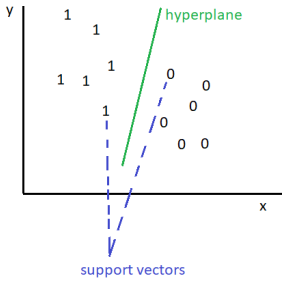


Figure 4. Support Vector Machine

- Support Vectors - The points of a data set that are closest to the dividing boundary as usually referred to as the "support vectors". These "support vectors" are considered to be the critical elements of a data set. If any of these points were removed, the position of the hyperplane would change significantly.
- What is a hyperplane? - Once the hyperplane has been set, if we introduce new testing data, the further the data point is from the hyperplane, the more confident we are that the point belongs to the class represented by that side of the hyperplane.
- How do we find the right hyperplane? - The margin is defined as: The distance between the hyperplane and the nearest data point. To achieve the best chance of new data being classified correctly, choosing a hyperplane with the greatest possible margin between the hyperplane and any point in the training set (Figure 5).

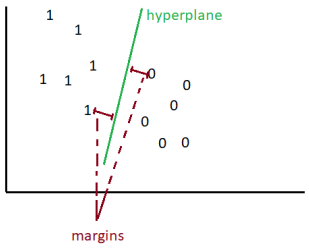


Figure 5. Support Vector Machine (hyperplane, margins)

IV. METHODOLOGY

Our research dataset is divided into two parts, two-thirds of the data is used as a training set, and one-third of the dataset is defined as a testing set to evaluate the performance of several classifiers. All classifiers were fitted to the same training and testing data. The specific process is shown in Figure 6.

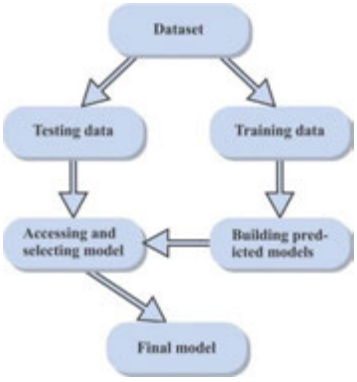


Figure 6. Methodology (flowchart)

A. *Dataset*

In order to analyze the factors contributing to the presence of diabetes, we sourced a dataset (Figure 7). This original dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset can be used to predict whether or not a patient has diabetes, based on certain diagnostics e.g. glucose, insulin levels in the dataset (Table I).

A	B	C	D	E	F	G	H	I
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0

Figure 7. The dataset (contains 768 lines)

The dataset consists of 8 medical independent variables (pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index – BMI, diabetes pedigree function and age) and one target dependent variable (outcome). "Outcome" is a binary target variable that has a value of 1 for diabetes and 0 for no diabetes. 268 of the 768 are 1, the others are 0. All records for patients in the dataset are females greater than 21 years old of Pima Indian heritage [15].

TABLE I. INDEPENDENT VARIABLES

1.	**Pregnancies**	Number of times pregnant
2.	**Glucose**	Glucose concentration a 2 hours in an oral glucose tolerance test
3.	**BloodPressure**	Diastolic blood pressure (mm Hg)
4.	**SkinThickness**	Triceps skin fold thickness (mm)
5.	**Insulin**	2-Hour serum insulin (mu U/ml)
6.	**BMI**	Body mass index (weight in kg/(height in m)^2)
7.	**DiabetesPedigreeFunction**	Diabetes pedigree function
8.	**Age**	Age (years)

B. *Tools used*

We used the S3 service from AWS that allow creating a bucket to store the dataset. In order to analyze the dataset we used Amazon SageMaker Machine Learning service. The tools used are Jupyter notebooks using Python code.

V. RESULTS OF CLASSIFICATION ALGORITHMS USING AWS

In order to see if our model gives good results at identifying 1's and 0's we check the confusion matrix (Table II and Table III).

TABLE II. CONFUSION MATRIX

True Negative	False Negative
False Positive	True Positive

The Receiver Operator Characteristic (ROC) curve is a graphical output which is used to analyse the accuracy of how good the model is at separating 1's and 0's. The higher the Area Under the Curve (AUC) value, the better the model is at predicting classification problems (predicting a 1 or 0 outcome). The area under the curve represents how 0s as 0s and 1s as 1s.

An excellent model has an AUC near to the 1. When the AUC is 0.5, it means that the model has no ability to separate 1's and 0's.

TABLE III. CONFUSION MATRIX (SIMPLER TERM)

Model predicts a 0: actually a 0 in the dataset	Model predicts a 0: actually a 1 in the dataset
Model predicts a 1: actually a 0 in the dataset	Model predicts a 1: actually a 1 in the dataset

Numerically, the area under the curve value equal to 0.9 represents a 90% probability that the model can distinguish between 1 and 0. The further the ROC curve is from the center straight line, the better in terms of accuracy. The area under this curve is a numerical measure of the test accuracy.

A. *Model 1: Logistic Regression Model*

The results of diabetes dataset for Logistic Regression Model classifier using AWS are shown in Table IV.

TABLE IV. MODEL 1 RESULTS

True Negative: 445	False Negative: 55
False Positive: 112	True Positive: 156

A more visual way to measure the performance of a binary classifier is the ROC curve (Figure 8). It is created by plotting the true positive rate (TPR) (or recall) against the false positive rate (FPR).

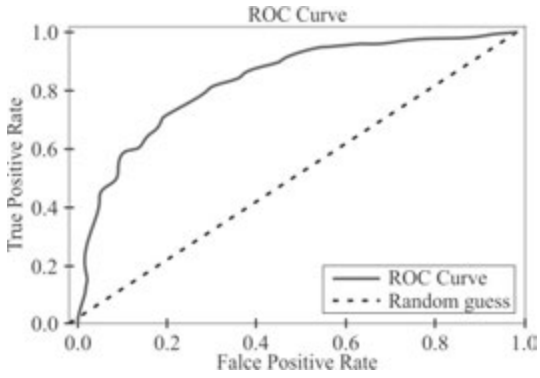


Figure 8. Model 1 - ROC Curve

As we can see the curve is close to the left-hand border and the top border. This indicates that it is quite good at predicting 1's as 1's and 0's as 0's.

We get the value for AUC of **0.83** which is good.

B. *Model 2: K-nearest Neighbours Algorithm*

The results of diabetes dataset for K-nearest Neighbors Model classifier using AWS are shown in Table V.

TABLE V. MODEL 2 RESULTS

True Negative: 442	False Negative: 58
False Positive: 93	True Positive: 175

Figure 9 shows the ROC curve for the K-nearest Neighbors Model.

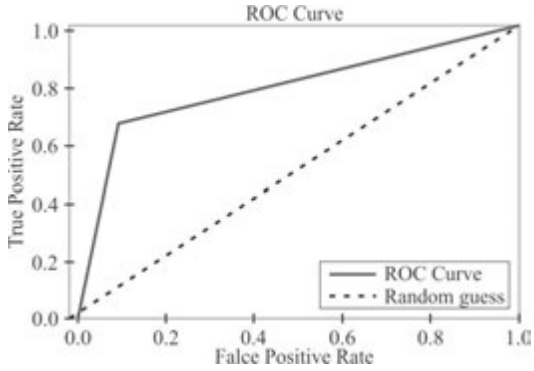


Figure 9. Model 2 - ROC Curve

At a first glance, it does not seem as accurate as the Logistic Regression model. We look at the AUC to check this numerically.

We get a value of **0.768** compared to our result for Logistic regression which yielded a value of 0.839 so this algorithm doesn't perform as well.

C. Model 3: Support Vector Machine Algorithm

The results of diabetes dataset for Support Vector Machine Model classifier using AWS are shown in Table VI.

TABLE VI. MODEL 3 RESULTS

True Negative: 500	False Negative: 0
False Positive: 0	True Positive: 268

The ROC curve for the model is shown in Figure 10. As for previous models, we will analyze the ROC curve and look at the AUC measurement.

We can see that the curve follows the left hand border and the top very closely which indicates excellent accuracy.

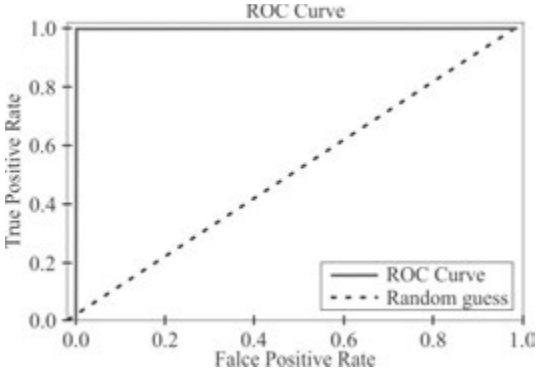


Figure 10. Model 3 – ROC Curve

We get a value of **1.0** which indicates perfect prediction.

VI. DISCUSSION

The results of applying different classification algorithms on diabetes dataset are shown in Table VII, which consists of correctly classified and incorrectly classified instances.

From this study we can see the performance of the three predictive models. For the testing dataset, the final comparative analysis results demonstrated that the Support Vector Machine algorithm proved the best, with an accuracy of 100%, and the results of the Confusion Matrix were True Negative: 500, False Negative: 0, False Positive: 0, and True Positive: 268. The Logistic Regression Model algorithm came out to be the second best with a classification accuracy of 83%, and the results of the Confusion Matrix were True Negative: 445, False Negative: 55, False Positive: 112, and True Positive: 156. Finally, the K-nearest Neighbors algorithm came last with an accuracy of 76.8%, and the results of the Confusion Matrix were True Negative: 442, False Negative: 58, False Positive: 93, and True Positive: 175.

In the results, the area under the receiver operating Characteristic (ROC) curve, Area Under The Curve (AUC) value of the SVM, Logistic Regression Model, and K-nearest Neighbors Algorithm were 1.000, 0.830, 0.768, respectively. The area under ROC curve for SVM method is 100%, showing a high reliability of classification capability among all the methods. Overall, the methodology presented in this paper has shown the good classification performance of health management control satisfaction of patients with diabetes.

VII. CONCLUSION

In this paper, we have classified the data of the diabetes dataset from the National Institute of Diabetes. We applied three classification techniques: Logistic Regression Model, K-nearest Neighbors Algorithm and Support Vector Machine. According to the results, Support Vector Machine is the best algorithm for classification. The dataset that we used in this study is small. In the following research work, we plan to apply algorithms to classify a larger dataset using three different platforms: Amazon Web Services, Microsoft Azure and Google Cloud. In this way we can get statistics, determine the time spent on analysis and compare algorithms.

TABLE VII. RESULTS OF DIFFERENT CLASSIFICATION ALGORITHMS

Algorithms	**Value of AUC**	**Confusion Matrix**	
Logistic Regression Model	0.830	True Negative: 445	False Negative: 55
		False Positive: 112	True Positive: 156
K-nearest Neighbours Algorithm	0.768	True Negative: 442	False Negative: 58
		False Positive: 93	True Positive: 175
Support Vector Machine Algorithm	1.00	True Negative: 500	False Negative: 0
		False Positive: 0	True Positive: 268

REFERENCES

[1] C. Razbonyalı and E. Güvenoğlu, "Traditional Data Storage Methods and the Big Data Concepts", International Research

Journal of Engineering and Technology (IRJET), vol. 03, pp. 2556-2561, June 2016.

[2] Y. Riahi, “Big Data and Big Data Analytics: Concepts, Types and Technologies”, International Journal of Research and Engineering, vol. 5, pp. 524-528, September 2018.

[3] J. R. Kalyvas, M. R. Overly, Big Data A Business and Legal Guide. CRC Press, Auerbach, 2015.

[4] P. Zikopoulos, D. Deroos, K. Parasuraman. T. Deutsch, D. Corrigan, J. Giles, Harness the Power of Big Data, The McGraw-Hill Companies, 2013.

[5] G. Shahi, B. Luu (Editors), Emerging Trends in Global Health, Global Health Review and GBI Books, 2008.

[6] K. Moscou, K. R. Snipe, Pharmacology for pharmacy technicians, Elsevier, 2018.

[7] K. Cooper, K. Gonsell, Foundations and Adulht Health Nursning, Mosby, 2015.

[8] A. Alahmar, E. Mohammed., R. Benlamri, “Application of data mining techniques to predict the length of stay of hospitalized patients with diabetes”, 4th International Conference on Big Data Innovations and Applications (Innovate-Data), Barcelona, Spain, pp. 38-43, 2018.

[9] A. Mir, S. N. Dhage, “Diabetes disease prediction using machine learning on big data of healthcare”, Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, pp. 1-6, 2018.

[10] D. Verma, N. Mishra, “Analysis and predicion of breast cancer and diabetes disease datasets using data mining classification techniques”, International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 533-538, 2017.

[11] S. Gollapudi, Getting Started with Greenplum for Big Data Analytics, Packet Publishing Ltd. 2013.

[12] J. Han, M. Kamber, Data Mining; Concepts and Techniques, Morgan Kaufmann Publishers, 2000.

[13] R. P. Anderson, R. Jin, and G. L. Grunkemeier, “Understanding logistic regression analysis in clinical reports: an introduction”, The Annals of thoracic surgery, vol. 75, pp. 753-757, March 2003.

[14] P. Cunningham and S. Delany, k-Nearest neighbour classifiers. Mult Classif Syst, 2007.

[15] (2017) Pima Indians Diabetes Database, [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database

Trends from Minimally Invasive to Non-invasive Glucose Measurements

M. Gusev*, L. Poposka**, E. Guseva***, M. Kostoska*, B. Koteska*,
M. Simjanoska*, N. Ackovska* and A. Stojmenski*
*Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering, Skopje, North Macedonia
**Ss. Cyril and Methodius University, University Clinic of Cardiology, Skopje, North Macedonia
***Innovation Dooel, Skopje, North Macedonia
E-mail: {marjan.gushev, magdalena.kostoska, bojana.koteska, monika.simjanoska, nevena.ackovska,
aleksandar.stojmenski} @finki.ukim.mk, lidijapoposka@gmail.com, ema.guseva@innovation.com.mk

***Abstract*—Approximately 7% of the elderly over 45 are diagnosed in several diabetes forms and it is believed that 3% of the population is undiagnosed. As the number of diabetes diagnoses has been increasing, so too has the technology for managing the disease. The traditional way of measuring glucose levels is to apply a blood drop to a chemically treated, disposable "test-strip" and insert it into an electronic device. In this study, we overview the latest technology for continuous glucose measurement, as they trend from minimally invasive to non-invasive glucose measurement methods.**

***Index Terms*—continuous glucose measurement, wearable devices, electrocardiogram, heart rate variability, machine learning, neural networks,**

I. Introduction

Diabetes is documented to cause eye, kidney and nerve damage when glucose levels get out of the normal range [1]. Therefore, it is extremely important for diabetes patients to control their glucose levels. Most of the existing glucose monitoring techniques are invasive, which is a burden when done on a frequent basis [2]. Specifically, a blood drop is required for reading blood glucose levels, so patients must use invasive means of getting a blood drop every time they want to measure [3].

In addition, routine finger pricking is troublesome for diabetic patients because it can lead to scarring. It is no surprise then that the idea that glucose measurement can be done in a non-invasive way surpasses the wildest dreams of diabetics, especially those on insulin therapy since they use glucometers every couple of hours.

The cost of each glucose reading encompasses the invasively-collected blood drop to include costs associated with the supplies they need for self-monitoring, including blood glucose testing strips (as high as $1 per strip), sensors for continuous glucose monitoring ($350 a month), glucagon ($280) and doctor visits [4]. Along with expensive insulin shots, the math adds up to average medical expenditures of $16,752 per person per year according to the American Diabetes Association [4].

Given the dream of managing diabetes in a cheap and non-invasive way, many scientific efforts have been oriented toward finding new precise non-invasive techniques for glucose level measurement. One of the fruits of this effort is Technology for Continuous Glucose Monitoring (CGM), which is a technique that allows medical devices to measure glucose levels [5], [6], [7], and it can be invasive, minimally invasive and non-invasive. The biggest benefiters of CGM are diabetics that need frequent glucose level monitoring. And they are likely to be the early adopters to wearable sensors in a market whose potential is estimated to be the size of the market for glucometers, which was 613$ million in 2016 and is expected to reach 915$ million by 2021 [8].

From a user point of view non-invasive measurement methods are superior to invasive or minimally invasive methods as they provide diabetics better tools to manage their disease with the added opportunity for remote real-time continuous monitoring.

Many recent review studies of available technology cover current trends and present minimally invasive and non-invasive methods [9], [10], [11], [12], [13], [14], [3], [15], [16], [17], or focus on recent technologies for continuous monitoring [18], [3], [7], [5], [6], [19] or use of wearable sensors [20], [19], [21], [16].

In our earlier paper [17], we gave a comprehensive overview of non-invasive glucose measurement methods, and in this paper, we focus on the analysis of recent trends on using wearable sensors for *continuous glucose monitoring* presenting the available methods and ongoing projects towards minimally invasive and non-invasive glucose measurement.

The paper is organized following the next structure. Section II defines the essential terms for continuous monitoring, wearable sensors, minimally invasive techniques and non-invasive methods, and gives background on the development of generations of medical devices for glucose measurement. Related work on similar review papers is presented in Section III. Available minimally invasive techniques and continuous glucose measurement and non-invasive methods are elaborated correspondingly in Section IV and Section V. Section VI concludes the review and gives future trends.

II. Background

Methods for glucose measurement can be invasive, minimally invasive and non-invasive [22], [23], [16]. Invasive glucose measurement medical devices are usually implanted in

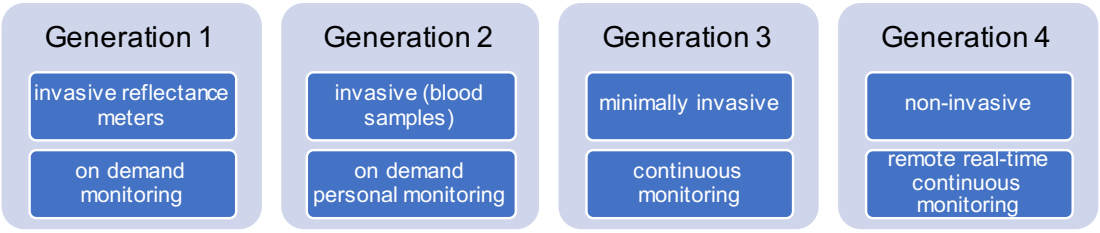


Fig. 1. Application domains and methods for glucose measurement generations

the human body to enable continuous measurement, or based on a blood sample, usually obtained by finger pricking. Clarke and Foster [22] analyze the evolution of blood glucose meters towards more sophisticated devices.

Klonoff [18] elaborates on the roadmap of continuous glucose measurement as an initiator of the next generation of minimally and non-invasive techniques. Minimal invasive methods can be portable or wearable devices, mostly realized with a set of small micro-needles attached to the skin or a small part inserted underneath the skin. Non-invasive methods do not invade the human body and are based on various techniques, including spectrometry, or analysis of other parameters correlated to glucose levels [13]. Several other non-invasive methods include analysis of certain physiological fluids, such as urine, sweat and saliva. In addition to these non-invasive methods, this paper presents new trends to detect glucose levels through analysis of electrocardiogram (ECG) or heart rate variability (HRV) parameters.

Abd Salam et al. [23] summarize four generations of glucose monitoring. The evolution of glucose measurement medical devices is illustrated in Figure 1 whereby different generations of measurement devices are classified according to used technology. Commercial glucose measurement commenced in the 1970s with the first-generation glucose meters, which used reflectance technology and were made as heavy devices that require relatively big amounts of blood. Second generation devices used a drop of blood and were made as smaller devices with affordable prices, that allowed personalized use. The third-generation devices started as minimally invasive devices that include an array of small needles on the skin and enable continuous monitoring.

Recently, a new generation is rising on the horizon, although it is still a kind of an alternative, rather than an actual application of this kind of medical devices, due to its early stage of development. Nevertheless, we will refer to it as the *fourth-generation of glucose monitoring* which includes wearable medical devices using non-invasive methods and provides an environment for remote and real-time continuous monitoring.

III. Related work of reviews on non-invasive glucose measurement

Even though invasive techniques are the most common and widely used method for glucose level measurement, a lot of research and development activities for non-invasive techniques is ongoing. Hence, this section presents a short overview of review articles that target non-invasive or minimally invasive methods of glucose measurement, commenting on their use for continuous measurement.

To begin with, glucose biosensor technology consists of point-of-care devices, continuous glucose monitoring systems and non-invasive glucose monitoring systems. Montagnana et al. [10] analyze self-monitoring of blood glucose, also known as point-of-care devices or portable glucose monitors, that play a key role in preventing complications in the short, medium and long term.

Laboratory techniques mainly use a drop of blood and sophisticated techniques based on chemical reactions, while the home-monitoring techniques are mainly self-monitoring (activated on demand) and CGM minimally invasive continuous glucose monitoring (CGM), which continuously measure glucose levels [16].

The devices that measure glucose do so by analyzing several psychophysiological properties and sensing chemical, electrical, optical, thermal, acoustic properties or any combination of them, as concluded by Garcia-Guzman et al.[24], or by analyzing autonomic dysfunction properties, concluded by Gusev et al. [17].

Generally, Villena Gonzales et al. [16] present a comprehensive overview of the progress of glucose measurement. Wearable Continuous Glucose Monitoring (CGM) sensors were overviewed by Cappon et al. [19] including wearable or minimally-invasive devices that measure glucose concentration by exploiting different physical principles, such as glucose-oxidase, fluorescence, or skin dielectric properties. An analysis of various optical non-invasive methods based on different spectrometry is overviewed in [12].

Bruen et al. [20] highlight recent advances towards non-invasive and continuous glucose monitoring devices, with a particular focus on wearable devices for monitoring glucose concentrations in physiological fluids alternative to blood. Similarly, Salam [23] classifies two categories of non-invasive measurements based on fluids (intermittent) using sweat, urine, saliva, and tears; and tissues (continuous) using skin, aqueous, humours of the eye, oral mucosa, tongue and tympanic membrane. In addition, the techniques are classified as optical (using near-infrared, mid-infrared, Raman, thermal emission or photoacoustic spectroscopy, or polarimetry), and transdermal (using reverse iontophoresis and bio-impedance). The classification introduced in this survey, as shown in Figure 2,

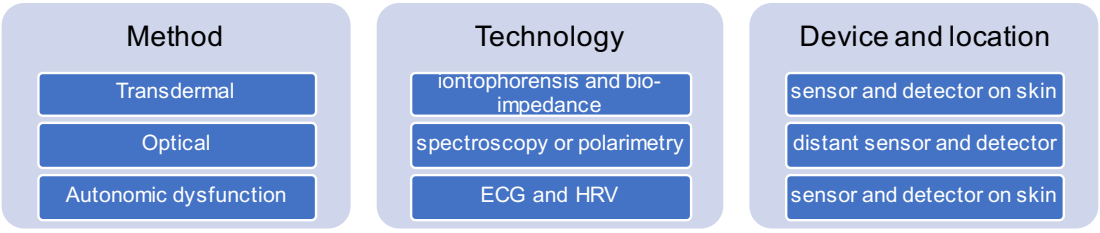


Fig. 2. Non-invasive glucose measurement

adds another class based on the detection of changes in the autonomous nerve system.

Notably, and different from the content of other reviews of non-invasive technology, Gusev et al. [17] present noninvasive glucose measurement using Machine Learning and Neural Network Methods pointing out the possibility to use the correlation with Heart Rate Variability. This research is not standalone. Maser and Lenhard [25] conclude that those with parasympathetic dysfunction have a high resting heart rate likely because of vagal neuropathy that results in unopposed increased sympathetic outflow, and that persons with a combined parasympathetic/sympathetic dysfunction have slower heart rates. They state that it is apparent that the determination of heart rate itself is not a reliable diagnostic sign, and that reduction in heart rate variability is the earliest indicator of diabetes.

Finally, we end our review of related works by analyzing existing limitations. Lin et al. [15] review past and current non-invasive glucose monitoring techniques as well as devices in progress, summarizing the major challenges to be confronted. They discuss the weakness in glucose specificity, sensitivity, physiological time lag, calibration process and human factors. Special attention is given to usability, user experience and applicability for home use. Similarly, Yoo and Lee [11] give an overview and conclude the need for significant improvements to this technology with several challenges to accurate and reliable glucose monitoring.

IV. Available Minimally Invasive and Continuous Glucose Measurements

The main technology developers are summarized in Table I, along with their sensors and methods of analysis. To our knowledge, there is no evidence of commercially successful solution [3]], only evidence of research activities to develop continuous blood glucose monitoring based on wearable sensors.

Most of these solutions are not continuous real-time medical devices, meaning that the measurements are activated on-demand, without real-time monitoring possibility. A more comprehensive measurement method will be supported by a software agent using Big Data concepts of incoming streams of health-related vital parameters, as a kind of a technological value proposition.

Given the importance of continuity of measurement, there are multiple systems that use Continuous Glucose Monitoring (CGM) as a method for continuous monitoring of glucose levels throughout the day. CGM systems take glucose measurements at regular intervals, 24 hours a day, and translate the readings into dynamic data, generating glucose direction and rate of change reports.

Mainly, today, CGM systems use minimally invasive techniques. Villena Gonzales et al. give a comprehensive report about the progress of glucose monitoring and review existing invasive, minimally invasive and non-invasive techniques [16]. They analyze glucose measurement devices using minimally invasive methods based on electric techniques (ultrasound, sonophoresis and reverse iontophoresis), and nanotechnology (surface plasmon resonance [26] or others, such as detecting the minimal concentration ranges [27]).

Techniques used for minimally invasive methods include:

- *Transdermal* method which uses skin as a media to sense various properties, including the following techniques:
  - *Reverse iontophoresis* measures the effect caused by a low electric current across the skin between two electrodes [28].
  - *Bio-impedance spectrometry* using a surface impedance probe that penetrates through the skin but not into the dermis to measure the impedance frequency spectrum [29].
  - *Skin suction blister* uses vacuum suction over a small skin area [30], and
  - *Sonophoresis* uses ultrasound to access interstitial fluid that transports glucose [31].

Iontophoresis uses electric current to transport ions into tissues through the skin while Phonophoresis uses acoustic energy (ultrasound) to drive molecules into the tissue.

V. Non-invasive Glucose Measurements

Recently, several ideas for non-invasive glucose measurement are patented, and only a few appeared on the market, but are yet to become commercially successful products [3]]. In particular, several R&D projects reported the use of these non-invasive methods for glucose monitoring. Non-invasive glucose measurement is mainly based on transdermal, optical and autonomic dysfunction methods (Figure 2).

The transdermal method using path-selective and specific glucose monitoring via a graphene-based platform was reported in [32], using Raman spectrometer suitable for home-use non-invasive glucose monitoring in [33], and a sensor using depth-selective photothermal IR spectroscopy of skin

TABLE I
COMPARISON OF AVAILABLE PRODUCTS FOR CONTINUOUS GLUCOSE MEASUREMENT

Available product	Accuracy	Features	Description	Requirements
Medtronic Guardian sensor 3 and CareLink line	13.6%	7-day lifetime, trend arrows, rate-of-change alerts, hyper and hypo alarms, direct integration with Medtronic insulin pumps	A continuous glucose monitoring system and software that allows patients to upload data from diabetes devices to a secure online database. The healthcare provider then downloads data for analysis.	Calibration on 12 h
Abbott FreeStyle Libre	11.4%	14-day lifetime, trend arrows, communication with a smart device	Eliminates the need for routine finger pricks and rather monitoring glucose levels through a sensor that can be worn on the back of upper arm for up to 14 days.	To read glucose values the sensor needs to be scanned with the receiver or the smartphone
DexCom G4 Platinum/ G5	9%	7-day lifetime, trend arrows, rate-of-change alerts, hyper and hypo alarms, remote monitoring, direct wireless communication with smart devices (up to 5 devices)	Is an externally-worn glucose sensors that continuously measure and display glucose values. The device supports trending information reporting and real-time capabilities for up to seven days.	Calibration on12 h
Agamatrix WaveSense Blood Glucose Meters		WaveSense Blood Glucose Meters are an invasive method for measuring glucose and are therefore more dangerousto the user. Morever the need for a constant re-supply of the test strips restricts the user	Is the first FDA-cleared blood glucose monitor to connect directly to the iPhone and it provides estimated haemoglobin A1c values as well.	
Senseonics Eversense	11.4%	90-day lifetime, trend arrows, rate-of-change alerts, hyper and hypo alarms, communication with a smart device	This system includes a pill-sized sensor implanted in the upper arm for 90 days, an on-body transmitter. Eversense's implant is enabling longer measuring time than the other competitors	The sensor needs to be inserted and removed in doctor's office

in [34]. Also, non-invasive ultrasound or spectroscopy (light) technology was used with GlucoTrack [35]. Glucose noninvasive monitoring using mid-infrared absorption spectroscopy based on a few wavenumbers was reported in [36]. Methods using sweat-based glucose monitor biosensors reported in [37] and [38]. A theoretical study of resonant photoacoustic spectroscopy for glucose noninvasive detection was reported in [39].

Optical (or electromagnetic) non-invasive methods are based on the effects that blood glucose levels produce on the skin and react on various electromagnetic waves by measuring the spectrum of

- reflections on radiated signal,
- scattering effects, and
- human body emission of electromagnetic waves.

The other non-invasive methods are those that analyze the electromagnetic properties of other matter effected by glucose concentration, including:

- Interstitial fluid,
- Urine,
- Sweat,
- Saliva,
- Ocular fluid, and
- Breath acetone.

In addition, Gusev et al. [17] analyze autonomic dysfunction as another class, as they conclude that the functionality of the autonomous nerve system which controls glucose levels can be measured by characteristics of an electrocardiogram (ECG) or various heart rate variability (HRV) parameters.

The non-invasive glucose measurement methods (classified in Figure 2) include the following methods:

- *Optical spectroscopy methods* which measures spectra of radiated signal fluctuations realized in the:
  - *Infrared* range [40],
  - *Near-Infrared (NIR)* range [9],
  - *MidInfrared (MIR)* range [41],
- *Scattering-based optical methods* that include:
  - *Scattering* of the reflected scattered waves [42],
  - *Occlusion spectroscopy* of the arterial flow [43],
  - *Raman spectroscopy*, that uses a specific laser [44],
  - *Photoacoustic spectroscopy* with ultrasonic pulsating waves [45],
  - *Ocular spectroscopy* of hydrogel-bound contact lens reflections [46],
- *Thermal-based methods* including:
  - *Thermal emission spectroscopy* of naturally emitted infrared signals [47],
  - *Temperature-regulated localized reflectance* of localized reflected light [48],
  - *Metabolic heat confirmation* with thermal and optical sensors [49],
  - *Far-Infrared (FIR) spectroscopy* based on skin temperature variations and MIR light scattering [41].
- *Electromagnetic sensing* realized by:
  - *Terahertz spectroscopy* of radiated absorption [50],
  - *Millimeter and microwave Infrared spectroscopy* of reflections by a NIR antenna [51],
  - *Ultrasound technology* that measures the time to propagate through the extracellular fluid [52].
- *Other optical methods* include the following:
  - *Polarimetry method* of optical rotary dispersion of polarized light [53]

  - *Fluorescence method* of fluorescence effects of tears [54],
  - *Optical coherence tomography* of the coherent light scattering [55],
  - *Kromoscopy* that uses human color perception of NIR light [56]
- *Electrical conductivity*
  - *Electromagnetic sensing* of the conductivity of dielectric parameters [57].
  - *Iontophoresis* measuring the electrical properties of the interstitial fluid [58], and
  - *Impedance spectrometry* of the tissue [59].
  - *Bio-impedance* based on the resistive and capacitive impedance of body [60]
- *Autonomic dysfunction*
  - *ECG and HRV-based method* based on correlating with the HRV parameters [61], [17].

VI. Conclusion

In this paper we have presented trends towards continuous and non-invasive techniques for glucose measurement. The problems initiated with constant finger pricking and need for continuous monitoring motivate researchers to find methods and companies to develop non-invasive solutions for reliable continuous monitoring with wearable devices.

We have presented the latest technology and products, which are mainly based on minimally invasive techniques. In addition, we have concluded that the status of medical devices based on non-invasive methods is still not sufficient, due to several problems, including delayed effects of glucose on the measured parameter, calibration of the device, precision, and reliability.

As a summary, the methods include transdermal methods that measure effects on the skin, interstitial fluid or other fluids produced by our body. The methods sense electromagnetic properties, such as spectrography of radiated, reflected or scattered waves, or electrical properties such as impedance and conductivity. Furthermore, some methods include fluorescence, perception, coherence, or autonomic dysfunction by analysis of ECG and HRV.

Current trends show maturity of transdermal and optical methods, although the commercially available products show a large need of calibration and customization in order to have accurate measurements with expected performance. Still there is a lot of research in the field of exploiting the electromagnetic and optical properties that will reveal sufficient information about glucose level. Most promising are the new attempts to use HRV to detect the autonomic dysfunction and ability to control the blood glucose level.

This field remains a hot research topic due to a number of unsolved issues and the increasing rate of diabetes diagnoses, and great effort is devoted to making glucose monitoring less painful for patients and less distracting in their everyday activities.

References

[1] Diabetes Control and Complications Trial Research Group, "The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus," *New England journal of medicine*, vol. 329, no. 14, pp. 977–986, 1993.

[2] National Diabetes Data Group and others, "Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance," *Diabetes*, vol. 28, no. 12, pp. 1039–1057, 1979.

[3] A. Nawaz, P. Øhlckers, S. Sælid, M. Jacobsen, and M. N. Akram, "Non-invasive continuous blood glucose measurement techniques," *Journal of Bioinformatics and Diabetes*, vol. 1, no. 3, p. 01, 2016.

[4] American Diabetes Association, "Economic costs of diabetes in the us in 2017," *Diabetes care*, vol. 41, no. 5, p. 917, 2018.

[5] D. Slattery and P. Choudhary, "Clinical use of continuous glucose monitoring in adults with type 1 diabetes," *Diabetes technology & therapeutics*, vol. 19, no. S2, pp. S–55, 2017.

[6] A. L. Carlson, D. M. Mullen, and R. M. Bergenstal, "Clinical use of continuous glucose monitoring in adults with type 2 diabetes," *Diabetes technology & therapeutics*, vol. 19, no. S2, pp. S–4, 2017.

[7] T. Danne, R. Nimri, T. Battelino, R. M. Bergenstal, K. L. Close, J. H. DeVries, S. Garg, L. Heinemann, I. Hirsch, S. A. Amiel *et al.*, "International consensus on use of continuous glucose monitoring," *Diabetes care*, vol. 40, no. 12, pp. 1631–1640, 2017.

[8] Kalorama Information. (2017) The global market for medical devices, 8th edition. [Online]. Available: https://www.kaloramainformation.com/Global-Medical-Devices-11198459/

[9] C. E. F. do Amaral and B. Wolf, "Current development in non-invasive glucose monitoring," *Medical engineering & physics*, vol. 30, no. 5, pp. 541–549, 2008.

[10] M. Montagnana, M. Caputo, D. Giavarina, and G. Lippi, "Overview on self-monitoring of blood glucose," *Clinica Chimica Acta*, vol. 402, no. 1-2, pp. 7–13, 2009.

[11] E.-H. Yoo and S.-Y. Lee, "Glucose biosensors: an overview of use in clinical practice," *Sensors*, vol. 10, no. 5, pp. 4558–4576, 2010.

[12] C.-F. So, K.-S. Choi, T. K. Wong, and J. W. Chung, "Recent advances in noninvasive glucose monitoring," *Medical Devices (Auckland, NZ)*, vol. 5, p. 45, 2012.

[13] A. Ciudin, C. Hernández, and R. Simó, "Non-invasive methods of glucose measurement: current status and future perspectives," *Current diabetes reviews*, vol. 8, no. 1, pp. 48–54, 2012.

[14] S. K. Vashist, "Non-invasive glucose monitoring technology in diabetes management: A review," *Analytica chimica acta*, vol. 750, pp. 16–27, 2012.

[15] T. Lin, A. Gal, Y. Mayzel, K. Horman, and K. Bahartan, "Non-invasive glucose monitoring: a review of challenges and recent advances," *Curr Trends Biomed Eng Biosci*, vol. 6, pp. 1–8, 2017.

[16] W. Villena Gonzales, A. T. Mobashsher, and A. Abbosh, "The progress of glucose monitoring - a review of invasive to minimally and non-invasive techniques, devices and sensors," *Sensors*, vol. 19, no. 4, p. 800, 2019.

[17] M. Gusev, L. Poposka, G. Spasevski, M. Kostoska, B. Koteska, M. Simjanoska, N. Ackovska, A. Stojmenski, J. Tasic, and J. Trontelj, "Non-invasive glucose measurement using machine learning and neural network methods and correlation with heart rate variability," *Journal of Sensors*, vol. 2020, 2020.

[18] D. C. Klonoff, "Continuous glucose monitoring: roadmap for 21st century diabetes therapy," *Diabetes care*, vol. 28, no. 5, pp. 1231–1239, 2005.

[19] G. Cappon, G. Acciaroli, M. Vettoretti, A. Facchinetti, and G. Sparacino, "Wearable continuous glucose monitoring sensors: A revolution in diabetes treatment," *Electronics*, vol. 6, no. 3, p. 65, 2017.

[20] D. Bruen, C. Delaney, L. Florea, and D. Diamond, "Glucose sensing for diabetes monitoring: recent developments," *Sensors*, vol. 17, no. 8, p. 1866, 2017.

[21] J. Kim, A. S. Campbell, and J. Wang, "Wearable non-invasive epidermal glucose sensors: A review," *Talanta*, vol. 177, pp. 163–170, 2018.

[22] S. Clarke and J. Foster, "A history of blood glucose meters and their role in self-monitoring of diabetes mellitus," *British journal of biomedical science*, vol. 69, no. 2, pp. 83–93, 2012.

[23] N. A. B. Abd Salam, W. H. bin Mohd Saad, Z. B. Manap, and F. Salehuddin, "The evolution of non-invasive blood glucose monitoring system for personal application," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 8, no. 1, pp. 59–65, 2016.

[24] J. García-Guzmán, N. González-Viveros, and H. H. Cerecedo-Núnez, "Comparative analysis of optoelectronic properties of glucose for non-invasive monitoring," in *Emerging Challenges for Experimental Mechanics in Energy and Environmental Applications, Proceedings of the 5th International Symposium on Experimental Mechanics and 9th Symposium on Optics in Industry (ISEM-SOI), 2015*. Springer, 2017, pp. 55–63.

[25] R. E. Maser and M. J. Lenhard, "Cardiovascular autonomic neuropathy due to diabetes mellitus: clinical manifestations, consequences, and treatment," *The Journal of Clinical Endocrinology & Metabolism*, vol. 90, no. 10, pp. 5896–5903, 2005.

[26] D. Li, J. Wu, P. Wu, Y. Lin, Y. Sun, R. Zhu, J. Yang, and K. Xu, "Glucose measurement using surface plasmon resonance sensor with affinity based surface modification by borate polymer," in *2015 Transducers-2015 18th International Conference on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS)*. IEEE, 2015, pp. 1569–1572.

[27] S. Zeng, D. Baillargeat, H.-P. Ho, and K.-T. Yong, "Nanomaterials enhanced surface plasmon resonance for biological and chemical sensing applications," *Chemical Society Reviews*, vol. 43, no. 10, pp. 3426–3452, 2014.

[28] R. O. Potts, J. A. Tamada, and M. J. Tierney, "Glucose monitoring by reverse iontophoresis," *Diabetes/metabolism research and reviews*, vol. 18, no. S1, pp. S49–S53, 2002.

[29] D. Kamat, D. Bagul, and P. Patil, "Blood glucose measurement using bioimpedance technique," *Advances in Electronics*, vol. 2014, 2014.

[30] N. Oliver, C. Toumazou, A. Cass, and D. Johnston, "Glucose sensors: a review of current and emerging technology," *Diabetic Medicine*, vol. 26, no. 3, pp. 197–210, 2009.

[31] J. Kost, S. Mitragotri, R. A. Gabbay, M. Pishko, and R. Langer, "Transdermal monitoring of glucose and other analytes using ultrasound," *Nature medicine*, vol. 6, no. 3, p. 347, 2000.

[32] L. Lipani, B. G. Dupont, F. Doungmene, F. Marken, R. M. Tyrrell, R. H. Guy, and A. Ilie, "Non-invasive, transdermal, path-selective and specific glucose monitoring via a graphene-based platform," *Nature nanotechnology*, p. 1, 2018.

[33] S. M. Lundsgaard-Nielsen, A. Pors, S. O. Banke, J. E. Henriksen, D. K. Hepp, and A. Weber, "Critical-depth raman spectroscopy enables home-use non-invasive glucose monitoring," *PloS one*, vol. 13, no. 5, p. e0197134, 2018.

[34] O. Hertzberg, A. Bauer, A. Küderle, M. A. Pleitez, and W. Mäntele, "Depth-selective photothermal ir spectroscopy of skin: potential application for non-invasive glucose measurement," *Analyst*, vol. 142, no. 3, pp. 495–502, 2017.

[35] T. Lin, Y. Mayzel, and K. Bahartan, "The accuracy of a non-invasive glucose monitoring device does not depend on clinical characteristics of people with type 2 diabetes mellitus," *Journal of drug assessment*, vol. 7, no. 1, pp. 1–7, 2018.

[36] R. Kasahara, S. Kino, S. Soyama, and Y. Matsuura, "Noninvasive glucose monitoring using mid-infrared absorption spectroscopy based on a few wavenumbers," *Biomedical optics express*, vol. 9, no. 1, pp. 289–302, 2018.

[37] S. Emaminejad, W. Gao, E. Wu, Z. A. Davies, H. Y. Y. Nyein, S. Challa, S. P. Ryan, H. M. Fahad, K. Chen, Z. Shahpar *et al.*, "Autonomous sweat extraction and analysis applied to cystic fibrosis and glucose monitoring using a fully integrated wearable platform," *Proceedings of the National Academy of Sciences*, vol. 114, no. 18, pp. 4625–4630, 2017.

[38] H. Lee, C. Song, Y. S. Hong, M. S. Kim, H. R. Cho, T. Kang, K. Shin, S. H. Choi, T. Hyeon, and D.-H. Kim, "Wearable/disposable sweat-based glucose monitoring device with multistage transdermal drug delivery module," *Science advances*, vol. 3, no. 3, p. e1601314, 2017.

[39] Y. Tanaka, T. Tajima, and M. Seyama, "Acoustic modal analysis of resonant photoacoustic spectroscopy with dual-wavelength differential detection for noninvasive glucose monitoring," *IEEE sensors letters*, vol. 1, no. 3, pp. 1–4, 2017.

[40] F. E. Barton, "Theory and principles of near infrared spectroscopy," in *Proceedings of the Korean Society of Near Infrared Spectroscopy Conference*. The Korean Society of Near Infrared Spectroscopy, 2001, pp. 1012–1012.

[41] C. D. Malchoff, K. Shoukri, J. I. Landau, and J. M. Buchert, "A novel noninvasive blood glucose monitor," *diabetes care*, vol. 25, no. 12, pp. 2268–2275, 2002.

[42] L. Heinemann, U. Krämer, H.-M. Klötzer, M. Hein, D. Volz, M. Hermann, T. Heise, and K. Rave, "Noninvasive glucose measurement by monitoring of scattering coefficient during oral glucose tolerance tests," *Diabetes technology & therapeutics*, vol. 2, no. 2, pp. 211–220, 2000.

[43] I. Fine, B. Fikhte, and L. D. Shvartsman, "Occlusion spectroscopy as a new paradigm for noninvasive blood measurements," in *Optical Diagnostics and Sensing of Biological Fluids and Glucose and Cholesterol Monitoring*, vol. 4263. International Society for Optics and Photonics, 2001, pp. 122–130.

[44] J. R. Ferraro, *Introductory raman spectroscopy*. Elsevier, 2003.

[45] G. Spanner and R. Niessner, "Noninvasive determination of blood constituents using an array of modulated laser diodes and a photoacoustic sensor head," *Fresenius' journal of analytical chemistry*, vol. 355, no. 3-4, pp. 327–328, 1996.

[46] A. Domschke, W. F. March, S. Kabilan, and C. Lowe, "Initial clinical testing of a holographic non-invasive contact lens glucose sensor," *Diabetes technology & therapeutics*, vol. 8, no. 1, pp. 89–93, 2006.

[47] O. S. Khalil, "Noninvasive photonic-crystal material for sensing glucose in tears," 2004.

[48] R. Fusman, R. Rotstein, K. Elishkewich, D. Zeltser, S. Cohen, K. Kofler, D. Avitzour, N. Arber, S. Berliner, and I. Shapira, "Image analysis for the detection of increased erythrocyte, leukocyte and platelet adhesiveness/aggregation in the peripheral blood of patients with diabetes mellitus," *Acta diabetologica*, vol. 38, no. 3, pp. 129–134, 2001.

[49] O. K. Cho, Y. O. Kim, H. Mitsumaki, and K. Kuwa, "Noninvasive measurement of glucose by metabolic heat conformation method," *Clinical chemistry*, vol. 50, no. 10, pp. 1894–1898, 2004.

[50] E. Alarousu, J. T. Hast, M. T. Kinnunen, M. Y. Kirillin, R. A. Myllyla, J. Plucinski, A. P. Popov, A. V. Priezzhev, T. Prykari, J. Saarela *et al.*, "Noninvasive glucose sensing in scattering media using oct, pas, and tof techniques," in *Saratov Fall Meeting 2003: Optical Technologies in Biophysics and Medicine V*, vol. 5474. International Society for Optics and Photonics, 2004, pp. 33–41.

[51] M. Nakamura, T. Tajima, K. Ajito, and H. Koizumi, "Selectivity-enhanced glucose measurement in multicomponent aqueous solution by broadband dielectric spectroscopy," in *2016 IEEE MTT-S International Microwave Symposium (IMS)*. IEEE, 2016, pp. 1–3.

[52] I. Harman-Boehm, A. Gal, A. M. Raykhman, J. D. Zahn, E. Naidis, and Y. Mayzel, "Noninvasive glucose monitoring: a novel approach," 2009.

[53] B. D. Cameron, J. S. Baba, and G. L. Coté, "Measurement of the glucose transport time delay between the blood and aqueous humor of the eye for the eventual development of a noninvasive glucose sensor," *Diabetes technology & therapeutics*, vol. 3, no. 2, pp. 201–207, 2001.

[54] J. C. Pickup, F. Hussain, N. D. Evans, and N. Sachedina, "In vivo glucose monitoring: the clinical reality and the promise," *Biosensors and Bioelectronics*, vol. 20, no. 10, pp. 1897–1902, 2005.

[55] K. V. Larin, M. S. Eledrisi, M. Motamedi, and R. O. Esenaliev, "Noninvasive blood glucose monitoring with optical coherence tomography: a pilot study in human subjects," *Diabetes care*, vol. 25, no. 12, pp. 2263–2267, 2002.

[56] L. A. Sodickson and M. J. Block, "Kromoscopic analysis: a possible alternative to spectroscopic analysis for noninvasive measurement of analytes in vivo." *Clinical chemistry*, vol. 40, no. 9, pp. 1838–1844, 1994.

[57] M. Gourzi, A. Rouane, R. Guelaz, M. Alavi, M. McHugh, M. Nadi, and P. Roth, "Non-invasive glycaemia blood measurements by electromagnetic sensor: study in static and dynamic blood circulation," *Journal of medical engineering & technology*, vol. 29, no. 1, pp. 22–26, 2005.

[58] G. Rao, R. Guy, P. Glikfeld, W. LaCourse, L. Leung, J. Tamada, R. Potts, and N. Azimi, "Reverse iontophoresis: Noninvasive glucose monitoring in vivo in humans," *Pharmaceutical research*, vol. 12, no. 12, pp. 1869–1873, 1995.

[59] A. Caduff, E. Hirt, Y. Feldman, Z. Ali, and L. Heinemann, "First human experiments with a novel non-invasive, non-optical continuous glucose monitoring system," *Biosensors and Bioelectronics*, vol. 19, no. 3, pp. 209–217, 2003.

[60] A. F. Kadjo, B. N. Stamos, C. P. Shelor, J. M. Berg, B. C. Blount, and P. K. Dasgupta, "Evaluation of amount of blood in dry blood spots: ring-disk electrode conductometry," *Analytical chemistry*, vol. 88, no. 12, pp. 6531–6537, 2016.

[61] B. Ballinger, J. Hsieh, A. Singh, N. Sohoni, J. Wang, G. H. Tison, G. M. Marcus, J. M. Sanchez, C. Maguire, J. E. Olgin *et al.*, "Deepheart: semi-supervised sequence learning for cardiovascular risk prediction," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Respiration Extraction from Single-Channel ECG using Signal-Processing Methods and Deep Learning

E. Merdjanovska*,† and A. Rashkovska*
* Department of Communication Systems, Jožef Stefan Institute, Ljubljana, Slovenia
† Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
E-mails:{elena.merdjanovska, aleksandra.rashkovska}@ijs.si

***Abstract*—The measured bio-potential on the surface of the body can give most of the information on the health status of an individual. Sensors that measure the ECG potential difference on the body surface could also provide information on other vital functions indirectly, like respiration, by a customized analysis of the ECG signal. Respiration is one of the most characteristic vital signs and can reflect the status of a patient or the progression of an illness. In this paper, we utilize signal-processing and deep learning methods for the extraction of the respiratory signal from the differential surface potential of a single-channel ECG. From signal processing, we investigate feature-based and filter-based methods, while from deep learning, an encoder-decoder architecture. Simultaneous measurements of a single-channel ECG and respiration have been obtained from 61 subjects before and after cardiac intervention in several positions of the body. We also investigate the power of the methods for respiration extraction depending on the period (pre-/post-operation) and the position of the body when the signal is obtained. The results show that the deep learning approach performs better than the filter-based methods but worse than the feature-based. Moreover, we conclude that different body positions do not influence respiration extraction significantly before and after the operation.**

***Keywords*—ECG; respiration; signal processing; deep learning**

I. Introduction

Respiration is one of the most informative vital signs for the physiological state of a person or the progression of an illness [1], [2]. Its continuous monitoring is valuable not only in hospital environments, but also during daily activities at home or in community settings [3], for example, for remote monitoring of chronic obstructive pulmonary disease (COPD) [4] or sleep apnoea [5]. Wearable devices for direct measurement of respiration parameters are usually uncomfortable when worn for a longer period of time [6]. However, wearable sensors for monitoring other vital signals, like electrocardiogram (ECG) and photoplethysmogram (PPG), i.e. pulse oximetry, are more unobtrusive. Medically-certified ECG sensors for long-term monitoring are already available on the market and are used in various areas of everyday life [7]. By utilizing the phenomena that respiration modulates the ECG and PPG signals, respiration can be extracted from these signals without using any additional sensors. Extracting new information from the primary signal supports the idea of a multi-functional device for monitoring various vital signals by combining a minimal number of body sensors with different functions into a single one [8], [9]. In this way, some of the vital signs are provided with direct measurement and others are derived from them by customized analysis.

The most utilized techniques for extraction of the respiratory signal or estimation of the respiratory rate (RR) from ECG or PPG are from the area of signal processing [10], [11]. The methods are usually based on three types of modulation: baseline wander (BW), amplitude modulation (AM), and frequency modulation (FM). Studies have shown that extracting respiration from ECG performs better than extracting it from PPG [12]. Besides signal processing, in recent years, deep learning (DL) [13], an area of machine learning, has become increasingly popular for a wide range of problems. Different types of deep neural network architectures have been developed for a wide range of problems and they were usually first used on images, and were later adjusted for other kinds of signals and their specific applications. Deep learning methods have been successfully used to solve many different tasks of ECG analysis, such as classification [14] and denoising [15], and also for the task of respiration extraction from PPG [16].

In this paper, the goal is to derive the respiratory signal from the differential surface potential of a single-channel ECG (ECG-derived respiration). Whereas most of the research in the domain of respiratory information extraction from ECG is focused on estimating the respiratory rate, extracting the respiratory signal as a pattern would be valuable for a more comprehensive evaluation of different respiratory problems [4], [5]. For this task, we utilize signal-processing algorithms and deep learning. Simultaneous measurements of a single-channel ECG and respiration have been obtained before and after cardiac intervention in several positions of the body. We further investigate the power of the methods for respiration extraction depending on the period (pre-/post-operation) and the position of the body when the signal is obtained. The

We acknowledge the financial support of the Slovenian Research Agency under Grant P2-0095 and the support of the EU H2020 project SAAM under Grant No. 769661.

paper is organized as follows: in Section II, we describe the experimental data, and the investigated signal-processing and deep learning methods. In Section III, we present and discuss the results. We conclude with the final remarks in Section IV.

II. Material and Methods

A. Experimental data

The experiments in this paper were conducted on a dataset which contains simultaneous ECG and respiratory signal recordings originally collected for the study in [17]. The respiration was recorded using a nose tip thermistor probe while two ECG signals were obtained: modified V2 and V5 leads. For this study, only the modified V5 lead ECG was used. The subjects in the study, 61 in total, were all cardiac patients who underwent coronary artery bypass surgery. The measurements were done both before and after the operation for the same patients, ranging from 2 days before to 28 days after. The patients were lying in three different positions during the measurements: supine (on the back), left lateral recumbent (on the left side) and right lateral recumbent (on the right side). Each recording is 10 minutes long and the signals are sampled at 450 Hz.

The study resulted in 867 simultaneous ECG and respiration recordings in total, since each of the 61 patients was recorded multiple times, on different days and positions. However, some of the recordings were excluded from our experiments due to different reasons, such as low signal quality and missing parts of the signal. In the end, the data we worked on consisted of 750 measurements. Table I gives the number of measurements in each group, depending on the position of the body and time when the signals are obtained.

TABLE I
Number of measurements in different groups: S (laying supine–on the back), L (laying on the left side), R (laying on the right side), Pre-op (before surgery), Post-op (after surgery)

Group	Number of measurements
S	331
L	149
R	270
Pre-op	271
Post-op	479
Total	750

B. Signal-processing methods

The signal-processing algorithms for extraction of the respiratory signal from ECG have been taken from the RRest project – Respiratory Rate Estimation project to develop and assess methods for automated respiratory rate (RR) monitoring, either from ECG or PPG [12]. The RRest project implements over 300 algorithms in Matlab, divided into three stages: extraction of respiratory signal, estimation of RR, and fusion of estimates. Considering the task in this work, we utilize only the algorithms for extraction of respiratory signal from the original ECG signal. These algorithms are divided into two groups: feature-based and filter-based. For investigation on our measurements, we have selected the best performing feature-based and filter-based algorithms based on the results from the main RRest study in [12]. From the feature-based, we have investigated the following algorithms:

- ELF_RSlinB_FMeam_FPt_RDtGC_EHF: difference between the amplitudes of troughs and proceeding peaks,
- ELF_RSlinB_FMebw_FPt_RDtGC_EHF: mean signal value between consecutive troughs [18],
- ELF_RSlinB_FMefm_FPt_RDtGC_EHF: time interval between consecutive peaks [19], [18].

From the filter-based, we have investigated the following algorithms:

- flt_Bfi: Band-pass filter between 4 and 60 breaths per minute [20],
- flt_Wam: The maximum amplitude of the Continuous Wavelet Transform (CWT) within plausible cardiac frequencies (30–220 beats per minute) [21],
- flt_Wfm: The frequency corresponding to the maximum amplitude of the CWT within plausible cardiac frequencies [21].

For each algorithm, common initial and final filtering to the signals have been applied. The initial filtering includes elimination of very low frequencies using a high-pass filter (-3 dB cutoff frequency of 4 breaths per minute). Intermediate steps depend on the extraction technique used: feature-based or filter-based, as described above. The final filtering includes elimination of non-respiratory frequencies using a band-pass filter (-3 dB cutoff frequencies of 4 and 60 breaths per minute). Moreover, signal quality is assessed using the algorithm described in [22]. The ECG signal is segmented into 32-second windows and the correlation between individual beats in a window and the window's average beat template is calculated. If the correlation is below an empirically determined threshold (here set to 0.66), then the segment is deemed to be of low quality and is excluded from the analysis.

C. Deep learning

1) Encoder-decoder model: The proposed deep learning network for respiration estimation is based on the topology of an encoder-decoder [23]. The encoder section takes $x^{(i)}$ as input and it produces feature vectors $z^{(i)}$. The decoder section uses $z^{(i)}$ as its input and produces the predicted respiratory signal $y^{(i)}_{pred}$ as output. These operations can be written as:

$$z^{(i)} = F_{enc}(x^{(i)}; \theta_{enc})$$

$$y^{(i)}_{pred} = F_{dec}(z^{(i)}; \theta_{dec})$$

where F_{enc} and F_{dec} are function representations of the encoder and the decoder with parameters θ_{enc} and θ_{dec}, respectively.

The neural network learns the parameters of both the encoder and the decoder (θ_{enc} and θ_{dec}) through back-propagation by minimizing the mean squared error. The encoder takes the ECG signal as input and outputs feature vectors, which are representations of the ECG signal. The

objective of the encoder part of the deep network is to find representations related to the respiration. In other words, the encoder needs to learn features from the ECG, using which the decoder could correctly estimate and output the respiratory signal. The decoder, on the other hand, aims to learn how the features output by the encoder are related to the respiratory signal.

2) Network Architecture: Both the encoder and the decoder of the network implemented in this work are fully convolutional. Due to this, the feature representations of the ECG signal are obtained by applying many different filters to the signal. The specific architecture used was originally applied to images for the task of image segmentation and is known as U-net [24]. Our implementation is a simplified version of U-net, with fewer layers and parameters, but the concept of shortcut connections is kept. Similar as in [16], where U-net has been implemented for extraction of the respiratory signal from PPG signal, we implement an U-net architecture for respiration extraction from ECG signal.

The network architecture is illustrated in Fig. 1. It consists of 13 convolutional layers in total. The ones in the first level have 8 filters, the ones in the second 16, and the ones in the third 32 filters. Each filter is a 1D convolution filter of length 27. In addition, each convolutional layer has a L2 kernel regularizer with a factor of 0.0005. All convolutional layers are followed by a ReLU activation and Batch Normalization (BN). The encoder part performs down-sampling of the signal using MaxPooling layers, while the decoder part performs up-sampling after each group of convolutional layers. In addition, there are two dropout layers, which reduce over-fitting. The dropout rate is set to 0.6. All these parameter settings resulted in 91,745 trainable neural network parameters. The batch size was set to 256 and the network was trained for 200 epochs, with a learning rate of 0.0005. The network was implemented using the Python library Keras and was trained on an Nvidia GeForce RTX 2080 Ti 11GB GPU.

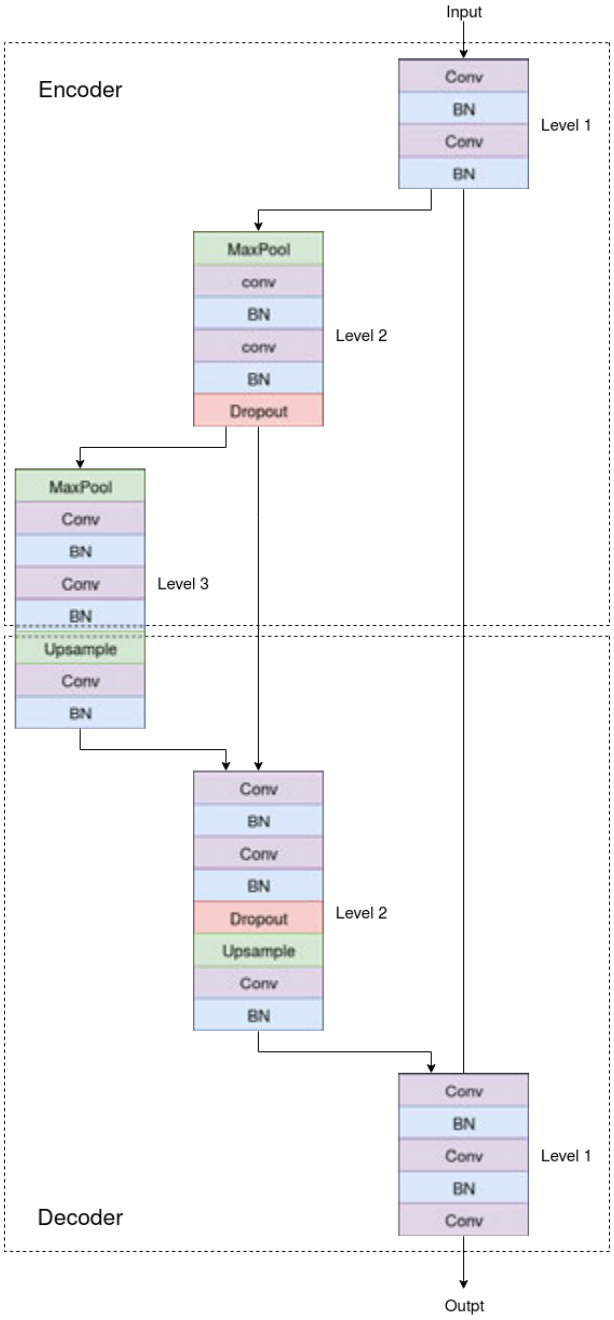


Fig. 1. U-net architecture

D. Evaluation technique

For the signal-processing methods, the input ECG signal was segmented into adjacent windows of 32 s. For each window, the respiratory signal was derived and compared to the measured one. For the neural network, a sliding window technique was used, where the full recording is split into 32-second long segments, with 16 seconds overlap. Then each of these windows was down-sampled to 1024 samples and fed as an input to the network. At the output, 32-second long estimation of the respiratory signal is obtained. After the training phase, the full respiratory signal is reconstructed.

For deep learning, we used the cross-validation evaluation technique. This is a common technique when training machine learning models and enables to have each record as a test sample exactly once. Using this type of evaluation, we ensure that our models are able to generalize well and the final results are an accurate estimate of the performance of the models for previously unseen examples. More specifically, 5-fold cross-validation was used, which means the models were always trained on 80% of the full dataset and tested on the remaining 20%. Furthermore, a random validation set consisting of 15% of the training set was used to select the best model for each fold.

To evaluate the methods quantitatively, we use the correlation coefficient (CC) metric calculated between the measured (actual) and derived respiratory signal. This metric can give an estimate of how similar two signals are in shape; however, it is not perfect in conveying how close one signal is to another. For each investigated method, we report the average CC on all measurements and all groups of measurements.

TABLE II
CORRELATION COEFFICIENT (CC) BETWEEN THE MEASURED AND ECG-DERIVED RESPIRATORY SIGNAL FOR THE RREST SIGNAL-PROCESSING METHODS AND THE U-NET

	Method	CC
Signal processing	ELF_RSlinB_FMeam_FPt_RDtGC_EHF	0.745765184
	ELF_RSlinB_FMebw_FPt_RDtGC_EHF	0.646964775
	ELF_RSlinB_FMefm_FPt_RDtGC_EHF	0.553050046
	flt_Bfi	0.46416707
	flt_Wam	0.433004984
	flt_Wfm	0.393748786
DL	U-net	0.485849124

III. RESULTS AND DISCUSSION

The CC values for the signal-processing methods and the U-net are given in Table II. Taking into consideration the different evaluation techniques for the RRest methods and the neural network, a direct comparison of their performance using the CC values cannot be taken as fair. Nevertheless, the results have shown that, from the signal-processing algorithms, the feature-based algorithms perform better than the filter-based, with the AM feature-based algorithm achieving the best result of 0.7458 average CC. This outcome is in agreement with the finding made in the original study of the RRest project [12], where conclusions have been made that feature-based methods in the stage of extraction of the respiratory signal perform the best. The 5-fold performance of the U-net is better than the performance of the filter-based methods, but worse than the one of the feature-based.

In order to compare the performance of the methods depending on when and how the measurements were obtained, the measurements have been divided into 3 groups according to the position of the body: S group (laying supine–on the back), L group (laying on the left side), R group (laying on the right side); and 2 groups according to the time of the measurement: pre-op (before surgery) and post-op (after surgery). Such investigation is feasible using the signal-processing methods, because the deep learning approach has an evaluation procedure that cannot be easily accommodated for such an experiment. The results for the CC of the signal-processing methods for the different groups are given in Table III. Graphical representation of the results is given in Fig. 2. The graph reveals that when the algorithms are compared to one another for different groups, again the feature-based are better than the filter-based. When compared between groups, it can be seen that for almost all algorithms, except one, the best performance is achieved for the measurements obtained before the operation. Between the groups with different body positions during measurement, such a conclusion cannot be made. Nevertheless, the performances for each group are relatively very close to each other and we can conclude that the position and the time of the measurement do not influence much the extraction process. This finding is in accordance with the conclusions made in the original study in which the measurements were obtained: different body positions do not influence breathing rates significantly before and after operation [17].

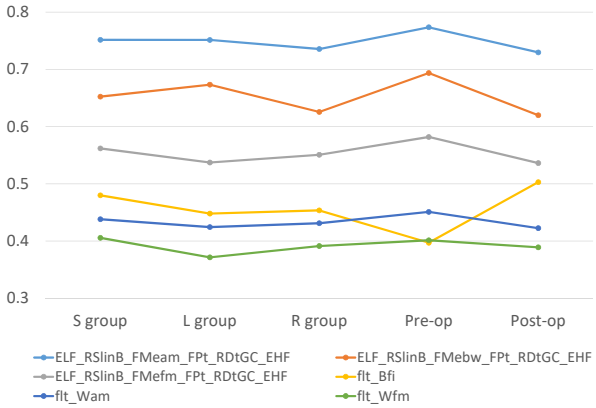


Fig. 2. Graphical plot of the correlation coefficient (CC) between the measured and ECG-derived respiratory signal for different groups: S group (laying supine–on the back), L group (laying on the left side), R group (laying on the right side), Pre-op (before surgery), Post-op (after surgery)

A more visual presentation of the results is given in Fig. 3. There, we compare the actual and derived respiratory signal for two examples with higher (CC = 0.92733) and lower (CC = 0.62352) correlation, alongside with the original ECG signal. Good overlap of the actual and derived respiration can be observed in the first example, whereas in the second example, more discrepancies are present.

IV. CONCLUSION

The goal of the paper is to assess the extraction of the respiratory signal from single-channel ECG measurements in 750 concurrent respiration-ECG recordings obtained from 61 subjects before and after cardiac intervention in different positions of the body. The study has demonstrated the feasibility of extracting the respiratory signal as a pattern from a single-channel ECG, which supports the multi-functional approach to monitor ECG and breathing simultaneously using a single bio signal. For this task, we investigated the performance of feature-based and filter-based signal-processing methods, as well as a deep learning approach using U-net neural network architecture. The experiments have shown that the best-performing method belongs to the group of feature-based methods and exploits the amplitude modulation of the ECG signal caused by respiration. The deep learning approach performs better than the filter-based methods, but worse than the feature-based. The results have confirmed that different body positions do not influence respiration extraction significantly before and after the operation. Future work includes the evaluation of the performance of respiration extraction on measurements acquired under a variety of conditions different from the ones in this study. Evaluating ECG-derived respiration techniques on measurements obtained during every-day activities could be a valuable task for health monitoring in home environments.

TABLE III

CORRELATION COEFFICIENT (CC) BETWEEN THE MEASURED AND ECG-DERIVED RESPIRATORY SIGNAL FOR DIFFERENT GROUPS: S GROUP (LAYING SUPINE–ON THE BACK), L GROUP (LAYING ON THE LEFT SIDE), R GROUP (LAYING ON THE RIGHT SIDE), PRE-OP (BEFORE SURGERY), POST-OP (AFTER SURGERY)

Method	S group	L group	R group	Pre-op	Post-op
ELF_RSlinB_FMeam_FPt_RDtGC_EHF	0.751655786	0.751340606	0.73546931	0.77352044	0.729584718
ELF_RSlinB_FMebw_FPt_RDtGC_EHF	0.652334101	0.673325195	0.625677164	0.693534856	0.619815843
ELF_RSlinB_FMefm_FPt_RDtGC_EHF	0.562002058	0.537340911	0.550930482	0.581908568	0.53622641
flt_Bfi	0.479953607	0.448183338	0.453742745	0.397227129	0.50301551
flt_Wam	0.438330829	0.424482607	0.431240009	0.450998764	0.422559344
flt_Wfm	0.405703995	0.371696278	0.391385664	0.401517963	0.389215863

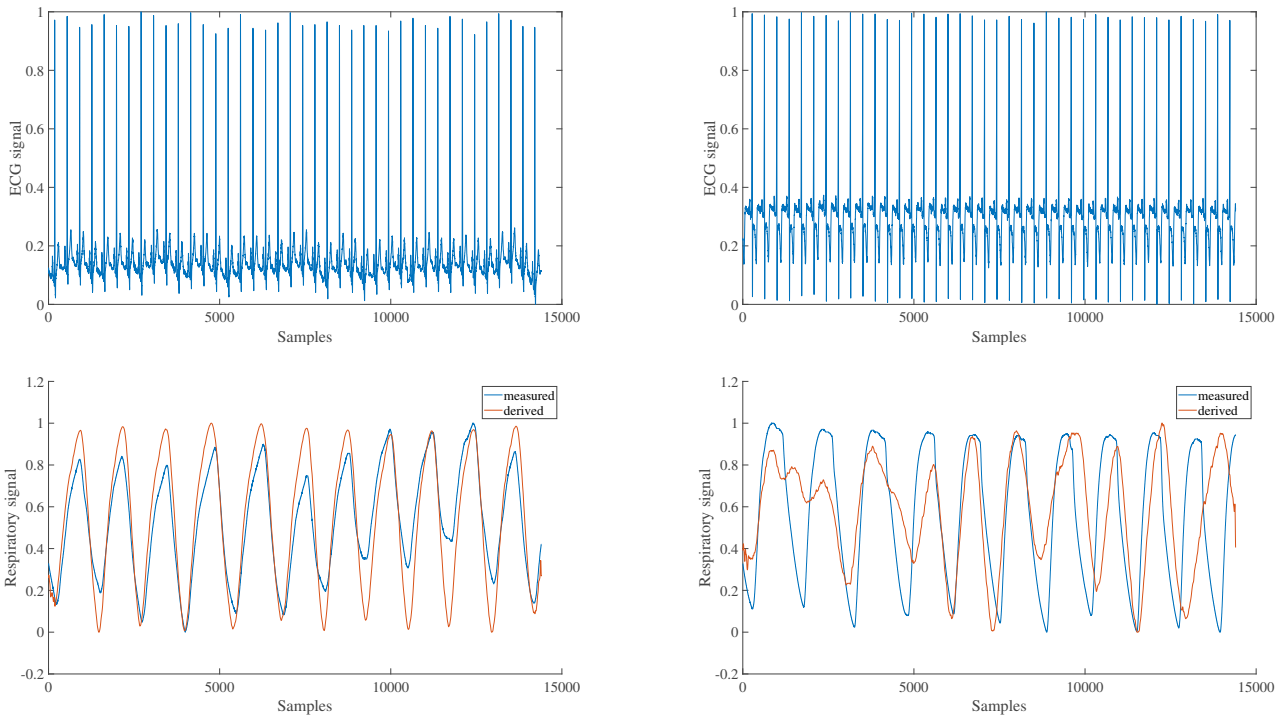


Fig. 3. Examples of measured and ECG-derived respiration with (a) high correlation (CC = 0.92733). (b) lower correlation (CC = 0.62352). The input ECG signal is displayed above.

REFERENCES

[1] S. Ridley, "The recognition and early management of critical illness," *Ann. R. Coll. Surg. Engl.*, vol. 87, p. 315–322, 2005.

[2] M. A. Cretikos, R. Bellomo, K. Hillman, J. Chen, S. Finfer, and A. Flabouris, "Respiratory rate: the neglected vital sign," *Med. J. Aust.*, vol. 188, pp. 657–659, 2008.

[3] P. Bonato, "Wearable sensors and systems," *IEEE Eng. Med. Biol. Mag.*, vol. 29, pp. 25–36, 2010.

[4] I. Tomasic, N. Tomasic, R. Trobec, M. Krpan, and T. Kelava, "Continuous remote monitoring of COPD patients—justification and explanation of the requirements and a survey of the available technologies," *Med. Biol. Eng. Comput.*, vol. 56, pp. 547–569, 2018.

[5] A. Malhotra and D. P. White, "Obstructive sleep apnoea," *Lancet*, vol. 360, pp. 237–245, 2002.

[6] C. Orphanidou, D. Clifton, S. Khan, M. Smith, J. Feldmar, and L. Tarassenko, "Telemetry-based vital sign monitoring for ambulatory hospital patients," in *Conf. Proc. IEEE Engineering in Medicine and Biology Society*, 2009, pp. 4650–4653.

[7] A. Rashkovska, M. Depolli, I. Tomašić, V. Avbelj, and R. Trobec, "Medical-Grade ECG Sensor for Long-Term Monitoring," *Sensors*, vol. 20, 2020, 1695.

[8] R. Trobec, V. Avbelj, and A.Rashkovska, "Multi-functionality of wireless body sensors," *The IPSI BgD transactions on internet research*, vol. 10, pp. 23–27, 2014.

[9] M. Depolli, V. Avbelj, R. Trobec, J. M. Kališnik, T. Korošec, A. P. Susič, U. Stanič, and A. Semeja, "PCARD Platform for mHealth Monitoring," *Informatica*, vol. 40, pp. 117–123, 2016.

[10] R. Bailon, L. Sornmo, and P. Laguna, "ECG-derived respiratory frequency estimation," in *Advanced Methods and Tools for ECG Data Analysis*, G. D. Clifford, F. Azuaje, and P. E. Mcsharry, Eds. London: Artech House, 2006, ch. 8, pp. 215–244.

[11] R. Trobec, A. Rashkovska, and V. Avbelj, "Two Proximal Skin Electrodes — A Respiration Rate Body Sensor," *Sensors*, vol. 12, pp. 13 813–13 828, 2012.

[12] P. H. Charlton, T. Bonnici, L. Tarassenko, D. A. Clifton, R. Beale, and P. J. Watkinson, "An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram," *Physiol. Meas.*, vol. 37, pp. 610–626, 2016.

[13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[14] A. Hannun, P. Rajpurkar, M. Haghpanahi, G. Tison, C. Bourn, M. Turakhia, and A. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, 2019.

[15] H.-T. Chiang, Y.-Y. Hsieh, S.-W. Fu, K.-H. Hung, Y. Tsao, and S.-Y. Chien, "Noise Reduction in ECG Signals Using Fully Convolutional Denoising Autoencoders," *IEEE Access*, vol. 7, pp. 60 806–60 813, 2019.

[16] V. Ravichandran, B. Murugesan, V. Balakarthikeyan, S. M. Shankaranarayana, K. Ram, S. P. Preejith, J. Joseph, and M. Sivaprakasam, "RespNet: A deep learning model for extraction of respiration from photoplethysmogram," 2019, arXiv:1902.04236v2 [eess.SP].

[17] J.-M. Kalisnik, V. Avbelj, R. Trobec, and B. Gersak, "Position-dependent changes in vagal modulation after coronary artery bypass grafting," *Comput. Biol. Med.*, vol. 37, pp. 1404–1408, 2007.

[18] W. Karlen, S. Raman, J. M. Ansermino, and G. A. Dumont, "Multiparameter respiratory rate estimation from the photoplethysmogram," *IEEE Trans. Biomed. Eng.*, vol. 60, pp. 1946–1953, 2013.

[19] C. Orphanidou, S. Fleming, S. Shah, and L. Tarassenko, "Data fusion for estimating respiratory rate from a single-lead ECG," *Biomed. Signal Process. Control*, vol. 8, p. 98–105, 2013.

[20] L. G. Lindberg, H. Ugnell, and P. A. Öberg, "Monitoring of respiratory and heart rates using a fibre-optic sensor," *Med. Biol. Eng. Comput.*, vol. 30, pp. 533–537, 1992.

[21] P. Addison and J. Watson, "Secondary transform decoupling of shifted nonstationary signal modulation components: application to photoplethysmography," *Int. J. Wavelets Multiresolution Inf. Process*, vol. 2, pp. 43–57, 2004.

[22] C. Orphanidou, T. Bonnici, P. Charlton, D. Clifton, D. Vallance, and L. Tarassenko, "Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring," *IEEE J. Biomed. Health Inf.*, vol. 19, pp. 832–838, 2015.

[23] W. Wang, Y. Huang, Y. Wang, and L. Wang, "Generalized Autoencoder: A Neural Network Framework for Dimensionality Reduction," in *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014, pp. 490–497.

[24] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241.

Simultaneous Phonocardiography and Electrocardiography Using Smartphone in Dogs, Cats and Horses

V. Kadunc Kos*, M. Brložnik**, A. Domanjko Petrič** and V. Avbelj***

*Clinic for Reproduction and Large Animals, Veterinary Faculty, University of Ljubljana, Slovenia
**Small Animal Clinic, Veterinary Faculty, University of Ljubljana, Slovenia
***Department of Communication Systems, Jožef Stefan Institute, Ljubljana, Slovenia

Corresponding author: viktor.avbelj@ijs.si

***Abstract* – A phonocardiographic device was constructed using a microphone connected to a stethoscope tube and a smartphone and synchronized with a wireless electrocardiographic sensor attached precordially to the animal's chest and connected to the smartphone via Bluetooth technology. Cardiac sounds and murmurs were recorded simultaneously with the animal's electrocardiogram on the smartphone in 18 animals (eight dogs, seven horses and three cats). Ten of these animals (six dogs, three horses and one cat) were cardiac patients with various murmurs detected on the phonocardiogram and nine were healthy animals. The quality and synchronization of recorded heart sounds and murmurs relative to the recorded electrocardiogram was assessed in order to analyse this prototype device in practice.**

***Keywords* -** smartphone phonocardiography, wireless electrocardiography, veterinary medicine.

I. INTRODUCTION

Heart sounds are produced by vibrations of cardiac structures and blood circulation, and can be interpreted by cardiac auscultation with a stethoscope or can be graphically presented and further analysed. Recordings of heart sounds and their graphical presentation – phonocardiography have been introduced in veterinary medicine in late 60s of 20th century [1] to analyse heart sounds and murmurs and to help the clinician reaching the diagnosis. Phonocardiography can also be used as a teaching tool and for legal and insurance purposes [2]. Despite the diagnostic capabilities of this modality, phonocardiography was rarely used in the last decades in small animal cardiology, although there are some scientific papers that rely upon that technique [3-7]. The recent underplaying of traditional phonocardiography is not only because echocardiography offers a more powerful means for cardiological diagnosis, but also because traditional phonocardiography is a rather cumbersome method with potentials of technical problems often causing artefacts and poor-quality recordings.

Synchronization of phonocardiogram (PCG) to electrocardiogram (ECG), which is a recording of electrical cardiac activity (P-QRS-T waves on surface ECG), is important since it allows for proper timing of individual heart sounds and murmurs, and permits observation of electrical systole (Q wave to the end of T wave), electromechanical systole (Q wave to the beginning of S2 heart sound) and electro-mechanical window (end of T wave to the beginning of S2 sound). Electromechanical window is used to predict arrhythmias in patients [8] and in computer simulations [9], and is today easily measured with small wireless sensors connected to smart devices.

Mobile phonocardiography has been used to classify between innocent and pathological murmurs in human foetus [9] and as medical students' teaching aid [10]. In veterinary medicine, acoustic ambulatory ECG monitor was used for continuous recordings of heart sounds and murmurs in 123 healthy horses [11].

In this paper a simple device, constructed from wireless ECG sensor, a stethoscope, a microphone and a smartphone, recording simultaneous electrocardiographic and phonocardiographic data, was used to make recordings in 18 animals. The quality of recorded heart sounds and murmurs relative to the simultaneous electrocardiographic data was assessed in order to analyse this prototype device in practice.

II. PRESENTATION

A. Materials and Methods

A phonocardiographic device was constructed using a microphone of a phone accessory connected to a stethoscope tube and a smartphone (Fig. 1). Mobile application WaveEditor for Android™ was used for recordings of phonocardiographic data. It was synchronized with a wireless ECG sensor (Savvy®, Saving d.o.o., Slovenia) that was attached to the animal's chest precordially and connected to the smart phone via Bluetooth technology (Fig. 2). Mobile application MobECG (Jožef Stefan Institute, V1.8.1) was used on the android smart device. The ECG sensor was used for electrocardiographic recording in dogs, cats and horses previously [12,13]. The construction of phonocardiographic device and its synchronization to ECG sensor is described in details in a separate paper [14].

Cardiac sounds and murmurs were recorded simultaneously with the animal's electrocardiogram in 18 animals (eight dogs, seven horses and three cats).

Ten of these animals (six dogs, three horses and one cat) were cardiac patients with various heart murmurs and eight were healthy animals.

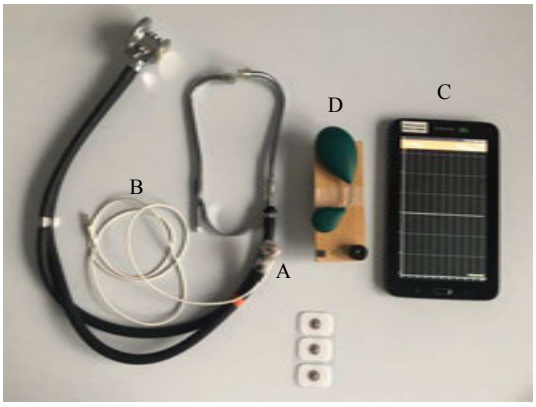


Figure 1: Device for phonocardiography was constructed from a microphone (A) connected to a stethoscope tube and via a wire (B) to a smart device (C). The ECG sensor Savvy (D) is attached to a device used for synchronisation.

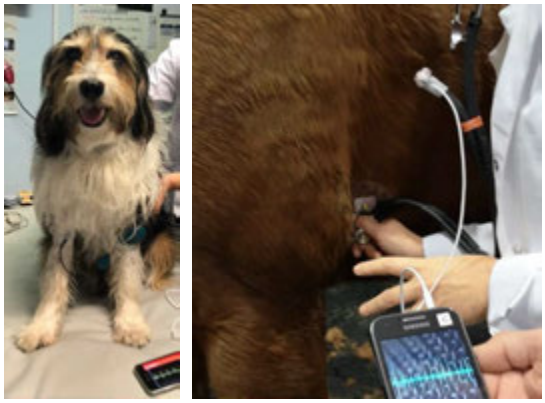


Figure 2. In the image on the left, ECG sensor is attached to dog's legs and ECG can be seen on the phone display. Simultaneous PCG and ECG recordings were made with a device seen to the right, while recording a horse; constructed from a stethoscope, a microphone, an ECG sensor (not visible) and a smartphone.

For each animal the quality of simultaneous recordings was assessed, and durations of electromechanical activation time (EMAT = Q–S1 interval), electrical systole (QT interval), mechanical systole (S1–S2 interval) and electromechanical window (EMW) were measured (Fig. 3). Each duration was measured manually on 5 successive heart beats to cover at least one respiratory cycle. For the end of T wave, we used the tangent method, which defines the end of T wave at the time, where a tangent drawn to the steepest slope of the last limb of the T wave intersects the local baseline.

If the start of S2 was later than the end of T wave, EMW interval had positive value. EMW was negative, if the start of S2 was earlier than the end of T wave (Fig. 4).

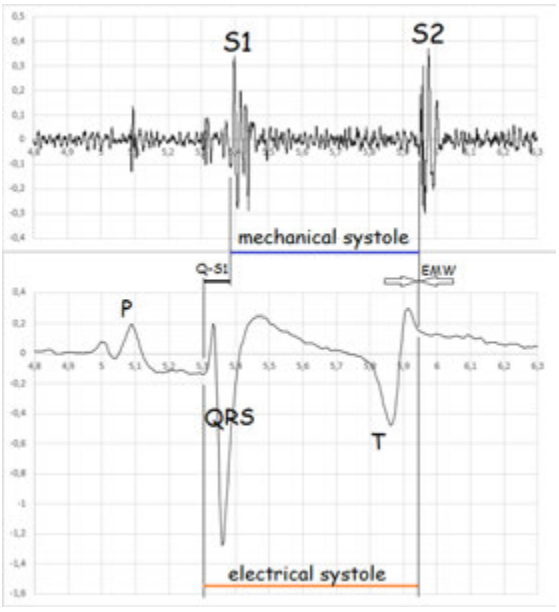


Figure 3. PCG and ECG of a healthy 2-year-old horse (H1). The interval between heart sounds S1 and S2 defines the mechanical systole, while the interval from the start of QRS to the end of T wave defines electrical systole. Electromechanical Window (EMW) is the interval from the end of T wave to the beginning of S2. EMW interval in this horse is near zero.

B. Results

In ten animals (six dogs, three horses, and one cat) the interval durations on 5 consecutive beats were obtained and their average values with standard deviations are presented in Table I. The start of S1 was less well identifiable and commonly more variable, while the start of S2 was well identifiable (1 ms) except in the case D5. QT variability and S1–S2 variability were low (< 10 ms). Negative EMW (Table 1, Fig. 4) was measured in two dogs and in the cat.

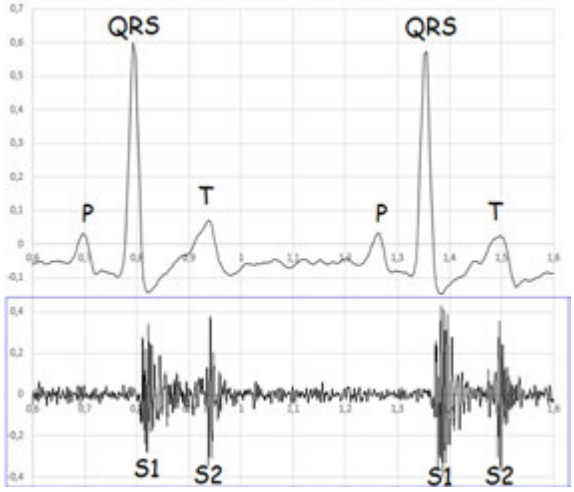


Figure 4. ECG and PCG of a dog, a 13-year-old male Pekingese (D4) with myxomatous mitral valve disease, a holosystolic murmur grade III/VI and pulmonary hypertension, with heart rate of 123 bpm. EMW is negative, S2 starts before the end of T wave (-32 ms).

For clear presentation of auscultated heart murmurs, the recorded signal was high-pass filtered for a display of only higher frequencies (Fig. 5, the lower graph).

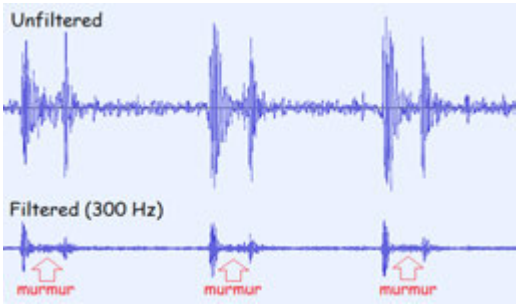


Figure 5. PCG of the dog in Fig. 4 (D4) with holosystolic murmur grade III/VI, which can be clearly seen if the recorded signal is high-pass filtered at 300 Hz – only the frequencies above 300 Hz are present. Such filtering augments the visibility of recorded murmurs.

TABLE I. AVERAGE VALUE AND STANDARD DEVIATION OF HEART RATE (HR), RR INTERVAL, ELECTROMECHANICAL ACTIVATION TIME (EMAT IS EQUAL TO Q-S1), ELECTRICAL SYSTOLE (QT), MECHANICAL SYSTOLE (S1-S2) AND ELECTROMECHANICAL WINDOW (EMW) IN 10 ANIMALS

Animal HR [bpm], RR [ms] Diagnoses	EMAT [ms]	QT [ms]	S1-S2 [ms]	EMW [ms]
D1 113, 530 ± 62 normal heart	*	176 ± 2	211 ± 2	*
D2 156, 385 ± 13 MVD, TVD, PH	27 ± 0	195 ± 7	144 ± 1	-23 ± 7
D3 143, 420 ± 13 MVD, PH	43 ± 1	188 ± 6	153 ± 3	9 ± 7
D4 123, 489 ± 49 MVD, PH	32 ± 3	179 ± 5	115 ± 3	-32 ± 5
D5 164, 367 ± 14 MVD, PH	22 ± 2	180 ± 4	213 ± 16 **	54 ± 16 **
D6 200, 300 ± 2 MVD	32 ± 1	157 ± 1	137 ± 2	11 ± 1
C1 228, 263 ± 6 normal heart	20 ± 5	171 ± 4	136 ± 5	-15 ± 4
H1 32, 1886 ± 23 normal heart	67 ± 3	631 ± 1	563 ± 2	-1 ± 2
H2 50, 1221 ± 18 colic, normal heart	37 ± 4	448 ± 1	474 ± 4	62 ± 1
H3 33, 1830 ± 27 MVD	***	519 ± 7	432 ± 2	27 ± 6

LEGEND: D = dog, C = cat, H = horse; bpm = beats per minute, MVD = mitral valve disease, TVD = tricuspid valve disease, PH = pulmonary hypertension, ms = miliseconds, * = ECG and PCG synchronisation failed and no electromechanical parameters could be determined, ** = low amplitude S2 and therefore the start of S2 cannot be reliably determined, *** = low amplitude S1 and therefore the start of S1 cannot be reliably determined.

For eight animals (8/18, 44%) the quality of recordings was poor, either synchronization was unsuccessful or one of the recordings was non-diagnostic. An example of a noisy ECG with low amplitude values hindering the determination of the end of T wave is presented in Fig. 6.

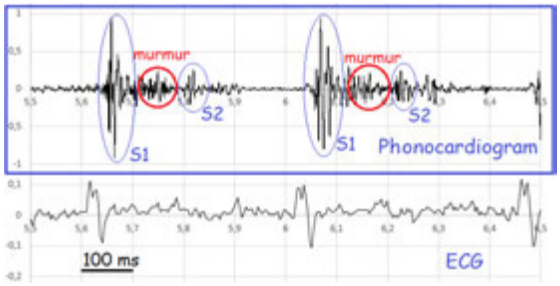


Figure 6. PCG of a dog with a holosystolic murmur grade IV/VI and heart rate of 138 beats/min with a simultaneous noisy ECG with low amplitude of QRS waves (0.2 mV).

III. DISCUSSION

A simple device constructed from the ECG wireless sensor, a stethoscope, a microphone and a smartphone proved to give excellent results in more than half of the animals. In others, synchronisation was not successful or one of the recordings was non-diagnostic. Precise synchronisation of phonocardiography to ECG is of crucial importance in order to determine event durations and for exact timing of heart murmurs and clicks.

In all recordings, QT variability (electrical systole) and S1–S2 variability (mechanical systole) were low (< 10 ms) except in one case (D5), where variability of mechanical systole (S1–S2) was 16 ms due to low amplitude of sound S2 causing inaccurate determination of S2 starting time. Low QT variability (electrical systole) was previously reported for humans [15,16]. It is well documented and established that changes in length of electrical systole may make the myocardial cells more vulnerable and unstable, and may also cause serious cardiac arrhythmias and even sudden death [17].

The electromechanical window (EMW), which is the time interval from the end of T wave to the closure of aortic and pulmonary valves, has been proposed as a risk marker for Torsade de Pointes of drug induced arrhythmias in a canine model [18]. A negative EMW would mean that the ventricular relaxation comes earlier than electrical repolarization and is most commonly associated with long QT syndrome in people. Another study suggested that EMW is no better marker to assess risk for arrhythmias than prolonged QT [19]. Measurement of Q–S1 interval or Electro Mechanical Activation time (EMAT) that is the interval from the beginning of electrical activation of LV (the onset of Q wave on ECG) to the onset of the first sound (mitral valve closure) is one of the parameters that can be used to evaluate LV function by modern acoustic cardiography.

In general, the RR interval is used to standardize the EMAT, namely EMATc (EMAT/RR), which represents a

proportion of EMAT in the cardiac cycle. EMATc reflects the time required by LV systole to produce a sufficient pressure to close the mitral valve, while an increased EMATc is associated with a decline in LV systolic function in heart failure patients [20]. Duration times of EMAT in our horses were in accordance with the ones measured with acoustic ambulatory ECG monitor [7]. In human medicine EMAT is used to predict discharge outcomes in patients hospitalized for acute heart failure syndrome [21] and was shown to be strongly associated with impaired LV contractility but had no association with LV filling pressures [22].

Start of S1 was less well identifiable and commonly more variable, while start of S2 was well identifiable (1 ms); therefore, accurate monitoring of electrical and mechanical systole duration was possible. Beside major heart sounds also heart murmurs and clicks were recognized on graphic display in 10 cardiac patients. Heart murmurs can be clearly seen when PCGs are filtered. Due to conventional auscultation's subjective nature and the fact that it is not easily shared, simple device, as presented in our case, enables excellent teaching aid for students. Modern medical technology allows us to optimize auscultatory findings by physically characterizing sounds through recordings, visualization and automated analysis systems [10,23-25].

Our simple low-cost device consisted of traditional acoustic stethoscope that operates on the transmission of sound from the chest piece via air-filled tube. Disadvantage of acoustic stethoscopes compared to electronic is the low sound level. Despite the fact that electronic stethoscopes might be superior in the detection of low-grade heart murmurs, experience in auscultation remains more relevant than the device [22,25-27]. However, students' performance on a cardiac auscultation test did not differ if an electronic or a conventional stethoscope was used during training [28]. Anyhow, analysis of PCG seems a beneficial additional tool for the diagnosis of canine cardiac murmurs for those with less cardiological experience [29]. In a study to differentiate innocent from pathological murmurs on PCGs, innocent murmurs had a significantly shorter "murmur-to-systole duration ratio" than the pathological ones [30]. In some electronic stethoscopes ECG is provided, however, synchronization of the recordings should be investigated in electronic stethoscopes and acoustic ambulatory ECG monitors if they are to be used to measure the durations of electromechanical events.

Today, smartphones, applications, heart sound databases and development of algorithms for heart sound evaluation enable not only easy use of phonocardiography [31-33], but also a simultaneous ECG that permits accurate measurements of electro-mechanical events.

IV. Conclusion

With a simple device, constructed from the Savvy® ECG wireless sensor, a stethoscope, a microphone and a smart device, synchronised good quality ECG and PCG recordings can be obtained. Audio recordings can be shared for listening or as PCGs, which is a valuable tool for teaching and diagnostics, and could be used in telemedicine. Synchronisation of ECG to PCG enables accurate measurements of EMAT, electrical and mechanical systole, and EMW. Results of our study are however only preliminary and further research is needed. Furthermore, the prototype of the device should be technologically improved and certified in order to enable wider use.

Acknowledgment

This work was partially supported by the Slovenian Research Agency (grant P2-0095; V. Avbelj).

References

[1] E. W. Fischer, "Heart sound recording", J Small Animal Pract 1967; 8: 171–173.

[2] C. Kvart and J. Häggström, "Cardiac auscultation and phono-cardiography in dogs, horses and cats", Sweden: TK i Uppsala, 2002: 124 pp.

[3] K. Vörös, I. Nolte, S. Hungerbühler, J. Reiczigel, J. P.Ehlers, G. Tater, R. Mischke, T. Zimmering and M. Scneider, "Sound recording and digital phonocardiography of cardiac murmurs in dogs by using a sensor-based electronic stethoscope", Acta Vet Hung 2011; 59: 23–35.

[4] I. Ljungvall, C. Ahlstrom, H. Höglund, P. Hult, C. Kvart, M. Borgarelli, P. Ask and J. Häggström, "Use of signal analysis of heart sounds and murmurs to assess severity of mitral valve regurgitation attributable to myxomatous mitral valve disease in dogs", Am J Vet Res 2009; 70: 604–613.

[5] H. D. Pedersen, J. Häggström, T. Falk, T. Mow, L. H. Olsen, L. Iversen and A. L. Jensen. Auscultation in mild mitral regurgitation in dogs: Observer variation, effects of physical maneuvers, and agreement with color Doppler echocardiography and phonocardiography", J Vet Intern Med 1999; 13: 56–64.

[6] F. H. Welker and W. W. Muir, "An investigation of the second heart sound in the normal horse", Equine Vet J 1990; 22: 403–407.

[7] N. Zuber, M. Zuber and C. C. Schwarzwald, "Assessment of systolic and diastolic function in clinically healthy horses using ambulatory acoustic cardiography", Equine Vet J 2019; 51: 391–400.

[8] R. M. A. Bekke, K. H. Haugaa, A. van den Vijngaard, J. M. Bos, M. J. Ackerman, T. Edvardsen and P. G. Volders, "Electro-mechanical window negativity in genotyped long QT syndrome patients: relation to arrhythmia risk", Eur Heart J 2015; 36: 179–186.

[9] E. Passini, C. Trovato, P. Morisette, F. Sannajust, A. Bueno-Orovio and B. Rodriguez, "Drug-induced shortening of the electro-mechanical window is an effective biomarker for in silico prediction of clinical risk of arrhythmias", Br J Pharmacol 2019; 176: 3819–3833.

[10] A. M. Amiri, G. Armano and A. Mohammad, "PhonoSys: mobile phonocardiography", Mobihealth 2015.

[11] L. J. Thoms, G. Colicchia and R. Girwidz, "Phonocardiography with a smartphone", Phys Educ 2017; 52: 1–4.

[12] M. Brložnik, Š. Likar, A. Krvavica, V. Avbelj and A. Domanjko Petrič, "Wireless body sensor for electrocardiographic monitoring in dogs and cats", J Small Animal Pract 2019; 60: 223–230.

[13] M. Brložnik, A. Domanjko Petrič, V. Kadunc Kos, A. Rashkovska and V. Avbelj, "Wireless electrocardiographic monitoring in equine medicine", 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp 279–283.

[14] V. Avbelj and M. Brložnik, "Phonocardiography and electro-cardiography with a smartphone", Proceedings of 43rd International Convention (MIPRO), Opatija, Croatia, 2020.

[15] V. Avbelj, R. Trobec and B. Gersak. Beat-to-beat repolarisation variability in body surface electrocardiograms. Med Biol Eng Comput 2003; 41: 556–560.

[16] M. Baumert, A. Porta, M. A. Vos, M. Malik, J. P. Couderc, P. Laguna, G. Piccirillo, G. L. Smith, L. G. Tereschchenko and P. G. A. Volders, "QT interval variability in body surface ECG: measurement, physiological basis, and clinical value: position statement and consensus guidance endorsed by the European Heart Rhythm Association jointly with the ESC Working Group on Cardiac Cellular Electrophysiology", Europace 2016; 18: 925–944.

[17] F. R. Breijo-Marquez and M. Pardo Rios, "Variability and diversity of the electrical cardiac systole", BMJ Case Rep 2009; 2009.

[18] H. J. van der Linde, B van Deuren, Y. Somers, B. Loenders, R. Towart and D. J. Gallacher, "The electro-mechanical window: a risk marker for Torsade de Pointes in a canine model of drug induced arrhythmias", Br J Pharmacol 2010; 161: 1444–1454.

[19] T. R. Stams, V. J. Bourgonje, H. D. Beekman, M. Schoenmakers, R. van der Nagel, P. Osterhoff, JM van Opstal and M. A. Vos, "The electromechanical window is no better than QT prolongation to assess risk of Torsade de Pointes in the complete atrioventricular block model in dogs", Br J Pharmacol 2014; 171: 714–722.

[20] J. Zhang, W. X. Liu and S. Z. Lyu, "Predictive value of electromechanical activation time for in-hospital major cardiac adverse events in heart failure patients", Hindawi Cardiovasc Ther 2020; 532596, 6 pp.

[21] T. F. Chao, S. H. Sung, H. M. Cheng, W. C. Yu, K. L. Wang, C. M. Huang and C. H. Chen, "Electromechanical activation time in the prediction of discharge outcomes in patients hospitalized with acute heart failure syndrome", Intern Med 2010; 49: 2031–2037.

[22] S. Efstratiadis and A. D. Michaels, "Computerized acoustic cardiographic electromechanical activation time correlates with invasive and echocardiographic parameters of left ventricular contractility", J Cardiol Fail 2008; 14: 577–582.

[23] E. Andrès, A. Hajjam and C. Brandt, "Advances and innovations in the field of auscultation, with a special focus on the development of new intelligent communicating stethoscope systems", Health Technol 2012; 2: 5–16.

[24] E. Andrès, "Advances and perspectives in the field of auscultation, with a special focus on the contribution of new intelligent communicating stethoscope systems in clinical practice, in teaching and telemedicine", eHealth and Remote Monitoring 2011, chapter 6: 97–116.

[25] S. Leng, R. S. Tan, K. T. C. Chai, C. Wang, D. Ghista and L. Zhong, "The electronic stethoscope", BioMed Eng OnLine 2015; 14: 66–103.

[26] C. Liu, D. Springer, Q. LI, B. Moody, R.A. Juan, F.J. Chorro, et al., "An open access database for the evaluation of heart sound algorithms", Physiological Measurement 2016; 27: 2181–2213.

[27] K. Vörös, A. Bonnevie and J. Reiczigel, "Comparison of conventional and sensor-based electronic stethoscopes in detecting cardiac murmurs of dogs", Tierarztl Prax Kleintiere 2012; 40: 103–111.

[28] H. Hoyte, T. Jensen and K. Gjesdal, "Cardiac auscultation training of medical students: a comparison of electronic sensor-based and acoustic stethoscopes", BMC Medical Education 2005; 5: 14.

[29] S. M. Marinus, H. van Engelen, and V. Szatzmári, "N-Terminal Pro-B-Type Natriuretic Peptide and phonocardiography in differentiating innocent cardiac murmurs from congenital cardiac anomalies in asymptomatic puppies", J Vet Intern Med 2017; 31: 661–667.

[30] N. Giordano and M. Knaflitz, "A novel method for the timing of heart sound components through digital phonocardiography", Sensors 2019; 19: 1868–1884.

[31] H. Mondal, S. Mondal and K. Saha, "Development of low-cost wireless phonocardiograph with a Bluetooth headset under resource-limited conditions", Med Sci 2018; 6. 117–125.

[32] R. Abasi-Kesbi, A. Valipour and K. Imani, "Cardiorespiratory system monitoring using a developed acoustic sensor", Health Tecnol Letters 2017; 5: 7–12.

[33] A. Sa-Ngasoongsong, J. Kunthong, V. Sarangan, Xinwei Cai and S. T. S. Bukkapatnam, "A low-cost, portable, high-throughput wireless sensor system for phonocardiography applications", Sensors 2012; 12: 10851–10870.

Phonocardiography and Electrocardiography with a Smartphone

V. Avbelj* and M. Brložnik**

*Department of Communication Systems, Jožef Stefan Institute, Ljubljana, Slovenia
**Small Animal Clinic, Veterinary Faculty, University of Ljubljana, Slovenia

viktor.avbelj@ijs.si, majabrloznik@gmail.com

Abstract – **Phonocardiogram, a graphic recording of auscultatory findings of the cardiac sound vibrations, was acquired through an add-on to a stethoscope: a microphone was inserted into the stethoscope tube and connected to a smart device. Phonocardiographic data were synchronised to electrocardiogram, a recording of cardiac electrical activity, that was obtained with a wireless body sensor. Simultaneous phonocardiographic and electrocardiographic data enable a much more detailed understanding of the cardiac function than auscultation of the heart sounds and murmurs with a stethoscope only. Furthermore, the possibility to study phonocardiogram synchronised to electrocardiogram seems a valuable diagnostic tool for the clinician and provides a great teaching material for students. We demonstrate a simple way to construct a device that records simultaneous phonocardiographic and electrocardiographic data, and hence enables a thorough insight into the cardiac electro-mechanical interactions. Cardiac activity was analysed in 9 volunteers, where timing relations of various electromechanical events were measured.**

Keywords **-** PCG, ECG, wireless body sensor, stethoscope, smartphone accessories.

I. INTRODUCTION

Since René Laënnec, a physician and a musician, who made the first stethoscope from a quire of paper rolled into a kind of a cylinder in 1816 [1, 2], the stethoscope has been used in every day's work of the physicians and has become a symbol of healthcare professionals. Before that, doctors placed their ear directly over the chest to hear the sounds made by the heart and lungs. The binaural stethoscope was developed in 1851 and commercial production started next year, with many refinements of the stethoscope made until nowadays. Cardiac auscultation provides a quick diagnostic information about the heart, while phonocardiography (PCG) allows the heart sounds and murmurs to be recorded and analysed as graphic information. Heart sounds are produced by pumping action of the heart, which causes vibrations of cardiac and other structures. Major heart sounds are first heart sound or S1 that occurs at the beginning of ventricular contraction and second heart sound or S2 that occurs at the end of the ejection phase of the systole and is related to the closure of the aortic and pulmonary valves [3]. If blood flow is turbulent due to abnormal narrow passages, heart murmurs are produced and they can be auscultated and documented in PCG.

The mechanical action of the heart is in close relation to the ionic currents through the membranes of the cardiac cells, and in their vicinity. These currents produce varying voltages or electric potentials over the body that can be measured in electrocardiography (ECG). In fact, the start of mechanical action (systole) is preceded by ionic currents for atrial action, which can be seen in ECG as P wave, and for ventricular action as QRS complex. The QT interval (from the start of Q wave to the end of T wave) is electrical systole, while S1-S2 interval (from the start of S1 sound to the start of S2 sound) is mechanical systole. The time relationship between mechanical (PCG) and electrical (ECG) events can be determined with accurate measurement of both phenomena at the same time. The Q-S1 interval (from the start of the Q wave to the start of S1 sound) is known as electromechanical activation time (EMAT) and is the time from the start of electrical activity in the ventricles to the closure of mitral valve that normally closes before the tricuspid. EMAT is used to predict cardiac adverse events and discharge outcomes in patients with heart failure [4,5], and was shown to be strongly associated with impaired LV contractility [6]. The interval from the end of T wave to the beginning of S2 sound is electro-mechanical window (EMW). EMW is used to predict arrhythmias: in patients [7], animal models [8], and computer simulations [9].

Phonopneumogram, which is a recording of broncho-pulmonary sounds, can be used for assessment of possible lung pathology. Furthermore, respiratory rate can be extracted from ECG measurements [10], which proved to be useful in patients with chronic obstructive pulmonary disease [11]. Therefore, PCG synchronized to ECG could evaluate both, cardiac and respiratory system. Automated computerized lung sound analysis and telemedicine applications have been described [12-17].

In our study, we used a conventional stethoscope with a microphone, an ECG sensor and a smart device (phone or tablet) rather than a commercial PCG+ECG instrument, because we did not want to be limited by the characteristics of a commercial device. We present the construction of the PCG device, it's synchronization with the wireless ECG, and measurements of timing relations of various mechanical and electrical events in nine people.

II. PRESENTATION

A. Materials and Methods

A phonocardiographic device was constructed using a conventional stethoscope and a microphone of a smartphone accessory inserted into the stethoscope tube (Fig. 1) and connected to a smart device (Samsung SM-T113). The signals were recorded by WaveEditor for Android™ (V1.82) application in WAV format with a 44.1 kHz sampling frequency and a 16-bit resolution. The ECG was recorded simultaneously on the same smart device with wireless ECG sensor (Savvy®, Saving d.o.o., Slovenia) and MobECG (Jožef Stefan Institute, V1.8.1) application with a sampling frequency of 256 Hz. Wireless electrocardiographic method was described in details previously [18].

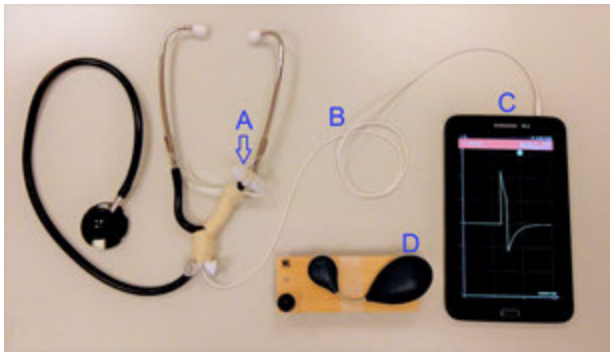


Figure 1: Phonocardiographic device was constructed from a microphone (A) inserted into a stethoscope tube and via a wire (B) connected to a smart device (C). Wireless ECG sensor (D) is attached to a device that generates acoustic and electric test signals (its schematics is shown in Fig. 2).

As each of the samplings, that is of PCG and ECG data, runs with its own free running oscillator, it was necessary to synchronize the recordings. We made a simple circuit (Fig. 2) that generates acoustic test signal for the PCG recording and an electrical one for the ECG. We recorded test signals before and after each measurement made in nine subjects. This allowed us to accurately synchronize PCG and ECG recordings with linear time interpolation. Without such correction, there would be a time lag of about 2 milliseconds per minute of recording, because relative difference of frequencies between the oscillators was around 34 parts per million. The synchronization signal for ECG and for PCG is shown in Fig. 3.

When the key in Fig. 2 was pressed, the acoustic test signal of damped oscillation started with positive deflection. At this moment, the membrane of the electromagnetic audio transducer (QMB-06, Mantech) first moved towards inside, therefore the membrane of the stethoscope, that was placed around 0.5 cm from the transducer, moved to the outside direction. When stethoscope was placed on the skin, the positive wave from the heart means that the skin moved to the inside direction of the body.

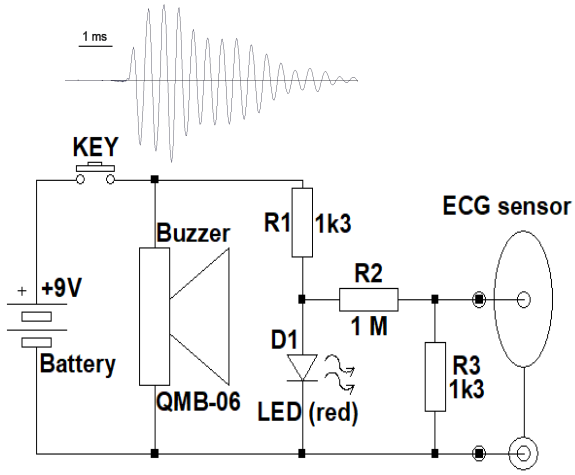


Figure 2: Circuit diagram of a test signal generator (below). When the KEY is depressed, the audio transducer QMB-06 generates an acoustic test signal of damped oscillation at a frequency of around 2.2 kHz (above). At key release, the waveform is similar but with opposite polarity.

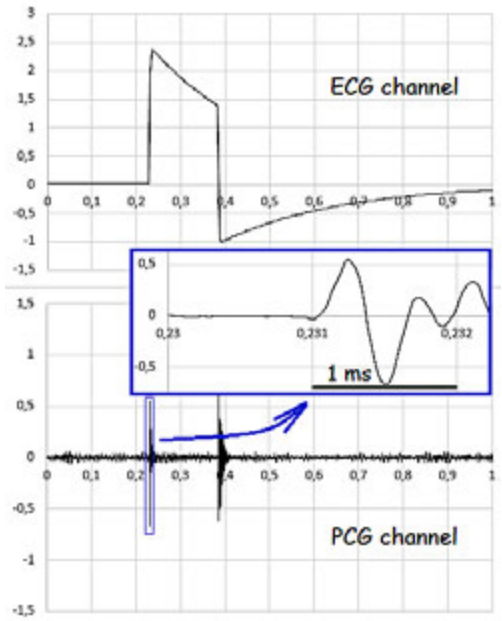


Figure 3: Synchronization signals for ECG and PCG channel from generator shown in Fig. 2 when the KEY was pressed for a short time interval of approximately 0.16 s. The high sampling rate of PCG channel enables precise localization (0.1 ms) of the start of the sound waveform.

Simultaneous PCG and ECG measurements were made in 9 volunteers: 4 females and 5 males, who were 23-64 (43±17) years of age. While volunteer was sitting comfortably, we presented the purpose of the measurements, asked for written consent and then we switched on the PCG and ECG devices, and recorded the synchronised signals as described earlier.

The wireless ECG sensor was attached to the chest with Skintact T-60 electrodes at an approximate position of V1 and V2 of standard 12-lead ECG; the positive electrode of the ECG sensor was on the position of V2. This non-standard ECG lead is bipolar and measures the V2-V1 difference of the standard ECG leads. Position of the stethoscope's head, where S1 and S2 heart sounds were best audible, was determined by auscultation of the area inferior to the ECG sensor. The stethoscope's head was kept as calm as possible for at least 10 seconds. The ECG sensor was then removed and the signals were synchronised again as described previously.

The time resolution of the ECG recordings was 4 ms due to 256 Hz sampling. Since this was not enough, filtering with 40 Hz low-pass filter and interpolation of ECG signal were used to increase the resolution to 1 ms before the time instants were determined. Measurements of electrical events were made with VisECG program (Jožef Stefan Institute, V1.2.20): beside the start of synchronisation signal at the beginning and at the end of the measurement, R wave peak (or S wave peak, if R wave was not pronounced), start of Q wave and end of T wave were recorded for 5 consecutive heartbeats. Tangent method was used to define the end of the T wave that was at the intersection of downslope tangent with baseline.

PCGs were analysed with Audacity program (V2.2.2): beside the start of synchronisation signal at the beginning and at the end of the measurement, start of S1 and start of S2 were obtained for 5 consecutive heartbeats.

For the following intervals average values of 5 measurements and their standard deviations were computed in Excel:

- electrical systole or QT interval (from the start of Q wave to the end of T wave),
- electromechanical activation time (EMAT) or Q-S1 interval (from the start of Q wave to the start of S1 sound),
- Q-S2 interval (from the start of Q wave to the start of S2 sound),
- mechanical systole or S1-S2 interval (from the start of S1 to the start of S2),
- electromechanical window or EMW (from the end of T wave to the beginning of S2 sound = (Q-S2) – QT,
- R-R interval (between the peaks of R waves),
- Q-Q, interval (between the starts of Q waves),
- S1-S1 interval (between the S1 sounds), and
- S2-S2 interval (between the S2 sounds).

B. Results

The complete results for the volunteer with the highest heart rate variability (P8, 27 years old) are shown in Table I and Table II. Although the variability of R-R intervals is large (SD of R-R interval in Table II is 68 ms), the variabilities of electrical systole (QT), mechanical systole (S1-S2), electromechanical activation time (Q-S1) and electromechanical window (EMW) are small (<5 ms). An example of S2 sound details is presented in Fig. 4.

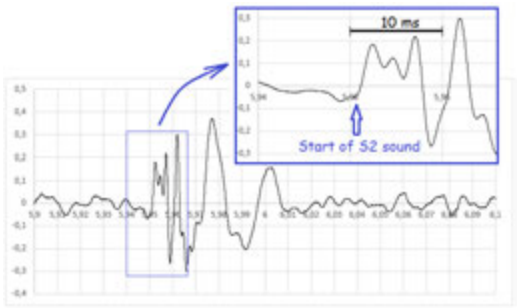


Figure 4: An example of cardiac sound S2 that is produced by aortic valve closure has a rapid rise in the PCG signal. This enables precise time localization of the start of S2 (< 1 ms).

TABLE I. ELECTRICAL SYSTOLE (QT), Q-S1 AND Q-S2 INTERVALS, MECHANICAL SYSTOLE (S1-S2) AND ELECTROMECHANICAL WINDOW (EMW) FOR 5 CONSECUTIVE HEARTBEATS, AVERAGE VALUE AND STANDARD DEVIATION

Beat No.	QT [ms]	Q-S1 [ms]	Q-S2 [ms]	S1-S2 [ms]	EMW [ms]
1	349	68	411	343	62
2	342	70	412	342	70
3	344	70	404	334	60
4	343	70	411	341	68
5	352	68	415	347	63
Average	**346.0**	**69.5**	**410.9**	**341.4**	**64.9**
SD	**4.3**	**1.1**	**4.0**	**4.7**	**4.2**

TABLE II. INTERVALS FROM BEAT TO BEAT, MEASURED BETWEEN CONSECUTIVE HEARTBEATS FROM DIFFERENT WAVES IN ECG AND SOUNDS IN PCG FOR 5 HEARTBEATS

Beat No.	R-R [ms]	Q-Q [ms]	S1-S1 [ms]	S2-S2 [ms]
1	1163	1162	1165	1165
2	1201	1201	1203	1202
3	1079	1080	1080	1072
4	1028	1029	1029	1036
5	1132	1131	1129	1135
Average	**1120.6**	**1120.6**	**1121.2**	**1122.0**
SD	**68.4**	**67.7**	**68.7**	**67.7**

The time intervals of all 9 volunteers are shown in Tables III and IV. In Table III there are two outliers in the column of Q-S1 parameter or electromechanical activation time (EMAT), where high variability is present in third and seventh volunteer (15 and 16 ms respectively). It is in these two recordings that the start of S1 sound could not be as well determined because of low PCG signal-to-noise ratio. Variability of measured intervals is associated with signal quality. The electro-mechanical window (EMW) had a positive value in all volunteers and measured 34 ± 20 ms. As can be seen in Table IV the intervals between beats, determined from the ECG or from the PCG recordings differ by only a few milliseconds (compare values in one row).

TABLE III. AVERAGE VALUES AND STANDARD DEVIATIONS (SD) FOR 5 INTERVALS AS IN TABLE I FOR THE 9 VOLUNTEERS

ID	QT [ms]	Q-S1 [ms]	Q-S2 [ms]	S1-S2 [ms]	EMW [ms]
1	388±6	61±3	407±4	346±7	19±7
2	336±4	47±3	351±2	304±2	15±4
3	336±2	39±15	395±6	356±10	59±5
4	350±9	46±6	363±9	317±10	13±9
5	386±4	74±4	431±4	357±7	45±7
6	313±6	56±1	334±4	278±5	21±6
7	355±5	70±16	377±8	307±21	22±10
8	346±4	70±1	411±4	341±5	65±4
9	313±2	70±2	359±3	289±3	46±2
all	336±34	59±13	381±32	329±29	34±20

TABLE IV. AVERAGE VALUES AND STANDARD DEVIATIONS (SD) FOR 4 INTERVALS AS IN TABLE II FOR THE 9 VOLUNTEERS

ID	R-R [ms]	Q-Q [ms]	S1-S1 [ms]	S2-S2 [ms]
1	1242±20	1242±22	1242±19	1242±23
2	856±17	856±16	855±17	855±18
3	970±12	970±12	968±19	969±13
4	927±20	925±20	928±21	926±22
5	1111±31	1110±27	1110±33	1108±25
6	749±15	750±14	750±16	747±12
7	802±27	803±28	796±32	803±21
8	1121±68	1121±68	1121±69	1122±68
9	859±51	859±51	860±50	860±52

III. DISCUSSION

With the PCG + ECG device presented in our paper accurate results were obtained for the electromechanical events. We did not expect that the mechanical and electromechanical intervals would be as stable for 5 consecutive beats, we were worried that the ECG with a sampling rate of 256 Hz would not be sufficient to give an exact insight into the dynamics of the heart, which would be possible by means of higher resolution. The presented device hence cannot be used for medical diagnosis, although we made every effort to make the results credible. We confirmed that the durations of electrical and mechanical activity are closely related. The variabilities of electrical systole (QT interval), mechanical systole (S1-S2 interval), electromechanical activation time (Q-S1 interval), and electromechanical window (T-S2) were all small, except for the two outliers where the start of S1 sound could not be well determined because of low PCG signal-to-noise ratio. Variability of measured intervals is associated with signal quality, as was shown for QT intervals previously [19,20].

Although the electrical activity of ventricles (start of QRS complex) always precedes the mechanical activity, the end of mechanical activity may be before or after the end of T wave. EMW is the time interval from the end of T wave to the closure of aortic and pulmonary valves and has positive value if the T wave ends before the valve closure and vice versa. Graphic presentation of EMW definition can be seen in a separate paper [21]. In our healthy individuals EMW was always positive and measured 34 ± 20 ms, which is in close agreement with previously reported values for healthy subjects, 26 ± 13 ms [22]. In drug development research and pharmacological industry, it is important to identify drug-induced cardiac adverse events in early stages. In computer simulations EMW has been shown as an effective biomarker for prediction of arrhythmias [9]. Negative value of EMW is most pronounced in patients with long QT syndrome with documented arrhythmic events [7], and has been proposed as a risk marker for drug induced arrhythmias in a canine model [8]. EMW can be measured invasively [8] or noninvasively either by ultrasound [7] or by simultaneous measurement of ECG and PCG, as was done in this study. The precision of the EMW measurements by the latter method depends on exact determination of the end of T wave in ECG and the beginning of S2 sound in PCG. We have shown that the onset of S2 can be determined to an accuracy of better than 1 ms, while the determination of the end of the T wave is more difficult, because the last part of the T wave is variable and its end difficult to determine. Stable results can be obtained by using the tangent method, as was done in our study, where we propose high resolution simultaneous ECG and PCG measurements that are a simple and cost-effective way to determine EMW, an indicator of cardiac electro-mechanical interactions.

Furthermore, the low-cost device constructed in our study can be used for evaluation of heart murmurs and lung sounds. It can also be beneficial for teaching and in telemedicine. Electronic stethoscopes that record PCG proved to be a valuable teaching aids for medical students [23,24]. Similarly, educational function has been presented for simple low-cost PCG systems constructed from stethoscope, microphone, and smart-phone [25]. Heart murmurs recorded by electronic stethoscopes can be sent via e-mail for telemedical evaluation [26]. Wireless PCG [27,28], developed acoustic sensors [29], and novel methods of simultaneous PCG and ECG [30] have been described, and modern technology that is nowadays so easily accessible, is certainly going to contribute to more remarkable evaluations of cardiac dynamics.

IV. Conclusion

Digital era brought us exciting novel methods of PCG and the possibility to study PCG synchronised to ECG provides a great teaching material. Students can gain the ability to differentiate various cardiac sounds and murmurs if auscultation is practiced diligently, because the sound phenomena are relatively short and silent. PCG devices are useful simple diagnostic tools, which can also be employed in telemedicine. With the advent of echocardiography, the art of cardiac auscultation has lost its glamour, however, a detailed insight into cardiac electro-mechanical interactions is nowadays possible with ubiquitous smart devices and precise multifunction sensors, be it either in the hands of researchers or medical practitioners.

Acknowledgment

This work was supported by the Slovenian Research Agency (grant P2-0095; V. Avbelj).

References

[1] R. T. H. Laënnec, "De l'Auscultation Médiate ou Traité du Diagnostic des Maladies des Poumons et du Coeur", Paris: Brosson & Chaudé, 1819; 1–14.

[2] R. Hajar, "The art of listening", Heart views 2012; 13: 24–25.

[3] M. A. Chizner, "Cardiac auscultation: rediscovering the lost art", Curr Probl Cardiol 2008; 33: 326–408.

[4] J. Zhang, W. X. Liu and S. Z. Lyu, "Predictive value of electromechanical activation time for in-hospital major cardiac adverse events in heart failure patients", Hindawi Cardiovasc Ther 2020; 532596, 6 pp.

[5] T. F. Chao, S. H. Sung, H. M. Cheng, W. C. Yu, K. L. Wang, C. M. Huang and C. H. Chen, "Electromechanical activation time in the prediction of discharge outcomes in patients hospitalized with acute heart failure syndrome", Intern Med 2010; 49: 2031–2037.

[6] S. Efstratiadis and A. D. Michaels, "Computerized acoustic cardiographic electromechanical activation time correlates with invasive and echocardiographic parameters of left ventricular contractility", J Cardiol Fail 2008; 14: 577–582.

[7] R. M. A. Bekke, K. H. Haugaa, A. van den Vijngaard, J. M. Bos, M. J. Ackerman, T. Edvardsen and P. G. Volders, "Electromechanical window negativity in genotyped long QT syndrome patients: relation to arrhythmia risk", Eur Heart J 2015; 36: 179–186.

[8] H. J. van der Linde, B. van Deuren, Y. Somers, B. Loenders, R. Towart and D. J. Gallacher, "The electromechanical window: a risk marker for Torsade de Pointes in a canine model of drug induced arrhythmias", Br J Pharmacol 2010; 161: 1444–1454.

[9] E. Passini, C. Trovato, P. Morisette, F. Sannajust, A. Bueno-Orovio and B. Rodriguez, "Drug-induced shortening of the electro-mechanical window is an effective biomarker for in silico prediction of clinical risk of arrhythmias", Br J Pharmacol 2019; 176: 3819–3833.

[10] R. Trobec, A. Rashkovska and V. Avbelj, "Two proximal skin electrodes – a respiration rate body sensor," Sensors (Basel) 2012; 12: 13813–13828.

[11] I. Tomasic, N. Tomasic, R. Trobec R, M. Krpan and T. Kelava, "Continuous remote monitoring of COPD patients–justification and explanation of the requirements and a survey of the available technologies", Med Biol Eng Comput 2018; 56: 547–569.

[12] R. L. Murphy, A. Vyshedskiy, V. A. Power-Charnitsky, D. S. Bana, P. M. Marinelli, A. Wong-Tse and R. Paciej, "Automated lung sound analysis in patients with pneumonia", Resp Care 2004; 49: 1490–1497.

[13] E. Andrès, A. Hajjam and C. Brandt, "Advances and innovations in the field of auscultation, with a special focus on the development of new intelligent communicating stethoscope systems", Health Technol 2012; 2: 5–16.

[14] A. Gurung, C. G. Scrafford, J. M. Tielsch, O. S. Levine and W. Checkley, "Computerized Lung Sound Analysis as diagnostic aid for the detection of abnormal lung sounds: a systematic review and meta-analysis", Respir Med 2011; 105: 1396–1403.

[15] R. L. H. Murphy, A. Vyshedskiy, V. A. Power-Charnitsky, D. S. Bana, P. M. Marinelli, A. Wong-Tse and R. Paciej, "Automated lung sound analysis in patients with pneumonia", Respir Care 2004; 49: 1490–1497.

[16] C. Hernandez, J. Mallow and G. L. Narsavage, "Delivering telemedicine interventions in chronic respiratory disease", Breathe 2014; 10: 198–212.

[17] B. A. Reyes, N. Reljin, Y. Kong, Y. Nam, S. Ha and K. H. Chon, "Towards the development of a mobile phonopneumogram: automatic breath-phase classification using smartphones", Ann Biomed Eng 2016; 44: 2746–2759.

[18] R. Trobec, I. Tomašić, A. Rashkovska, M. Depolli and V. Avbelj, "Body sensors and electrocardiography", Springer Briefs in Applied Sciences and Technology 2018, ed. J. Kacprzyk. Springer, Basel, Switzerland: 107–114.

[19] V. Avbelj, R. Trobec and B. Gersak. Beat-to-beat repolarisation variability in body surface electrocardiograms. Med Biol Eng Comput 2003; 41: 556–560.

[20] M. Baumert, A. Porta, M. A. Vos, M. Malik, J. P. Couderc, P. Laguna, G. Piccirillo, G. L. Smith, L. G. Tereschchenko and P. G. A. Volders, "QT interval variability in body surface ECG: measurement, physiological basis, and clinical value: position statement and consensus guidance endorsed by the European Heart Rhythm Association jointly with the ESC Working Group on Cardiac Cellular Electrophysiology", Europace 2016; 18: 925–944.

[21] V. Kadunc Kos, M. Brloznik, A. Domanjko Petrič and V. Avbelj, "Simultaneous phonocardiography and electrocardiography using smartphone in dogs, cats and horses", Proceedings of 43rd International Convention (MIPRO), Opatija, Croatia, 2020.

[22] H. Boudoulas, Y. H. Sohn, W. O'Neill, R. Brown and M. Weissler, "The QT>QS2 syndrome: a new mortality risk indicator in coronary artery disease" Am J Cardiology 1982; 50: 1229–1235.

[23] M. E. Legget, M. Toh, A. Meintjes, S. Fitzsimons, G. Gamble and R. N. Doughty, "Digital devices for teaching cardiac auscultation – a randomized pilot study", Med Educ Online 2018; 23: 1–7.

[24] H. Hoyte, T. Jensen and K. Gjesdal, "Cardiac auscultation training of medical students: a comparison of electronic sensor-based and acoustic stethoscopes", BMC Medical Education 2005; 5: 14.

[25] L. J. Thoms, G. Colicchia and R. Girwidz, "Phonocardiography with a smartphone", Phys Educ 2017; 52: 1–4.

[26] L. B. Dahl, P. Hasvold, E. Arild and T. Hasvold, "Heart murmurs recorded by a sensor based electronic stethoscope and e-mailed for remote assessment", Arch Dis Child 2002; 87: 297–301.

[27] H. Mondal, S. Mondal and K. Saha, "Development of low-cost wireless phonocardiograph with a Bluetooth headset under resource-limited conditions", Med Sci 2018; 6. 117–125.

[28] A. Sa-Ngasoongsong, J. Kunthong, V. Sarangan, X. Cai and S. T. Bukkapatnam, "A low-cost, portable, high-throughput wireless sensor system for phonocardiography applications", Sensors 2012; 12: 10851–10870.

[29] R. Abasi-Kesbi, A. Valipour and K. Imani, "Cardiorespiratory system monito-ring using a developed acoustic sensor", Health Technol Lett 2017; 5: 7–12.

[30] N. Giordano and M. Knaflitz, "A novel method for the timing of heart sound components through digital phonocardiography", Sensors 2019; 19: 1868–1884.

Detection of Uninterpretable ECG Signal Segments

E. Ajdaraga Krluku and M. Gusev
Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering, Skopje, North Macedonia
Email: era.ajdaraga@students.finki.ukim.mk,
marjan.gushev@finki.ukim.mk

***Abstract*—Remote diagnosis represents one of the fundamental reasons for the introduction of telemedicine services. Specialized wearable health monitoring devices collect large amounts of data, which are transmitted to cloud collection centers for further monitoring and interpretation. However, the presence of noise corrupts the ECG signals, especially in wearable sensors, due to physical activities and movements. This significantly decreases the diagnosis accuracy and performance. Therefore, timely noise detection and identification of uninterpretable ECG segments are crucial for wearable devices.**

In this paper, we present results from our research to detect noisy segments in ECG signals without a goal to eliminate them and improve the QRS detection, but to detect where QRS detection would be impossible and avoid detection and interpretation mistakes. Our work includes two algorithms and multiple related variables that add to the success of the proposed algorithms. Finally, we achieved high performance for detecting signals where the signal to noise ratio is lower than 6 dB, with sensitivity and a positive predictive rate of over 90%.

***Index Terms*—Cloud Computing, ECG, Noise Detection**

I. Introduction

The ubiquity of IoT devices and the growing processing capabilities of cloud technologies allowed the implementation of computer science in many other fields of study. In the field of medicine, these advancements have allowed for the introduction of continuous monitoring devices for numerous vital parameters. Lee et al. [1] predict that the number of IoT devices will reach 26 billion units by 2020. The authors note that by utilizing big data it is possible to completely transform the healthcare product industry.

Furthermore, Hiremath et al. [2] predict that by 2024 wearable IoT devices will transform healthcare practices "for which the cost of treatments is low; early detection of illnesses is possible; efficient treatment is easy to administer remotely; and patients do not need to see their physicians so often". They note that there is significant demand for wearable IoT technologies for the management of chronic diseases.

Such a solution is offered by ECGalert [3], which is a specialized monitoring system of ECG signals. The system records the ECG of the user through the electrodes of a special recording device called Savvy [4], which in turn communicates with the user's mobile phone via Bluetooth. Through the mobile phone the data is relayed to a cloud server, which is responsible for the rest of collection, processing and diagnostic operations.

The analysis performed by the cloud is the most crucial step, but at the same time, it is a process that is very vulnerable to noise and other anomalies in the ECG signal. Noise corruption may result in an incorrect diagnosis, which in turn may cause serious anomaly indicators to go unnoticed.

Issues with noise corruption are not limited only to telemedicine devices such as ECGalert. This is a problem that also affects more conventional ambulatory devices, such as the Holter.

Because of the implications that noise corruption has on the accuracy of the provided diagnosis, it is evident that a solution is needed. In the majority of these cases, corruption is caused by motion interference resulting from movement of the patient or the electrode. Another type of noise mostly appears when the electrodes of the ECG recording device are not properly placed, or have come into contact with sweat or other elements that may interfere with the skin-to-electrode electrical transmission. In addition, artifacts are QRS-like signals with abnormalities which are not generated by the heart. They may appear as peaks with similar QRS-like shape but located on an unexpected distance of the previous heartbeat, or their shape is distorted.

The first step in eliminating noise corruption is to identify when it appears - preferably in real-time so that the user can be immediately notified and guided in how to avoid noise corruption in the future. Ultimately, the goal is to not only identify this noise, but also be able to extract the original ECG from the distorted signal.

In this paper, we introduce two algorithms based on a combination of a filter and expected threshold value from a typical ECG signal. Our research considers the performance of the following algorithms by their success in detecting noise in ECG signals.

This paper is organized as follows, Section II gives insight on published related work, Section III gives details of the applied methodology, while Section IV presents the results. A discussion is given in Section V, and finally, a conclusion can be found in Section VI.

II. Related work

Rodrigues et al. [5] propose an agglomerative clustering approach to the problem of artifact detection in ECG signals. The authors present the usage of multiple methods, such as standard deviation, zero-crossing rate, peak rate, and peak distance to extract valuable information from the signal. With the use of clustering, clean records can be differentiated from the noise. However, the authors note that the success of the algorithm decreases as the length of the record increases.

Lee et al. [6] propose a two-step detection method where they utilize the first-order intrinsic mode function (F-IMF).

After the IMF is normalized, they compare each segment against three defined thresholds: Shannon entropy, mean, and variance. If all three thresholds are exceeded, the segment is deemed as noisy. Once detected, the corrupted segments are excluded from further ECG analysis.

Varon et al. [7] propose a robust solution to detecting ECG artifacts, which they gear toward the more precise detection of sleep apnea through ECG signals. Their solution uses the Fourier transform, against traditional methods. They claim the algorithm aims to identify artifacts and not study beat abnormalities or rhythm changes, the use of the Fourier transform is appropriate. They achieve satisfactory results, but the authors note the algorithm cannot be utilized in a real-time environment.

Moeyersons et al. [8] propose the use of decision trees as a tool for identifying artifacts in ECG signals. Similarly to the previous [7] study, this algorithm also makes use of the autocorrelation function (ACF). The authors recommend a 5s segment duration, because this approach is a better indicator of a lack of similarity between segments. The authors note that this algorithm can be applied in a real-time scenario, however, the training dataset only consisted of files that were recorded during sleep, so a drop in performance is expected.

We have performed an initial study on analyzing the differential noise detection filter in ECG signals [9]. Furthermore, in this paper, we continue with research and present more compact results with comparative analysis of two different filters.

III. Methodology

In this section, we explain the applied methodology and give insight into the analyzed dataset and realized experiments. To make the explanations more comprehensible, we have divided the information into the following four subsections Algorithms, Evaluation methodology, Dataset, Definition of BASE and Experiments.

A. Algorithms

We introduce two algorithms based on measuring the energy of the input signal as the area under the squared magnitude of ECG signal.

Our idea for the first algorithm denoted by Simple Differential Filter (SDF) is based on evaluating the value of the input data sample and comparing it to the previous sample, so it is essentially a simple differential filter as defined by (1).

$$y[n] = |x[n] - x[n-1]| \tag{1}$$

As the data is gathered by the Savvy [4] wearable sensor, in essence, what is measured is the voltage of the electrical signal that is emitted by the heart when it pumps blood. The wearable sensor has an integrated AD converter which translates the analog values into a fixed set of digital values. The size of the digital value set is called bit resolution. The bit resolution plays a significant role in the definition of the next reviewed algorithm.

The Mean Differential Filter is defined in equation (2), where this time the differentiation is between the current sample and a zero-value signal which is predetermined depending on the bit resolution. For example, if the bit resolution is 10 bits (meaning the possible values range between 0 and 1023), the zero-value signal is precisely the middle of this scale, i.e. $\overline{X}$ takes on the value 511.

$$y[n] = |x[n] - \overline{X}|, \tag{2}$$

The second step of the algorithm realizes a summation of the output elements to evaluate the integral (energy) part of the analyzed segment. This operation represents the measured energy contained within the electrical activity of the heart up to that point.

Our research also takes into consideration a set of different filter lengths that are multiples of the sampling frequency (in our case 125 Hz). Sampling rate (or sampling frequency) is the rate at which a measurement sample is generated by the AD converter, (in our case 125 samples per second).

The implemented summation algorithm generates an output with a value defined in (3).

$$\sum_{n=0}^{len-1} y[n] \tag{3}$$

where len is the predefined filter length. Because the filter capacity is predetermined, this sum is recalculated each time a new element replaces the oldest value in the filter.

B. Evaluation methodology

Test results are matched for standard statistical analysis that checks if a sample belongs to a detected noise segment that is annotated in the reference file. This evaluation includes evaluation of true positive (TP) for the correct detection of noise; true negative (TN) for the correct detection of a non-noisy sample; false positive (FP) for incorrect detection of a noisy sample and false negative (FN) for incorrect detection of a non-noisy sample.

To evaluate the performance we use standard statistical measures: sensitivity (SEN) as a portion of TPs in regards to the total number of noisy samples, specificity (SPC) as a portion of TNs in regards to the total number of non-noisy samples and positive predictive value (PPV) as a portion of TPs in regards to all detected noisy samples.

The best algorithm depends on achieving maximal values for SEN, SPC. and PPV at the same time. However, a typical performance may fluctuate between proportions of these parameters, such that if SEN scores a higher value, it is to the detriment of SPC and PPV and vice-versa.

To estimate an overall performance we use F1 score defined by (4) as a harmonic mean of SEN and PPV.

$$\frac{1}{F1} = \frac{1}{2}\left(\frac{1}{SEN} + \frac{1}{PPV}\right) \tag{4}$$

C. Dataset

For the purpose of this research, we have used the NST database [10] available on the Physionet website [11]. This database is construed upon two records (118 and 119) from the standard MITDB Arrhythmia database [12]. Six files are constructed for both these records by adding noise to the signal in specific ECG segments, such that the Signal to Noise Ratio (SNR) equals to $24dB$, $18dB$, $12dB$, $06dB$, $0dB$ and $-6dB$ resulting in a total of 12 ECG files that contain ECG signals in the NST database. The noise in this dataset is specially created and covers a wide range of noise, including a large set of different artifacts, or QRS-like peaks with high degree of deformations, such as randomly located shorter intervals between QRS-like peaks, random false peaks, white noise with different spectrum etc.

Our analysis showed that our QRS detector [13] based on improved Hamilton code [14] performs well in case the noise level is below $6dB$ when compared to the signal. Therefore, our goal is to identify noise segments in files with SNR below $6dB$, such as is the case with files $118e_6$, $119e_6$, $118e00$ and $119e00$ from the NSTDB.

D. Definition of BASE

One of the crucial elements in the proposed algorithms is the threshold. This threshold value is separately calculated for each combination of the pair (algorithm, filter length). Every sum obtained in the sliding-window manner is then checked against this threshold value. If the sum is higher than the threshold, the algorithm classifies the sample as noisy.

The threshold calculation follows the following steps:

- A noise-free file (in our case, file 118 from the MITBIH database [12]) is processed with each of the two reviewed algorithms and each of the nine filter lengths. The output for each pair is a file containing the entirety of the computed sums in a sliding window manner, where the offset is equal to 1, i.e. the sum is recalculated each time a new element is added to the filter and the oldest element is removed.
- A single value (threshold) is computed for each combination. This value is computed using the following formula:

$$\frac{\sum_{n=0}^{len/50} y[n*i]}{len/50}, i \in (0, ..., 50) \tag{5}$$

  where *len* is equal to the total number of sums contained within the corresponding file generated in the previous step. y[i*n] represents the sum of values from n*i to (n*i+*filter length*). In short, this step generates a value that is the average of the generated sums computed in a sliding window manner, where the offset is equal to 50. So for example, if the filter length is 250 we take into this average the sums 0 to 250, 50 to 300, 100 to 350, and so on.
- We repeat this process for each (algorithm, filter length) pair, and subsequently each processed file from the NST database is matched to the corresponding pair for further processing.

To keep matters concise, we will denominate the threshold value computed in the previous manner with BASE. It is important to note that each (algorithm, filter length) pair has its own individually-computed value for BASE. In our experiments, we have considered three variations which contain the BASE value as a threshold. T_1 is equal to the original value of BASE, T_2 is equivalent to 2*BASE, and finally, T_3 is computed as BASE+(BASE*2.5/7).

E. Experiments

In our research, we have chosen to work with data sampled at 125 Hz, because it is the frequency at which our wearable sensor collects information. However, the derived conclusions theoretically apply to any sampling frequency, without losing generality.

In this study, we have experimented with the following filter lengths: 250 (2s), 375 (3s), 500 (4s), 625 (5s), 750 (6s), 875 (7s), 1000 (8s), 1125 (9s), 1250 (10s).

These values were chosen because a normal heartbeat is 60 to 80 BPM, which corresponds to one QRS per second, or 0.75 QRS in one second respectively. By this principle, extending the filter length to 2 seconds and beyond, allows us to capture more QRS complexes. Therefore, we can more easily establish a 'normal' or mean value, which allows us to more accurately detect noisy segments.

IV. Results

From the realized experiments, the following results have been obtained. Table I provides an important overview of the maximal success rate for each algorithm, as well as the optimal filter length. In the cases where multiple filter lengths have scored the same results, the smallest filter length has been chosen as optimal because of the reduced resources the algorithm would need to use. Furthermore, it is important to note that T_3 has proven to be superior compared to T_1 and T_2.

Table II holds important information regarding the average F1-Score for each reviewed threshold and algorithm analyzed on a per-file basis. It is important to note that the averages are computed across all reviewed filter lengths. From the contained data we can conclude that for file $118e_6$ the best results are achieved when using the SDF algorithm and T_3 as a threshold.

For file $118e00$ the highest F1-Score is achieved for algorithm SDF and T_3 as a threshold value.

File 119e_6 also gives the best results when analyzed with SDF algorithm and the threshold set at T_3, while $119e00$ performs best when processed with SDF and T_1.

The optimal threshold value has been deduced experimentally. We have extended our test range for multiple values - from BASE, up to BASE * 2 with a step of 0.01. Through our experiments, we have found that for optimal F1-Score results, a threshold of BASE * 1.2 should be used. However, if we wish to achieve a higher PPV, then a threshold of BASE * 1.4 is recommended. Through these two experimental conclusions, it can logically be deduced that if we wish to obtain an optimum

TABLE I
MAXIMUM PER-FILE F1-SCORE BY ALGORITHM AND THRESHOLD

Algorithm	Optimal Length	Max F1_Score	Optimal Threshold
SDF	375	85%	T_3
MDF	750	68%	T_3

TABLE II
AVERAGE PER-FILE F1-SCORE BY ALGORITHM AND THRESHOLD

Algorithm	Filename	SNR	T_1	T_2	T_3
SDF	118e_6	-6	75%	76%	80%
SDF	118e00	0	72%	75%	80%
SDF	119e_6	-6	81%	75%	81%
SDF	119e00	0	87%	72%	85%
MDF	118e_6	-6	64%	70%	68%
MDF	118e00	0	64%	57%	65%
MDF	119e_6	-6	61%	70%	68%
MDF	119e00	0	61%	52%	65%

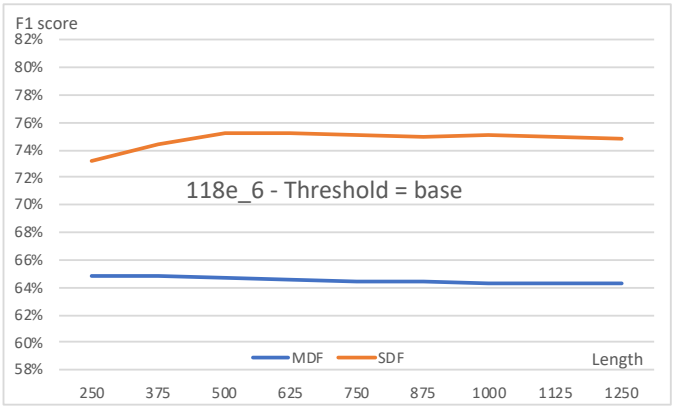


Fig. 1. F1 score results with Threshold=1.00 * BASE on NSTDB record 118e_06.

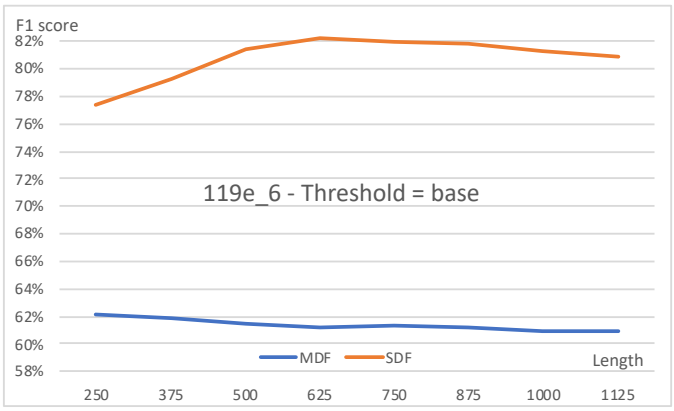


Fig. 2. F1 score results with Threshold=1.00 * BASE on NSTDB record 119e_06.

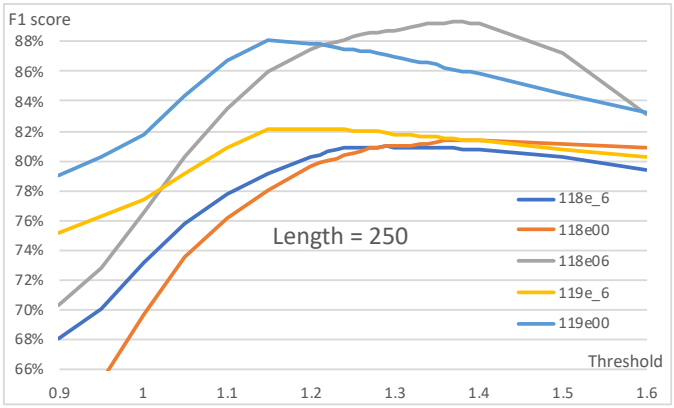


Fig. 3. F1 score results for Length = 250 on NSTDB records 118e_06 and 119e_06.

of peak values for both parameters, a threshold of BASE * 1.3 should be applied.

V. DISCUSSION

From the results, it is evident that the SDF algorithm vastly outperforms MDF. However, it is important to note that from the three reviewed thresholds, optimal results vary depending on which statistical parameter the judgment is based on.

The typical behavior of SDF and MDF algorithms is presented in Fig. 1 and Fig. 2, where the threshold is equal to the BASE value on the signal records $118e_6$ and $119e_6$ correspondingly. The SDF algorithm outperforms the MDF algorithm.

Note that the best performance of the SDF algorithm in the presented examples is achieved for a filter length of 500 and 625 correspondingly. However, this does not apply when optimizing for the F1 score. To demonstrate the overall spectra of results, we present a typical distribution of the F1 score for all records for the fixed length of 250 in Fig. 3, where the x-axis presents all tested threshold values.

The records $118e06$ and $119e06$ achieved better F1 scores, which is expected, since these recordings for which the SNR is significantly higher. All of the analyzed combinations in our experiments show an optimal value reached for a specific threshold value depending on the analyzed records. All optimal values are found for threshold values in the region between $1.15 * BASE$ and $1.4 * BASE$.

In order to find the optimal value which can reveal the best results for all analyzed signals in the database, we have used a harmonic mean of all results to find a relevant F1 score representation. The achieved results are presented in Fig. 4. It shows a clearer behavior and extraction of optimal values.

The results show that the optimal threshold value is between $1.27 * BASE$ and $1.28 * BASE$. We plotted another diagram to find the optimal filter length in Fig. 5. This analysis reveals that the best performance is achieved for filter length of 375, although 250 and 500 are also good candidates.

Finally, we present Fig. 6, which is the extraction of Fig. 4 showing only the performance of the experiments with length 250, 375, and 500. For threshold values smaller than 1.2*BASE the filter length 500 achieves better performance, and for threshold values over 1.2*BASE it is the filter length of 375.

A deeper analysis and performance evaluation is presented

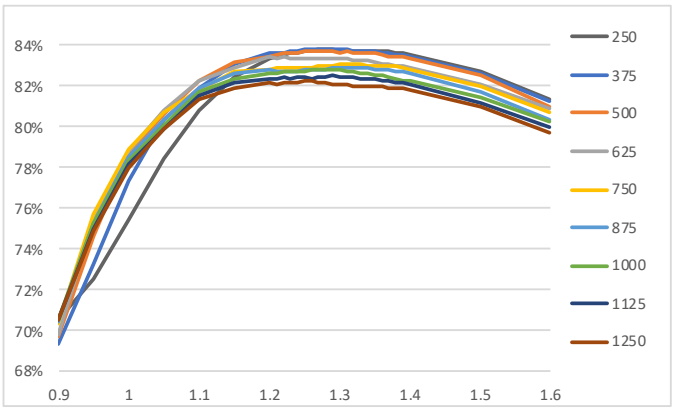


Fig. 4. F1 score results for harmonic mean of test signals on NSTDB.

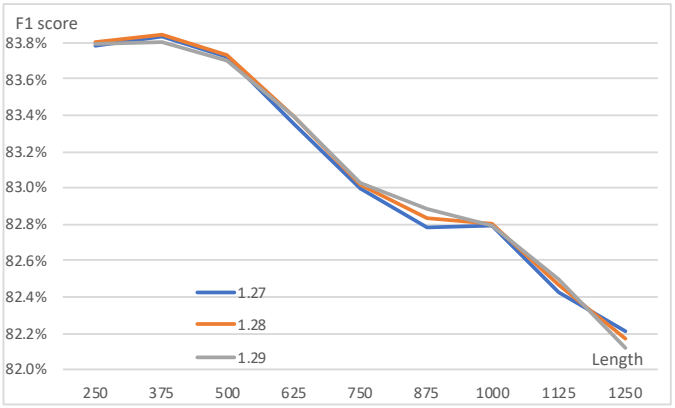


Fig. 5. F1 score results for harmonic mean of test signals on NSTDB.

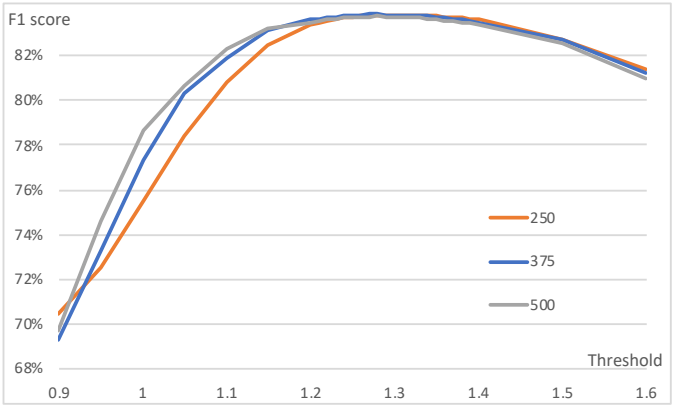


Fig. 6. F1 score results for harmonic mean of test signals on NSTDB.

in Fig. 7, where the optimal values are extracted to be a filter length of 375 and a threshold value of 1.28 * BASE.

From the obtained results, the following can be concluded:

- PPV reaches its peak (99% and 55% respectively) for algorithms SDF and MDF when T_2 is used as a threshold,
- F1 SCORE attains its maximum (89% and 68% respectively) for algorithms SDF and MDF when working with

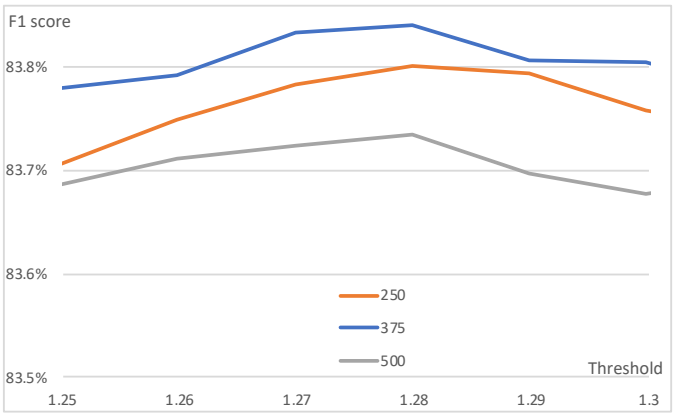


Fig. 7. F1 score results for harmonic mean of test signals on NSTDB.

T_3

Therefore, we may conclude that both T_2 and T_3 are adequate thresholds to be used when applying these algorithms.

VI. Conclusion

We presented two robust algorithms for the detection of uninterpretable (noise-corrupted) segments in ECG signals, which allow us to provide users with a more accurate diagnosis. Through experimental tests, we conclude that SDF is an optimal algorithm. In cases where accuracy and low memory requirements are both priorities, SDF would be the most adequate choice.

T_3 outperforms the remaining two proposed thresholds, but only in cases where the F1 score is the determining statistical parameter. We have offered a more thorough explanation in section V.

To conclude, there are multiple advantages to applying these algorithms as a pre-processing step, such as:

- Low processing and memory requirements for the algorithm execution
- Successful detection of noise segments
- Fast execution
- Real-time calculation

Future work will include more in-depth analysis of the reviewed parameters, as well as more advanced noise-detection with more complex algorithms, such as non-linear windowing, wavelet, correlation or similar.

References

[1] I. Lee and K. Lee, "The internet of things (iot): Applications, investments, and challenges for enterprises," *Business Horizons*, vol. 58, no. 4, pp. 431–440, 2015.

[2] S. Hiremath, G. Yang, and K. Mankodiya, "Wearable internet of things: Concept, architectural components and promises for person-centered healthcare," in *2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*. IEEE, 2014, pp. 304–307.

[3] M. Gusev, A. Stojmenski, and A. Guseva, "Ecgalert: A heart attack alerting system," in *International Conference on ICT Innovations*. Springer, 2017, pp. 27–36.

[4] A. Rashkovska, M. Depolli, I. Tomašić, V. Avbelj, and R. Trobec, "Medical-grade ECG sensor for long-term monitoring," *Sensors*, vol. 20, no. 6, p. 1695, 2020.

[5] J. Rodrigues, D. Belo, and H. Gamboa, "Noise detection on ecg based on agglomerative clustering of morphological features," *Computers in biology and medicine*, vol. 87, pp. 322–334, 2017.

[6] J. Lee, D. D. McManus, S. Merchant, and K. H. Chon, "Automatic motion and noise artifact detection in holter ecg data using empirical mode decomposition and statistical approaches," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 6, pp. 1499–1506, 2011.

[7] C. Varon, D. Testelmans, B. Buyse, J. A. Suykens, and S. Van Huffel, "Robust artefact detection in long-term ecg recordings based on autocorrelation function similarity and percentile analysis," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 3151–3154.

[8] J. Moeyersons, C. Varon, D. Testelmans, B. Buyse, and S. Van Huffel, "Ecg artefact detection using ensemble decision trees," in *2017 Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.

[9] E. Ajdaraga and M. Gusev, "Analysis of a differential noise detection filter in ecg signals," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2019, pp. 295–300.

[10] G. B. Moody, W. Muldrow, and R. G. Mark, "A noise stress test for arrhythmia detectors," *Computers in cardiology*, vol. 11, no. 3, pp. 381–384, 1984.

[11] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[12] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[13] E. Domazet and M. Gusev, "Improving the QRS detection for one-channel ECG sensor," *Technology and Health Care*, vol. 27, no. 6, pp. 623–642, 2019.

[14] P. Hamilton, "Open source ecg analysis," in *Computers in cardiology*. IEEE, 2002, pp. 101–104.

A Programmatic Approach for Development of the ViewHRV Service Platform with Accurate and Reliable Results

E. Shaqiri* and M. Gusev**
*Innovation Dooel, Skopje, North Macedonia
**Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering, Skopje, North Macedonia
E-mail: ervin.shaqiri@innovation.com.mk, marjan.gushev@finki.ukim.mk

***Abstract*—There are dozens of available packages and libraries that claim to calculate HRV. This paper aims at comparing the results from the calculation of a single array of intervals between normal beats, including the most popular open source Python HRV measurement packages available on GitHub. Furthermore, the same array was ran through the *Kubios* software and compared to the previous results. In order to compare the accuracy of the results, as a benchmark we used the C programs provided by *Physionet*. The results showed a huge difference in the results with almost all the indices, in fact the simplest measurement that of Standard Deviation of NN Intervals showed to be incorrect in *Kubios* and in most of the Python packages. Results like these are the reason we decided to develop our own package to calculate HRV. Finally, the goal of this paper is to present details on developing a publicly available web service platform ViewHRV with guaranteed precision obtaining accurate and reliable results.**

Keywords – *ECG signal*; *Heart Rate Variability*; *Python Library*; *Physionet*

I. Introduction

Heart Rate Variability (HRV) has been subject to a lot of research in the past decade. As a result, many different methods of calculating HRV indices have come about. Development of cloud computing and small Internet of Things devices have brought up the need to calculate these parameters programmatically with the requirements of accuracy, speed, and easy to understand codebase.

HRV refers to the change in the time intervals between adjacent heartbeats [1] and has been associated with the incidence of new cardiac events (angina pectoris, myocardial infarction, coronary heart disease death, or congestive heart failure) assessed with proportional hazards regression analyses [2]. In retrospect, HRV and Heart Rate (HR) have been used early on in human history dating back to 300 B.C. [3]. It was until 1996 when it was standardized and wide-spread acceptance of HRV as a clinical tool for the evaluation of cardiac autonomic changes in patients [3].

Cardiovascular disease (CVD) is the leading cause of morbidity and mortality in both men and women. This is particularly true in developed countries [4] [5] [6]. The National Heart, Lung, and Blood Institute of the US National Institutes of Health list eight risk factors for heart disease and stroke, six of which are modifiable. Three of these modifiable risk factors are associated with what could be called biological factors. They are high blood pressure (hypertension), diabetes, and abnormal cholesterol. Three others listed as modifiable could be considered lifestyle factors and include tobacco use (smoking), physical inactivity (exercise), and obesity. The two non-modifiable risk factors are age and family history of early heart disease or stroke. It is interesting to note that there is at least some data to suggest that each of these risk factors is associated with decreased HRV [4].

Most commonly Autonomic dysfunction is observed to exist in cardiac disorders or other diseases including myocardial infarction, syndrome of multiple organ dysfunction severe head and brain injury and sepsis [7]. HRV indices are found to be high for Premature Ventricular Contraction (PVC), Sick Sinus Syndrome (SSS), Atrial Fibrillation (AF) and low for Congestive Heart Block (CHB), Left Branch Bundle Block (LBBB), Ischemic/ dilated cardiomyopathy.

By the early 1960s, ambulatory electrocardiography was increasingly used to detect "phantom arrhythmias and evanescent cardiac abnormalities" to help explain clinical syndromes described by patients [8]. Holter ECG monitoring devices for long-term (one or two days) measurements that detect arrhythmia and automatically transmit asymptomatic and symptomatic events are the current standard in healthcare institutions around the world. These devices acquire data from patients both passively and with active patient participation. Abnormal detectable rhythm events may include atrial fibrillation, tachycardia, bradycardia, and pauses [9].

The Holter ECG measurement devices measure bio-signals very accurately. However, this device is heavy and bulky and it makes obstacles to physical movements of the patient. With the turn of the century and advancements in technology and especially in that of data communication, the consumer world has seen companies try to develop an apparatus that will be as effective as the Holter itself [9]. Georgiou et al. [10] have indicated that wearable devices can accurately measure bio-signals with the limitation of extreme exercise.

These wearable devices produce a lot of continuous data for even longer periods (expressed in more days and months) [11], [12] and it needs to be stored and processed by some server. Although these devices are preferred for commodity and easy to wear feature in everyday life activities, physical movements create problems in detection of heartbeats, such

various types of artifacts [13], [10] and therefore calculation of HRV indices.

With all that said, the goal of this paper is to process a single data set of time intervals between consecutive normal beats (NN intervals) through different open source Python packages available online, in order to get the HRV computations. Secondly, the idea is based on processing the same data set through a third party standalone software *Kubios* a software package that has gained wide spread popularity over the past years. Thirdly, in order to gain a point of reference to benchmark the results, the C programs offered by *Physionet* to calculate HRV and its indices are also fed the previously mentioned data set. With this we hope to shed some light on the state of the art regarding the Python implementation of HRV calculations, this is important since without the correct calculations then the clinical value of these wearables will be close to none. Finally, the comparison if the results from these different tools will give an indication of what is implemented correctly and what is not. If there are no consistent readings then that would mean that there is no reliable sowftware package and as a result a new one will have to be developed, as we have realized with our ViewHRV platform.

The rest of the paper is organized according to the following structure. Section II presents the related work and findings in the specific field. Section III presents the methods to accurately calculate HRV and identified obstacles, along with methods to compare the calculated HRV parameters from other software. The results are presented in Section IV by presenting the software architecture and functional description of the new developed ViewHRV platform. The evaluation results from conducted comparison and elaboration of our software performance are discussed in Section V. Finally, the conclusions and future work are elaborated in Section VI.

II. Related work

In this research, a multitude of packages were analyzed and evaluated among each other for the purpose of identifying a package that provides all the necessary tools to calculate HRV. To narrow down the exhaustive list of packages available, it was decided to exclude those packages whose repositories in GitHub were not active for a very long time and those packages which did not have research backing their algorithms or the authors behind them had no experience in the field. The following software packages stood out the most:

- hrvanalysis [14]
- HRV [15]
- pyHRV [16]
- hrv [17]
- pypyphysio [18]

Additionally two alternative tools were taken into consideration for analysis:

- Physiolyze (web service) [19]
- Kubios (third party standalone application) [20]

pyHRV is developed as a part of a master thesis by Pedro Miguel Caridade Gomes [16]. It is currently maintained by the original author however, it is not that active of a GitHub repository. This means that not much of the bugs are reported or solved, furthermore even new features take a long time to be released. *pyHRV* offers signal processing with peak detection, which is an extension of BioSPPY [21] a Python library focused on processing biosignals. *pyHRV* builds on top of this package by adding HRV calculations.

Hrvanalysis [14], is an open-source project developed by the company Aura-Healthcare. The package provides tools for evaluating the time domain, frequency domain and nonlinear measurements. Plots are also provided, however it is missing the Detrended Fluctuation Analysis plot. Regarding the frequency domain, it gives the user to select which method to use whether that be Welch or Lomb-Scargle. One downside of this package is that it has very limited documentation additionally all the sections for explanation have a ”coming soon” text, on the other hand, the repository has not seen activity for almost two years.

Hrv [17], is a package developed by Rhenan Bartels. Similarly to *Hrvanalysis*, the package provides tools to calculate the appropriate Indices, with exception to Detrended Fluctuation Analysis. Unlike the other packages mentioned in this paper, *hrv* has its own class of how the RR interval object should look like. Additionally, it only takes in RR intervals and runs its own algorithm to detect peaks making it impossible to take annotation files or already processed NN Intervals. Despite all this, *hrv* has a pretty active repository on GitHub with solid documentation for all the functions. Finally, it is the only package that reads Polar® (.hrm) files.

HRV [15], is a package which calculates every HRV parameter, including that of Detrended Fluctuation Analysis as opposed to the previously mentioned packages. The only issue with this package is that it has seen no updates for three years now and since then python as a language has had many updates and a new version making the package quite buggy for new systems. Similarly to *hrv*, this package also does not accept annotation files and tries to detect the QRS Complex on its own, the algorithm for the QRS Complex detection is based on Pan-Tompkins.

Pyphysio [18] is an open-source library, written in Python, for biosignal processing. In fact, it is developed by the same author behind Physiolyze, a web service for HRV and other related parameters related to biosignals. However, upon more detailed research it turns out that the web service uses only the signal processing from Pyphysio and additionally utilizes the HRV calculation methods from *pyHRV*. This ultimately means that the accuracy of the calculations is based on *pyHRV*.

Finally, *Kubios* is a commercially available software for HRV measurements and monitoring [20]. The software offers both signal processing with the ability to detect QRS Complexes and in turn provides calculations for time domain, frequency domain, and nonlinear measurements.

Physionet is the most credible source when it comes to calculating and manipulating biosignal data. In this particular experiment, it provided all the appropriate tools from signal processing to QRS Complex detection and even HRV com-

putation. However, the package is programmed in C whilst the benefit that the calculation is done at lightning speed there is a downside to it when it comes to integrating it with other software. Additionally, it proves to be a real hassle for developers who have very little experience with the C programming language.

III. Methods

In this section, we present the identified obstacles and challenges of methods to accurately calculate HRV. that motivated us to develop methods for own solution ViewHRV [22] and, in addition, we present methods to evaluate other existing software packages and compare with our solution by analyzing the performance on real existing ECG data.

A. Methods to accurately calculate HRV and access the results

Three methods are used to calculate HRV parameters [23] as follows:

- **Time-domain**: *SDNN* is the standard deviation of consecutive NN intervals, *SDANN* represents the standard deviation of the average NN intervals, *ASDNN* presents the mean of the standard deviations of all the NN intervals, *NN50* shows the number of successive NN intervals that differ by more than 50ms, *pNN50* represents the percentage of successive NN intervals that differ by more than 50ms, *rMSSD* is the root mean square of successive NN interval differences
- **Frequency domain**: *ULF* is the frequency band between 0Hz and 0.00333Hz (inclusive), *VLF* represents the frequency between 0.00333Hz and 0.04Hz (inclusive), *LF* presents the band between 0.04Hz and 0.15Hz (inclusive), *HF* shows the frequencies between 0.15Hz and 0.4Hz (inclusive), *LF/HF* represents the ration between Low Frequency and High Frequency readings, *Total Power* is the sum of all the frequencies
- **Nonlinear measurements**: *SD1* Poincaré plot standard deviation perpendicular the line of identity, *SD2* Poincaré plot standard deviation along the line of identity, *SD1/SD2* Ratio of SD1-to-SD2, *DFA* $\alpha 1$ Detrended fluctuation analysis, which describes short-term fluctuations, *DFA* $\alpha 2$ Detrended fluctuation analysis, which describes long-term fluctuations

Developing a web app alongside a web service for calculating Heart Rate Variability is one of the most obvious challenges from the start of the development is figuring out how the algorithms are going to look like. In fact, this problem was the one that took the most time and effort. It is of the utmost importance to get the correct algorithms on calculating HRV parameters as the entire purpose of the web app and service is to offer the most accurate and reliable calculations.

To solve the problem, a plethora of literature and text on the topic were reviewed. The sheer number of texts on the subject and the different methods of calculating differed substantially throughout the literature, this meant that we had to plow through the different solutions to find the most optimal and sufficiently accurate calculations.

Firstly, time domain calculations came to be basic statistical calculations which meant that they were pretty straightforward. The same came about when calculating Poincare features. However, when it came to calculating frequency domain indices, there was some inconsistency in the medical community. Whilst many of the software out there were calculating the frequency domain through Welch method, *Physionet* who is the most experienced organization dealing with HRV calculations firmly stood by calculating frequency domain through Lomb-Scargle method. Upon further literature reading it was obvious that the way to go was calculating by using Lomb-Scargle method, this is so because Lomb-Scargle calculates the frequency domain of a signal that is not evenly spaced which is exactly the signal type used to compute HRV as opposed to Welch method which deals with evenly spaced signals.

Secondly, there was very little emphasis on Detrended Fluctuation Analysis (DFA) in the majority of the literature. DFA is a measure of the short-term and long-term fluctuations, which comes quite in handy when analyzing HRV. However, there was no out of the box solution in python to calculate this parameter, which meant coming up with a custom solution. This is where the Python package Numpy [24] came into play. Numpy is a package for mathematical computations, it has proven to be a fast and reliable solution throughout all different disciplines of mathematics [25].

Thirdly, after figuring out the basic algorithms for calculating all HRV parameters there was a need for a reference point to see how correct they were and where they needed updating or maybe even changing completely. This is where *Physionet* and its tools alongside its resources came in handy. *Physionet* extensively explains these different calculations and even provides open-sourced software to use, written in C. Of course the goal of this paper is to provide a Python solution to all these calculations to easily implement and use with other applications. Furthermore, the same data sets were ran between *Physionet* C program and *ViewHRV* calculations. Once we had acceptable results we moved on to developing the web app, web service, and Application Programming Interface (API).

B. Performance evaluation

The software libraries and publicly available online solutions aiming at calculation of HRV were evaluated with the software tools offered by *Physionet* [26]. *Physionet* has gained a reputation for over a decade as the go-to place for biological signal processing and especially for gaining insights from those signals. It has been around even before *Kubios* and is completely written in C, which makes it very fast and very reliable. Because of the reputation that *Physionet* boasts, it was decided that the results from their programs be used as the gold standard for all calculations.

The methods used in this research consists of running the same data set through all these packages and software and then comparing the results. *Physionet* offers a database on ECG recordings MIT-BIH arrhythmia database [27]. Although the database contains 48 records of 30 minute ECG signal, we have chosen the record id 100 to present the results in this anal-

ysis, as a person diagnosed with a Normal Sinus Rhythm and some extrasystoles (33 premature atrial contractions (PAC) and one premature ventricular contraction (PVC). The other patients were diagnosed with more frequent arrhythmia episodes and the difference between the corresponding results is higher.

A simple experiment was conducted to compare the time-domain HRV parameters SDNN, ASDNN, SDANN, NN50, pNN50 and RMSSD. Since the record 100 contains several extrasystoles, the calculation of consecutive NN intervals needs to extract them from the evaluation. Record 100 incorporates the lowest number of arrhythmia when compared to other records, and is usually treated as a healthy patient record.

IV. ViewHRV software architecture

ViewHRV [22] is our own developed platform to accurately calculate HRV paramaters. The technology had to be: python friendly, support for web service development, support for API building, effortless scalability, and able to handle congruent requests. Python has quite some packages for web development however Django [28] and Flask [29] fit the previously mentioned requirements. When it came to deciding on which one to use, it was obvious that Django was more of a mass scale enterprise with longer development time and more complex code. This made Django excessive and unreasonable to use and thus it was decided to use Flask.

This section explains all the functional components related to the web app, web service, and application programming interface. Figure 1 presents the software architecture and the functional units. From a birds-eye view, the system is divided into three major components:

- *Access*, describing the ways a user may access the system
- *Units*, describing the functional units
- *Results*, describing the format of the results

Access to the system is granted through two ways, one being through a user interface which in this case is a website and the second way programmatically through web requests. The user can access the website and whilst there can be informed about the different possibilities of the system. The second method exists so that users who are developing a new software application can easily integrate the calculations of *ViewHRV*.

There are three main modules (Web App, Web Service, and API) and one module with a collection of all the core units which are shared between the main modules (Core Units). The Core Units module is a collection of nine units which are crucial to the system:

- *time domain*, encapsulates the functions and methods for calculating time domain indices
- *frequency domain*, contains all the functions and methods for calculating frequency domain Indices
- *nonlinear measurements*, has all the functions and methods for calculating Nonlinear indices
- *Poincare plot*, contains the appropriate code for producing the Poincare Plot
- *Lomb-Scargle periodogram*, contains the needed code for producing the Lomb-Scargle Periodogram

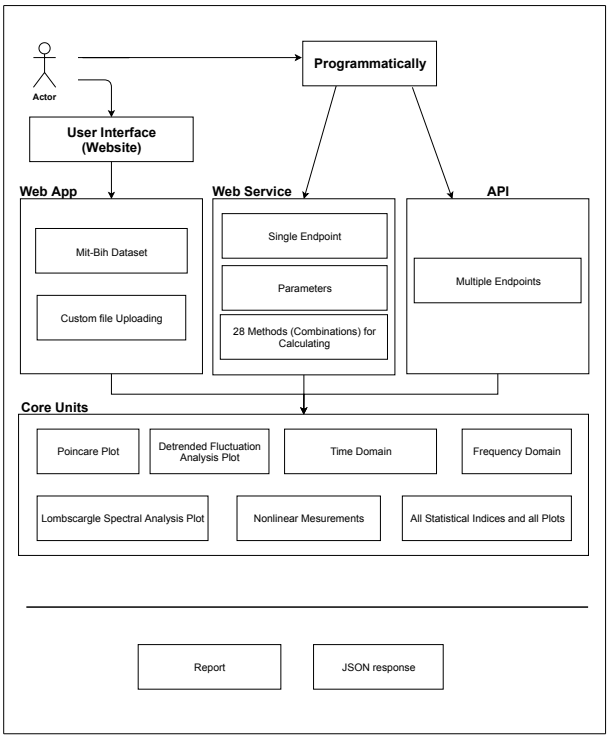


Fig. 1. Software architecture of the ViewHRV platform

- *Detrended Fluctuation Analysis plot*, contains the source code for producing the Detrended Fluctuation Analysis plot
- *All statistical indices and plots*, summarizes all the above mentioned calculations and plots

The Web App module is a collection of two units specific only to the web application: MIT-BIH dataset and custom file uploading. The MIT-BIH dataset unit handles the part of the web application dealing with the displaying and processing of the publicly available MIT-BIH Arrhythmia Database provided by *Physionet*. The custom file uploading unit provides the user with the ability to upload a custom annotation file to calculate HRV parameters. Additionally, the unit provides flexibility in the type of file to be uploaded and even prompts for the contents inside the file (for faster calculations).

Web Services in the module encapsulating three units which are accessible through a single endpoint. These units are Single Endpoint, Parameters, 28 Methods (Combinations) for Calculating. The first unit, Single Endpoint, provides access to all the calculations through a single endpoint and accepts the payload data from the user. The second unit, Parameters, takes care of the input parameters which in this case are Frequency and Method. It then processes the requests and allows the third and final unit to take action. 28 Methods (Combinations) for Calculating, contains all the different combinations between statistical computations and graphing figures. It facilitates the coupling of the different combinations of the Core Units.

The API module contains only one unit named Multiple

TABLE I
COMPARISON OF SOFTWARE ON PATIENT 100 FROM MIT-BIH ARRHYTHMIA DATABASE.

Comparison of time domain calculations							
Indice	hrvanalysis	HRV	hrv	pyhrv	Kubios	physionet	ViewHRV
SDNN	36.191	36.183	36.191	36.187	37.851	34.661	36.183
ASDNN	N/A	N/A	N/A	32.314	N/A	30.962	32.074
SDANN	N/A	N/A	N/A	16.877	N/A	16.416	15.978
NN50	127	84	127	110	228	103	110
pNN50	5.762%	3.811%	5.762%	4.993%	10.062%	4.749%	4.991%
rMSSD	28.574	28.567	28.574	28.580	50.376	26.910	28.574

Endpoint. This unit provides a single endpoint to every core unit separately. It accepts a payload from the user in a JSON format containing annotation data and returns the requested parameter whether that be a statistical calculation or a graphical visualization. It does not except a parameter for sampling frequency and by default is set to 128Hz.

The final component shows the formats of the results a user should expect. There are a total of two formats:

- *Report*
- *JSON response*

For the user accessing the calculation through the Web Application, the resulting calculation will be displayed in a Report inside the Web Application. On the other hand, users accessing the calculations programmatically the results are returned in a JSON format. It was deliberately chosen to return the results in a JSON format since it is much easier to develop on top of these formats given the constant format. The plots in the JSON response are Base64 Encoded Strings, this was chosen as it is compatible with every other programming language thus enabling the user complete flexibility when choosing their development environment.

V. DISCUSSION

When it came to computing HRV indices in the Python world, there are a plethora of packages available. However, many of the packages are either in development or are extremely buggy. But still, there were a few that tackled the challenge appropriately such as *pyHRV*. On the other hand, when compared to *Physionet's* calculation there were a number of discrepancies.

This is the main motivation behind the development of *ViewHRV*, which aims to be a package that will accurately compute HRV indices and do so reliably and consistently. Our current project Gluco [30] aims to detect the correlation between HRV parameters and blood glucose level, we need a reliable source of calculating the HRV out of an ECG signal obtained by a wearable one-channel ECG sensor.

Table I presents the evaluated HRV parameter results for the record 100 of MIT-BIH arrhythmia database. In fact, *Kubios* had the most incorrect readings when compared to Python packages. As mentioned before, *Physionet*'s C programs are used as the benchmark. Throughout the different Python packages there is very little difference, matter of fact *ViewHRV* (internally developed package) and *HRV* (open source package) came the closest to *Physionets* calculations.

Our investigation of existing software has shown that the existing available software sources are not reliable or available online. Our software *ViewHRV* has proven to be a more reliable and accurate package compared to the already open source packages available online. It comes closest to the *Physionet* results throughout all Domain calculations. Of course, this does not mean that the other packages were that far off, for example when it came to time domain calculation *pyHRV* and *ViewHRV* were neck in neck and even showed the same result for NN50. However, for the other time domain indices *ViewHRV* came closer to *Physionets* calculation compared to *pyHRV*. Concerning frequency domain, *ViewHRV* continued to prove reliable and accurate results. On the other hand, *pyHRV* showed the furthest results even when compared with the other packages despite those packages not having proper results for time domain indices. Involving nonlinear Measurements all Python packages showed very similarly to almost identical results.

One important observation here is that the commercially available software under the name *Kubios*, provided the most inaccurate calculations compared to any other tool. In fact, it did not even come close to the *Physionet* calculations. When it came to frequency domain calculation, surprisingly it uses Welch's method instead of Lomb-Scargle even though the latter is the preferred method since HRV data is not evenly spaced data.

VI. CONCLUSION

We have evaluated available software libraries and existing software solutions for their ability to accurately calculate HRV parameters. As the findings showed different results for calculated HRV parameters, in this paper we describe how we have created our own platform ViewHRV by elaborating the design of the software architecture and functional descriptions.

To sum it up, our evaluation of existing online software showed that our software product *ViewHRV* is on the right track and that it provides the most consistent results throughout all different domain calculations. *ViewHRV* [22] comes the closest to *Physionets* results and even provides better results than *Kubios* with its hefty pricing. *ViewHRV* provided reliable calculations and if one were to start a new project or were to take on the processes of monitoring and calculating HRV, then this package would be the right choice.

We plan to use this software in our research and development project Gluco and detect correlation and appropriate ML

algorithm to predict the ability of the human body to regulate the glucose level based on a wearable ECG sensor.

REFERENCES

[1] R. McCraty and F. Shaffer, "Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk," *Global advances in health and medicine*, vol. 4, no. 1, pp. 46–61, 2015.

[2] H. Tsuji, M. G. Larson, F. J. Venditti, E. S. Manders, J. C. Evans, C. L. Feldman, and D. Levy, "Impact of reduced heart rate variability on risk for cardiac events: the framingham heart study," *Circulation*, vol. 94, no. 11, pp. 2850–2855, 1996.

[3] G. E. Billman, "Heart rate variability–a historical perspective," *Frontiers in physiology*, vol. 2, p. 86, 2011.

[4] J. F. Thayer, S. S. Yamamoto, and J. F. Brosschot, "The relationship of autonomic imbalance, heart rate variability and cardiovascular disease risk factors," *International journal of cardiology*, vol. 141, no. 2, pp. 122–131, 2010.

[5] C. J. Murray, A. D. Lopez, R. Black, C. D. Mathers, K. Shibuya, M. Ezzati, J. A. Salomon, C. M. Michaud, N. Walker, and T. Vos, "Global burden of disease 2005: call for collaborators," *The Lancet*, vol. 370, no. 9582, pp. 109–110, 2007.

[6] S. Yusuf, S. Reddy, S. Ôunpuu, and S. Anand, "Global burden of cardiovascular diseases: part i: general considerations, the epidemiologic transition, risk factors, and impact of urbanization," *Circulation*, vol. 104, no. 22, pp. 2746–2753, 2001.

[7] L. Khadra, A. Al-Fahoum, and H. Al-Nashash, "Detection of life-threatening cardiac arrhythmias using the wavelet transformation," *Medical and Biological Engineering and Computing*, vol. 35, no. 6, pp. 626–632, 1997.

[8] E. Corday, V. Bazika, T.-W. Lang, S. Pappelbaum, H. Gold, and H. Bernstein, "Detection of phantom arrhythmias and evanescent electrocardiographic abnormalities: use of prolonged direct electrocardiocording," *Jama*, vol. 193, no. 6, pp. 417–421, 1965.

[9] J. M. Pevnick, K. Birkeland, R. Zimmer, Y. Elad, and I. Kedan, "Wearable technology for cardiology: an update and framework for the future," *Trends in cardiovascular medicine*, vol. 28, no. 2, pp. 144–150, 2018.

[10] K. Georgiou, A. V. Larentzakis, N. N. Khamis, G. I. Alsuhaibani, Y. A. Alaska, and E. J. Giallafos, "Can wearable devices accurately measure heart rate variability? a systematic review," *Folia medica*, vol. 60, no. 1, pp. 7–20, 2018.

[11] R. Trobec, I. Tomašić, A. Rashkovska, M. Depolli, and V. Avbelj, "Commercial ECG systems," in *Body Sensors and Electrocardiography*. Springer, 2018, pp. 101–114.

[12] A. Rashkovska, M. Depolli, I. Tomašić, V. Avbelj, and R. Trobec, "Medical-grade ECG sensor for long-term monitoring," *Sensors*, vol. 20, no. 6, p. 1695, 2020.

[13] A. R. Pérez-Riera, R. Barbosa-Barros, R. Daminello-Raimundo, and L. C. de Abreu, "Main artifacts in electrocardiography," *Annals of Noninvasive Electrocardiology*, vol. 23, no. 2, p. e12494, 2018.

[14] R. Champseix, "hrvalysis," 2019. [Online]. Available: https://github.com/Aura-healthcare/hrvanalysis

[15] S. Pickus, "Hrv," 2017. [Online]. Available: https://github.com/pickus91/HRV

[16] P. Gomes, P. Margaritoff, and H. Silva, "pyHRV: Development and evaluation of an open-source python toolbox for heart rate variability (HRV)," in *Proc. Int'l Conf. on Electrical, Electronic and Computing Engineering (IcETRAN)*, 2019, pp. 822–828.

[17] R. Bartels, "hrv," 2020. [Online]. Available: https://github.com/rhenanbartels/hrv

[18] A. Bizzego, A. Battisti, G. Gabrieli, G. Esposito, and C. Furlanello, "pyphysio: A physiological signal processing library for data science approaches in physiology," *SoftwareX*, vol. 10, p. 100287, 2019.

[19] A. Bizzego, M. Mina, C. Zarbo, G. Esposito, and C. Furlanello, "Physiolyze: a galaxy-based web service for heart rate variability analysis with online processing," in *2014 8th Conference of the European Study Group on Cardiovascular Oscillations (ESGCO)*. IEEE, 2014, pp. 97–98.

[20] M. P. Tarvainen, J.-P. Niskanen, J. A. Lipponen, P. O. Ranta-Aho, and P. A. Karjalainen, "Kubios hrv–heart rate variability analysis software," *Computer methods and programs in biomedicine*, vol. 113, no. 1, pp. 210–220, 2014.

[21] C. Carreiras, A. P. Alves, A. Lourenço, F. Canento, H. Silva, A. Fred *et al.*, "BioSPPy: Biosignal processing in Python," 2015–, [Online; accessed ¡today¿]. [Online]. Available: https://github.com/PIA-Group/BioSPPy/

[22] Innovation Dooel, "ViewHRV," 2020. [Online]. Available: https://viewhrv.com/

[23] M. Malik, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: Task force of the european society of cardiology and the north american society for pacing and electrophysiology," *Annals of Noninvasive Electrocardiology*, vol. 1, no. 2, pp. 151–181, 1996.

[24] T. E. Oliphant, *A guide to NumPy*. Trelgol Publishing USA, 2006, vol. 1.

[25] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The numpy array: a structure for efficient numerical computation," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, 2011.

[26] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[27] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[28] J. Forcier, P. Bissex, and W. J. Chun, *Python web development with Django*. Addison-Wesley Professional, 2008.

[29] F. A. Aslam, H. N. Mohammed, J. M. Mohd, M. A. Gulamgaus, and P. Lok, "Efficient way of web development using python and flask," *International Journal of Advanced Research in Computer Science*, vol. 6, no. 2, 2015.

[30] Innovation Dooel, "A doctor in your pocket measure ECG & glucose levels with a small, non-invasive, wearable monitor," 2018. [Online]. Available: http://glyco.innovation.com.mk/

Impact of Subthreshold Transcutaneous Auricular Vagus Nerve Stimulation on the Heart Rate Variability and Atrial Arrhythmias

R. Trobec*, M. Lindén** and M. Šinkovec***
* Department of Communication Systems, Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
** Mälardalen Univeristy/Embedded Sensor Systems for Health, Västerås, Sweden
*** Department of Cardiology, University Medical Centre, Zaloška 7, Ljubljana, Slovenia
roman.trobec@ijs.si; maria.linden@mdh.se; matjaz.sinkovec@kclj.si

Abstract – Dysfunction of cardiac intrinsic autonomic nervous system could be one of possible triggers for atrial fibrillation. Different studies show that it is possible to modulate the cardiac intrinsic autonomic activity with electrical stimulation through the skin in the ear, innervated by auricular branch of the vagus nerve (tVNS). The aim of our study was to determine, weather tVNS reduces the burden of atrial arrhythmias in patients with paroxysmal atrial fibrillation. A randomised double-blind study, with one-hour daily subthreshold tVNS, over three-day blocks of sham or true stimulation, was performed. A continuous single-channel ambulatory ECG was recorded. Significant technical assistance to computer-based ECG interpretation was necessary for highest possible interpretation quality. The obtained results were compared between periods with true and sham stimulation to discover eventual differences. We have successfully validated the designed methodology. The subthreshold tVNS has a detectable impact on heart rate variability. However, there was no significant impact of tVNS on the burden of atrial arrhythmias. A 3-day tVNS is probably too short for such evaluation.

Keywords – electrical stimulation, tVNS, ECG body sensor; auricular vagal nerve; heart rate variability, HRV; AF burden; atrial arrhythmias;

I. INTRODUCTION

Unbalanced sympathetic/parasympathetic activity has been recognized in a number of clinical conditions, such as heart failure [1,2], sudden cardiac death [3] and paroxysmal atrial fibrillation (AF) [4, 5]. In patients with heart failure, the invasive cervical vagus nerve stimulation (VNS) failed, probably due to abundant sympathetic fiber and low cardiac vagus nerve content at that location [6]. Animal and human studies have demonstrated that the cutaneous auricular branch of the vagus nerve (ABVN), which innervates a region of external ear, is connected to the central vagal projections [7]. It has been reported that transcutaneous stimulation of the ABVN (tVNS) can exert more selective efferent vagal activity and was recognized as a promising therapeutic approach [5,7,8].

One of the possible triggers of AF [9] is dysfunction or imbalance of intrinsic cardiac autonomic nervous system [5]. Several recent studies indicate, that it is possible to reduce the burden of AF with tVNS. However, different stimulation protocols and sites have been used, which complicate a direct comparisons and conclusions about a real efficiency of this non-invasive method.

Several investigations have been published analyzing different stimulation parameters. The effects of low (10 Hz) and high (100 Hz) frequency transcutaneous electrical nerve stimulation (TENS) on heart rate variability (HRV) were analyzed in healthy subjects, with stimulation applied in the paravertebral region for 30 minutes [10]. The stimulation current was adjusted to a subjectively comfortable level. After 10-Hz stimulation, there was decrease of LF components of the HRV and increase of HF (parasympathetic modulation), while the 100-Hz stimulation had an opposite effect (sympathetic modulation).

The tVNS effects on cardiac vagal activity and subjective perception of different stimulation intensities (0.5, 1.0 and 1.5 mA) in 61 healthy participants was tested in [11]. Stimulation electrodes were placed in the left cymba conchae and sham electrodes on earlobe. The marker of vagally-mediated HRV was the root mean square of consecutive RR interval differences (*RMSSD*). An increase in *RMSSD*, i.e. in cardiac vagal modulation, was found from resting to stimulation phases. However, the increase of HRV was not dependent on stimulation intensity, on the contrary to subjective perception.

Recently, an extensive overview of the neuroanatomy of the ABVN has been published with

respect to different aspects of tVNS [12]. Authors concluded that at present it is reasonable to surmise that concha and inner tragus are suitable locations for tVNS and vagal modulation. However, earlobe is not appropriate for sham stimulation.

Effects of short (10 minutes) and prolonged (one hour) tVNS on the HRV were investigated in 30 heathy participants [13]. Stimulation intensity was in the comfort level with stimulation frequency of 25 Hz. The stimulation protocol was 30-second stimulation phase followed by 30-second pause. Authors found that only the right-ear stimulation significantly increases standard deviation of the RR intervals (*SDNN)* compared to baseline. HRV changes could be more consistent in women than in men. Non-uniform results could be explained by the study limitations and biases, e.g. no double-blind design, relatively small sample size, short stimulation periods, short abstain period (3 h) for caffeine, alcohol and smoking.

The aim of our study was to determine, whether tVNS affects HRV and the burden of atrial arrhythmias in patients with paroxysmal AF. The outcome of such investigations can be severely biased by placebo effect, unfairness of investigators, biased reports of patients, etc. Therefore, we used a randomised double-blind design, where neither patients nor the assisting personal were aware of the time of true and false (sham) stimulation, in addition to imperceptible subthreshold level of stimulation. Therefore, our patients could not distinguish between periods of true and sham stimulation. An additional reason for testing sub-threshold tVNS was the applicability for everyday ambulatory use.

In order to verify the ability and efficiency of tVNS, the following two hypotheses have been tested: *(H1)* - imperceptible electrical stimulation of auricular vagal branch impacts HRV and atrial arrhythmias and *(H2)* – the applied methodology is appropriate for a larger study of tVNS impact on the burden of AF.

II. Equipment and Methods

The study was designed as a prospective randomized double-blind clinical study to compare HRV and rhythm parameters during true and sham tVNS. We excluded patients with significant associated cardiac or other diseases and patients after cardiac surgery. The study was approved by National Ethics Committee of the Republic of Slovenia (KME 77/01/16). An informed consent was obtained from all participants prior to investigations.

Thirteen volunteers have been recruited in the study of both genders with average age of 59 years (39 to 76 years), who were on a waiting list for catheter ablation of paroxysmal AF at the Cardiology Department, University Medical Centre Ljubljana (see Table 1). Four of them, marked in gray were excluded from the study because of persistent irregular atrial rhythm (IAR) during the measurements. We analyzed ECG recordings obtained in the period from January to June 2017 [14].

TABLE 1. DEMOGRAPHIC DATA OF THIRTEEN PARTICIPANTS U1 TO U13. FOUR OF THEM: U3, U5, U7, U9 ARE EXCLUDED BECAUSE OF IAR.

Patient ID	*gender*	*age*	*medicaments*	*AF prevalence*	*Reason for exclusion*
U1	M	63	ar	d	
U2	M	54	ar	m	
U3	M	76	ac	d	IAR
U4	W	74	ar, ac	m	
U5	W	47	ar	m	IAR
U6	M	61		w	
U7	M	39		m	IAR
U8	W	65	ar, ac	d	
U9	M	53		d	IAR
U10	M	66	ac	y	
U11	M	65	ar, ac	y	
U12	W	54	ar	m	
U13	M	54	ar, ac	w	

Legend: ar- antiarrhythmic drugs, ac-anticoagulants, d-daily, w-weekly, m-monthly, y-yearly.

We recorded 72 hours of baseline ECG, followed by two 3-day block (72 hours of ECG) of true or sham one-hour daily tVNS, for each patient. Based on the preliminary tests on fifteen volunteers, average maximal imperceptible electrical current was 0.13 mA. The stimulation was performed with 1 ms square pulse of the constant electrical current (0.13 mA) with frequency of 20 Hz, intermittently 1-second stimulation and 1-s pause. Three consecutive days belong either to the stimulation or sham stimulation.

The one-hour tVNS was performed in a doctor's office in the afternoon between 15 and 20 pm, during rest in lying position. We used adapted devices for cutaneous electrical nerve stimulation (TENS, Medikoel d.o.o., Slovenia [15]) with hidden inner switch, which turns the stimulation on or off. For tVNS, we placed a clip with one electrode in cavum conchae of both ears (see Figure 1). The ECG recordings were performed by one-channel ambulatory ECG recorder (Savvy, Saving d.o.o., Slovenia [16]), implemented as a bipolar wireless body sensor [17]. Its electrodes were fixed on the chest above the atrium, at positions of standard V1 and V2 leads. The recording and stimulation starts after initial rest of approximately 15 minutes after sensor fixation procedures. The Savvy sensor wirelessly transmits the recorded ECG stream to a smart phone where a mobile application MobECG is storing it in a file with a unique name (Sensor MAC address, and date and time of recording start).

Intensity of the current was under the threshold of perception. The most important reason for such a choice was our aim to keep the investigation blind. The

order of stimulation or sham stimulation phases was determined randomly by leading investigator, using the hidden stimulation switch. In this way the stimulation order was not known to any other involved medical personal, patients and investigators, until the final discussion of obtained results.

Before the analysis of ECG recordings a visual examination of all data was made to confirm the minimal required values of amplitudes and signal-to-noise ratios. The ECG rhythm analysis (CER-S software, AMPS llc, Italy [18]) is able to identify noisy segments or disturbances, to count atrial and ventricular premature beats and to detect more complex atrial and ventricular rhythm disturbances, e.g. couplets, multigeminy, tachycardia, flutter and fibrillation.

The AMPS software finally calculates the following well known HRV parameters:

nRR, RR Mean, PNN50, SDNN, RMSSD,

for each 5-minute time slot from a continuous ECG recording of up to 3 days. *nRR* denotes the number of detected RR intervals (beats), *RR Mean* is the mean of all detected RR intervals, *PNN50* is number of successive difference of RR intervals which differ by more than 50 ms expressed as a percentage of the total number of ECG cycles analyzed, while *SDNN,* and *RMSSD* have been defined in the previous section. The HRV analysis was performed for the first hour of ECG measurements, while patients are resting in laying position. The HRV parameters are also calculated for each block of 72 hours with either true or sham stimulation.

The rhythm parameters are calculated for hours, days and nights. We analyzed again the first hour during true or sham stimulation of each 3-day block, and the whole 3-day phase. We do not show all rhythm parameters, because mostly no differences have been detected between true and sham stimulation. The presented rhythm parameters are:

HR Min, Brady, AFib %, nV Beats, nS Beats,

where *HR Min* is the minimum heart rate, *Brady* denotes the number of bradycardia sequences, *nV Beats* and *nS Beats* denote the number of ventricular or supraventricular beats, and *AFib %* denotes the percentage of analyzed time with atrial irregular rhythm.

The HRV and rhythm parameters are either summed by hours or days, e.g. *nRR*, or calculated as an average, weighted by the number of detected beats per analyzed period. For example, if we denote the number of analyzed beats *nRR* and *PNN50* of day *k* by *nRR_k* and *PNN50_k*, respectively, then the by-days weighted average of *PNN50* for Phase 1 (day 4, 5 and 6) is:

$$PNN50 = (PNN50_4 * nRR_4 + PNN50_5 * nRR_5 + PNN50_6 * nRR_6) / (nRR_4 + nRR_5 + nRR_6).$$

The statistical analysis of all obtained parameters was performed with one-tailed *t*-test for paired samples. Statistical significance was determined at $p < 0.05$.



Figure 1. A photo of TENS stimulator and fixation of electrodes (The figure is adapted from [14]).

III. RESULTS

In Table 2 summarized HRV parameters that are shown for all analyzed patients and for both phases of measurements are presented. Two sub-tables above are for by-day analysis and below for by-hour (the first hour of each day) analysis, either during true stimulation (two left sub-tables) or sham stimulation (two right sub-tables). One-tail *p* value is shown in the last line of each sub-table. In the left sub-tables *p* was calculated for all considered participants and in the right sub-tables *p* was calculated only on the subset of all participants without excluded "outlying" participants. The exclusion reasons are listed below.

In by-day analysis (right sub-table above) participants marked with gray color: U7, U10, U12, U13 are excluded because:

- U7- Significant amount of AF in sham stimulation results in high HRV parameters.
- U10 - More than 20% of AF in sham stimulation. HRV is not applicable.
- U12 - Significant amount of AF in real stimulation results in high HRV (HRV is also high in sham stimulation?).

- U13 - More than 30% of AF in real and sham stimulation. HRV is not applicable.

In by-hour analysis (right sub-table below) participants marked with gray color: U7, U12 are excluded because:

- U7 - Significant amount of AF in the first hour of sham stimulation results in false high HRV.
- U12 – Only one hour of recording available in the sham stimulation phase because of extremely low signal-to-noise ratio (weak electrode-skin contact?).

Looking on data from Table 2 (*p*-all) we see that there is no statistical difference between HRV by-day and by-hour parameters if all participants are considered. However, there is a tendency in increased HRV during true stimulation. After exclusion of outlying participants, because of previously described reasons, we detect a statistical difference in *PNN50*, which is higher for by-day and by-hour true stimulations. This result could indicate increased vagal activity and is in accordance with previous findings [1, 11].

In Table 3 a subset of summarized rhythm parameters is shown in the same way as for HRV. Left sub-tables are for true stimulation and right for sham stimulation. Upper two tables are for by-day analysis and lower two tables for by-hour analysis. Again, for the same reasons, the same participants have been excluded as in the HRV analysis. The results indicate a significant statistical difference between by-day true and sham stimulation in *HR Min* only (two sub-tables above), which is lower during true stimulation. There are no statistical differences in the by-hour rhythm parameters (two sub-tables below).

We noticed a trend of decreased duration of AF in by-day stimulation but the difference did not reach statistical significance. Also the number of premature atrial or ventricular beats did not change after the true stimulation.

TABLE 2. HRV PARAMETHERS OF DAYS AND HOURS ANALYSIS. LEFT – TRUE STIMULATION, RIGHT – SHAM STIMULATION. P-ONE TAIL VALUE IS SHOWN IN THE LAST LINE OF EACH SUB-TABLE. LEFT – FOR ALL PATICIPANTS, RIGHT - FOR PRTICIPANTS WITH EXCLUDED: U7, U10, U12, U13 (ABOVE) AND FOR PRTICIPANTS WITH EXCLUDED: U7, U12 (BELOW). EXCLUDED PARTICIPANTS ARE MARKED WITH GRAY COLOR.

True Stimulation					
Days	***nRR***	***RR Mean***	***PN N50***	***SDNN***	***RMS SD***
U1 day 4-5	190564	1218	4,1	161,8	31,0
U2 day 7-9	305448	793	4,0	159,5	24,7
U4 day 7-9	172651	1042	22,8	130,5	71,0
U6 day 4-5	281833	799	4,0	144,1	26,7
U7 day 4-5	224930	1105	16,1	184,7	44,8
U8 day 4-5	220229	1018	12,2	157,0	39,7
U10 day 4-5	125753	1075	8,7	131,1	34,6
U11 day 4-5	287946	741	3,4	80,8	21,8
U12 day 4-5	134921	1030	21,9	186,4	84,0
U13 day 4-5	287754	754	17,9	112,0	96,0
p-all	*0,49*	*0,26*	*0,28*	*0,40*	*0,27*

Sham Stimulation					
Days	***nRR***	***RR Mean***	***PN N50***	***SDNN***	***RMS SD***
U1 day 7-9	199369	1278	2,2	147,7	25,5
U2 day 4-5	330226	756	1,9	146,8	20,3
U4 day 4-5	222515	910	19,1	153,7	67,9
U6 day 7-9	278194	789	3,5	132,7	27,2
U7 day 7-9	211487	1124	21,0	205,9	53,2
U8 day 7-9	235584	898	5,2	133,2	30,5
U10 day 7-9	112326	1034	13,5	135,1	95,5
U11 day 7-9	263178	783	2,2	86,4	23,9
U12 day 7-9	102632	1118	16,9	178,4	71,7
U13 day 7-9	273500	734	21,6	142,6	99,9
p-U1, U2, U4, U6, U8, U11	*0,15*	*0,18*	***0,02***	*0,23*	*0,06*

Hours	***nRR***	***RR Mean***	***PN N50***	***SDNN***	***RMS SD***
U1 hour 4-5	7969	1314	1,4	23,8	20,8
U2 hour 7-9	12386	853	2,3	36,1	18,9
U4 hour 7-9	10666	960	17,3	60,9	54,5
U6 hour 4-5	12075	884	5,5	49,4	27,6
U7 hour 4-5	9361	1150	15,7	70,8	40,3
U8 hour 4-5	9398	1139	6,6	33,1	34,3
U10 day 4-5	8323	1111	4,2	27,1	25,3
U11 day 4-5	14609	729	1,2	17,2	13,6
U12 day 4-5	8806	1080	2,9	29,7	32,8
U13 day 4-5	13231	806	0,1	19,2	8,4
p-all	*0,18*	*0,42*	*0,14*	*0,34*	*0,25*

Hours	***nRR***	***RR Mean***	***PN N50***	***SDNN***	***RMS SD***
U1 hour 7-9	8252	1307	1,1	18,8	19,1
U2 hour 4-5	12621	822	0,3	30,5	11,7
U4 hour 4-5	10309	1042	14,3	47,6	52,3
U6 hour 7-9	11921	904	1,1	40,0	19,0
U7 hour 7-9	8840	1203	18,6	75,4	45,3
U8 hour 7-9	10631	1000	1,7	33,9	21,6
U10 day 7-9	8961	1122	2,3	32,0	25,1
U11 day 7-9	13676	786	0,1	13,8	8,5
U12 day 7-9	3823	1079	7,1	42,6	46,2
U13 day 7-9	12571	801	0,2	22,1	10,7
p-U1, U2, U4, U6, U8, U10, U11, U13	*0,45*	*0,48*	***0,01***	*0,08*	***0,02***

TABLE 3. RHYTHM PARAMETHERS OF DAYS AND HOURS ANALYSIS. LEFT – TRUE STIMULATION, RIGHT – SHAM STIMULATION. P-ONE TAIL VALUE IS SHOWN IN THE LAST LINE OF EACH SUB-TABLE. LEFT – FOR ALL PATICIPANTS, RIGHT - FOR PRTICIPANTS WITH EXCLUDED: U7, U10, U12, U13 (ABOVE) AND FOR PRTICIPANTS WITH EXCLUDED: U7, U12 (BELOW). EXCLUDED PARTICIPANTS ARE MARKED WITH GRAY COLOR.

True Stimulation					
Days	***HR Min***	***Brady***	***Afib %***	***V Beats***	***S Beats***
U1 day 4-5	36	457	0,00	47	377
U2 day 7-9	47	0	0,00	991	197
U4 day 7-9	44	1	0,00	49	683
U6 day 4-5	51	0	0,00	16	1827
U7 day 4-5	34	1005	0,01	7	432
U8 day 4-5	41	34	0,00	64	713
U10 day 4-5	38	121	0,00	21	364
U11 day 4-5	60	0	0,00	275	52
U12 day 4-5	34	237	0,00	143	614
U13 day 4-5	48	0	33,00	800	159
p-one tail	***0,00***	*0,16*	*0,21*	*0,30*	*0,35*

Sham Stimulation					
Days	***HR Min***	***Brady***	***Afib %***	***nV Beats***	***nS Beats***
U1 day 7-9	38	633	0,09	59	269
U2 day 4-5	49	0	0,00	510	134
U4 day 4-5	48	0	10,00	2457	399
U6 day 7-9	53	0	0,00	8	867
U7 day 7-9	35	1109	0,23	31	663
U8 day 7-9	50	0	0,00	26	1095
U10 day 7-9	42	32	27,53	38	929
U11 day 7-9	61	0	0,00	217	67
U12 day 7-9	33	337	0,00	105	1315
U13 day 7-9	53	0	22,00	375	275
p-U1, U2, U4, U6, U8, U11	***0,02***	*0,24*	*0,18*	*0,25*	*0,20*

Hours	***HR Min***	***Brady***	***Afib %***	***V Beats***	***S Beats***
U1 hour 4-5	43	84	0	1	64
U2 hour 7-9	59	0	0	107	0
U4 hour 7-9	49	0	0	25	32
U6 hour 4-5	57	0	0	2	9
U7 hour 4-5	42	6	0	0	16
U8 hour 4-5	43	0	0	5	38
U10 day 4-5	45	0	0	4	74
U11 day 4-5	68	0	0	4	4
U12 day 4-5	46	0	0	1	51
U13 day 4-5	64	0	0	2	4
p-one tail	*0,33*	*0,31*	*0,17*	*0,13*	*0,20*

Hours	***HR Min***	***Brady***	***Afib %***	***nV Beats***	***nS Beats***
U1 hour 7-9	39	32	0	3	34
U2 hour 4-5	62	0	0	16	1
U4 hour 4-5	50	0	0	0	4
U6 hour 7-9	58	0	0	0	37
U7 hour 7-9	41	22	0,40	2	28
U8 hour 7-9	52	0	0	0	21
U10 day 7-9	43	6	0	3	81
U11 day 7-9	68	0	0	12	5
U12 day 7-9	45	0	0	2	22
U13 day 7-9	63	0	0	1	5
p-U1, U2, U4, U6, U8, U10, U11, U13	*0,27*	*0,21*	*N.A.*	*0,12*	*0,26*

IV. DISCUSSION AND CONCLUSION

Based on our results the main contributions of our study are listed below:

- bilateral imperceptible tVNS, i.e. 0.13 mA, demonstrated some heart rate responses, possibly through efferent vagal activity, which open up opportunities for clinical testing,
- position of the electrodes in cavum conchae and stimulation frequency (20 Hz) are appropriate for tVNS,
- for reliable stimulation, a constant current stimulation device should be used to compensate variations in skin-electrode contact resistance,
- episodes of AF or frequent premature beats should be excluded from HRV analysis,
- the AF burden analysis showed a favorable trend, however, requires larger group of patients and lengthy measurements, one week or even more, to accumulate sufficient AF episodes.

We found out that the tVNS increases HRV, in particular *PNN50*. Further, through the whole post-stimulation day, higher *RMSSD* and lower *HR Min* were demonstrated. This indicates a protracted cardiac vagal activity even with the use of subthreshold imperceptible stimulation. This finding is not well known and should be further investigated.

From our ECG recordings, it is not possible to speculate about effect of tVNS on the burden of atrial fibrillation or other atrial arrhythmias. Consequently, our hypotheses *H1* is not confirmed.

H2 was approved with the tested and validated methodology, however, several improvements have been identified for further research. We found out that also the imperceptible stimulation has an expected and protracted effect, which is important in designing methodologies for clinical testing. We propose: *(i)* larger number of participants with more precise inclusion/exclusion criteria, *(ii)* longer measurement periods (a week or more) should be performed in order to cover sufficient AF episodes for reliable AF burden evaluation, and *(iii)* even higher personal engagement in ECG interpretation is necessary for more accurate computer classification of heart rhythm and its variability.

ACKNOWLEDGMENT

The author R.Trobec acknowledges the financial support from the Slovenian Research Agency under the grant P2-0095. The author M. Lindén acknowledges the financial support from the Swedish Knowledge Foundation for the grant 20180158 (ESS-H+). The author M. Šinkovec acknowledges the financial support from the grant 20170039, University Medical Centre Ljubljana. We are thankful to Tadeja Sotlar and Lucija Steklačič for their help in collecting the data.

REFERENCES

[1] Sabbah HN. Electrical Vagus Nerve Stimulation for the Treatment of Chronic Heart Failure. Cleve Clin J Med. 2011; 78:doi:10.3949/ccjm.78.s1.04.

[2] De Ferrari GM at al. Chronic vagus nerve stimulation: a new and promising therapeutic approach for chronic heart failureEuropean Heart Journal 2011; 32:847–55.

[3] Vasgui M, Shivkumar K. The role of the autonomic nervous system in sudden cardiac death. Prog Cardiovasc Dis. 2008; 50: 404-19.

[4] Stavrakis S, at al. Low-level transcutaneous electrical vagus nerve stimulation suppresses atrial fibrillation. J Am Coll Cardiol. 2015; 65(9): 867–75.

[5] Murray AR, at al. The strange case of the ear and the heart: The auricular vagus nerve and its influence on cardiac control. Autonomic Neuroscience: Basic and Clinical 2016; 199:48–53.

[6] Zannad F, et al. Chronic vagal stimulation for the treatment of low ejection fraction heart failure: results of the Neural Cardiac TherApy foR Heart Failure (NECTAR-HF) randomized controlled trial. Eur Heart J. 2015; 36: 425-33.

[7] Frangos E, Ellrich J, and Komisaruk BR. Non-invasive Access to the Vagus Nerve Central Projections via Electrical Stimulation of the External Ear: fMRI Evidence in Humans. Brain Stimulation 2015; 8:624e636.

[8] Clancy JA, at al. Non-invasive Vagus Nerve Stimulation in Healthy Humans Reduces Sympathetic Nerve Activity. Brain Stimulation 2014; 7(6):871-77.

[9] Chugh SS, at al. Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study. Circulation. 2014; 129(8):837-47.

[10] Stein C, at al. Transcutaneous electrical nerve stimulation at different frequencies on heart rate variability in healthy subjects. Autonomic Neuroscience: Basic and Clinical 2011; 165:205–08.

[11] Borges U, Laborde S, and Raab M. Influence of transcutaneous vagus nerve stimulation on cardiac vagal activity: Not different from sham stimulation and no effect of stimulation intensity. PLoS ONE 2019; 14(10): e0223848.

[12] Butt MF, Albusoda A, Farmer AD, Aziz Q. The anatomical basis for transcutaneous auricular vagus nerve stimulation. J Anat. 2019; doi: 10.1111/joa.13122. (e-pub ahead of print).

[13] De Couck M, at al. Effects of short and prolonged transcutaneous vagus nerve stimulation on heart rate variability in healthy subjects. Autonomic Neuroscience: Basic and Clinical 2017; 203:88–96

[14] Sotler T, and Steklačič L. Vpliv električnega draženja kožne veje vagusa na breme atrijskih aritmij pri bolnikih s paroksizmalno atrijsko fibrilacijo. Delo za Prešernovo nalogo, mentor: Šinkovec M. Medicinska fakulteta Ljubljana, 2017.

[15] https://www.medikoel.com/elektrostimulacija/aparati/elektronski-stimulator-tens-z-ogrevalno-funkcijo-se35b, last visited on February 2020.

[16] http://www.savvy.si/en/, last visited on February 2020.

[17] Tomašić I, and Trobec R. Electrocardiographic systems with reduced numbers of leads - synthesis of the 12-lead ECG. IEEE Rev Biomed Eng. 2014; 7:126-42.

[18] https://www.amps-llc.com/, last visited on February 2020.

Clinical Decision Support Systems in Practice: Current Status and Challenges

A. Jović*, I. Stančin*, K. Friganović*, and M. Cifrek*
* University of Zagreb Faculty of Electrical Engineering and Computing, Unska 3, 10 000 Zagreb, Croatia
Corresponding author: alan.jovic@fer.hr

Abstract - **Decision support systems (DSS) are computer programs based on artificial intelligence methods that contribute to reaching a correct decision in an often-narrow domain of interest. Clinical decision support systems (CDSS) are such DSSs that may be used by medical professionals in clinics and hospitals. They are used for diagnosis, treatment protocol recommendations, treatment outcome predictions and other tasks. CDSS are constructed based on symbolic and machine learning (including deep learning) approaches to represent and infer medical knowledge. The aim of this work is to provide an overview of past and current methods in designing a successful CDSS. The study considers the systems that were claimed to be implemented in clinical practice. Currently, the development of a CDSS is mostly pursued in two directions: 1) a more traditional approach based on rules, ontologies, probabilistic models, and the use of standards; 2) machine learning based approach. Both approaches may be used complementary within a healthcare information system. This work seeks to provide an objective view on the advantages and limitations of the approaches as well to discuss future research avenues that could lead to more accurate and trustworthy CDSS and improved healthcare.**

Keywords - decision support system, clinical decision support system, healthcare, artificial intelligence, deep learning

I. INTRODUCTION

Artificial intelligence (AI) is used nowadays in a variety of application domains, with varying levels of success. Its application in healthcare is of paramount importance for transforming and improving the traditional medical workflows. The improvement needs to be made in a sense that it would benefit all the stakeholders (i.e. patients, medical personnel, clinic) in the process. The application of AI in medicine may be performed in different ways. According to [1], virtual and physical branches of AI may be distinguished. The virtual branch includes any informatics approaches, ranging from electronic health records (EHR), deep learning (DL), control of health management systems, and active guidance of physicians in their treatment decisions. The physical branch is best represented by robots and other hardware systems used to assist patients or a surgeon in a surgery room.

Clinical decision support system (CDSS) is a well-known name that refers to the program (virtual AI) that implements one or more AI methods into clinical practice in order to enhance medical decisions. Medical personnel may use CDSS for diagnosis, prevention, treatment protocol recommendations, care-coordination, treatment outcome predictions, and other tasks [2]. While there has been much progress reported in the efficiency of the artificial intelligence algorithms on various medical problems [3,4], still, the integration of AI solutions in a CDSS that performs well in clinical practice is rare and has proven to be notoriously difficult [5].

The development of CDSSs has been proceeding gradually, following the course of the improvements in AI algorithms, knowledge representation methods, software architectures and clinical data availability. While there are many AI approaches one could use for developing a CDSS, most of them can be classified into one of two classes: symbolic and machine learning (ML, including DL).

In this work, we consider the research approaches that were reported to be implemented in clinical practice or that were tested in real-clinical settings. We also briefly elucidate the principles of the underlying AI methods that were used in these systems. Since there has been many research works covering different aspects of CDSS in recent years, we stipulate the aim of this study as twofold: 1) a critical synthesis of current literature regarding the possibilities and challenges in implementing a successful CDSS, and 2) an overview of the best CDSS practices and challenges that, when overcome, have the potential to significantly improve healthcare.

Our work is structured as follows. In section II, we provide a critical review of the related work, with a focus on more general review studies that are related to CDSS. In section III, we provide an overview of the successful examples of CDSSs in practice. Then, in section IV, we delve into the AI technologies that underpin contemporary CDSSs. Section V is devoted to open challenges in designing and implementing a CDSS. We discuss some of the debatable CDSS topics in section VI. Finally, section VII concludes the paper.

II. RELATED WORK

The topic of our paper resembles several studies in recent years that pursue the goal of systematization of knowledge about CDSSs or AI methods in healthcare and raise awareness about the challenges related to these topics. In continuation, we provide a brief critical assessment of several recent important scientific review papers that deal with CDSSs.

Castaneda et al. [6] provided a general review paper on CDSS in 2015. They stipulated the important role of EHR in employing CDSS, however they also pointed out that EHR adoption is only the first step towards improving patient healthcare, since its use may not lead to improved clinical outcomes if it is not supported by models

constructed on large datasets, possibly derived from multiple databases. They consider that, currently, the use of CDSS in practice is limited, since they are mostly used to generate alerts, reminders and summaries. As a marked example of large-scale AI integrated CDSS, they mention IBM Watson, but also question its ability to improve treatment outcomes. The study also offers a valuable input on several real-world CDSS implementations.

The study of Middleton et al. [2] from 2016 offers a holistic 25-year retrospective and a 25-year vision in CDSSs. The study is an important source of information on evolvement of CDSSs in the USA, which includes mentioning of multiple CDSS implementations, some of which were used successfully in clinical practice. It describes six axes in CDSS development: data, knowledge, inference, architecture and technology, implementation and integration, and users, and provides elaboration along these axes. Several topics that adversely affect CDSSs are also discussed.

A review study by Dwyer et al. [7] from 2018 concerns the use of machine learning approaches in governing decisions in the field of clinical psychology and psychiatry. The review provides an insight into the potential for ML methods to augment clinical practice (diagnosis, prognosis and treatment) in the field, with description of the principles of these methods. The authors admit that there are very few successful CDSS implementations available and point to several reasons why this is the case, such as: cultural norms in clinical practice, validity of diagnostic and prognostic labels, representativeness of training data, mechanistic understanding of detected patterns, practical implementation, and others. Although CDSS implementations are not covered, the study is a relevant resource for understanding the AI algorithms that can be used in clinical psychology and psychiatry.

Topol [8] presented a recent review study dealing with the topic of high-performance medicine, in which he highlights the recent advances in AI, especially in DL and the applications in various medical domains. The author shares his opinion that the clinical practice is only at the starting point of using advanced DL approaches. The study is highly valuable as a harbinger of DL applications in CDSS, as it describes numerous important implementations, both of peer-reviewed publications of AI algorithms in healthcare and of FDA approvals of AI algorithms implementations. The study acknowledges the pitfalls of clinical implementations, with the example of IBM Watson, which in recent years made several erroneous recommendations that could have threatened patient safety. The author considers the current lack of prospective studies as a downside of DL-based CDSSs.

The study of Picardi et al. [9] deals with the significant topic of assurance in CDSSs. Especially, ML-based black-box models suffer from the issue of weak interpretation capabilities and generally low validity when the whole process is not closely controlled. The authors induce an assurance case pattern based on an important case study of retinal disease diagnosis. They suggest that the assurance pattern for a model should describe: a clinical setting (context of use), a benchmark (against a gold clinical standard), an ML model choices explanation, training and validation data sufficient to describe the clinical setting, an independent test data, and a thorough ML process explanation.

Kelly et al. [5] focus on key challenges for delivering clinical impact with artificial intelligence. Their study points to numerous existing challenges, including the ones that are inherent to ML methods, implementation and logistical difficulties, as well as the existing sociocultural barriers of adoption. They conclude that, based on the available reports, AI metrics such as accuracy, area-under-curve, and others often do not reflect clinical applicability, and especially any beneficial change in patient care. On that topic, they consider that clinicians need to be able to understand how the AI algorithms work, to improve patient care within a workflow they are familiar with, which is a topic often neglected in scientific papers. Regarding AI algorithms interpretability, they report that the recent European Union's General Data Protection Regulation (GDPR) legislation mandates the right to explanation for algorithmically generated user-level predictions that have the potential to significantly affect users [10].

The review study by Wang and Preininger [11] considers the state-of-the-art, challenges and future directions of AI applications in healthcare. They categorized the studies published in the period 2014 – 2019 into five separate categories according to the data type analyzed. These include multi-omics (genomics, proteomics, transcriptomics, etc.), clinical, behavioral, environmental, and pharmaceutical research and development data. They described each data type category, associated challenges and practical implications that have emerged over the last years. They stress out that both AI models' and data challenges exist that currently hamper the success of developed CDSSs.

Shu et al. [12] present the applications of AI methods in the field of pediatrics. The paper is valuable because of the multiple descriptions of AI research papers covering several medical branches in pediatrics (e.g. cardiology, neurology, radiology), as well as for a discussion on multi-omics research in CDSSs, including the consideration of cloud, clusters, and grid computing for high-precision medicine. Nevertheless, the work does not cover many successful practical implementations of CDSS, but instead focuses more on future perspectives and limitations that need to be overcome.

Krittanawong et al. [13] offer a practical view on DL-based AI methods applied to cardiology. They focus on explaining the properties, benefits and downsides of using these methods, and provide a guide to implementation options (e.g. hardware considerations, software packages, modeling strategies). Their work suggests that the use of DL in cardiology is still of limited extent in the number of conducted studies and that the studies may suffer from a positive publication bias, which remains to be discovered by future meta-analysis studies. Their work does not report on real-world CDSS implementations of DL.

III. An Overview of Successful Implementations of Clinical Decision Support Systems

According to Wright and Sittig [14], the evolution of CDSS architectures may be divided into four phases:

- Phase 1: Standalone decision support system
- Phase 2: Decision support integrated into clinical systems
- Phase 3: Standards developed for sharing content between clinical decision support systems
- Phase 4: Web service models for decision support systems

We consider that, from the AI perspective, a fifth phase, as a continuation of the fourth phase, may be distinguished, which includes big biomedical data analytics from the cloud and grid platforms using ML or DL approaches. The timeline presented in Fig. 1 illustrates some of the significant CDSSs either implemented in real-time clinical practice or demonstrated to be able to work in such a setting. The references for the shown CDSSs may be found in the studies [2,8,11,15–19]. The years listed in the figure are those of published research.

From Fig. 1, it can be observed that there is currently an abundance of high-quality research into ML and DL-based approaches, covering a great diversity of medical issues. It should be mentioned that only some of the most significant recent CDSS research works and implementations are depicted in Fig. 1, while there exist many others, especially those that use ML or DL methods in clinical settings.

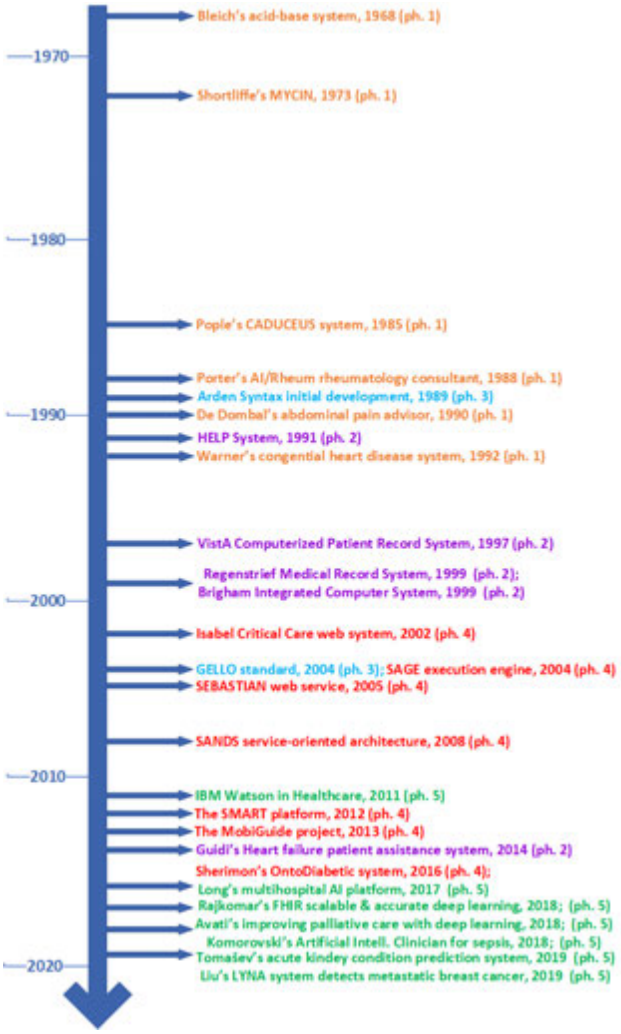


Figure 1. Significant CDSSs implemented in clinical settings or proven to work with real-time clinical data. Orange / violet / turquoise / red / green entries are for phase 1 / phase 2 / phase 3 / phase 4 / phase 5, respectively.

Aside from the clinical implementations of various CDSSs, in the last several years (i.e., from 2014), there has been a number of US Food and Drugs Administration's (FDA) approvals of proprietary AI algorithms and the accompanying platforms in healthcare (especially since FDA issued a fast-track approval plan for the AI algorithms in 2018). Some of these algorithms may be used for personal health assessment (e.g., via smart wearable devices), in primary care, or in clinical settings. Although a full list is not openly available, according to several resources, there are currently more than 50 such AI algorithms available [8,20].

IV. Clinical Decision Support System Technologies

A typical CDSS consists of a user interface (also called shell), an inference engine (also called core) and a data repository. As we have shown in the previous section, all CDSS components may be deployed in different settings, depending on the system architecture. The core usually implements one or more of the knowledge representation and reasoning schemes: symbolic and/or ML (also DL).

The symbolic (or declarative) approach includes all methodologies that deal with the representation of knowledge as a collection of symbols, including the relations among symbols. Such methodologies include formal mathematical logic (e.g., propositional, predicate, temporal), rule-based expert systems (they are an efficient subset of logic-based systems), imperfect knowledge inference systems (e.g. fuzzy rule-based systems, certainty factors, etc.), and Bayesian (probabilistic, belief) reasoning systems (e.g., Bayesian networks). The symbolic approach also includes biomedical computer ontologies, as a specific subtype of formal mathematical logic knowledge representation scheme (as ontologies are mostly based on description logic) [21].

The most common types of symbolic methods used in CDSSs are rule-based systems, ontologies and Bayesian networks [2,6]. A rule-based system usually contains an external (as a part of data repository) set of IF-THEN rules and valid facts. The reasoning may be performed either by forward or backward chaining. In forward chaining, the core makes the inference by matching the IF side of the rules with facts, resolves the conflict set (when more than one rule has its IF side satisfied), and activates (triggers, fires) the one selected rule, after which the reasoning cycle is repeated. Some of the advantages of using rule-based systems include:

- direct natural representation and the use of empirical knowledge acquired from an expert in a specific (narrow) domain
- modularity of knowledge collection and its simple maintenance
- efficient execution in narrow domains

- high interpretability of the automatic reasoning process

On the other hand, there are also some significant disadvantages of using rule-based systems:

- the rules acquired from the experts are heuristic and fragmentary ("knowledge nuggets") and do not explain the holistic relations in the process
- sometimes the rules may be contradictory, which is why maintaining consistency in a rule set is important
- the rules in a classical rule-based system are not robust and do not include uncertain and incomplete knowledge (this may be mitigated by using imperfect knowledge-based systems such as fuzzy rules)
- rule-based system stops being useful at the edge of the domain, while instead, like an expert, it should become monotonically weaker at problem solving
- clarifications, which result from the representation of chaining, do not provide causal connections in the process (there is no connection of ground or root causes with effects)

Unlike rule-based systems, ontologies are not intended primarily for executing actionable decisions. Instead, they are used as a mean to store hierarchical classification of terms and important relations among terms pertaining to a biomedical domain. They can be visualized as rather complex (meta)thesauri, covering different aspects of a biomedical topic or a more general biomedical knowledge domain. Contemporary ontologies use the open world assumption (which contrasts some important earlier knowledge organization schemes such as Frames [22]), under which no negation of a term's existence may be inferred unless its non-existence is explicitly stated. This somewhat limits their potential for constructing actionable rules [23]. Ontologies are mostly well-developed in the fields of biology and bioinformatics; however, they are still not well-integrated with other resources in implemented CDSSs [6]. Some of the reasons for this include:

- the complexity of biomedical vocabulary which limits the development of large ontologies
- integration of different ontologies is difficult
- slowness of the reasoning process for large ontologies

An example study by Gordon and Weng [24] used a semi-automated method to acquire data from diverse knowledge sources in order to evaluate and improve bacterial clinical infectious diseases ontology.

Bayesian networks may help medical experts in reaching accurate decisions [25]. They are based on probabilistic relations among nodes that represent variables that cover a topic. Both the structure of the network (the dependencies among variables) and the probabilities (either the a priori or the conditional ones) required by a network need to be derived from an expert's knowledge of a problem or from a dataset that representatively describes the covered topic. This, as well as the issue of interpretability of the reached decisions, severely limits their usefulness in CDSSs.

Machine learning techniques may be used in order to reach decisions based on available domain data. They cover a range of options, depending on the goal of the medical analysis: classification, regression, clustering, semi-supervised learning, etc. [7]. A simple division of ML techniques may be given in two categories: 1) black-box techniques, and 2) techniques that provide comprehensible models. Black-box ML techniques include support vector machines, artificial neural networks, ensembles of models (e.g. random forests) and other, nonlinear function-based models. Comprehensible models include linear and logistic regression models, decision trees and induction rules.

Deep learning can be considered as a specific subtype of machine learning that includes artificial neural network architectures consisting of multiple hidden layers with various transfer functions. Generally, DL methods are good at obtaining low-level feature representations from raw data, which enables them to be used out-of-the-box, i.e. without the significant human effort put into the engineering of higher-level features [11]. On the downside, all DL architectures may be considered as black-box models, since the distribution of weights in a network does not allow for an easy way of understanding why a decision is reached. The most common types of DL architectures used in CDSSs in the biomedical domain include:

- Convolutional neural networks (CNN) – by far the most common architecture type for biomedical imaging applications [8], consisting of multiple hidden convolutional and pooling layers
- Recurrent neural networks (RNN) – used most often for physiological time series modeling such as in cardiology or neurology, but also in predicting events onset [26]. Here, several subtypes of RNNs are used often, such as long-short term memory (LSTM) networks [27] and gated recurrent unit (GRU) networks [26]
- Deep belief networks (stacked restricted Boltzmann machines) [28, 29]
- Autoencoders – for unsupervised learning tasks, including dimensionality reduction and representation learning [29]

V. Clinical Decision Support Systems Open Challenges

The successful use of CDSSs in practice still suffers from some challenges that need to be properly addressed. In Table I, we compile a list of some frequently mentioned challenges discussed in the literature, which we categorize into major and minor ones, depending on our consideration about their relevance and the ability to be overcome. In continuation, we reflect on the major challenges.

AI model transparency (comprehensiveness) should be sought where possible in CDSSs. Due to the black-box nature of most ML and DL models, human needs to be present when reaching the final decision [11]. Additionally, providing a model interpretation is needed in healthcare, which is also enforced by GDPR [5].

Clinical workflows lack standardization for insertion points that can be used by developed CDSSs [2]. While

TABLE I. OPEN CHALLENGES FOR CLINICAL DECISION SUPPORT SYSTEMS

Challenge type	Short description	Reference
Major	AI model transparency	[5, 7, 11, 13]
Major	Standardization of DCSS integration into clinical workflow	[2, 7, 11]
Major	AI algorithms effectiveness	[5, 11, 13]
Major	Standardization of EHR data	[11, 30]
Major	EHR data are heterogeneous, sparse and noisy which hinders analysis	[11]
Major	Efficient external validation	[5, 7]
Major	Data integration from multiple sources	[5]
Major	Secure and scalable sharing of clinical data	[31]
Major	Data representativeness is lacking	[7, 11, 13]
Major	Regulation and rigorous quality control of AI devices is needed	[5]
Minor	Currently, too few prospective studies	[5]
Minor	Increasing trend of proprietary algorithms development	[11]
Minor	Positive publication bias in medical studies of ML and DL	[13]
Minor	Adversarial attacks as security risk	[5, 11]
Minor	Computationally expensive analysis of physiological data	[11]
Minor	Integrative analysis of physiological with other clinical data is lacking	[11]
Minor	Rapid development of new technology – many legacy systems	[2]
Minor	Behavioral data are heterogeneous, difficult to obtain ground truths	[11]
Minor	Environmental data linking with EHR data is needed for precision medicine	[11]
Minor	Heterogeneity of multi-omics data impedes its analysis and integration	[11]
Minor	Large scale knowledge engineering and management	[2]
Minor	Better understanding of human-computer interaction	[5]
Minor	Implementation issues: algorithms unavailable, too much training needed, not commercially viable	[7]
Minor	Data quality and diversity needs to be considered in addition to quantity when assessing ML and DL methods	[13]
Minor	Replication of study should be possible – models, data and code should be made openly available	[13]
Minor	Reporting of model results should be more elaborate with respect to metrics used	[13]

current practice inserts a CDSS where needed in the physician's workflow [11], the clinical practice would greatly benefit from having a more theoretically well-founded integration procedure.

Peer-reviewed randomized control trials need to be pursued where possible, since current practice shows that higher-accuracy AI systems do not necessarily result in better patient outcomes due to clinical inapplicability. Several researchers point to the necessity of improving patient outcomes as the ultimate goal of using CDSSs [5,13] Also, state-of-the-art performance of AI algorithms in many health applications is far from perfect, which leads to the question of the advanced AI algorithms effectiveness in practice. Currently, in some cases, results for DL vs simple logistic regression may not be favorable for DL [11]. It should be noted that due to the overfitting problem, sometimes simpler methods are more efficient and more practical [13]. Some of the unfavorable results may stem from too few dedicated DL model architectures specifically designed for medical images, which is mostly due to too few shared medical datasets. In these cases, transfer learning from natural images is used, which may be suboptimal [11].

Logistical problems that are related to data integration for efficient model construction abound. Namely, most of the healthcare data is not prepared for ML, as there are many image archival systems, EHRs, electronic prescribing tools and insurance databases available [5]. In this respect, unified data formats such as FHIR [30] could be beneficial; however, EHR data is still heterogeneous and has inconsistent semantic coding, which prevents data preparation for ML [11].

For efficient external validation of the model used by CDSS, site-specific adaptations need to be considered for generalizations (external dataset curation). A problematic fact reported is that very few (only about 6%) of medical studies perform external validation on datasets from outside of current institution [5]. The data representativeness is the problem of not having appropriate samples to construct a widely applicable model (e.g., there were too many light-skinned and too few dark-skinned images available for skin lesions classification [11]). One of the solutions to this problem is to increase the dataset size to include all the needed sample variations in a significant amount.

Scalable sharing of clinical data is a major challenge for the fifth CDSS architecture phase, since the abundance of medical data prevents previous solutions for data storage and sharing. In order to handle this challenge, cloud-based solutions seem to be a viable option, but this must be pursued on the scale of multiple hospitals in order to ensure CDSS validity. Secure and scalable information sharing may use blockchain or similar technology [31].

VI. DISCUSSION

It is a common assumption that a person working in a partnership with an information resource, such as a CDSS, is better than the same person unassisted [2,5]. However, a physician should take care when interpreting the results of CDSS recommendations. Any user, including the physician and the patient, needs to understand the CDSS benefits and limitations, and needs to be clear about which questions one can ask of the CDSS.

Currently, the development of AI methods by clinicians on-site instead of expertly designed models is still unsatisfactory due to the complexity of the AI methods in CDSS systems, but shows some advancements [32].

Explainable AI has seen a recent resurgence as a hot research topic in DL [11]; however, its results are still unsatisfactory for healthcare. We suggest that a viable option would be to use black-box or weakly understandable models only for image/signal analysis, and interpretable models for all other CDSS considerations (such as, for example, medical guidelines-based recommendations).

VII. CONCLUSION

This study has revealed that the successful implementations of CDSSs in practice are still uncommon and that there are many challenges that need to be overcome before CDSSs become ubiquitous. Since medical institutions have always strived to instill the highest level of trust in the patients that are admitted and treated there, the question of using AI-algorithms guided CDSS to improve patient outcomes is still debatable. Nevertheless, we consider that future, carefully conducted studies will show that the use of accurate and reliable CDSS is possible and that clinical practice would greatly benefit from its introduction into medical workflows.

REFERENCES

[1] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," Metabolism, vol. 69, s36–s40, 2017; doi:10.1016/j.metabol.2017.01.011.

[2] B. Middleton, D. F. Sittig, and A. Wright, "Clinical Decision Support: a 25 Year Retrospective and a 25 Year Vision," Yearb Med Inform, S103–S116, 2016; doi:10.15265/IYS-2016-s034

[3] H. Zheng, P. Zheng, L. Zhao, et al., "Predictive diagnosis of major depression using NMR-based metabolomics and least-squares support vector machine," Clinica Chimica Acta, vol. 464, pp. 223–227, 2017; doi:10.1016/j.cca.2016.11.039

[4] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," Nature Medicine, vol. 25, pp. 65–69, 2019; doi: 10.1038/s41591-018-0268-3

[5] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," BMC Medicine, vol. 17, p. 195, 2019; doi:10.1186/s12916-019-1426-2

[6] C. Castaneda, K. Nalley, C. Mannion, et al., "Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine," Journal of Clinical Bioinformatics, vol. 5, p. 4, 2015; doi:10.1186/s13336-015-0019-3

[7] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine Learning Approaches for Clinical Psychology and Psychiatry," Annu. Rev. Clin. Psychol., vol. 14, pp. 91–118, 2018; doi:10.1146/annurev-clinpsy-032816-045037

[8] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, vol. 25, pp. 44–56, 2019; doi:10.1038/s41591-018-0300-7

[9] C. Picardi, R. Hawkins, C. Paterson, and I. Habli, "A Pattern for Arguing the Assurance of Machine Learning in Medical Diagnosis Systems," in A. Romanovsky, E. Troubitsyna, F. Bitsch, Eds., SAFECOMP 2019, Lecture Notes in Computer Science, vol. 11698. Springer, Cham, pp. 165–179, 2019; doi: 10.1007/978-3-030-26601-1_12

[10] European Union General Data Protection Regulation (GDPR), May 23, 2018. https://gdpr-info.eu/, Accessed: Feb. 7, 2020.

[11] F. Wang and A. Preininger, "AI in Health: State of the Art, Challenges, and Future Directions," Yearb Med Inform, vol. 28, no. 1, pp. 16–26, 2019; doi:10.1055/s-0039-1677908

[12] L. Q. Shu, Y.-K. Sun, L. H. Tan, Q. Shu, and A. C. Chang, "Application of artificial intelligence in pediatrics: past, present and future," World Journal of Pediatrics, vol. 15, no. 2, pp. 105–108, 2019; doi:10.1007/s12519-019-00255-1

[13] C. Krittanawong, K. W. Johnson, R. S. Rosenson, et al., "Deep learning for cardiovascular medicine: a practical primer," Eur Heart J, vol. 40, no. 25, pp. 2058–2073, 2019; doi: 10.1093/eurheartj/ehz056

[14] A. Wright and D. F. Sittig, "A Framework and Model for Evaluating Clinical Decision Support Architectures," J. Biomed. Inform., vol. 41, no. 6, pp. 982–990, 2008; doi:10.1016/j.jbi.2008.03.009

[15] M. Sordo, A. A. Boxwala, O. Ogunyemi, and R. A. Greenes, "Description and Status Update on GELLO: a Proposed Standardized Object-Oriented Expression Language for Clinical Decision Support," in: Studies in Health Technology and Informatics, vol. 107: Proceedings of MEDINFO 2004 Conference, pp. 164–168, 2004; doi: 10.3233/978-1-60750-949-3-164

[16] P. C. Sherimon and R. Krishnan, "OntoDiabetic: An Ontology-Based Clinical Decision Support System for Diabetic Patients," Arab. J. Sci. Eng., vol. 41, pp. 1145–1160, 2016; doi: 10.1007/s13369-015-1959-4

[17] N. Tomašev, X. Glorot, J. W. Rae, et al. "A clinically applicable approach to continuous prediction of future acute kidney injury," Nature, vol. 572, pp. 116–119, 2019; doi:10.1038/s41586-019-1390-1

[18] Y. Liu, T. Kohlberger, M. Norouzi, et al., "Artificial Intelligence–Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists," Arch Pathol Lab Med, vol. 143, pp. 859–868, 2019; doi: 10.5858/arpa.2018-0147-OA

[19] E. Long, H. Lin, Z. Liu, et al. "An artificial intelligence platform for the multihospital collaborative management of congenital cataracts," Nature Biomedical Engineering, vol. 1, p. 0024, 2017; doi: 10.1038/s41551-016-0024

[20] The Medical Futurist, "FDA Approvals For Smart Algorithms In Medicine In One Giant Infographic," The Medical Futurist Institute, https://medicalfuturist.com/fda-approvals-for-algorithms-in-medicine/; published: June 6, 2019; Accessed: February 10, 2020

[21] F. Baader, "Description Logics," in: Reasoning Web: Semantic Technologies for Information Systems, 5th International Summer School 2009, Lecture Notes in Computer Science, vol. 5689, pp. 1–39. Springer–Verlag, 2009.

[22] M. Minsky, "A framework for representing knowledge," Psychology of Computer Vision, P. Winston, Ed., pp. 211–281, McGraw-Hill, New York, USA, 1975

[23] A. Jovic, M. Prcela, and D. Gamberger, "Ontologies in Medical Knowledge Representation," Proceedings of the ITI 2007 29th International Conference on Information Technology Interfaces, V. Luzar - Stiffler, V. Hljuz Dobric, Eds., Zagreb : IEEE Region 8, pp. 535–540, 2007; doi: 10.1109/ITI.2007.4283828

[24] C. L. Gordon and C. Weng, "Combining expert knowledge and knowledge automatically acquired from electronic data sources for continued ontology evaluation and improvement," J Biomed Inform, vol. 57, pp. 42–52, 2015; doi:10.1016/j.jbi.2015.07.014

[25] O. Pourret, P. Naïm, and B. Marcot (eds), Bayesian Networks: A Practical Guide to Applications, John Wiley and Sons, 2004

[26] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," J. Am. Med. Inform. Assoc., vol. 24, no. 2, pp. 361–370, 2017; doi: 10.1093/jamia/ocw112

[27] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "DeepCare: A Deep Dynamic Memory Model for Predictive Medicine," in: J. Bailey, L. Khan, T. Washio, G. Dobbie, J. Huang, and R. Wang, Eds., Advances in Knowledge Discovery and Data Mining. PAKDD 2016. Lecture Notes in Computer Science, vol. 9652, Springer, Cham, pp. 30–41, 2016

[28] W. Pinaya, A. Gadelha, O. Doyle, et al., "Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia," Sci Rep, vol. 6, p. 38897, 2016; doi: 10.1038/srep38897

[29] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," Brief. Bioinform., vol. 19, no. 6, pp. 1236–1246, 2018; doi: 10.1093/bib/bbx044

[30] A. Rajkomar, E. Oren, K. Chen, et al., "Scalable and accurate deep learning with electronic health records," npj Digital Med, vol. 1, p. 18, 2018; doi: 10.1038/s41746-018-0029-1

[31] P. Zhang, J. White, D. C. Schmidt, G. Lenz, and S. T. Rosenbloom, "FHIRChain: Applying Blockchain to Securely and Scalably Share Clinical Data," Comput Struct Biotechnol J, vol. 16, pp. 267–278, 2018; doi:10.1016/j.csbj.2018.07.004

[32] L. Faes, S. K. Wagner, D. J. Fu, et al., "Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study," The Lancet Digital Health, vol. 1, no. 5, e232–e242, 2019; doi: 10.1016/S2589-7500(19)30108-6

MOVIDA.eros : an eHealth Solution for Cardiac Rehabilitation Programs

R. Fonseca-Pinto[1,2], E. Silva[1,2], R. Martinho[2], R. Rijo[2], F. Januário[3], A. Antunes [3]
[1] ciTechCare – *Center for Innovative Care and Health Technology*, Leiria, Portugal
[2] Polytechnic Institute of Leiria, Leiria, Portugal
[3] CHL – Centro Hospitalar de Leiria, Leiria, Portugal
rui.pinto@ipleiria.pt

***Abstract* - MOVIDA.eros platform is a rehabilitation program emerging from a partnership between academia and a central hospital in Portugal. The platform combines a mobile application and a web backoffice platform to presceibe and monitoring physical activity. The system also allows changes to be made to the prescription, depending on physical performance and physiological response. These changes are automatically updated in the mobile app.**

The results regarding the adhesion and usability of the platform are encouraging. So far, the integration of the MOVIDA.eros platform into a classic cardiac rehabilitation program has allowed an increase in the number of patients under supervision, and has contributed to increasing the feeling of security during physical training. This program, in 1.5 years, has allowed the number of patients being monitored and treated to currently be 150, in a continuous monitoring regime (the classic program was designed for 5 patients every 3 months).

In the near future, the platform will be able to be easily integrated into Cardiac Rehabilitation Programs in other hospital centers and contribute to the increase in the number of patients in rehabilitation, which is a central strategy in the management of heart diseases, with special emphasis on myocardial infarction.

Keywords – ehealth, cardiac rehabilitation, mobile computing

I. INTRODUCTION

Cardiac rehabilitation constitutes as a central part in cardiovascular health programs and has demonstrated effective results in cases of acute Myocardial Infarction (MI), Heart Failure (HF), after angioplasty or post-cardiac surgery [1]. Although the widly known and comproved benefits, health systems are prone to provide adequate number of Cardiac Rehabilitation Programs (CRPs) and those that exist are not evenly distributed. This is the case of south Europe countries (e.g Portugal) [2]. CRPs are key in the optimization of physical, psychological and social capability (seen as secondary prevention) in post-myocardial infarction, heart failure or after coronary or valvular surgery.

The number of CRPs has been increasing, however, it is necessary to continue developing strategies so that all these multidisciplinary and integrating care programs be accessible at broad range level. Therefore, the use of mobile technology is a tool with the potential to spread these rehabilitation programs globally (mostly for low-risk patients, in particular in phase II and phase III of CRPs).

Besides the problem of accessibility, another of the issues identified is adherence to the CRPs. The distance from the place of residence to the rehabilitation center is indicated as one of the main constraints [3]. In addition to adherence, there are other factors that contribute to the sparsity of CRPs in some countries when compared with other countries in northern Europe, which are related to cultural aspects of weak adherence to the practice of physical activity, and also related to the scarcity of financial resources [2]. Accordingly, considering that accessibility and financial costs are two main barriers that hinder patients' adherence to PCR [4], the proposal for outpatient programs that address these problems is an emerging inevitability for health systems [1]. It was in this context that the program MOVIDA.eros emerged, being integrated into an interdisciplinary health promotion program at the community, supported by mobile technology, bigdata and artificial intelligence – the MOVIDA project [5, 6].

Technology is a constant in the day-to-day health care practice, and cardiovascular health and rehabilitation is no exception [7]. Nowadays, there is a trend in the use of technology in everyday life to access relevant information in terms of monitoring and health. Thus, the democratization of the use of smartphones and the increase of technological literacy in the population allows creating strategies based on technology and in the ability of data processing to develop customized health programs. The widespread use of mobile devices (i.e., smartphones, tablets, etc.) facilitates remote monitoring, and given the large amount of data it generates, it also allows for intelligent-based decisions supported by artificial intelligence algorithms. This is an important question for patients and physicians, as it allows increasing adherence and monitoring of several patients at the same time.

There are already some technology-based solutions based on mobile applications (Apps) for smartphones and tablets that allow the auto-management of the training program recommended by the CRPs and the registration of training data with subsequent submission to the health professional. Examples of such solutions are the HeartMapp and Heart Failure Health Storylines [8, 9]. The main difference regarding MOVIDA is the underneath concept of supervised prescription and real-time follow-up. After an initial patient assessment, the cardiac rehabilitation team prescribes physical activity through the web platform, which is sent automatically to the user's smartphone. Whenever the user accomplishes their training program, the records are automatically sent to the platform, enabling the physician to monitor effort perception, heart rate, and adverse events.

In this work, a new version of the platfom already presented in previous work [10] is described. Herin, the figures after one year of implementation of the program in 3 hopsital centers are reported.



Figure 1: Presciption and Exercise Management Platform

II. MATHERIALS AND METHODS

CRPs are divided into 3 main phases. Phase I is performed in the hospital, phase II is performed in a mixed model in an outpatient setting, and phase III is performed outside the hospital mostly at home. The MOVIDA.eros was created to be used by patients integrated into a CRPs and allows the establishment of a continuous relationship between prescription and monitoring, which is fundamental to obtain the confidence required to preserve adherence levels in the outpatient system. This platform was designed to be integrated in CRPs (mostly in phases II and III) and was conceived by a multidisciplinary team with experience in the prescription of physical exercise in rehabilitation programs. The requirements were defined taking into account the objectives, certified guidelines, and experience of the clinical team, transmitted to a team with experience in the development of mobile applications in health.

The program comprises a web platform and a mobile App and was developed for smartphones (Android And IOS), which makes it accessible to a wide variety of users. To use the platform, CRP professionals need a computer with access to the internet, and the patients need a smartphone to install the App and perform the prescribed training program. The platform also works in offline mode.

A. Users access to the platform

In terms of access to the system, for each rehabilitation center (RC) a local administrator (AdminRC) is designated. The AdminRC is responsible to create the user acoocunt for each health professional. After this initial step, each user will menage is own patients at the platform (insert data, precribe, send messages).

Each user registered by the AdminRC accesses the prescription platform through a web browser, where all the individual records of physical activity are/will be stored. At the registration procedure, an email is sent to the patient with a direct link to download the App (for IOS and Android), or alternatively, a code is generated to install the App after being downloaded to the patient's smartphone. After the initial registration, the platform is ready to be exploited. Each time the patient accomplishes the prescribed training program, the data will be automatically updated on the Web platform to be analysed by the CRP professionals.

B. Prescription and monitoring on the Web platform

Following the initial registration, at the Web browser, four separators are available to interact: Prescription in progress, Prescription history, Messages, and Profile. By default, the Prescription in progress folder is the main page, where is possible to view and monitor the physical activity of the patient when using the App, edit prescriptions and change them according to their performance (i.e. in a customized and adaptive way). The prescription is valid for a period of time, with the recommendation of target heart rate, a target Borg scale and type of exercise (walking or biking).

The messaging tab was implemented in a message-sending module that works bi-directionally. Thus, whenever patients need to report signs or symptoms they consider alarming, is possible to use the App-specific module to communicate directly with the platform. This tool can also be used by the CRP team to send motivational messages to patients feel accompanied, even remotely. At the Prescription history tab, is possible to see all the past prescriptions, as well as the performed changes to the initial prescription. The profile tab contains a summary of the clinical information that the CRP team initially obtained in the assessment at the initial evaluation of the patient.

The web platform was conceived to be fast-read and to be usable in the prescription and monitoring of registrations simultaneously. Accordingly, whenever possible, infographic signs for key indicators were applied, allowing for decision-making more intuitive and fast. In Figure 1 it is possible to observe an example of the activity of a patient concerning the ongoing prescription (3 in 5 visible days of activity. At the 3rd and 5th day Borg scale was above the maximum prescribed. At day 5th the duration was achieved)

For the first prescription, data from the initial evaluation based on complementary diagnostic exams (i.e. stress test) of the patient is considered. The prescription is based on the resting heart rate and reserve heart rate, which is, for each patient, customized by an adjustment factor. This formulation for the target heart rate is performed through the Karvonen formula, which is automatically implemented at the platform. For each patient a physician is associated to supervise the ongoing prescription, which is composed by a period of validity (which can be altered), a type (walking or static bike), a duration in minutes (which can be broken down into two periods on the same day), a Borg scale (to assess subjective perception of individual effort) and the maximum heart rate defined by the Karvonen formula.

There are also other infographic elements that allow a summary evaluation of the active prescription (percentage of the accomplished prescription) and specific information of each training session (maximum heart rate, distance traveled, average velocity, Borg scale, and alert symptoms). An example of these records for a given workout is shown in Figure 2.



Figure 2:Paltform view of workout setions

C. *APP Features*

Concerning the mobile application, the layout comprises a main screen and five tabs: Exercises, Prescriptions, Diet, Message and Profile. The App allows receiving the prescription as well send information regarding the activity effectively accomplished. The application automatically communicates with the prescription platform and is also running at the offline mode. Patients can also access their training history as well as record their evolution in terms of body mass index, and heart rate during the exercise. All these information is available graphically in the App.

One of the tabs in the App is relative to nutritional counseling that is established at the initial assessment and can be easily accessed from this tab. Figure 3 presents some examples of APP navigation screens.

III . RESULTS

The platform is now being applied in three hospitals in Portugal, integrated into the ongoing CRPs. The use of the platform in the real environment was previously validated by the Health Ethics Committee (HEC) of the hospital Center of Leiria.

As mentioned, the platform was integrated into the CRPs as an auxiliary tool for phases II and III (recovery at the hospital or at home respectively). Regarding the integration in phase III, the patients continue to perform their training program outside the hospital or rehabilitation center, and the monitoring task is made by the App. By using the App, it is possible to maintain these patients within the CRP remote supervision, which was not possible in the traditional programs due to the involved resources (human and physical). These patients continue having supervision through the mobile App by the CRP team, as in the previous 3 months (when at phase II), maintaining regular monitoring of the physical activity and making the appropriate adjustments whenever necessary. Still, after 6 months, these patients are again evaluated at the hospital by the CRP team.



Figure 3: App Screens

For patients in phase II where the platform is also included, the participants perform training sessions two times a week at the hospital, and in these sessions, they are monitored by telemetry and supervised by the CRP team. In this case, the MOVIDA.eros will be used to track training sessions in the remaining days of the weak, outside the hospital.



Figure 4: Example of one implemented prescription

The platform is being applied in three hospitals in Portugal, having 16 Medical Doctors and around 75 patients registed. All the technical support is kept by a computer science engineer.

Although the limited time span since the beginning of the program to draw solid conclusions, this work allows to report the experience that the CRP team has accumulated (from the beginning of the technological development to the phase of enrollment of real users) and to highlight the potential tool for using mHealth strategies in the context of CRPs.

Based on the data obtained from the platform is possible to select some examples to illustrate the platform functioning. Returning to the example in Figure 4, it is possible to verify that since the initial prescription, the patient did not use the App in 2 from the possible 7 days

already passed. The target duration of the prescription was allways reached and the maximum prescribed heart rate was not achieved in 2 days. Regarding Borg scale, in 3 out 5 days of exercice, the maximum range wasn't achieved) (visually it is possible to verify that the respective column is emphasized with a red line contour). Moreover, it is still possible to verify that one message was sent by the user to report adverse symptoms. In face of these records, the PRC team may change the prescription, and the information is automatically send to the App and a message is also send to the CRP user to be aware of the changes. This message module is key, especially for patients in phase III, as they are performing their training plan in their residency area.

The exchange of messages is, on the one hand, for the patient to have a tool that allows him to feel accompanied in case of doubts, and on the other hand, allows to send motivational messages by the CRP team. In case it is necessary to call the patient for an in site revaluation, this functionality can also be used as a communication channel.

The examples wherein presented aims to illustrate the potentialities of the MOVIDA.eros, and many other examples having clinical interest can be further reported in decision-making criteria regarding prescription changing.

There are naturally some constraints already identified, whose resolution will improve the experience of use and interest of the program. One of the identified problems is related to heart rate measurement. In the current version of the platform, these values are manually entered by the patient in the App. In fact, during the classic CRP, the patients attend to educational sessions where they learn how to register the carotid or radial pulse, and this problem is already somehow mitigated, but may not be the ideal solution. There is also the possibility of using wearable devices to register heart rate and send the information via Bluetooth directly to the platform (this functionality is already implemented in a new version of the platform).

III . CONCLUSIONS

This work presents a platform for prescription and monitoring of physical activity that can be used as a tool to overcome some of the problems identified in CRPs using a mHealth solution: the MOVIDA.eros. Moreover, some examples of real user cases were also presented and discussed. The platform was developed by a multidisciplinary team (Clinical and Engineering) and can be seen as an example in which technology can be used to optimize resources and create synergies that allow health services to provide personalized care, contributing to improving quality of life.

In the near future the results of an ongoing study using qualitative analysis techniques (content analys and focous groups) reporting the experience of patients and health professional will be also reported elsewer.

ACKNOWLEDGMENT

This work was supported by Fundação para a Ciência e Tecnologia FCT- Portugal, under the scope of MOVIDA project: 02/SAICT/2016 – 23878 and UID/05704/2020

REFERENCES

[1] American Heart Association, What is Cardiac Rehabilitation?, 2018. Available at: http://www.heart.org/HEARTORG/Conditions/More/CardiacRehab/What-is-Cardiac-Rehabilitation_UCM_307049_Article.jsp. [Consulted: 16-Feb-2020].

[2] M. Mendes, cardiac rehabilitation in Portugal. Point of the situation in 2013 - 2014 Portuguese Journal of Cardiology. 2016; 35 in: 637-718.

[3] H. M. Dalal, A. Zawada, K. Jolly, T. Moxham, and R. S. Taylor. Home based versus center based cardiac rehabilitation: Cochrane systematic review and meta-analysis. BMJ. 2010; 1-15.

[4] V. M. Conraads, Christi Deaton, Ewa Piotrowicz, Nuria Santaularia, Stephanie Tierney Massimo F. Piepoli et al. Adherence of heart failure patients to exercise: Barriers and possible solutions, Eur. J. Heart Fail., 2012 14:451-458.

[5] Fonseca-Pinto, R., Rijo, R., Assunção, P., Seco, M. A., Guarino, M. P., Braga-Pontes, C., ... & Pires, G. (2019, May). Prescribe and Monitor Physical Activity Through a Community-Based eHealth Program: MOVIDA Platform. In International Conference on Medical and Biological Engineering (pp. 13-19). Springer, Cham.

[6] Emanuel Silva, Rui Rijo, Ricardo Martinho, Pedro Asunción, Alexandra Seco, Rui Fonseca-Pinto, A Cardiac Rehabilitation Program Supported by mHealth Technology: The MOVIDA. Eros Platform, Procedia Computer Science. 2018; 138, 119-124.

[7] S. Hamine, E. Gerth-Guyette, D. Faulx, B. B. Green, and A. S. Ginsburg. Impact of mHealth Chronic disease management on treatment adherence and patient outcomes: A systematic review. Journal of Medical Internet Research. 2015.17 (2): E52.

[8] Lola Koktysh. Top 4 heart failure apps to reduce readmissions and power self-management 2017. Available: Https://www.scnsoft.com/blog/top-4-heart-failure-apps. [consulted: 10-Feb-2019].

[9] A. Pai, "Tablet app reduces CHF patient readmissions by 53 percent," 2015. Available at: https://www.mobihealthnews.com/42273/tablet-app-reduces-chf-patient-readmissions-by-53-percent. [consulted: 16-Feb-2020].

[10] Silva, E., Rijo, R., Martinho, R., Assunção, P., Seco, A., & Fonseca-Pinto, R. (2018). A Cardiac Rehabilitation Program Supported by mHealth Technology: The MOVIDA. eros Platform. Procedia computer science

Smart Healthcare Ecosystem for Elderly Patient Care

A. Abugabah* and N.Nizamuddin*
* Zayed University, Abu Dhabi, United Arab Emirates
Ahed.abugabah@zu.ac.ae, Nishara.nizamuddin@zu.ac.ae

Abstract - **The healthcare sector is one of the rapidly growing service-based sectors in the world. Constant technological adoptions have been implemented to provide enhanced patient care and other healthcare-related services. Recently, the rapidly growing Internet and sensing technologies, cloud platforms, and remote healthcare monitoring have paved a way to build smart healthcare ecosystems. However, special attention needs to be given for delivering high-quality clinical services for the growing elderly population and critically ill patients who are finding it difficult to reach out for professional medical help either due to terminal illness or because of their remote geographical location. In this paper, we propose a service model for a smart healthcare ecosystem where the patient data is collected via medical IoT sensors connected to the patient, sensor's data is stored in cloud infrastructure, and is analyzed by an expert from a remote telemedicine center. Moreover, an authorized telemedicine infrastructure's person can regularly monitor the activities of the caregiver and interact with the patient without having the patient to visit the hospital. This proposed healthcare service model aims to improve trust, reliability, and cost-effectiveness of the overall healthcare service delivery.**

Keywords - Cloud data storage; Home-based healthcare service delivery; IoT; Medical sensors; Telemedicine (keywords)

I. INTRODUCTION

With the advancement of healthcare technology, the average life expectancy of an individual has increased. The growing elderly population has caused concern for the government in terms of providing pensions, quality healthcare support, providing safety and protection. Soon, many countries are likely to handle situations related to social and economic pressures from a growing older population [1]. According to United Nations Organization's "World Population Prospects", by the year 2050, 1 in 6 people in the world would be above the age of 65, which will amount to 16% of the total population [2]. Even the developed countries that provide a good quality of life and healthcare are now framing new strategies and healthcare policies to cope up with providing the aging population, better healthcare, home health care benefits and pension schemes [1] [2] [3]. Due to the advanced industrialization, most elderly people live an isolated life away from their loved ones and families which makes them vulnerable. Especially, elderly people living in the rural side, developing countries, and those who are critically ill and need emergency healthcare services might not be always able to travel to hospitals to get the required medical aid [3]. The elderly population often needs medical assistance for age-related health issues, or sometimes for other ailments such as depression and anxiety. In most cases, those living in remote rural areas are unable to get access to quality healthcare.

The advancement in IoT technology and Telemedicine can be exploited to provide critically ill and elderly patients a better home-based or self-help healthcare service [4] [5] [6] [31]. Internet-of-Things technology has influenced many industries following the development of the Internet and web technologies. The healthcare industry has also started embracing the IoT technology and other new innovative, hi-tech technologies to improve healthcare service delivery to patients [7]. Modern healthcare technology has harnessed the benefits of IoT technology by employing IoT powered sensors, which collects a huge amount of data from the patient. This, in turn, helps the medical professionals to diagnose, analyze, and provide high-quality healthcare services. Figure 1 shows the typical lifecycle of data in the IoT environment. On the other hand, cloud computing is constantly evolving and is promptly becoming a necessity in the medical field [8] [9] [10]. The on-demand services offered by the cloud platforms enable the service providers to find this technology to be the best choice for their industry as it is scalable, provides elasticity, and is cost and time-efficient [13] [14]. All the requirements and features including the network infrastructure, databases, software, immunity to local hardware failure, and network convenience to store, and retrieve patient electronic health records (EHR) made cloud technology to be efficient in handling patient data. With the increased number of connected medical devices and regular collection of data from IoT sensors connected to patients, it has become an essential factor for healthcare organizations to store clinical data as per HIPAA-compliance and make it easy for authorized medical professionals to access patient data at any time [11].

With such technological advancements, the healthcare industry can now focus on high-quality healthcare service delivery rather than concentrating on infrastructure, storage, and network-based issues. The development of Internet technology and its predominant existence in the healthcare industry concerning patient information management, appointment scheduling, centralized data storage, and improved data access across various sections of the medical field has provided voluminous benefits for the healthcare industry. Telemedicine is a field that has started gaining traction since the time of its discovery. It is a ground-breaking service in the medical field as it offers remote delivery of healthcare services, diagnosis, video-based consultations, and treatment [4] [5] [12]. Patients

can avail of high quality and expert healthcare services even from remote places in the world. Due to the increased influence of digitization on the healthcare industry and also as this industry deals with huge volumes of sensitive data, there is a need to build a robust system that benefits the patient as well as the healthcare service provider (hospitals).

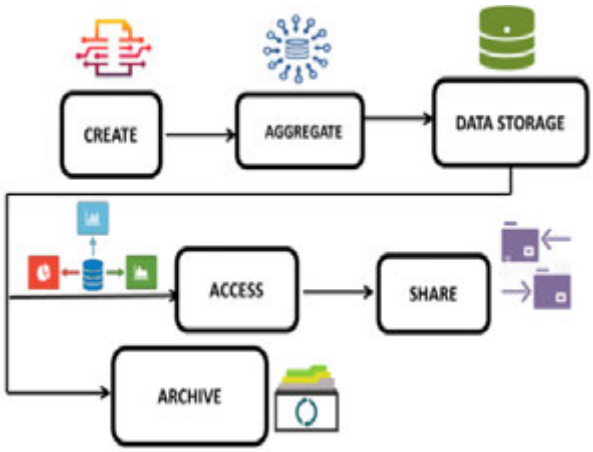


Figure 1. Typical IoT Data Lifecycle.

In this paper, we present an IoT based framework for remote elderly care. We demonstrate a framework that is powered with medical IoT sensors connected to the critically-ill elderly patient that collects data from various medical sensors connected to the patient. This data is then stored in the public cloud which can be then accessed by the medical professional from a remote telemedicine center via a web-based user interface. Our framework also helps the doctors/experts in a remote telemedicine center to monitor the health condition of the patient with the help of data collected from sensors, view the records, track the nurse/caregiver's behavior, interact with the patient and prescribe drugs depending upon the patient's condition. The remainder of this paper is organized as follows. Section 2 presents related work. Section 3 illustrates the proposed smart healthcare framework and its important aspects. Section 4 concludes the paper.

II. Related Work

In this section, we present the work available in the current literature for proposed methods and techniques for healthcare data and ecosystem management using modern technology and IoT sensors. For instance, [17] focuses on the impact that IoT and sensor technology will have on the healthcare industry and its applications to track patient's medical history, future appointments, nutrition, and medicinal demands, blood pressure, and heart rate monitoring, physical activity and future treatment requirements. Further the authors [17] discuss the state of the art facility of the healthcare industry, and to combat the future challenges that have to be handled by the healthcare service industry.

On the other hand, Domingos et al [18] proposed a stochastic workflow method to assess the reliability and quality of IoT based healthcare services. The model proposed [18] can analyze the alternatives at design time to select suitable participants, monitor, and execute the process. This model was based on Ambient Intelligent healthcare systems. Integration is one of the key factors to be considered when delivering healthcare services, as patients these days tend to visit more than one physician for their ailments. Utilizing IoT technology to identify the patient, track the patient's medical history, and collect data from sensors connected to the patient, and AI-based technology that improves decision-making practices of the healthcare service provider [19]. The advanced growth in the healthcare sector and the organization's inclination towards a distributed healthcare management system including, Telehealth, e-health, and telemedicine has improved because of the technological advances offered by IoT and sensor technologies. However, the authors [19] claim that the healthcare industry is still in its infancy to make a mark with IoT technologies and standards need to be revised.

Industries have started to embrace the ICT revolution and have started harnessing the benefits for their businesses. Cloud computing and IoT have benefitted many industries including the supply chain, logistics, the digital publishing industry, and data-centric industries such as retail, banking, healthcare [23]. Darwish et al [20] present an extensive analysis of current literature on integrating cloud computing and IoT technologies to solve problems of distributed data and resource management, effective handling of drug prescription, and improved remote healthcare service delivery. The proposed system [20], combining Cloud and IoT technology, aims to resolve the problems existing in integrating highly complex healthcare structures. The paper also discusses the challenges involved in integrating both these technologies.

IoT is becoming a dominant technology in the digital healthcare domain since its initiation and is now the number of connected IoT devices that are expected to reach 161 million by 2020 [21]. A huge amount of data is collected from millions of connected devices that lead to an increased challenge in developing a holistic approach for achieving data interoperability. Mavrogiorgou et al [21] proposed a mechanism that efficaciously addressed the integration challenges of huge volumes of data derived from different devices. The raw data obtained from devices are then cleaned, to ensure and capture the quality of each dataset. This data is then evaluated and is checked to achieve more than 90% accuracy.

Security and integrity play a significant role when dealing with sensitive medical data and several data encryption and anonymization techniques have been employed across the healthcare industry [22] [23]. Elhoseny et al [22] propose a hybrid security model that integrates the "*two-dimensional wave transform level steganography technique*" with an encryption scheme. This proposed model is built by combining many well-known encryption algorithms and begins by encrypting secret data and conceals the details of the data. The proposed model [23] proved to safeguard the sensitive patient data, by converting the data into a cover image with enhanced security and minimal deterioration of the image.

IoT based healthcare devices are more susceptible to data hacks, attacks, and security threats when compared to any other network devices. Though the state of the art facility of the connected devices has a reliable level of security, there is still a huge threat of collusion attack and data leakage. Luo et al [24] proposed a framework that protected patient privacy and prevented the occurrence of such attacks. The authors [24] present a secret sharing algorithm, that uses multiple, distributed cloud servers that conceals the identity of the patient and maintains uncompromised security. The patient's details are shared with the physicians by a shared patient data access control method, where multiple clouds collaborate to provide the encrypted data without revealing the identity of the patient to any participant within the cluster.

Technology has counterbalanced the growing needs of data management in the healthcare sector. The recent innovation in technologies such as mobile communications, m-health, e-health, Telehealth, AI-based health monitoring platforms, and the Internet, made it possible to detect, assist and respond to epidemic outbreaks such as Zika, Ebola, SARS and the recent outbreak of novel – coronavirus in various African countries and China [27] [28] [29]. Though technological innovation cannot combat the outbreak directly, several measures including proper communication about the epidemic outbreaks, data collection and management from wearable devices, and distributed countermeasure manufacturing can help the medical fraternity to fight an epidemic outbreak [28] [29].

The advancement in Information and Communication Technologies has improved human lives on a larger scale [26]. The use of IoT technologies and wearable sensor devices has brought physicians much closer to patients irrespective of the geographical location. This technology helps the physicians to monitor patients and their activities daily and was proved effective when implemented on the growing elderly population to monitor their general wellbeing. For critically ill elderly patients, the readings collected from wearable sensor devices and other vital signs monitoring IoT devices helped to track every movement of the patient regularly. Subasi et al [25] presented a system that was based on human activity recognition supported by IoT technology and by applying data mining techniques. The body movements are recorded regularly and the data mining model uses this dataset to identify vital signs to deliver mobile health care services with enhanced accuracy and increased patient safety.

Industrialization, distributed availability of medical services, and recent innovations in the healthcare sector has led to the increased life expectancy of the elderly population [1] [2] [26] [30]. Governments have started giving a lot of insight into these details like a huge sum of amount is being invested for healthcare expenditure in the form of insurance for critically ill patients. Dozet et al [30] discuss two different cases of technology diffusion for treating patients that includes both young and elderly population. Certain cases such as coronary bypass surgery and dialysis are becoming very common among both sides of the population due to lifestyle changes and stress at the workplace. The study [30] explores the fact that the dissemination of technology could impact the per capita health care expenditure among the older population.

III. PROPOSED FRAMEWORK

In this section, we propose our solution that works on the combination of three technologies – IoT, public cloud storage infrastructures, and the Telemedicine center. Figure 2 presents our proposed framework of smart healthcare ecosystem for elderly care.

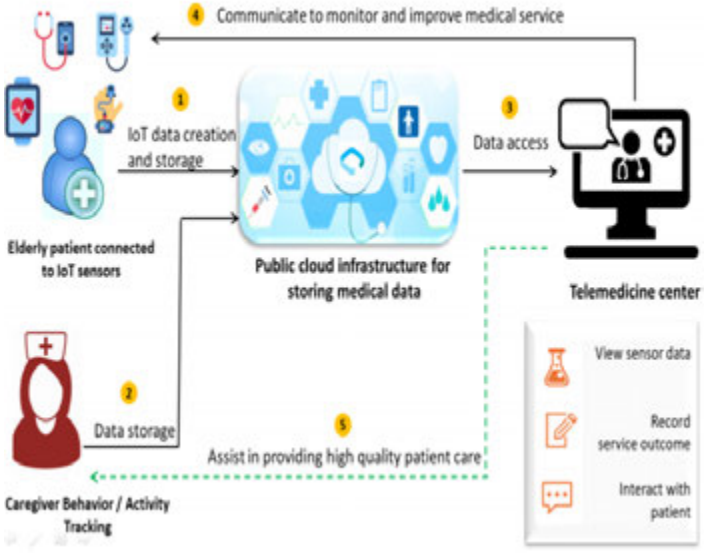


Figure 2. Smart Healthcare Ecosystem for elderly healthcare.

The elderly person / critically ill patient is connected with several life-saving devices, and medical sensors to detect their health condition. These sensors include sensors to detect blood sugar, body temperature, and blood pressure monitoring sensors, heart rate, breathing rate, fall detection, and body posture sensors. These sensors continuously monitor the condition of the patient and store all of the sensor data on a public cloud storage infrastructure. The cloud storage is usually comprised of a data server that the user – in this case, the caregiver uploads the files or it is uploaded from the IoT sensors connected to the patient that sends the collected medical data to the cloud either via Wi-Fi, Bluetooth, low-power wide-area networks (LPWAN), or by directly connecting to the Internet via Ethernet [13]. The data stored on the cloud infrastructure is accessible by the authorized medical expert in a remotely located telemedicine center via a web-based user interface. The medical expert in the telemedicine center monitors the patient records, sensor readings, any fluctuations in the patient's health history via the web-based user interface. The doctor will be then able to diagnose the ailment and prescribe any new drugs if necessary. The medical expert, on the other hand, would be also able to make a video consultation with the patient. Such a framework is not only time-efficient but also is cost-efficient as the patient need not travel to the healthcare center to get quality healthcare. Moreover, the telemedicine center's authorized personnel can also monitor the behavior and activity of the patient's caregiver

or nurse. The activity of caregiver and telemedicine center-based healthcare professional is recorded and is monitored regularly which makes the entire framework to be transparent and reliable.

For a critically ill patient or an elderly patient, it is extremely hard to plan a travel to the hospital to get medical assistance and it is equally tough for the family members to accompany them throughout the journey when the patient is connected with therapeutic gadgets for monitoring their wellbeing. Telemedicine technology juxtaposed with cloud computing and medical IoT sensors can enable to monitor and improve the facility provided to the patient as well as supports the loved ones and professional caregivers of the patient. Remote video monitoring and consultation will not only reduce the cost of the travel for the patients but also help them to get treated by expert professionals in the medical field who are sometimes not available to deliver their services in developing countries or rural communities. Further, utilizing cloud-based infrastructure will not only improve scalability but will ensure on-demand service, effective data storage, and reduced costs for storing data. This reduced cost on data storage infrastructures will aid the healthcare service providers to focus on superior healthcare service delivery. In this paper, we aimed to propose a framework for smart healthcare service delivery for critically ill and elderly patients.

This framework can be further extended for enhancing care given to emergency cases, general health, enhances legal standing evidence for the doctors as the video calls and consultations are on record, and for remote video-based counseling. The healthcare service provider needs not to worry about data storage, network infrastructure, and updated software as it will be dealt by the cloud platforms. Further, as a part of future development the hospitals can prefer to build their private cloud infrastructures to gain more control over the medical data storage, but deploying a private cloud might not be cost-effective, and also requires expert staff for monitoring and maintenances [14]. However, there are concerns and debates about the safety and security of sensitive medical data stored on cloud servers remotely. The storage of a hospital's patient data on the public cloud might make some of the service providers as well as the patients uncomfortable. There is an increased vulnerability for cloud storage to be hacked, as the Internet itself is not safe. In such cases, as a part of future work, we plan to implement a telemedicine-based patient monitoring system with the help of the decentralized blockchain technology where the data gets stored in the interplanetary file system (IPFS) and a hash code will be stored in the shared digital ledger (i.e., smart contract). Any illegal access, modification, or corruption of medical data will be immediately notified to all the participants involved within the network thereby ensuring enhanced security and data privacy of sensitive medical records. Decentralized ledger technology will further help in managing the medical lab diagnosis, medical records management, appointment scheduling, improving quality service [15] [16] [32]. Figure 3 shows the upcoming trends that will influence the healthcare industry in the near future as proposed by Forbes [33].

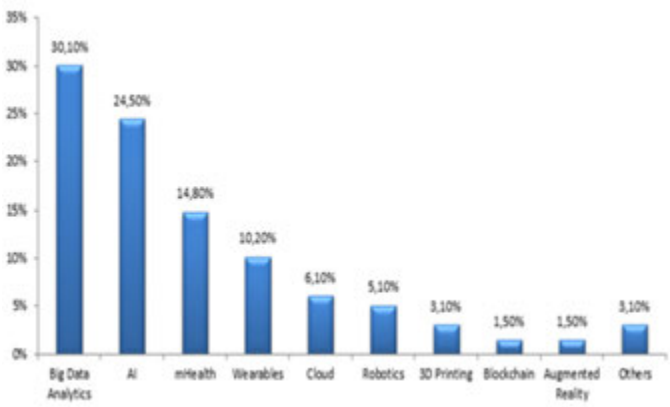


Figure 3.Future Technologies that will impact healthcare [33]

Big data analytics, on the other hand, has started to gain momentum in the field of healthcare service delivery [34] [35]. The humongous volume of data generated from devices and EHRs have paved the way for several big data applications for the medical field. A data-driven approach can help in medical drug research; and to identify disease trends and health strategies based on geography, demographics, socio-economic factors; data mining techniques to analyze and identify the cause of illness, and to support exclusively hyper-personalized healthcare for critically ill and elderly patients.

IV. Challenges and Future Work

The digital era has seen the IoT technology to be identified as one of the most dominant trends in the industry. The growth of Internet technology, distributed systems, and IoT has led to the development of smart homes, smart digitalized healthcare services, autonomous vehicles, and smart cities. However, there are several challenges to be addressed before an extensive digitalized implementation is possible [36-39]. Table 1 shows the security challenges for implementing IoT systems.

Table 1. Challenges for implementation of IoT systems in healthcare.

Identified Challenges / Threats	Impact
Outdated hardware and software. [38] [39] [40]	The constant change in the name of innovation has increased the security of new IoT based systems [38] [39].
Malware & ransomware attacks [38] [39]	Smart homes and offices installed with IoT enabled cameras are prone to cyber risks and attacks [38][39]
Patient Data collection & management [36] [38] [40]	Patient data is collected from various devices and sensors and is exchanged between users and service providers via the Internet, where there is a high risk of data breach [36] [38]

Data protection and security challenges [36] [38] [40]	Lack of compliance with regulations and laws leads to security incidents [37] [38].
Integration with multiple devices [37] [40]	There is no synchronous protocol that is available for data aggregation [37]
Data accuracy [37] [40]	IoT sensors collect data without any precise accuracy.
Cost [37] [38] [39 [40]	Cost remains to be a greater challenge when implementing IoT technology for healthcare services.
A large amount of DataAccumulation [39] [40]	IoT sensors are implemented across several services in the medical field including tracking assets, managing emergency rooms, and patient data. The large flood of data created by sensors and other medical gadgets challenges the organization's capability in handling the huge volume of data [39] [40].

As a part of future work, we plan to implement a fully functional IoT based remote healthcare service framework thereby making the monitor the health condition, consult, and provide remote healthcare service for elderly and terminally ill patients, who cannot afford to travel to a healthcare service center

V. CONCLUSION

Technology powered healthcare solutions and service deliveries are being implemented precisely by most of the healthcare and medical service-based organizations. In this paper, we proposed a framework for an IoT and Cloud infrastructure powered telemedicine framework for improving the elderly healthcare service. Our proposed model focuses on building a transparent, cost-efficient, and robust system that monitors and records patient data continuously with the help of medical IoT sensors connected to the patients. The sensors store the data on cloud infrastructure which is then accessed by authorized medical personnel with the help of a web-based interface at the remote telemedicine center for diagnosis, video consultation, and treatment. This proposed model aims to improve the quality of healthcare service provided to the elderly and critically ill patients. On the other hand, this gives a chance for the doctors to support the caregivers, monitor the activity of the caregiver or the nurse, and have proof in case of legal standing.

ACKNOWLEDGMENT

This research is supported by Zayed University Under Grant No R18088.

REFERENCES

[1] K. Lawler, "Coordinating housing and health care provision for America's growing elderly population", 2001.

[2] United Nations, "World Population Prospects 2019", Available [online]: https://population.un.org/wpp/, Accessed: 25 January 2020.

[3] R. T. Goins, U. Kategile, and K.C. Dudley, "Telemedicine, rural elderly, and policy issues", Journal of aging & social policy, 2002, vol. 13(4), pp.53-71.

[4] H. Xie, X. Zhang, N. I. Xuefeng, H. Jiang, and L.I. Dongjing, "Application of telemedicine in the elderly population", Chinese Journal of General Practitioners, 2016, vol. 15(10), pp.805-808.

[5] J. Boudy, J. L. Baldinger, F. Delavault, M. Muller, I. Farin, R. V. Andreao, S. Torres-Müller, A. Serra, D. Gaiti, F. Rocaries, F. and C. Dietrich, "Telemedicine for an elderly patient at home. In Smart Homes and Beyond", 4th International Conference on Smart Homes and Health Telematics (ICOST 2006), 2006, June, vol. 19, pp. 74-81.

[6] P. P. Ray, "Home Health Hub Internet of Things (H 3 IoT): An architectural framework for monitoring the health of elderly people". In 2014 International Conference on Science Engineering and Management Research (ICSEMR), 2014, November, IEEE, pp. 1-3.

[7] S. J. Park, M. Subramaniyam, S. E. Kim, S. Hong, J. H. Lee, C. M. Jo, and Y. Seo, "Development of the elderly healthcare monitoring system with IoT". In Advances in Human Factors and Ergonomics in Healthcare, 2017, Springer, Cham, pp. 309-315.

[8] O. Shimrat, "Cloud computing and healthcare". San Diego Physician. org, 2009, pp.26-29.

[9] N. Sultan, "Making use of cloud computing for healthcare provision: Opportunities and challenges". International Journal of Information Management, 2014, vol. 34(2), pp.177-184.

[10] S. P. Ahuja, S. Mani, and J. Zambrano, "A survey of the state of cloud computing in healthcare", Network and Communication Technologies, 2012, vol.1 (2), p.12.

[11] B. C. Drolet, J. S. Marwaha, B. Hyatt, P. E. Blazar, and S. D. Lifchez, "Electronic communication of protected health information: privacy, security, and HIPAA compliance", The Journal of hand surgery, 2017, vol. 42(6), pp.411-416.

[12] G. C. Doolittle, and R. J. Spaulding, "Defining the needs of a telemedicine service", In Introduction to Telemedicine, second edition, CRC Press, 2017, pp. 79-92.

[13] P. Mell, and T. Grance, "The NIST definition of cloud computing", 2011.

[14] M. Sajid, M., and Z. Raza, "Cloud computing: Issues & challenges". In International Conference on Cloud, Big Data and Trust, 2013, December vol. 20(13), pp. 13-15.

[15] M.P. Mcbee and C. Wilcox, "Blockchain Technology: Principles and Applications in Medical Imaging", Journal of digital imaging, 2020, pp.Journal of digital imaging, 02 January 2020.

[16] Y. Chen et al., "Blockchain-Based Medical Records Secure Storage and Medical Service Framework", Journal of Medical Systems, 2019, 43(1), pp.1–9.

[17] D. Domingos, A. Respício, and R. Martinho, "Reliability of IoT-aware BPMN healthcare processes", In Virtual and Mobile Healthcare: Breakthroughs in Research and Practice, 2020, IGI Global, pp. 793-821.

[18] S. Ansari, T. Aslam, J. Poncela, P. Otero, and A. Ansari, "Internet of Things-Based Healthcare Applications", In 2020 IoT Architectures, Models, and Platforms for Smart City Applications, IGI Global, pp. 1-28.

[19] T. Adhikary, A. D. Jana, A. Chakrabarty, and S. K. Jana, "The Internet of Things (IoT) Augmentation in Healthcare: An Application Analytics", 2019, January, In International Conference on Intelligent Computing and Communication Technologies, Springer, Singapore, pp. 576-583.

[20] A. Darwish, A. E. Hassanien, M. Elhoseny, A. K. Sangaiah, and K. Muhammad, "The impact of the hybrid platform of the internet of things and cloud computing on healthcare systems: opportunities, challenges, and open problems", Journal of Ambient Intelligence and Humanized Computing, 2019, 10(10), pp.4151-4166.

[21] A. Mavrogiorgou, A. Kiourtis, K. Perakis, S. Pitsios, and D. Kyriazis, "IoT in Healthcare: Achieving Interoperability of High-Quality Data Acquired by IoT Medical Devices", Sensors, 2019, 19(9), p.1978.

[22] M. Elhoseny, G. Ramírez-González, O. M. Abu-Elnasr, S. A. Shawkat, N. Arunkumar, and A. Farouk, "Secure medical data transmission model for IoT-based healthcare systems", IEEE Access, 2018, 6, pp.20596-20608.

[23] N. Nizamuddin, and R. Pandey, "Enhancing Security in Public Clouds using Data Anonymization Techniques", International Journal of Computer Applications, 2015, 975, p.8887.

[24] E. Luo, M.Z.A. Bhuiyan, G. Wang, M.A. Rahman, J. Wu, and M. Atiquzzaman, "Privacyprotector: Privacy-protected patient data collection in IoT-based healthcare systems", IEEE Communications Magazine, 2018, 56(2), pp.163-168.

[25] A. Subasi, M. Radhwan, R. Kurdi, and K. Khateeb, "IoT based mobile healthcare system for human activity recognition", 2018, February, IEEE, In 2018 15th Learning and Technology Conference (L&T), pp. 29-34.

[26] A. Faulkner, "Medical technology into healthcare and society: A sociology of devices, innovation, and governance", 2008, Springer.

[27] E. Tambo, A. Kazienga, M. Talla, C. F. Chengho, and C. Fotsing, "Digital technology and mobile applications impact on Zika and Ebola epidemics data sharing and emergency response", J. Health Med. Informat, 2017, 8, p.254.

[28] J. Hamilton, and R.S. Hopkins, "Using Technologies for Data Collection and Management", Centers for Disease control and prevention, Available [online]: https://www.cdc.gov/eis/field-epi-manual/chapters/data-collection-management.html, Accessed: 08 February 2020.

[29] Medical Futurist, "How Digital Health Technology Can Help Manage the Coronavirus Outbreak", Available [online]: https://medicalfuturist.com/how-digital-health-technology-can-help-manage-the-coronavirus-outbreak/ Accessed: 08 February 2020.

[30] A. Dozet, C. H. Lyttkens, and P.Nystedt, "Health care for the elderly: two cases of technology diffusion", Social Science & Medicine, 2002, 54(1), pp.49-64.

[31] R. M. Logue, "Self-medication and the elderly: How technology can help", AJN The American Journal of Nursing, 2002, 102(7), pp.51-55.

[32] M. Mettler, "Blockchain technology in healthcare: The revolution starts here", In 2016 IEEE 18th international conference on e-health networking, applications and services (Healthcom) IEEE 2016, September, pp. 1-3.

[33] R. Das, "Top Five Digital Health Technologies in 2019", Available [Online]: https://www.forbes.com/sites/reenitadas/2019/02/04/the-top-five-digital-health-technologies-in-2019/#994702b6c0f9, Accessed: 08 February 2020.

[34] B. Feldman, E. M. Martin, and T. Skotnes, "Big data in healthcare hype and hope", Dr. Bonnie, 2012, 360, pp.122-125.

[35] W. Raghupathi, and V. Raghupathi, "Big data analytics in healthcare: promise and potential", Health information science and systems, 2014, 2(1), p.3.

[36] A. Abugabah, N. Nizamuddin, and A. Abuqabbeh, "A review of challenges and barriers implementing RFID technology in the Healthcare sector". ANT/EDI40, 2020.

[37] K.A.M Zeinab, and S.A.A. Elmustafa, "Internet of things applications, challenges, and related future technologies", World Scientific News, 2017, 2(67), pp.126-148.

[38] V. Shah, "9 Main Security Challenges for the Future of the Internet Of Things (IoT)", Available [online]: https://readwrite.com/2019/09/05/9-main-security-challenges-for-the-future-of-the-internet-of-things-iot/, Accessed 13 July 2020.

[39] M.U Farooq, M. Waseem, S. Mazhar, A. Khairi, and T. Kamal, "A review on internet of things (IoT)", International Journal of Computer Applications, 2015, 113(1), pp.1-7.

[40] Joyia, G.J., Liaqat, R.M., Farooq, A., and Rehman, S., 2017. Internet of Medical Things (IOMT): applications, benefits, and future challenges in the healthcare domain. J Commun, 12(4), pp.240-247.

High-throughput Virtual Screening Web Service Development for SARS-CoV-2 Drug Design

V. Miletić *, M. Ašenbrener Katić * and Ž. Svedružić **
* University of Rijeka, Department of Informatics, Rijeka, Croatia
** University of Rijeka, Department of Biotechnology, Rijeka, Croatia
{vmiletic, masenbrener}@inf.uniri.hr, zeljko.svedruzic@biotech.uniri.hr

Abstract - **The available structures of viral proteins and RNA molecules related to SARS-CoV-2 are used to screen and design a new set of drugs using the commercial databases and molecular docking protocols. The selected molecules are then studied further using molecular dynamics. Based on our earlier experiences we can target proteases, enzymes in DNA and RNA metabolism, and protein—protein interactions. In this paper we describe the planned research and development efforts for efficient screening and design of new drugs. Prior to the screening campaign, we will develop new open-source computational infrastructure, with two major outcomes. A new database containing all commercially available small-molecule ligands will be developed. A docking server with a web-based user interface will be developed and interfaced with the compound database. The docking server will use the database for sourcing of the molecules for the high-throughput virtual screening. Our approach offers major advantages that can bypass the problems that have traditionally plagued the pharmaceutical industry: our protocols are faster, cheaper, versatile, and offer minimal risks. We are developing new drugs using commercial databases, which allows us to buy the lead compounds for affordable prices that can bypass expensive and slow organic synthesis protocols.**

Keywords - database modeling; high-throughput virtual screening; molecular dynamics; high performance computing

I. INTRODUCTION

Viral epidemics pose an ongoing threat to public health and will continue to pose an unknown risk both to the human well-being and robustness of the world economy. Even with all scientific advancement and improvements of private, national and global healthcare systems during the 20th and early 21st century, humankind still lacks the meaning of dealing with emergent viral strains. Both human immune systems and the global pharmaceutical industry have means to neutralize only the viral strains that previously emerged in the population, and even then, quarantine and various other population isolation systems might be the only solution. One such example is the SARS-CoV-1 virus, the precursor of the SARS-Cov-2 virus responsible for the current epidemic and global shutdown. As of today, there is still no vaccine or drug that affects the SARS-CoV-1 virus, a virus that caused the SARS epidemic that officially ended on June 5th, 2003.

Viral pandemics regularly occurred during the 21st century, the most notable ones being the SARS epidemic (2002-2003), the swine flu epidemic (2009-2010), Middle East respiratory syndrome outbreaks (2012-ongoing), and current COVID-19 (2019-ongoing) pandemic. Viruses will continue to mutate and new epidemics in the future are imminent and unavoidable. Furthermore, novel vaccines cannot be developed prior to specific viral mutation and vaccine development is a process that takes time. During that time, a part of the population will inevitably be affected by whichever viral strain emerges, which is the exact current situation with the COVID-19 pandemic. No matter whether vaccine development is successful or not, and how fast its development is, there will be a large number of patients that will be in dire need of an efficient and specific drug for their ailment. Therefore, viral drug discovery remains one of the most important fields of medicinal chemistry and molecular biology, both in the academic and industrial settings.

A novel coronavirus (SARS-CoV-2) has caused 8 242 999 confirmed cases and 445 535 confirmed deaths globally, according to the COVID-19 situation report from WHO on June 18th, 2020 [1], and the numbers are still growing. Similar to the SARS-CoV outbreak in 2002, SARS-CoV-2 causes severe respiratory problems – fever, coughing, difficulties in breathing and shortage of breath are common symptoms. Human-to-human transmission occurs rapidly, and even human-to-feline contact was observed. To treat infected patients and slow this pandemic down, rapid and efficient development of highly specific antiviral drugs is of the highest urgency [2].

SARS-CoV-2 binds to a specific enzyme found on the surface of human cells. The enzyme, named angiotensin-converting enzyme 2, or ACE2 for short, lines endothelial cells in the lungs' blood vessels. The enzyme happens to be the same target for SARS-CoV-1, the virus that caused SARS.

Like other kinds of coronaviruses, SARS-CoV-2 forms a series of "spike" proteins that protrude from its core, giving it the shape of the sun's corona. Those spike proteins latch onto ACE2, which allows the virus to fuse its membrane to the cell membrane. Once inside, SARS-CoV-2 acts like any other virus – it hijacks the human cell, forcing the cell to produce many copies of the virus, spurring an infection.

A team of researchers designed a Michael acceptor inhibitor N3 using computer-aided drug design (CADD),

which can specifically inhibit multiple CoV Mpros, including those from SARS-CoV and MERS-CoV [3] – [6]. They constructed a homology model of COVID-10 Mpro and used molecular docking to see if N3 could target this new CoV Mpro. A docking pose showed that it could fit inside the substrate-binding pocket. To assess the efficacy of N3 for COVID-19 Mpro, a kinetic analysis was performed. A progress curve showed that it is a time-dependent irreversible inhibitor of this enzyme. The shape of this curve supports the mechanism of two-step irreversible inactivation. The inhibitor first associates with COVID-19 Mpro with a dissociation constant Ki; then, a stable covalent bond is formed between N3 and Mpro (E-I). The evaluation of this time-dependent inhibition requires both the equilibrium-binding constant Ki and the inactivation rate constant for covalent bond formation k3. However, N3 exhibits very potent inhibition of COVID-19 Mpro, such that measurement of Ki and k3 proved difficult. When very rapid inactivation occurs, kinetic constant kobs/[I] was utilized to evaluate the inhibition as an approximation of the pseudo-second-order rate constant (k3/Ki) [3]. The value of kobs/[I] of N3 for COVID-19 Mpro was determined to be 11,300±880 M-1s-1, suggesting this Michael acceptor has potent inhibition.

To elucidate the inhibitory mechanism of this compound, they determined the crystal structure of COVID-19 Mpro in complex with N3 to 2.1-Å resolution [7]. The crystal structure is available in the PDB protein databank as 6LU7 entry [8] (shown on Fig. 1). All residues (residues 1–306) are visible in electron density maps. Each protomer is composed of three domains. Domains I (residues 8–101) and II (residues 102–184) have an antiparallel β-barrel structure. Domain III (residues 201–303) contains five α-helices arranged into a largely antiparallel globular cluster and is connected to domain II through a long loop region (residues 185–200). COVID-19 Mpro has a Cys–His catalytic dyad, and the substrate-binding site is located in a cleft between Domain I and II. These features are similar to those of other Mpros reported previously [4], [6], [9]–[10]. The electron density map shows that N3 binds in the substrate-binding pocket in an extended conformation (Fig. 2, Fig. 3), with the inhibitor backbone atoms forming an antiparallel sheet with residues 164–168 of the long strand 155-168 on one side, and with residues 189–191 of the loop linking domains II and III.

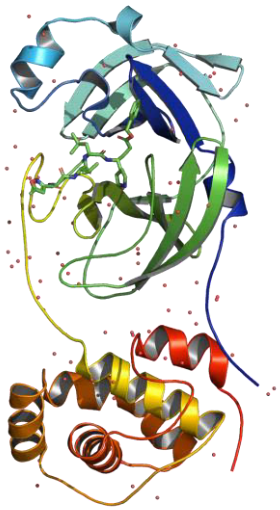


Figure 1. 6LU7 - The crystal structure of COVID-19 main protease in complex with an inhibitor N3 visualized using PyMOL [11]

II. OBJECTIVES

A. High-throughput virtual screening campaign of SARS-CoV-2 targets with commercially available compounds

N3 forms multiple hydrogen bonds with the main chain of the residues in the substrate-binding pocket, which also helps lock the inhibitor inside the substrate-binding pocket (Fig. 2).

The structure of SARC-CoV-2 Mpro in complex with N3 provides a model for rapidly identifying lead inhibitors to target SARS-CoV-2 Mpro through in silico screening. To achieve this, the existing and new molecular docking software will be combined to screen the compounds efficiently.

Virtual screening (VS) is a computational technique used in drug discovery to search libraries of ligands in order to identify those protein receptors which are most likely to bind to a drug target. VS can be followed by molecular dynamics, a computer simulation method for studying the movements of protein-ligand complexes in space over time. High-throughput virtual screening (HTVS) is a type of VS where the amount of dedicated computational resources is very high so many ligands can be checked at the same time.

B. Virtual screening of bioactive peptide and synthetic peptides against SARS-CoV-2 and ACE2 protein-protein interface

Recent crystallographic studies of the SARS-COV-2 receptor-binding domain (RBD) and full-lengths human ACE2 receptor revealed key amino acid residues at the contact interface between the two proteins and provide valuable structural information that can be leveraged for the development of disruptors specific for the SARS-CoV-2/ACE2 protein-protein interaction (PPI) [12], [13]. Small-molecule inhibitors are effective in binding to targeted binding sites, such as the aforementioned SARS-CoV-2 and N3 complex binding site but are less effective at disrupting extended protein binding interfaces [14]. Peptides, on the other hand, offer a synthetically

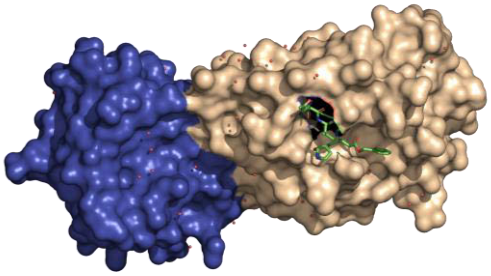


Figure 2. 6LU7 - The surface view of COVID-19 main protease in complex with an inhibitor N3 visualized using PyMOL [11]

accessible solution to disrupt protein-protein interactions by binding at interface regions containing multiple contact "hot spots" [15]. Analyzing the RBD-ACE2 co-crystal structure, we found that SARS-CoV-2-RBD/ACE2 interface spans a large elongated surface area, as is common for protein-protein interactions.

We hypothesize that disruption of the viral SARS-CoV-2-RBD and host ACE2 interaction with peptide-based binders will prevent virus entry into human cells, offering a novel opportunity for therapeutic intervention. To accomplish this, all proven bioactive peptides will be screened on the SARS-CoV-2-RBD domain using the aforementioned methodology, and the best peptides will serve as a model for the development of novel, synthetic peptides that will specifically target SARS-CoV-2-RBD surface. The peptides will then be evaluated using molecular dynamic simulations to observe specific polar and non-polar protein—protein interactions between the functional domains of SARS-CoV-2-RBD and ACE2.

C. *All-atom molecular dynamic simulations of single- and double-anti-viral therapy effect of SARS-CoV-2 binding to ACE2 target*

The co-crystal structure of SARS-CoV-2-RBD with ACE2 (PDB: 6M17, Fig 3.) will be used for the initial structure for molecular dynamic (MD) simulation for peptide binders development. Standard MD simulation procedures will be used during this work, particularly conformational exploration simulations of protein-protein docking, model refinement and testing of homology modeled SARS-CoV-2 proteins, addition and removal of ligands from HTVS campaign, mutation and modification of specific peptide chains, application of mechanical force to characterize strength of protein-protein interactions, and conformational changes of protein-protein supramolecular complex with various ligands and synthetic peptides designed to disrupt said complex.

The same MD simulations will be used for SARS-CoV-2 Mpro in complex with N3 compound (PDB: 6LU7) for the same purpose. Additionally, MD simulations of Mpro with screened ligands will be run, using the same research methodology. At the very end of this research work, MD simulations with both small-molecule inhibitors of Mpro and peptide binders on the SARS-CoV-2-RBD/ACE2 PPI region will be run to assess the effect of double anti-viral therapy.

III. RESEARCH AND DEVELOPMENT METHODOLOGY

A. *Design and development of the compound database*

The potential inhibitors, sourced from vendors of commercially available compounds, will be stored in a compound database. The existing databases were evaluated for the purpose: the freely accessible databases such as ChemSpider [16] by Royal Society of Chemistry and PubChem [17] by U.S. National Library of Medicine are not open source software that can be extended, interfaced, and changed according to the research requirements of the various projects and in particular the research and development work proposed here. Therefore, the compound database that satisfies the following requirements will be designed and developed:

- Able to hold millions of compounds
- Searchable according to name, structure, and molecular descriptors
- Able to be updated with new data from commercial databases

The compound database will be interfaced with the docking server described in the following to allow the easy screening of selected compounds.

B. *Design and development of the docking server*

The engine of the screening workflow that will be used is RxDock (formerly RiboDock(r) [18] and rDock [19]), a fast, versatile, and open-source program for docking ligands to proteins and nucleic acids developed by Vernalis R&D, University of York, University of Barcelona, RxTx, and others [20]. RxDock and its predecessor rDock were specifically designed for high-throughput virtual screening usage and are already used for HTVS of compounds that could bind to SARS-CoV-2 proteins, in particular by Galaxy [21], [22] and COVID.SI [23].

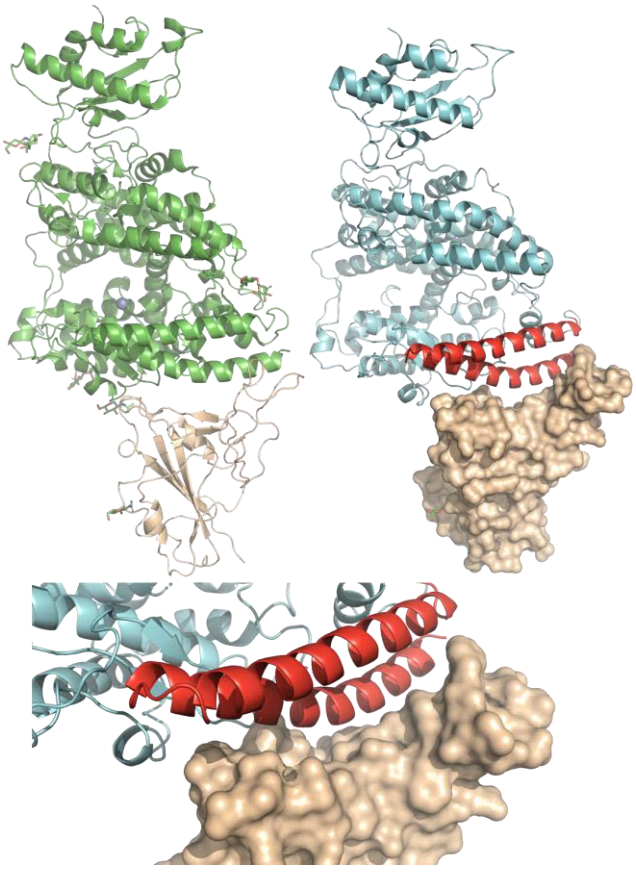


Figure 3. (Top left) SARS-CoV-2-RBD/ACE2 protease domain (PD) cryo-EM structure (PDB code: 6M17) in cartoon mode. ACE2 is colored in green, while SARS-CoV-2-RBD is colored in wheat color. (Top right) SARS-CoV-2-RBD/ACE2 protease domain (PD) cyro-EM structure. Selection of peptide fragments making key contacts colored in red. SARS-CoV-2 RBD is shown in surface mode, colored in wheat color. (Bottom) SARS-CoV-2-RBD/ACE2 key contact close-up. (All figures visualized using PyMOL [11])

Gorgulla et al. recently had a paper accepted in Nature [24] that described VirtualFlow, an approach to running HTVS campaigns on supercomputers managed by a batch system such as SLURM. Two types of tasks are performed by VirtualFlow: ligand preparation (checking the validity of the input files, a transformation of the input files to supported formats, cutting input files into pieces for multi-process execution, etc.) and virtual screening. The approach of VirtualFlow will be used for running HTVS campaigns with RxDock on Bura supercomputer. Instead of running jobs directly via scripts, a custom docking server with a web interface will be used. The interface will allow the user to specify a set of ligands (selected from the compound database) and the target site on the protein and will use a connection to Bura for scheduling docking jobs and gathering results after the jobs have finished executing for presenting them back to the user (Fig. 4).

The docking server and the compound database will be deployed at the University of Rijeka, Department of Informatics for usage by the researchers working on this problem and by the wider research community. Additionally, both the docking server and the compound database software developed in the scope of this work will be released under open-source licenses so they can be studied, improved, and deployed elsewhere by other researchers for other purposes.

C. *Drug repositioning*

Using newly developed infrastructure, a high-throughput virtual screening campaign of all FDA approved drugs will be performed. Crystal structures of SARS-CoV-2 Mpro structures (6LU7, 6YB7, 6M17, and any other relevant Mpro crystal structure or in silico model) be prepared accordingly before HTVS campaign itself. Hydrogens will be added, and structures will subsequently be minimized using the custom force field. The protonation state of His residues and the orientation of hydroxyl groups, Ans residues, and Gln residues will be optimized as well. Compounds from FDA approved drugs database will be protonated and 3D structure will be generated using the MMFF94s force field.

Pharmacophore approaches are one of the crucial tools in drug discovery [25]. IUPAC defines pharmacophore as "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response" [26]. A pharmacophore model can be constructed in two ways:

- ligand-based: by overlapping a set of active molecules and deriving prevailing chemical features that are essential for their bioactivity
- structure-based: by probing possible interaction points between the macromolecular target and ligands.

Structure-based pharmacophore modeling will be performed to obtain potential structures and a subset of the database matching the structure will be used for screening.

Once the protein structures are prepared, binding pockets will be defined and calculated using the cavity search engine of RxDock. Molecular docking itself will be carried out using the RxDock docking engine, considering the ligands as flexible but treating the receptor as a rigid structure. A post-docking minimization of the most promising docking results according to the docking score will be performed using the GROMACS software suite. The best candidates will be used in all-atom MD simulations of their disruption effects of SARS-CoV-2/ACE2 enzyme binding. The systems will be explicitly solvated to perform incrementally increasing MD simulation – starting simulations with initial binders will be evaluated using 20 ns all-atom simulations, with an increase to 50, 100 and 200 ns simulations in each passing run – using GROMACS on University of Rijeka's Bura supercomputing cluster.

D. *HTVS campaign of commercially available small-molecule ligands*

Using the same methodology as previously described, commercially available small-molecule ligands from all major and relevant commercial vendors will be evaluated in the HTVS campaign. Three specific subsets of ligands will be defined before HTVS campaign itself:

- Molecules with similar SAR properties of known SARS-CoV-2 ligands (such as N3)
- Molecules that comply with the Lipinski Rule of Five
    - Molecules that comply with the Lipinsky Rule of Three will be specially favored
- Molecules that comply with Ghose filter.

Each subset will be evaluated independently. The best candidates from each group will be used in all-atom MD simulations of their disruption effects of SARS-CoV-2/ACE2 enzyme binding using the same MD protocol explained previously.

E. *Peptide binder design*

Bioactive peptide structures and descriptors will be stored in newly developed database. For simulating these

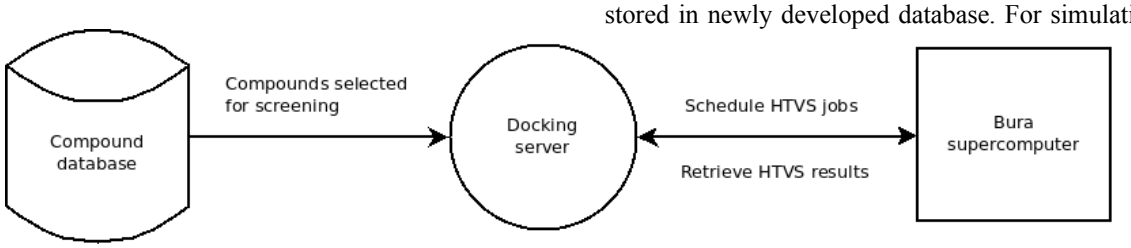


Figure 4. The interactions between the compound database, the docking server, and the Bura supercomputer

peptides inside a complex with SARS-CoV-2-RBD domain in GROMACS, usage of a99SB-disp and CHARMM36m force fields will be considered for their ability to model coil-to-structure transitions [27], [28]. Synthetic helical peptide sequence of spike-binding peptide 1 derived from the α1 helix of ACE2 peptidase domain (ACE2-PD) and synthetic peptide sequences similar to the bioactive peptides that performed best in MD simulations will also be evaluated (example MD simulation of protein—protein interactions from prior work.

IV. EXPECTED RESULTS

We are confident that we can perform full HTVS campaign of the major commercial databases on RNA-dependent RNA polymerase from SARS-CoV-2 virus, SARS-CoV-2 spike protein, protease of SARS-CoV-2 and other emerging targets an also develop the software infrastructure to make future endeavors of the same type easier. Our expected results out of HTVS campaign of FDA approved drugs and commercially available compounds (several million compounds) are couple of dozens of compounds with binding affinity lower of 1 μM.

Every molecule sequence that shows homology with the SARS-CoV-2 genes will be generated and the resulting model will be minimized using coarse-grain and all-atom MD protocols.

The effect of single- and double-anti-viral therapy will be evaluated. The effect of small-molecule ligands and peptide binding to SARS-CoV-2 targets and disruption of SAR-CoV-2/ACE2 complex (single-anti-viral therapy) and combined small-molecule ligand and peptide disruption of SARS-CoV-2/ACE2 complex will be ranked.

V. CONCLUSION AND FUTURE WORK

The authors of this paper are proposing a research and development of a novel computational infrastructure and research model for viral drug discovery and development. Reasons for proposing this research and development are: The first problem we want to address is the issue of chemical synthesis during the drug discovery phase. Chemical synthesis of novel compounds is a time-consuming process with a variable rate of success and during an ongoing epidemic/pandemic, it is a step that would best be avoided. Our solution for that is the development of a database that queries all major commercial compound databases for compounds that are available for purchase at the starting moment of a HTVS campaign.

The second issue we want to address is a lack of proper and well-documented open-source docking platforms that can easily be scaled on any hardware, from basic workstations to complex supercomputers in any of the major HPC centers, and deployed at any time by any research group without any licensing issues. We will be using RxDock as a novel docking platform specifically designed for HTVS campaigns and open-source any novel features we develop. In that way, any progress we make on developing novel drugs for the SAR-Cov-2 virus can be deployed in the future for any project by any research group on whatever (super)computer they have access to. The key component we will develop is a docking server built on top of RxDock, interfaced with compound database for sourcing of ligands and the Bura supercomputer for performing molecular docking. The web interface to the docking server will make the setup of large HTVS campaigns straightforward regardless of the device used by the user for accessing the interface.

The third and final problem we plan to solve is the discovery of novel SARS-CoV-2 drugs using infrastructure we develop. Since the main targets for viral infections are various proteases, we will utilize our extensive research experience in protease mechanism research [29]–[34] to simulate ligand-protein interactions of the most promising hit candidates from HTVS campaigns. The simulations will be run on local HPC Bura infrastructure. Insights gained during this research and development will be applied for all subsequent work with the infrastructure we develop ready to deploy at any time.

Since all parts of the proposed infrastructure are and are going to be open-source, resulting infrastructure will feature the following characteristics:

- reproducibility – all results generated on our software infrastructure will be easily repeated by any research group
- speed of setup – once developed, all software solutions will be easily deployed within a working day on any hardware by any research group or private company
- collaboration – multiple research groups will be able to collaborate using a single server setup.

Furthermore, the novel drug discovery infrastructure we develop is target-agnostic and will retain its full usefulness after this work is completed, and for any subsequent and inevitable viral epidemic. Some of the ongoing projects that will utilize this infrastructure include the development of DNA methyltransferase inhibitors, ubiquitin iso-peptidase inhibitors, and γ-secretase modulators.

REFERENCES

[1] World Health Organization, “Coronavirus disease 2019 (COVID-19) Situation Report - 150.” [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200618-covid-19-sitrep-150.pdf?sfvrsn=aa9fe9cf_2.

[2] G. Zhang, S. Pomplun, A. R. Loftis, A. Loas, and B. L. Pentelute, “The first-in-class peptide binder to the SARS-CoV-2 spike protein,” Bioengineering, preprint, Mar. 2020. doi: 10.1101/2020.03.19.999318.

[3] H. Yang et al., “Design of Wide-Spectrum Inhibitors Targeting Coronavirus Main Proteases,” PLoS Biol., vol. 3, no. 10, p. e324, Sep. 2005, doi: 10.1371/journal.pbio.0030324.

[4] X. Xue et al., “Structures of Two Coronavirus Main Proteases: Implications for Substrate Binding and Antiviral Drug Design,” J. Virol., vol. 82, no. 5, pp. 2515–2527, Mar. 2008, doi: 10.1128/JVI.02114-07.

[5] Z. Ren et al., “The newly emerged SARS-Like coronavirus HCoV-EMC also has an “Achilles’ heel”: current effective inhibitor targeting a 3C-like protease,” Protein Cell, vol. 4, no. 4, pp. 248–250, Apr. 2013, doi: 10.1007/s13238-013-2841-3.

[6] F. Wang et al., “Michael Acceptor-Based Peptidomimetic Inhibitor of Main Protease from Porcine Epidemic Diarrhea Virus,” J. Med. Chem., vol. 60, no. 7, pp. 3212–3216, Apr. 2017, doi: 10.1021/acs.jmedchem.7b00103.

[7] Z. Jin et al., “Structure of Mpro from COVID-19 virus and discovery of its inhibitors,” bioRxiv, p. 2020.02.26.964882, Mar. 2020, doi: 10.1101/2020.02.26.964882.

[8] R. P. D. Bank, “RCSB PDB - 6LU7: The crystal structure of COVID-19 main protease in complex with an inhibitor N3.” http://www.rcsb.org/structure/6LU7 (accessed Apr. 09, 2020).

[9] K. Anand, J. Ziebuhr, P. Wadhwani, J. R. Mesters, and R. Hilgenfeld, “Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs,” Science, vol. 300, no. 5626, pp. 1763–1767, Jun. 2003, doi: 10.1126/science.1085658.

[10] H. Yang et al., “The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor,” Proc. Natl. Acad. Sci., vol. 100, no. 23, pp. 13190–13195, Nov. 2003, doi: 10.1073/pnas.1835675100.

[11] W. L. DeLano. "Pymol: An open-source molecular graphics tool." CCP4 Newsletter on protein crystallography, vol. 40, no. 1, pp. 82–92, Mar. 2002. http://www.ccp4.ac.uk/newsletters/newsletter40.pdf (accessed Apr. 09, 2020).

[12] D. Wrapp et al., “Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation,” Science, vol. 367, no. 6483, pp. 1260–1263, Mar. 2020, doi: 10.1126/science.abb2507.

[13] R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, and Q. Zhou, “Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2,” Science, vol. 367, no. 6485, pp. 1444–1448, Mar. 2020, doi: 10.1126/science.abb2762.

[14] M. C. Smith and J. E. Gestwicki, “Features of protein-protein interactions that translate into potent inhibitors: topology, surface area and affinity,” Expert Rev. Mol. Med., vol. 14, p. e16, Jul. 2012, doi: 10.1017/erm.2012.10.

[15] K. Josephson, A. Ricardo, and J. W. Szostak, “mRNA display: from basic principles to macrocycle drug discovery,” Drug Discov. Today, vol. 19, no. 4, pp. 388–399, Apr. 2014, doi: 10.1016/j.drudis.2013.10.011.

[16] H. E. Pence and A. Williams, “ChemSpider: An Online Chemical Information Resource,” J. Chem. Educ., vol. 87, no. 11, pp. 1123–1124, Nov. 2010, doi: 10.1021/ed100697w.

[17] S. Kim et al., “PubChem Substance and Compound databases,” Nucleic Acids Res., vol. 44, no. D1, pp. D1202–D1213, Jan. 2016, doi: 10.1093/nar/gkv951.

[18] S. D. Morley and M. Afshar, “Validation of an empirical RNA-ligand scoring function for fast flexible docking using RiboDock®,” J. Comput. Aided Mol. Des., vol. 18, no. 3, pp. 189–208, Mar. 2004, doi: 10.1023/B:JCAM.0000035199.48747.1e.

[19] S. Ruiz-Carmona et al., “rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids,” PLOS Comput. Biol., vol. 10, no. 4, p. e1003571, Oct. 2014, doi: 10.1371/journal.pcbi.1003571.

[20] “RxDock,” rxdock.org. https://www.rxdock.org/ (accessed Apr. 16, 2020).

[21] Anton Nekrutenko et al., galaxyproject/SARS-CoV-2: second biorxiv release. Zenodo, 2020.

[22] “Cheminformatics | COVID-19 analysis on usegalaxy.” https://covid19.galaxyproject.org/cheminformatics/ (accessed Apr. 05, 2020).

[23] “Skupnostna znanost in boj proti koronavirusu | COVID.SI,” COVID.SI. https://covid.si/ (accessed Apr. 09, 2020).

[24] C. Gorgulla et al., “An open-source drug discovery platform enables ultra-large virtual screens,” Nature, pp. 1–8, Mar. 2020, doi: 10.1038/s41586-020-2117-z.

[25] S.-Y. Yang, “Pharmacophore modeling and applications in drug discovery: challenges and recent advances,” Drug Discov. Today, vol. 15, no. 11, pp. 444–450, Jun. 2010, doi: 10.1016/j.drudis.2010.03.013.

[26] C.-G. Wermuth, C. Robin Ganellin, P. Lindberg, and L. A. Mitscher, “Chapter 36 - Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1997),” in Annual Reports in Medicinal Chemistry, vol. 33, J. A. Bristol, Ed. Academic Press, 1998, pp. 385–395.

[27] P. Robustelli, S. Piana, and D. E. Shaw, “Developing a molecular dynamics force field for both folded and disordered protein states,” Proc. Natl. Acad. Sci., vol. 115, no. 21, pp. E4758–E4766, May 2018, doi: 10.1073/pnas.1800690115.

[28] J. Huang et al., “CHARMM36m: an improved force field for folded and intrinsically disordered proteins,” Nat. Methods, vol. 14, no. 1, Art. no. 1, Jan. 2017, doi: 10.1038/nmeth.4067.

[29] Ž. M. Svedružić, I. Odorčić, C. H. Chang, and D. Svedružić, “Substrate Channeling via a Transient Protein-Protein Complex: The case of D-Glyceraldehyde-3-Phosphate Dehydrogenase and L-Lactate Dehydrogenase,” bioRxiv, p. 2020.01.22.916023, Jan. 2020, doi: 10.1101/2020.01.22.916023.

[30] V. Miletić, I. Odorčić, P. Nikolić, and Ž. M. Svedružić, “In silico design of the first DNA-independent mechanism-based inhibitor of mammalian DNA methyltransferase Dnmt1,” PLOS ONE, vol. 12, no. 4, p. e0174410, Nov. 2017, doi: 10.1371/journal.pone.0174410.

[31] P. Nikolić, V. Miletić, I. Odorcić, and Ž. M. Svedružić, “Chapter 5 - In Silico Optimization of the First DNA-Independent Mechanism-Based Inhibitor of Mammalian DNA Methyltransferase DNMT1,” in Epi-Informatics, J. L. Medina-Franco, Ed. Boston: Academic Press, 2016, pp. 113–153.

[32] Z. M. Svedruzic, “Mammalian Cytosine DNA Methyltransferase Dnmt1: Enzymatic Mechanism, Novel Mechanism-Based Inhibitors, and RNA-directed DNA Methylation,” 2008. https://www.ingentaconnect.com/content/ben/cmc/2008/00000015/00000001/art00007 (accessed Apr. 13, 2020).

[33] Ž. M. Svedružić, K. Popović, and V. Šendula-Jengić, “Modulators of γ-Secretase Activity Can Facilitate the Toxic Side-Effects and Pathogenesis of Alzheimer’s Disease,” PLoS ONE, vol. 8, no. 1, Jan. 2013, doi: 10.1371/journal.pone.0050759.

[34] Ž. M. Svedružić, K. Popović, and V. Šendula-Jengić, “Decrease in catalytic capacity of γ-secretase can facilitate pathogenesis in sporadic and Familial Alzheimer’s disease,” Mol. Cell. Neurosci., vol. 67, pp. 55–65, Jul. 2015, doi: 10.1016/j.mcn.2015.06.002.

Drug screening with the Autodock Vina on a set of kinases without experimentally established structures

D. Tomić*, D. Davidović*, V. Janđel**, J. Mesarić***, K. Skala*, T. Lipić*
* Institut Ruđer Bošković / Centre for Informatics and Computing, Zagreb, Croatia
** KBC Zagreb / Department of Gynecological Surgery and Urology, Zagreb, Croatia
*** Faculty of electrical engineering and computing, Zagreb, Croatia

Abstract - **Virtual drug screening is one of the most widely used approaches for finding new drugs candidates. The process consists in selecting one or more chemical compounds with the highest binding free energy to target proteins. Given that the empirical space of chemical compounds is extremely large and estimated to has over 50 millions of them, finding the most effective drug is computationally challenging. Furthermore, the vast majority of proteins still lack the experimentally obtained 3D structures, making it hard to accurately calculate their binding free energies with chemical compounds. With this in mind, the aim of our study was to investigate the accuracy of the Autodock Vina program in a virtual drug screening on a set of proteins that do not have experimentally determined structures. To do this, we performed a virtual drug screening with the Autodock Vina on a large set of drug-kinase pairs taken from the IDG-Dream Drug-Kinase Binding Prediction Challenge. The results obtained show that the Autodock Vina can be used effectively in such unstructured environments.**

Keywords - Autodock Vina; virtual drug screening; drug discovery; binding free energy; unstructured proteins, kinase inhibitors

I. INTRODUCTION

Virtual drug screening is a computer technique often used in finding new drugs [1]. Virtual drug screening approaches, including ligand based ones, have been used in the discovery of Janus kinase 2 (JAK2) inhibitors. Developing JAK2 inhibitors has become a significant focus for small molecule drug discovery programs in recent years because the inhibition of JAK2 may be an effective approach for the treatment of polycythemia vera, primary myelofibrosis and essential thrombocythemia. Yang and colleagues reported a strategy using virtual screening approaches that can be a starting point for development of novel JAK2 inhibitors with high potency and selectivity [2]. A promising results were reported by Viegas and associates, emphasizing importance of virtual drug screening as a useful strategy to prospect active compounds, showing satisfactory pharmacokinetic and toxicological profiles They have presented novel anti Herpes simplex virus 1 (HSV-1) compounds which can be used for development of new drugs for the treatment of HSV-1 infections with different modes of action than acyclovir [3]. Results from a study by Lim et al. showed also encouraging results in the field of Hepatitis C virus (HCV) helicase inhibitors, revealing 10 potential drugs discovered via new docking optimization protocol with better docking accuracy [4]. Their findings could contribute to the discovery of novel HCV antivirals, serving as an alternative approach of in silico rational drug discovery.

In doing so, libraries [5] containing the structures of small-molecules with molecular weights less than 900 Daltons (in further text referred as compounds), are searched in order to identify those compounds that have the highest free energy of binding with target molecules. These target molecules are most commonly proteins and enzymes. While some compound libraries like PubChem already contain molecular structures [6], others are based on make-on-demand creation of compound structures, e.g. from 130 well-characterized chemical reactions [7]. However, in both cases, a fast and effective search and processing of libraries consisting of thousands and even tens of millions of compounds is required. This poses a significant challenge, both in terms of the required computing resources and the accuracy of the results obtained [8]. Although artificial neural networks coupled with deep-learning techniques promote the dramatic reduction of computing resources required for virtual drug screening [9], there is still a work to match their accuracy with a structure-based tools. The computation of the free binding energy between protein and compound is usually performed in three steps. In the first step protein and compound structures are prepared for docking and docking site on the protein surface is identified. As the next, docking between protein and compound is performed. Final step is scoring, where binding free energy between protein and compound is calculated [10]. There are several dozens of docking tools available, with various implementations of their scoring function [11]. The accuracy of docking and scoring varies from tool to tool, and is also conditioned by the structure of the protein and compound [12]. Higher accuracy in the computation of binding free energy can be achieved through the use of molecular dynamics (MD) computer programs [13]. However, due to the high demand on computational resources, their use in virtual drug screening on a large data sets is still prohibitive [14]. Another problem with structure-based virtual drug screening is that docking tools demand the existence of

well-defined 3D protein structures (in the further text referenced as structured proteins). RCSB portal [15] is usually searched for these structures. But, there are much more proteins without experimentally determined 3D structure (in the further text referenced as unstructured proteins), then structured proteins. During the preparation of this paper, and according to the statistics on RCSB portal, there was near 150.00 protein structures available. However, many of them are redundant, as they describe the same protein in complex with different compounds. Because of that, the total number of structured proteins is smaller and is currently near 10.000. This is by far less than the estimated number of different proteins in the human genome, which according to some sources equals to about 80.000 [16]. In order to perform virtual drug screening on unstructured proteins, one needs to predict their structures from their respective amino acid sequences [17]. Furthermore, some protein structures are complete as they have spatial coordinates defined for every atom in the protein, while others have coordinates for only a fraction of them. This is called the completeness. In addition, different structures for the same protein may have different resolution levels dependent on which method for obtaining the structure was applied (NRM spectroscopy, X-ray crystallography, X-ray free electron lasers, 3D electron microscopy). Besides, the number of structures describing proteins without compounds is relatively small compared to the number of structures describing proteins in complex with one or more compounds. And because compounds cause a deformations on protein surface, it is of the advantage to take structures without compounds. Therefore, for the best possible accuracy of virtual drug screening, one needs to take protein structures with highest completeness and resolution, and with the lowest number of compounds. Afterall, if compounds are present in the structure, one needs to remove them. The next step is to prepare the protein and compound structures. The most crystal structures on RCSB portal are missing residues [18], so they need to be added before docking and scoring. Furthermore, hydrogen atoms in protein and compound structures must be added, partial charges assigned, missing loops filled, and the protein structures minimized to relieve steric clashes [19] [20]. On the other hand, if we have unstructured proteins, it is advisable to predict their structures as accurately as possible. And once a virtual drug screening is launched, it should be performed without interruption and human intervention, until the very end. All relevant information should be recorded for the later analysis and possible corrections. A good example of how a virtual drug screening process should be conducted is found in [21].

II. MATERIALS AND MEHODS

We used the Vini in silico model of cancer [22] for testing the accuracy of Autodock Vina performing docking on large datasets with unstructured proteins. The main reason for this choice is the fact that the Vini model has already implemented virtual drug screening workflow, and the software necessarry for selecting the best protein structures, preparing them, and predicting the high-quality structures for unstructured proteins. The VINI model uses UCSF Chimera [23], MGLTools [24], Autodock Vina [25], and Open Babel [26] for virtual drug screening. In addition, it has its own modules to support virtual drug screening workflow. These are create_completeness_list module for selecting and retrieving the best RCSB protein structures, prepare_protein module for preparing protein structures for docking, and predict_protein_structure module for selecting and retrieving predicted protein structures from the SWISS-MODEL repository[27]. For the testing purposes, we used part of the experimentally obtained data from the IDG-Dream Drug-Kinase Binding Prediction Challenge [28], aimed at accurately predicting the binding intensity between protein kinase inhibitors [29] and target proteins. We obtained this information in the form of a list of 5329 pairs of protein kinase inhibitors and target proteins, with experimentally measured values of binding free energy expressed with pkd values, where pkd is defined as:

$$pkd = -1000log(Kd) \quad (1)$$

Kd in (1) is the dissociation constant, measuring the propensity of a larger object to separate (dissociate) reversibly into smaller components. The relation between the binding free energy ΔG and the dissociation constant is defined as [30]:

$$\Delta G = RTlog(Kd) \quad (2)$$

where R is gass constant and has a value 8.3145 J/mol·K, while T is the temperature at which the experiments were performed and is 293 Kelvin. By including Kd from (1) into (2), it follows:

$$\Delta G = - (pkd\ R\ T) / 1000 \quad (3)$$

Further on, we decided to use mean absolute error (MAE) instead of root mean square error (RMSE) in our analysis. MAE provides us with physically more understandible interpretation of data then it RMSE does. The advantage of using MAE over RMSE in evaluating the mean performance of the model is given in [31]. For the total of M pairs with structured proteins and N pairs with unstructured proteins, MAE can be expressed as

$$MAEe = \frac{1}{M}\sum_{m=1}^{M}|\Delta Ge(m) - \Delta Gc(m)| \quad (4)$$

for the experimentally obtained protein structures, and

$$MAEp = \frac{1}{N}\sum_{n=1}^{N}|\Delta Ge(n) - \Delta Gc(n)| \quad (5)$$

for the predicted protein structures, where ΔGe(m) and ΔGc(m) are measured and computed binding free energy values for pairs with structured proteins, and ΔGe(n) and

ΔGc(n) are measured and computed binding free energy for pairs with unstructured proteins. In the case of accurate prediction of all protein structures, MAEe equals MAEp, but due to the limited prediction accuracy of SWISS-MODEL, the following inequality applies:

$$MAEe < MAEp \quad (6)$$

From eqs. (4) and (5), and considering eq. (6), one can define the average quality of the N predicted structures as:

$$q(N) = 100\frac{MAEe}{MAEp} \text{ ; subject to } MAEe < MAEp \quad (7)$$

Defined in such a way, *q(N)* will have a value of 100 if all (N) protein structures are predicted accurately (MAEp = MAEe). Also, it will decrease with lower prediction accuracy (MAEp > MAEe), and has a value near zero if the prediction is bad (MAEp >> MAEe). Equation (7) was used to evaluate the accuracy of binding free energy calculations using predicted protein structures *versus* binding free energy calculations using experimentally established protein structures. The Vini model carried out virtual drug screening without human intervention from start to end, in three separate stages. First, protein and compound structures were prepared, then binding and binding free energy calculation for each pair was performed. Finally, MAEe and MAEp were computed. In addition to calculating the free binding energy for protein-compound pairs given by the IDG-Dream Drug-Kinase Binding Prediction Challenge, the Vini model also calculated the free binding energy for all other possible protein-compound pairs. In total, the Vini model calculated the binding energy for 34,748 protein-compound pairs using 96 processor cores over a 27-hour period.

The virtual drug screening process for each protein-compound pair is shown schematically in Figure 1.

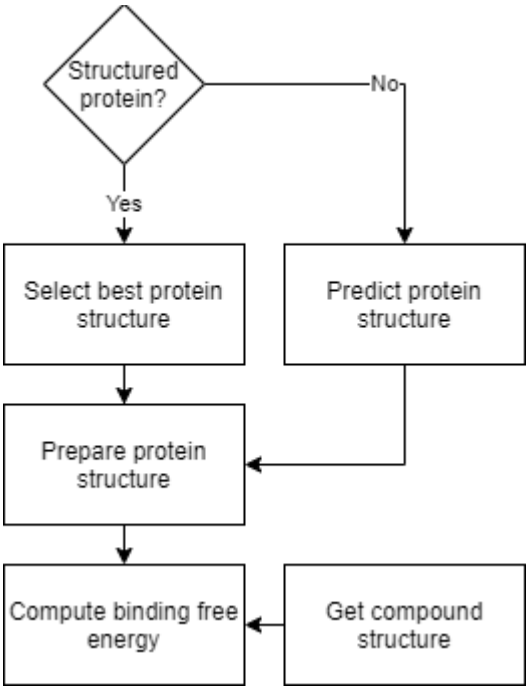


Figure 1. Overview of a virtual drug screening procedure for a particular protein-compound pair.

The Vini model repeated this procedure for each protein-compound pair, in two stages. In the first stage, it selected and prepared the experimentally determined protein structures. In case the protein structure did not exist, the model performed the structure prediction by using the SWISS-MODEL [27]. In the next stage, the Vini model performed binding between proteins and compounds, and the binding free energy calculations.

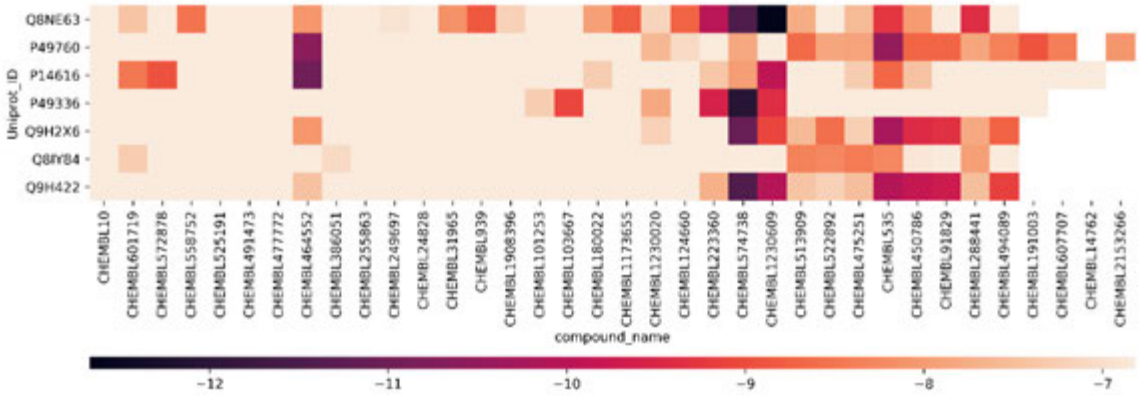


Figure 2. Heatmap of experimental value for failed protein-compound pairs

a) Preparation of protein structures

In this initial stage of virtual screening, the Vini model tried to select and prepare protein structures, and also 238 compound structures from the CHEMBL base [32] were prepared by converting their smiles to pdb format, and by adding hydrogen atoms to them. The Vini model identified 151 proteins from the Uniprot base [33] among 5329 protein-compound pairs. Among them, 136 were

structured, and 25 unstructured proteins. For unstructured proteins, the Vini model attempted to find and retrieve predicted structures in the SWISS-MODEL repository with the highest estimated quality using the QMEAN rating [34]. For five unstructured proteins, the Vini model could not find predicted structures, and they omitted from further analysis. In case of a structured proteins, the Vini model attempted to select and obtain from the RCSB database for each protein its 3D structure with the highest completeness. The next step was to remove water molecules and ligands from that structure, and then to add partial charges [35], hydrogen atoms to polar bonds [36], and missing residues. In doing so, the Vini model used the Dock Prep module of the UCSF Chimera software, the prepare_receptor4 module of the MGLTools software package, and the Open Babel software. If an error occurred during the preparation of the protein structure, the Vini model attempted to prepare the next structure with the next lower QMEAN. The process was repeated until some structure was sucesfully prepared for docking, or until the preparation of all available structures failed. If no structure could be prepared for docking, the Vini model additionally tried to prepare the predicted structure for docking. If this also failed, the VINI model dropped from further analysis all the protein-compound pairs containing that protein. For 231 from a total of 5329 protein-ligand docking pairs, the simulation could not be performed due to the limitation of used tools for seven protein structures with Uniprot_ID's P49760 (35), P14616 (35), P49336 (33), Q8NE63 (32), Q9H2X6 (32), Q9H422 (32) and Q8IY84 (32). Experimental values for failed pairs are presented in Figure 2. The average experimental (affinity) value for the 'failed' docking pairs is -7.439 kcal/mol with a standard deviation of 1.127, while the median is -6.817 with 148 pairs having the same experimental value.

b) Binding and binding free energy calculations

At this stage, the Vini model performed docking between proteins and compounds, and calculated free binding energies betwen them. The size of the simulation box was set to 40 Angstroms for all pairs. The spatial coordinates of the docking positions were calculated using the AutoGrid4 module from MGL Tools. Docking simulations and binding free energy calculations were performed with the Autodock Vina. In case the docking between protein and compound failed, the Vini model dropped that pair from further analysis.

Finally, the Vini model calculated the MAEe value of binding free energy for pairs with structured proteins by using eq. (4), and the MAEp value of free binding energies for pairs with unstructured proteins by using eq. (5).

III. RESULTS AND DISCUSSION

In total, 5329 protein-compound pairs from the Grand Challenge data set were analyzed. Binding free energy was calculated for 3979 pairs with structured proteins, for 1119 pairs with unstructured proteins, while the binding free energy of 231 pairs was not calculated, either as they contained proteins whose predicted structure was not found in the SWISS-MODEL repository, or the binding simulation failed. For pairs with structured proteins, the mean experimentally determined binding free energy was -7.612 kcal/mol and the mean calculated binding free energy was –7.210 kcal/mol. The MAEe value of 0.401 kcal/mol was calculated. For pairs with unstructured proteins, the mean of experimentally determined binding free energies was -7.580 kcal/mol and the mean of computed binding free energies was -6.990 kcal/ mol. MAEp of 0.590 kcal/mol was calculated. In Figure 3, we show for each protein-compund pair its corresponding MAE. In addition, in Figure 4 we show distributions of corresponding MAE for experimentally determined (E) and predicted protein structures (P) with median values of 0.897 and 0.931, respectively. The mean quality factor q(N) of 25 predicted protein structures calculated by eq. (7) was 67.9. This value is in a good agreement with the average IDDT test result [37] on the CAMEO portal [38] of 65.5 for 65 protein structures predicted by SWISS-MODEL.

Significant deviation of the computed binding free energy relative to the experimentally determined value has been obtained for multiple pairs, and the largest absolute error (8.70 kcal/mol) was identified for the pair between the Q15349 protein and the CHEMBL191003 compound,.

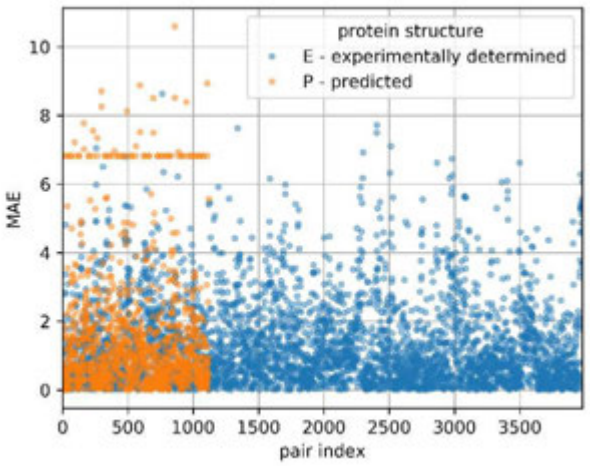


Figure 3. Mean absolute error distribution for pairs with experimentally determined (E) and predicted (P) protein structure

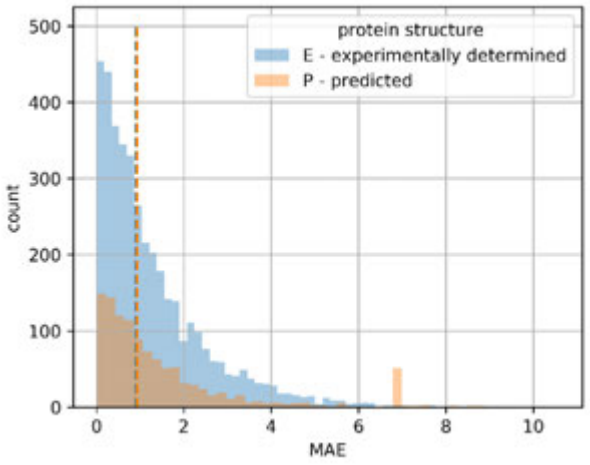


Figure 4 Distributions of mean absolute error for pairs with experimentally determined (E) and predicted (P) protein structure

Besides the virtual drug screening of 5329 pairs, we let the Vini model to perform simulation run on first 1000 pairs, and perform it 10 times. The rationale behind was to find out how well the results for the same pairs dissipate around a certain mean value. The results of this runs are shown in Figure *5*.

The relatively small MAE was found both for pairs with structured and unstructured proteins, thus indicating that the Autodock Vina is a good tool for initial virtual drug screening on a large data sets. On the other hand, large and relatively frequent absolute errors indicate that Autodock Vina gives (in some cases) a false estimate of the binding free energy. For this, it is advisable to re-check all drug candidates resulting from initial virtual drug screening on a large datasets with more accurate molecular MD tools. However, it should be kept in mind that MD tools put high demand on computer resources. Therefore, in the final screening, the number of drug candidates must be appropriate to the available computer resources. The q(N) value of 67.9 obtained in this experiment overhelmes the accuracy of the ten docking programs, including the Autodock Vina [39]. Therefore, it seems appropriate to use the predicted protein structures in the virtual drug screening process. On the other hand, their use by much more accurate MD tools seems still to be prohibitive and deserves the future research.

Regarding some very high absolute errors found in this experiment, one possible cause may be that many of the compounds in the experimental dataset were large molecules with a molecular mass having > 900 Daltons, and Autodock Vina is not built and optimized to work with a large ligands. Another possible cause of such major errors is that many compounds in the test datasheet are biotech drugs, and Autodock Vina was not trained on them. Notwithstanding the relatively large individual errors in the calculation of free binding energies, the test results show that the predicted structures contribute very little to reducing the accuracy of the results. Furthermore, the results obtained show that SWISS-MODEL predicts protein structures that may be used in docking programs.

The possible cause of relatively high absolute errors between experimental and computed values may also rest in 3D structures of compounds of the experimental dataset. Concretely, the IDG-Dream Drug-Kinase Binding Prediction Challenge delivered the description of the compounds in SMILES [40] data format. SMILES were converted to 3D structures with Open Babel and delivered to us. Based on our previous research and the experience, conversion from SMILES format to 3D structures by Open Babel is not the best solution and may introduce a certain amount of errors in obtained structures. However, this part is beyond the research scope of this paper. Based on the results obtained, our conclusion is that it is justified to use the predicted 3D protein structures in virtual drug screening of a large datasets containing unstructured proteins. We expect that with a more accurate 3D structures of compounds, the absolute error values will be considerably lower. This will be a part of our future research.

ACKNOWLEDGMENT

This research was partially supported by the European Regional Development Fund under the auspices of KK.01.1.01.0009 (DATACROSS) and the Ministry of Science and Education of the Republic of Croatia with the support of 533-19-15- 0007 (Centre of Research Excellence for Data Science and Cooperative Systems). All simulations in this research were performed on the supercomputer Bura, which was procured under the project "Development of research infrastructure at the University campus in Rijeka", co-funded by the European Regional Development Fund (ERDF).

REFERENCES

[1] Rester U (July 2008). "From virtuality to reality - Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective". Current Opinion in Drug Discovery & Development. 11 (4): 559–68. PMID 18600572.

[2] Yang M, Tao B, Chen C, et al. Machine Learning Models Based on Molecular Fingerprints and eXtreme Gradient Boosting Method Lead to the Discovery of JAK2 Inhibitors. J Chem Inf Model. 2019 Dec 23;59(12):5002-5012. doi: 10.1021/acs.jcim.9b00798.

[3] Viegas DJ, Edwards TG, Bloom DC, et al. Virtual screening identified compounds that bind to cyclin

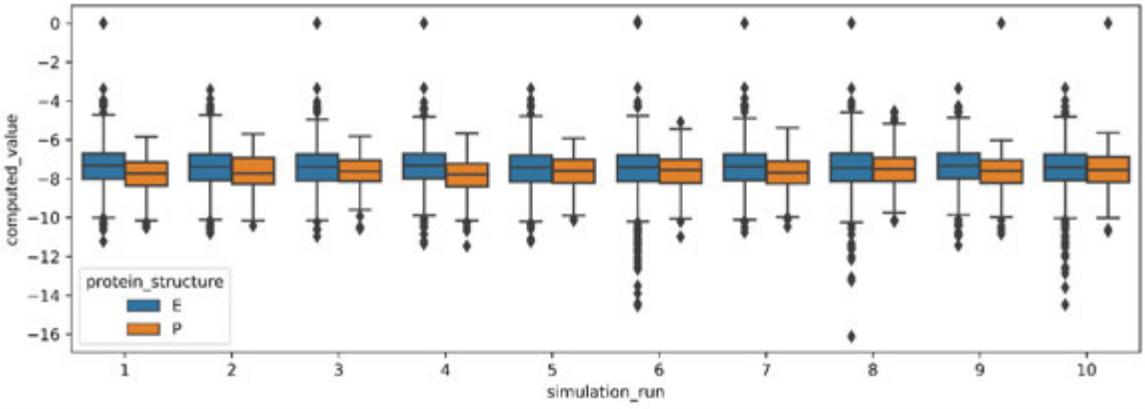


Figure 5. Box plot of computed affinity values for first 1000 protein-compound pairs across 10 simulation runs

dependent kinase 2 and prevent herpes simplex virus type 1 replication and reactivation in neurons. Antiviral Res. 2019; 172:104621 (ISSN: 1872-9096)

[4] Lim SK, Othman R, Yusof R, Heh CH. Rational Drug Discovery of HCV Helicase Inhibitor: Improved Docking Accuracy with Multiple Seeding in AutoDock Vina and In Situ Minimization. Curr Comput Aided Drug Des. 2017;13(2):160-169. doi: 10.2174/1573409912666161130122622

[5] Wishart DS. Chapter 3: Small molecules and disease. PLoS Comput Biol. 2012;8(12):e1002805. doi:10.1371/journal.pcbi.1002805

[6] Xie XQ. Exploiting PubChem for Virtual Screening. Expert Opin Drug Discov. 2010;5(12):1205–1220. doi:10.1517/17460441.2010.524924

[7] Lyu, J., Wang, S., Balius, T.E. et al. Ultra-large library docking for discovering new chemotypes. Nature 566, 224–229 (2019). https://doi.org/10.1038/s41586-019-0917-9

[8] E. Yuriev, Challenges and advances in structure-based virtual screening, Future Med Chem. 2014 Jan;6(1):5-7. doi: 10.4155/fmc.13.186

[9] Carpenter KA, Cohen DS, Jarrell JT, Huang X. Deep learning and virtual drug screening [published online ahead of print, 2018 Oct 5]. Future Med Chem. 2018;10(21):2557–2567. doi:10.4155/fmc-2018-0314

[10] Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov. 2004;3(11):935–949. doi:10.1038/nrd1549

11] Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. Biophys Rev. 2017;9(2):91–102. doi:10.1007/s12551-016-0247-1

[12] Castro-Alvarez A, Costa AM, Vilarrasa J. The Performance of Several Docking Programs at Reproducing Protein-Macrolide-Like Crystal Structures. Molecules. 2017;22(1):136. doi:10.3390/molecules22010136.

[13] Aldeghi M., Heifetz A., Bodkin MJ., et all. Accurate calculation of the absolute free energy of binding for drug molecules, Chemical Science, vol. 7, no. 1, Chemical Science, 2015, 207–18.

[14] Mobley DL, Gilson MK. Predicting Binding Free Energies: Frontiers and Benchmarks. Annu Rev Biophys. 2017;46:531–558. doi:10.1146/annurev-biophys-070816-033654.

[15] Rose PW, Prlić A, Altunkaya A, et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res. 2017;45(D1):D271–D281. doi:10.1093/nar/gkw1000

[16] Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760–1774. doi:10.1101/gr.135350.111

[17] Deng H, Jia Y, Zhang Y. Protein structure prediction. Int J Mod Phys B. 2018;32(18):1840009. doi:10.1142/S021797921840009X.

[18] Djinovic-Carugo K, Carugo O. Missing strings of residues in protein crystal structures. Intrinsically Disord Proteins. 2015;3(1):e1095697. Published 2015 Oct 23. doi:10.1080/21690707.2015.1095697.

[19] Lionta E, Spyrou G, Vassilatis DK, Cournia Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. Curr Top Med Chem.2014;14(16):1923–1938. doi:10.2174/1568026614666140929124445

[20] Sastry GM, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. J Comput Aided Mol Des. 2013;27(3):221–234. doi:10.1007/s10822-013-9644-8

[21] Forli S, Huey R, Pique ME, Sanner MF, Goodsell DS, Olson AJ. Computational protein-ligand docking and virtual drug screening with the AutoDock suite. Nat Protoc. 2016;11(5):905–919. doi:10.1038/nprot.2016.051.

[22] Tomic D, Skala K, Kranjcevic L, Pirkic B, Stifter S, et al. (2018) Evaluation of the Efficacy of Cancer Drugs by Using the Second Largest Eigenvalue of Metabolic Cancer Pathways. J Comput Sci Syst Biol 11: 240-248.

[23] Pettersen EF, Goddard TD, Huang CC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605–1612. doi:10.1002/jcc.20084.

[24] G. Morris, R. Huey, W. Lindstrom, M. Sanner, R. Belew, D. Goodsell, and A. Olson. 2009. Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity. Journal of Computational Chemistry. 30(16): 2785-91.

[25] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455–461. doi:10.1002/jcc.21334.

[26] O'Boyle, N.M., Banck, M., James, C.A. et al. Open Babel: An open chemical toolbox. J Cheminform 3, 33 (2011). https://doi.org/10.1186/1758-2946-3-33.

[27] Bienert, S., Waterhouse, A., de Beer, T.A.P., Tauriello, G., Studer, G., Bordoli, L., Schwede, T. The SWISS-MODEL Repository - new features and functionality. Nucleic Acids Res. 45, D313-D319 (2017).

[28] https://www.synapse.org/#!Synapse:syn15667962/wiki/583305

[29] Kannaiyan R, Mahadevan D., A comprehensive review of protein kinase inhibitors for cancer therapy, Expert Rev Anticancer Ther. 2018 Dec;18(12):1249-1270. doi: 10.1080/14737140.2018.1527688.

[30] Du X, Li Y, Xia YL, et al. Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. Int J Mol Sci. 2016;17(2):144. Published 2016 Jan 26. doi:10.3390/ijms17020144

[31] C.J. Willmott, K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Climate Research Vol. 30, No. 1 (December 19 2005), pp. 79-82

[32] Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012;40(Database issue):D1100–D1107. doi:10.1093/nar/gkr777

[33] UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015;43(Database issue):D204–D212. doi:10.1093/nar/gku989

[34] Benkert P, Biasini M, Schwede T., Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 27, 343-350 (2011), doi:10.1093/bioinformatics/btq662

[35] Kramer C, Spinn A, Liedl KR. Charge Anisotropy: Where Atomic Multipoles Matter Most. J Chem Theory Comput. 2014;10(10):4488–4496. doi:10.1021/ct5005565

[36] Jensen W.B. The Origin of the "Delta" Symbol for Fractional Charges. J. Chem. Educ. 2009, 86, 5, 545, doi:10.1021/ed086p545

[37] Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013;29(21):2722–2728. doi:10.1093/bioinformatics/btt473

[38] Haas J, Roth S, Arnold K, et al. The Protein Model Portal--a comprehensive resource for protein structure and model information. Database (Oxford). 2013;2013:bat031. Published 2013 Apr 26. doi:10.1093/database/bat031

[39] Wang Z, Sun H, Yao X, et al. Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. Phys Chem Chem Phys. 2016;18(18):12964–12975. doi:10.1039/c6cp01555g

[40] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of chemical information and computer sciences, 1988, 28, 1, 31-36.

[

CTI

International Conference on

TELECOMMUNICATIONS & INFORMATION

Steering Committee

Chairs: Željka Car, University of Zagreb, Croatia
Stjepan Golubić, Zagreb, Croatia
Dragan Jevtić, University of Zagreb, Croatia
Branko Mikac, University of Zagreb, Croatia

Members: Slaviša Aleksić, Vienna University of Technology, Austria
Krešo Antonović, Ministry of Sea Transport and Infrastructure, Zagreb, Croatia
Gordan Gledec, University of Zagreb, Croatia
Darko Huljenić, Ericsson Nikola Tesla, Zagreb, Croatia
Oliver Jukić, Virovitica College, Croatia
Ozren Jureković, VERN University, Zagreb, Croatia
Erich Leitgeb, Graz University of Technology, Austria
Velimir Švedek, Croatian Regulatory Authority for Network Industries, Zagreb, Croatia

Program Committee

Members: Željka Car, University of Zagreb, Croatia
Vlado Delić, University of Novi Sad, Serbia
Saša Dešić, Ericsson Nikola Tesla, Zagreb, Croatia
Renato Filjar, Zagreb University of Applied Sciences, Croatia
Tihana Galinac Grbac, Juraj Dobrila University of Pula, Croatia
Stjepan Golubić, Zagreb, Croatia
Vlasta Hudek, Rezultanta, Zagreb, Croatia
Robert Inkret, Ericsson Nikola Tesla, Zagreb, Croatia
Dragan Jevtić, University of Zagreb, Croatia,

Admela Jukan, Institute of Computer and Network Engineering, Braunschweig, Germany
Marko Lacković, IBM Croatia, Zagreb, Croatia
Dražen Lučić, Croatian Chamber of Economy, Zagreb, Croatia
Branko Mikac, University of Zagreb, Croatia
Mladen Sokele, Zagreb University of Applied Sciences, Croatia
Dina Šimunić, University of Zagreb, Croatia
Darko Švelec, University of Zagreb, Croatia
Miroslav Vrankić, E-Glas d.o.o., Rijeka, Croatia
Antonio Teixeira, University of Aveiro, Portugal
Mario Weber, Croatian Regulatory Authority for Network Industries, Zagreb, Croatia

INVITED PAPERS

Cyber Security Risk is a Board-Level Issue

Biljana Cerin
Ostendo Consulting, Zagreb, Croatia
biljana.cerin@ostendogroup.com

Abstract - This presentation will give an overview of the contents that should be presented to the Board, ways to gather and structure the information on cyber security risks exposure, and methods of presenting it in relatively short time given at the Board meetings, in a way that is understandable to the Board members, in order for them to make informed decisions in alignment with their Board responsibilities. It will also examine the prerequisites for having such a successful cyber risk management system that can deliver the requested outcome at any moment in time - a clear picture of the cyber security risks exposure and expected trends, for all the relevant stakeholders.

For years have the cyber security professionals struggled to bring up to the management's attention the importance of proper and timely dealing with the cyber security risks. They were busy overcoming obstacles such as problems with understanding of the business versus technology terminology, understanding of the exact nature of the cyber security risks by the non-infosec people, explaining to business what may be the consequences of having the cyber security risks materialize, and what appropriate actions the responsible business owners should take in order to mitigate these risks. They have established the understanding, with a help of compliance requirements and of the cyber security, critical network and infrastructure and personal data protection regulations, although the most responsible companies have never had to wait for the laws and regulations in order to set their risk management systems properly.

Based on the interviews conducted with 200 CISOs in UK and USA in 2019 [1], the following findings illustrate an example of the progress made in having the cyber security positioned on the top of their organizations' agendas:

- 96% either slightly or strongly agreed that senior executives have a better understanding of cyber security than they did five years ago.
- 67% said their businesses prioritize cyber security above all other business considerations.
- 76% indicated that cybersecurity risk has become important enough to businesses that CISOs will begin to be named as CEOs.

The professional community has done a good job, however, it is time to move forward, or one level up – cyber security now more than ever needs to be seen not only as the regular agenda point on management's meetings, but also discussed as an important Board-level issue too. And the CISOs – Chief Information Security Officers, need to be prepared to step up once they are called to the Board meeting.

Taking into account that the average cost of a data breach today is $3.9 million inclusive of legal fees, fines, lost productivity, crisis response efforts, remediation, and that the loss of intellectual property, competitive insights, or consumer trust can often be the greatest source of long-term damage in the wake of a data breach [2], and having in mind that the Board owns the strategy of an organization, it is of no question that cyber security related risks need to be clearly understood by the Board members.

As defined by the US National Association of Corporate Directors (NACD), the following should be considered as the cyber risk oversight principles [3] the responsible Boards should be driven by:

1. Boards need to understand and approach cybersecurity as an enterprise-wide risk management issue, not just an IT issue.
2. Boards should understand the legal implications of cyber risks as they relate to their company's specific circumstances.
3. Boards should have adequate access to cybersecurity expertise, and discussions about cyber-risk management should be given regular and adequate time on board meeting agendas.
4. Board directors should set the expectation that management will establish an enterprise-wide cyber-risk management framework with adequate staffing and budget.
5. Board-management discussion about cyber risk should include identification of which risks to avoid, which to accept, and which to mitigate or transfer through insurance, as well as specific plans associated with each approach.

However, the issues lay within the current capabilities of both the Board members and the CISOs to meet these expectations.

Typical Board members are former business executives who have led a variety of operational functions and have an orientation toward financial metrics; for them, cyber security risk and its technical concepts and vocabulary often feel foreign; their average age is 63, and most probably they did not encounter cyber security risk during the course of their careers [4].

On the other side, typical CISOs come from an IT or audit background, often do not think in the Board-level terms, have a primarily tactical focus, and are often struggling to secure the organization from cyber threats and stay in compliance with the frameworks and complex regulatory requirements. CISOs also often do not have all the necessary information about the strategic initiatives Board is concerned with, and most importantly, they do not communicate the issues with the Board directly, but through the CEO or executive management that are usually present at the Board meetings.

The question arises, how to close these gaps? By identifying the common language both the Board and the CISO understand, and that is the risk management language. This includes clearly setting the risk appetite, establishing objective risk metrics, monitoring the key risk indicators, and setting the adequate risk management strategy and required actions that the CISO can execute through the cyber security program, that is understood and supported by both the management and the Board.

The Board will typically want to know what are the top cyber security risks for the organization, how efficient is the cyber security program in addressing them, is the organization equipped with the right skills and technologies to defend against evolving threats, have there been any significant incidents or data breaches, how efficient is the security spending, and what is the state of compliance with the cyber security related regulations. They would also like to be informed about the cyber security risks associated with any important business initiatives, with the supply chain, and these associated with eventual mergers and acquisitions.

In getting ready for the meeting with the Board, the CISO should address all these concerns in the format relevant to the audience and easy to understand. It is not the responsibility of the Board to become IT experts, but the Board must know what questions to ask the IT departments. In addition, boards must provide the leadership and the commitment necessary – by proactively overseeing and holding management and the C-suite responsible - to make protecting the organization from cyber-attack a priority [3]. The top cyber security risks the organization is facing and its potential impacts, whenever possible, should be expressed in the financial terms the Board usually understands the best.

CISO should also use this opportunity to efficiently present the major cyber security improvement projects and challenges, the current state of the cyber security program, its influence, and its efficiency, using objective metrics wherever possible.

Although the research conducted in February 2020 [5] found that the Board does take cyber security seriously, with 47 percent saying that cyber security is a “great” concern to them, and that they are more likely than CISOs to think that cyber threats are a “high” or “very high” risk to their business (90% vs 66%), and are also aware of the high-pressure nature of the CISO’s job, with 74 percent saying they believe their security team to be moderately or tremendously stressed, many still hold the CISO responsible for a breach and expect them to deliver more value to the business.

By following the approach presented herein, a deeper Board’s understanding of the CISO’s work, and a deeper Board’s involvement with informed insight of what it takes to deal with the cyber security risks, will be achieved. This will enable much stronger management support at all levels in the organization for the cyber security initiatives, will reduce the organization’s exposure to the cyber security threats, and will make the CISOs’ lives much less stressful.

REFERENCES

[1] https://www.helpnetsecurity.com/2019/09/24/ciso-role/

[2] https://www.ibm.com/security/data-breach

[3] https://www.oas.org/en/sms/cicte/docs/ENG-Cyber-Risk-Oversight-Handbook-for-Corporate-Boards.pdf

[4] https://insights.diligent.com/cyber-risk

[5] https://www.securitymagazine.com/articles/91652-new-survey-reveals-ciso-stress-and-the-toll-it-takes

Govorna komunikacija čovjek-stroj X & Y vs. Z & α generacije

Vlado Delić
Fakultet tehničkih nauka Univerziteta u Novom Sadu
Novi Sad, Srbija
vlado.delic@uns.ac.rs

Abstract - Human-machine speech communication is an area of artificial intelligence based on machine learning. Recent advances in automatic speech recognition and text-to-speech synthesis based on deep neural networks have enabled dialog systems in a number of languages, including South Slavic. The experience of touchscreen and speech communication with smart devices are something new generations (Z & α) grow up with. It is much easier and more natural for Gen Z to engage in speech communication with devices in their environment than previous generations (X & Y-Millennials), so it is very important that they can do so in their own language. Speech technologies, unlike other new technologies, are language dependent and must be developed extensively for each language. The lecture will present the achievements in the development of speech technologies for South Slavic languages and will look at perspectives of human-machine speech communication in the near and far future, especially for different generations of people who grow up with different experiences in communication with people and things.

Današnji tinejdžeri (rođeni na prijelazu u ovo tisućljeće, Gen Z) uvelike koriste mogućnost diktiranja poruke ili razmjene govorne ili video poruke. Djeca koja sad tek kreću u školu (Gen α) su, kad su prohodali, pokušavali koristiti ekran televizora kao zaslon na dodir jer su to već radili na tabletima. A najmlađi, koji tek progovaraju, odrastaju uz govorne asistente i imaju rano iskustvo govorne komunikacije s tabletima i mobilnim telefonima: nastoje da svojim nemuštim jezikom izazovu odgovor ljudskim glasom iz tih pametnih uređaja; taj način komunikacije s pametnim uređajima, multimedijskim igračkama i knjigama im se urezuje u sinapse – iskustvo koje nisu imale prethodne generacije. Što možemo očekivati od ove i narednih generacija mladih? Oni neće imati samo društvene mreže nego i mreže svojih internet stvari (*Personal Network of Internet of Things* – P@IoT ili MyIoT): npr. automobil, razni uređaji u pametnoj kući, senzori na tijelu, osobna računala i osobni asistenti, digitalni novčanik-račun u banci itd. Primjeri komunikacije čovjek-stroj u vlastitom mreži stvari MyIoT su da prije ulaska u vozilo čovjek traži da se vozilo uključi kako bi se ugrijalo, i da ugrije sjedalo, ili kad čovjek traži od hladnjaka da provjeri što od namirnica nedostaje za ručak/večeru te da onda zamoli kućnog robota da to naruči preko interneta uz komunikaciju s odgovarajućim računom u banci... Prethodna generacija, Gen Y (rođeni 80-tih i 90-tih) odrasta u digitalno doba uz društvene mreže i drugačije poimanje posla i autoriteta od Gen X iz 60-tih i 70-tih koja je rasla još u doba analogne tehnike, i za njih je prijelaz sa analogne telefonije i analogne televizije na digitalnu bio krupan korak. Ovi mlađi (Gen Z & α) ne mogu ni zamisliti svijet bez interneta i mobilne telefonije. Dakle, nove generacije će imati ne samo društvene mreže za komunikaciju i poslovanje s drugim ljudima već i svoje mreže interneta stvari s kojima će pričati. Postavlja se pitanje: Na kojem jeziku? Da bi to bili u mogućnosti na svom materinjem jeziku, nužno je da ustrajemo u razvoju govornih tehnologija za svoje jezike.

Na ovom predavanju će biti analiziran napredak u razvoju govornih tehnologija na osnovu nove paradigme dubinskog učenja [1], koja je omogućila pojavu digitalnih govornih asistenata na mnogim jezicima, uključujući i južno-slavenske. Bit će predstavljeni govorni i jezični resursi koji su razvijeni za srpski i hrvatski jezik i njihove dijalekte u državama zapadnog Balkana, a koji su neophodni za daljnji razvoj multimodalnih dijaloških sustava. Uz to će biti opisane prve aplikacije govornih tehnologija koje su našle primjenu u ovom području. Na kraju su analizirani trendovi razvoja i primjene u ovom području računajući na brze promjene i iskustva s kojim odrastaju nove generacije u ovom tisućljeću.

REFERENCES

[1] Vlado Delić et al: "Speech technology progress based on new machine learning paradigm", *Computational Intelligence and Neuroscience*, Special Issue "Advanced Signal Processing and Adaptive Learning Methods", Hindawi, June 2019. DOI: 10.1155/2019/4368036 (Review Article)

[2] LeCun Y., Bengio Y., Hinton G., "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015. DOI: 10.1038/nature14539.

[3] Deng L., Hinton G.E., Kingsbury B., "New types of deep neural network learning for speech recognition and related applications: an overview," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013, pp. 8599-8603.

[4] Graves A., Jaitly N., "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of The 31st International Conference on Machine Learning*, Beijing, China, 2014, pp. 1764-1772.

[5] Nagamine T., Seltzer M.L., Mesgarani N., "On the role of nonlinear transformations in deep neural network acoustic models," in *Proceedings of Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, San Francisco, CA, USA, 2016, pp. 803-807.

[6] Kalchbrenner N., Grefenstette E., Blunsom P., "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, MD, USA, 2014, pp. 655-665.

[7] Mišković D. et al., "Hybrid methodological approach to context-dependent speech recognition," *International Journal of Advanced Robotic Systems*, vol. 14, no. 1, 2017. DOI: 10.1177/1729881416687131.

[8] Popovic B., Pakoci E., Pekar D., "End-to-end large vocabulary speech recognition for the Serbian language," in *Proceedings of 19th International Conference Speech and Computer (SPECOM)*, Hatfield, UK, 2017, pp. 343-352. DOI: 10.1007/978-3-319-66429-3_33.

[9] Pakoci E., Popović B., Pekar D., "Improvements in Serbian speech recognition using sequence-trained deep neural networks," *SPIIRAS Proceedings*, vol. 3, no. 58, pp. 53-76, 2018.

[10] Swain M., Routray A., Kabisatpathy P., "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93-120, 2018. DOI: 10.1007/s10772-018-9491-z.

[11] Han K., Yu D., Tashev I., "Speech emotion recognition using deep neural network and extreme learning machine," in *Proceedings of Interspeech, 15th Annual Conference of the International Speech Communication Association*, Singapore, 2014, pp. 223-227.

[12] Fayek H.M., Lech M., Cavedon L., "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60-68, Aug. 2017. DOI: 10.1016/j.neunet.2017.02.013.

[13] Deng J., Xu X., Zhang Z., Frühholz S., Schuller B., "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 31-43, 2018. DOI: 10.1109/TASLP.2017.2759338.

[14] Delic V., Bojanic M., Gnjatovic M., Secujski M., Jovicic S., "Discrimination capability of prosodic and spectral features for emotional speech recognition," *Elektronika ir Elektrotechnika*, vol. 18, no. 9, pp. 51-54, 2012. DOI: 10.5755/j01.eee.18.9.2806.

[15] Wu Z., Watts O., King S., "Merlin: an open source neural network speech synthesis system," in *9th ISCA Speech Synthesis Workshop*, Sunnyvale, CA, USA, 2016, pp. 202-207. DOI: 10.21437/SSW.2016-33.

[16] Wu Z., King S., "Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1255-1265, 2016. DOI: 10.1109/TASLP.2016.2551865.

[17] Lorenzo-Trueba J. et al., "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," *Speech Communication*, vol. 99, pp. 135-143, 2018. DOI: 10.1016/j.specom.2018.03.002.

[18] Siniša Suzić, Tijana Delić, Darko Pekar, Vlado Delić, Milan Sečujski "Style Transplantation in Neural Network Based Speech Synthesis", Acta Polytechnica Hungarica, Vol. 16, No. 6, Sep. 2019, pp. 171-189, DOI: 10.12700/APH.16.6.2019.6.11

[19] Ping W. et al., "Deep Voice 3: scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2018.

[20] Skerry-Ryan R. et al., "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, 2018, pp. 4693-4702.

[21] Chen H., Liu X., Yin D., Tang J., "A survey on dialogue systems: recent advances and new frontiers," *SIGKDD Explorations*, vol. 19, no. 2, pp. 25-35, 2017. DOI: 10.1145/3166054.3166058.

[22] Bengio Y., Ducharme R., Vincent P., Janvin C., "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, no. 6, pp. 1137-1155, 2003.

[23] Serban I.V., Lowe R., Henderson P., Charlin L., Pineau J., "A survey of available corpora for building data-driven dialogue systems," *arXiv preprint arXiv:1512.05742*, 2015.

[24] Xing C., Wu Y., Wu W., Huang Y., Zhou M., "Hierarchical recurrent attention network for response generation," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, 2018, pp. 5610-5617.

PAPERS

TOWARDS 5G NETWORK

LTE and 5G NR – Coexistence and Collaboration

Teodor B. Iliev*, Grigor Y. Mihaylov **, Ivaylo S.Stoyanov *** and Elena P. Ivanova*
* University of Ruse/Department of Telecommunication, Ruse, Bulgaria
** University of Telecommunications and Post/Department of Telecommunication, Sofia, Bulgaria
*** University of Ruse, Department of Electrical and Electronic Engineering, Ruse, Bulgaria
e-mail address: tiliev@uni-ruse.bg, gregmihaylov@gmail.com, stoyanov@uni-ruse.bg, epivanova@uni-ruse.bg

***Abstract* - The initial introduction of new generation of mobile communications technologies is typically done in high traffic density areas and with high requirements for new service capabilities. This is followed by gradual further construction, which may be more or less rapid depending on the strategy of the operator. During this, a combination of new and legacy technologies will provide ubiquitous coverage of the operator network, with devices constantly moving out and out of the areas covered by the new technology. Therefore, seamless transmission between new and outdated technology is a key requirement.**

Keywords - Fifth-Generation (5G), New Radio network, Mobile communication systems

I. INTRODUCTION

During the recent years, there has been a tremendous surge in the demand for mobile and wireless connectivity, which is forecasted to grow exponentially within the foreseeable future. Over the last 40 years, the world has witnessed four generations of mobile communication. However, the Fifth-Generation (5G) mobile communication system [1] will not only have to cope with an increasing demand of traffic volume, but also provide a broader range of applications with new and not yet supported requirements in terms of reliability, availability and efficiency. The main goal for communication technologies have always been to enable people to talk to each other more flexible and convenient. Now that this goal seems to be achieved and the foundation is set (product acceptance and convenient service coverage for mobile devices) new goals have been set for the next generation 5G. The upcoming next generation mobile network offers much higher data rates in terms of Gigabits per second for every user connected to the 5G NR (New Radio) network. [2, 3, 4]

5G wireless access is provided to enable the network community, where information can be accessed and shared anywhere, anytime, by anyone and anything. To enable a truly networked society, there are three major challenges:

- Huge growth in the number of connected devices.
- Huge growth in traffic volume.
- Wider range of applications with different requirements and characteristics.

To meet these challenges, 5G wireless access requires not only new features, but also significantly more spectrum and wider bandwidths.

Most users will migrate to new devices, which support the latest technologies within a few years. For example, let's take a look at South Korea – it is, maybe, the most technological-minded country, with about 8 million 5G subscribers. The Korean government continues to improve measures to accelerate 5G adoption through direct and indirect measures. Shortly after Korea premiered 5G commercial services, it reached 2 million subscribers in just four months, faster than what was anticipated by mobile operators and outperforming previous 4G launches. During its initial rollout phase, the first 5G service defaults to 4G, even in an area where 5G is available since most of these 5G networks are built on top of 4G LTE. When 5G tasks like HD gaming, AR or VR are initiated, they require upswitching. This slows speeds at the beginning of such tasks, and can also affect battery life. Upswitching is happening in South Korea and it will happen in all future 5G regions until standalone 5G networks are running.

It is not just humans who will benefit from 5G, machines are also about to get smarter. 5G's connectivity is the game changer that could make self-driving cars more prevalent. The 4G wireless technology does not have enough capacity or responsiveness to support communication among hundreds of thousands of vehicles all at once, whereas 5G can join the dots in any traffic situation.

However, a limited amount of legacy devices can remain for a long time. This is compounded by the growing number of mobile devices that are not directly used by individuals, but are an integrated part of other equipment such as parking meters, card readers, surveillance cameras, and more. Such equipment can have a life span of 10 years or more. It is expected, these to remain connected during this life span. This is, in fact, one important reason why many 2G GSM networks are still operational, although both 3G and 4G networks have been implemented. [5]

To this end, we structure the paper as follows. The next section discusses the new radio (NR) access technology and the deployment of a 5G NR network. Next, in Section III and Section IV we present the

interaction between 5G NR and LTE, the dual connectivity and signaling. Section V concludes the paper.

II. COEXISTENCE

Although LTE is a very capable technology, it is not possible to respond to LTE or its evolution. In addition, the development of technology over 10 years since the start of LTE has enabled more advanced technical solutions. To meet these requirements and harness the potential of new technologies, 3GPP has initiated the development of a new radio access technology known as NR (New Radio). A scope determination workshop was held in the fall of 2015 and technical work began in the spring of 2016. The first version of the NR specification was available at the end of 2017 to meet the commercial requirements for early implementation of 5G. NR uses many of the structures and characteristics of LTE. However, being a new radio access technology means that NR, unlike LTE evolution, is not limited by the need to maintain backward compatibility. [6, 7, 8]

The deployment of a 5G NR network should start from an existing 4G LTE network that already has good coverage. 5G NR will need to coexist and interact with LTE for many years to come, not only as a way to reduce time to market but also to provide good coverage and mobility. With the introduction of 5G NR, new use cases will emerge and 5G devices will hit the market. We see the introduction of standalone 5G NR when devices are more widely available, new use cases (e.g., ultra-reliable and low-latency communications, industrial IoT) are beginning to pick up speed, and NR has more access to both new and existing spectrum. Ultimately, 5G NR is likely to become the core cellular technology used to handle multiple use cases across multiple industries. [8]

Current cellular systems operate below 6 GHz. A large amount of spectrum is available in the millimeter wave band (30-300 GHz); however, today there is no commercial mobile communication system operating in millimeter wave frequencies.

4G LTE is only designed for frequencies below 6 GHz. 5G new radio (NR) will operate at frequencies below 1 GHz up to 100 GHz. 5G NR will unveil new frequencies and new functionality to support ever-growing human-centric and machine-centric applications. [9]

III. DUAL-CONNECTIVITY

The reason the roll-out of 5G requires not only updated antennas, but also considerably more of them, has to do with the different wavelength signals. That means the 5G network can drop to 4G, depending on proximity to the base stations. The interaction between 5G NR and LTE goes far beyond simply allowing seamless transmissions and handover between these two technologies and allowing their parallel deployment. 5G NR can be deployed in the same spectrum as LTE in such a way that the total spectrum capacity can be dynamically allocated between the two technologies. In addition, the 5G NR allows dual connectivity to LTE, which implies that the devices can have a simultaneous connection to both LTE and NR.

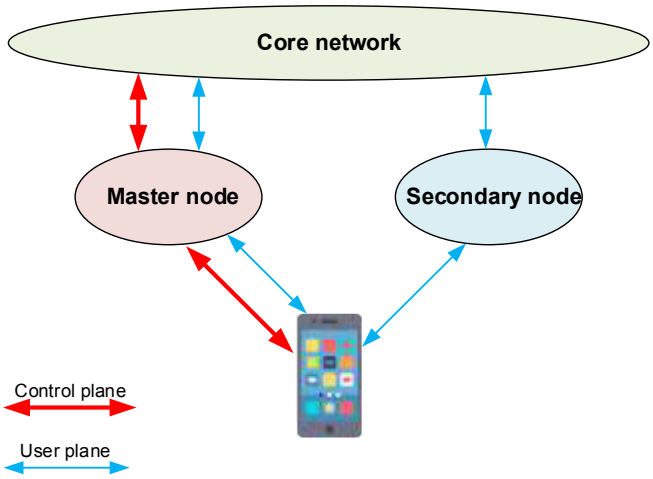


Figure 1. Principle of dual connectivity

The basics of dual connectivity are in that, a single device has simultaneous connections to few nodes in the radio access networks, for example eNB in LTE and gNB in 5G NR, (Fig. 1). Radio access control plane is controlled by one master node, which also handles the entire RRC-based (Radio Resource Control) configuration of the device. One or more secondary nodes provide additional user plane connections for the device. [10, 11]

A. *Different Locations of Base Station*

In this case, LTE and 5G NR network base stations (eNB and gNB respectively) are located in different geographical locations, as shown on Fig.2. The user device can be registered on the smaller 5G cell and a much larger LTE cell range, at the same time. [12, 13]

The LTE macro layer will be responsible to ensure the master nodes. Thus, the control plane is maintained even if the connection to the higher-frequency smaller 5G cell layer is temporarily lost. In this case, the 5G NR layer provides very high bandwidth and very high data rates, while dual connection to the lower-frequency LTE cell provides additional connection strength. [14]

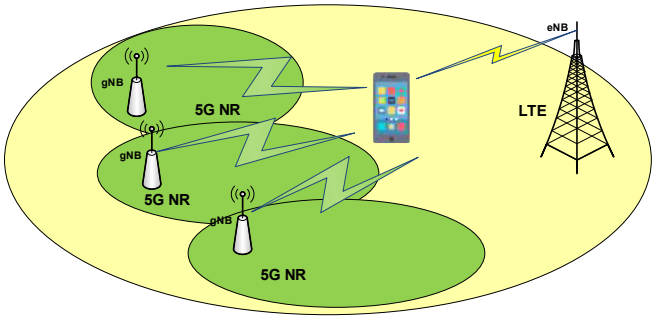


Figure 2. Dual connectivity for different locations

B. *Co-sited Base Stations*

In the case of joint base stations (eNB and gNB), double connections can be achieved (Fig.3). In this way, dual connectivity allows higher data rates for the end user by aggregation of bandwidth of 5G NR and LTE media. This is only possible when 5G NR operates in the same (or close) frequency spectrum as LTE. [13]

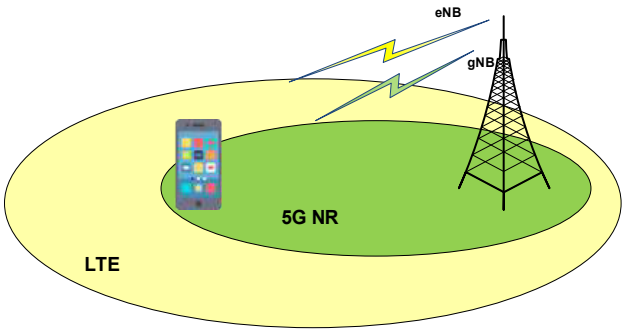


Figure 3. Co-sited base stations, same spectrum

C. *Co-sited Base Stations, Different Spectrum*

Co-location can be used when the two technologies operate in different spectrum, including 5G NR transmission in mm-Wave range. 5G NR may not be able to ensure connectivity over the entire cell. However, the 5G NR portion of the network can still capture much of the total traffic, thus allowing the LTE portion to focus on providing services to devices in poorly covered locations (Fig. 4).

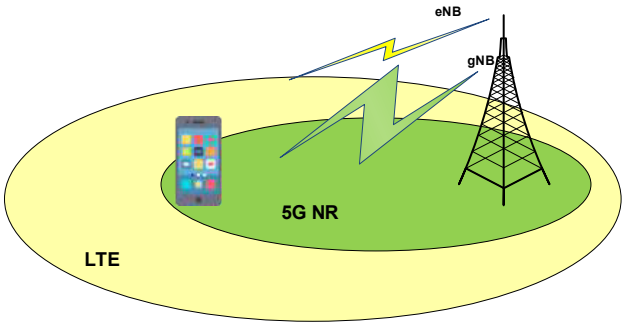


Figure 4. Co-sited base stations, different spectrum

In this scenario, 5G NR media typically has a much wider bandwidth than LTE. While in coverage, the NR carrier will provide much more higher data rates than LTE. The main advantage of dual connectivity is the increased durability for higher frequency deployment.

IV. SIGNALING

In dual connectivity between LTE and NR, there will be multiple uplink carriers (at least one LTE uplink carrier and one 5G NR uplink carrier), transmitted from a single device. Due to the nonlinearity in the radio frequency circuit, simultaneous two-carrier transmission will create intermodulation products at the output of the transmitter. Depending on the specific carrier frequencies of the transmitted signals, some of these intermodulation products may fall within the receiver range of the device, causing "self-interference", or also called intermodulation distortion (IMD). IMD will add noise to the receiver and cause the receiver's sensitivity to deteriorate. IMD impact can be reduced by granting stricter linearity requirements on the device. On the other side, this will lead to a negative impact on device' cost and power consumption. [15]

To reduce the impact of IMD, a Single-TX operation must be implemented – this will ensure that, there will be no simultaneous transmission of LTE and 5G NR media upstream into the device, even though the device operates with a LTE / NR dual connection. LTE and 5G NR planners must prevent jointly the simultaneous transmission of LTE and NR upstream carriers in the case of Single-TX operation. This requires additional coordination between eNB and gNB. 3GPP specifications include explicit support for the exchange of standardized messages between base stations for this purpose. A Single-TX operation essentially results in time multiplexing between LTE and 5G NR uplink transmissions in the device. It is desirable that, full usage of the respective downlinks carriers can be maintained.

V. WORKING TOGETHER

Most low-frequency spectra is already occupied by contemporary technologies, mainly LTE and Wi-Fi standards. Embedding 5G NR in lower frequency spectrum needs to be performed in the same (or near) spectrum, already used by LTE. The most convenient way to perform this task is static frequency domain sharing, where part of the LTE spectrum migrates to 5G NR. The main drawback will be that the bandwidth is less accessible for each technology, which will result in a reduced data rate.

The more desirable solution is to share dynamically the same spectrum to 5G NR and LTE. In this way, it will keep the full bandwidth and corresponding peak data rates for both technologies. In addition, the total capacity of the spectrum can be dynamically set to match the conditions of each technology.

There are two main scenarios for LTE / 5G NR coexistence – both in uplink and downlink transmissions or coexistence in uplink only (Fig.5).

In general, coexistence in uplink direction is more reliable, compared to downlink transmission and can be supported through planning coordination.

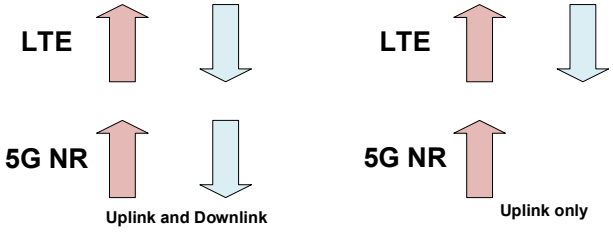


Figure 5. LTE/5G NR coexistence

5G NR and LTE uplink planning must be coordinated, in order to avoid collisions. In addition, downlink should use scheduling coordination to avoid collisions between scheduled LTE and 5G NR broadcasts.

VI. CONCLUSION

In the pursuit of higher data rates and lower latencies beyond the capabilities of the 4G network, the next generation of wireless mobile communications must embrace revolutionary ways to use the frequency spectrum. A new era is emerging, which requires the development of evolved network technology. The infrastructure must be ready and face the next-generation services and demands. We not only need to develop our inheritance system to be more competitive, but we also need new troublemaking ideas to secure the 5G market and ensure growth in the future. In fact, we need to take an active position to be ready for making history of 5G.

REFERENCES

[1] Iliev T.B., Gr. Mihaylov; T. Bikov, E. Ivanova, I. Stoyanov, D. Radev. LTE eNB traffic analysis and key techniques towards 5G mobile networks. 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatia, Croatia, 2017, pp. 497 – 500

[2] Iliev, T., Gr. Mihaylov, E. Ivanova, I. Stoyanov, Power control schemes for device-to-device communications in 5G mobile network. 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatia, Croatia, 2017, pp. 416 – 419

[3] E.P. Ivanova, T. Iliev, G. Mihaylov, I. Stoyanov, F. Tsvetanov, E. Otsetova, D. Radev, Computer Modeling and Simulation of Teletraffic Models for 5G Networks, 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatia, Croatia, 2015, pp. 479-482

[4] E. Dahlman, S. Parkvall, J. Sköld, 4G LTE-Advanced Pro and The Road to 5G, Academic Press, 2016.

[5] E. Dahlman, S. Parkvall, J. Sköld, 5G NR: The Next Generation Wireless Access Technology, Academic Press, 2018.

[6] O. Teyeb, G. Wikström, M. Stattin, T. Cheng, S. Fáxer, H. Do, Evolving LTE to fit the 5G future, Ericsson Technology Review.

[7] 3GPP TR 36.913, Requirements for Further Advancements for Evolved Universal Terrestrial Radio Access (E-UTRA) (LTE-Advanced) (Release 9).

[8] A. Zaidi, F. Athley, J. Medbo, et al., 5G Physical Layer: Principles, Models and Technology Components, Academic Press, 2018.

[9] 3GPP TS 23.402, Architecture Enhancements for Non-3GPP Accesses.

[10] 3GPP TS 23.501, System Architecture for the 5G System.

[11] 3GPP, TS 38.211 NR Physical channels and modulation, June 2018.

[12] 3GPP, TS 38.213 NR Physical layer procedures for control, June 2018.

[13] 3GPP, TS 38.214 NR Physical layer procedures for data, June 2018.

[14] 3GPP TS 38.101-3, NR; User Equipment (UE) Radio Transmission and Reception. Part 3. Range 1 and Range 2 Interworking Operation with Other Radios.

[15] 3GPP TS 38.101-4, NR; User Equipment (UE) Radio Transmission and Reception. Part 4. Performance Requirements.

Decoding of LTE Turbo Codes Initialized with the Two Recursive Convolutional Codes

Dejan Spasov
Faculty of Computer Science and Engineering, Skopje, Republic of North Macedonia
dejan.spasov@finki.ukim.mk

***Abstract* – Turbo codes were the first error-correcting codes that demonstrated reliable communications near the channel capacity with practically feasible hardware. Due to their excellent error-correcting capability, they are part of many modern communication technologies, like 3G, 4G, LTE, etc. An LTE turbo encoder is a rate-1/3 systematic encoder made of a first 8-state recursive convolutional encoder and a second 8-state recursive convolutional encoder. Recursive convolutional encoders are identical, coupled in parallel concatenation scheme, and connected with pseudo-random QPP interleaver. The decoding of LTE Turbo codes is iterative. One iteration starts with decoding of the first convolutional code and continues with decoding the second convolutional code. In order to achieve higher decoding speeds, recursive convolutional codes are decoded with the MAX-Log-MAP algorithm. In this paper, we decode LTE turbo codes with two decoding processes. In one of the decoding processes iterations start with the first convolution code, wherein in the other decoding process iterations start with the second convolution code. We combine the results of both decoding processes to improve turbo decoding process. In channel coding in many communications systems, the speed of decoding is the main design goal. Proposed decoding technique may be used to reduce the number of iterations of LTE turbo decoders, thus improving the decoding speed.**

Keywords* – *Turbo codes; convolutional codes; forward error-correction; Long Term Evolution.

I. INTRODUCTION

Channel coding plays an important part in modern communication technologies. In [1] Shannon gave probabilistic proof that we can communicate with an arbitrary small probability of error as long as the communication rate is below the channel capacity. Error-correcting codes are a constructive solution of the Shannon's theorem. Many error-correcting codes were developed to provide low-power and reliable communication over unreliable channels [2]. In fact, communication near the channel capacity can be achieved with any long code [3]. The bigger problem was the decoding of the long codes. One of the first error-correcting codes with the potential to be successfully decoded were the low-density parity-check (LDPC) codes [4]. However, semiconductor technology of the time was not adequate for practical implementation of the decoding algorithms for the LDPC codes.

Funded by a project from the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University.

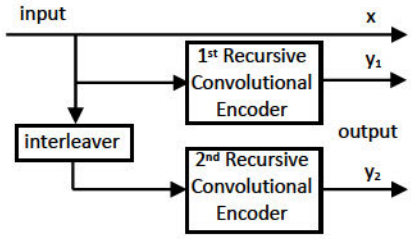


Figure 1. Turbo encoder

Invented by a hardware engineer, turbo codes were the first error-correcting codes with feasible hardware complexity that demonstrated reliable communications with a bit error probability of 10^{-5} for a signal-to-noise ratio of 0.7 dB [5]. In their almost 30-year life span, turbo codes played an important role in communication technology. Since their inception, they were immediately included in many standards, e.g. 3G, 4G, LTE, WiMAX, etc. Moreover, they have motivated scientists to search and discover new efficiently decodable error-correcting codes with error-correcting capability near the Shannon's limit. Thus, LDPC codes were re-discovered in 1996 [6], and polar codes were introduced in 2008 [7].

Turbo codes are convolutional error-correcting codes. Fig. 1 show an ordinary turbo encoder. Originally, turbo codes were envisioned as rate-1/3 systematic convolutional codes comprising two recursive convolutional codes coupled in a parallel concatenation scheme via random interleaver. The turbo encoder is configured to output the original information sequence x in addition to the parity bit sequences y_1 and y_2. Information sequence x and the parity sequence y_1 are referred as the 1st convolutional code, while the information sequence x and the parity sequence y_2 are referred as the 2nd convolutional code. Thus, we may say that the turbo code is parallel concatenation of the 1st convolutional code and the 2nd convolutional code. The purpose of the random interleaver is to provide random permutation of the information sequence x to the 2nd recursive systematic convolutional encoder. Higher code rates can be achieved by puncturing the parity bits from the recursive convolutional codes. The 1st recursive convolutional encoder and the 2nd recursive convolutional encoder are configured to output the parity bit sequences y_1 and y_2, accordingly. In general, a turbo encoder is defined as any combination of two or more preferably identical (recursive) convolutional encoders connected with random interleavers [8].

At the receiver's end, a turbo decoder is configured to decode both the 1st convolutional code and the 2nd convolutional codes and to output bit estimates. The 1st

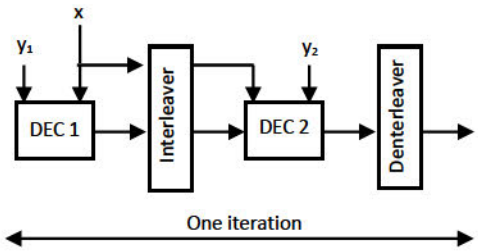


Figure 2. Turbo Decoder

convolutional code and the 2nd convolutional code are decoded sequentially one after another in alternating fashion, starting with the 1st code. The original turbo decoder decoded the 1st and the 2nd convolutional code individually with the Bahl-Cocke-Jelinek-Raviv (BCJR) decoding algorithm [9]. LTE turbo codes, on the other hand, are decoded with the MAX-Log-MAP algorithm [11]. In modern applications, the MAX-Log-MAP algorithm is preferred choice because it achieves greater decoding speeds.

In [10] we have published a turbo decoding algorithm that improves (decreases) signal to noise ratio (SNR), while achieving same bit error rate (BER) as the original turbo decoder. In this paper, we extend this idea to propose a turbo decoding algorithm where individual recursive convolutional codes are decoded with the MAX-Log-MAP algorithm. To measure improvements of the proposed algorithm we apply it for decoding the LTE turbo codes and report improvements. In Section II the turbo decoding algorithm is introduced. In Section III the MAX-Log-MAP decoding algorithm is described. In Section IV the LTE turbo codes are outlined. In Section V a decoder for decoding LTE turbo codes is proposed. In Section VI the proposed decoder is compared with regular turbo decoder for decoding LTE turbo codes. Section VII concludes the paper with discussion on reported improvements in decoding the LTE turbo codes.

II. The Turbo Decoding Algorithm

Decoding of turbo codes is serial process. Consider an ordinary turbo code that consists of two recursive convolutional codes in parallel concatenation, as in fig. 1. Fig. 2 shows a turbo decoder for the ordinary turbo code. The turbo decoder consists of two decoders DEC 1 and DEC 2 for decoding convolutional codes in serial coupling. Each decoder DEC 1 or DEC 2 is configured to decode one of the recursive convolutional codes. DEC 1 may be configured to decode the 1st convolutional code, while DEC 2 may be configured to decode the 2nd convolutional code. DEC 1 and DEC 2 may be configured to perform soft decoding based on the BCJR algorithm [9] or on the Max-Log-MAP algorithm [11]. For each decoded bit x_i in the information sequence x, both decoders are configured to output Log Likelihood Ratio (LLR). For a bit x_i, the Log Likelihood Ratio $LLR[x_i]$ is a real number proportional to the probability that the bit x_i is logical 1

$$LLR[x_i] = Log\left(\frac{\Pr\{x_i = 1|observation\}}{\Pr\{x_i = 0|observation\}}\right) \quad (1)$$

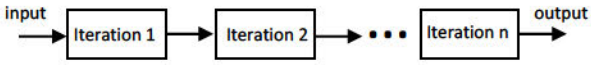


Figure 3. Iterative turbo decoding

DEC 1 is coupled to receive input sequence x and parity sequence y_1 masked with channel noise, i.e. $\hat{x} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ and $\hat{y}_1 = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$. If available, DEC 1 is coupled to receive a priori estimates for each bit x_i. DEC 1 is configured to output $LLR[x_i]$ for each bit. The second decoder DEC 2 is coupled to receive interleaved input sequence $\hat{x}$, and parity sequence y_2 masked with channel noise, i.e. $\hat{y}_2 = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$. Additionally, DEC 2 is coupled to receive interleaved output from DEC 1 as a priori estimates (LLR) for decoded bits x_i. DEC 2 is also configured to output $LLR[x_i]$ for each bit.

Fig. 3 shows the iterative nature of the turbo decoding process. The turbo decoder on fig. 2 may be considered as one iteration of the iterative turbo decoder on fig. 3. LLR estimates from one iteration are fed as a priori estimate for the next iteration. Ideally, an iterative turbo decoder may be implemented with separate hardware for each iteration and for each convolutional decoder DEC 1 and DEC 2. The ideal turbo decoder may offer best performance in terms of decoding speed, but at the cost of highest hardware complexity. In order to decrease hardware complexity, parts of the iterative turbo decoder may be reused and may be time-shared by more than one iteration. For example, one hardware decoder may perform the function of DEC 1. Then, the same hardware decoder may perform the function of DEC2. This hardware decoder may be reused for each iteration. Such an implementation of the turbo decoder may offer lowest hardware complexity, but slowest decoding speed. Thus, when designing turbo decoders, hardware designers need to balance between decoding speed and hardware complexity.

III. The MAX-Log-MAP Decoding Algorithm

A recursive convolutional encoder with m memory cells may be considered as finite state machine with 2^m states. Transitions from one state to another through the time is known as trellis diagram. When producing a sequence of parity bits y_1 or y_2, the encoder may be considered to move through the trellis. Thus, a sequence of parity bits corresponds to a path through the trellis. The MAX-Log-MAP algorithm is configured to decode convolutional code generated with a recursive convolutional encoder. The MAX-Log-MAP tries to find the most probable path through the trellis. In one embodiment, DEC 1 and DEC 2 may implement the MAX-Log-MAP algorithm. The MAX-Log-MAP algorithm may be envisioned to move through the trellis diagram of the recursive convolutional encoder in same direction as the encoder, and in the opposite direction of the encoder. The former direction is known as *forward α recursion*. The later direction is known as *backward β recursion*.

In the forward α recursion the decoder associates a probability estimate $\alpha_i(s)$ with each vertex of the trellis diagram, where index i refers to the time $t = i$, and

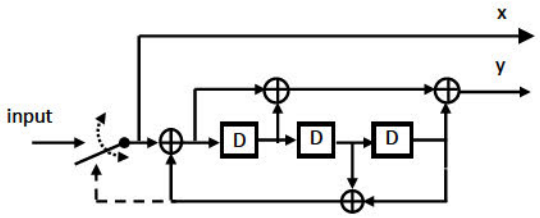


Figure 4. RSC encoder for LTE turbo code

argument s refers to the state of the FSM machine. Following the direction of the movement through the trellis, at a time $t = i + 1$, probability estimate $\alpha_{i+1}(s')$ for a state s' is recurrently computed

$$\alpha_{i+1}(s') = \max\left(\alpha_i(s) + \gamma_i(s, s')\right) \tag{2}$$

where maximization is over all states $s = 0,1, \ldots, 2^m - 1$, and $\gamma_i(s, s')$ denotes branch probability, which depends on the transition probabilities of the trellis diagram and the transition probabilities of the transmission channel. Initial conditions for the recurrence relation (2) are defined as

$$\begin{aligned} \alpha_0(s) &= 0, \quad for\ s = 0, \\ \alpha_0(s) &= -\infty, \quad for\ s \neq 0. \end{aligned} \tag{3}$$

In the backward β recursion the decoder associates a probability estimate $\beta_i(s)$ with each vertex of the trellis diagram, where index i refers to the time $t = i$, and argument s refers to the state of the FSM machine. Following the direction of the movement through the trellis, at a time $t = i - 1$, probability estimate $\beta_{i-1}(s')$ for a state s' is recurrently computed

$$\beta_{i-1}(s') = \max\left(\beta_i(s) + \gamma_i(s, s')\right) \tag{4}$$

where maximization is over all states $s = 0,1, \ldots, 2^m - 1$, and $\gamma_i(s, s')$ denotes branch probability, which depends on the transition probabilities of the trellis diagram and the transition probabilities of the transmission channel. Initial conditions for the recurrence relation (4) are defined as

$$\begin{aligned} \beta_0(s) &= 0, \quad for\ s = 0, \\ \beta_0(s) &= -\infty, \quad for\ s \neq 0. \end{aligned} \tag{5}$$

Solving the recurrence relations (2) and (4), the logarithm of likelihood ratio ($LLR[x_i]$) for each bit x_i is computed as

$$\begin{aligned} LLR[x_i] = \max\left(\alpha_i(s) + \beta_i(s_1) + \gamma_i(s, s')\right) - \\ \max\left(\alpha_i(s) + \beta_i(s) - \gamma_i(s, s')\right). \end{aligned} \tag{6}$$

IV. LTE TURBO CODES

LTE turbo codes are specified as channel coding technique in the 3GPP technical specification [13]. A turbo encoder, configured to output LTE turbo code, is implemented as parallel concatenation of two recursive systematic convolutional codes, as shown on fig. 1. Each of the recursive convolutional encoders is implemented with three-bit shift register, as shown on fig. 4, which implies

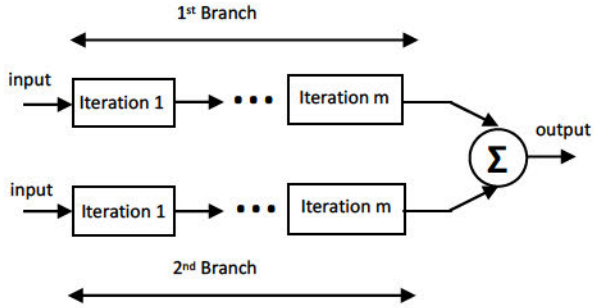


Figure 5. Proposed turbo decoding for LTE turbo codes

that the encoder is with 8 states. The transfer function of an 8-state LTE recursive convolutional encoder is given as

$$G(D) = \left[1 \quad \frac{g_1(D)}{g_0(D)}\right] \tag{7}$$

where $g_0(D) = 1 + D^2 + D^3$ and $g_1(D) = 1 + D + D^3$. Initially, before it starts to encode, the shift register is initialized with all zeros. After an input sequence is encoded, the 3GPP technical specification mandates trellis termination to be performed. Trellis termination adds additional bits to the information bits that drive the RSC encoder to state zero. Fig. 4 shows the trellis termination with intersected lines. The pseudo-random interleaver needed in a turbo encoder is also part of the 3GPP technical specification. Let input bits to a turbo-code interleaver be denoted as $x_1, x_2, \ldots, x_n$, and let the pseudo random permutation of $x_1, x_2, \ldots, x_n$ provided to the output of the interleaver be denoted as $x'_1, x'_2, \ldots, x'_n$, wherein $x'_i = x_{\Pi(i)}$. The 3GPP technical specification mandates the pseudo random permutation $\Pi(i)$ to be computed as

$$\Pi(i) = (f_1 i + f_2 i^2)\ mod\ n \tag{8}$$

where f_1 and f_2 are given in the table 5.1.3-3 in the technical specification [13].

V. PROPOSED TURBO DECODING ALGORITHM FOR THE LTE TURBO CODES

A weak side of turbo codes is the speed of decoding. Modern standards for communications require decoding speeds above 20 Gbps [14]. However, serial nature of the iterative decoding of turbo codes, as on fig. 3, implies that it would be difficult to reach such decoding speeds. Many decoders have been proposed to meet the 20 Gbps requirement, e.g. PMAP, FPMAP, XMAP, UXMAP [12]. An important feature of all of the proposed turbo decoders designed for decoding at high speeds is that they are all using the MAX-Log-MAP decoding algorithm to decode individual recursive convolutional codes. Algorithms simpler than the MAX-Log-MAP algorithm have been proven to be ineffective.

In order to improve decoding speed of the LTE turbo codes, we propose a turbo decoding algorithm with reduced number of iterations. The proposed turbo decoding algorithm is shown on fig. 5. The proposed algorithm consists of two branches named the 1st branch and the 2nd branch. The 1st branch is a regular turbo decoding algorithm

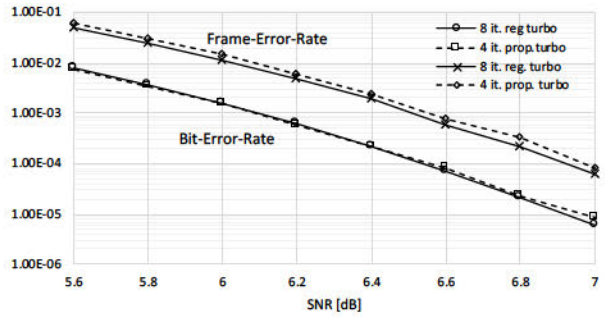


Figure 6. Performance comparison between regular and proposed turbo decoders for frame length N=128 bits

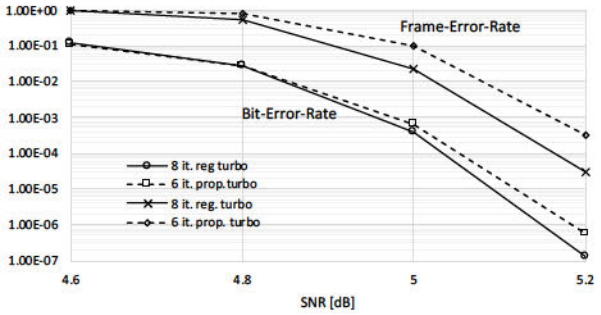


Figure 7. Performance comparison between regular and proposed turbo decoders for frame length N=6144 bits

as on fig. 3, where each iteration starts with the 1st recursive convolutional code. The 2nd branch is also a regular turbo decoding algorithm as on fig. 3, where each iteration starts with the 2nd recursive convolutional code. After m iterations, LLR outputs from both branches are combined into one LLR estimate.

VI. Experimental Results

The 3GPP technical specification mandates that the length of the information sequence in LTE turbo codes is between 40 bits and 6144 bits. The most frequently tested information sequence lengths in academic papers are 128 bits and 6144 bits. Fig. 6 and fig. 7 show bit error rate and frame error rate of the proposed turbo decoder and the regular turbo decoder for various noise levels. Information sequence on fig. 6 is 128 bits. Information sequence on fig. 7 is 6144 bits.

Regarding fig. 6, the regular turbo decoder performs 8 iterations, which is frequently used number of iterations. The proposed turbo decoder performs four iterations. Fig. 6 shows almost identical error performance between the proposed turbo decoding algorithm and the regular turbo decoding algorithm, although the proposed decoder reduces the number of iterations by half. The number of iterations is reduced by half implies that the decoding speed may be doubled compared to the regular iterative turbo decoder shown on fig. 3, while the hardware complexity remains the same.

Regarding fig. 7, the regular turbo decoder performs eight iterations, while the proposed turbo decoder performs six iterations. Fig. 7 shows that same error rate both decoders achieve for SNR difference less than 0.05 dB. However, the proposed decoder performs six iterations, which makes it faster than the regular turbo decoder.

VII. Conlusion

We have demonstrated a turbo decoding technique that improves the decoding speed of turbo codes. For smaller lengths of the information sequence, the proposed technique doubles the decoding speed of the turbo codes, while hardware complexity and bit error performance remain the same as in the regular turbo decoding technique. Even though, LTE turbo codes are used as a test bed, this decoding technique may be used for decoding any turbo code. Moreover, this decoding technique may be applied to any modern turbo decoding algorithm, like PMAP, FPMAP, XMAP, and UXMAP [12].

The proposed turbo decoder may be considered to be obtained by rearranging the building blocks of a regular iterative turbo decoder. Therefore, a natural question that arises is if some new rearrangement of the building blocks of the regular turbo decoder may further improve the decoding speed or the error floor of the turbo codes.

References

[1] C. E. Shannon, "The mathematical theory of communication," Bell System Technical Journal, vol. 27, pp. 379–423, July 1948.

[2] F. J. MacWilliams, N.J.A. Sloane, "The Theory of Error-Correcting Codes," North Holland, Amsterdam, 1977.

[3] A. Barg, "Complexity Issues in Coding Theory," Electronic Colloquium on Computational Complexity, Volume 4, 1997.

[4] R. G. Gallager, " Low-Density Parity-Check Codes," PhD Thesis, MIT University Press, 1963.

[5] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes," in Proc. ICC, Geneva, Switzerland, May 1993.

[6] D. J. C. Mackay, R. M. Neal, "Near Shannon limit performance of low density parity check codes," Electronic Letters, vol. 33, iss. 6, pp. 457-458, 1997.

[7] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes," ISIT 2008, pp. 1173-1177.

[8] S. Benedetto, D. Divsalar, G. Montorsi, F. Pollara, "Serial Concatenation of Interleaved Codes: Performance Analysis, Design, and Iterative Decoding," IEEE Trans. Information Theory 44(3): 909-926 (1998).

[9] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," IEEE Trans. Inform. Theory, vol. 20, no. 3, pp. 284–287, Mar. 1974.

[10] D. Spasov, "Parallel Decoding of Turbo Codes," ICT Innovations 2018, Ohrid, N. Macedoina, pp. 100-106.

[11] A. J. Viterbi, "An Intuitive Justification and a Simplified Implementation of the MAP Decoder for Convolutional Codes," IEEE Journal on Selected Areas in Communication, vol. 16, pp. 260–264, Feb. 1998.

[12] S. Weithoffer, C. A. Nour, N. Wehn, C. Douillard, and C. Berrou, ''25 years of turbo codes: From Mb/s to beyond 100 Gb/s,'' in Proc. IEEE 10th Int. Symp. Turbo Codes Iterative Inf. Process. (ISTC), Dec. 2018, pp. 1–6.

[13] ETSI Technical Specification, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding," 3GPP TS 36.212 version 8.8.0 Release 8, Jan. 2010.

[14] R.G. Maunder, "The 5G Channel Code Contenders," AccelerComm White Paper, pp. 1-3, August 2016.

Evaluation of Commercial Off-The-Shelf LTE Antennas for Use on Unmanned Aerial Vehicles

K.Kainrath*, H.Flühr*, W. Stocksreiter**, R. Findenig**, E. Leitgeb*** and M. Gruber*
* Institute of Aviation, FH JOANNEUM GmbH, Graz, Austria
** Institute of Electrical Engineering, FH JOANNEUM GmbH, Graz, Austria
*** Institute of Microwave and Photonic Engineering, TU Graz, Graz, Austria
klaus.kainrath@fh-joanneum.at

***Abstract*— In this study, different commercial off-the-shelf (COTS) Long Term Evolution (LTE) antennas were evaluated for use on Unmanned Aerial Vehicles (UAVs). Due to the existing infrastructure, the mobile radio network would be suitable for the use of UAVs. Therefore, antenna gain, reflection parameter as well as the radiation pattern of different, rather small and lightweight LTE antennas were measured. It was also investigated how the antenna pattern changes when two identical antennas are used and if a protective cover made out of glass fiber has an influence on the characteristic of the antennas.**

***Index Terms—LTE, antenna,* pattern, UAV**

I. INTRODUCTION

The European Union Aviation Safety Agency (EASA) has published common European rules on Drones (UAVs) in June 11th, 2019 [1]. These regulations must be implemented by the national authorities by June 2020. In Austria, the Austro Control [2] is responsible for this. There will be three different categories, open, specific and certified [3]. Especially regarding the use of UAVs of the specified and certified category, a reliable data link becomes essential. There are already surveys on this topic [4] and it has emerged that the mobile radio network is considered a suitable technology for this.

At the Institute of Aviation of the FH JOANNEUM a UAV for beyond visual line of sight (BVLOS) operations has been developed as part of a research project. Within the upcoming regulations this UAV will count to the category certified. According to current status (Oct 2019), there are no mandatory technical guidelines and regulations for beyond visual line of sight (BVLOS) operation of UAVs. The system designed by FH JOANNEUM was therefore designed on the basis of experimental approval. To ensure the safe operation, all avionics in the UAV were designed redundantly. In terms of data communication, three systems were used in parallel, each of which could be used to fully control the UAV. These include two air-model-making systems, one at 868MHz and another at 2.4GHz and a system based on mobile radio. These systems were all mounted together in the fuselage of the UAV [Fig.1]. The whole UAV is made of fiberglass, including balsa wood reinforcements. Thus, the material has little influence on the reception property. Nonetheless, disruptive influences of the radio system on the overall performance occurred. This means that in range tests the desired range was not achieved when all subsystems in the UAV were in operation. Therefore, among other things, an Electromagnetic Compatibility (EMC) test of the entire system was performed.



Figure 1. Fuselage of the used UAV including avionics

Since the air-model hardware is especially important for takeoff and landing purposes, the arrangement of the 2.4GHz and 868MHz antennas were subsequently optimized. The control over mobile radio should happen only in the flight phase. Commercial off-the shelf (COTS) components were used for the data connection. The transceiver module used is a HUAWEI E3372h LTE USB modem [5]. This supports all necessary LTE frequencies in Austria. The E3372 has built-in patch antennas. The values according to the data sheet for the transmission power for LTE are given with power class 3 (23dBm) and for the reception power accorded with 3GPP TS 36.101 (R9) [6].

On the one hand to avoid unwanted interactions with the neighboring systems and to improve the transmission and reception quality on the other hand, external antennas for the Huawei E3372 were used. This device has two CRC9/TS9 connectors for the use of two external antennas. To avoid exposing of the antennas directly to the airflow of the UAV, protective covers have been built.

The LTE antennas were mounted on the top of the UAV in the first attempt. Due to the circular polarized propagation of mobile radio, a satisfactory result was thus expected.

After successful test flights, an adaptation and, subsequently, an optimization of the arrangement of the LTE antennas is being considered. For this reason, suitable LTE antennas were measured in the radio frequency (RF) measuring chamber of the FH JOANNEUM. From these findings, the installation of the antennas should be improved in the future. In this article, several measurement results of small and lightweight Commercial off-the-shelf (COTS) LTE antennas for the usage on this UAV are presented. For this purpose, suitable LTE antennas were measured according to their receiving characteristics, possible positioning was considered, and the influence of the protective housing examined.

II. SELECTED ANTENNAS AND OPTIONAL MOUNTING

LTE antennas must be suitable for multiple frequencies, depending on their geographical location. For Austria, according to the broadcasting authority "Rundfunk und Telekom Regulierungs-GmbH" (RTR), this would be the bands 1,3,7,8 and 20 which correspond to the frequencies 800MHz (band 20), 1800MHz (3), 2100MHz (1), 2600MHz (7) and 900MHz (8) [7].

During the first test flights with a commercial UAV, no external mobile radio antennas were used. The connection quality was sufficient, there were no dropouts. For use in the target UAV, however, it was decided to rely on external antennas due to the many installed hardware and the associated possible EMC problems. COTS LTE antennas in a compact design, ideally with a CRC9/TS9 connector, were used. For the protection of the antennas as well as for the aerodynamics, protective covers made of glass fiber are to be used on the UAV.

The first model tested is an LTE rod antenna designed for Universal Serial Bus (USB) LTE dongle. The model is called TD-LTE FDD-LTE 4G antenna and is marked Antenna 645 668 SMA. The second antenna that was tested was an Alda PQ antenna called GXS404 SMA. The third antenna tested was a DELOCK 12479. The data are given in the following table I.

TABLE I. ANTENNA MODELS

Model	Antenna parameter		
	Frequency range	*Gain*	*Max. power*
645 668 SMA	0.8 - 2.6GHz	2.5dBi	5.5W
GXS404 SMA	0.8 - 2.4GHz	2dBi	2.5W
DELOCK 12479	0.69 - 0.96 & 1.7 - 2.7GHz	3dBi	5W

These antennas are dipole antennas with an internal matching network to provide the necessary reception characteristics at different frequencies [8]. This consists of either discrete components or microstrip filters. A balun is installed at the output. Details could not be found, because in the data sheets very little information about the used antennas was provided.

The measurements were carried out in the RF chamber of the FH JOANNEUM. This is a full absorber chamber which is suitable for the measurement of free space propagation. On the one hand the transmitting power, the receiving power, the far field effect as well as the reflection parameters S11 were measured. The following measurement setup was used to measure the antenna gain [Fig.2]. A signal generator SMIQ from Rhode and Schwarz was used for signal generation. The transmission power for the antenna was set at 10 dBm. In order to eliminate the cable attenuation of the leading line, signal strength was tapped via a socket before the device under test (DUT).



Figure 2. Measuring set up for antenna gain

The receiving antenna was mounted at a constant distance of three meters. It was connected to a Rhode and Schwarz FSIQ spectrum analyzer. The attenuation of the line was taken into account. The gain of the antenna was determined from this measurement system by simple addition (gain - antenna, signal generator) and subtraction (attenuation - line, Free Space Path Loss (FSPL)) of the individual components.

With this measuring setup, all three antennas were measured with regard to their radiation pattern. Subsequently, different positions and orientations were tested for two antennas of type 645 668 SMA. For this purpose, the antennas were mounted on a fuselage of the UAV specially designed for this purpose [Fig.3].

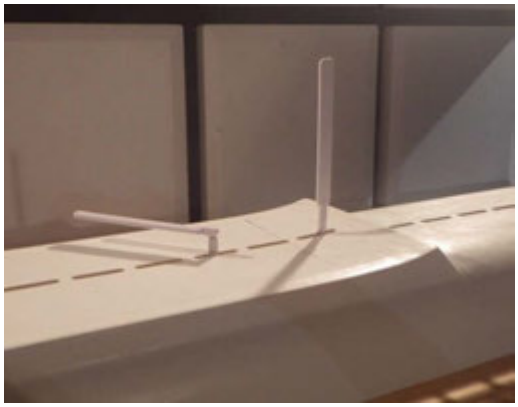


Figure 3. Antenna arangement on UAV fuselage

In the second step the antenna 645 668 SMA was measured with an EMSCAN RFX nearfield scanner. The antenna was examined at 800MHz. For this purpose, one

antenna was measured by itself and then in combination with a second antenna at a distance of λ/4. A spectrum analyzer FSIQ26 was used to measure the received power. All antennas were aligned in 90° elevation and rotated in 5° steps from 0° to 360° in azimuth. The transmission power was set to 10dBm. In two of the three models tested, a fixed cable was mounted on the antenna, which influenced the measurement results because the attenuation of the cable could not be measured explicitly. In addition to the radiation measurements, the return loss S11 was measured with a network analyzer RS ZNB8. At the GXS404 antenna, the influence of the protective cover on the reflection parameter was measured, as shown in figure 4 below:



Figure 4. Antenna with protective aerodynamic fiber glass cover

III. MEASUREMENT RESULTS

The measurement results are divided into results for transmitting power, antenna radiation pattern, receive power (Equivalent Isotropically Radiated Power (EIRP) – Gain) and reflection parameters S11.

The measurement of the transmission power of the antennas was carried out for each antenna for an elevation from 0°, 45° and 90°. The values given are for 90° elevation. In the following table the highest and lowest dBi values can be found for the according azimuth angle.

TABLE II. TRANSMISSION POWER COMPARISON AT 90°

Model	Antenna parameter		
	Frequency (GHz)	*Max dBi @ azimuth*	*Min dBi @ azimuth*
645 668 SMA	0.8	3.2 @ 320°	2.7 @ 30°
GXS404 SMA	0.8	2.4 @ 190°	-2.4 @ 0°
645 668 SMA	1.8	3.3 @ 0°	2.5 @ 280°
GXS404 SMA	1.8	-0.5 @ 350°	-10.7 @ 260°
645 668 SMA	2.1	4.1 @ 30°	3.3 @ 260°
GXS404 SMA	2.1	-2.9 @ 70°	-5.1 @ 300°
645 668 SMA	2.6	1.4 @ 345°	0.1 @ 160°
GXS404 SMA	2.6	-4.6 @ 30°	-8.0 @ 265°

The measurement of the antenna pattern was carried out for the antenna 645 668 SMA. In the first run only one [Fig.5], in the second two antennas were measured. The measurement result for one antenna at 800 MHz is shown graphically in the following figure 5:

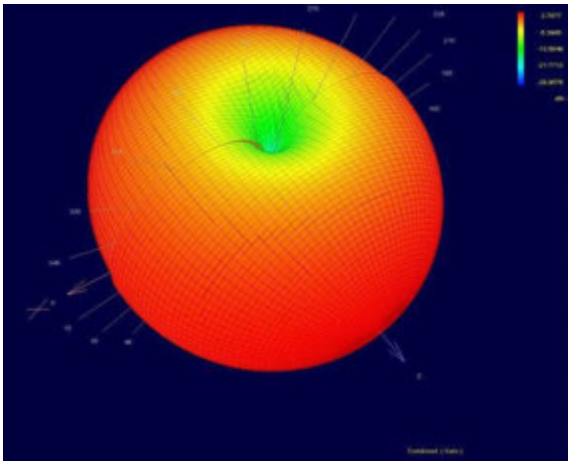


Figure 5. 3D radiation pattern of one 645 668 SMA antenna

The results for the received power measurement are shown in figure 6. Markers are set at the relevant positions for the LTE bands 20, 3, 1 and 7.

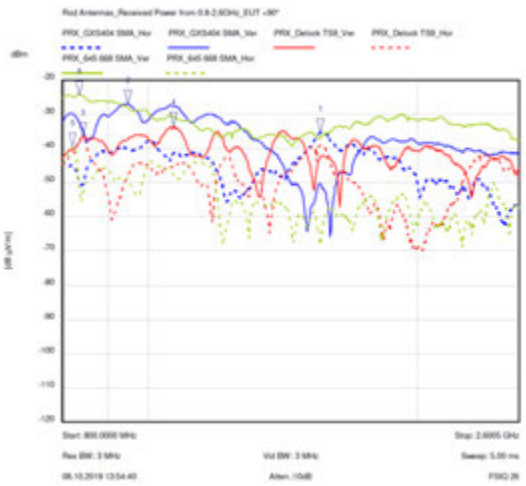


Figure 6. Received power from 0.8 – 2.5GHz

The measurement of the reflection parameters S11 of all antennas in the same range from 800MHz to 2.7GHz yielded the following results [Fig.7]:

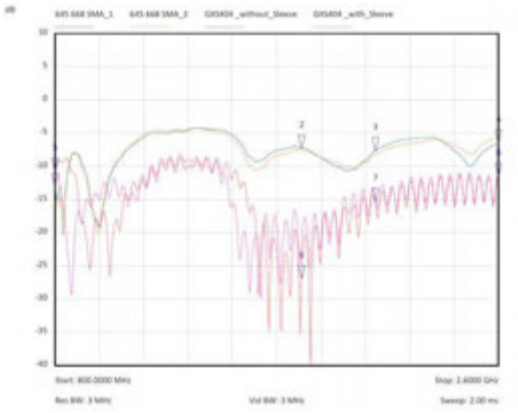


Figure 7. Reflection parameter from 0.8 – 2.5GHz

IV. Conclusion and Outlook

The examination of the antennas yielded very different results: Especially the antenna 645 668 SMA proved to be good over the whole frequency range. Although it is not optimally matched, as shown in Figure 7, it had the best gain of all tested antennas in the LTE bands used [Fig.6].

Particularly interesting is the use of 2 antennas of this type. You can change the antenna pattern by the physical arrangement according to the requirements. The radiation characteristic changes from a sphere to an ellipsoid according to the arrangement of the two antennas. By changing the position and the distance the shape of the ellipsoid can be optimized for the UAV. This so-called beamforming will be part of further investigations.

The GXS404 proved to be good in the frequency range up to about 1.3GHz, but its gain is too low at the higher frequencies. This can be related to the approximately 2.5m long connection cable, the housing material or the internal signal processing. Compared to the 645 668 SMA, this model is much better matched [Fig.7] but has worse radiation properties [Table II and Fig.6]. This model was also used for measurements with and without the glass fiber cover. It was noticeable that the radiation characteristics changed significantly in some cases. This can be attributed to the fact that the antenna does not fit exactly in the envelope. The resulting gaps cause different dielectric transitions between the materials and thus lead to different radiation characteristics. The resulting capacitive influences increase with frequency, making the antenna less efficient at higher frequencies. This could be counteracted by filling one of the antennas with a homogeneous material in the envelope, such as an adhesive. This will be evaluated in further studies.

The DELOCK 12479 proved to be unsuitable during the measurements. This can be due to a defect of the antenna or to a too high attenuation of the SMA to TS9 adapter. The results were always 20dB below the expected values. There were two antenna models as well as four adapters available for testing and all showed the damping effect. Hence, the results were not filled in Table II. This model was therefore not taken into account in further measurements. Nevertheless, the Antenna was used in the first test flights before the measurements and seamed to perform properly. An option would be, to remove the TS9 connector and replace it by an SMA connector just for measurement purposes. However, soldering on another connector could also lead to inaccurate measurement results.

The measurements show that a gain of 2 up to 4dBi can be achieved with small and lightweight COTS antennas. Not all models show the necessary efficiency over the whole frequency range. The antenna pattern can be optimized for use at the UAV by arranging the antennas differently. Alternatively, one could use more expensive antennas from the automotive sector.

Investigations to optimize the antenna positions will be continued. Further measurements are also planned. The results obtained so far shall be verified and new models shall be considered. Furthermore, the DELOCK 12479 antenna with soldered SMA connector will be tested again.

[1] Europan Union Aviation Safety Agency, https://www.easa.europa.eu/newsroom-and-events/press-releases/eu-wide-rules-drones-published, online, last seen 16.01.2020

[2] Austro Control, LBTH67, Betrieb von unbemannten Luftfahrzeugen, https://www.austrocontrol.at/luftfahrtbehoerde/lizenzen__bewilligungen/flugbewilligungen/unbemannte_lfz , online, last seen 16.01.2020

[3] [1] Europan Union Aviation Safety Agency, Civil drones, https://www.easa.europa.eu/easa-and-you/civil-drones-rpas, online, last seen 16.01.2020

[4] K. Kainrath, "Evaluation of Mobile Radio Links as Command-And-Control Links for Unmanned Aerial Vehicles," IEEE, 2016

[5] HUAWEI E3372, technical documentation, https://consumer.huawei.com/at/mobile-broadband/e3372/specs/, online, last seen 16.01.2020

[6] 3GPP TS 36.101 (R9), LTE;Evolved Universal Terrestrial Radio Access (E-UTRA);User Equipment (UE) radio transmission and reception, technical specification, 2010, https://www.etsi.org/deliver/etsi_ts/136100_136199/136101/09.03.00_60/ts_136101v090300p.pdf, online, last seen 16.01.2020

[7] Austrian Regulatory Authority for Broadcasting and Telecommunications, https://www.rtr.at/de/tk/Frequenzen, online, last seen 16.01.2020

[8] K. Rothammel, A. Krischke, "Rothammels Antennenhandbuch", 12th ed, DARC Baunatal, 2001, pp745-749

Techno-Economic Analysis of Deployment Options for Converged 5G Wireless-Optical Access Networks

Maurice Bongard*, Kai Grunert** and Slavisa Aleksic***
* Deutsche Telekom Technik GmbH, Cologne, Germany
** Detecon International GmbH, Cologne, Germany
*** Leipzig University of Technology (HfTL), Leipzig, Germany
Maurice.Bongard@telekom.de, Kai.Grunert@detecon.com, Aleksic@hft-leipzig.de

Abstract - **Currently, a lot of efforts are put in place to design and implement a high-capacity, flexible and highly efficient 5G network infrastructure. A very important consideration that should be kept in mind when designing future 5G networks is developing an adequate network architecture and high-capacity and future-proof backhaul and fronthaul links. Especially novel network concepts such as the cloud radio access network (C-RAN) set very high requirements on the backhaul and fronthaul connectivity and call for a solution that ensures high-capacity, flexibility, practicability and future readiness while providing an easy and economical migration path.**

This paper examines deployment options for converged optical-wireless access networks in urban areas regarding their cost efficiency and implementation flexibility. It presents results of a case study for a coordinated planning of optical fixed access and wireless backhaul solutions for 5G networks by considering following options: i) roll-out of a converged fiber-to-the-home (FTTH) network infrastructure that provides both fixed internet access and integrated cellular backhaul, ii) use of the existing fiber-to-the-curb (FTTC) network infrastructure to provide the backhaul for cellular networks, and iii) deployment of a dedicated optical fiber network to act as backhaul of cellular networks. A techno-economic study was carried out to compare the economic viability of the three deployment options under realistic assumptions.

Keywords - Wireless-Optical Convergence; Network Planning; Techno-Economic Analysis; Access Networks; 5G

I. INTRODUCTION

In the age of digitalization and with the advent of 5G, a powerful and flexible optical-wireless access network is needed. 5G will be capable of providing enhanced mobile broadband coverage (emBB) and ultra-reliable low-latency communications (URLLC) to meet the requirements of future applications and use cases. To support the requirements, 5G will soon use additional frequency spectra including the mm-wave band. This makes significantly higher bandwidths available. However, millimeter wavelengths do not provide the same coverage as low-frequency radio waves. For this reason, a densification of the radio access network by using small cells is necessary, which will require a roll-out of more fiber-optic network infrastructure to connect the large number of base stations to the network core. Traditionally, fixed and mobile access infrastructures has mostly been planned and implemented separately. The implementation of a new infrastructure is very cost-intensive, so that in the future, a fixed and mobile convergence in the access network could be a solution to reduce both the capital expenditure (CAPEX) and the operational expenditure (OPEX) by a common fiber access [1]. Indeed, many telecommunications operators have started to follow the combined network planning approach and implement technologies such as cloud radio access network (C-RAN) to reduce the CAPEX and OPEX of a modern mobile access network [2].

In this paper, we present results of a techno-economic study that analyses the CAPEX of three deployment options for fiber-based access infrastructure for future 5G networks.

II. RELATED WORK

In recent years, telecommunications operators have massively deployed fiber-to-the-home (FTTH) or intermediate solutions such as fiber-to-the-curb (FTTC) or fiber-to-the-building (FTTB) in the fixed access network. Provision of FTTH network access for residential users is mostly based on passive optical network (PON) technologies - a point-to-multipoint (P2MP) architecture with optical splitters. In contrast, dedicated point-to-point (P2P) fibers are still often implemented for both the fixed network access and backhauling of the mobile network [1,3].

Due to the broad penetration of point-to-multipoint FTTH networks, especially in urban areas, PON access has become increasingly common. The next generation PON solutions, such as the NG-PON2, are capable to deliver much higher capacities by using higher data rates and wavelength-division multiplexing (WDM), making this technology attractive for implementing a converged network that can be used to transport both fixed and mobile traffic over a common physical network [1,4].

Much of the research on converged networks, e.g. [4,5,6,7,8,9], focuses on a converged infrastructure that assume 5G mobile network with an optical backhaul. A

recent study by the FTTH Council Europe [4] as well as a webinar by CommScope [9] has underlined the attractiveness of the convergence of fixed FTTH and mobile transport networks for minimizing CAPEX and OPEX. A special focus of many research works is placed on the migration path, which is a complex, strategic problem of network planning. There exist several approaches proposed to minimize the migration complexity and costs [10]. Independently of the approach chosen, the use of powerful planning tools is essential. Modern network planning tools support optimized and automated planning processes to achieve a high-capacity and future-proof design that ensures an easy migration path while minimizing costs. Most of these tools are based on either geometric or geographical models [11]. While geometric models are well suited for theoretical studies and are easy to use, they, however, can lead to inaccurate results because of their inability to take into account geographical and operational constraints as well as location dependent variation of the distribution of demand points and data traffic. Geographical models, on the other hand, ensure realistic results due to the availability of exact geographical and operational data as well as constraints at the street and building level. For the case study presented in this paper we used a tool based on a geographical model in order to estimate the required equipment and installation effort to cover a particular urban area. Three deployment scenarios for a converged wireless-optical access network have been considered and analyzed on the basis of realistic assumptions. The scenarios are compared with each other regarding their economic viability.

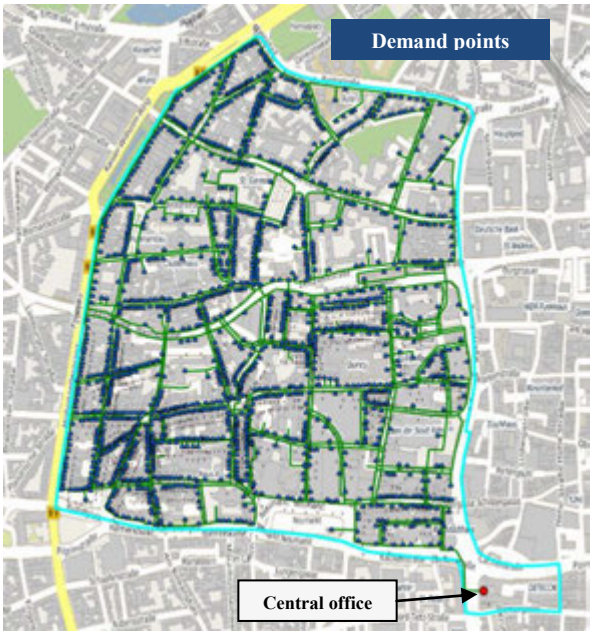
III. MODELLING APPROACH

The planning tool "NetWorks Fixed Access" from Detecon International GmbH is used to determine the effort required for the roll-out of the converged network infrastructure. For this case study, we consider a network roll-out in an urban development area with a high cell density. The number of demand points (see Figure 1 a)) results from the locations of the buildings. The locations of macro and small base stations are determined such that an adequate coverage of the exemplary development area within the city of Cologne is provided (see Figure 1 b)).

For an automated planning procedure, a uniform material concept is necessary and decisive for the implementation of cable infrastructures. It should be pointed out that we assumed that costs are based on approximate but realistic values and only passive infrastructure is considered for the cost calculation.

The first step in planning the passive infrastructure for optical access networks is to identify the demand points for the deployment options as well as potential routes for civil engineering works. The required demand points were grouped together to form a distribution area according to the capacity of the distributor as can be seen in Figure 2. The planning tool applies optimization algorithms to calculate the ideal size and shape of the distribution area. Cable protection duct routes and cable connections were then determined based on the placement and the size of the distribution areas. After the first step in planning, a balance of the planning was made from the Bill of Quantities (BoQ).

a)



b)

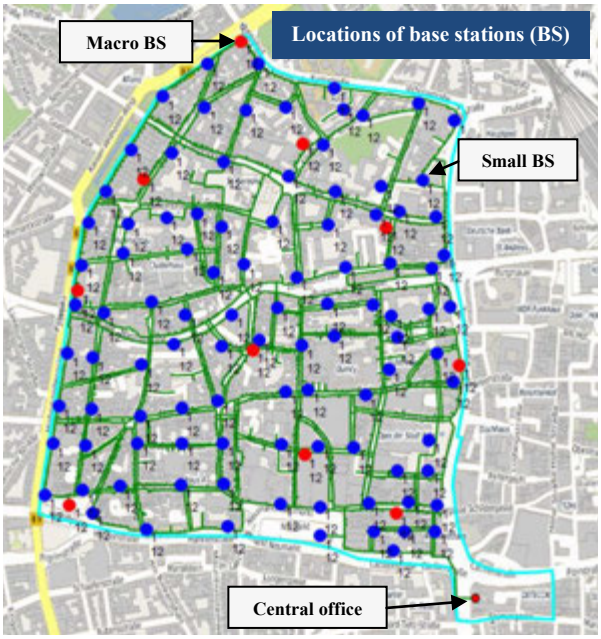


Figure 1. Demand points (a) and locations of base stations (b) in the considered exemplary urban coverage area.

The cost assumptions for various components and working processes are shown in Table I. Note that the cost values for the underground work and the installation of cables and ducts are given per meter, while all other values represent the total costs of a component or a subsystem.

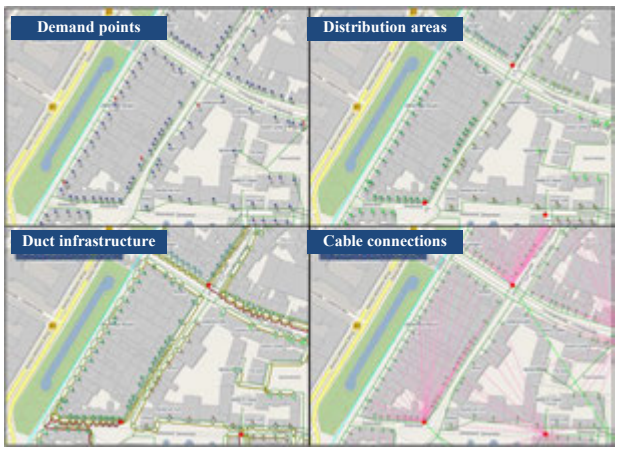


Figure 2. The main steps carried out during the planing procedure.

TABLE I. COST ASSUMPTIONS

Component	Unit	Cost
Central Office		€ 60,000
House feed		€ 300
Fiber Distribution Hub		€ 1,500
Construction work	[m]	€ 100
Micro-Duct 12x10mm	[m]	€ 8
Micro-Duct 7x16mm	[m]	€ 10
LWL-Cable 144F/288F	[m]	€ 3/ € 5
LWL-Cable 12/24F	[m]	€ 0.5 /€ 1
APL-24F		€ 50
Optical Distribution Frame (ODF)		€ 15,000
Optical Splitter 1:32		€ 50

IV. DEPLOYMENT OPTIONS

In this section, the considered deployment options are presented. The roll-out of new network infrastructure has always been very expensive and time-consuming. Since telecommunications operators have already put a lot of effort to roll-out fiber optic networks to various extents and using different technological options, a flexible expansion strategy is being sought to realize the backhaul for 5G. This strategy should be as cost efficient and future-oriented as possible. In this study, we consider the following three deployment options, which we examine for their flexibility and economic viability.

A. Converged optical-wireless network with FTTH as a cellular backhaul (C-FTTH)

In the first deployment option, we consider an area-wide FTTH deployment that includes both fixed access and backhaul for cellular network. Here, fiber-optic cables reach all buildings in the residential and business areas. The distribution hubs are supplied from the central office (CO) in a star configuration. The buildings are then directly connected to the fiber optic distribution hubs. Additionally, the FTTH network infrastructure also provides direct links to mobile base stations. Each fiber optic connection is implemented in a P2P manner to provide maximum flexibility.

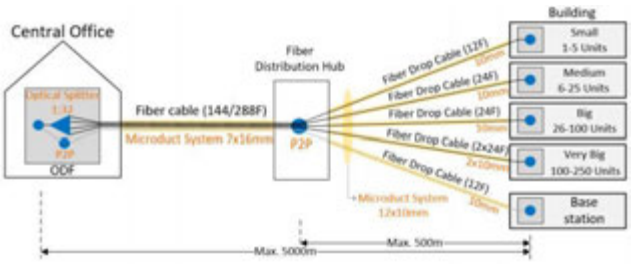


Figure 3. Schematics of the C-FTTH deployment option (P2P: Point-to-Point, ODF: Optical Distribution Frame)

B. Use of the existing FTTC network infrastructure to provide the backhaul for cellular networks (E-FTTC)

The second deployment option assumes the use of an existing FTTC network to connect mobile stations to the core network. In this scenario, mobile base stations are directly connected by optical cables to the existing FTTC distribution hubs (see Figure 4). We assume here that the distribution hubs have sufficient capacity, enough space and free ports to connect base stations.

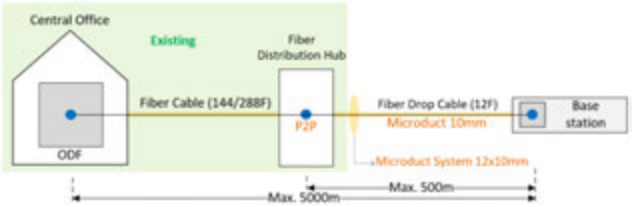


Figure 4. Schematics of the E-FTTC deployment option (P2P: Point-to-Point, ODF: Optical Distribution Frame)

C. Dedicated optical network as a backhaul for cellular networks (D-FTTH)

For the purpose of comparison with the converged deployment options, a dedicated optical network for the mobile base stations is considered and modeled. We assume here direct fiber connections of mobile station to a dedicated fiber network. The fixed internet access is provided through a different FTTH network. It is of particular interest to determine whether a converged infrastructure or dedicated networks is more preferable regarding both the implementation flexibility and cost-efficiency.

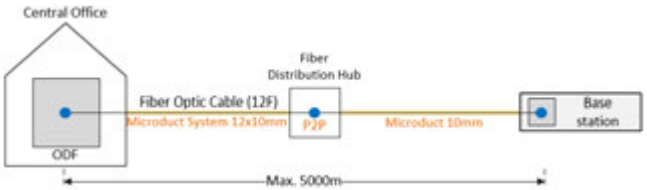


Figure 5. Schematics of the D-FTTH option (P2P: Point-to-Point, ODF: Optical Distribution Frame)

V. RESULTS AND DISCUSSION

A. *Analysis of the roll-out of the converged wireless-optical network (C-FTTH)*

We first analyze the roll-out of a converged P2P-FTTH network infrastructure. The modeled network has 41 distribution hubs within the considered development area that provides certain flexibility in planning fiber connections. The feeder network between the central office and the distribution hub comprises 47 pipes. The length of the feeder section is estimated to be about 38 km, which requires putting a large effort in construction work to lay fiber cables. Consequently, the main cost of the FTTH roll-out is related to the construction work and contributes by 66.8% to the overall implementation cost as indicated in Figure 6. Other significant cost factors are the cost of optical cables with 14.6%, the ducts costs with 12.1% and that of implementing ODFs with 2.8%. The residual components account for additional costs that lie below 1%, giving that they have no potential for eventual cost savings. Assuming 13,833 demand points within the coverage area, the estimated total implementation cost is €5.94 million so that the cost per subscriber accounts to €430.

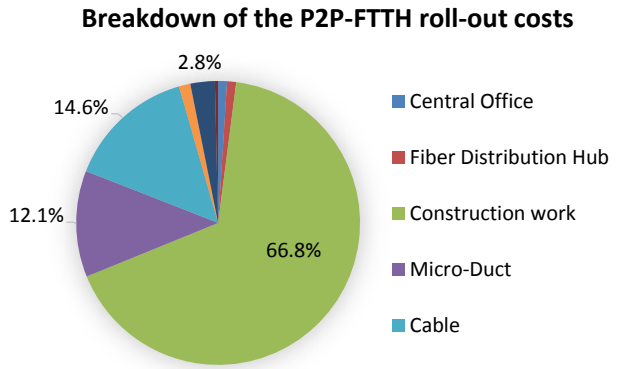


Figure 6. Cost breakdown of the converged P2P-FTTH roll-out

Although an area-wide roll-out of a converged P2P-FTTH network can potentially offer maximum flexibility and performance because dedicated fibers are used, this implementation option requires a high management effort in the CO. As an alternative to the P2P-FTTH network, we also modeled a P2MP FTTH-Network using optical splitters with a 1:32 split ratio. The use of optical splitters in distribution hubs has proven to be more efficient and a certain level of flexibility can still be kept in the distribution network. As a result, the relative cost of ODFs can reduced to less than 1%. The cost breakdown of the P2MP-FTTH roll-out is shown in Figure 7. The use of optical splitters has a strong effect on the number of required ducts and cables in the feeder network. The cost of the ducts can be consequently reduced from 12.1% to 7.1% and that of the cables from 14.6% to 4.7%. The cost of the construction work remains the same, but now it accounts for 83.3% of the total cost.

As a consequence, the total cost of the converged FTTH roll-out can be reduced from €5.94 million to €4.76 million by using optical splitters in the distribution hubs. Thus, a total savings potential of 20% can be achieved as indicated in Figure 8.

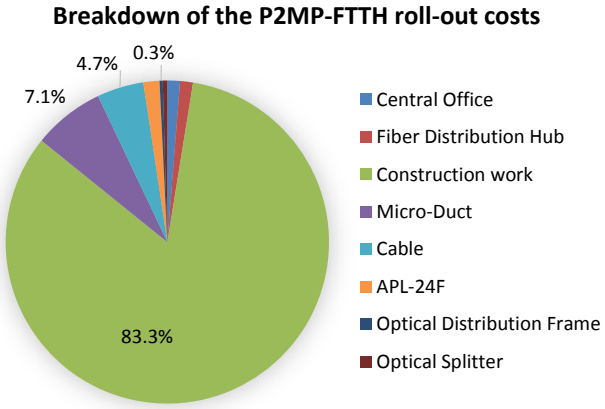


Figure 7. Cost breakdown of the converged P2MP-FTTH roll-out

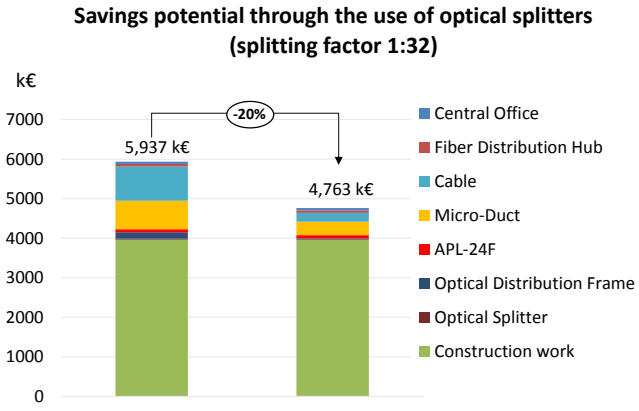


Figure 8. Total savings potential of the converged P2MP-FTTH roll-out in comparison to the P2P-FTTH.

B. *Comparison between C-FTTH, E-FTTC and D-FTTH options*

As already mentioned above, in a traditional network planning, the fixed and mobile access infrastructures has mostly been planned and implemented separately. To investigate whether the combined planning of fixed and mobile network infrastructure has potential to reduce the CAPEX of a common fiber access, we analyzed the deployment options for fiber-based access networks in separate and combined manners and considering an implementation of optical backhaul for future radio access networks deploying both macro and small cells

The results show that using the combined planning approach instead of a separate deployment the implementation of the backhaul for macro cells promises a cost reduction of 8% and for small cells 19% (see Figure 9). As expected in both cases, the highest costs are those of the construction work. It becomes clear that in the converged network, there is less need for construction work because the infrastructure is already available almost everywhere through the fixed network. This becomes clearer as the number of demand points increases.

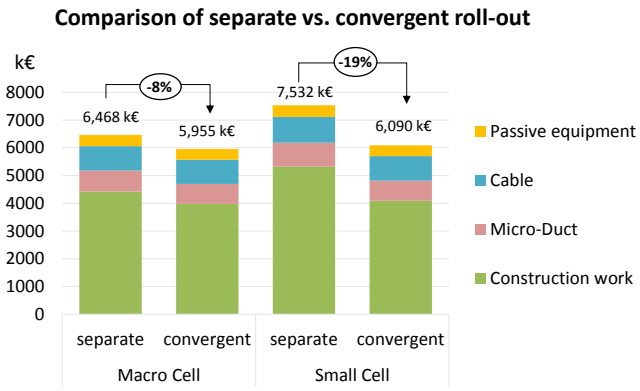


Figure 9. Comparison of the separate and convergent roll-out of FTTH networks considering the backhaul for macro and small base stations.

Finally, we investigated how much additional effort is needed to implement the backhaul for mobile base stations in comparison to an implementation of a fixed FTTH access only. Our results have shown that that the cost of connecting not only macro cells but also small cells using the C-FTTH option is much lower in comparison to the two other options (E-FTTC and D-FTTH). As can be seen from Figure 10, all cost factors can be reduced by a combined planning of the FTTH infrastructure. The converged infrastructure also provides high flexibility and is future ready. It is open for upgrades to higher capacities and a potential efficient operation that can be achieved through an easier reconfiguration and resource allocation between the fixed and mobile parts as well as different distribution areas.

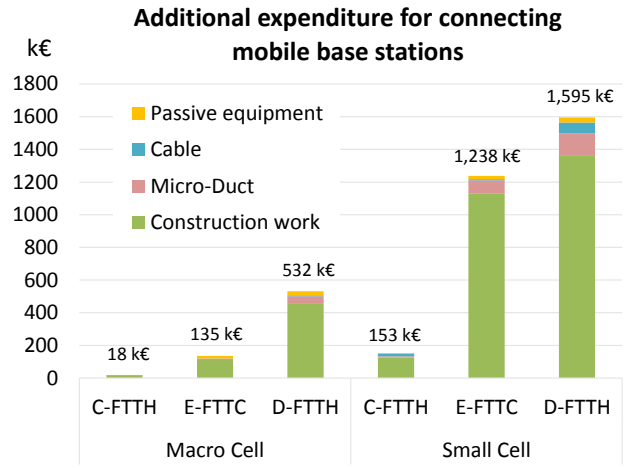


Figure 10. Comparison of the additional expenditure for connecting mobile base stations

VI. Summary and Conclusion

This paper examines deployment options for converged optical-wireless access networks. The analysis is based on the results of a case study for an exemplary urban development area within the city of Cologne. Although some other studies have already considered converged networks and combined planning of fiber-to-the-home (FTTH) and 5G radio access infrastructures, extended fiber-to-the-curb (FTTC) networks and greenfield deployments are not sufficiently investigated. For this reason, we considered three different deployment options including i) a converged optical-wireless point-to-point (P2P) and point-to-multipoint (P2MP) FTTH network, ii) an extension of an existing FTTC network to connect 5G base stations, and iii) a dedicated FTTH network as a backhaul for 5G radio access network. These three deployment options were analyzed with each other with regard to their economic viability and keeping in mind the requirements of 5G applications.

The results show that the convergence in future access networks has the potential to reduce both CAPEX and OPEX. Convergence also provides an improvement in flexibility. Traditionally, cost-effectiveness has always been a primary goal in planning and deploying access networks. However, current trends such as the high requirements for capacity and flexibility driven by novel network technologies and concepts such as network virtualization and softwarization, network programmability, edge networking, and cloud radio access network (C-RAN) as well as the continuous increase of the number of base stations have led to the need for many new fiber optic connections and more flexibility of the fiber optic infrastructure. For this reason, fiber-optic access networks must be dimensioned in a flexible and future-oriented manner, as passive network infrastructures will need to provide not only cost-efficiency but also maximum flexibility to support future developments in order to make the high investments worthwhile.

The additional costs for connecting mobile base stations are lowest in the C-FTTH option, especially if the number of demand points increases. This becomes clear when considering the case of connecting small cells, in which the additional costs are only 3%. This means that a high CAPEX is not to be expected in the future when implementing additional services such as smart city with a converged FTTH network, as there will be the need for relatively little construction work and this is a major cost driver. Furthermore, the infrastructure can be used by several operators and applications so that the CAPEX can be divided among several cost units. To further increase the flexibility of the network, the additional cost of approx. 20% for a throughout P2P infrastructure may be considered. This would provide absolute flexibility to implement arbitrary network configuration or application in the future and would eventually enable a significant long-term reduction in the OPEX of the network infrastructure.

References

[1] R. Gaudino, "Overview on Optical Fronthauling Technologies for Fixed-Mobile Convergence", STAMPA. - (2018), pp. 99-108, Online available https://www.5gitaly.eu/2018/wp-content/uploads/2019/01/5G-Italy-White-eBook-Fronthauling-technologies.pdf

[2] FTTH Council Europe, "Fixed- Mobile Network Convergence: The Key Role of Fibre", white paper by the deployment & operations committee, pp. 1 – 29, March 2019

[3] S. Gosselin et al, "Fixed and Mobile Convergence: Which Role for Optical Networks?", Opt. Commun. Netw., vol. 7, no. 11, pp. 1075-1083, Nov. 2015

[4] J. Salgado, R. Zhao & N. M. P. Inovação, "New FTTH-based technologies and applications", FTTH Council Europe, A White

Paper by the Deployment & Operations Committee, pp. 1 – 16, 2014.

[5] S. Aleksic, „Towards fifth-generation (5G) optical transport networks" (invited), 17th International Conference on Transparent Optical Networks (ICTON 2015), Budapest, Hungary, paper We.A2.2, July 2015

[6] D. T. Eckstein, N. Zhelev, I. Salihbegovic und E. Hadzimujagic. "Unleashing PON full potential – connectivity for all use cases", Detecon International GmbH, pp. 1 – 28, Feb. 2019

[7] C. Behrens, S.Krauß, E.Weis, D. Breuer, "Technologies for Convergence of Fixed and Mobile Access: An Operator's Perspective", Journal of Optical Communications and Networking, 10(1), A37-A42, Jan. 2018

[8] S. Aleksic, M. Deruyck, V. Vereecken, W. Joseph, M. Pickavet, and L. Martens, "Energy Efficiency of Femtocell Deployment in Combined Wireless/Optical Access Networks", in Computer Networks Journal, ELSEVIER, Vol. 57, No. 5, pp. 1217–1233, April 2013

[9] P. Torres, R. Meersman, W. Shi, "5G's Fiber Foundation: The Costs and Opportunities of Convergence", CommScope Webinar, Sep. 2019

[10] S. K. Patri, E. Grigoreva, W. Kellerer and C. Mas Machuca, "Rational Agent-Based Decision Algorithm for Strategic Converged Network Migration Planning," in IEEE/OSA Journal of Optical Communications and Networking, vol. 11, no. 7, pp. 371-382, July 2019

[11] A. Shahid and C. M. Machuca, "Dimensioning and Assessment of Protected Converged Optical Access Networks", in IEEE Communications Magazine, vol. 55, no. 8, pp. 179-187, Aug. 2017

Resource utilization – QoS isolation trade-off in 5G networks considering network slicing reconfiguration interval

Saša Demirović, Paula Kiš, Jasna Janković, and Željko Ilić
University of Zagreb, Faculty of electrical engineering and computing, Croatia

e-mail: {sasa.demirovic, paula.kis, jasna.jankovic, zeljko.ilic}@fer.hr

***Abstract* - 5G network is expected to support a wide range of use cases classified into three service types, eMBB, URLLC, and mMTC, with possibly conflicting QoS requirements and different traffic characteristics. The network slicing concept enables physical and logical separation of network resources and, therefore, QoS isolation. Network slices guarantee QoS requirements by allocating resources in a way that balances both demand and availability while considering resource scarcity. However, it is challenging to achieve effective end-to-end QoS isolation of network slices while maintaining resource utilization efficiency, which we call resource utilization – QoS isolation trade-off. Resource utilization is a performance metric that gives feedback on how efficiently network resources are allocated to slices. Considering the QoS policy of each slice, QoS isolation describes the ability of each slice to fulfill its QoS requirements. In this article, the goal is to describe how different network slicing reconfiguration interval affects the described trade-off and how to adapt it for the mentioned service types. We show through simulations that maximizing resource utilization decreases the QoS isolation and vice versa. As slice resources are granted for a fixed time window, i.e. reconfiguration interval, our article shows how its variations impact both resource utilization and QoS isolation.**

Keywords - 5G, network slicing, resource utilization efficiency, QoS isolation

I. INTRODUCTION

Smartphone usage trends and the appearance of smart devices forming massive IoT networks prompted an urging demand for dedicated mobile services provisioned on a flexible multi-service network. The existing networks based on "one-size-fits-all" architecture cannot accommodate diversified services of very distinct and possibly conflicting requirements. Therefore, the fifth generation of mobile networks (5G) is expected to support a wide range of use cases and vertical industries. A new concept, called network slicing, enables administering multiple virtual networks on top of a single shared physical infrastructure. Thus, network slices provision services of different requirements, and to fulfill those, network resources need to be efficiently allocated among them [1].

Network resources are allocated to slices considering various service requirements and time-varying traffic characteristics. As limited resources are available, especially in the radio access network, resource overprovisioning is impossible, so there is a need for effective allocation, management, and control of resources. Consequently, optimal resource allocation for network slices is the key to balance traffic loads, improve network performance, and guarantee demanding Quality of Service (QoS). This requires innovative algorithms and solutions that will allocate resources among slices while at the same time attempting to maximize the overall network utilization, whereby the allocation reconfiguration interval could be of significant impact.

To analyze the impact of the reconfiguration interval, we consider resource allocation to both slices and users in each slice. Intra-slice resource sharing is based on proportional fairness rule while inter-slice resource sharing considers static and dynamic reconfigurations. Dynamic resource provisioning of network slices has caught interest over the static solutions as 5G network architecture tends to improve the overall network efficiency. Therefore, slicing schemes should be flexible and elastic to accommodate unpredictable and varying traffic flow. Resource utilization and QoS isolation are the main performance metrics when evaluating different proposals of resource provisioning in the network slicing environment [2]. The challenge is to maximize resource utilization whilst achieving agreeable QoS isolation between slices. A disbalance between user demands and allocated resources can possibly happen, which results in ineffectively utilized slices. To fully exploit network slicing with dynamic resource provisioning it is necessary to optimize resource allocation reconfiguration interval. Slice resources are reconfigured in each reconfiguration interval according to the traffic load and resource availability. Frequent reconfigurations cause high provisioning complexity, and thus higher cost, but at the same time improve resource allocation performance and increase utilization as well as isolation. The question is what is the optimal solution, better performance or less provisioning complexity?

In this paper, we investigate the relationship between resource allocation efficiency and reconfiguration interval, whereby efficient resource allocation means maximized user and satisfying inter-slice isolation. The rest of this

paper is organized as follows. Section II provides a review of related work on resource allocation reconfiguration interval. The proposed allocation schemes and evaluation performance metrics are described in Section III. Finally, simulation results in the form of graphical representations are given in Section IV and the paper conclusions are summarized in Section V.

II. RELATED WORK

Resource allocation in mobile networks has been thoroughly investigated research topic, especially in the last years with the new services and demands imposed by the 5G network. Here we overview the recent works considering allocation reconfiguration interval.

Network slicing for enhanced Mobile BroadBand (eMBB), massive Machine-Type Communication (mMTC), and Ultra-Reliable Low-Latency Communication (URRLC) is investigated in [3]. Therein, eMBB is characterized by large traffic load, low latency, and moderate reliability of Packet Error Rate PER<10^{-3}. Its requirements include maximum data rate and a stable connection. On the other hand, mMTC requires diverse connectivity support and relatively low reliability of PER<10^{-1}. mMTC traffic consists of many intermittent transmissions of small data payloads by Internet of Things (IoT) devices which are sporadically active. Like mMTC, URLLC also supports intermittent transmissions with small traffic load. Moreover, URLLC requires low latency, time localized transmission, diverse connectivity, and huge reliability of PER<10^{-5}. Because of the diverse service requirements, the authors conclude that the resources must be shared in random processes.

Mixed timescale Radio Resource Management (RRM) framework described in [4] is based on two timescales: slow, second/minute timescale, and fast, millisecond timescale. The proposed mixed timescale RRM outperforms both mentioned timescales in all traffic conditions resulting in reduced delay and relatively low complexity.

The most related work to ours reviews the trade-off between service customization, resource management efficiency, and system complexity [5]. System complexity is the consequence of advanced dynamic allocation algorithms that achieve higher efficiency. The conducted analysis specifies cases in which a specific resource sharing algorithm is worth the sacrifice in efficiency, customization, or isolation. Resource management is observed in two real network infrastructures with 38 different services popular in late 2016. Four use cases are represented, the first is worst-case setting with strict slice specification and no reconfiguration, second is also static allocation but with slice specification impact, the third case is dynamically re-allocated slices and finally impact of a varying number of slices. In the worst-case efficiency is extremely low, approximately 15%. These are the most useful conclusions from paper [5]. Slicing in higher network layers provides effective resource sharing with no additional complexity, but full customization cannot be offered. *Traffic direction is a factor*, meaning that downlink has better efficiency for the same QoS requirements. Reconfiguration interval in a case with the dynamically allocated resource is 30 minutes long, however, multiplexing efficiency is decreased as the resource assignment interval grows. Significant gain is acquired when services with similar requirements are aggregated.

III. SYSTEM DESIGN

We consider a system model which consists of three main entities, varying slice load, slice QoS requirements, and resource provisioning and reconfiguration.

The slice load defines the amount of resources generated and needed by users of a slice. It can represent any type of resources in the mobile network. Furthermore, as resources are completely abstracted, our model of resource allocation can be applied to any part of the network, whether radio or core. As an example, users could need processing resources in a MEC (Multi-Access Edge Computing) scenario or a share of spectrum for wireless transmission. The total slice load is generated based on traffic characteristics of each user. We envision users of same traffic characteristics and QoS requirements to be grouped in the same slice. User traffic characteristics are devised based on 5G service types, thus possible slices being eMBB, mMTC, and URLLC [6]. Therefore, we model each 5G service type traffic load with packet arrival time distribution and packet size following the traffic model in [7]. The expected total system load is then aggregated based on eMBB, mMTC, and URLLC user traffic load and arrival intensity. QoS requirements are demands on service quality provided by slice to each user. As well, requirements are predefined and known in advanced. Usually, they can be defined in terms of maximal processing time of a task, user data rate or minimal number of users provided with the service. As our system model is abstract and resource demands can be interpreted in any suitable way, we define QoS requirements as time spent from the moment the request for resources appeared until the moment the resources were fully provided.

The goal of resource slicing is to divide resources among slices in a way that best satisfies users QoS constraints. For the resource provisioning and reconfiguration procedure we first consider non reconfigurable, i.e. static resource sharing, meaning that resource reconfiguration interval is infinite. In static resource allocation each slice has constant share of resources. Initially, when no real traffic data is considered, the amount of resources provisioned for each slice is calculated by expected traffic load of each slice. In the both static and dynamic slicing scenario, resources allocated to a slice in each interval cannot be used by another slice even if the resources are not used by the slice they were allocated to. Intra-slice resource allocation is based on a proportional fairness rule, the idea is that every packet within a slice is given an equal amount of resources to ensure fairness among users of a slice. Therefore, resource availability changes with arrival and provision of packages. To cope with the traffic variability, reconfigurable allocation scheme is necessary.

Reconfigurable allocation or the dynamic resource allocation implies readjustment of the share of resources among the slices in fixed interval, repeated periodically. Figure 1 shows our example of dynamic allocation based on fairness rule.

The goal is to provision resources among the slices efficiently and to achieve accurate prediction of future network traffic. In each reconfiguration interval we employ an algorithm to approximately predict the traffic load of each slice for the next allocation interval in order to fairly and efficiently divide the available resources among slices for the next interval. To anticipate the upcoming traffic in each slice for the next interval and to accomplish resource allocation efficiency, we combine average expected traffic load with the traffic provisioned in the previous interval to determine new reconfiguration decision. Then, we investigate the reconfiguration interval impact on resource utilization and QoS isolation.

Due to variety of heterogeneous services, it is challenging to fulfill users QoS requirements. QoS isolation measures the ability of network to maintain those QoS requirements in presence of other traffic. We devise QoS isolation of a slice from the slice QoS performance. QoS performance for each packet achieved with dynamic provisioning, $P_{p,d}$, is defined as the ratio between packet provisioning time and maximal packet delay:

$$P_{p,d} = \frac{t_{e,p} - t_{a,p}}{D_p} \qquad (1)$$

where $t_{a,p}$ is the packet arrival time, $t_{e,p}$ the packet provisioning end time, and D_p the maximal packet delay. Maximal delay is QoS requirement predefined for each user. QoS performance is therefore satisfying if the ratio is less than 1, meaning that required maximal delay is greater than the total provisioning time of the packet. Further, QoS isolation I_p is based on QoS performance and defined as normalized difference of QoS performance achieved with static and dynamic provisioning:

$$I_p = \begin{cases} 1, & if\ P_{p,d} \geq P_{p,s} \\ 1 - \frac{P_{p,s} - P_{p,d}}{P_{p,s}}, & otherwise \end{cases} \qquad (2)$$

where $P_{p,s}$ is the QoS performance achieved with static provisioning. Therefore, if the dynamic provisioning QoS performance is better or equal to the static provisioning one, the QoS isolation is interpreted as full, 100% isolation of resources. QoS isolation implies that service performances are preserved, meaning that QoS of one slice is not affected by the other slice [8]. Resource utilization U describes the efficiency of resource allocation or how well the resources are utilized. It is calculated as percentage of resources utilized by all slices in an allocation interval on average:

$$U = \frac{\sum_{s \in S} r_{p,s}}{r t_p} \qquad (3)$$

where $r_{p,s}$ are all resources provisioned to slice $s \in S$, r the resources available in each allocation interval, and t_p the provisioning time.

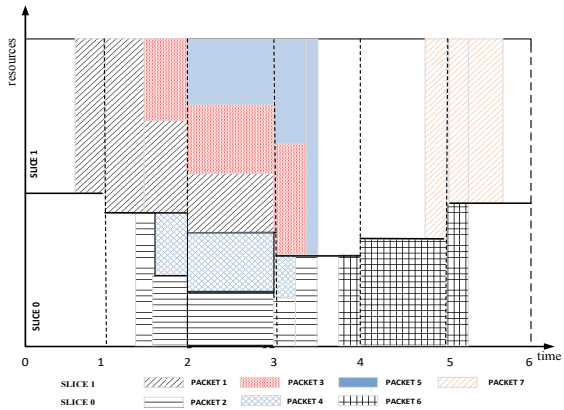


Figure 1 An example of dynamic resource provisioning to slices based on fairness rule

IV. Simulation Results

In our simulation scenario, we examine the behavior of the resource utilization–QoS isolation trade-off by changing the reconfiguration interval from static slicing case where there is no reconfiguration to highly responsive reconfiguration interval in order of time less than *ms*. In this section, we present the simulation results to examine the described trade-off.

A. Algorithm consideration

In this subsection, we evaluate the influence of different resource slicing algorithms on overall performance of the presented dynamic resource allocation. To optimize resource slicing decision, we considered four algorithms of different approaches. The main difference between these algorithms is that they are predictive or past oriented. The first algorithm, named *Load*, allocates the resources among slices in the next reconfiguration interval based on current load of each slice. The percentage of allocated resources for each slice in the next reconfiguration interval is calculated as the load share of the slice in total load of all slices in current interval. If there is no traffic load in current interval, the percentage is calculated as in static allocation. The second algorithm, *LoadExpected* algorithm, includes predefined traffic demands of each slice, the load division among slices for the next reconfiguration interval depends on the current load and average expected load in the next interval. The third and fourth algorithm, *Predict* and *LoadPredict,* respectively, simulate predictive algorithms. The percentage of resources allocated to each slice in *Predict* algorithm is calculated as the slice load share of traffic in the upcoming reconfiguration interval for each slice[1]. In *LoadPredict* algorithm, the percentage of allocated resources to each slice depends on upcoming and current load of each slice. To evaluate each algorithm's efficiency, we monitor the behavior of the previously explained metrics P_p, I_p, and U. There are two slices in the system with 100 users, where each slice provisions 50 users of similar service characteristics. The services are grouped into three service types: eMBB, mMTC, and URLLC. Their typical parameters are shown in Table 1.

[1] It is possible to know the traffic load in advance because it is simulation, in real system a real predictive algorithm should be employed.

TABLE I. TYPICAL PARAMETERS FOR EMBB, MMTC AND URLLC SERVICE TYPE

Service type	Packet size	Arrival rate	Max. delay
eMBB	100 bytes	100 packet/s	10 ms
mMTC	10 bytes	10 packet/min	100 ms
URLLC	10 bytes	10 packet/min	2 ms

First, we examine the scenario of eMBB slices with packet arrival rate λ = 0.1 *packet/ms* and 17700 abstract *resources/ms* provisioned in the system and being divided for both eMBB slices. In each simulation, the service type (i.e. eMBB) and the overall resources available are the same but the resources are differently allocated among slices in intervals based on described algorithms. We have to remark that the utilization of these algorithms can cause additional complexity and may interfere isolation among slices.

Figure 2 shows how resource utilization changes with length of the reconfiguration interval (RI). Resource utilization for static allocation (shown with red dashed line) is acquired over total provisioning time, and since the algorithm does not affect static allocation, it remains constant at 0.215. We can conclude that resource utilization for shorter reconfiguration intervals is higher. It can be noted that all algorithms excluding the first one, which is based only on the current traffic load, behave in the same way and show better resource utilization than the static allocation. The *Load* algorithm is not able to accurately reconfigure resources in intervals longer than 5*ms*. We confirm the assumption that the best utilization is achieved for the algorithm which combines the future and the current traffic load. The legend shown on Figure 2 applies to Figures 3 and 4.

Similar behavior can be observed in Figure 3, where the average QoS performance of slices is shown. QoS performance equal to 1 or below means the packet arrived before the deadline delay on average. For longer intervals the QoS performance decreases. Again, the algorithm which combines the future and the current traffic load shows the best overall performance. QoS performance of the algorithm *LoadPredict* goes below 50% for intervals shorter than 15*ms*. As displayed in Figure 3, the QoS performance of the first algorithm, *Load*, goes up to 4.5, and that is extremely bad performance. For all intervals between 5 and 250 *ms* the QoS performance of the first algorithm is worse than using only static allocation, with the delay being double or more.

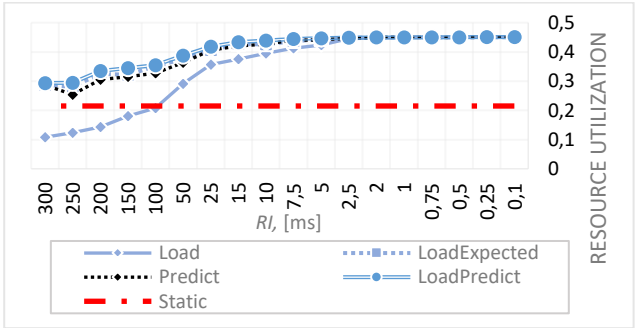


Figure 2 Resource utilization for different slicing algorithms (eMBB traffic with λ=0.1 $\frac{packet}{ms}$)

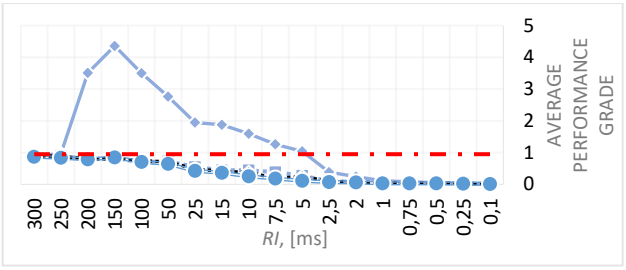


Figure 3 Performance behavior for different slicing algorithms (eMBB traffic with users λ=0.1 $\frac{packet}{ms}$)

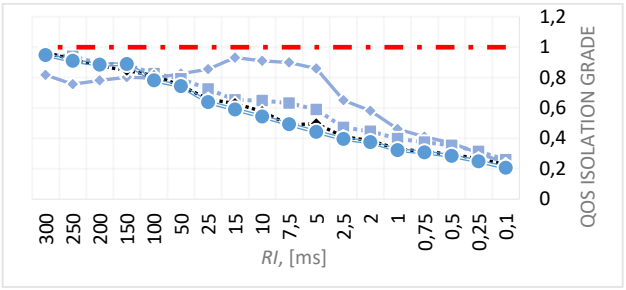


Figure 4 QoS isolation grade for different slicing algorithms (eMBB traffic with λ=0.1 $\frac{packet}{ms}$)

Figure 4 offers an overview of QoS isolation metric. The only difference from the previous graph is the fact that for large intervals the *Load* algorithm achieves better isolation than other algorithms but as the intervals get smaller its ability to adjust again sinks and QoS isolation gets worse. The performance of the *Load* algorithm does not deviate from other algorithms as for resource utilization and QoS performance. QoS isolation for each algorithm is worse than the QoS isolation of static slicing.

B. Traffic consideration

In this subsection, we describe how resource allocation effects the slice performance based on different service provisioned. In each case we use the already mentioned *Predict* algorithm. First, we examine eMBB service with different traffic loads which change with arrival rate. Arrival rate λ describes how many packets arrive in a time unit. Figures 5 – 7 offer a performance overview of traffics with 5 different λ. It is important to note that the usual λ for eMBB service is considered 0.1 *packet/ms*, meaning that 0.1 packets per millisecond arrive on average, i.e. 100 packets arrive in a second. The other arrival rates considered are λ = 0.025, λ = 0.05, λ = 0.25, and λ = 0.5 *packets/ms*. By increasing λ, the traffic load increases too, meaning that more resources are needed for higher λ. Resources provisioned for the usual eMBB service of λ = 0.1 *packet/ms* are 17700 and we adapt the amount of resources to match different traffic loads proportionally.

Figure 5 shows resource utilization over time interval for five eMBB traffics with different λ. We can see that resource utilization gets better when the reconfiguration interval is shorter, meaning that with smaller interval the allocation decision is better adapted to traffic. When RI = 0.1 *ms*, all eMBB traffic has the same resource utilization equal to 0.45, that is the maximal resource

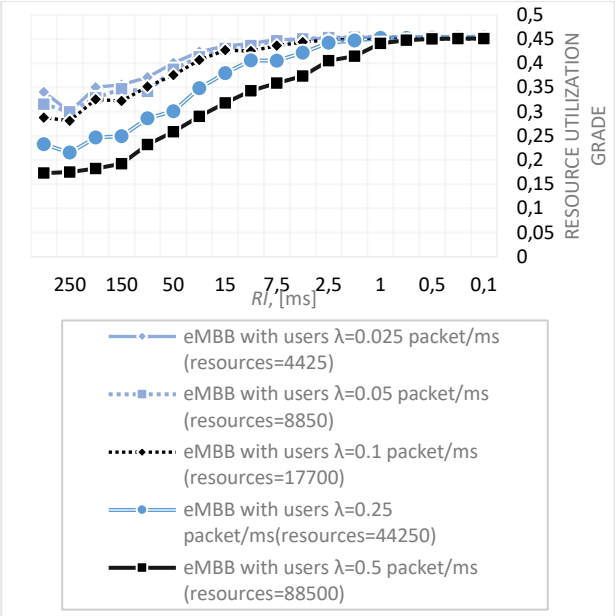


Figure 5 Resource utilization metric of eMBB traffic with different arrival rate

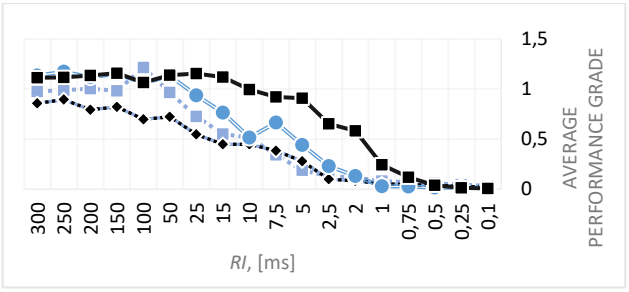


Figure 6 Average performance metric of eMBB traffic with different arrival rate

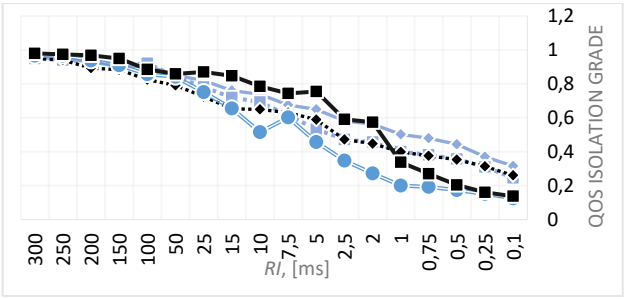


Figure 7 QoS isolation metric of eMBB traffic with different arrival rate

utilization, meaning that 45% of each interval is utilized on average. As we increase the reconfiguration interval towards 500 *ms*, the dynamic allocation starts to behave like static allocation because the reconfiguration frequency decreases. For longer reconfiguration intervals the resource utilization is constant. The utilization gets better as λ gets lower, i.e. utilization is better for smaller traffic load. Longer reconfiguration interval improves more the slow-varying eMBB traffic. The best QoS performance has the eMBB traffic with λ = 0.025 *packets/ms*. The resource utilizations on Figure 5 confirm the intuition that the network is better utilized when we reconfigure the allocation decision more often. The traffic adjusts worse for very long reconfiguration intervals. When considering short reconfiguration interval, complexity of the resource provisioning must be considered since new provisioning decision must be made rapidly. The legend shown on Figure 5 applies to Figures 6 and 7.

Figure 6 portrays average QoS performance over time reconfiguration interval. QoS performance is satisfied if it is below 1, and we can see that the QoS performance for all traffic after RI = 150 *ms* oscillates around 1, so we conclude that QoS performance in static allocation is not so satisfactory. The worst QoS performances are achieved for the traffic with the highest arrival rate. With very short reconfiguration intervals (<1 *ms*) resources are very well adjusted and the average performance gets better with average provisioning time being more than five times smaller than the expected deadline. The steepest QoS performance behavior has the traffic with λ = 0.25 *packets/ms*.

The QoS isolation results are illustrated in Figure 7. As the reconfiguration interval increases the QoS isolation weighs toward 1, meaning that it gets equal to the QoS isolation of static slicing. For very small reconfiguration interval, QoS isolation is very low, for λ = 0.5 *packets/ms* QoS isolation is 0.15, and for λ = 0.025 *packets/ms* QoS isolation is 0.35. Hence, as the arrival rate and the traffic load increase, the QoS isolation gets lower. It is noticeable that the reconfiguration interval change affects more traffic with higher λ, i.e. the traffic with greater load.

As it can be observed from Fig. 5 and Fig. 7 there is a clear relation between QoS isolation and resource utilization considering different reconfiguration intervals, which can be described as trade-off because QoS isolation increases with longer intervals while resource utilization grade decreases for longer reconfiguration intervals. To achieve optimal network performance, it is crucial to find the appropriate reconfiguration interval for which both are satisfying. For different eMBB traffics we estimate the reconfiguration interval of 5 *ms* to be feasible to adequately balance resource utilization and QoS isolation. At RI = 5 *ms*, average resource utilization is 0.44, meaning that 44% of available resources per each interval are used, and average QoS isolation for RI = 5 *ms* is 0.6, after which the average QoS isolation starts to fall very fast. Also, very long and very short intervals can be good solution for slow-varying traffic, but for fast-varying traffic we need to consider mean values of reconfiguration intervals for higher overall performance. Figure 8 shows common QoS performance and QoS isolation grade of averaged URLLC and averaged mMTC traffic with equal amount of resources allocated using *Predict* algorithm.

The resources required to achieve similar performance for mMTC and URLLC are substantially smaller than the resources required for eMBB due to the fact that the eMBB packets and arrival rate are larger. Performance metrics indicate that the QoS performance of mMTC is greater than the URLLC QoS performance. URLLC service requires low latency, meaning that allowed delay is very low, and on the other hand, mMTC has larger delay allowed. The smaller the allowed service delay, the worse the performance. As already stated, QoS performance gets

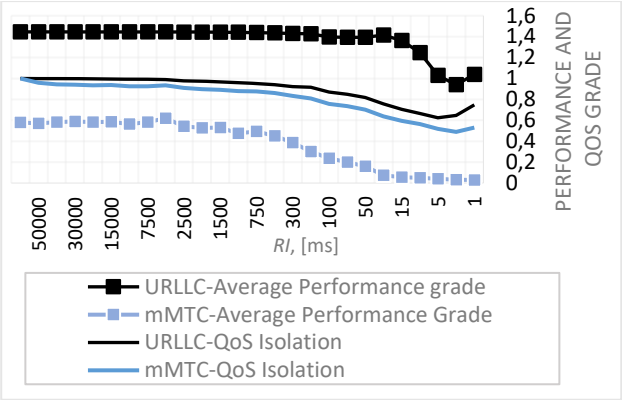


Figure 8 Average performance and QoS isolation of averaged mMTC and averaged URLLC traffic

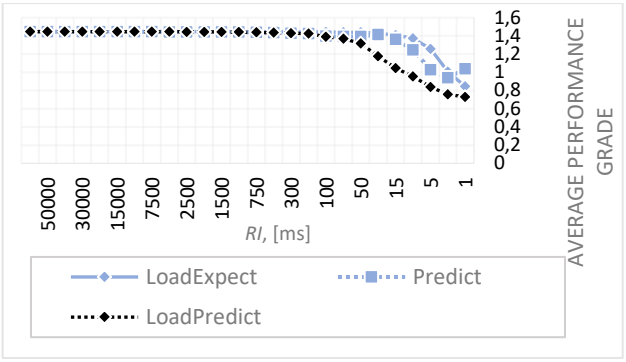


Figure 9 Different slicing algorithms for URLLC traffic

better for shorter reconfiguration intervals. For longer intervals the performance of both, mMTC and URLLC, is constant. mMTC QoS performance is constant until RI = 15000 *ms*, but URLLC QoS performance is constant even until RI = 300 *ms*, and that implies that URLLC slice is overprovisioned. The behavior of QoS isolation is analogical to the QoS performances of mMTC and URLLC but the difference between the performances is quite small due to the calculation algorithm which evaluates any dissatisfactory performance with 1 meaning the large and small excess are evaluated the same. In order to explain the behavior of URLLC QoS performance at RI = 5 *ms*, we need to reconsider Figure 9. Figure 9 shows behavior of average QoS performance of URLLC service for different algorithms, *LoadExpect*, *Predict*, and *LoadPredict*. The *Predict* performance bends at RI = 5*ms*, and the performance is the lowest at that reconfiguration interval, but then, unexpectedly, gets higher at RI = 1 *ms*. The *Predict* algorithm cannot efficiently allocate resources in short reconfiguration intervals because it operates on fact that the past traffic is resolved and only future traffic is considered, but because the interval is too short for packets to be fully provisioned in the interval they arrived in, the algorithm does not perform well, the packets are piling up and causing traffic congestion. *LoadExpect* and *LoadPredict* performances fall exponentially for shorter reconfiguration intervals.

V. CONCLUSION

In this work, we investigated the resource utilization – QoS isolation trade-off in network slicing environment considering resource allocation reconfiguration interval. Proposed resource allocation simulation is based on completely abstract resources and as such it can be applied to any part of network. Simulation results show that resource utilization increases as length of reconfiguration interval decreases, but QoS isolation decreases as the reconfiguration interval gets shorter. The optimal reconfiguration interval considering resource utilization and QoS isolation trade-off, for different eMBB traffic, is 5 *ms* which adequately balances slice performance and resource usage. The overall performance is improved if we consider both the upcoming and the current traffic load, but the computational complexity is then increased.

REFERENCES

[1] S. Zhang, "An overview of network slicing for 5G," IEEE Wirel. Commun., vol. 26, no. 3, pp. 111-117, June 2019.

[2] M. I. Kamel, L. B. Le, and A. Girard , "LTE wireless network virtualization: dynamic slicing via flexible scheduling," in 2014 IEEE 80th Vehicular Technology Conference, Vancouver, BC, Canada, 2014.

[3] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: a communication-theoretic view," IEEE Access, vol. 6, pp. 55765-55779, September 2018.

[4] F. Teng and D. Guo, "Resource management in 5G: a tale of two timescales," in 49th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 2015.

[5] C. Marquez, M. Gramaglia, A. Banchs, M. Fiore, and X. C. Perez, "How should I slice my network? A multi-service empirical evaluation of resource sharing efficiency," in MobiCom '18, 2018.

[6] ITU, "Recommendation ITU-R M.2083-0 IMT vision – framework and overall objectives of the future development of IMT for 2020 and beyond," Geneva, Switzerland, 2015.

[7] 3GPP, "TR 37.868 V11.0.0, Study on RAN improvements," 2011.

[8] G. Wang, G. Feng, T. Q. S. Quek, S. Qin, R. Wen, and W. Tan, "Reconfiguration in network slicing – optimizing the profit and performance," IEEE Trans. Netw. Service Manag., vol. 16, no. 2, pp. 591 - 605, June 2019.

Bounding Reliability in Service Function Chaining

Anna Engelmann, Wolfgang Bziuk and Admela Jukan
Technische Universität Braunschweig, Germany
Email: {a.engelmann, w.bziuk and a.jukan}@tu-bs.de

***Abstract*—In Network Function Virtualization (NFV), the reliability of Service Function Chain (SFC) is one of the most critical performance parameters. Since each SFC has different composition, placement, configuration and deployment requirements, the exact analysis of SFC reliability is a challenge. In this paper, we address this challenge by analyzing the end-to-end service reliability as a combined function of SFC placement strategy over Inter- and Intra-DCNs and failures of hardware and software components. The analysis introduces a novel reliability benchmark, the so-called *Affordable Component Failure* which can take into account the interdependency of different component failures, i.e., failure of data centers, racks, servers and virtual network function as well as their impact on availability of other SFC components and the overall SFC reliability.**

***Index Terms*—NFV, SFC, reliability, availability, backup VNF**

I. Introduction

Today, NFV enables decoupling of network functions from the physical hardware utilizing virtual network functions (VNFs), a software instances running on virtual machines (VM) hosted by physical machines. In NFV, an end-to-end service is presented by a chain of VNFs distributed over Inter- and Intra- Data Center Networks (DCN), referred to as Service Function Chain (SFC). The reliability of a SFC is a critical performance parameter affected by processing vulnerability, failures caused by hardware or software, which can break down an entire service chain. Depending on reliability requirements, an SFC needs to consider failure protection, based on expensive redundant configurations and network traffic redirection. However, assessing an SFC reliability is a major challenge, as one needs to consider co-location and sharing of hardware components, heterogeneity of software and hardware and complex interdependencies between them [1].

Due to complexity of the problem, previous methods to analyze SFC reliability considered reliability of individual system components, such as servers, VNFs, or switches, with the goal of optimizing the SFC reliability value by solving the *reliable SFC mapping problem* [2]–[5]. In addition, the VNF placement optimization problem in combination with reliability analysis was used without considering the interdependency between failed and remaining available components [6], [7]. However, failures of individual components involved in SFC are not independent, whereby ignoring of components' interaction during reliability prediction can be fatal as any single failure can become a cause of failure of entire end-to-end service. For instance, when a requested SFC is placed inside the same rack, even if on different servers, the loss of connectivity to a Top-of-Rack (ToR) switch results in failure of the entire SFC and, thus, service interruption.

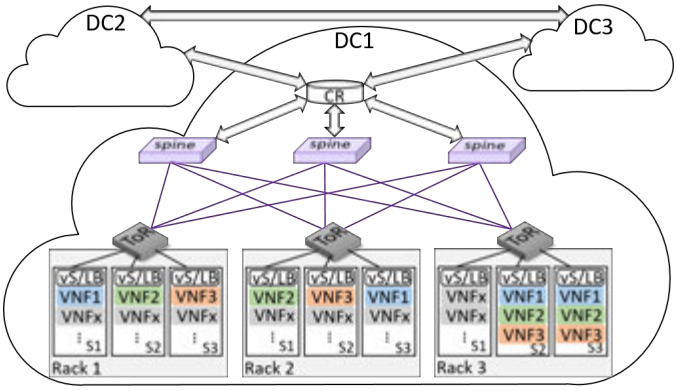


Fig. 1. DCN based on leaf-spine architecture based on [8], [9].

In this paper, we analyze the SFC reliability as a function of VNF placement strategies in generic inter- and intra- Data Center Networks (DCNs) with backup protection. For the first time, we consider VNF chaining where active and backup racks, servers, VNFs and even entire Data Centers (DCs) can interdependently fail and thus impact the availability of other components and the overall end-to-end service reliability. To this end, considering heterogeneity of components involved, we first introduce the level of disjointedness and the amount of backups for each component type and analyze the so-called *Affordable Component Failures* as a reliability benchmark of the system. Finally, we derive the SFC reliability comprehensively taking into account all possible combinations of VNF placements over inter- and intra-DCNs. The results show that our SFC reliability analysis can help find the best VNF placement strategy that can be achieved for a required service reliability value with the lowest level of redundancy.

II. Preliminaries

A. Reference Architecture of Data Center Networks

The reference DCN architecture is presented in Fig 1. The core router (CR) provides connectivity to other remote DCs and to other parts of own data center. The switch interconnect can be realized as a CORD's leaf-spine architecture, where a series of leaf switches, i.e., Top-of-Rack (ToR), are fully-meshed with spine switches. ToR switches handle forwarding within a rack, while forwarding across racks is performed by spine switches. Generally, each rack incorporates its own subnet of servers. Each server hosts a virtual switch (vS) and multiple VNFs of the same or different types. All VNFs in a certain server are connected to a vS, which provides a

load balancer (LB). The main task of vS/LB is to select a legitimate VNF instance for any arrived request and to provide a connectivity to ToR. In Fig. 1, the ordered sequence of VNFs, i.e., SFC (VNF1-VNF2-VNF3), can be built in the same server, e.g., in S2 in Rack 3, or distributed over different servers, e.g., over S1-S3 in Rack 1 or over S1-S3 in Rack2. There are multiple options to build SFCs over different racks as well, e.g., VNF1 can be utilized in S1(Rack1), VNF2 and VNF3 in servers S1(Rack2) and S2(Rack3), respectively.

B. Component failures

Without loss of generality, we consider VNFs, servers, racks and DCs as atomic components used to build an SFC. Whereas we assume that failure of one component of a certain type does not impact other components of the same type, the failures of one component type can result in failures of other component types. This is a reasonable assumption, since for instance, a failed connection to one DC does not cause connectivity failure to another DCs. In contrast, a failed server will result in failure of all VNFs installed on it. Given these assumptions, the following failures are considered:

1) DC: The failure of a data center can be caused by a failure of connectivity to DC through CR, failure of CR or the connection between CR and all spine switches. DC failure results in failure of all racks, servers and VNFs relevant for building and maintenance of SFC.

2) Rack: A Rack failure can be a result of a failure of its ToR switch, failures of all spine switches or of all links between the spine switches and its ToR. Any rack failure will cause failure of all servers and VNFs within it.

3) Server: Any server failure is a result of hardware failures, i.e., power supply, memory, etc., failed link between server and ToR or due to failure of vS/LB. The failed server causes a failure of all VNFs installed on it.

4) VNF: VNF is equivalent to a virtual machine (VM). Since we assume that each VM allocates one VNF only, the failure of one VNF causes no failure of any other components.

C. Failure Protection

To protect one active SFC, $r \geq 1$ backup SFCs can be used. As shown in [2], to provide a highly reliable service, there are different strategies for an effective deployment of active and backup VNFs. Here, we do not explicitly consider any backup deployment strategy. Instead, we assume an SFC as available, if at least one out of $n = 1+r$ VNFs of any type is available during service run-time. Thus, one of the backup VNFs replaces a failed active VNF of the same type. Any backup component can fail during service run-time. This is the reason to consider $r \geq 1$ backup SFCs, which allow to further increase the reliability of the 1:1 backup strategy used in practice. Generally, we allow hardware and software components to fail, as long as they do not impact the availability of the entire SFC. This assumption is of practical importance, as it can determine the amount of failed components pre-reserved for protection and maintenance of a SFC to keep the same SFC always available. Thus, we refer to the number of failed hardware or software components pre-reserved for protection and maintenance of SFC as *Affordable Component Failures*, if these failures do not lead to interruption of the service. It is obvious that when no failure protection is applied, there are no affordable component failures and all active components, i.e., DCs, racks, servers and VNFs, must stay available during service run-time. In contrast, with a certain level of redundancy, we can estimate how many components are free to fail, without interruption of the end-to-end service.

TABLE I
NOTATION

p_v	VNF reliability value;
p_s	server reliability value;
p_r	rack reliability value;
p_d	DC reliability value;
Ψ	number of VNFs in SFC;
ψ	number of VNFs of the same type allocated together in one server;
r	total number of backup SFCs;
n	total number of pre-reserved SFCs, i.e., $n = 1 + r$;
r_d	total number of DCs with backup VNFs of a certain type;
n_r	number of racks inside a DC containing VNFs of a certain type;
r_r	total number of racks with backup VNFs of a certain type;
n_s	number of servers inside a rack containing VNFs of a certain type;
r_s	total number of servers with backup VNFs of a certain type;
N_r	amount of backup component types;
Δ	number of disjoint component types for a certain VNF type;
A_c	affordable failures of component of type c, $c \in \{d, r, s, v\}$;
Λ_c	number of remaining available components of type c, $c \in \{d, r, s, v\}$;

III. Modeling and Analysis

We define the SFC reliability as a probability that a traffic can successfully traverse one out of $n = 1 + r$ SFCs to complete the requested service, where $r \geq 0$. Since each SFC consists of Ψ different VNFs, at least Ψ VNFs of any type have to be available. For the evaluation of the SFC reliability as a function of placement strategy, we assume that the component reliability is the same for all active and backup components of certain type. The used notations are introduced in Tab. I.

A. SFC and VNF Placement

In our model, the racks and servers inside DCs are allocated to any SFC in the same way so that active and backup SFCs and their VNFs are placed in similar fashion. To place all $n = 1 + r$ VNFs of a certain type, e.g., all VNF1, one active and $r_d \leq n - 1$ backup DCs can be used. Inside each DC, $1 \leq n_r \leq n$ racks, and $1 \leq n_s \leq n$ servers inside any rack are required. For a total number of r backup VNFs of a certain type and a chosen placement strategy, there is a need for a $0 \leq r_d \leq r$ DCs, and $0 \leq r_r \leq r$ backup racks and $0 \leq r_s \leq r$ backup servers, where any server allocates $1 \leq \psi \leq n$ VNFs of the same type. Thus, an active VNF can be replaced by a backup VNF of same type located in the same server or in a different server, rack and in a different DC.

For example in Fig. 1, a SFC consists of $\Psi = 3$ VNFs. Let us first assume $n = 2$ SFCs, i.e., $r = 1$ backup SFC, and only consider the Rack 1 and Rack 2 (Rack 3 is not deployed). In this case each server allocates only $\psi = 1$ VNF of same type, e.g., in Rack 2 one VNF1 in S3, one VNF2 in S1 and one VNF3 in S2. Furthermore, let us assume that the active SFC is placed in Rack 2, and that the backup SFC is placed in the

same fashion in Rack 1. However, all active and backup VNFs of a certain type, e.g., VNFs of type 1 or type 2 or type 3, are placed in the same DC, i.e., $r_d = 0$ backup DCs. Furthermore, each out of $n = 2$ SFCs are placed in different racks, i.e., $n_r = n = 2$ racks utilized, where $r_r = r = 1$ backup racks (Rack1). The active VNFs of certain type are allocated by $n_s = 1$ server requiring in total $r_s = r = 1$ backup servers for any VNF type, i.e., in total $r_s\Psi = 3$ backup servers per SFC. Overall, active VNF1 in S3 (Rack2) is protected by backup VNF1 allocated in S1 (Rack1), S3 in Rack2 is protected by backup server S1 in Rack1 and Rack 2 is protected by Rack 1. In contrast, let us assume that only Rack 3 is deployed in the DC shown in Fig. 1. In this case VNFs of different types are placed together in a single server to build a SFC, i.e., the active SFC in S2 and the backup VNF in S3, both in Rack 3. Please note, that our model requires VNFs placed in similar fashion, and, thus, requires that the VNFs are assigned such as in Rack 1 and Rack 2 or as in Rack 3, but not a mixture.

To reduce the amount of backup resources and to provide a level of disjointedness, different strategies are possible. For latency reduction, VNFs of different types can be placed together to build a SFC. If VNFs of the same type are placed together the control and management of Service Function Paths are simplified [10]. To evaluate the robustness of each placement strategy, we introduce *the amount of backup component types* N_r, $1 \leq N_r \leq 4$, and *the level of disjointedness* Δ, $1 \leq \Delta \leq 4$, whereby all VNFs inside the same server are disjoint due to their allocation in separated VMs.

The amount of backup component types N_r is defined regarding a certain VNF type and shows how many different component types are used to place VNFs of the same type disjointedly from each other. When active and backup VNFs of a certain type are placed in different DCs and, thus, in different racks, and servers, all component types, i.e., data center (DC), rack (R), server (S) and VNF, are protected by related backup components resulting in $N_r = 4$. An allocation of all VNFs of a certain type in the same server requires one DC and one R and reduces the number of backup component types to $N_r = 1$, i.e., only VNFs are protected. For example, in Fig. 1, considering the SFCs build by Rack1 and Rack2 only, we have $N_r = 3$, because active and backup VNFs of any certain type are allocated in different racks and server, but in the same DC. Similar, if the SFCs are only build in Rack3 and a whole SFC is allocated to a single server (S2 and S3), the number of backup component types is reduced as $N_r = 2$.

The level of disjointedness Δ describes the number of component types used to place a certain VNF type disjointedly from other VNF types, e.g., all VNFs2 from all VNFs3. Due to assumption, that at most one VNF is allocated by any VM disjointedly from other VNFs and each server can allocate multiple VMs, the minimal level of disjointedness is $\Delta = 1$, e.g., multiple VNFs of the same type are allocated in the same server. The maximal level of disjointedness $\Delta = 1 + 3$ can be reached when all VNFs of a certain type are additionally allocated in its own server, rack and DC disjointedly from other VNF types. In Fig. 1, for instance, when the SFCs are build by Rack1 and Rack2 only, the level of disjointedness is $\Delta = 2$, as each VNF type (VNF1, VNF2, VNF3) is placed in different server, e.g., in S1 (VNF2), S2 (VNF3) and S3 (VNF1) in Rack2. In contrast, if only Rack3 is used for the SFCs, the level of disjointedness is $\Delta = 1$, as all VNFs of different type are placed in the same server.

B. Affordable Component Failures (ACF)

Generally, $f_v \leq r$ VNFs, $f_s \leq r_s$ servers, $f_r \leq r_r$ racks and $f_d \leq r_d$ DCs can fail without any effect on the service maintenance. However, there is a need to take into account VNF placement strategy and the interrelation between different component types. To this end, we derive ACF classified as 1) Affordable Data Center Failures (ADCF) A_d; 2) Affordable Rack Failures (ARF) A_r; 3) Affordable Server Failures (ASF) A_r; and 4) Affordable VNF Failures (AVF) A_v. Since there are $r_d \geq 0$ backup DCs allocating backup racks, servers and VNFs, the maximal number of DCs, that can fail without causing service interruption is r_d. Thus, the ADCF is

$$A_d = r_d \tag{1}$$

Then, the number of failed DCs can vary as $0 \leq f_d \leq A_d$, whereby any f_d out of $n = 1+r_d$ DCs can fail. The failure of f_d DCs causes a failure of $f_d \cdot n_r$ racks reducing ARF, which can be calculated by taking into account all backup racks r_r and the failed racks due to DC failure, i.e.,

$$A_r(f_d) = r_r - f_d n_r \equiv A_r \tag{2}$$

As a result, the number of available racks, which can still fail independently from DC failures without causing any service interruption, can vary as $0 \leq f_r \leq A_r$. Also failure of f_r racks cause failure of $f_r n_s$ servers, i.e., the total number of failed servers due to failure of f_d DCs and f_r racks is $(f_d n_r + f_r)n_s$. Any server failures are, generally, compensated by the available out of r_s backup servers, whereby ASF is

$$A_s(f_d, f_r) = r_s - (f_d n_r + f_r)n_s \equiv A_s \tag{3}$$

Thus, the number of pure server failures f_s is bounded as $0 \leq f_s \leq A_s$. Generally, there are r backup VNFs of any type to protect an active VNF of the same type, while at most A_v out of n VNFs of a certain type can fail due to failures of servers, i.e., $f_s\psi$ VNFs, racks, i.e., $f_r n_s \psi$ VNFs, and DCs, i.e., $f_d n_r n_s \psi$ VNFs. Thus, AVF can be calculated as follow

$$A_v(f_d, f_r, f_s) = r - ((f_d n_r + f_r)n_s + f_s)\psi \equiv A_v, \tag{4}$$

where f_v, $0 \leq f_v \leq A_v$, additional VNFs are allowed to fail.

Additionally, we introduce and determine variable Λ, which describes the number of remaining DC (Λ_d), remaining racks (Λ_r), remaining servers (Λ_s) and remaining VNFs (Λ_v) of certain type, which are available and not affected by failures of any other components, but can also fail during service run-time The amount of available components (Λ_d, Λ_r, Λ_s, Λ_v), which can fail, is generally a function of already failed components (f_d, f_r, f_s, f_v) and derived below. There are $1 + r_d$ DCs allocating active and backup racks, servers and VNFs. DCs

can not be affected by failures of other component types, i.e., by failures of racks, servers and VNFs, resulting in

$$\Lambda_d = 1 + r_d = const, \tag{5}$$

whereby any f_d out of Λ_d DCs can fail during service run time. Since failure of f_d DCs causes a failure of $f_d \cdot n_r$ racks, the number of available backup racks is a function of utilized Λ_d and failed f_d DC, and can be calculated as

$$\Lambda_r(\Lambda_d, f_d) = (\Lambda_d - f_d)n_r \equiv \Lambda_r \tag{6}$$

Failure of f_r racks causes failure of $f_r n_s$ servers additionally to failure of $f_d f_r n_s$ servers due to f_d failed DCs, while the total number of remaining servers is

$$\Lambda_s(\Lambda_r, f_r) = (\Lambda_r - f_r)n_s \equiv \Lambda_s \tag{7}$$

After failures of f_d DCs, f_r racks and f_s servers the remaining amount of VNFs of any type can be determined as follow

$$\Lambda_v(\Lambda_s, f_s) = (\Lambda_s - f_s)\psi \equiv \Lambda_v \tag{8}$$

C. Service Reliability

Assuming first, that only VNFs can fail, i.e., $p_d = p_r = p_s = 1$, the service reliability is independent of VNF placement and determined as a probability that at least one over all $n = r+1 = \Lambda_v$ VNFs of each type, i.e VNF$i, i = 1, ..., \Psi$, do not fail and at least one SFCs can be composed by Ψ VNFs of each type to serve a client request. The reliability R_{v_i} of a certain VNF type i, e.g., VNF1, is independent of another types, e.g., VNF2, whereby the component reliability of any VNF is the same, i.e., $p_{v_1} = ... = p_{v_i} = ... = p_{v_\Psi}$. Thus, the service reliability of the SFC follows a serial reliability model given by $R = \prod_{i=1}^{\Psi} R_{v_i} = [R_v]^\Psi$. Furthermore, the reliability of a VNF of a certain type depends on the number of backup VNFs of same type, which, in this simple case, is modeled by a parallel reliability model determined by $R_v = \sum_{f_v=0}^{A_v} \binom{\Lambda_v}{f_v} p_v^{\Lambda_v - f_v}(1-p_v)^{f_v}$, where $A_v = r$. When SFC consists of $\Psi = 1$ VNF only, the defined service reliability results in $R = \sum_{f_v=0}^{A_v} p(\Lambda_v, f_v, p_v)$, where, to simplify notation, we introduce a probability mass function of binomial distribution as $p(\Lambda_v, f_v, p_v) = \binom{\Lambda_v}{f_v} p_v^{\Lambda_v - f_v}(1-p_v)^{f_v}$.

Assuming next, that $\Psi = 1$ and any component, i.e., DC, R, S and VNF, can fail, i.e., $0 < p_d < 1$, $0 < p_r < 1$, $0 < p_s < 1$, $0 < p_v < 1$ and $p_d \neq p_r \neq p_s \neq p_v$. We generalize the notation for the probability mass function as $p(\Lambda, f, p)$, i.e., $p(\Lambda, f, p) = \binom{\Lambda}{f} p^{\Lambda - f}(1-p)^f$, where variable Λ is equivalent to Λ_d, Λ_r, Λ_s, Λ_v and describes the number of available components (DC, Rack, etc.), which were not affected by occurred failures of any other components, but can still fail during service run-time. Similar, the variable f can be substituted by the number of failures, where f_d DCs, f_r racks, f_s servers and f_v VNFs of a certain type can fail, and p is equivalent to p_d, p_r, p_s, or p_v. The service reliability is now a function of any possible failure combinations, and thus, a function of ACF. With Eqs. (1), (2), (3), (4) and (5), (6), (7), (8), the end-to-end service reliability for $\Psi = 1$ is

$$\begin{aligned} R = \sum_{f_d=0}^{A_d} p(\Lambda_d, f_d, p_d) \sum_{f_r=0}^{A_r} p(\Lambda_r, f_r, p_r) \sum_{f_s=0}^{A_s} \\ p(\Lambda_s, f_s, p_s) \sum_{f_v=0}^{A_v} p(\Lambda_v, f_v, p_v), \end{aligned} \tag{9}$$

where the bounds of summations are functions of the amount of backup and failed components as derived in Sec. III-B. With these bounds, we further simplify notation of Eq. (9) as follows

$$R = P(f_d)P(f_r)P(f_s)P(f_v), \tag{10}$$

where $P(f_c) = \sum_{f_c=0}^{A_c} p(\Lambda_c, f_c, p_c)$ and index c can be replaced as $c \equiv d$, r, s, or v. The terms $P(f_d), ..., P(f_v)$ are ordered in decreasing hierarchical order, whereby a failure of a component at a higher hierarchy results in failure of all allocated components of lower hierarchy, e.g. failure of a rack result in failure of all allocated servers and VNFs.

To extend Eq. (10), i.e., Eq. (9), to arbitrary SFC lengths, i.e., $\Psi \geq 1$, there is a need to consider the VNF placement and resulting level of disjointedness Δ. If any VNF type is placed in different DCs and, thus, in different racks and servers, i.e., $\Delta = 4$, the SFC reliability is determined as

$$R_{\Delta=4} = [P(f_d)P(f_r)P(f_s)P(f_v)]^\Psi \tag{11}$$

If different VNF types are placed in the same DC but different racks and servers ($\Delta = 3$), the service reliability is

$$R_{\Delta=3} = P(f_d)[P(f_r)P(f_s)P(f_v)]^\Psi \tag{12}$$

As example, let us assume, that each out of $1+r$ DCs provides Ψ racks, where each rack support one server which hosts one VNF, so that, inside each DC, one SFC can be build. This placement is described by a serial reliability model, i.e., $P(f_d)[P(f_r)P(f_s)P(f_v)]^\Psi$, whereby r out of $1+r$ DCs can fail affecting all Ψ VNF types, but without causing a service interruption as described by term $P(f_d)$, which is a serial reliability term. However, the number of failed DCs has an impact on the components of lower hierarchy described by the variables A_c and Λ_c used in $P(f_c)$ for $c = r, s, v$.

If different VNF types are placed in the same rack, and different servers, i.e., $\Delta = 2$, the service reliability is

$$R_{\Delta=2} = P(f_d)P(f_r)[P(f_s)P(f_v)]^\Psi \tag{13}$$

In case of Ψ different VNF types are placed in the same server ($\Delta = 1$), the service reliability is

$$R_{\Delta=1} = P(f_d)P(f_r)P(f_s)[P(f_v)]^\Psi \tag{14}$$

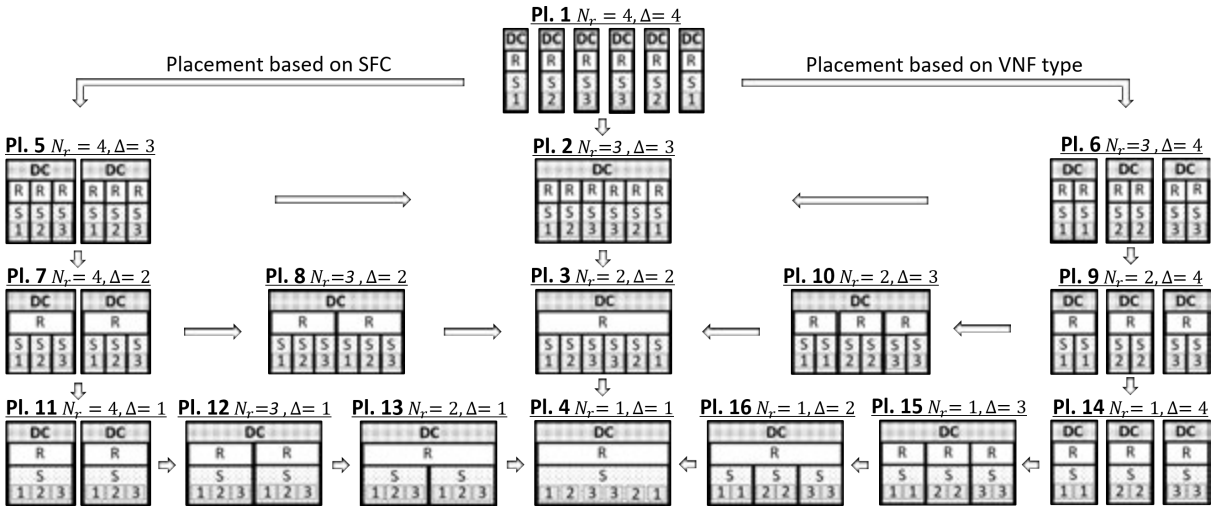


Fig. 2. Inter- and Intra-DCN placement of an active and a backup SFCs ($n = 1 + 1$) and the variable N_r, i.e., the amount of component types, i.e., data center (DC), rack (R), server (S) and VNF (presented by numbers), protected by backups, i.e, backup DC, R, S and VNF, respectively.

TABLE II
AMOUNT OF UTILIZED DCS, RACKS, SERVERS, NUMBER OF PROTECTED COMPONENT TYPES N_r AND LEVEL OF DISJOINTEDNESS Δ

	r_d	n_r	r_r	n_s	r_s	ψ	N_r	Δ
Pl. 1	r	1	r	1	r	1	4	4
Pl. 2	0	$1+r$	r	1	r	1	3	3
Pl. 3	0	1	0	$1+r$	r	1	2	2
Pl. 4	0	1	0	1	0	$1+r$	1	1
Pl. 5	r	1	r	1	r	1	4	3
Pl. 6	0	$1+r$	r	1	r	1	3	4
Pl. 7	r	1	r	1	r	1	4	2
Pl. 8	0	$1+r$	r	1	r	1	3	2
Pl. 9	0	1	0	$1+r$	r	1	2	4
Pl. 10	0	1	0	$1+r$	r	1	2	3
Pl. 11	r	1	r	1	r	1	4	1
Pl. 12	0	$1+r$	r	1	r	1	3	1
Pl. 13	0	1	0	$1+r$	r	1	2	1
Pl. 14	0	1	0	1	0	$1+r$	1	4
Pl. 15	0	1	0	1	0	$1+r$	1	3
Pl. 16	0	1	0	1	0	$1+r$	1	2

IV. NUMERICAL RESULTS

We now evaluate the placement and component dependent SFC reliability following example from Fig. 2 with $n = 2$ SFCs and $\Psi = 3$ VNF types, where all 16 possible placement strategies to symmetrically distribute one active and r backup VNFs of any type over Inter- and Intra-DCNs are illustrated. Table II summarizes the configuration for any placement strategy (Pl.). For example, in Pl. 1, each VNF is placed in a different DC, where each DC requires $n_r = 1$ rack (R), $n_s = 1$ server (S) and $\psi = 1$. Since all component types, i.e., DC, R, S and VNFs, are protected by r backups, there are $r_d = r$ backup DCs, $r_r = r$ backup R, $r_s = r$ backup S, and thus the number of protected component types is $N_r = 4$.

Also the level of disjointedness is $\Delta = 4$, since any VNF of a certain type is allocated in their own DC, R, S disjointedly from other VNF types. In contrast, Pl. 2 allocates all VNFs in the same DC while in different R, S, requiring $r_d = 0$, $r_r = r$, $r_s = r$. Here, DC is not protected against failure resulting in $N_r = 3$, also the disjointedness level is reduced to $\Delta = 3$ as VNFs of different types are allocated in the same DC. Generally, we can distinguish two VNF placement strategies, which are based either on SFC or VNF type. In the case of "VNF placement based on SFC", VNFs of different types are placed together to build SFC, e.g., Pl. 5, Pl. 7, Pl. 8, Pl. 11, Pl. 12 and Pl. 13. In contrary, the placements Pl. 6, Pl. 9, Pl. 10, Pl. 14, Pl. 15 and Pl. 16 are "based on VNF type", i.e., VNFs of the same type are placed together. The presented example can be extended to any value of r and Ψ.

For reliability evaluation, we consider two scenarios: 1) Use Case 1 relates to practice, where $p_d = 0.99999$, $p_r = 0.9999$, $p_s = 0.999$, $p_v = 0.99$; and 2) Use Case 2 is idealized scenario, where $p_d = p_r = p_s = p_v = 0.99999$. We investigate the end-to-end service reliability over one active SFCs with $\Psi = 4, 8$ VNFs and $r = 0, 1, 100$ redundant SFCs. We verified analytical results calculated with Eqs. (11), (12), (13) and (14) by Monte-Carlo simulations. Since the simulation results overlapped with analytical results (with 95% confidence interval), we present only analytical results in a tabular form for clarity, where N_r and in most cases SFC reliability values are sorted in the descending order.

Tab. III shows the SFC reliability for Use Case 1 and Use Case 2 as a function of the amount of redundancy r, SFC length Ψ and different VNF placement strategies defined in Fig. 2. Without backup protection ($r = 0$), the SFC reliability is independent from N_r, since there are no backup components. However, as expected, for a fixed N_r the SFC reliability decreases with increasing level of disjointedness Δ and increasing SFC length Ψ (for fixed Δ but different $N_r = 1, .., 4$ the reliabilities are identical). As expected, Use Case 2 provides higher SFC reliability due to higher component reliability defined, but reflects the same functional dependency. Without any protection, a disjointedness of $\Delta = 1$ will provide the highest end-to-end service reliability. This disjointedness can be provided by different VNF placement strategies as given by Pl. 11, Pl. 12, Pl. 13 and Pl. 4.

Now, we consider the usual 1+1 backup protection, i.e., $r = 1$. At first glance, the SFC reliability decreases with decreasing amount of backup component types N_r, and an increasing level of disjointedness Δ, e.g., for Use Case 1 with $\Psi = 4$. Intuitive, that seems to be a practical rule of thumb.

TABLE III
THE PLACEMENT DEPENDENT SFC RELIABILITY IN USE CASE 1 AND USE CASE 2 VS. REDUNDANCY r AND SFC LENGTH Ψ

N_r	Δ	Pl.	Use Case 1						Use Case 2					
			$\Psi = 4$			$\Psi = 8$			$\Psi = 4$			$\Psi = 8$		
			$r = 0$	$r = 1$	$r = 100$	$r = 0$	$r = 1$	$r = 100$	$r = 0$	$r = 1$	$r = 100$	$r = 0$	$r = 1$	$r = 100$
4	1	11	0.9595298551	0.9995123444	$\rightarrow 1$	0.9217205502	0.9990295230	$\rightarrow 1$	0.9999300030	0.9999999963	$\rightarrow 1$	0.9998900055	0.9999999935	$\rightarrow 1$
4	2	7	0.9566541431	0.9995075495	$\rightarrow 1$	0.9152878303	0.9990157556	$\rightarrow 1$	0.9999000045	0.9999999948	$\rightarrow 1$	0.9998200153	0.9999999900	$\rightarrow 1$
4	3	5	0.9563671756	0.9995073731	$\rightarrow 1$	0.9146473210	0.9990150262	$\rightarrow 1$	0.9998700078	0.9999999939	$\rightarrow 1$	0.9997500300	0.9999999879	$\rightarrow 1$
4	4	1	0.9563384848	0.9995073585	$\rightarrow 1$	0.9145832976	0.9990149596	$\rightarrow 1$	0.9998400120	0.9999999936	$\rightarrow 1$	0.9996800496	0.9999999872	$\rightarrow 1$
3	1	12	0.9595298551	0.9995031488	0.9999899999	0.9217205502	**0.9990210788**	0.9999899999	0.9999300030	0.9999899976	0.9999899999	0.9998900055	0.9999899956	0.9999899999
3	2	8	0.9566541431	0.9994984114	0.9999899999	0.9152878303	0.9990074399	0.9999899999	0.9999000045	0.9999899967	0.9999899999	0.9998200153	0.9999899935	0.9999899999
3	3	2	0.9563671756	0.9994982407	0.9999899999	0.9146473210	0.9990067233	0.9999899999	0.9998700078	0.9999899964	0.9999899999	0.9997500300	0.9999899928	0.9999899999
3	4	6	0.9563384848	0.9994682561	0.9999600006	0.9145832976	0.9989367949	0.9999200028	0.9998400120	**0.9999599970**	**0.9999600006**	0.9996800496	**0.9999199956**	**0.9999200028**
2	1	13	0.9595298551	0.9994111840	0.9998900009	0.9217205502	0.9989366284	0.9998900009	0.9999300030	0.9999799988	0.9999800001	0.9998900055	0.9999799976	0.9999800001
2	2	3	0.9566541431	0.9994070213	0.9998900009	0.9152878303	0.9989242748	0.9998900009	0.9999000045	0.9999799985	0.9999800001	0.9998200153	0.9999799969	0.9999800001
2	3	10	0.9563671756	0.9991072291	0.9995900640	0.9146473210	0.9982252375	0.9991902879	0.9998700078	**0.9999499994**	**0.9999500010**	0.9997500300	**0.9999100004**	**0.9999100036**
2	4	9	0.9563384848	0.9990772562	0.9995600766	0.9145832976	0.9981553639	0.9991203467	0.9998400120	**0.9999200012**	**0.9999200028**	0.9996800496	**0.9998400088**	**0.9998400120**
1	1	4	0.9595298551	0.9984906149	0.9988901109	0.9217205502	0.9980912785	0.9988901109	0.9999300030	0.9999699999	0.9999700003	0.9998900055	0.9999699995	0.9999700003
1	2	16	0.9566541431	0.9954981375	0.9958964363	0.9152878303	0.9911255646	0.9919188219	0.9999000045	**0.9999400011**	**0.9999400015**	0.9998200153	**0.9999000037**	**0.9999000045**
1	3	15	0.9563671756	0.9951995179	0.9955976973	0.9146473210	0.9904319848	0.9912246871	0.9998700078	**0.9999100032**	**0.9999100036**	0.9997500300	**0.9998300128**	**0.9998300136**
1	4	14	0.9563384848	0.9951696622	0.9955678297	0.9145832976	0.9903626567	0.9911553034	0.9998400120	**0.9998800062**	**0.9998800066**	0.9996800496	**0.9997600268**	**0.9997600276**

However, some placement strategies with low N_r provide higher SFC reliability (marked bold) than some placements with a higher N_r. In Use Case 1 with a SFC length of $\Psi = 8$, Pl. 12 with $N_r = 3$ outperforms Pl. 7, Pl. 5 and Pl. 1 with $N_r = 4$. That is because the SFC in these placement strategies will be most likely interrupted due to VNF failures caused by the longer VNF chain. Then, if we consider the VNF components of a SFC only, its reliability relates to $(p_v)^{\Psi}$, i.e., the lower reliability of the longer SFC chain, i.e., $(p_v)^8 \approx 0.92274 < (p_v)^4 \approx 0.9606$, can not be compensated enough by the backup DC, R and S. Thus, the amount of other backup (hardware) components, i.e., N_r, can be reduced.

Next, we investigate Use Case 2, where the placement strategies based on VNF type are marked bold. In this example we observe, that a placement based on SFC with lower N_r outperform a placement based on VNF type, e.g., Pl. 13 outperforms Pl. 6. Moreover, also other placement strategies based on VNF type (marked bold) provide the lowest SFC reliability for any value of r and Ψ and it is preferable to use a placement strategy based on SFC.

Next, let us consider the impact of increasing redundancy on resulting SFC reliability in both studied use cases, where $r = 100$ can be viewed as a good approximation of an upper bound. First, a more detailed rule can be observed for both use cases, which also reflects the case without protection: *For a given value of N_r, the reliability decreases with increasing level of disjointedness Δ, and, for a fixed value of Δ the reliability increases with increasing N_r.* For instance, for a $\Delta = 1$ we use Pl. 11, Pl. 12, Pl. 13 and Pl. 4 corresponding to descending $N_r = 4...1$, and, as can be seen for all r and Ψ, the higher N_r results in the higher reliability. Furthermore, the reliability increase with increasing redundancy, i.e., 0.9, 0.999, 1 with $r = 0, 1, 100$, respectively, for Pl.11 in Use Case 1. In contrast, Pl. 13 and Pl.4 with $N_r = 2$ and $N_r = 1$ provide SFC reliability of "3-" and "2-"nines almost independently from value $r \neq 0$, respectively. As discussed above and verified by the results in Tables III, the highest SFC reliability in any configuration can be reached by Pl. 11, where each SFC is allocated in the same server, i.e., $\Delta = 1$, but VNFs of the same type are placed in different DCs allowing protection of DC, rack, server and VNFs, i.e., $N_r = 4$. But for Use Case 1 and Pl. 11 the VNF reliability of $p_v = 0.99$ is too low to guarantee a "5-"nines, which can only be reached at the impractical bound ($r = 100$). Also for the upper bound $r = 100$, the placements Pl. 12, Pl. 13 and Pl. 4 can not reach the "5-"nines in both Use Cases.

V. CONCLUSION

This paper analyzed the bounds of SFC reliability as a function of SFC placement strategy over Inter- and Intra-DCNs. For the first time, we considered failures of racks, servers, VNFs and DCs, which can interdependently fail, thus impacting availability of other components and the end-to-end service reliability. Based on our analysis the most suitable placement strategy can be found for a high SFC reliability demand with lowest redundancy.

REFERENCES

[1] A. Immonen and D. Pakkala, "A survey of methods and approaches for reliable dynamic service compositions," *Service Oriented Computing and Applications*, vol. 8, no. 2, pp. 129–158, Jun 2014.

[2] H. Nakamura et al., "ETSI GS NFV-REL003: NFV; Reliability; Report on Models and Features for End-to-End Reliability."

[3] W. Ding et al., "Enhancing the reliability of services in nfv with the cost-efficient redundancy scheme," in *IEEE ICC*, 2017, pp. 1–6.

[4] A. Hmaity, M. Savi, F. Musumeci, M. Tornatore, and A. Pattavina, "Virtual network function placement for resilient service chain provisioning," in *RNDM*, Sept 2016, pp. 245–252.

[5] Z. Y. et. al, "Joint topology design and mapping of service function chains for efficient, scalable, and reliable network functions virtualization," *IEEE Network*, vol. 30, no. 3, pp. 81–87, May 2016.

[6] L. Qu, C. Assi, K. Shaban, and M. Khabbaz, "Reliability-aware service provisioning in nfv-enabled enterprise datacenter networks," in *CNSM*, Oct 2016, pp. 153–159.

[7] T. A. Nguyen, D. Min, E. Choi, and T. D. Tran, "Reliability and availability evaluation for cloud data center networks using hierarchical models," *IEEE Access*, vol. 7, pp. 9273–9313, 2019.

[8] L. Peterson and S. Das, "Trellis: Cord network infrastructure," Oct. 29 2017. [Online]. Available: https://wiki.opencord.org/pages/viewpage.action?pageId=2557405

[9] A. Singh et. al, "Jupiter rising: A decade of clos topologies and centralized control in googles datacenter network," in *Sigcomm*, 2015.

[10] J. M. Halpern and C. Pignataro, "Service Function Chaining (SFC) Architecture," RFC 7665, Oct. 2015. [Online]. Available: https://rfc-editor.org/rfc/rfc7665.txt

INTERNET DEVELOPMENT & ENERGY EFFICIENCY

IoT cloud-based services in network management solutions

O. Jukić, I. Heđi, E.Ciriković
Virovitica College, ICT department, Virovitica, Republic of Croatia
{ oliver.jukic | ivan.hedi | enes.cirikovic } @vsmti.hr

Abstract – **Main goal of research in this paper is to improve resilience of network management in networks with IoT (Internet of Things) devices. To set up a network with many IoT devices in the constrained environment can be a challenge. IoT objects refer to a wide variety of devices which are most often equipped with sensors. They can generate a large amount of real time data which can serve as an input to network management systems. If we consider the functional architecture of the TMN (Telecommunications Management Network), it is obvious that the QAF (Q Adapter Function) and MF (Mediation Function) will serve a great purpose in monitoring the above-mentioned network architectures. These two functions enable the integration of data from multiple sources, in which we include the data from the other control systems. Collecting data from devices most often requires developing custom-made applications which imply time and cost consumption. If devices are connected to Internet there are many cloud providers offering connectivity between devices and the cloud. One of them is AWS IoT (Amazon Web Services for Internet of Things). AWS IoT services can collect data from many different devices and connect them to endpoints for other tools, like network management solutions. Authentication and authorization are covered. Also, rules engine can filter and uniform data. The rule is triggered when a message that matches some filter is received. This paper describes some aspects of implementation of such system. Performance is evaluated by connecting small, single-board device, like Raspberry Pi to cloud service through different scenarios.**

Keywords – Internet of Things; Cloud services; Amazon IoT; Network management; TMN; MQTT; Raspberry Pi

I. INTRODUCTION

If problem appears in some part of network, network will generate large number of unsolicited events called alarms. That alarms are typically delivered in one center called network and service operating center. In that case we are talking about centralized network management concept. Alarms are carrying information about failure or some other types of malfunction. Detection of network alarms is called fault management. They can be classified in categories [1]. Sequence of that alarms can be recognized as a global problem with some root-cause. In addition, isolation and correction of unusual operational behaviours of telecommunication network and its environment are also called fault management [2]. Services provided by telecom operators are implemented in telecommunication network in such a way that they are dispersed through network resources [3] (Figure 1).

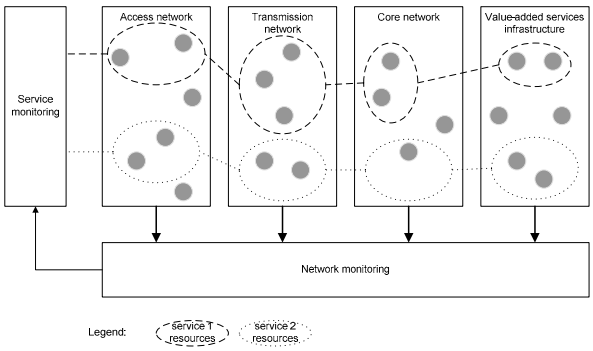


Figure 1. Service spreading over network resources

Knowing how service is implemented using network resources, it is possible to detect potential problems that can degrade quality of service that is guaranteed to customers. In functional architecture of the TMN (Telecommunications Management Network) [2] (Figure 2), we can consider that the QAF (Q Adapter function) and MF (Mediation function) will serve a great purpose

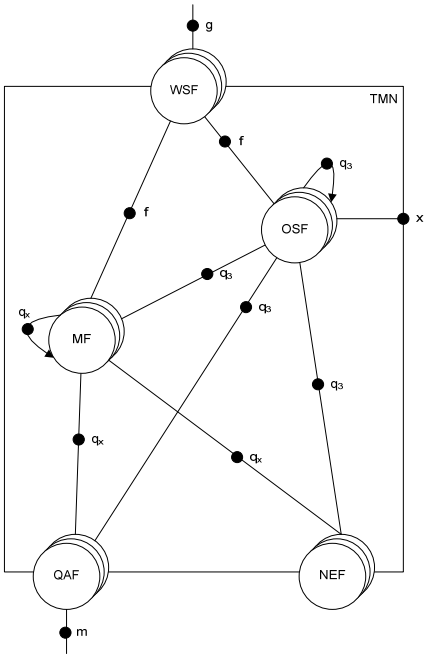


Figure 2. TMN functional architecture

in monitoring data from other sources [3], like data from IoT (engl. Internet of Things) devices in the constrained environment. QAF is a gateway to bridge the non-TMN systems to TMN systems through the interface (point called “m” on figure 2). Mediation function is custom function that performs data filtering and adjustment of data from other sources. Other blocks on figure 2 are standard TMN functional blocks.

Because all services depend on state of certain network element, it is obvious that network problems will have impact on quality of service which is offered to customer. IoT devices equipped with sensors can generate a large amount of real time data which can serve as an input to network management systems. In centralized network management systems real network resource is represented with network object. Status of this object is recalculated periodically using fault and performance management data. In our model, status of these objects can be enriched with additional information. Every network element has its own attributes. For example, in GSM network architecture base station controllers (BSC) controls number of base station transceivers (BTS). Example of attributes for base stations controller is number of operational base station transceivers. In our experimental architecture temperature in base station transceiver is measured with IoT device equipped with temperature sensor. IoT device is connected to Amazon AWS IoT (Amazon Web Services for Internet of Things) core service which collecting temperature values. Information is propagated to network management system.

Paper is structured as follows: first we will present short introduction; then, we are going to present the network model overview. Next, we have touched basic implementation aspects of model presented, such as data collection requirements and object's status visualizations.

A. Related work

Network management is very known term described in many papers, there is lack of management integration of data from other management systems and sources. In [4] is presented integrated view on telecommunication network in which relevant data sources are fault management data, performance management data, end-to-end testing results, customer complaints and other sources. In that model irrelevant data sources can decrease reliability and usefulness of model because number of alarm events coming from telecommunication network is huge. Authors are focused primarly on fault management and performance management with real time data. Term other sources is used for obtaining data by other network management systems through some interface like northbound interface. In our previous work [8] introduction to cloud-based services is presented on Amazon AWS IoT case study. In [5] similar model is used. Data from other sources are used for management data between two operaters in case when service is delivered to customers using other operators infrastructure. Similar approach is presented in [15] also but with Google cloud services. Main focus is to achieve energy conservation and efficient data management in IoT sensing applications. Authors implemented proposed GraphQL-based approach and compared it with solutions based on REST architecture. REST architecture and its benefits is described in many papers [19], [20], [21], [22]. In this paper we are not focused on energy consumption so we used REST architecture. In [16] authors described challenges in developing IIoT (Industrial Internet of Things) solutions. Such solutions typically heavily rely on interoperability requirements between different systems. The problem of data quality and the detecting of incorrect data is described in [17]. Authors proposed predictive temperature control algorithm for fault tolerant detection of a large number of IoT nodes. Management of many heterogeneous things in the Internet of Things network can be challenge. In [18] authors proposed using existing standards for configuring IoT network automatically. There are several open source IoT management platforms using the most popular IoT network management protocols. Performance evaluation study of open source IoT management platforms like ODL (OpenDayLight) is presented in [23].

II. Model Overview

A. IoT

IoT is known term and refers to the Internet of Things, system with computing devices with ability to transfer some kind of data to the Internet but there is no universal definition. By the [6] IoT is network that interconnects objects also known as Internet of Objects. When devices are connected to the Internet they can communicate with other devices or deliver information to certain endpoint. These devices can be connected to Internet directly using standard technology like 3G, 4G or they can connect to local area network which is connected to Internet. On other hand, devices can form M2M (Machine to Machine) networks in which devices are connected using radio technology communication standards and protocols like Wi-Fi (based on the standard IEEE 802.11), Bluetooth (based on the IEEE 802.15.1), Zigbee (based on the standard IEEE 802.15.4) or 6LowPAN over Zigbee (IPv6 over Low Power Personal Area Networks). In most cases application layer protocols are used for handling communication. The most representative application layer protocols are CoAP (Constrained Application Protocol), MQTT (Message Queue Telemetry Transport), XMPP (Extensible Messaging and Presence Protocol), RESTFUL Services (Representational State Transfer), AMQP (Advanced Message Queuing Protocol) and websocket [7].

For collecting and analyzing data generated from sensors attached to the IoT devices most often requires developing custom-made applications which imply time and cost consumption. There are many cloud providers offering connectivity between devices and the cloud, more than ten. One of them is AWS IoT. AWS IoT services can collect data from a large number of different devices and connect them to endpoints for other tools, like network management solutions. In our previous work [8] cloud-based services aimed for the connectivity, monitoring, and management of the IoT devices are presented on Amazon AWS IoT case study. Number of

cloud providers offering IoT services is constantly growing. What is common to all is possibility to connect IoT devices using some protocol to handle connection and store data generated by that devices. In addition to common functionalities each cloud provider offers its own set of features [9]. Some advantages why we used AWS IoT are fast and easy implementation of simple model case study, very intuitive device management interface, easy of integration to other services and opportunities for hosting web applications for presenting IoT devices data. Despite author's choice, used functionalities are common to all top-rated IoT cloud providers.

B. MQTT

MQTT is a M2M (Machine to Machine) Internet of Things lightweight connectivity protocol. Protocol is application layer protocol, released by IBM. Small code footprint makes it suitable to implement in small devices (e.g., 8-bit, 256KB RAM controllers). Protocol fulfills requirements for low power consumption, low bandwidth consumption and low latency. It uses the publish and subscribe pattern for transmitting and receiving messages between devices and applications (e.g., gateways or servers). Communication between devices and AWS IoT platform relies on this protocol. It is described in many papers [7], [8], [10], [11].

Connected devices are known under term „clients", which can communicate with an applications referred to as the „broker". Broker handles data transmission between clients. When client wants to distribute data, it will publish data to a certain topic. In that case client is „publisher". Broker then sends this data to any clients that have subscribed to that topic. All clients which can receive data are known as „subscribers" to certain topics (Figure 3) [8]. On figure 3 is presented communication between client equipped with temperature sensor and data presenting application layer. Client is publisher to topic under name „temperature". Client push values about temperature to given topic, client is publisher. Presenting application layer reads data from given topic, presenting application layer is subscriber.

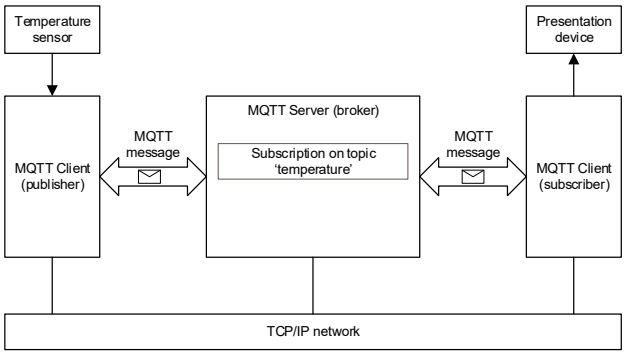


Figure 3. MQTT protocol messages

Messages on each topic can be retained, which means that each topic can have one retained message that a client automatically receives when it subscribes. MQTT protocol has libraries in many programming languages, so implementation is very simple.

Protocol is based on Message Oriented Middleware (MOM) approach. It is kind of architecture in which messages travel between entities rather than function calls. It is most useful in heterogeneous and high performance systems. In MQTT protocol each client registers its interests to broker, publisher or subscriber. Benefit of that architecture is separation of identities for different kinds of clients.

Amazon AWS IoT supports MQTT over the websocket protocol to enable browser-based and remote applications to send and receive data from AWS IoT-connected devices using AWS credentials. It means that MQTT broker places the MQTT message into a websocket message, and sends it to the client. The client unpacks the MQTT message from the websocket message and then processes it as a normal MQTT message. Every browser can be both publisher and subscriber. Websocket support is available on TCP port 443, which allows messages to pass through most firewalls and web proxies [12].

C. Amazon Web Service IoT platform

IoT platform is multi-layer technology which is used to manage IoT devices. If devices are connected to Internet there are several cloud providers offering connectivity between devices and the cloud. One of them is AWS IoT (Amazon Internet of Things). The primary function of the IoT platform is to act as middleware layer to connect devices or applications from one end to another end. IoT concept contains a variety of functions like sensors and controllers, gateway devices, software for data analyzing and end application services. IoT cloud platform can handle huge data volume from sensors, devices, applications, and take actions to give a real-time response. According to [12] Amazon AWS IoT platform is one of the most popular platforms in 2020. It provides communication between IoT objects and the cloud in both directions, it collects data from devices or enable users to control devices remotely. All devices or applications that are supported by AWS IoT cloud platform are called "thing". Thing is connected to AWS IoT by device gateway which serves as an entry point for all things. Device gateway handles connection and security. It supports a couple of protocols used for IoT networks: Message Queue Telemetry Transport (MQTT), MQTT over websocket, MQTT over the secure websocket and HTTP protocols. In one project, communication with devices can be realized using one or more available protocols. Security is covered using X.509 certificates. AWS IoT can generate a certificate for you or you can use your own X.509 certificate. In either case, both types of certificates must be registered and activated in AWS IoT device management portal. All connected devices must undergo an authentication and authorization process. Authorization process determines which action device can perform. All information about device is stored in registry as well as certificates used. By the registry, each device is given a unique identifier. The state of each device is stored in device shadow part.

Device Shadows is virtual version of each device that includes latest state so that other devices or applications can evaluate messages and communicate with the device. It is service allows not only retrieve state of device but also change state using RESTful (representational state transfer) API (application program interface) or specialized MQTT topics (Figure 4).

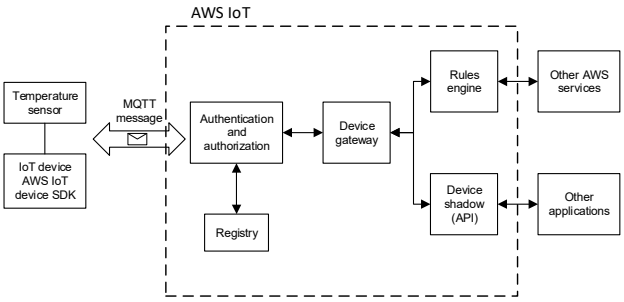


Figure 4. AWS IoT architecture components

To easily and quickly connect device to AWS IoT, the AWS IoT device SDK (software development kit) is needed. The AWS IoT Device SDK include open-source libraries and developer guides with samples. The Rules engine makes it possible to build custom IoT applications that analyze and act on data generated by connected devices without having to manage any infrastructure.

III. IoT and Network Management Solution Use Case

Based on model overview mentioned above, we have integrated AWS IoT data into concept of existing network management system solution used in GSM network architecture. Presentation layer of network management system solution shows object for base station controller (BSC) which controls number of base station transceivers (BTS). Example of attributes for base stations controller is number of operational or number of non-operational base station transceivers calculated from fault management data (Figure 5).

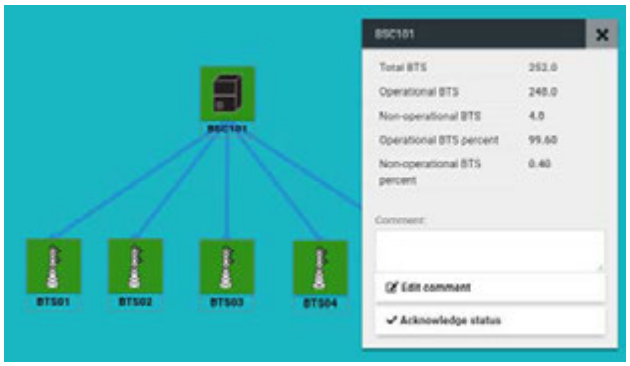


Figure 5. Graphical user interface for service monitoring

In our experimental architecture these set of fault management attributes is enriched with additional data coming from different sources, like AWS IoT cloud service. Temperature in base station transceiver is measured with IoT device equipped with temperature sensor. To collect data from these IoT devices it is necessary to develope custom-made application, like access module in network management system, with implemented application layer protocol for handling connection, like MQTT. Hence, IoT device is connected to Amazon AWS IoT core service using for collecting temperature values. Information is propagated to network management system.

As an IoT device Raspberry Pi 4 Model B is used. Device is single board computer with integrated 2.4 GHz and 5.0 GHz IEEE 802.11ac wireless module, bluetooth 5.0 module and gigabit ethernet module for connecting to network. For measuring temperature, Raspberry Pi must be equipped with some temperature sensor. DS18B20 digital temperature sensor is used which communicates with Raspberry Pi using 1-wire method (Figure 6).

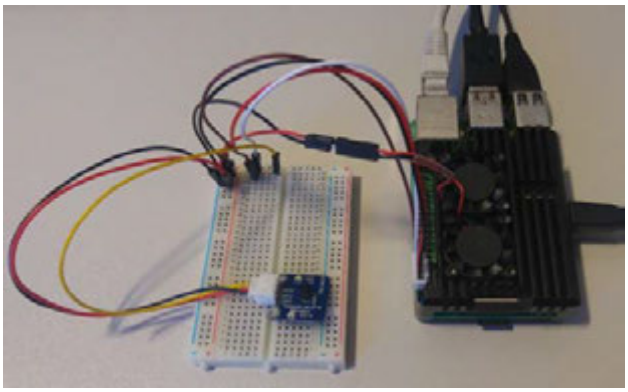


Figure 6. Experimental hardware architecture

Device is not coming with preinstalled operating system. There are number of operating systems optimized for Raspberry Pi, we used Raspbian which is based on Debian operating system but optimized for Raspberry Pi board.

Device is registered in AWS IoT as a thing and got unique identifier called Amazon Resource Name (ARN). After that appropriated certificates must be created and downloaded to IoT device. At the end there is need to set up authorizations which means to define which actions a device can perform. It is JSON structure containing policies:

```
{
   "Version": "2020-02-02",
    "Statement": [{
      "Effect": "Allow",
         "Action": [
               "iot:Connect", "iot:Publish"
          ],
         "Resource": "arn:aws:iot:…:topic/temp"
    }]
}
```

Statement consists of three parameters: Effect consists of values allow or deny, Action is ability to connect to AWS IoT cloud and publish message to certain topic and Resource consists of device ARN and topic.

To easily and quickly connect device to AWS IoT, the AWS IoT device SDK is needed. The AWS IoT Device SDK enables end devices to connect, authenticate, and exchange messages with AWS IoT Core using the

MQTT, HTTP or websocket protocols. The AWS IoT Device SDK supports C, Javascript, and Python programming languages. We used Javascript programming language in Node.js environment and MQTT protocol. Part of source code for creating instance of device ready for authentication and authorization to AWS IoT:

```
var oAws = require('aws-iot-device-sdk');
var oDevice = oAws.device({
   keyPath: <PrivateKeyPath>,
  certPath: <CertificatePath>,
    caPath: <RootCACertificatePath>,
  clientId: <UniqueClientIdentifier>,
      host: <CustomEndpoint>
  });
```

To send MQTT message it is only necessary to call function for publishing data to certain topic:

```
setInterval(function){
   var fTemp = GetTemperatureValue();
   var oDate = new Date();

device.publish('temperature',JSON.stringify({
       value:fTemp,
       datetime: oDate.getTime();
   }));
},30000)
```

Published message look like:

```
{

   "value": 50,
   "datetime": 1581700066476
}
```

IoT device publish message (temperature value) every 30 seconds. AWS IoT Rules engine gives device ability to interact with other AWS services adding rules. Rules are analyzed and actions are performed based on the MQTT topic. Presentation layer of network management system solution has RESTful API interface to set additional temperature status attribute value. Every rule created on AWS IoT platform has its own query statement in form of SQL query statements. For example:

```
   SELECT value FROM 'temperature' WHERE value
>= 50
```

where value is part of MQTT published messages and temperature is MQTT topic. Rule is triggered when temperature value rises 50 degrees Celsius. Rules execute one or more actions like insert a message into DynamoDB table, send a message as an SNS push notification or Send a message to a Lambda function.

AWS Lambda service can run some code without provisioning or managing servers. For creating Lambda function, it is necessary to choose a runtime. A runtime is a version of a programming language or framework that can be used to write Lambda functions. AWS Lambda supports runtimes for the following languages: C#/PowerShell, Go, Java, Node.js, Python and Ruby. We used Node.js that calls RESTful API interface on presentation layer of network management system solution. On figure 7, object BTS01 has changed status and color for warning notification because temperature value is 54 degrees Celsius. Object is still operational.

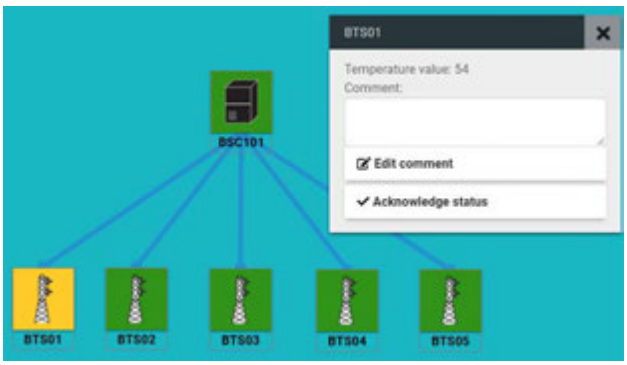


Figure 7. Graphical user interface for service monitoring

AWS IoT services can collect data from many different devices and connect them to endpoints for other tools, like network management solutions. One scenario when temperature rises 50 degrees is shown with message sequence chart on figure 8.

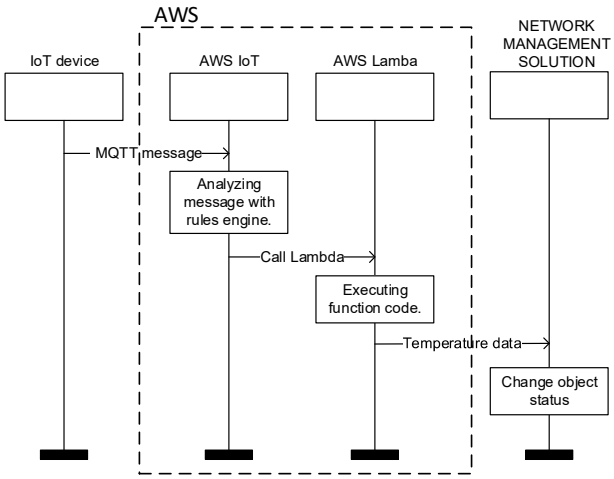


Figure 8. Messages exchanged in one scenario

IV. CONCLUSION

In this paper, we have presented an integration of IoT cloud services with network management solutions. Mobile network monitoring system is enriched with additional data collected by IoT devices. Main goal of research in this paper is to improve resilience of network management in networks with IoT devices. Collecting data from devices most often requires developing custom-made applications which imply time and cost consumption. If devices are connected to Internet there are several cloud providers offering connectivity between devices and the cloud. One of them is AWS IoT. In our experimental architecture temperature in base station transceiver is measured with IoT device equipped with temperature sensor. IoT device is connected to Amazon AWS IoT core service that collects temperature values. Information is propagated to network management system. We have briefly described architecture and components of the AWS IoT. AWS IoT services can collect data from many different devices and connect them to endpoints for other tools, like network management solutions.

LITERATURE

[1] O. Jukic, M. Špoljaric and V. Halusek, "Low-level alarm filtration based based on alarm classification", Proceedings of the 51st International Symposium ELMAR 2009, pp. 143-146, Zadar, September 2009.

[2] Udupa, K.D., TMN – Telecommunications Management Network, McGraw-Hill Telecommunications, New York, 1999.

[3] O. Jukić and I. Heđi, "Service monitoring and alarm correlations", 4th International congress on ultra modern telecommunications and control systems, pp. 330-334, Saint Petersburg, 2012.

[4] O. Jukić and M. Kunštić, "Integrated view on telecommunication network status", Proceedings of the 34rd International Convention MIPRO 2011, pp. 141-145, MIPRO, Opatija, 2011.

[5] O. Jukić and I. Heđi, "Some aspects of network management system for video service", Proceedings of the 39th International Convention MIPRO 2016, pp. 805-808, MIPRO, Opatija, 2016.

[6] F. Xia, L. T. Yang, L. Wang and A. Vinel, "Internet of Things", International Journal of Communication Systems, Vol. 25, pp. 1101-1102, 2012.

[7] V. Karagiannis, P. Chatzimisios, F. Vazquez-Gallego and J. AlonsoZarate „A Survey on Application Layer Protocols for the Internet of Things", Transaction on IoT and Cloud Computing 2015.

[8] O. Jukić, I. Špeh and I. Heđi, "Cloud-based services for the Internet of Things", Proceedings of the 41st International Convention MIPRO 2018, pp. 407-412, MIPRO, Opatija, 2018.

[9] T. Pflanzner, A. Kertesz, "A Survey of IoT Cloud Providers", Proceedings of the 39th International Convention MIPRO 2016, MIPRO, Opatija, 2016.

[10] P.P. Ray, "A survey on Internet of things architectures", Journal of King Saud University – Computer and Information Sciences, pp. 1319-1578, 2016.

[11] D. Kang, M. Park, H. Kim, D. Kim, S. Kim, H. Son and S. Lee, "Room Temperature Control and Fire Alarm/Suppression IoT Service Using MQTT on AWS", Platform Technology and Service (PlatCon), 2017 .

[12] https://docs.aws.amazon.com/iot/latest/developerguide/mqtt-ws.html

[13] https://www.softwaretestinghelp.com/best-iot-platforms/

[14] https://docs.aws.amazon.com/iot/latest/developerguide/aws-iot-how-it-works.html

[15] R. Khan, A. Noor Mian, "Sustainable IoT Sensing Applications Development through GraphQL-Based Abstraction Layer" Electronics 2020, 9, 564, 2020.

[16] N. Santos, F. Morais, H. Rodrigues, R. J. Machado, "Systems Development for the Industrial IoT: Challenges from Industry R&D Projects" In: Mahmood Z. (eds) The Internet of Things in the Industrial Sector. Computer Communications and Networks. Springer, 2019.

[17] R. Casado-Vara, Z. Vale, J. Prieto, J. M. Corchado, "Fault-Tolerant Temperature Control Algorithm for IoT Networks in Smart Buildings", Energies 2018, 11, 3430, 2018.

[18] Sahlmann, K., Scheffler, T., Schnor, B., "Ontology-driven Device Descriptions for IoT Network Management", 2018 Global Internet of Things Summit (GIoTS), Bilbao, 2018.

[19] A. Bhattacharyya, S. Agrawal, H. K. Rath and A. Pal, "Improving Live-Streaming Experience for Delay-Sensitive IoT Applications: A RESTful Approach", 2018 IEEE Globecom Workshops (GC Wkshps), Abu Dhabi, United Arab Emirates, 2018.

[20] R. K. Kodali and V. S. K. Gorantla, "RESTful Motion Detection and Notification using IoT", 2018 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 2018.

[21] K. Aloufi and O. Alhazmi, "Secure IoT Resources with Access Control over RESTful Web Services", Jordan Journal of Electrical Engineering, Volume 6, Number 1, 2020.

[22] B. N. Silva et al., "RESTful Web of Things for Ubiquitous Smart Home Energy Management", 2020 International Conference on Computing, Networking and Communications (ICNC), Big Island, HI, USA, 2020.

[23] J. de C. Silva, P. H. M. Pereira, L. L. de Souza, C. N. M. Marins, G. A. B. Marcondes and J. J. P. C. Rodrigues, "Performance Evaluation of IoT Network Management Platforms" 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018.

Innovation Platfirm – A Novel Energy Service Utility

A. Viskovic*, D. Simunic** and V. Franki***
* Energy Platform Living Lab, Zagreb, Croatia
** University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
*** University of Rijeka, Faculty of Engineering, Rijeka, Croatia
alfredo.viskovic@gmail.com

Abstract –

The plan of a novel Energy Service Utility (ESU) traced the evolution lines towards a modern positioning of a multi-utility environment that is oriented towards the required topics of green and smart city, a leader in the environmental context, in smart networks and new energy models. This strategy has triggered numerous interventions with the main aim to mobilize citizens and especially young population in this field. After the analysis, our goal is to try to translate the plan into suggestions for improvement and business ideas, based on the emerging ICT. The main point of changing the classical „platform" to „platfirm" was to contribute to the ideas, to the tips regarding the levels of the plan and, for the classified groups to its implementation.

Keywords – energy service utility; platform; platfirm; smart city; emerging ICT; 5G

I. INTRODUCTION

Smartization process of a country incorporates a capability of integrating data and the corresponding cutting-edge information and technologies for planning, management and operations on multiple sectors and territorial areas to generate sustainable and shared public value. In 2018, the United Nations published projections that show the increase of the world's population in urban areas by another 2,5 billion people, from current 55% to 68% in 2050 [1]. In this context, smart cities reconfigured as urban ecosystems where places, people and objects are connected and interconnected in every aspect of daily life to create a totally integrated system, to increase the productivity of resources, to improve the use of energy and raising the quality of life of citizens. This is the basis of the currently planned Society 5.0.

The EU policy on smart cities incorporates the European innovation partnership on smart cities and communities (EIP-SCC), as an initiative supported by the European Commission that brings together cities, industry, small business (SMEs), banks, research and others. In the field of energy, EU is especially active [2], [3],[4],[5]. The most important goal is related to the sustainability of all the solutions and to the integration of the solutions. The easiest way to perform integration is to connect all the vertical lines, currently separated by different policy areas, by a strong and unique integrator: horizontal line. The Fourth industrial revolution [6] leans on the digital solutions that are and will furthermore support the exponential growth of different kinds of data collections and communications. For the case of the electric grid, it is evolving together with the whole smartization philosophy to the smart grid. This means that the digital utilities and their infrastructure support the two way communication that is the basis for the introduction of the smart grid, as well as for the smart city. Smart grid represents a true revolution with the property of incorporation possibility of renewable energy sources to the grid [7], and its distribution and demand in a bidirectional based way of communication that is based on the newest Information and Communications Technologies (ICT). The smart grid is revolutionary, because the mostly applied traditional electric energy systems are providing unidirectional power delivery system with the minimum flow of data information. Addition of a possibility of a bi-directionality enhances the Energy Efficiency (EE) of a smart grid. The EE improves not only on the level of a home or a city, but on the country and/or region level. Therefore, the smart grid is a crucial and unavoidable part of any smart city and, in fact, of any modern country of 21st century. The true smart grid can be run only by a smartly designed platform, presented in this paper.

The ESU platform is an integrator, based on the ICT. Since ICT progresses by an enormous pace, the most significant evolution will take place by already happening introduction of 5G. When 5G will come to the end of its technical possibilities, the world of the research will be oriented toward implementation of 6G. The move to 5G will be a considerable leap. However, the 6G, currently assumed for 2030, will be definitely a huge leap. This can be thought of especially when considering that 5G is already in the strong discontinuity with the present up-to-4G technologies. The 5G technologies will be able to reach the maximum speed of up to 5 Gigabits per second (on average 10 times higher than today), with a latency of a few milliseconds (in average 5 milliseconds compared to the current 30-40ms), but with the tendency toward 1 ms. The high data rates and the reduced communication latency of the 5G are incredibly important for the ESU due to the full reliable concept of the next industrial revolution in terms of Industry 4.0. Industry 4.0 relies on sensors, development of smart products and assets. Thus,

Industry 4.0 can move from a factory-focused initiative to a company-wide business strategy, where the energy sector and its corresponding ESUs is one of the most important applications. This is the background of the relationship between the most profound modern technical knowledge and technologies and the related business models. Nowadays, we do hope that the capacity of the system will be able to manage up to a million connections simultaneously, and at the same time to generate an energy consumption lower than the currently deployed communications systems (e.g., 4G). The results of the ongoing studies will show the relation between the consumption of 5G systems and the increase of energy savings in the whole smart city and the smart country system [8]. However, as it will be later explained, the transformation of a platform to a platfirm is possible only with the well-balanced combination of the thoroughly studied business models and environment and deep knowledge of the modern ICT.

The emerging technologies, coming along with 5G can address exactly city-specific challenges of different vertical lines. Thus, the ICT already serves as a foundation of the smart city, with the prospect of even higher engagement for it. Some of the examples in the field of smart grid multi-utility services based platform are given here.

II. PLATFORM EXAMPLES

Liu Jianming et al [9] proposed a smart grid Multi-utility Services Platform based on Power Fiber to the Home which adopts a new type of composite cable named Optical Fiber Composite Low-voltage Cable that achieves transmission of energy and information over the same cable. Increased demands for energy and thus also for energy saving in the home environment has evolved also into technologies of monitoring and controlling energy consumption through smart utility network. Another example is given by Sangjae Lee et al [10] who implemented an IEEE 802.15.4g wireless communication in the platform incorporated in the smart utility network. Another kind of usefulness of a grid-based service portal was shown by Babafemi O. Odusote et al [11] in the design and implementation of grid-based service portal for evolving computing platforms. A. Vukojevic et al [12] developed and optimized an ecosystem by an open interoperable distributed intelligence platform on an integrated utility microgrid test site that is a need for nowadays existing siloed proprietary software protocols, communication networks and hardware devices, being the core part of the smart grid devices having the properties of digital signal processing, data analytics, and communication capabilities. This shows that for an operation of the smart city, it is crucial to plan the platform properties in advance, in order to enable utilities to realize their full financial and operational benefits. Since smart grid is based on ICT, it inherits all the ICT vulnerabilities. Therefore, Mohammed Saleh et al in [13] proposed setting up a mobile modular lab for cyber testing of smart grid end point devices in order to identify and quantify the security risk in these devices and suggest recommendations. Jun Ma et al. in [14] described the SMART Infrastructure Facility of University of Wollongong in Australia that is a development of an infrastructure dashboard, with the aim of providing infrastructure planners, designers, developers and regulators with a cross-organization information platform for better understanding the infrastructure development in a specific area and the community requirements. The idea of the development of such a platform is a result of the research that indicates that the residential sector, incorporating typically uncertainty and periodicity, generally accounts for 1/6 to half of the total national energy consumption, whereas the average figure worldwide is approximately 30% [15]. This is another proof that the development and application of ESU platform can contribute to the sustainable and more efficient utility services.

Finally, the industrial example of such ESU is a company A2A [16], from Milan, Italy. The company traced the lines of its evolution towards a modern positioning of a multi-utility oriented towards the topics of a green and smart city, a leader in the environmental context, in smart networks and new energy models. The company states that “Digital is the future of Utility A2A, both in the process of transformation and continuous progress of the business practice”. The A2A operational perimeter defines also an ad hoc company, A2A Smart City, Smart City Lab, that is responsible for the realization of innovative Internet of Things digital technologies to be applied to the management of services dedicated to the territory.

III. SMART CITY POLICY

The smart city policy is based on the

- sustainable urban mobility
- sustainable districts and built environment
- integrated infrastructures and processes in energy, information and communication technologies and transport
- citizen focus
- policy and regulation
- integrated planning and management
- knowledge sharing
- baselines, performance indicators and metrics
- open data governance
- standards
- business models, procurement and funding

Energy is the most important factor that contributes to functioning of any industry, mobility, all the systems of the city and community. Nowadays, the knowledge is that still a big part of the energy comes from the burning of fossil fuels and results in the release of Carbon Dioxide (CO2) into the atmosphere [17]. The ICT with all the accompanying hardware and software can contribute not only to the analysis, but also to the increase of the overall energy efficiency, as well as to the combination of various utilities (i.e., multi-utility environment).

The examples of the ICT and energy applications include:

- buildings (with the smart building management systems and sensor networks)
- smart grid (helping to reduce peak demand)
- homes (with smart meters and smart appliances)
- smart city (by integration of energy generation and supply with transportation and mobility systems).

At the base of the realization of this digital energy system there is a predisposition of technological infrastructures like robust hyper-fast connection systems, creating smart networks capable of operating with extremely low latency based on the emerging technologies, like 5G [18] or even, 6G [19],[20] technologies. There are many areas of activities that are useful for the transformation of a smart city, but one of the most important is the ESU.

The thematic challenges of innovation technology PLATF**O**RM to a novel energy PLATF**I**RM are:

• How to improve the image of utility in the eyes of the public?

• How to enhance customer experience (satisfy its needs and stimulate new ones)?

• How to get to know internally who we are as a company and as individuals?

• How to develop new business models for utility?

• How to trigger change in the company culture to develop greater entrepreneurship and accountability among citizens?

• How to simplify and/or revolutionize business processes?

• How to do more with less?

• How to make a digital utility with new technologies?

The answers will bring the ESU to the innovations of the energy producers and suppliers. The planned innovations are in the areas of energy production, energy transport, energy delivery and of offered services. Therefore, users will be the centre of the transformation toward the smart city and towards overall smart environment. The innovations can help the city and the country (or the region, if we consider the EU as a multinational region) to decrease an import and to increase an energy export, which can again come back to the support of creation of new jobs and thus, industry growth for the separate countries and thus, to the whole EU.

IV. Energy Platform Phases

A. Platform Design

The first phase of platform design had an initial moment of "envisioning" which was attended by the future contributors who have been sensitised, in the presence of top management and who are themselves strongly involved in the project, to the benefits and opportunities of the new collaborative platform of open innovation. Through a multimedia envisioning event, we described and stimulated the organization and the participants towards new participatory paradigms, on the logic of collaboration, informal network dynamics and on employee engagement and co-creation. This phase allowed us to create the right setting of expectations regarding the required commitment, rules and logic of governance, the risks and the levers that were necessary for the project team to adopt in order to succeed. The project team shared an engagement approach and a platform approach too (and not just for communication or training), focusing efforts on creating a more appropriate environment of the activation. Then, we focused to "behaviour design activities" in relation to the interactions that we had imagined and expected to be enabled by the social innovation and idea management platform.

In the specific case, it means designing the interactions that must involve participants and contributors called upon to propose, comment, enhance and vote on creative ideas for solutions, services and innovative products for ESU. Recalling the canvas model of platform construction, the value of exchange was the idea that is produced and commented on, while the exchange currency was identified for the company's reputation (symbolic coins) that the participants earned by their commitment and contribution to the platform. The important passages of this phase of the behaviour design were carried out in a logic of co-design, involving a subgroup of participants:

- definition of a challenge (to which the participants will be called), selected according to the priorities and the interests of the participants
- interactions, expected collaborative behaviour, and filtering mechanisms will be defined: the process of developing ideas (funnel), the format of the ideas, the possibility to vote and comment on the ideas, the further possibility for the author of an idea to promote a comment to a "significant contribution", maybe even a wiki system to trace the progressive evolution of the idea and to incorporate the collected "significant contributions" in the original proposal
- gamification mechanism will be calibrated to build charts of ideas/challenges and of the participants representing the collaborative dynamics of the system (the dynamics and mechanics of gamification, specially designed, are relevant in order to keep the participants active, involved and stimulated as regards their presence on the platform and the strength of their contributions).

B. Platform Building

In the second phase of platform building, we designed the journey of the participants on the platform by identifying the key moments (access, possible actions, interactions between the different profiles) and the touchpoints (desktop, mobile, physical meetings) and user experience with minimal customization of the layout and the setting of key features.

The most appropriate platform of the idea management was selected and configured. The choice was addressed towards highly usable solutions with attention to mobility. We proceeded to configure all the key elements of the platform with filter mechanisms. In this case, the filter mechanisms were, for example, the actions of the participants that assigned relevance to a comment (a significant contribution), the classification of challenges by the relevant subject that allowed the participants to be oriented in the challenges, systems and practices of idea tagging.

The system of governance was built by sponsorship (C-level), experts and community managers defined and trained for each challenge (for care, customization and to address interactions and exchanges between participants).

A test with a closed group of users allowed the validation of the governance and technology system and permitted the system to be populated with the first content. The testing also served to verify the contributor journey in relation to different personas (or typical profiles imagined for participants: creator, commentator, facilitator, expert) and to touchpoints (desktop, mobile, etc.).

C. Platform Deployment

The launch of the system was accompanied by an appropriate communication activity, through a dedicated blog and video production of documentation and storytelling (the pull activity proved relevant in order to stimulate and activate access to the platform).

A team dedicated to the community management handled user support, reporting, animating (with the involvement of experts when it was necessary to intervene on the content) and the implementation of corrective actions when needed. Analytics were fundamental to monitor the evolution of the participation along the way, its critical moments or the most effective initiatives.

The platform was able to progressively capture the collective intelligence of the interactions managing to outline the profiles of the participants in an emerging manner (those who are more creators, how many of them prefer to comment on other people's ideas rather than to propose their own, and those who are the best at stimulating others to contribute).

Some live moments permitted us to better examine and to improve the proposals, as well as to articulate the specific themes of interest from an educational perspective (techniques of creativity, business case and evaluation of feasibility, digital evolution etc.). In the final deployment phase, when difficulties associated with digitalisation efforts are surpassed together with the resolved technical and regulatory constraints, energy utilities can start their operations as platforms.

D. Platform Operation

The operation of the utility as the platform has at least two profound reasons, as its basic support:

- In a liberalised energy sector, energy suppliers should be registered as separate legal entities with regard to energy producers. Electricity is produced by a power producer, transmitted by a transmission system operator, distributed by a distribution system operator and then supplied to end consumers. This means that a utility selling electricity to end consumers can act only as a procurator of a commodity. Its job would then be to offer a reliable and affordable supply through the use of various transactions. If a utility can function only as a commodity procurator and a service provider, without owning energy related assets, having a platform would be essential in differentiating among other suppliers, especially considering the slim margins that the energy sector is currently facing. A platform can provide a better service for clients, enable easier and more transparent monitoring of consumption, optimise energy management models and simplify bill handling processes.

The fact that there is a single type of electricity, meaning that it is always the same type of good for selling, the major distinction between different energy suppliers is, in fact, cost related. The business is based on providing a service rather than a particular type of good. Exactly this fact allows platforms to be the most adequate tool. Furthermore, a platform that entails an open innovation model procures capabilities from outside companies. These capabilities raise the platforms level of service and inevitably provide a source of added value for the entire ecosystem, including end consumers.

E. Platform Enables an Eco-system

The utility platform must function as a connective tissue that connects various players of the energy ecosystem and enables a digitally controlled, decentralised flow of energy throughout the system. Unfortunately, unlike other sectors, where platform companies can often act simply as matchmakers between supply and demand, the story in the energy sector is quite different and much more complicated. This is because of the specificities of electricity as a commodity and the gravity of the energy system itself. In order to be successful, a platform operating in the energy arena needs to be able to draw real time data from countless physical components of the system. It then needs to utilise these data to optimise system's operation (again in real time) without endangering the reliability of supply. When numerous consumers, regulatory restrictions and technical issues are added to the equation, an enormous task of reconciling these constraints with needs and desires is put in front of us and of a platform operation.

F. Time for Energy Disruptors is Now!

While following a platform strategy, it is of extreme importance to adopt an open-minded approach and to embrace new partnerships. An innovative mindset that supports an open innovation model is a prerequisite of a successful, digitally-based company. This will prove to be particularly true in the case of the energy industry where we see the utility serving as glue connecting different players across the energy supply chain. A platform based utility with an open innovation model will also engage new entrants. They will contribute to the enhancement of the system, while at the same time profit from the

ecosystem created by the utility. Building a product or service and creating a consumer is the focus of a firm's operating system. The digital transformation in progress has been questioning the very nature of what we have accustomed to when referring to a company. A platform firm (platfirm) is structured around an organisational platform and an operating system. It bases its operation on the cloud services, external connectors and identity management protocols. At the heart of the platfirm lies a hardware and software architecture that functions as a hub used for facilitating connections, resources and transactions. By doing so, a successful platform represents an ecosystem enabling a synergy of various participants all with the goal of creating value.

G. The Case of the Entrant Utility

The first steps of a new entrant utility are crucial to the entire endeavour – detailing a concrete business model and plan of actions is essential.

H. Utility Start-up First Moves

A new entrant utility needs to address the high level assessment of the key issues.

The first step is to ensure an affordable and reliable provision of energy. It is essential to leverage the high volatility of the energy market with a medium-term or a long-term arrangements. Bilateral agreements guarantee a stable, acceptable price for energy as the basis for the start of operation and acquiring customers. Every utility sells the same type of product. The only differentiation between suppliers is the price and the service. It is of a paramount importance to be able to offer a premium customer experience that gives an added value to the end consumers. In this context, a digital platform is the only one that can make a true difference. Obtaining all the necessary permits and surpassing licensing procedures is necessary to allow the legality of an energy transfer. Furthermore, the construction and operation of a district heating network and a connection to the existing heating system is of a primary importance. Smartization of the supply chain should follow, as it is the first step towards a platform based utility.

After the flow of reliable, affordable and sustainable energy has been ensured, the utility needs to acquire a customer base. Offering a premium service paired with a competitive price should facilitate gaining commercial and household consumers. During the initial stages it is crucial to establish good communication and prove as a proactive, trustworthy provider.

V. CONCLUSION

At the base of the realization of the digital energy system there is the predisposition of technological infrastructures that can support the exchange of huge amounts of data at any time, to ensure that people and objects are constantly interconnected and that is able to produce and exchange information.

We studied phases of the platform starting from the platform design, platform building, platform deployment and platform operation, while also discussing platform as an enabler of an eco-system, platform as an energy disruptor, platform for the case of an entrant utility and, finally, the utility start-up first moves.

To achieve this, the realization requires robust hyper-fast connection systems, creating smart networks capable of operating with extremely low latency to satisfy all the business and technical requirements of the fully developed ESU platforms. The systems can be a 5G, or in the near future 6G, ICT connection that will be largely implemented in the world.

REFERENCES

[1] „68% of the world population projected to live in urban areas by 2050, says UN", Retrieved June 13, 2020, from https://www.un.org/sustainabledevelopment/blog/2018/05/68-of-the-world-population-projected-to-live-in-urban-areas-by-2050-says-un/, published on 16 May 2018

[2] European Commission (EC), 2008. Communication from the Commission. 20 20 20 by 2020: Europe's climate change opportunity. Brussels, Belgium: COM; 30 final

[3] European Commission (EC), 2007. A European strategic energy technology plan (SET Plan) – towards a low carbon future. Brussels, Belgium: COM; 723 final

[4] European Commission (EC), 2015. Communication from the Commmission: Towards an Integrated Strategic Energy Technology (SET) Plan: Accelerating the European Energy System Transformation. Brussels, Belgium: C(2015) 6317 final

[5] European Commission (EC), 2011. Communication from the commission to the European Parliament, the council, the European economic and social committee and the committee of the regions. A roadmap for moving to a competitive low carbon economy in 2050. Brussels: European Commission

[6] Britannica, Retrieved June 14, 2020, from https://www.britannica.com/topic/The-Fourth-Industrial-Revolution-2119734

[7] S. S. Bhalshankar and C. S. Thorat, "Integration of smart grid with renewable energy for energy demand management: Puducherry case study," 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, 2016, pp. 1-5

[8] D. Simunic, „Preliminary study of the energy efficiency of 5G", ongoing study, 2020

[9] L. Jianming, Zh. Bingzhen and Zh. Zichao, „The smart grid multi-utility services platform based on power fiber to the home", 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, pp. 17-22, 2011.

[10] S. Lee, B. Kim, M.-K. Oh, Y. Jeon and S. Choi, „Implementation of IEEE 802.15.4g wireless communication platform for smart utility service", 2013 IEEE Third International Conference on Consumer Electronics Berlin (ICCE-Berlin), pp. 287-289, 2013.

[11] B. O. Odusote, Matthew O. Adigun, „Technologically enabling resource-constrained enterprises in developing nations through the implementation of E-Commerce on-demand service portals for a grid utility computing platform", 2014 IEEE 6th International Conference on Adaptive Science & Technology (ICAST), 2014.

[12] A. Vukojevic, S. Laval and J. Handley, „An integrated utility microgrid test site ecosystem optimized by an open interoperable distributed intelligence platform", 2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), 2015.

[13] M. Saleh, Th. Khdour, A. Ferrah, M. Qasaymeh and M. Togher, „Analysis of Digital Utility Endpoints in Smart Grid using Modular Computing Platform", 2019 Advances in Science and Engineering Technology International Conferences (ASET), 2019.

[14] J. Ma, R. Wickramasuriya, M. Safadi, T. Davies and P. Perez, „A conceptual method for modeling residential utility consumption using complex fuzzy sets", 2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), pp. 1227-1232, 2013.

[15] L. G. Swan and V. I. Ugursal, "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques, " Renewable and Sustainable Energy Reviews, vol. 13, pp. 1819-1835, 2009.

[16] a2a, Retrieved June 14, 2020, from https://www.a2a.eu/en
[17] V. Valentic, A. Viskovic and V. Franki, “Market position simulation of an independent power producer on the South East Europe electricity market,” 2014 IEEE International Energy Conference (ENERGYCON), 13-16. May 2014, IEEE, pp. 508–513, 2014.
[18] S. Parkvall et al., "5G Wireless Access - Trial Concept and Results," 2015 IEEE Global Communications Conference (GLOBECOM), San Diego, CA, pp. 1-6, 2015.
[19] M. Katz, P. Pirinen and H. Posti, Towards 6G: Getting Ready for the Next Decade, 2019 16th International Symposium on Wireless Communication Systems (ISWCS), Conference Paper, pp. 714-718, 2019.
[20] M. Matinmikko-Blue, S. Yrjölä and P. Ahokangas, "Spectrum Management in the 6G Era: The Role of Regulation and Spectrum Sharing," 2020 2nd 6G Wireless Summit (6G SUMMIT), Levi, Finland, pp. 1-5,2020.
[21] D. Rossetti di Valdalbero, B. Birnbaum,,”Towards a new economy – Co-creation and open innovation in a trustful Europe,” in W. Vassallo (ed.), Crowdfunding for Sustainable Entrepreneurship and Innovation, IGI Global, 2016.
[22] A. Founoun and A. Hayar, "Evaluation of the concept of the smart city through local regulation and the importance of local initiative," 2018 IEEE International Smart Cities Conference (ISC2), Kansas City, MO, USA, pp. 1-6, 2018.
[23] ISO/IEC JTC 1 Smart cities Preliminary Report (2014)
[24] World Energy Council. Deciding the future: energy policy scenarios to 2050. London: World Energy Council; 2007.
[25] World Energy Council. Decoding new signals of change. London: World Energy Council; 2020.
[26] International Energy Agency (IEA), 2010. Energy technology perspectives 2010: scenarios and strategies to 2050: complete edition, vol. 2008. Paris: OECD–Organisation for Economic Cooperation and Development/International Energy Agency
[27] GCCSI, March 2011. Economic Assessment of Carbon Capture and Storage Technologies. Global CCS Institute, Canberra ACT, Australia (2011 Update).
[28] GCCSI, November 2019. Targeting Climate Change. Global CCS Institute, Canberra ACT, Australia.

Modeling of Energy Consumption of Sensor Nodes

F. Tsvetanov*, I. Georgieva **
* South-West University/Department of Communication and computer technic and technology, Blagoevgrad, Bulgaria
** South-West University/Department of Electrotechnic, electronic and automation, Blagoevgrad, Bulgaria
ftsvetanov@swu.bg

***Abstract* - Power consumption is a significant problem in Wireless Sensor Networks. Most applications require wireless sensor networks to operate reliably over an extended period. It is essential to know the energy consumption of each sensor node as the life of the sensor network has a strong functional dependence on the life of the sensor node. Mechanisms for the more prolonged operation of wireless sensor networks by energy management are known. Or hardware solutions for sensor unit using renewable energy sources. The generation of an energy model for sensor nodes that can accurately reveal the energy consumption of nodes is a significant part of the development of protocols and the design of wireless sensor networks and assessment of performance in WSNs. Research shows that there is currently no standard research process to meet all the requirements of these networks.**

This article presents a model for determining the energy consumption of sensor nodes in the network using mesh topology, as energy is determined depending on the number of hops of sensor nodes, the size of the sensor communication range of sensors, and their electrical parameters. Determining the residual energy of the sensor nodes involved in the construction of the network topology allows us to predict the probability of failure of routes that build the communication in the network.

Keywords - sensor networks, energy consumption, sensor notes, simulation.

I. INTRODUCTION

Currently, wireless sensor networks have become the new paradigm for the development of network technology. Such networks consist of a large number of autonomous nodes, consisting of a transceiver, a microprocessor, a sensor device, and a power supply (battery). Network nodes, located at a small distance from each other, collect information about the environmental parameters and pass them along the chain: from node to node to the point of collection data information in storage and processing. The number of nodes in these networks can vary from several to several thousand. The fields of application of sensor networks are extensive: environmental control of large areas, protection of various objects, monitoring of production processes, control of complex engineering structures, collecting information about the status of agricultural land, monitoring of housing, and many others. Sensor nodes unevenly distributed in the net, which affect the connectivity of the network. Unevenly distributed notes may be due to the peculiarities of the problem solved. Sensor node servicing and battery replacement require considerable cost, mainly when nodes located in hard-to-reach areas. In the process of operation, the sensor nodes themselves organized into a communications network where each node uses only those functions that are necessary to solve the assigned task. Because each node has a limited radius of radio communication, the low density of the sensor distribution can lead to the separation of the network into groups whose link is missing. Even in the case of a high density of nodes, terrain features (water bodies, natural barriers, buildings, etc.) can lead to the isolation of certain groups of nodes from the main part of the network. Energy consumption is also one of the most important metrics in WSN because of the limitations in power supply to the sensor nodes. Rapid consumption of energy sensors leads to their "dying" disconnection and demand for other alternative routes from neighboring sensors, which in most cases require higher energy consumption and reduce the life of the sensor network. A major in the field of sensor networks considered the standard IEEE 802.15.4. To ensure continuous operation of the wireless sensor network to apply different tools and techniques for power management and development of the sensor unit with a self-powered mechanism. WSN applications characterized by limited resource power (batteries) and a requirement for longer life. These two so different characteristics are becoming a significant problem in the design of WSN.

Summary of Contributions. In this paper is proposed a model for determination of the energy consumption by the sensor nodes in a wireless sensing network with a mesh topology. The sensor network model for the power consumption of each node, depending on the number of sensor nodes hops, the sensor network size, the communication range of the sensors, and their electrical parameters.

II. RELATED WORKS

In sensor networks, some new issues are apparent every day, and many researchers and specialists from the industry are looking for new methods to solve them. Many researchers from around the world have been working on a study of the energy consumption of sensor nodes.

At work [1], a stochastic model presented to determine the expected power consumption of the sensor nodes.

Here, each cycle operated by the timing of occurrence of an external event (i.e., activity at the node level) and the energy consumed by a Multiple Access Control / Interference Loss Control Protocol (i.e., retransmissions, listening to channels and other network-level activities). The stochastic energy model used to calculate the expected life and life cycle of the sensor node.

In work [4], the authors present energy models of the main components of the sensor node. In proposed models, including processors, RF modules, and sensors, revealing the energy correlations between the components of the node and then establishing the nodal energy model based on the mechanism trigger by events. Thus, the proposed model enables the analysis and evaluation of WSN energy consumption under different communication protocols [2]. Proposed an energy model for WSNs based on the physical layer and the parameters of the MAC layer for determining the energy consumption of each bit of data that has passed without an AWGN channel error. It shows how to select transmission power to achieve energy-efficient AWGN channel communications.

The analysis of the studies conducted at work [5] found that efforts to solve the problem by various authors directed to the application of energy efficiency techniques and design optimization of WSN platforms, as well as the use of system software and protocols. Simulation tools and communication protocol modeling technologies such as OPENET, NS2, SHAWN, SensorSim, EmStar, OMNet, GloMoSim, and TOSSIM have been researched and studied. WSN applications can't be unattended and need to monitored to fulfill the essential function of collecting, processing, and transmitting data. Due to the low accuracy caused by errors or inaccurate models of systems, current platforms (tools) to simulate WSN seriously affected the development of the protocol, system design, and analysis of the performance of the WSN system and thus hampered the industrialization of WSN.

The authors [6] are reviewing a stochastic model of Wireless Sensor Networks (WSNs) in which each sensor node accidentally and alternatively remains inactive or sleep mode. The active mode consists of two phases, called the full active phase and the half-active phase. When an appropriate sensor node is in the active phase of the active mode, it can receive packets and forward received packets. However, when the phase of the sensor node passes from the full active phase to the semi-active phase, it is only capable of transmitting data. When the corresponding node of the sensor is in sleep mode, it does not interact with the outside world. Based on the above understanding of the operating modes of sensor nodes, authors developed a stochastic model for the WSN sensor node and offer an explicit expression of the stationary distribution of the number of packet data in the sensor unit. They also analyze some important performance measures, including the energy of the sensor node, the power consumption of the transmission, the energy consumption of the sensor operations, and the average power consumption of the active node sensor in sleep mode. Not consider the model to determine the energy consumption of sensor nodes in a wireless sensor network with a mesh topology.

III. MODEL TO DETERMINE THE ENERGY CONSUMPTION OF SENSOR NODES IN A WIRELESS SENSOR NETWORK WITH MESH TOPOLOGY

WSN's using mesh topology is dynamically self-organizing and self-configuring networks whose network nodes automatically create an ad hoc network and maintain connectivity of the devices [7]. Mesh networks differ from other network topologies in that individual nodes can connect through multiple hops. There are two types of topology: Full-Mesh and Partial-Mesh. At the full mesh topology, communication takes place at the nodes "each with each," and there is a great deal of redundancy. This type of network used when there can be no interference in communications, for example, at a nuclear power plant. With the partial mesh network, is reducing the number of connections and reducing links are accomplished by using routing protocols to determine optimal routes when transmitting data from sensor nodes to the coordinator. A typical architecture of a sensor mesh topology shown in Figure 1.

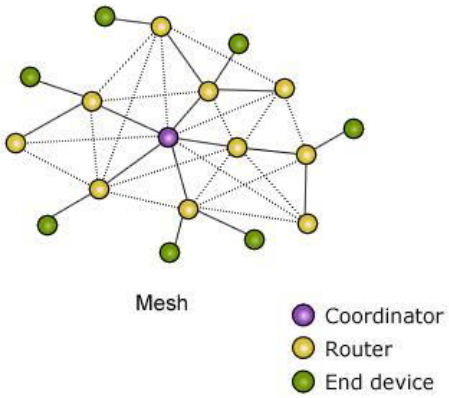


Figure.1. The wireless sensor mesh architecture

WSN used Mesh topology have advantages over other wireless networks. Easy deployment, high reliability, self-configuration, self-diagnosis and scalability, wireless transport channels, access to "everyone with everyone" network resistance in some nodes, and thus reduce the cost and complexity of fully connected topology.

A feature of mesh networks is routing protocols which, through special algorithms, allow access points to create a list of network subscribers, perform traffic level control, maintain optimal dynamic routing of traffic, or search for the best quality road between neighbor devices. When a device for any reason failed and not transmit data, traffic automatically redistributed on another route, ensuring not only data transmission but also transmission with minimal delay. Wireless sensor mesh networks built from four types of devices:

- *A mesh coordinator* who, when dropping out of the network, is replaced by another full-function device on the network;
- *Mesh routers* that form a wireless highway network and provide dynamic routing of

information packets. Practically, mesh-routers are stationary nodes. To access other networks is used as a mesh gateway;

- *Mesh clients* appear to end devices on the network and can be fixed or mobile;
- *The mesh access point* is a node that allows different devices to connect to the network.

The main tasks that arise when designing and building wireless mesh networks related to traffic optimization, the energy efficiency of sensor nodes and communication efficiency that directly related to the length of the network routes [3],

Due to the high interest in the BSM using mesh topologies further, improvement is of interest to researchers. And in this work, we provide a model of the sensor network using mesh topology to conduct simulation studies related to the energy consumption of these networks [8].

It assumed that n sensor devices are located in a two-dimensional coordinate system and have the same maximum radio range. The network modeled as a connected UDG count, G (V, E), where V is the set of nodes, and E is a set of edges. An edge $\{u, V\} \in E$, only if the distance between nodes u and V is within the communication range. It assumed that the Euclidean length $d_{i,j}$ between two vertices $i, j \in V$ is the physical distance between the nodes represented as matrix E. Finding the optimal path between two nodes, determine the shortest distance according to equation one by applying Dijkstra algorithm by the grShortPath function included in the Matlab grTheory toolbox.

$$[dSP, spi, j] = grShortpath\left(E, ID_i ID_j\right) \quad (1)$$

Data transmission in mesh networks is using a multi-hop. The routing table contains all possible paths and intermediate nodes to reach the destination. To maximize energy efficiency, it is important that the routes are as short as possible. The Greedy algorithm used to find the shortest path as it strives to find the shortest path with the least steps to the coordinator in the network. It assumed that the mesh network built with ZigBee devices operating on the CSMA/CA algorithm and the main mode of operation includes data reception, data transmission, channel listening, and sleep mode, is essential for sensor nodes. ZigBee packages have a maximum size of 128 bytes, including protocol overhead. When real-time data transmitted to ZigBee, it is possible to define higher message priority. It achieved by using the guaranteed time mechanism so that high priority messages can be sent as quickly as possible. During sleep, devices only support the timer, which virtually allows a reduction in power consumption below 1 μA. Inactive mode, each query receives an ACK acknowledgment that informs about the existence of a two-way connection. All soo contains information on whether the control node is waiting for a message. If there is a message, the device does not go into sleep mode but sends a Roll Request message. Between the Roll Request, the module includes the radio system and switches to the *Idle* mode where the current consumption is about 15 mA. We assume that a node can wake up for data transmission at any time, but can only receive data in its active time slot. We assume that the power consumption per node determined according to dependence 2 [8].

$$E_{kons,n_i} = E_{rx} + E_{tx} + E_{listen} + E_{edle} + E_{sleep} \quad (2)$$

Where: E_{rx} is the energy to receive the data, E_{tx} is the data sending energy, E_{listen} energy to listen to the channel, E_{edle} energy in standby, E_{sleep} energy needed for sleep mode.

IV. Simulation Research of Sensor Network with Mesh Topology

The research carried out on WSN built with XBee ZB devices using a mesh topology. With Matlab, we developed program code to carry out simulation studies, which allows practical experiments with an unlimited number of sensor nodes, areas, and communication range of the sensor devices. Simulation tests with the simulation parameters shown in Table 1.

TABLE 1. SIMULATION PARAMETERS OF STUDY SCENARIOS

Simulation parameters	***Scenario 1***	***Scenario 2***	***Scenario 3***	***Scenario 4***
Area of mesh, meters in x and y	150x150	200x200	200x200	300x300
Total number of sensor nodes	25	35	45	45
Location of the sink	X=75 m Y=75 m	X=100 m Y=100m	x=150 m; y=150m	x=50 m; y=50m;
Sensor communication range	80 m	60 m	60 m	60 m

A 3V battery with a power of 675 mJ, powered by the communication modules designed to operate in the wireless network with mesh topology, used to conduct the research. It envisaged that the energy consumption for the receiving, transmitting, *idle,* and devices in sleep mode, accordingly mJ 6,3, 9,04 mJ, 1mJ, and 0,005 mJ. Time interval, same for all nodes. When determining the communication range of the devices, the actual parameters for devices from the XBee ZB specification for outdoor mesh topology that are up to 120 meters taken into consideration, the modeling of the topology accomplished using a graph formed by points in the Euclidean length bounded within the range of the communication range. Each node sends a packet to the coordinator node, and each transmission performed over a one-time slot. Determine the residual energy for each node that presented in the energy table. The simulation time is of great importance, but for the simplicity of the model, it assumed that the transmission of the data packet from the starting node to the final target takes one simulation second, and then the time interval is updated.
When simulating the sensor networks, the location of the nodes in the net is stochastic. Examples of such a sensor placement shown in Figures 2, 4, 6, 8, according to the scenarios shown in Table 1.

The experimental sensor network in Figure 2 contains 25 sensors, and the network coordinator is located approximately in the middle of the network with the number 15. The communication range of the sensors is 80 meters. There are 261 possible connections on the network.

Figure 4 shows an experimental sensor network with 35 sensors, the location of the coordinator is also approximately in the middle of the network, the communication range of the sensors is 60 meters, and the number of possible connections is 247.

The experimental network of Figure 6 simulated at 45 sensors, a communications radius of 60 meters, and a coordinator node 21 located at the middle of the network, possible connections is 505.

Figure 8 shows the experimental sensor network with 45 nodes, but the coordinator positioned in the lower right corner with number 28, and possible connections 393.

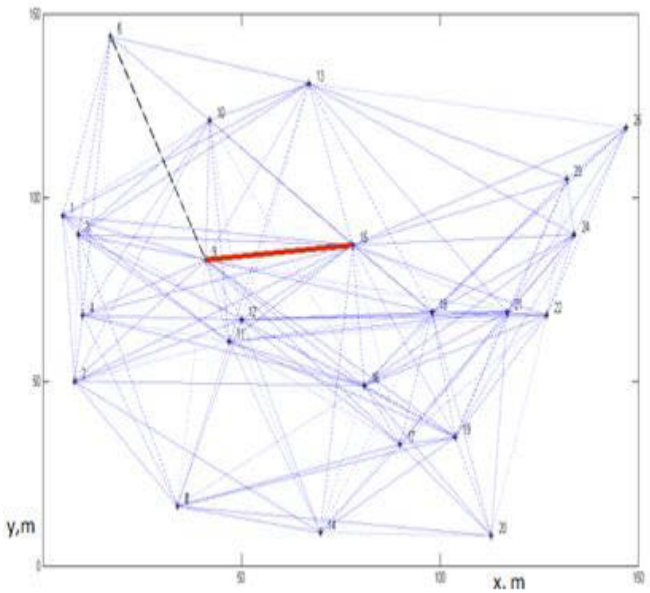


Figure 2. Location of the sensors in an experimental network of 25 sensors

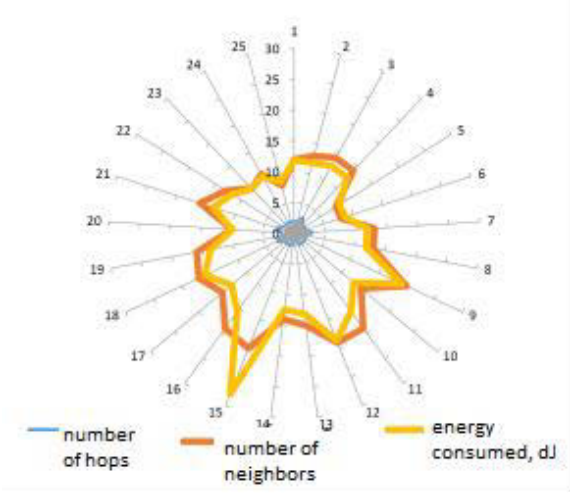


Figure 3. Results of network simulation studies with scenario 1

The graphs in Fig. 3, 5, 7, 9 visualize the simulation results of each sensor network for the number of hops to reach the coordinator from each sensor node, number of neighbors of each node, energy consumed by each node for transfers.

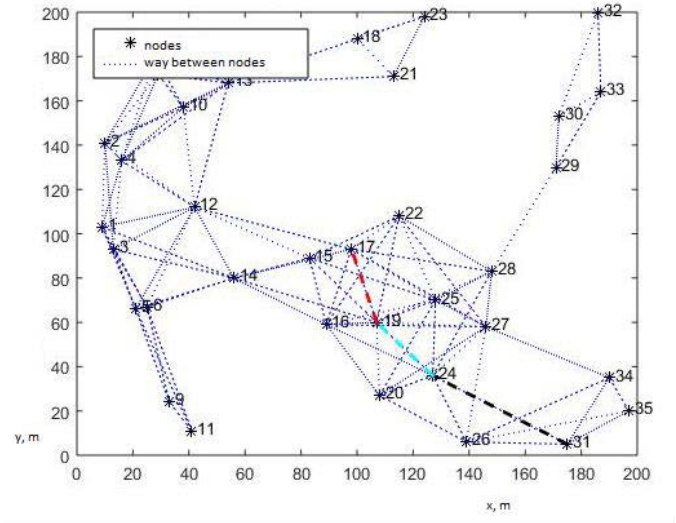


Figure 4. Sensor positioning with a network of 35 sensors

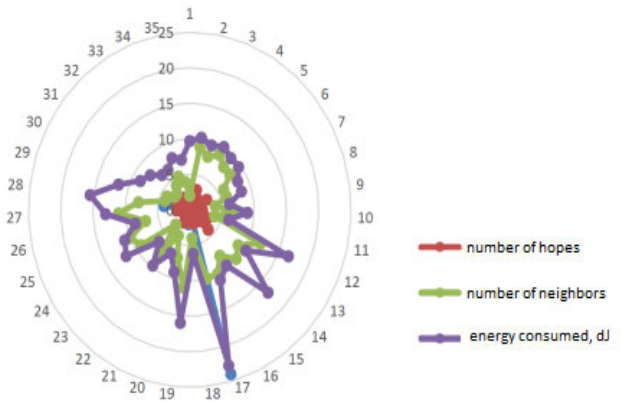


Figure 5. Results of simulation studies on network scenario 2

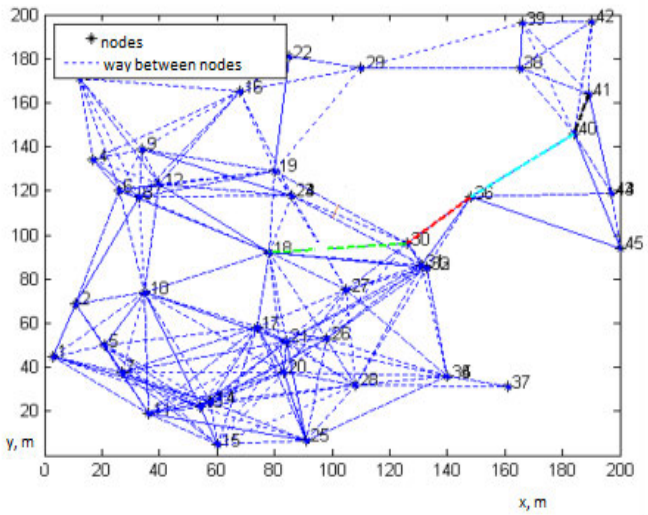


Figure 6. Sensor positioning with a network of 45 sensors and a coordinator in the middle of the network ID 18

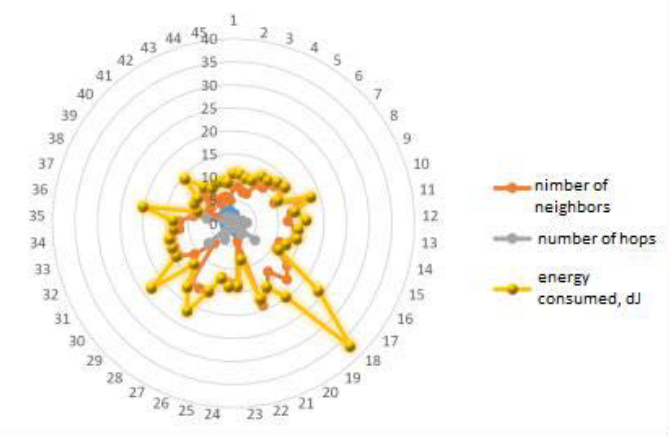


Figure 7. Simulation results of network with scenario 3

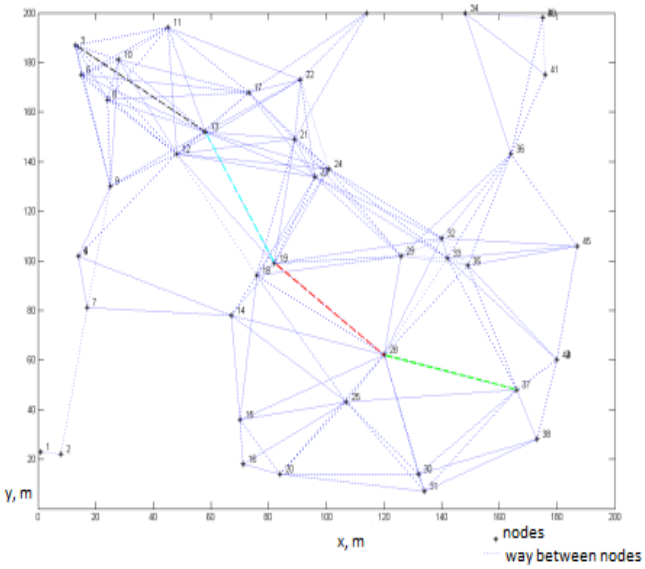


Figure 8. Location of 45 nodes and coordinator in the lower right corner - ID 28

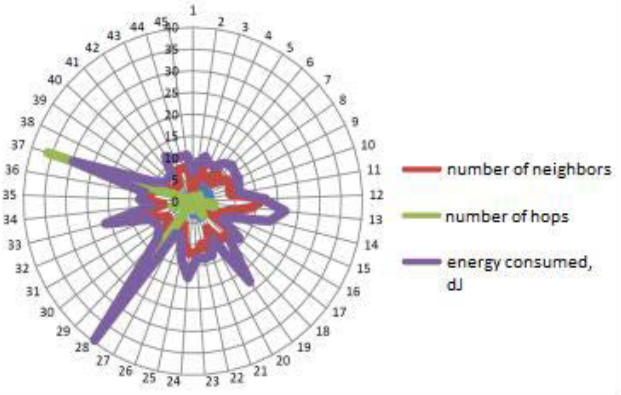


Figure 9. Results of simulation studies on network scenario 4

Table 2 shows the results of simulation studies regarding energy sensors and metrics for assessing the existence of minimum and maximum energy consumption of devices to sensing data.

TABLE 2. METRICS FOR MINIMUM AND MAXIMUM ENERGY CONSUMPTION OF SENSOR NODES

Metrics	***Scenario 1***	***Scenario 2***	***Scenario 3***	***Scenario 4***
E_{min}	327,7	285,5	447,4	475,4
E_{max}	580,7	524,4	612,7	619,0
$E_{average}$	454,2	404,95	530,05	547,2
Number of network connections	261	247	505	393

Analysis of the results for the conducted experiments of the research energy consumption of the nodes in the network shows:

- The more large number of sensors in the network, the higher the energy consumption of sensor nodes for data transmission because of the presence of more links and possible routes to reach the coordinator;
- When the coordinator located in the middle of the sensor network, the number of hops influenced most heavily on energy consumption;
- For networks with the coordinator at the edge of the network, the sensor's energy resources are exhausted more quickly due to the increased number of hops and longer routes to the coordinator;
- The most energy-efficient network is when the coordinator located in the middle of the network;
- It is necessary precisely to select a radio communication range of devices to reduce the excesses of connections and more efficient use of energy resources available for each device in the mesh network.

V. CONCLUSION

This paper presents the results of simulation studies, describing the energy consumption of sensors in a sensor network with a mesh topology.

Experimental studies established a link between the number of sensors, the location of the coordinator, the number of possible connections, and the number of possible routes. On the metrics to assess the energy consumption of the sensors on the network, the chosen topology gives a good result.

An important challenge for scientifically proven experimental results is the proper planning of the experiment, the selection of the number of sensors in the network, the radiocommunication range of the sensors, and the correct location of the coordinator in the network.

REFERENCES

[1] V. Agarwal, R.A. DeCarlo; L.H. Tsoukalas, "Modeling Energy Consumption and Lifetime of a Wireless Sensor Node Operating on a Contention-Based MAC Protocol", Published in IEEE Sensors Journal, Volume: 17, Issue: 16, p.p. 5153 – 5168, Aug.15, 2017.

[2] A. Ali, M. Abo-Zahhad, M.Farrag, “Modeling of Wireless Sensor Networks with Minimum Energy Consumption”, Arabian Journal for Science and Engineering, Volume 42, Issue 7, pp 2631–2639, July 2017.

[3] F. Bokhari, Zaruba G., “The use of smart ants for efficient routing in wireless mesh networks,” International Journal of Wireless & Mobile Networks, Vol. 4, No. 2, p.p 113 – 134, 2012.

[4] H. Y. Zhou, D.Y.Luo, Y. Gao, D.C. Zuo, “Modeling of Node Energy Consumption for Wireless Sensor Networks,” Wireless Sensor Network” 3, p.p.18-23, 2011.

[5] S. Chen, L.Chen, R. Shu, “Energy Consumption Model for Wireless Sensor Networks,” Advanced Materials Research, Vols. 588-589, pp. 664-667, 2012.

[6] Y. Zhang., Wei Li., “Modeling and energy consumption evaluation of a stochastic wireless sensor network,” EURASIP Journal on Wireless Communications and Networking, Article number: 282, volume 2012.

[7] F Tsvetanov., K Tsvetanov., I Georgieva., E. Ivanova., (2014), Modelling and Simulation of Communication Efficiency in Low-Speed Networks, Proceedings of the International conference, TELFOR’2014 p.p. 178-182, 2014.

[8] F. Tsvetanov, Investigation of Low-Speed Communication Networks, Ph.D. theses, University of Rousse Publishing, 2014.

[9] T. Iliev, Bikov T., Mihaylov G., Ivanova E., Stoyanov I., Keseev V., Algorithms for Congestion Control in LTE Mobile Networks, Advances in Intelligent Systems and Computing book series (AISC, volume 763), p.p 474-484, 2018.

Čimbenici adresiranja i povezivanja za IoT usluge i aplikacije

Darko Švelec, Mislav Blažeković
HT d.d., Osijek, Hrvatska
darko.svelec@t.ht.hr, mislav.blazekovic@t.ht.hr

***Sažetak* - Za IoT se očekuje da će biti jedan od ključnih stupova 4IR. U svojoj infrastrukturnoj osnovici imaju M2M kao već poznatu ICT uslugu prijenosa podatka i 5G kao ICT infrastrukturnu osnovicu prijenosa podatka koja tek dolazi. Obje ove usluge imaju u svojoj pozadini korištenje prirodno ograničenih općih resursa, brojeva i adresa za M2M i IoT te frekvencijskog spektra za 5G. Ovi resursi nisu neiscrpni, a jedan od ključnih elemenata 4IR je masovno korištenje novih i modernih ICT usluga i potreba za velikom količinom krajnjih komunikacijskih točaka koje treba adresirati i međusobno povezati s visokom razinom kvalitete komunikacije.**

Rad razmatra načine rješavanja problema korištenja tih ograničenih resursa, s naglaskom na izazove koji tek dolaze, osobito u području 5G.

Ključne riječi - IoT, M2M, 5G, Resursi, Aplikacije

I. UVOD

Četvrta industrijska revolucija (4IR) postala je to kada su se dogodile dvije bitne evolucije. Jedno je evolucija mobilne mreže, pojavom LTE koncepta koji je otvorio vrata daljem, iz sadašnje perspektive, gotovo neograničenom razvoju. Taj razvoj, dodatno, jednako ide u pogledu tehničkih mogućnosti mreže kao i u pogledu funkcionalnih osobina mreže. Druga evolucija je evolucija M2M komunikacije prema IoT konceptu umreženog povezivanja. IoT koncept umreženog povezivanja usmjerio je razvoj i drugih funkcionalnih komponenti 4IR, dok sam razvoj mreže osigurava ključnu infrastrukturnu podršku. Konačno, sve je u međusobnom povezivanju, koje prelazi granice mrežnog povezivanja. Koncept IoT osigurava metodologiju, a mreža transportnu osnovicu.

Za uspješnu funkciju povezivanja bitne su dvije komponente: adresiranje terminalnih odredišta (ne nužno i ne samo uređaja) i mogućnosti transportne osnovice. I jedno i drugo određeno je osobinama prirodno ograničenih općih resursa, brojevima odnosno adresama za adresiranje točaka komunikacije i frekvencijskim spektrom potrebnim za prijenos podatka.

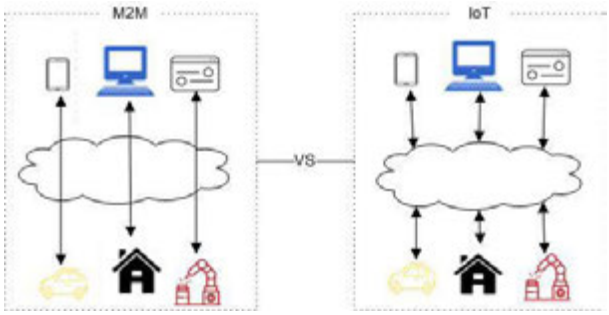
II. M2M I IOT

M2M usluga ima svoju dugu povijest i počinje još od telemetrije i općenito se odnosi na razmjenu podataka između različitih uređaja putem komunikacijske mreže bez ljudskog sudjelovanja. Mogućnosti digitalnih komunikacijskih mreža, osobito mobilnih, dale su kasnije toj usluzi novu dimenziju i potaknule njen veliki razvoj. Vrhunac M2M usluge dosežu početkom ovog stoljeća razvojem GSM mreža. M2M usluga je u osnovi usluga komunikacije stroj-stroj i općenito se odnosi na razmjenu podataka između različitih uređaja bez ljudskog posredovanja. Osnovno je obilježje te komunikacije da se radi o komunikaciji točka-točka. Pri tome treba imati u vidu da to ne znači da je to istovremeno i povezivanje isključivo između dva uređaja (stroja). Može se raditi o povezivanju više uređaja. Međutim, to je vremenski odijeljeno, komunikacija je uvijek između dviju točaka, a komunikacijom upravlja neki automat ili lokalna programska aplikacija. Razvojem tehnologije i ovdje se događaju promjene u smislu udaljenog ili adaptivnog upravljanja komunikacijom, ali ona ostaje orijentirana kao točka-točka.

Paralelno s razvojem mrežnih mogućnosti i ICT rješenja razvija se i kompleksnost usluge, njena primjena i prisutnost i značaj u svijetu ICT usluga. Razvojem interneta kao mrežne osnovice povezivanja krajnjih točaka dolazi do evolucije M2M usluga u dva smjera. Prvi je dalji i ubrzani razvoj unutar samog područja M2M komunikacije i kroz internet, a drugi je prelazak tehnološke granice i pojava IoT usluga koje se temelje na M2M konceptu. Iako se na prvi pogled može činiti da se radi o sinonimima, pa čak mnogi autori i promatraju jedinstveno M2M i IoT, IoT ima bitnu karakteristiku koja ih razlikuje i definira kao različite tehnologije (slika 1). Kao jedan prethodnik IoT-a, M2M podrazumijeva interakciju dvaju ili više strojeva bez ljudske intervencije. S druge strane, jedan od komunikacijskih temelja ideje IoT usluga je upravo u M2M konceptu usluga, ali ima za cilj pružiti puno više funkcionalnosti od samog prijenosa podataka te se stoga stalno mijenja i razvija. Brzina kojom se internetska tehnologija razvija zadnjih desetak godina uzrokovala je da M2M koncept ustupa mjesto IoT konceptu koji postaje općenito primarni koncept koji oblikuje način na koji razmišljamo o budućnosti poslovanja i našem svakodnevnom životu.

Stoga je razumijevanje razlika između Interneta stvari i komunikacije stroj-stroj (M2M), što je temeljni koncept koji je stvorio IoT kakav znamo, bitno za razumijevanje potreba M2M i IoT na adresiranje točaka komunikacije.

Sam prelazak na Internet ne znači automatski i promjenu M2M usluge u IoT usluge. Internet kao komunikacijska osnovica omogućava i M2M i IoT usluge, a razlika je i dalje u načinu komunikacije. U pogledu komunikacije kod M2M to je komunikacija točka-točka



Slika 1. Komunikacija M2M vs IoT

koja u IoT uslugama prelazi u komunikaciju više_točaka-više_točaka (slika 1).

Uz razlike u načinu na koji se daljinski pristupa uređajima, postavlja se i pitanje skalabilnosti. IoT je svojstveno skalabilniji, jer obuhvaća niz najsuvremenijih rješenja za povezivanje kako bi povezao, čak i integrirao, različite tehnološke grupe. Zauzvrat, njih obično pokreću M2M platforme. U tom pogledu, M2M može se smatrati lošijom ili ogoljenijom verzijom IoT-a, što nudi ograničene mogućnosti integracije jer uređaji koje povezuje moraju imati zajednički komunikacijski standard.

U svakom slučaju, može se smatrati da su IoT i M2M usluge tehnološki slične po tome što obje nude rješenja za prikupljanje, pohranu i razmjenu podataka između uređaja pod minimalnim ljudskim nadzorom. Ono što ih također snažno povezuje je činjenica da su oni u vezi upravljanja uređajima sve bliži jedni drugima. Također, razvojem M2M protokola za internetsko povezivanje razlike između M2M i IoT postaju sve nejasnije. Prikaz osnovnih razlika M2M i IoT koncepta ICT usluga dan je Tablicom I. Stoga bi moglo biti pitanje u skoroj budućnosti da li će oba termina steći isto značenje i da li će vremenom postati sinonimi. Odgovor na ovo pitanje je opet u komunikaciji, kako se to i vidi iz slike 1.

TABLICA I. RAZLIKE IZMEĐU M2M I IOT

M2M	***IoT***
Komunikacija uređaj-uređaj, uobičajeno s upravljačkim programima na strani klijenta	Velika (masovna) povezivanja, pristup „svega svemu“
Povezivanje izoliranih sustava uređaja koji koriste iste standarde	Integrirano povezivanje uređaja, podataka i aplikacija, korištenje različitih standarada
Ograničene mogućnosti skalabilnosti	Skalabilnost integralni dio povezivanja
Žična ili bežična mreža je komunikacijska osnovica	Internet je komunikacijska osnovica
Komunikacijska osnovica je pasivni element povezivanja	Komunikacijska osnovica je aktivni sudionik povezivanja; uređaji traže aktivnu Internet mrežu
Velika baština povijesnih aplikacija	Kontinuirani razvoj novih aplikacija temeljenih na M2M korijenima i uz primjenu modernih tehnologija

III. NUMERIRANJE I ADRESIRANJE

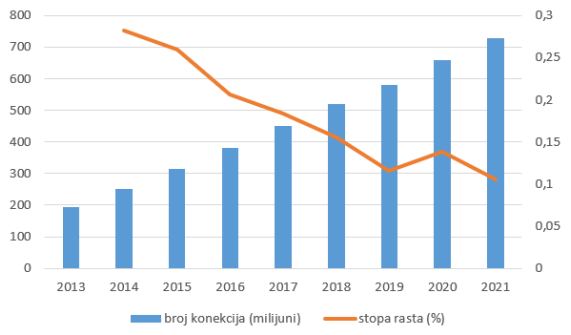
Svako "adresiranje" koristi brojeve. Ovi brojevi dolaze iz dogovora o načinu korištenja. Taj dogovor može ima dvije pojavnosti, protokol ili plan (numeriranja ili adresiranja). Iako su obje takve adrese brojke (znamenke), nisu oba broja prirodno ograničena dobra. Prirodno ograničeno dobro su samo brojevi koji dolaze iz plana numeriranja jer je ograničenje upotrebe posljedica količine dostupnih brojeva za potrebe adresiranja. Plan unaprijed definira skupine brojeva dostupne za korištenje.

A. M2M numeriranje

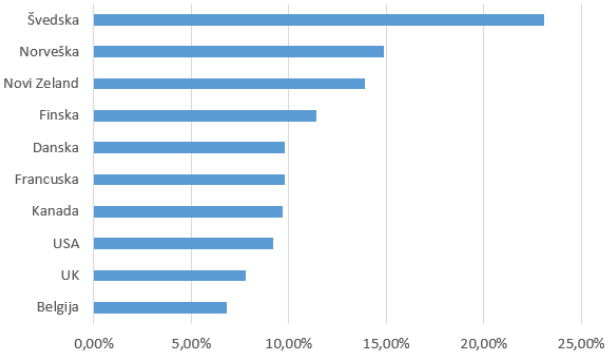
Pojednostavljeno, M2M se obično oslanja na komunikaciju koja se pruža putem mobilne ili, sada već puno rjeđe, fiksne mreže. Stoga M2M usluge za svoju komunikaciju za adresiranje krajnjih točaka komunikacije i dalje koriste brojeve povezane s tradicionalnim komunikacijskim mrežama.

M2M usluge u svijetu interneta koriste i adrese iz IP svijeta svrsishodne Internetu. Ali je pitanje možemo li ih promatrati kao čiste M2M usluge ili kao dio IoT spektra usluga. Odgovor, barem za sada i još neko vrijeme, proizlazi iz načina komunikacije. Prostor M2M-IoT je veoma fluidan i čiste M2M usluge lako postaju IoT usluge, i obrnuto. Međutim, prostor za adresiranje krajnjih točaka komunikacije je isti.

U području čistih M2M usluga u velikoj mjeri je to i dalje načelno isti tip usluge te se povezivanje krajnjih točaka komunikacije odvija kroz tradicionalnu telekomunikacijsku mrežu, što je samo po sebi prilično ograničavajuće. Stoga se postavlja i pitanje perspektive takvih usluga. Međutim, količina i raznovrsnost M2M orijentiranih usluga se povećavaju. Pri tome je taj porast oko 2010. g bio veoma brz. Kasnije se je rast usporio, ali je i dalje respektabilan, a stopa rasta je stabilna. Slika 2. daje prikaz promjene broja konekcija M2M usluga u milijunima konekcija [1] i pokazuje stabilan rast. Tehnološki, u pružanju M2M usluga dominira LTE tehnologija i očekuje se da će tako biti i dalje. S obzirom da se i dalje radi o tradicionalnoj telekomunikacijskoj mreži kao transportnoj osnovici komunikacija je i dalje točka-točka. Stoga i povijest razvoja M2M usluga kao i dalji očekivani razvoj daju naglasak na brojevima kao prostoru povezivanja krajnjih točaka komunikacije. U nekim zemljama, osobito u Europi, M2M postaju iznimno značajan dio ukupne komunikacije [2], kako se to može vidjeti sa slike 3. S obzirom da se radi o 2013. godini i da u pogledu M2M komunikacije dominira mobilna mreža



Slika 2. M2M konekcije u svijetu



Slika 3. Udio M2M konekcije u ukupnom prometu

pojavila se zabrinutost mogu li postojeći planovi numeriranja zadovoljiti rastuće potrebe M2M usluga.

Ovakav porast M2M usluga i osobito predikcije rasta oko 2010. godine otvorile su pitanje mogućnosti adresiranja krajnjih točaka komunikacije u postojećim telekomunikacijskim mrežama. M2M usluge kao područje adresiranja krajnjih točaka svoje komunikacije koriste postojeće nacionalne planove numeriranja. Po definiciji, Plan numeriranja uređuje ukupnost svih mogućih kombinacija sastavnica adresiranja uz pomoć znamenaka radi jedinstvenog prepoznavanja osoba, računalnih procesa, strojeva, uređaja ili elektroničke komunikacijske opreme koja je uključena u postupak ostvarivanja veze. Ovo je osobito dolazilo do izražaja u zemljama s jače reguliranim pitanjem planova numeriranja, recimo skandinavskim zemljama. Dodatno je to pitanje potencirano naglom i značajnom migracijom M2M usluga u mobilne mreže. Za razliku od elektroničkih usluga u fiksnoj mreži za koje su planovi numeriranja već poznavali neku segmentaciju po vrstama usluga, za elektroničke usluge u mobilnoj mreži u pravilu nije postojala nikakva segmentacija. M2M usluge rade s korištenjem brojeva iz aktualnih nacionalnih planova

TABLICA II. M2M NUMERACIJSKI RESURSI

	M2M	
zemlja	***NDC***	***N(S)N duljina***
Bugarska	430	12
Austrija	not spec.	13
Belgija	77	13
Bosna i Hercegovina	85	9
Hrvatska	89xy/890xy	10
Danska	37	12
Finska	49	11
Francuska	7000 to 7004	13
Njemačka	15020	11
Mađarska	71	12
Island	352	9
Irska	88	12
Italija	nije određeno	12
Luksemburg	nije određeno	12
Srbija	72	12
Slovenija	91	12

numeriranja. Korištenje ovih postojećih numeričkih resursa u uvjetima brzog rasta potreba za brojevima za M2M usluge može dovesti do situacije da resursi brojeva za adresiranje krajnjih točaka, općenito, ne samo za M2M usluge, postanu nedovoljni. To posebno može utjecati na postojeće komunikacijske usluge. S druge strane, usluge M2M zahtijevaju povezivanje na velike daljine, uključujući povezivanje preko državnih granica, visoku pokrivenost te fleksibilnost i skalabilnost resursa pri realizaciji. Stoga je izazov pronaći posebno rješenje za učinkovito i održivo adresiranje krajnjih točaka M2M usluge. Kao odgovor na tu situaciju u dijelu svijeta, osobito u Europi, pristupilo se formiranju posebnih numeracijskih područja namijenjeno isključivo M2M uslugama. M2M brojevi i adrese sve su češće posebno navedeni u nacionalnim planovima i u različitim formatima kako bi bili dovoljno perspektivni za nadolazeće usluge. Zbog potrebe široke uporabe i jednostavne primjene, pronađeno je rješenje u preporuci E.164 na način da maksimalno poveća količinu brojki (znamenki) u broju. Posebni rasponi brojeva za M2M brojeve uspostavljaju se putem nacionalnih planova numeriranja. Tablica II prikazuje shemu numeriranja i raspon M2M za neke zemlje. U nekim zemljama je raspon za M2M jasno definiran, neke države ne poznaju raspon M2M, a u nekim je zemljama plan numeriranja osmišljen tako da bude neovisan o posebnim zahtjevima prema namjeni ili usluzi. Broj znamenki (NSN duljina) određuje maksimalni kapacitet adresiranja. Njihov broj varira od zemlje do zemlje, od 9 do 13, što također ovisi i o duljini nacionalnog prefiksa. Takvo rješenje je globalno održivo sljedećih 15-20 godina, pretpostavljajući rast od 20% godišnje. Većina planova numeriranja za M2M brojeve uzima veću duljinu NSN-a nego li je to uobičajeno. I to pokazuje razinu očekivanja M2M usluga.

B. IoT adresiranje

IoT za transportnu osnovicu koristi Internet mrežu, stoga je i adresiranja drugačije nego za M2M usluge. IoT je komunikacijska infrastrukturna platforma koja omogućuje povezivanje ljudi i stvari u bilo koje vrijeme, bilo gdje, sa bilo čim i bilo kime, u idealnom slučaju pomoću bilo koje mreže i bilo koje usluge. Komunikacija utemeljena na konceptu IoT ima integrirajuću ulogu. Promatrajući E2E povezanost kao temelj IoT-a, postoje tri osnovne vrste IoT komunikacije:

1. komunikacija uređaja s ljudima (uključujući umjetnu inteligenciju)

2. komunikacija između uređaja

3. komunikacija stroj-stroj, ili M2M

Sve ove vrste IoT komunikacije mogu se povezati s IoT okvirom 4 uobičajena modela komunikacije: uređaj na uređaj, uređaj na oblak, uređaj na aplikaciju, aplikacija na bazu podataka [3].

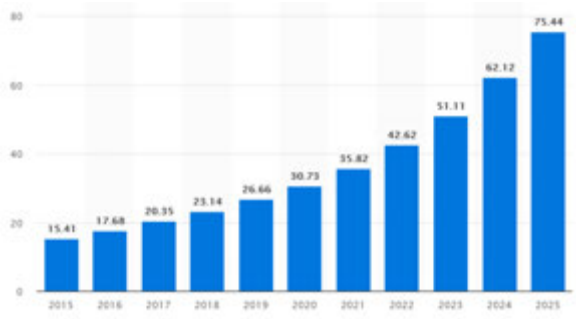
Međutim, razmatranje putem vrsta komunikacije umjesto preko komunikacijskog modela je primjerenije ako se E2E IoT komunikacija promatra s gledišta točke komunikacije prezentirane brojem, dakle sa stajališta brojeva kao prirodno ograničenog dobra.

U IoT društvu pametni telefoni, tableti i slično postaju ključni komunikacijski uređaji. U isto vrijeme, oni postaju uređaji za prilagođavanje oblicima komunikacije koji korisniku najviše odgovaraju. Način uspostavljanja komunikacije ovisi o vrsti usluge i korisničkom sučelju.

IoT je vodeća tehnologija 4IR i vodeći pokretač razvoja modernog društva. Predviđanja su da će se u sljedećih oko 7 godina broj priključenih uređaja udvostručiti, preko 75 milijardi [4], kao što je prikazano na slici 4. Gotovo svaki drugi IoT uređaj 2018. godine instaliran je u Europi, a 2020. taj će broj porasti na dvije trećine [5]. Međutim, postoje i znatno optimističnije prognoze. Mnoge od njih predviđaju da će 2020. godine širom svijeta postojati više od 50 milijardi IoT povezanih uređaja.

S obzirom na takvu situaciju i takva očekivanja "adresiranje" komunikacijskih točaka moglo bi biti važno pitanje. Međutim, komunikacija koju IoT usluge koriste između krajnjih točaka (E2E) uključuje i mnoge druge točke unutar mreže. Stoga treba razlikovati potrebe i uvjete „adresiranja“ komunikacijskih točaka s gledišta njihovog položaja unutar E2E komunikacije - na krajnjim točkama komunikacije ili unutar same komunikacije.

Međutim, kako je komunikacijska platforma Internet, "adresiranje" komunikacijskih točaka unutar mreže oslanja se na sam internet protokol, IPv4 i višeslojne mrežne protokole poput TCP (RFC 793), HTTP (RFC 2616) i DNS (RFC 1035). Iako se internet protokol oslanja na brojeve, metode i postupke koje koristi su takvi da zapravo nije prirodno ograničeno dobro iako ima ograničenja na adresiranje i mogućnosti adresiranja unutar IPv4 nisu dostatni za rastući promet. IPv4, kao trenutni protokol, dizajniran je za sve potrebe i koristi se u daleko većoj mjeri nego što je prvotno planirano. Međutim, postoji problem s podkapacitiranim mogućnostima adresiranja unutar IPv4. Rješavanje ovog problema nalazi se u nekoliko načina adresiranja, temeljenih na privatnim dinamičkim ili statičkim IP adresama i javnim dinamičkim IP adresama. Sve to komplicira komunikaciju, ali rješava pitanje "adresiranja" osobito krajnjih točaka komunikacije, ali i nekih drugih pitanja poput visokih troškova održavanja IP adresa i sigurnosnih rizika stalnih javnih IP adresa. Sam protokol IPv4 vrlo je uspješan protokol. Nadmašio je ciljeve svog originalnog dizajna, s obzirom na njegovu svrhu i opseg korištenja. S obzirom na takav uspjeh, očito je da brojevi kao osnova „adresiranja“ ne predstavljaju takvu prepreku da se u ovom slučaju tretiraju kao prirodno ograničeno dobro. S druge strane, IPv6



Slika 4. Količina IoT 2015 - 2025 (milijarde)

protokol („sveti gral“ komunikacije) zaslužuje da se i ovdje spomene. Taj protokol još nije sazrio. Različiti izvori izjavljuju da se koristi u 5-25% slučajeva, a uskoro se ne očekuje široko korištenje, iako IoT ne može čekati. No, s obzirom na relevantnost korištenja brojeva, ima iste značajke kao i IPv4.

Točke unutar komunikacije su sve "stvar", dok su krajnje točke komunikacije često ljudi. Stoga „adresiranje“ krajnjih točaka komunikacije mora biti prilagođeno ljudima. IPv4 protokol, posebno zbog njegovih varijacija metoda adresiranja, nije usklađen s ljudskom prirodom i zapravo nije namijenjen ljudima.

S obzirom na načela adresiranja i točke komunikacije postoje dvije osnovne vrste adresiranja - neizravno „adresiranje“ i izravno "adresiranje". Ako razmatranje usmjerimo na brojeve, neizravno adresiranje koristi brojeve na sekundarni način, dok se primarno koristi protokol komunikacije koji upravljaju brojevima. Komunikacijski protokoli koriste brojeve kao dio algoritama upravljanja i oni su podložni manipulacijama i transformacijama. Pri izravnom adresiranju brojevi se koriste izravno, komunikacijski protokoli koji koriste brojeve ne obrađuju ih i ne transformiraju ih.

Brojevi iz plana numeriranja koriste se izravno, njihovu upotrebljivost ne mogu povećati nikakvi postupci, metode ili algoritmi. Uostalom, planovi numeriranja, za razliku od mrežnih protokola, prvenstveno su orijentirani na ljude. Kako su ljudi također sastavni dio i IoT usluga, dio IoT adresnog sustava mora se i prilagoditi ljudima. S druge strane, veliki dio uređaja u internetskoj eri također koristi brojeve iz plana numeriranja za njihovo međusobno povezivanje, dijelom kao dio komunikacijskog nasljeđa, a dijelom jer je ljudima to pristupačniji način adresiranja. Stoga su oba načina adresiranja, izravni i neizravni, i dalje veoma važni i za M2M i za IoT usluge.

IV. FREKVENCIJSKI SPEKTAR I PRIJENOS INFORMACIJA

IoT i M2M u velikoj se mjeri oslanjaju na transportnu osnovicu mobilne mreže i svojim razvojem povećavaju pritisak na transportne mogućnosti mobilne mreže. Osobito tome doprinosi IoT, kao vodeća 4IR tehnologija. Razvoj i primjena IoT aplikacija, očekivana ekspanzija količine aplikacija te njihova kompleksnost i zahtjevi na kvalitetu s jedne strane, te ekspanzija broja korisnika i složenost korisničke strukture, zahtijeva povećanu propusnost i veći kapacitet. IoT aplikacije, osobito, zbog velike relativne količine povezanih uređaja, sa visokim zahtjevima na propusnost i kvalitetu usluge, postavljaju i velike zahtjeve za spektar i rade pritisak na uporabu RF spektra. Međutim, RF spektar kao dio elektromagnetskog spektra koji omogućava bežični prijenos i prijem poruka na većim ili manjim udaljenostima ograničen je svojim svojstvima, odnosno kao takav je prirodno ograničeno dobro. Stoga je i narastao naglo strah da spektra jednostavno neće biti dovoljno.

Iako se IoT često povezuje s 5G, 5G nije preduvjet za IoT, posebno ne za M2M. U okviru četvrte generacije mobilne mreže, LTE, već postoje rješenja koja podržavaju neke usluge IoT, kao Cat-NB1 i Cat-M1, koji su prvenstveno bili namijenjeni za M2M. Međutim, uvođenje

takvih rješenja ne ide brzo i jednostavno. Nadalje, rješenja za IoT i M2M unutar LTE-a se stalno traže i razvijaju.

Izazovi koje IoT dugoročno postavlja na mobilne mreže kreću se u pravcu povećanja propusnosti (brzine prijenosa podataka), kapaciteta (količine prenesenih podatka) i povećanja kvalitete komunikacije (gubici, kašnjenje). Od mobilnih se mreža očekuje da povećaju kapacitet prijenosa podataka za faktor 1000, brzine za faktor 500, a smanjenje kašnjenja za faktor 50. Sve to čini ogromne zahtjeve na dizajn 5G sustava. Kada se sve te brojke stave u kontekst spektra i mogućnosti spektra jasno je da korištenje spektra postaje jedan od vodećih izazova za IoT. Tome svakako doprinosi i dosadašnji, statični i dedicirani, način korištenja spektra koji se ne može jednostavno dokinuti. Kako povećanje spektra nije moguće, jedan je logičan korak ulaganje u nova tehnološka rješenja koja učinkovitije koriste spektar. Međutim, takav je razvoj relativno skup, osobito s obzirom na ograničenost mogućih rješenja te i dalje prisutnu nedostatnost spektra za nove (osobito IoT) usluge. Stoga i jeste prvi način povećanja mogućnosti spektra kroz regulatorne korake, kombinirano s tehnološkim inovacijama, osobito onima koje oslobađaju dijelove spektra za neke druge namjene. Zapravo, tu je prisutna jedna povratna sprega tehnologije i regulative.

Međutim, to nije dovoljno, svakako ne dugoročno. Stoga, dalji koraci idu u dva smjera:

- povećanje iskoristivog dijela spektra i primjena specifičnih osobina novih dijelova spektra,
- Skalabilno korištenje i dinamičko upravljanje spektrom.

Svaki pristup traži primjenu novih tehnoloških rješenja.

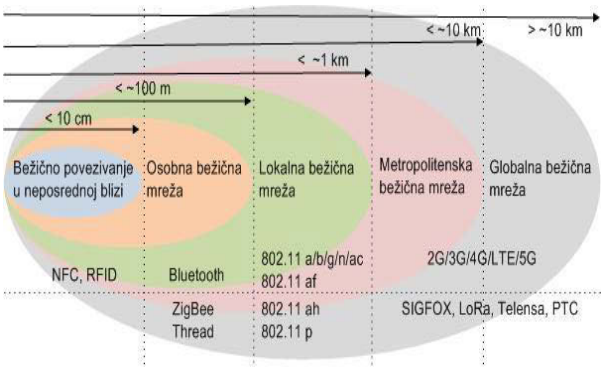
Očekuje se da će 5G povećati korištenje spektra korištenjem milimetarskih valova, omogućujući barem djelomično korištenje i spektra viših frekvencija, pa i iznad 30 GHz. Licencirani i nelicencirani spektar na tim frekvencijama dostupan je u mnogim dijelovima svijeta. Te su frekvencije često pogodne za male stanice jer zahtijevaju manje antene, a specifična orijentacija signala može povećati spektralnu učinkovitost.

S druge strane, ovi valovi su još manje primjenjivi kada je potrebno pokrivanje većih površina i lošije im je prodiranje u objekte. Međutim, to opet može biti prednost u pogledu ponovne upotrebe frekvencija. Raspoloživi spektar na određenom području pokrivanja dijeli se u određeni broj dijelova radi formiranja ćelijske strukture. Svaka ćelija koristi drugačiji frekvencijski raspon od susjednih kako bi se izbjegle interferencije. Isti frekvencijski raspon može se ponoviti tipično u svakoj trećoj ćeliji. Međutim, s obzirom da frekvencije višeg spektra teže prodiru kroz barijere to ponavljanje bi moglo biti i brže.

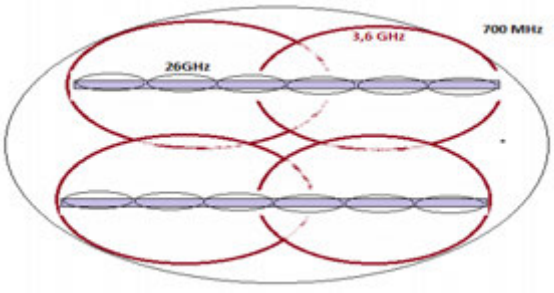
Spektar ispod 6 GHz podržava cijeli raspon uslužnih zahtjeva, izvrsnu i sveprisutnu pokrivenost, rad s niskom potrošnjom energije, osiguravajući odgovarajuću propusnost, sigurnu i jeftinu komunikaciju i zajamčenu isporuku, pa je to trenutno od velikog interesa. Međutim, i nižim dijelovima frekvencijskog spektra, a posebno ispod 6 GHz, postoji vrlo malo mogućnosti korištenja spektra. Neki od frekvencijskih raspona trebali bi postati dostupni već u vrlo bliskoj budućnosti, poput 700Mhz. Također, može se očekivati da će razvoj tehnologija digitalnog prijenosa signala osloboditi još neke dijelove spektra, ali to je neizvjesna budućnost. Ipak, potreban je veći spektar u pogledu ciljeva širine opsega i kapaciteta koji opravdavaju razvoj i primjenu 5G - sadašnja spektralna raspodjela sigurno neće biti dovoljna čak i sa sposobnošću agregiranja manjih blokova spektra. Iako nacionalni i nadnacionalni regulatori širom svijeta rade na rješavanju problema nedostatka spektra za 5G, procjenjuje se da su trenutne sposobnosti spektra dovoljne za neki naredni period.

Skalabilno korištenje spektra je pristup korištenju spektra s obzirom na strukturu i količinu usluga koje se pružaju, strukturu i količinu korisnika usluga, te uvjete pružanja usluga. Skalabilnost se postiže unutar 5G tehnologije ili kombiniranjem više bežičnih tehnologija [7]. IoT je orijentiran na korisnike i usluge i kao takav je neovisan o tehnologiji pružanja usluge. Stoga je korištenje heterogenog skupa bežičnih tehnologija optimizirajući način pružanja IoT usluga s obzirom na mogućnosti spektra (slika 5). Ovakav pristup kombinira, s ciljem optimizacije, strukturu i količinu usluga i korisnika te lokacijske i spektralne uvjete. Sa stanovišta spektra, a osobito s aspekta povećanja mogućnosti dijela spektra kojeg baštine povijesne tehnologije te razvoja novih tehnologija bežične komunikacije. Značajno za te nove bežične tehnologije je da su potaknute upravo IoT uslugama i njihovom konceptualnom neovisnošću o tehnologiji. Primjeri nekih takvih novih tehnologija potaknutih upravo IoT aplikacijama su ZigBee, Thread, 802.11ah, 802.11p, LTE-MTC, SIGFOX, LoRa, Telensa, PTC.

Unutar 5G tehnologije skalabilnost korištenja spektra orijentirana je kombiniranju s ciljem optimiziranja korištenja pojedinih dijelova frekvencijskog spektra, odnosno grupa frekvencija regulatorno dodijeljenim 5G mrežama. 5G tehnologija trenutno je orijentirana prema trima osnovnim grupama frekvencija: 700 MHz, 3,6 GHz i 26 GHz. Pri tome je potrebno razlučiti da prve dvije grupe zapravo pripadaju „tradicionalnom“ dijelu frekvencijskog spektra, dok je treća novi dio spektra koji bi se tek mogao koristiti. Bazna stanica u pojasu 700 MHz može pokriti veliki teritorij (ima veliki prostorni domet), uz vrlo dobro prodiranje u objekte i teško dostupna



Slika 5. Skalabilnost frekvencijskog spektra kroz tehnologije



Slika 6. Skalabilnost 5G frekvencijskog spektra

mjesta, ali s obzirom na relativno mali broj blokova od 5MHz raspoloživi spektar nije velik. Također su i raspoložive brzine manje. Pogodno je za ruralna područja, more i priobalje, otvorene ceste i slično. Sa 3,6 GHz pokrivamo manje površine, ali uz slabije prodiranje u objekte i na teško dostupna mjesta. Također su raspoloživi blokovi od 5MHz, ali je raspoloživi spektar značajno veći. Pogodno je za urbana područja. 26 GHz ima blokove po 200MHz (uz mogućnost i manjih blokova). Glavna obilježja su veliki kapacitet, mala latencija i kratki domet. Uz obilježja kvalitete same komunikacije koji taj dio spektra donosi tu je i jednostavnost izgradnje same mreže postavljanjem baznih stanica na uličnu infrastrukturu, bočne zidove zgrada i slično. Potrebe za velikim brzinama i malom latencijom za veliki broj korisnika pokrivamo upravo gustim rasporedom 26GHz baznih stanica, koje mogu imati vrlo male dimenzije i mali domet, ali zato se ponavljanjem istih frekvencija na manjim udaljenostima raspoloživi spektar bolje koristi. Nedostatak je slabo prodiranje u građevine i teško dostupna mjesta.

Koristeći ove osobine spektralnih grupa skalabilnost i optimiziranje korištenja spektra postiže se shemom pokrivanja određenog prostora mrežom raznovrsnih baznih postaja (slika 6).

Dinamičko upravljanje spektrom je još jedan način efikasnijeg korištenja spektra kao prirodno ograničenog dobra. To uključuje dijeljenje istog raspona spektra na različite vrste usluga, tehnološki, vremenski ili zemljopisno definirano, koristeći spektar različite namjene za realizaciju iste usluge, distribuciju geografskog spektra i dijeljenje spektra od strane operatora. Ovaj pristup je proširenje skalabilnog korištenja spektra, a osim tehnološkog izazova nosi sobom i druge izazove. Prvo je tu regulatorni izazov, zatim izazov suradnje te izazov upravljanja u realnom vremenu. Upravljanje spektrom u ovom slučaju teži algoritamskom. Spektar se koristi sekundarno, baš kao i brojevi u Internet adresnim protokolima.

V. ZAKLJUČAK

Prirodno ograničeno opće dobro svakako jeste jedan od komunikacijski orijentiranih izazova 4IR u tehnološkom, pa i ekonomskom, razvoju suvremenog komunikacijski orijentiranog društva. Ostavljajući po strani razmatranja represivnih regulatornih mjera temeljenih na administrativnim ograničenjima, kaznama za neučinkovito korištenje resursa i visoke cijene, pametno upravljanje prirodno ograničenim resursima može podržati razvoj i osigurati postojanje širokog spektra M2M i IoT aplikacija te razvoj 4IR općenito. Upravljanje uključuje planiranje, razvoj i provedbu načina njihove upotrebe.

Dosadašnja iskustva pokazuju da se postojeći planovi numeriranja na temelju E.164 mogu dovoljno prilagoditi novim potrebama i u kombinaciji s povećanjem procesorske moći te distribuiranih usluga spremanja i obrade podataka (u oblaku) mogu udovoljiti potrebama M2M usluga i velikim dijelom IoT uslugama, barem u razdoblju do potpunog prelaska na algoritamske načine adresiranja, poput internetskih protokola, koji brojevi koriste sekundarno.

U slučaju RF spektra, izazovi su znatno veći. Aktualni razvoj rješenja upravljanja frekvencijskim spektrom je slojevit i kompleksan, ali je i obećavajući, prvenstveno jer pomiče granice načina korištenja spektra. Međutim, učinkovita i održiva rješenja se tek očekuju i ona će uključivati i bitno drugačije razmišljanje svih dionika komunikacije. . Osiguranje dovoljno spektra za 5G ovdje je ključno pitanje. Statičke, regulatorno-administrativne metode ne mogu više samostalno davati rezultate. Ipak, radi se o resursima koji se ipak mogu još proširiti, ali i to je tek dio mogućeg rješenja. Dijelom se problem oskudnog spektra može riješiti i preraspodjelom korištenja samog spektra. Međutim, pravo rješenje je u posve novom načinu korištenja spektra, koji izlazi iz dosadašnjih okvira upravljanja spektrom kao takvim. Umjesto upravljanja spektrom radi se o upravljanju primjene spektra. To su metode upravljanja skalabilnošću spektra i dinamičko upravljanje implementacijom spektra. Korištenje skalabilnosti već nalazi svoju primjenu. Međutim, pravi izazov i pravo rješenje će biti dinamičko upravljanje spektrom. To pak znači visoku kompleksnost i usklađivanje više heterogenih područja poput tehnologije prijenosa, opreme, regulative, poslovne politike, itd, a osobito programsku implementaciju algoritama upravljanja. To je izazov sam po sebi.

LITERATURA

[1] https://ovum.informa.com/resources/product-content/lte-to-dominate-m2m-market-after-2021; svibanj 2020.

[2] https://www.statista.com/statistics/295754/m2m-connections-as-a-percentage-of-total-connections-leading-countries/; svibanj 2020.

[3] Santosh Kulkarni, Sanjeev Kulkarni: Communication Models in Internet of Things: A Survey, IJSTE - International Journal of Science Technology & Engineering, Volume 3, Issue 11, May 2017.

[4] https://www.statista.com/statistics/1101442/iot-number-of-connected-devices-worldwide/; svibanj 2020

[5] https://www.businesstoday.in/sectors/telecom/new-13-digit-numbering-sim-based-m2m-devices-only-not-regular-mobile-phones/story/271290.html; svibanj 2020.

[6] Georgios P. Koudouridis, Pablo Soldati: Trading off Network Density with Frequency, MDPI, Basel, prosinac 2018.

[7] https://www.itu.int/en/ITU-D/Regional-Presence/AsiaPacific/Pages/Events/2018/CoESmartcityoct2018/IOT_SmartCity.aspx: svibanj 2020.

SMART CITIES & CHALLENGES

Establishing the utility charges spatial database using digital twin technology

Viktor Mihoković * Luka Zalović* i Verica Zalović*,
* Geo-centar d.o.o./Čakovec, Hrvatska
viktor@geocentar.com

Abstract - **Digital twin technology is getting more attention when it comes to smart city management. Digital twins or 3D models of cities are created by utilizing 3D laser scanning, unmanned airborne vehicles (UAV) and mobile mapping systems. Photogrammetry based UAVs are increasingly used for land surveying and infrastructure mapping. In comparison to other techniques, photogrammetry can be used to collect large amount of geospatial data in relatively short time. Current techniques used to maintain and manage utility charges database often relies on long lasting field measurements and manual data collection. Such approach may result in incorrect, unreliable and outdated information about building area and floor count, which significantly affect the utility charge amount. This paper presents a methodology of establishing geospatial database by utilizing digital twin technology, photogrammetry and a fixed wing UAV with precise GNSS (Global Navigation Satellite System) also called GPS (Global Positioning System) receiver. Mentioned UAV has the ability of taking oblique images of terrain, buildings and infrastructure, which are then used to reconstruct 3D model (digital twin) of the city. It is presented how data extracted from 3D model is used to establish digital geospatial database containing building layouts, floor count and area. These information were compared to official data, which showed all the benefits of digital twin technology, but also all the issues of current utility charge system. Finally, the focus is turned to other applications of digital twin technology and its involvement in Industry 4.0, together with autonomous vehicles, 5G networks and IoT.**

Key words – digital twin technology, utility charges, UAV, 3D city model, GIS, geospatial database

I. INTRODUCTION

Nowadays, with the rapid technological development, local governments are starting to rely on modern technologies in order to run the city infrastructure more efficiently. That technology allows not only to lower costs through optimization of managing different processes, but it can also expand the functionalities of current systems. One example of that technology is digital twin, which allows creation of up to scale 3D virtual clone of buildings and infrastructure that can also be upgraded by adding attributes to each individual object. By using detailed replica of the city, it is possible to digitize and optimize different processes like traffic or spatial planning, and emergency services like fire department or police. Also, upgrading the digital twin model with data from different sensors and IoT devices would result in creating a smart clone of the city. That clone can not only monitor current state of the city, but if we would feed historical and current data into a machine learning software, it would be possible to model the behavior of the city and simulate what would happen in different emergency situations. Furthermore, the twin could be used for 5G network infrastructure planning. Since the 5G technology performance is limited by obstructions, and it requires line of sight between the antenna and the device for optimal performance, 3D data from digital twin would make it easy to create visibility maps and plan the best positions for antennas. In addition, by utilizing advanced neural networks, it would be possible to optimize the city traffic and predict the effect of future infrastructural and transportation projects, and by doing so, reduce traffic jams and speed up the public transportation.

Although there are a lot of different use cases for this technology and its benefits are obvious, the implementation process can be challenging. That is the reason why in this paper we decided to use the digital twin for a simple, yet effective method of creating spatial database for utility charges. It is a great example because such registers depend on spatial information like building areas and location. Moreover, the reliability of such data is often questionable since there are cases where current system is not up to date. Unfortunately, Croatian local governments quite frequently use nothing more than a simple Excel based table system, from which it is not possible to determine where the buildings are placed, what is their shape and size, what is in their neighborhood and has anything changed with the building (for example has it been expanded or demolished). It is not possible to know how the building area was defined or how many floors does it have from a simple table. What is more, there are some cases where payers themselves are the ones who reported the building areas, so it is doubtful if reported areas can be trusted. Despite that, those registers are used as official area and position information for utility charges calculation. All of this shows there is a need for upgrade, and digital twin technology may be the way to go. All the necessary data for the spatial database can be extracted precisely because it is high resolution, up to scale and visually representative model of reality.

A. Utility charges

According to the law of utility management, utility charges is a financial public charge used for maintenance of utility infrastructure [1]. Utility charges resources are meant to be used for: maintaining of public surfaces, maintenance of unclassified roads, public lighting, maintenance of cleanliness, drainage, maintenance of cemeteries and crematoriums, and the rest utility activities of the local significance [2]. Utility charge is a monthly fee that the real estate owner or user is required to pay. Monthly fee depends on three parameters [1]:

1. Zone coefficient (Kz)
2. Usage coefficient (Kn)
3. Utility charge point value (B)

Zone coefficient (Kz) has a different value based on a location (zone) of the building. Different zones are determined by the level of development of utility infrastructure [1]. For the first zone the coefficient value has the highest value of 1.00. Usage coefficient (Kn) depends on the type of real estate and what it is used for. For the residential and garage area the value is 1.00. Utility charge point value (B) is determined in Croatian kunas (HRK) by the square meter (m²) of net usable residental area in the first zone [1]. Utility charge point value can vary, and its amount is set by the local government. The value of utility charge is defined with the formula (1)

$$K(kn) = Kz \times Kn \times B \quad (1)$$

The most important parameter in Utility charge point value is net usable area of residential area. Net usable area defined by law as gross residential area reduced by the thickness of the walls [3].

Also, the net usable area does not include basement, attic, staircase and balcony areas which are not equipped enough to be used as residential or office spaces. Those areas and elements are multiplied by defined coefficient before adding to the net usable area. When everything is taken into consideration, net usable area will always be smaller than gross area. Lisjak [4] estimates this difference to be around 15%. One of the problem is that there is no regulation how the area is measured and calculated, or what register can be used as an official source for net usable building area. It is quite often the case to use the area reported in building permit, land registry or cadaster. In some cases, local government relies on the net usable areas and floor count reported by the payers themselves. The last system is used mostly in areas where the record is old, outdated or even nonexistent. Besides that, some local governments keep their record in a form of simple table. It makes it difficult to simply and unambiguously determine which building is the record referring to, where the real estate is located and how the building area is defined. For previously mentioned reasons those utility charge records are quite often based on inaccurate data. In these situations, benefits of the new technologies for mass spatial data collection like photogrammetry come to play.

II. Data acquisition and digital twin reconstruction

Photogrammetry is a science and technology of acquiring reliable data of Earth, its surroundings and other physical objects by using imagery, without direct contact with objects [5]. Although it is a very useful technique, its usage was not widely spread because the process required a specialized, expensive equipment and a lot of manual labor. Development of modern photogrammetric software and high-performance CPU-s and GPU-s vastly contributed photogrammetry as a technique to gain popularity, and is

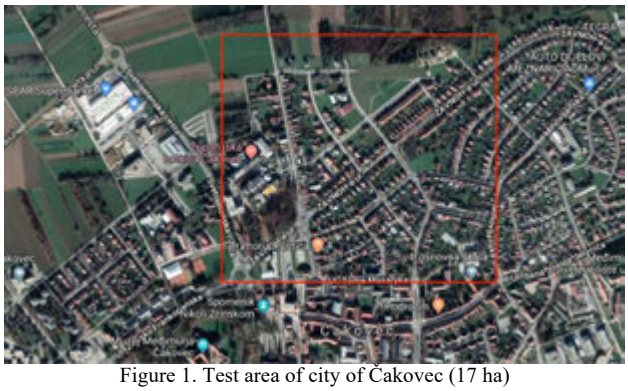


Figure 1. Test area of city of Čakovec (17 ha)

now one of few key methods for mass spatial data collection. [6][7]. One of the biggest advantages is the ability to reconstruct 3D models (digital twins) of any object using nothing more than digital photos. This ability makes it really useful in different industries like construction, agriculture, precise engineering, movie and gaming industries etc.

The development of unmanned aerial vehicles caused a revolution in acquiring spatial data for large areas. UAVs are designed to perform flights without any pilot on board, therefore they are controlled with remote control, by preprogrammed flight plan or autonomously by autopilot [8]. By mounting imaging sensor like photographic camera, GNSS receiver and inertial measurement unit (IMU) on UAV, it can be converted into photogrammetric mobile mapping unit that can acquire large amount of data necessary for reconstruction of spatial information in short time. It is systems like this that enable utilization of full potentials that modern photogrammetric algorithms provide.

In order to create digital twin model of the city and establish spatial database for calculation of utility charge amount, photogrammetric mobile mapping unit was used. As a test polygon we chose north part of the city of Čakovec, Croatia, with area of 17 ha (Figure 1). On our request, Local utility management department provided us official area information used for certain amount of residential buildings. That areas should represent net usable areas of the buildings, so we are expecting a mismatch between them and gross buildings area (which will be derived from digital twin of the city). Regardless, the mismatch should be quite small. If not, we can assume there is an error in one of compared areas.

A. Equipment used

The photogrammetric measurements were done on 10th of July 2019. Data acquisition was made with Topcon (Intel) Sirius Pro UAV (Figure 2). It is a fixed wing type of UAV with quite small weight relative to its size (2.7 kg including camera) and flight autonomy of 50 min on a single battery [9]. What distinguishes this UAV from all the other drones is integrated GNSS RTK (*Real Time Kinematics*) module, which enables centimeter grade positioning accuracy of every taken photo in the moment of exposition. In other words, Sirius Pro UAV allows us to collect the data needed for rendering 3D model with high positional accuracy without using ground control points (GCP). One of the most important aspects of any UAV used



Figure 2. Topcon Sirius Pro UAV



Figure 3. GeoSLAM mobile scanner

for accurate 3D reconstruction is the camera it can carry and use. For purposes of this paper Sony Alpha 6300 was used. It is equipped with 24 MP sensor, and it supports usage of so called "City mapping" module. The "City mapping" module allows capture of images in such way that not only the rooftops are visible, but also walls of the buildings. The module is a key component of the system since images taken with regular cameras can only reconstruct the rooftops and the ground.

Data for specified area was acquired in one 20 min automated flight. In order to create flight plan necessary for automated flight, we used MaVinci Desktop software. Flight altitude was set to 60m and the GSD (Ground sampling distance) which defines the final resolution of the digital twin was set at 3cm. Front and side overlaps were set at 80% to ensure good coverage and high reliability of photogrammetric reconstruction, which also ensured high accuracy of 3D city model.

B. Digital twin rendering

After successful flight, data post processing and 3D reconstruction was made. First step was the process of aerotriangulation, where images are aligned and exact positions and orientations of the images sensor in the moment of expositions are determined. It is followed by 3D reconstruction process, where software automatically renders 3D mesh and then transfers the texture/color from corresponding images onto the mesh (Figure 4). With all that, the creation process of digital twin model is done. For the purposes of this paper Topcon ContexCapture photogrammetric software was used.

C. Digital twin accuracy

In order to analyze positional accuracy of the created model, eight points were marked on the different parts of the test area. Position of the points was determined using high precision GNSS (GPS) receiver and RTK method of positioning. Measured coordinates were compared with the ones determined from the 3D model. Comparison was made separately along each coordinate axis. In addition, minimal error, maximal error, range, mean error and RMS (*Root Mean Square*) were calculated (table 1). The data shows high accuracy and consistency of position determined with RTK module. RMS values of only 2 cm and error range bellow 7 cm along each axis indicate that high accuracy was achieved.



Figure 4. Digital twin of the city

TABLE 1. DIGITAL TWIN POSITION ACCURACY

	Easting	**Northing**	**Height**
Min [m]	-0,030	-0,022	-0,031
Max [m]	0,033	0,047	0,013
Range [m]	0,063	0,069	0,043
Mean error [m]	0,013	0,015	0,011
RMS [m]	0,019	0,020	0,019

In addition, it was decided to measure part of the test area by using mobile 3D laser scanner GeoSLAM ZEB-Horizon (Figure 3). The scanner technology is based on SLAM (Simultaneous Localization and Mapping) algorithm. Since the accuracy of GeoSLAM system is proved to be within 1 to 3 cm [8][10] and can be used to cover large areas in short time span, the system is ideal for photogrammetric data and building gross area quality control. In both datasets the gross area is derived from outer building boundaries at the height of 1.4 m from the ground. By doing so, consistency was assured for the comparison between two datasets. The results are presented in table 2. The data shows minimal mismatch that can be attributed to human error in determining break points used in area measurements. It is worth mentioning that it wasn't possible to enter the backyards in order to scan the back of the buildings. When everything is taken into consideration, it is possible to conclude that the accuracy of the areas measured on digital twin model is adequate for determining the utility charge amount.

III. CREATION OF SPATIAL DATABASE

Since the created digital twin model is a representative and up to scale recreation of the real-life city, it is fairly easy to extract geometrical or spatial information like position, shape, area, volume etc. for any object. In order to make a database with consisted data, all of the building layouts were based on a section of the digital twin done at exactly 1.4 m above the ground. The height of 1.4 m was determined empirically by trying to eliminate all objects that are not buildings, like cars, fences and vegetation. In other words, the digital twin model was sliced at constant distance from the ground by taking in consideration the irregularities of the ground surface (Figure 5a). On created cross section, edges of the buildings were well defined, so to add buildings in the spatial database it was just the matter of vectorizing the boundaries in CAD software. (Figure 5b).

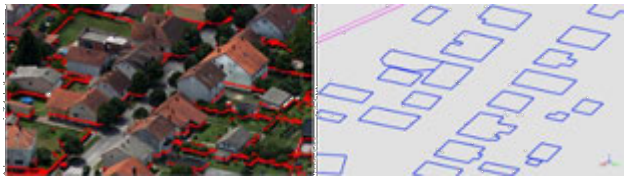


Figure 5. (a) Cross section at the constant height from the ground; (b) Vectorized building boundaries

TABLE 2. COMPARISON OF PHOTOGRAMMETRY AND LASER SCANNER DERIVED AREAS

Building ID	Area (digital twin) [m^2]	Area (Laser scanner) [m^2]	Difference [m^2]
1	62	488	692
2	211	275	30
3	207	273	32
4	101	443	340
5	-	83	-
6	-	186	-
7	134	312	132
8	129	386	200
9	62	332	435
10	194	351	81
11	146	346	138
12	86	210	145
13	136	301	121
14	158	286	80
15	477	1095	130
16	144	265	84

TABLE 3. COMPARISON BETWEEN SOME OF OFFICIAL AND PHOTOGRAMMETY DERIVED AREAS

Object ID	Official area [m^2]	Digital twin area [m^2]	Percentage of deviation [%]
1	62	488	692
2	211	275	30
3	207	273	32
4	101	443	340
5	-	83	-
6	-	186	-
7	134	312	132
9	62	332	435
17	214	223	4
23	227	256	13
24	158	256	62
25	205	422	106
26	216	238	10
27	195	224	15
35	483	510	6

Vectorized building layouts are then imported into GIS software (QGIS) as polygon shapefile layers, where each building became separate database object with its ID as a primary key. In that way we have created spatial database where each object is defined by its position, shape and attributes placed in the attribute table. Since the goal of this paper was to create a spatial database for billing utility charges, columns for floor count, building area, zone coefficient, usage coefficient, and calculated utility charge amount were added to the attribute table. Every building was connected with corresponding database object in original (official) utility charges database in order to cross reference the data and analyze the differences.

The floor count had to be input manually for every building, but thanks to the high-resolution digital twin, the process was pretty straightforward. Number of floors was counted on a twin by counting only those undoubtedly functional and in use. If it wasn't possible to determine whether building had a basement or functional attic, those floors were not taken into account. Usage coefficient was also manually input with the use of digital twin, where the operator would visually determine how the building was used (residential, office, garage or undeveloped area). The usage coefficient was also compared to the one registered in the official utility charges database. The Zone coefficient is defined by utility charge directive which defines the boundaries of the zones. Those boundaries were acquired in a form of a shapefile polygon, and the zones were simply transferred by visual inspection in GIS software. It the end, the building gross area was calculated using built in algorithm for polygonal area calculation. Final utility charge amount was calculated for each building using the official equation (1). Of course, it is always possible to expand the database by adding any additional attributes to the buildings, like house number, street name, municipality etc.

IV. RESULTS, DISCUSSION AND OTHER IMPLEMENTATIONS

Data extracted from digital twin were compared to official data provided by Local utility management department. Due to the fact that official data should represent net area values, while those calculated by using digital twin should represent gross area values, it was expected to see some differences between compared areas. Comparison is shown in table 3. It was decided to express the differences with percentages, with intention to show the proportion of deviations. By examining the results, it is obvious that there is a problem with current system of utility charges. Not only did we discover two objects which were not listed in official database, but the average percentage of deviation between official and digital twin data shows significant value of 122%. Potential reasons for such results could be found when inspecting the 3D city model. Many properties have garages which seem to be unregistered. Furthermore, many adaptations in form of terraces were found. On the other hand, areas calculated for newer buildings correspond to official data, which shows that such areas were calculated and recorded correctly. Such examples show all the benefits and efficiency of proposed methodology.

Nowadays, digital twin technology is much more than interesting news. It can be argued that it is highly efficient and advanced concept which could lead to some significant

time and cost reduction in many industries. One example is telecommunication industry, where this technology can be implemented to create 3D models of telecommunication towers. Such models could be used for inspection purposes, which would reduce time and risk, since it is still mainly done manually by climbing the tower. Moreover, digital twin could be used to extract important measurements like azimuth and tilt angles of antennas mounted on the top of the towers. Such approach provides detailed and more accurate results, while significantly reducing tower's down-time, even up to 5x [11]. Another example is inspection and monitoring of objects like bridges, dams or power plants. Digital twin technology could be used to detect potential threats and damaged parts. Thermal and RGB imagery provide added value to such inspection. Finally, it is possible to detect geometrical deformations, which could be critical due to the fact that such deformations could potentially cause the object to collapse.

V. Conclusion

Primary aim of this paper was to show how geospatial database can aid utility charges and make the complete system more efficient and functional. It was done by utilizing digital twin technology and photogrammetry based mobile mapping system. Proposed methodology turned out to deliver results of high accuracy and could easily reduce time and cost of creating and maintaining traditional utility charges database. Such approach can help to quickly determine area and floor count of every building, that can be used to identify objects with inaccurate data or no data at all. Although the sample size for this research was relatively small (35 buildings), the results seem to show some significant issues in current utility charges system. Such issues definitely have major financial impact on local government. Based on data derived from this research, we roughly estimated that, due to inefficient and outdated utility charges system, local governments experience losses counted in millions of Croatia kunas (HRK). This paper showed that these problems could potentially be solved by implementing digital twin technology.

An example we presented here is not the only way of utilizing 3D city models. This technology has wide application and we can expect to see its further expansion over next few years. It can be argued that many industries require high-resolution, up to scale and representative digital copy of real-life objects. Digital twin technology plays significant role in asset and infrastructure digitization with purpose to optimize resource management in smart cities. It can act as a geospatial foundation for autonomous vehicles since it can be used to derive 2D and 3D maps for navigation and positioning purposes. Furthermore, digital twins can be used for advanced and highly effective 5G networks planning, which are key aspect of IoT implementation. By connecting digital twin technology and IoT sensor data, it is possible to recreate the lifecycle of city in real time and, based on collected data, perform simulations that could optimize traffic and spatial planning. To sum up, it is undoubtedly true that digital twin technology will have major impact on many Industry 4.0 technologies. It will aid their successful implementation, which will consequently improve the overall life quality in cities.

References

[1] Law for Utility Charges (in Croatian), https://www.zakon.hr/z/319/Zakon-o-komunalnom-gospodarstvu

[2] Vodnjan, http://www.vodnjan.hr/hr/komunalna-naknada/sto-je-komunalna-naknada-i-tko-je-placa?

[3] Law on ownership and other real rights (in Croatian), https://www.zakon.hr/z/241/Zakon-o-vlasništvu-i-drugim-stvarnim-pravima

[4] J. Lisjak, V. Cetl, D. Markovinović, "Establishing spatial database for utility charges – integration of different needs for geodetic data and services (Cadaster of buildings, real estate tax)" (in Croarian), 9. Symposium of charted engineers of geodesy, Doing business in Croatia, Opatija, pp. 34-42, October 2016.

[5] D. Gajski, "Photogrammetry lessons", University of Zagreb, Faculty of Geodesy, Zagreb, October 2014.

[6] T. Medić, "Inaccessible terrain survey using UAV" (in Croatian), University of Zagreb, Faculty of Geodesy, Zagreb, September 2015.

[7] L. Zalović, "Testing of SLAM technology application for creating as-built documentation of a building" (in Croatian), Master's thesis, University of Zagreb, Faculty of Geodesy, Zagreb, June 2018.

[8] Policy on Unmanned areal vehicles (in Croatian), https://narodne-novine.nn.hr/clanci/sluzbeni/2018_11_104_2040.html

[9] Topcon positioning, https://www.topconpositioning.com/sites/default/filesproduct_files sirius_solutions_catalog_7010_2162_revb_sm_0.pdf

[10] E. Nocerino, F. Menna, F. Remondino, I. Toschi, P. Rodríguez-Gonzálvez, "Investigation of indoor and outdoor performance of two portable mobile mapping systems", SPIE Digital library, Trento, 2017.

[11] V. Zalović, L. Zalović, V. Mihoković, "Usage of UAV for surveying and inspection of telecommunication antenna towers" (in Croatian), 11. Symposium of charted engineers of geodesy, Geodesy and others, Opatija, pp. 46-52, October 2018.

Smart home IVR-based system with south Slavic language integration

Mustafa Spahic, Alma Secerbegovic, Vedad Mesic, Haris Hadzic, Amir Hasanbasic and Omar Jahic
Faculty of Electrical Engineering, University of Tuzla
Tuzla, Bosnia and Herzegovina
alma.secerbegovic@untz.ba

***Abstract* - The smart home concept is rapidly becoming a key component in the emerging Internet-of-Things (IoT) society. Home automation systems help customers by improving energy-efficiency, allowing for security monitoring and convenience with simplified control over smart IoT devices. However, it has been determined that the older population has difficulties interacting with complex technical devices. Possible solution to this problem would be involving Interactive Voice Response (IVR) machine, which would enable intelligent smart home control based on the information it gathers from voice-based commands. We explore the concept of a smart home with the integration of voice over IP (VOIP) and IVR technologies, along with smart IoT devices and cloud-based services. The presented smart home concept uses voice-assistance which allows for fluent and intuitive interaction. We have modified existing solutions applied for the English language and accommodated them to work for south Slavic languages. The design and implementation of the prototype for the simple IVR-based smart-home system are explained.**

Index terms -smart home, IVR, VOIP, IoT.

I. INTRODUCTION

Smart home systems and Internet-of-Things (IoT) platforms are getting more sophisticated, with future trends to anticipate the needs of the owners and occupants. The sustained growth of the smart home market, along with the development of IoT devices, is expected to continue in the following decade [1][2]. With the rising number of the elderly, the smart home concept is inevitably being adopted by the older population as well [3]. The quality of life of the elderly can be significantly improved with information and communication technologies while ensuring them to keep their independence at home instead of moving to a care facility.

Interactive voice assistants or interactive voice response (IVR) systems present the points of communication between the user and all connected devices in the home environment. Their integration involves a wired or wireless connection with smart appliances, as well as existing commercial cloud services for proper execution. Voice over Internet Protocol (VoIP) has been known as a useful technology for transmission of voice by using digital or analog telephones. With the existing open-source VoIP platform, the smart home concept can be easily integrated to operate with the voice-based IVR system. Previous studies have analyzed VoIP and IoT integration [4][5]. However, presented solutions have focused only on using VoIP as a service to transfer the data necessary for controlling the smart home appliances.

In this paper, we have presented simplified smart home system, which uses open-source technologies for transmission, recognition, and analysis of voice-based commands. The special contribution of this paper includes the integration of south Slavic languages in the smart home system, which has not been previously implemented, according to the author's knowledge. The remaining of the paper is organized as follows. Section II explains the VOIP-based smart home concept. In Section III, the proposed system is described, along with the implementation details for the integration of south Slavic language, as well as discussion about the benefits and disadvantages of the proposed concept. Conclusions are drawn in Section IV.

II. VOIP-BASED SMART HOME CONCEPT

Smart home, also known as home automation system, presents a residence equipped with smart electric appliances and Internet of Things (IoT) sensors, that allow for remote control, monitoring and access through a communication network. Special attention has been given to voice-controlled systems, where items and appliances around the house can be controlled with a simple voice command. Improvements such as energy savings when the house is unoccupied or the ability to control the devices inside the house from any location provide the individual with calmness and peace of mind. The elderly or people with disabilities can greatly benefit from the described smart home concept, which provides a comfortable lifestyle and simplification of ordinary tasks, as seen in Fig. 1.



Figure 1. Concept of smart home voice-controlled system

This work is supported by the project grant of University of Tuzla in collaboration with Federal Ministry of Education and Science in Bosnia and Herzegovina.

Home automation systems consist of sensors and electronic programmable controls, that are integrated with lighting, heating, cooling and entertainment devices with a wireless or wired communication network. The integration of IVR with the home automation system enables the design of interactive and convenient tools for monitoring and control of smart house devices, which is especially suitable for the elderly [6]. Most IVR technology scenarios support touch tones as an interactive way of communicating with a caller, e.g. choosing to turn on the light by pressing a certain number. On the other hand, voice recognition has become part of IVR implementation, where users can give a command or respond to a prompt by using spoken language. This proves to be a more intuitive alternative than simple touch-pad systems, where customers could only indicate how they wanted to proceed by listening for all menu options and then pressing the corresponding key on their phone. However, other less common languages are still not supported in commercially available smart home IoT platforms. For example, south Slavic languages are especially complex and their automatic speech recognition presents a difficult task [7]. Designing an interactive voice response assistant in south Slavic languages for the smart home automation system is presented, with the main focus on integrating voice-based commands in south Slavic languages as an input to the sensor-controlled home automation system.

The proliferation of always-on technology has created a culture both comfortable with and anxious for more automation. The average home is still not connected, and most commercial systems are built around a proprietary system making it difficult to connect products from various vendors. Another challenge of popular commercial systems is that they are completely relying on Internet connectivity for executing smart home control. By using VoIP-based open-source technology, users can communicate with their home devices by simply making an analog telephone call and giving voice commands. In this paper, we have attempted to solve the above-listed challenges. The design and implementation of the proposed smart home system are explained in the following sections.

III. SYSTEM IMPLEMENTATION

Simplified VoIP-based smart home is described below. Our private branch exchange (PBX) is created using the open-source software implementation Asterisk [8]. PBX was implemented in the Raspberry Pi platform [9]. The processing module and IVR are implemented on the Raspberry Pi platform, which also serves as an IoT gateway. This implies that it also can provide control over home devices and appliances. In order to implement real-time home device monitoring, we have employed Google Firebase [10]. This type of database is a not only SQL (NoSQL) database, known for providing the mechanism for storage and retrieval of data. The presented real-time database contains information about all the rooms and appliances available in those rooms. To link Asterisk to our Raspberry PI application, written in Python, Asterisk Representational State Transfer (REST) application programming interface (API) was used. This allows for the control of PBX enabling the record and playback options during a call to the smart home system. The implementation is called Interactive Voice Response (IVR). Using the Dialplan scripting language, we have implemented an extension that a user can call and interact with an interactive voice response assistant on south Slavic languages, called Yugo.

The presented application interacts with the user through prerecorded voice messages. Initially, it asks the user for an authorization phrase. This is a very important aspect of the application because home security is crucial. The method of operation is as follows; the user sets his security phrase in the client application, e.g. "The grass is green." When making a call to the smart home, the IVR requires the user to speak their security phrase. If the phrase is incorrect, the call is terminated. It is possible to detect attacks based on a large number of failed attempts and block the number. If the security phrase is correct, further interaction with the device is allowed. Afterward, the interaction between the IVR and the user is continued, where the IVR continuously asks the user what is the command to be executed in the smart home.

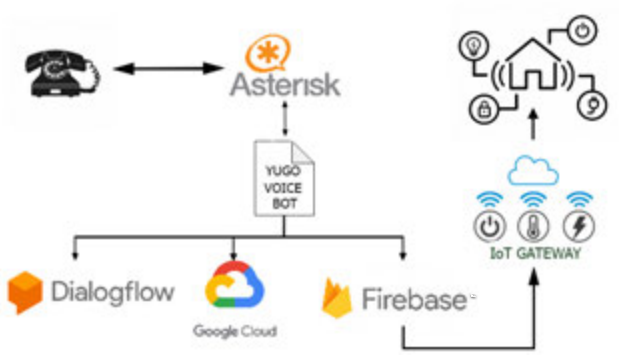


Figure 2. System implementation of VoIP-based smart home

When a user submits his request in the form of human speech, IVR records the message and forwards it to a free Google speech-to-text service. This service processes the request and returns to IVR a text transcript of the conversation in the Croatian language. The acquired text transcript is sent to Google Dialogflow. Dialogflow uses provided training data for building a machine learning model specifically designed for this purpose. Since this is a simplified solution, we have manually generated voice phrases for the machine learning model. E.g. if the user wants to turn on the light in the living room, the combination of words "living-room-light-on" is taken into account. In response to Dialogflow, our IVR receives a JavaScript Object Notation (JSON) object, containing the information about the user's intent and further performs the necessary actions to update the real-time database and change the condition of the device in the home. As a confirmation, the voice message is released informing about the completion of the task. The call will be terminated when the user does not require additional actions.

Table 1. shows the list of implemented actions and smart home locations. For example, it is not possible to turn on the stove in the living room, since these devices

are located in the kitchen. The proposed system carefully analyzes the user commands, meaning that lights are only turned on or off after the user declares in which room this action needs to be performed.

TABLE I. LIST OF AVAILABLE ACTIONS FOR SPECIFIC DEVICES IN SMART HOME LOCATIONS

Location	Devices				
	AC	Window	Light	Stove	Boiler
Hallway	x	Open / close	Turn on/off	x	x
Bedroom	Switch on/off	Open / close	Turn on/off	x	x
Bathroom	x	Open / close	Turn on/off	x	Switch on/off
Living room	Switch on/off	Open / close	Turn on/off	x	x
Kitchen	x	Open / close	Turn on/off	Switch on/off	x

With the presented VoIP-IVR system, adding another home appliance is as easy as adding a few training phrases and an intention keyword. The history of user interactions is logged, which grants the ability to make corrections and improve the system's accuracy. Sound files that are played to the user are 8 kHz PCM Wav files, which means they can be transmitted using a regular PSTN network.

The use of open-source PBX (Private Branch Exchange) enables us to develop and test this system in the local network in a relatively small amount of time. Further integration of this system enables us to connect an IVR to an external public switched telephone network PSTN) and then, later on, using an external Direct Inward Dialing (DID) code, to gain control of our house from anywhere around the world. To test this system, we created a model of a smart house and an IoT gateway by using an esp8266 WiFi System on a Chip (SoC), which connects to the IVR and provides the control of the house. To make it all operate together, Google's Firebase service is utilized. The diagram in Fig. 3 shows the flow chart of the entire IVR application. At every interaction between the user and IVR, Asterisk records the user's speech, and our IVR application passes it on to external services (Speech to Text, Dialogflow) that process the acquired speech, convert it to text, and then gather the user's intended intention. IVR takes all the necessary steps to complete this action by updating or reading the database, depending on the type of intent desired. The call ends when the user indicates to the IVR that there are no additional requests, or if the user terminates the call himself.

Fig. 4 shows a screenshot of the implemented database in Google Firebase service. A JSON object containing the key-value of a pair is used to view and place data in the database. This database can be used by other client applications in parallel with the present one via personal computer or phone, users can control their smart home.
Potential improvements of the proposed system include integration with an open-source gateway such as OpenHAB [11], which is a modular design architecture allowing the integration of more than 200 different technologies and systems and thousand of devices. This would permit the integration of different commercially available IoT devices. The synthesis of the OpenHAB framework and the proposed system would require incorporating the OpenHAB REST application interface into the IVR application. Another possible upgrade of the proposed system would be the deployment of the system to the cloud platform, which would eliminate the cost of IoT gateway.

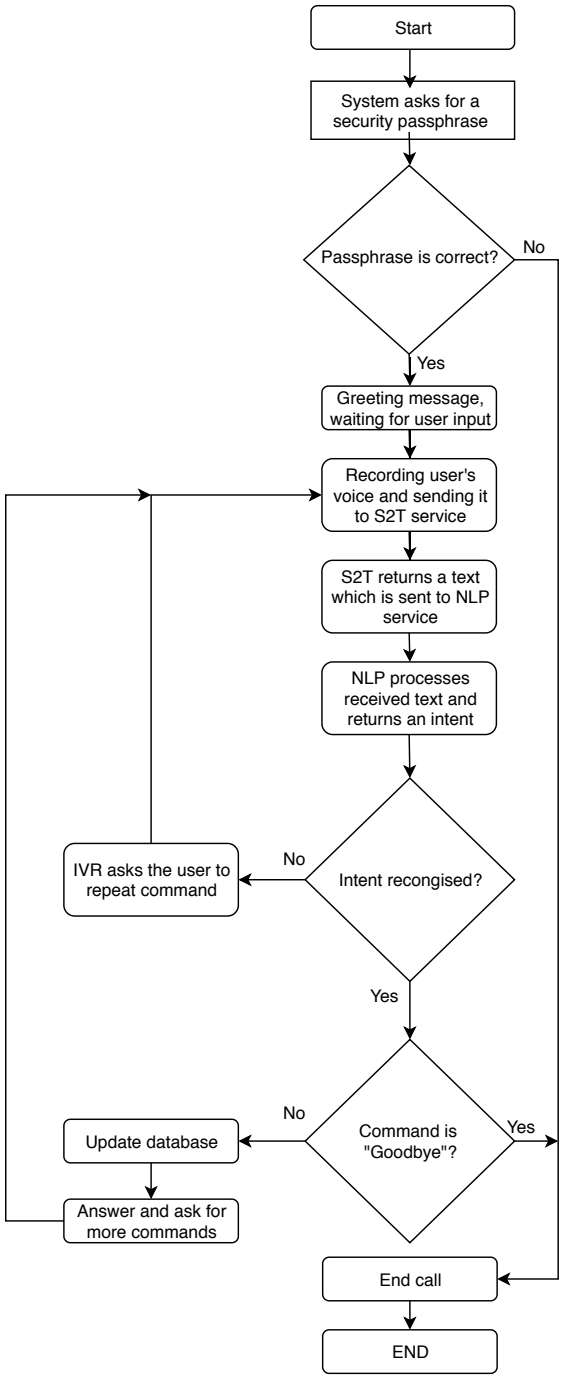


Figure 3. State diagram flow of presented IVR smart home system

The resulting product could be potentially offered as a service, leading to simple integration with existing smart home systems. A docker container could be used to

encapsulate the whole system, allowing for easy hosting on a cloud service. In order to demonstrate the operation of the system, we have created a prototype model of a smart home that uses the WiFi system on a chip (SoC) ESP8266 as a controller (Fig 5.).

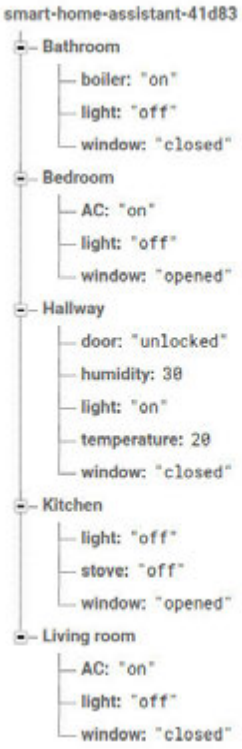


Figure 4. Screenshot from system's database in Google Firebase service



Figure 5. Presented smart home mock-up model with LED lights

It connects to the WiFi network and monitors the state of the real-time database while accordingly updating the state of the devices in the home. Through serial-parallel shift registers, illumination of light-emitting diodes (LED) signals the voice-generated commands. It also includes the controlling of temperature and humidity sensors. This system is overly simplified solution, that can benefit from available technologies such as text-to-speech and speech-to-text modules for south Slavic languages [12], where the manual creation of commands and natural language processing model can be omitted. In addition, security aspects of the proposed solutions could be further improved, with speaker recognition modules and caller ID options.

IV. Conclusion

The smart home concept has become widely recognized and adopted, with predictions of getting more attention in the future. While commercially available products are mostly based on widely used languages, integration of smart home automation with voice recognition assistant in south Slavic language is still missing. The proposed smart home VoIP-based system and the smart home IVR on south Slavic languages have been successfully developed and tested. Open-source platforms such as Google Cloud, Firebase and Dialogflow have been used for natural speech processing and recognition of control commands for five devices in five different locations. Light switches, temperature and humidity sensors have been integrated into the system to demonstrate its feasibility and effectiveness.

References

[1] B. L. R. Stojkoska and K. V. Trivodaliev, "A review of Internet of Things for smart home: Challenges and solutions," Journal of Cleaner Production, vol. 140, pp. 1454-1464, 2017.

[2] M. Kaneko, K. Arima, T. Murakami, M. Isshiki and H. Sugimura,"Design and implementation of interactive control system for smart houses," IEEE Int Conf on Consumer Electronics (ICCE), pp. 283-284, 2017.

[3] S. Pinto, J. Cabral, and T. Gomes, "We-care: An IoT-based health care system for elderly people." 2017 IEEE Int Conf on Industrial Technology (ICIT), pp. 1378-1383, 2017.

[4] F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon, "Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects," Personal and Ubiquitous Computing, vol. 17, no. 1, pp. 127-144, 2013.

[5] P.P. Gaikwad, J.P. Gabhane, and S.S. Golait, "A survey based on Smart Homes system using Internet-of-Things, " IEEE Int Conf on Computation of Power, Energy, Information and Communication , pp. 330-335), 2015.

[6] G. Aljabari, "Integrating VoIP Systems with The Internet of Things, " The 4th Palestinian International Conference on Computer and Information Technology (PICCIT), 2015.

[7] V. Delić, et al. "Speech technology progress based on new machine learning paradigm." Computational intelligence and neuroscience 2019 (2019).

[8] Asterisk, https://www.asterisk.org/

[9] Raspberry Pi, https://www.raspberrypi.org/

[10] Google Firebase, http://firebase.google.com/

[11] OpenHAB, https://www.openhab.org/

[12] T. Delić, M. Sečujski and S. Suzić, " A review of Serbian parametric speech synthesis based on deep neural networks", Telfor Journal, vol. 9, no. 1, pp. 32-37, 2017.

Software Prototype Based on Augmented Reality for Mastering Vocabulary

Mija Begic, Marko Cirimotic, Ivona Farkas, Ivan Skoric, Zeljka Car, Ivana Rasan, Matea Zilak
University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
{mija.begic; marko.cirimotic; ivona.farkas; ivan.skoric; zeljka.car; ivana.rasan; matea.zilak}@fer.hr

***Abstract* – Interactive experience of Augmented Reality that is created by deploying virtual objects in a real user environment, and its augmentation by audio-visual elements as well as the possibility of touch, provides added value for learning, especially regarding content multimodal presentation, increasing learner's motivation and level of engagement. This is very promising, especially in the field of education and rehabilitation, where current research slightly takes focus in this technological direction. To conduct this type of research, the prerequisite is to have a quality prototype with features, easy to understand and use by users of different cognitive skills, and yet sophisticated enough to carry out the research. The paper presents software prototype based on Augmented Reality aimed for mastering vocabulary for children with complex communication needs. The solution was developed in the multidisciplinary cooperation of students and teachers from the technical field and professionals from the education and rehabilitation field. The overall development process and initial evaluation are described, as well as the developed prototype from the technological and user perspective.**

Keywords – Augmented Reality; virtual objects; learning; education; rehabilitation; children; mastering vocabulary

I. INTRODUCTION

The use of Augmented Reality (AR) in education is still in early stages but shows great potential. The very purpose of AR is seamlessly blending digital information and entertainment with everyday surroundings, providing additional content to certain real-life problems without shutting out the real world. This makes it an excellent tool for learning, especially with today's younger generations who are being raised in a world dominated by smart devices. Everyday advances in digital technologies have had an impact not only on the way how children learn, but also on the way how children communicate with their family, friends, and peers at school. The advent of mobile touch-screen devices (e.g. tablets and smartphones) has also had an impact on the Augmentative and Alternative Communication (AAC) methods used to support the communication process of people with Complex Communication Needs (CCN) who also include children with motor, language, cognitive, and/or sensory perceptual impairments resulting in different developmental disabilities. Children with CCN may particularly benefit from using tablet devices with appropriate AAC applications because of its many advantages [1]. Despite the advances in AAC domain regarding technologies, researchers emphasize that future research and development is required around other technological innovations, including AR, that might improve AAC technology solutions for children with CCN [2]. Everything mentioned motivated members of ICT-AAC research group[1] which has been conducting intensive research in the AAC domain for over a decade and has developed over 30 AAC applications for most popular platforms, such as Android, iOS and Web[2], to conduct this type of research where the potential and possible benefits of AR-based solution for mastering vocabulary for children with CCN is explored.

To develop a functional prototype of AR-based software aimed for children with CCN to master the vocabulary, the multidisciplinary project was initiated. It involved students and teachers from the technical field and professionals from the education and rehabilitation field. Given the target user group, professionals from education and rehabilitation field were reached out to provide expertise and to evaluate the project's stages of development. The objective of the developed software prototype is to serve as a proof of concept, evolutionary prototype, and foundation for development of similar solutions or future upgrades, and for further verification of the applicability of such software by abovementioned professionals.

The paper presents developed software prototype from the technological and user perspective as well as the overall development process. Selection process for development environment is presented. Also, initial prototype evaluation of functionality, simplicity, design, usefulness, and technical quality of content is described, and the results are presented.

The rest of the paper is organized as follows: Section II looks into the principles of AR and its value in education, Section III gives a description of technology used for prototype development as well as development phases, Section IV describes the resulting prototype and its use, Section V brings the initial prototype user evaluation results, and, finally, Section VI concludes the paper.

II. AUGMENTED REALITY FOR LEARNING

Augmented Reality is an interactive experience of a real environment in which objects from the real world are complemented by computer generated information. In general, the additional digital information can be

[1] http://lab.ict-aac.hr/

[2] http://www.ict-aac.hr/index.php/en/applications

constructive, meaning it is simply added to the environment and is its continuation, or destructive, in which case it conceals and replaces our reality, be it just visually or through more senses. It is important to note the difference between augmented and virtual reality (AR and VR). Although they both lie on the so-called "reality-virtuality continuum", they do not present the same technological concept. While the technology of VR relies on destructive supplementation, with the purpose of blocking our real-world perception and placing us in a completely virtual world, AR makes use of constructive supplementation, working to add information about our surroundings and allow us to interact with the digital world without taking anything away from our perception of reality [3-6].

VR is commonly used in simulation software, video games and other entertainment products, and most often requires a VR headset with a display, speakers, and movement-tracking technology, possibly in combination with special remote controllers to be used. In terms of supported devices, AR is inarguably more versatile, with various implementations for smartphones and tablets, TV broadcasts, as well as AR-specific devices such as HUDs (head-up displays), HMDs (head- or helmet-mounted displays) and headsets (e.g., *Microsoft HoloLens*). There are many uses for AR today. Some have been evolving since the distant years of the Second World War, such as the aforementioned HUDs and HMDs that serve as digital aid for pilots and other vehicle crew members. Others are much younger but just as widespread, including special animations in TV broadcasts, control and guidance software in industrial plants, entertaining and informative smartphone applications, educational software, and even AR-powered computers such as the *Microsoft HoloLens* [4-5] [7-15].

Many interactive educational applications with AR elements can be found. The *Virtual Teacher* series' applications [3] includes AR applications for learning mathematics, physics, human anatomy, programming, and geography. *Octagon Studio* published the whole product line designed to make STEM learning simpler and fun with AR. With different sets of *4D+ Flashcards*[4], the users can learn about animals, space, dinosaurs, occupations, human anatomy, cars, history of aircrafts etc. from the facts and information in AR mode. Focusing on AR in educational software, several applications were studied more in-depth in the pre-development stage of the research to explore the established software requirements for similar products. *AR Anatomy 4D+*[5] is a free AR-based application used to learn about the human anatomy. It uses markers to display 3D models of organs or organ systems which can be interacted with. The products from the US company *Alive Studios*[6] that are designed specifically for use in classroom has also been examined. They utilize AR to educate students in several different fields, including reading and math. The individual applications also make use of markers to generate digital objects or trigger certain actions.

Overall, AR already has a large presence in mobile applications. Given the versatility of today's smart devices and the way AR can harness many of their existing elements, the mobile world on its own is a big playing field for this technology. Though the more complex interactions that exist famously in VR, such as hand tracking and hand gestures in front of a camera, are possible on mobile devices, there are still improvements to be made (though this will likely happen now that smartphones are being shipped with depth sensors and multiple-lensed cameras). That is why many applications use simpler touch commands or even voice recognition which are the interactions that most smartphone users are familiar with, and there is not a huge difference between typing a text message and clicking on models of distant stars that appear on your camera feed, as is the case in *SkyView® Lite*[7] and *Star Walk 2 Free - Identify Stars in the Night Sky* [8] applications. Just like hand tracking and gestures, visually processing a device's surroundings is still imperfect or simply too slow, so, to trigger certain actions or digital renderings, markers such as QR codes are still a popular way to track the "reality" part of "augmented reality" [4][16]. Nevertheless, environment scanning is used in a significant number of cases, such as the *IKEA Place* application[9].

III. Development of AR Prototype

The prototype development is based on an evolutionary model of the software development lifecycle, where the required prototype has been developed in several successive iterations. An evolutionary model involves continuous work with user representatives whose feedback on each product version can result in new features or functionalities for the next version. This development method is typical when working with new technologies. In this case, experts from education and rehabilitation field were the one who defined the requirements, i.e. features and design of an AR prototype. As mentioned before, the technical team involved students enrolled in the project and the main goal for this 3-month-project was to develop a functional AR solution prototype with features, easy to understand and use by users of different cognitive skills, and yet sophisticated enough to carry out the AR based research.

In the first phase of the project, that lasted for a week, the development team members and communication channels were defined. The second phase, with the duration of three weeks, mainly consisted of planning and choosing the right technology for the development. Because of its simplicity and accessibility, Unity game engine[10] (version 2018.4.14f) was chosen as a development platform for the prototype. Unity is used for compiling the project into an

[3] http://arkids.cards/en
[4] https://octagon.studio/products-and-services/4d-flashcards/
[5] https://play.google.com/store/apps/details?id=com.DanikTM.ARAnatomy
[6] https://alivestudiosco.com/
[7] https://play.google.com/store/apps/details?id=com.t11.skyviewfree
[8] https://play.google.com/store/apps/details?id=com.vitotechnology.StarWalk2Free
[9] https://play.google.com/store/apps/details?id=com.inter_ikea.place
[10] https://unity.com/

application for Android[11], which has been selected due to its leading market share and the project team's ability to test the application more reliably on a representative sample of Android devices. Unity offers a great range of AR software development kits (SDKs) to its users. The following SDKs were analyzed: ARCore[12], Maxst[13], Wikitude[14], ARKit[15] and Vuforia[16]. Maxst and Wikitude have both shown to be unacceptable, as it would take up to eight days to obtain a license for use. ARCore, ARKit and Vuforia were simple to set up, but there are some differences between them. ARCore is Google's, and ARKit is Apple's platform, which means that with these SDKs, applications can only be developed for Android or iOS devices, while Vuforia supports application development for both types of these devices. Every smartphone has a camera that supports the use of Vuforia and most importantly, Vuforia is marker-based. Markers are real-world images or 3D objects that Vuforia can detect and track and serve as a backdrop for AR (3D model, image, video, animation etc.). While scanning the space with a mobile device, Vuforia tries to detect markers by comparing the camera image with a database that contains all the markers used in the application. Once the marker is detected, Vuforia will monitor if it is at least partially in the camera's field of view. After reviewing and testing these SDKs, Vuforia was selected because it is easy to use and work with and it also worked very well with scanning markers which was necessary for our prototype. Smart Recorder (SmartMob)[17] application for recording sound and the default camera application on Android phones were used for creating audio and video content in the prototype (e.g. audio on the first level, video on the second level). The last technology that was used is Adobe Photoshop CC[18] for creating custom icons for the prototype (e.g. back button, info button etc.). All audio and video content, as well as the buttons in the prototype, were made by the student project team.

The third phase of the project, i.e. the implementation phase, lasted for ten weeks. Throughout the development there had been four versions of the prototype. The initial version of the prototype was an AR application whose main goal was to show the possible interactions in mobile AR, more specifically to show two types of content when scanning the marker. First option enabled the user to scan a predetermined marker after which a 3D object displayed over the same marker. Second option displayed a video (over the marker or in full screen mode) after the marker was scanned. This prototype version was shown to the experts from education and rehabilitation field, after which they filled out the prepared questionnaires and gave their feedback. This was the input for the next phase of the prototype development. Main functionalities that were defined based on experts' feedback are the following: the application will have three levels, each containing different AR content that is shown to the user upon scanning the marker. The next prototype version included implemented three levels for only one object defined for mastering vocabulary. Further direction of development was discussed through weekly meetings of students and teachers. The third version of the application, i.e. the alpha version, had implemented three levels of the application for all five objects the experts defined to be used for mastering vocabulary. The first level contains 3D models, the second contains videos and the third level contains a quiz after which the bonus level is unlocked if the user answers correctly to all questions in a quiz. For the user to access the content contained in each level, a marker needs to be scanned. The marker is placed somewhere in their environment, where the educational content is applicable. For example, the marker denoting educational content about the toothbrush should be placed in the bathroom, near the user's actual toothbrush. Using any level of the prototype, the device's camera view is opened, enabling the user to scan the marker designated for the desired object and view its educational content. The content used in the third version of the prototype was reviewed by the experts from education and rehabilitation field. The alpha version of the application was evaluated by students of the Faculty of Electrical Engineering and Computing. Firstly, they used the application and then they filled out a questionnaire in which they could grade each part of the application. Their feedback helped to upgrade the prototype and develop the fourth. i.e. the final version of the prototype, in which bugs were fixed and design improved.

The next step of the project will be sending the prototype application to experts who work with children with complex communication needs. This round of testing will be longer and more important than previous evaluations since the prototype will be in proper use with its target audience for a longer period. Using the first-hand feedback, the application can be further improved or adapted, and finally slated for public release.

IV. AR Prototype Description

When the user runs the application for the first time, the application requires the setup of a few options. The first option is to enter the user's name (i.e. a child's name who will be using the application). This step can be skipped, and the user's name can be set in the settings of the application. The second option is to adjust the theme of the application, i.e. to change the background and text color if better contrast level is needed for those with visual impairments (Figure 1). The last setup step is the option to enable or disable the in-application sound. After the setup, these options can always be changed in the settings menu (Figure 2). Besides the user's name, application theme and sound option, the user can access the description of the application via information button in the settings. Besides the settings button, the home screen of the application contains exit, information and play buttons (Figure 3). The information button on a home screen leads to all the markers that are used in the application, as well as the prizes won at the bonus level.

[11] https://www.android.com/
[12] https://developers.google.com/ar
[13] http://maxst.com/#/
[14] https://www.wikitude.com/
[15] https://developer.apple.com/augmented-reality/
[16] https://developer.vuforia.com/
[17] https://play.google.com/store/apps/details?id=com.andrwq.recorder&hl=hr
[18] https://www.adobe.com/sea/products/photoshop.html

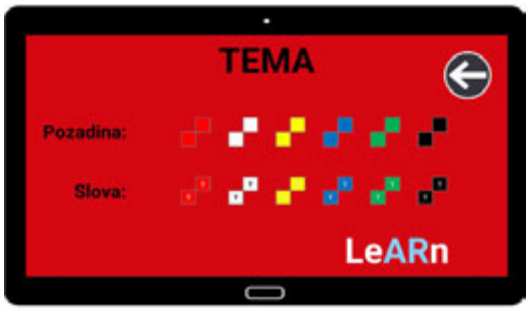


Figure 1 Setup of application theme



Figure 2 Settings screen of the application



Figure 3 Home screen of the application

The play button leads to a new screen where the user can choose one of the three different levels and read short information about each level (Figure 4). The fourth (bonus) level can be reached from the prizes button on this screen, too. On the first level, after scanning one of the five different object markers, a 3D model of the scanned object is shown on the marker in the real world (e.g. after scanning the toothbrush object marker, a 3D model of a toothbrush appears on the marker like it is showed in Figure 5). Also, an audio recording is played with a short description of the scanned object. After scanning one of the object markers on the second level, a new screen with a play button is shown which can be selected to play an informative video about the scanned object. In the video, the child can learn the purpose of the object and how to use it (Figure 6). The video can be paused and played again once it is over. By pressing a button for the next level, the third level screen is shown. The third level is a 3-question quiz about the scanned object. The quiz starts by scanning one of the object markers. After scanning the marker, a question appears on the right side. To answer a question, the user clicks on an eye button to reveal a camera window on the left after which the user scans the "yes" or "no" answer marker (Figure 7).



Figure 4 Application screen with three levels, prizes, and levels information buttons

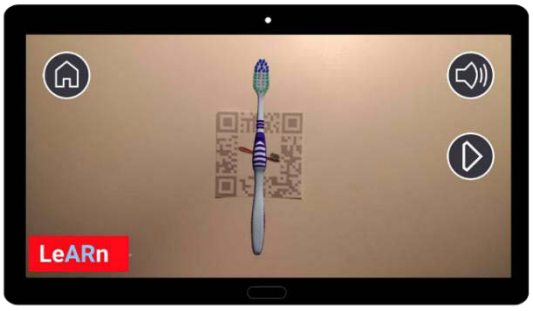


Figure 5 3D model on the object marker

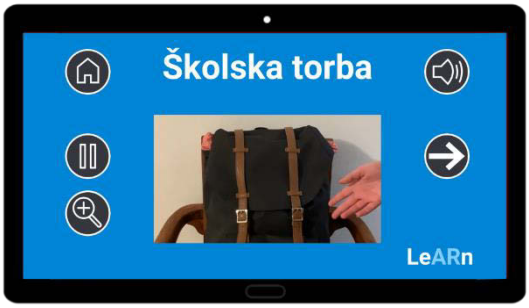


Figure 6 Informative video about the scanned object

If the question is answered correctly, the next question is shown. If the question is answered incorrectly, the same question is shown repeatedly until it is answered correctly. After answering correctly to all three questions about an object, a new 3D model is unlocked as a prize to be used in the bonus level. Once the prize for an object is unlocked, the child must scan the same object marker on the bonus level, and they will be able to see their prize object (Figure 8).



Figure 7 The quiz screen of the application with the exemplary question

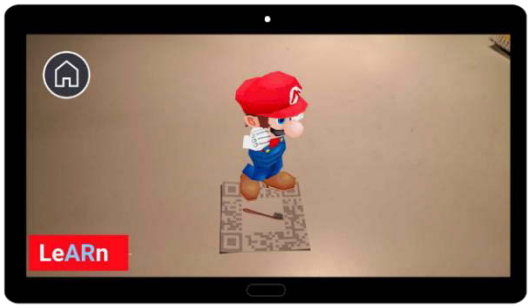


Figure 8 Prize model shown on the marker in the bonus level

V. AR PROTOTYPE USER EVALUATION

In this section, the questionnaire is described used for evaluation of functionality, simplicity, design, usefulness, and technical quality of content used in the prototype. The questionnaire consisted of seventeen questions and was fulfilled by student population outside the project. Sixteen out of seventeen questions had answers in the form of a *Likert scale*, i.e. statements that could be rated from 1 to 5, 1 being 'strongly disagree' and 5 meaning 'strongly agree'. The first five questions were based around the simplicity and functionality of the prototype. Other questions determined the quality of the content used, and lastly if the prototype has a clear and interesting use. Fourteen participants answered the questionnaire in time for the evaluation. Table 1 shows sixteen questions used in the user satisfaction questionnaire and average mark calculated for each question.

TABLE I. USER SATISFACTION QUESTIONNAIRE AND AVERAGE MARKS

No	*Question*	*Avg. mark*
1	The purpose of the application is clear.	4.64
2	The application was easy to use.	4.43
3	Navigating through the application was intuitive.	4.21
4	The application has no technical flaws.	4.14
5	The application design is suitable for target users – children	4.71
6	The order of levels of different difficulty is appropriate.	4.57
7	The use of markers for each level is clear.	4.93
8	The 3D models that appear on markers are appropriate, considering the purpose of the application.	4.86
9	The narration in the application is comprehensible.	4.71
10	The narration in the application is educational.	4.57
11	The videos in the application are comprehensible and educational.	4.86
12	The questions in the quiz are useful and appropriate.	4.07
13	The bonus level is a good incentive to use the application.	4.29
14	The use of AR enhances the user experience.	4.71
15	The application was interesting to use.	4.36
16	Overall, I am satisfied with the application.	4.29

The results showed an overall positive feedback, with few neutral answers. When asked if the purpose of the prototype was clear, 85,7% of students gave the highest grade, or they 'strongly agreed'. When asked about the ease of design and use, on average 48,23% of the students 'strongly agreed' and 42,88% of the students 'mostly agreed' with the rest giving neutral answers. In the second part, 65,18% of the students 'strongly agreed' when asked about the functionality of the prototype and practical use in targeted groups of users, 28,8% 'mostly agreed' while the rest gave neutral feedback. In the final part, where the students were asked if the prototype was interesting and their overall impression, feedback consisted of 47,6% students answering, 'strongly agree', 50% 'mostly agree' and only a small percentage of neutral answers. This questionnaire mostly helped with fixing bugs and design or navigation flaws since the evaluators are not the targeted users for the prototype. The evaluation with rehabilitation and educational experts is next step that is currently in the progress.

VI. CONCLUSION

Today's high technology provides numerous opportunities for creating educational content for all generations. It can contribute to the creation of an interesting solutions that enables and encourage the communication process, not only for children with developmental disabilities or CCN, but also for typically developing children. When it comes to AR which implies an interactive experience of a real environment, the impact of using it cannot be negligible. The AR-based software prototype aimed for mastering vocabulary for children with CCN provides insight into the phases that are necessary to be followed and implemented in the developing process of the software prototype based on augmented reality. The multidisciplinary approach is one of the most important preconditions in the process of developing high tech solutions designed for children with CCN. The developing process based on the project form approach consist of the following phases: defining the project team members, defining the channels for communication, planning, choosing the technology and developing phase – different kind of versions were tested before the final one was developed. The development of such solutions implies a significant investment of time and human resources in the context of inputs of experts from various fields. Described software solution indicates that the possibilities of a joint action - including students and professionals from different research fields could be very productive and socially responsible. The developing process described in this paper with the accent on the multidisciplinary approach, besides good practical example related to the development process in the field of augmented reality, can also serve as a strong motivation base for all students and their mentors working together.

REFERENCES

[1] D. McNaughton and J. Light, "The iPad and Mobile Technology Revolution: Benefits and Challenges for Individuals who require Augmentative and Alternative Communication," Augmentative and Alternative Communication, vol. 29, no. 2, pp. 107-116, 2013.

[2] J. Light, D. McNaughton, D. Beukelman, S. K. Fager, M. Fried-Oken, T. Jakobs and E. Jakobs, "Challenges and opportunities in augmentative and alternative communication: Research and technology development to enhance communication and participation for individuals with complex communication needs,"

Augmentative and Alternative Communication, vol. 35, no. 1, pp. 1-12, 2019.

[3] J. Carmigniani and B. Furht, "Augmented Reality: An Overview," in Handbook of Augmented Reality, Springer, 2011, pp. 3-46.

[4] I. Skorić, Augmented Reality, Zagreb: University of Zagreb, Faculty of Electrical Engineering and Computing, 2019.

[5] J. Werner, "Augmented Reality | John Werner | TEDxAsburyPark," YouTube, 3 August 2017. [Online]. Available: https://www.youtube.com/watch?v=RDvBowq3ed8. [Accessed 12 May 2019].

[6] (unknown), "Augmented reality," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Augmented_reality#Workplace. [Accessed 12 May 2019].

[7] B. Debiesse, H. Marzuoli and C. Tourdjman, "Gilbert Klopfstein, The father of the Head Up Display (HUD)," HeadUpFlight.net, [Online]. Available: http://www.headupflight.net/gilbert/ArticleKlopCorr.htm. [Accessed 12 May 2019].

[8] BAE Systems, "The evolution of the Head-Up Display," BAE Systems, [Online]. Available: https://www.baesystems.com/en/feature/our-innovations-hud. [Accessed 12 May 2019].

[9] BAE Systems, "Products & Services: Striker II Digital Helmet-Mounted Display," BAE Systems, [Online]. Available: https://www.baesystems.com/en-uk/product/striker-ii-digital-helmet-mounted-display. [Accessed 12 May 2019].

[10] B. Opall-Rome, "IronVision Helmet Provides Sight Through Armored Tanks," Defense News, 8 June 2016. [Online]. Available: https://www.defensenews.com/land/2016/06/08/ironvision-helmet-provides-sight-through-armored-tanks/. [Accessed 12 May 2019].

[11] BMW Group Press Club, "Absolutely real: Virtual and augmented reality open new avenues in the BMW Group Production System," BMW Group, 9 April 2019. [Online]. Available: https://www.press.bmwgroup.com/global/article/detail/T0294345EN/absolutely-real:-virtual-and-augmented-reality-open-new-avenues-in-the-bmw-group-production-system?language=en. [Accessed 12 May 2019].

[12] Boeing, "Boeing Tests Augmented Reality in the Factory," Boeing, 19 January 2018. [Online]. Available: https://www.boeing.com/features/2018/01/augmented-reality-01-18.page. [Accessed 12 May 2019].

[13] D. Dasey, "IKEA Place," IKEA, 2017. [Online]. Available: https://highlights.ikea.com/2017/ikea-place/. [Accessed 12 May 2019].

[14] Microsoft, "HoloLens 2," Microsoft, [Online]. Available: https://www.microsoft.com/en-us/hololens?icid=VRMRCat_Nav1_Sydney_022419#coreui-feature-jf1uj33. [Accessed 13 May 2019].

[15] C. Jee, "US Army soldiers will soon wear Microsoft's HoloLens AR goggles in combat," MIT Technology Review, 29 November 2018. [Online]. Available: https://www.technologyreview.com/f/612490/us-army-soldiers-will-soon-wear-microsofts-hololens-ar-goggles-in-combat/. [Accessed 12 May 2019].

[16] P. Mealy, »Designing Augmented Reality Apps: Interacting with Objects,« Dummies, [Online]. Available: https://www.dummies.com/software/10-augmented-reality-mobile-apps/. [Accessed 27 February 2020].

[17] M. Abraham and M. Annunziata, "Augmented Reality Is Already Improving Worker Performance," Harvard Business Review, 13 March 2017. [Online]. Available: https://hbr.org/2017/03/augmented-reality-is-already-improving-worker-performance. [Accessed 12 May 2019].

The Perception of the Fake News Phenomenon on the Internet by Members of Generation Z

Lordan Prelog, Ljubica Bakić - Tomić
Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
lprelog@unizg.h, ljbakictomic@bak.hr

Abstract - **Studies from last year conducted by the Reuters Institute and Eurobarometer indicate that most Europeans are often confronted with fake news and disinformation in the media, mostly on the Internet. Members of Generation Z (born in the mid-90s to 2010) are a younger demographic group which, by the end of 2020, will make a third of the world's population. In the near future, members of Generation Z will be an influential group of young citizens whose habits, in the context of media consumption and the perception of the phenomenon of fake news and misinformation, will become very respectable. The aim of this study is to examine how members of Generation Z consume and evaluate the information they receive through various media, including the Internet, their attitudes toward misinformation and fake news and to determine to what extent their attitudes differ in comparison with older citizens. The survey was conducted in 2019 on the sample of 152 students from the University of Zagreb. Attitudes were found to differ between the members of Generation Z and older generations.**

Keywords: Generation Z, fake news, disinformation, news, media credibility

I. INTRODUCTION

Members of Generation Z are the largest age group in the world, and according to Worldometer data[1], it is estimated that in January 2020 more than 23% of the total world population was between the ages of 10 and 24 which means that there is slightly less than 2 billion members of this generation of people in the world. These are young people born from the mid-1990s to 2010 and although there is no consensus on the initial years of birth of now the second youngest demographic group in the world (generation Alpha is considered as children born since 2010), Pew Research Center defines Generation Z as people born after 1997.[2] Cambridge Dictionary[3] states that Generation Z has been called 'Silent Generation' because of the time they will spend online“ and Collins Dictionary[4] defines them as „confident users of new technology“.

In addition to 'Silent Generation', Generation Z is also described as 'iGeneration', 'YouTube Generation', 'Post millennials' and, most commonly, as 'Digital Natives'. Accordingly, it is logical that members of this generation, compared to the previous (Millennials, people born between 1981 and 1996)[5], are much more tech savy than any generation born before them because they are 'constantly online' and the vast amount of information is available to them from an early age. Specifically, as millennials grew up, they went through technological changes in the sense that they had to adapt to them. Members of 'Silent Generation' began to emerge two years after Berners-Lee designed the World Wide Web. Generation Z literally grew up with the Internet which they use for consuming the news as well as a platform for socialization and communication. Numerous aspects of the digital revolution and Web 2.0 have undeniably influenced the Generation Z and the way how they live and communicate.

According to American technologist and writer Marc Prensky[6], who coined the term 'Digital Natives' back in 2001, „Our students today are all “native speakers” of the digital language of computers, video games and the Internet“.

Members of Generation Z are 'on the go' on cell phones and are quicker to absorb and find information and content of interest on the Internet compared to past generations because it is natural to them. Generation Z ‘were raised' on social networks and online sources of information they use in order to save time (searching, informing) as well as for 'wasting time’ (playing games, listening to music, watching movies). A 2018 PEW survey 'Teens, Social Media and Technology'[7] found that as many as 95% of American teens have access to a smartphone at home, with 45% of respondents claiming to be 'almost constantly' online (as opposed to 25% respondents from a similar survey conducted in 2014/2015). When asked about the impact of social media, only 16% of respondents said that social media makes it easier for them to access news or information.

Regarding the usage of social networking sites among young Americans, results of the survey[8] conducted by the world's largest professional services firm Ernst & Young in 2019 on a sample of more than 2,000 teenagers and young people, indicate that Generation Z members use social networks primarily to connect with family and friends (80% of respondents). In addition, as many as 84% of respondents confirmed that they have an open profile on some social network, even though less than half of them uses internet to inform themselves (41%). In this study, only 22% of survey participants stated that they

shared some personal opinion or attitude through social networks.

Another survey conducted by IPSOS in 2018 shows that younger people, members of Generation Z, have less confidence in the news than older people[9]. Furthermore, as many as 67% members of this pro-active generation of individuals stated that 'people cannot be trusted', which is an attitude that certainly indicates that this demographic group is 'additionally careful' when confronted with misinformation and fake news they see on the internet.

On the other hand, a scientific study „Students' Civic Online Reasoning“ conducted from July 2018 to May 2019 on a representative sample of 3,446 US high school students found that Generation Z members had serious difficulties evaluating digital information sources on the Internet. Specifically, according to a report released in November 2019 by Stanford History Education Group[10], two-thirds of respondents were unable to determine the difference between a paid ad and a post or an article on the popular website Slate.com, while more than half of high school students believed the video they saw was an example of manipulation with election results (video was filmed in Russia) and 'strong evidence' that voters were duped in the 2016 US election.

Regarding the credibility of media, in a 2019 report „How Young People Consume News and Implications for Mainstream Media“[11] conducted in London and New York by the Reuters Institute for the Study of Journalism, among members of both genders of Generation Z (18 to 24 years) and Generation Y (25 to 35 years), the researches have concluded that young people find misinformation more 'nuisance' than a problem, and that the decline of trust in the media for the mentioned generations is not so worrying as it seems in the public discourse.

The situation in Croatia, as expected, is somewhat different from those in more developed countries. Eurobarometer survey “Fake news and disinformation online”[12] from 2018 showed that most Croats, as well as Europeans, are often confronted with fake news in the media, predominantly on the Internet. The survey was conducted on 1,005 respondents from Croatia and it showed that 47% of surveyed citizens consider they encounter fake news every day. In addition, 35% of the subjects claimed it is difficult to differentiate between fake and real news. Overall, 86% of respondents in Croatia perceive fake news as a problem, and 80% of Croats think fake news are a problem for democracy in general.

According to the Croatian Bureau of Statistics[13], around 642,000 people between the ages of 10 and 24 were registered in Croatia in mid-2018. It is a large and in the near future influential group of citizens whose traits and habits, especially in the context of media consumption and perception of the phenomenon of fake news and misinformation, have not been studied in detail in Croatia so far.

In that context, it is interesting to look at the results of a survey “Media Habits in the Republic of Croatia”[14] conducted by IPSOS in March 2019 where Generation Z represents the population between the ages of 18-29. The survey showed that Generation Z is still heavily using the Internet both to communicate via email and to search for useful information or information on current affairs. Also, the survey showed that this generation uses social networks to communicate far more than any other age group. According to the same source, in 2017 the Internet surpassed television in Croatia in its perception of reliability as the most trusted source that delivers reliable and accurate information.

Another recent survey "Social Online Experiences and Mental Health of Young People"[15] conducted in Croatia on a sample of 1772 high school students by the Polyclinic for the Protection of Children and Youth Health of the City of Zagreb revealed that 95.1% of adolescents access the internet via smartphone, with every fifth respondent using the social network for more than 5 hours a day. As in the rest of world, our high school students mostly use social network YouTube (97.7% of respondents), and the popular messaging application WhatsApp is used by 88.7% of high school students. In comparison, according to data from the 2019 Reuters Digital News Report[16], 88% of Croats are online every day and in January 2019 the respondents shared a post on Facebook at least once.

How members of Generation Z consume and evaluate news and what is their opinion on misinformation and fake news that are ‘invading’ the Internet are some of the questions that this research will try to answer.

II. RESEARCH METODOLOGY

The aim of the study was to find out what young people today (Generation Z) think of the phenomenon of fake news and how they deal with it as well as how aware are they of their own intentions.

The following hypotheses were made:

H1: Students at the University of Zagreb (Generation Z) consume the news and are familiar with the concept of "fake news" and they can recognize them in the media.

H2: Students at the University of Zagreb (Generation Z) tend to inadvertently share fake news on social networks.

H3: The dominant social network for communication between University of Zagreb students, members of Generation Z, is Facebook.

The research method used in the study was a survey for which a questionnaire was created based on the experiences and insights of different authors who dealt with the phenomenon of fake news. The questionnaire

consisted of 13 questions: the questions about the habits of searching for the information in the online media and forwarding it to others; which are the respondents' favorite media, portals, and social networks for communicating and sharing information and other.

The research was conducted in April 2019 at the Student Center of the University of Zagreb. To reduce the baseline, we selected only the Z generation for research participants. The sample was convenient. The questionnaire was filled out by 152 respondents of the Z generation born in the period from 1995 to 2001, namely 82 persons of female gender and 70 persons of male gender.

III. ANALISYS OF RESULTS WITH DISCUSSION

Table 1 shows the structure of the survey sample. According to the statistical indicators of the structure of newborns in the Republic of Croatia for the mentioned period, the sample of this research has no major discrepancies, so although it is a suitable sample, the results may be acceptable as preliminary research and a guide to similar researches in the future.

As many as 86.80% of surveyed students, members of generation Z, responded positively to the question 'Do you follow the news on daily basis?' and 62.40% of them stated that they are familiar with the phenomenon of 'fake news'. Interestingly, 40% students who completed the survey said they do not think they could recognize 'fake news' in the media. We can conclude that hypothesis H1 "Students at the University of Zagreb (Generation Z) consume the news and are familiar with the concept of "fake news" and they can recognize them in the media" is partially confirmed.

As a rule, 59.20% of surveyed students do not check the credibility of suspicious news on another source of information, while 40.80% does it regularly. This minority represents the more conscientious and more

YEAR OF BIRTH OF RESPONDENTS, students of University of Zagreb	NUMBER OF RESPONDENTS (N: 152); 82 (F), 70 (M)
1995.	22
1996.	29
1997.	31
1998.	9
1999.	18
2000.	28
2001.	15

Table 1: Structure of the surveyed sample of respondents, students of University of Zagreb, members of Generation Z

curious part of the Generation Z students, we might call them 'truth seekers'. Furthermore, as many as 25% of respondents believe they have never shared 'fake news' on social networks, while 21.70% of them were unsure whether they might have sent 'fake news' to friends unknowingly or sometimes. It is worrying to know that 53.30% of the students who participated in the research were aware that there were 'fake news' among the other news they shared with peers, but they nevertheless decided to share them with others. This last data leads to the conclusion that the hypothesis H2 which reads "Students of the University of Zagreb (Generation Z) tend to inadvertently share fake news on social networks" is considered confirmed.

According to our research, 57.20% of students, members of Generation Z, mostly use WhatsApp for communication with their peers although majority of them have previously opened a profile on various and multiple social networks and communication platforms. The world's biggest social network Facebook is in the second place, chosen as dominant by 34.20% of students. Social network Instagram is used for communication by only 7.90% of students at the University of Zagreb. It can be concluded that hypothesis H3 "The dominant social network for communication of students of the University of Zagreb, members of Generation Z, is Facebook" has not been confirmed.

Disinformation and false or fake news that emerge in the media can be verified, but we often do not have enough time to check them. How to distinguish fake news from 'real' news? In part, our experience and intuition can help us, because we can assume which information is false and which is not. But, it is only a conjecture, a belief, even though after our evaluation we may again check the credibility of news (or not). What happens if we recognize certain information as interesting, humorous or interesting, even if it is not true? We often share such information with the intention of amusing other people. Or vice versa, what if we perceive the information as negative or malicious and then forward it to our "enemies" so they would be negatively affected? We are in a sphere of intent that is central to the impact of fake news on individuals.

IV. CONCLUSION

The young people we call members of Generation Z have a strong interest in new media, especially the Internet, where they seek information, but they also use it to communicate through messages on various platforms. Among the numerous social networks for communication and information sharing, WhatsApp is the most popular social network among the students of University of Zagreb and Facebook is the second most popular application. In this paper we tried to find out more about the media consumption habits and the perception 'fake news' phenomena of students of University of Zagreb, members of demographic group Generation Z. We find very interesting and worrying the fact that majority of students who participated in the research do not check the information they find on the internet. Students are aware that there are false information among the ones they have shared with their friends, but they do not consider it as a problem or a reason for concern. The question that arises

is: what does Generation Z think about truth in general and their responsibility for spreading 'fake news'? This is a crucial issue that needs to be explored in detail so we could determine more clearly whether Generation Z is different from other generations in using the Internet and social networks.

REFERENCES

[1] Worldometer, https://www.worldometers.info/demographics/world-demographics/ (12.12.2019.)
[2] Pew Research Center, „Defining generations: Where Millennials end and Generation Z begins", January 2019, https://www.pewresearch.org/fact-tank/2019/01/17/where-millennials-end-and-generation-z-begins/ (17.1.2019.)
[3] Cambridge Dictionary, https://dictionary.cambridge.org/dictionary/english/generation-z (12.12.2019.)
[4] Collins Dictionary, https://www.collinsdictionary.com/dictionary/english/generation-z (12.12.2019.)
[5] Pew Research Center, „Defining generations: Where Millennials end and Generation Z begins", January 2019, https://www.pewresearch.org/fact-tank/2019/01/17/where-millennials-end-and-generation-z-begins/ (17.1.2019.)
[6] Marc Prensky, Digital Natives, Digital Immigrants, On the Horizon, NCB University Press, Vol. 9, No. 5, October 2001
[7] Pew Research Center, „Teens, Social Media & Technology 2018", May 2018, https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/ (11.1.2020.)
[8] Ernst & Young LLP, „Gen Z Segmentation Study", January 2020, https://www.ey.com/en_us/news/2020/01/new-gen-z-data-brings-businesses-a-critical-question-whatsyour-plan-z (5.2.2020.)
[9] IPSOS, „Ipsos Thinks Gen Z: Beyond Binary", https://thinks.ipsos-mori.com/trusting-cynical/ (18.1.2020.)
[10] Stanford History Education Group, „Students' Civic Online Reasoning", November 2019, https://stacks.stanford.edu/file/gf151tb4868/Civic%20Online%20Reasoning%20National%20Portrait.pdf , (1.2.2020.)
[11] Reuters Institute for the Study of Journalism, „How Young People Consume News and Implications for Mainstream Media", 2019, https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-08/FlamingoxREUTERS-Report-Full-KG-V28.pdf(25.1.2020.)
[12] Flash Eurobarometer 464 „Fake news and disinformation online", April 2018, https://ec.europa.eu/digitalsingle-market/en/news/final-results-eurobarometer-fake-news-and-online-disinformation (9.12020.)
[13] Croatian Bureau of Statistics, Croatia in numbers 2019, Zagreb 2019, https://www.dzs.hr/Hrv_Eng/CroInFig/croinfig_2019.pdf (7.1.2020.)
[14] IPSOS, "Media Habits in the Republic of Croatia", March 2019, https://showcase.24sata.hr/2019_hosted_creatives/medijske-navike-hr-2019.pdf (18.1.2020.)
[15] Polyclinic for the Protection of Children and Youth Health of the City of Zagreb, „Social Online Experiences and Mental Health of Young People", December 2019., https://www.poliklinikadjeca.hr/istrazivanja/rezultati-nacionalnog-istrazivackog-projekta-drustvena-online-iskustva-i-mentalnozdravlje-mladih/ (27.2.2020.)
[16] Reuters Digital News Report 2019, http://www.digitalnewsreport.org/survey/2019/overview-key-findings-2019/ (26.1.2020.)

TABLES

Table 1: Structure of the surveyed sample of respondents, students of University of Zagreb, members of Generation Z

Smart Cities as an Opportunity and Challenge for People with Disabilities

Darko Švelec*, Nada Bjelčić**, Mislav Blažeković***
*HT d.d., Osijek, Croatia, **HT d.d., Zagreb, ***Croatia HT d.d., Osijek, Croatia
*darko.svelec@t.ht.hr, **nada.bjelcic@t.ht.hr, ***mislav.blazekovic@t.ht.hr

***Abstract* - The paper analyzes the possibilities and potential for future IoT Smart City applications which could improve the living and working conditions of people with disabilities (PwD) and considers several key factors for achieving this goal. It shows the presence of PwD in the general population, their actual social significance, and the specific necessities for their integration into regular life and work. Next part relates to the capabilities of the communication and information infrastructure related to this specific objective. Furthermore, the important factor is development of IoT Smart City applications obeying the principles of universal design. The appliance of Digital Twin Technology could play a significant role in this development. In conclusion, the paper provides a vision of an approach to the sustainable development of IoT Smart City applications that are highly tailored to the needs of PwD.**

Keywords: smart city, people with disabilities, elderly population, infrastructure, applications, digital twin.

I. INTRODUCTION

People with disabilities (PwD), and the elderly, should be observed together because they share similar needs, and need a higher level of various adaptations to their abilities. These adaptations can go in two directions. One is towards the environment and the other towards the PwD itself. Adaptation to the environment means changing the usual ways and procedures of communication and action of the environment towards the PwD with the aim that the PwD can be included in the world with its limitations. Experience shows that such adaptation is relatively difficult for the environment. Adapting PwD to the environment involves relying on some kind of interface that will adapt their possibilities to the capabilities of the environment. There are two forms of such an interface, technical aids including devices and targeted software products, and personal assistants. Devices and software products are known helpers but of very limited reach. The price of development and the products themselves is particularly problematic. Therefore, their application is relatively underrepresented among PwD.

However, with the appearance of the IoT concept, the situation is changing significantly. The IoT concept aims to facilitate daily routines as well as more advanced activities during the day, at work, at home, in social activities, in health needs, etc. Furthermore, IoT applications in their development have an open approach and tend to universal design. This means that the impact on their development as well as the possibilities of use are oriented to as many users as possible (both active and potential). Smart city services and applications as part of IoT are of particular interest to PwD. In cities the PwD population is relatively more represented than in areas outside cities. On the other side, PwD, due to the nature of their constraints, expect much more from the city's public services than other people. Also, they have a lot more benefits from expanding and improving that set of services than other people. Therefore, the concept of a smart city as a derivative of the IoT concept provides them with a chance like never before. It should also be considered that the urban population of PwD is more oriented towards technological assistants than personal assistants, so they are already significantly oriented towards technological solutions.

This paper discusses the structure and concept of connecting the needs of PwD and Smart City applications through two levels, from adapting and using applications, to managing of their development. The first part of the paper gives an analysis of the structure and needs of the PwD population. The second part gives a model of defining applications for PwD in Smart City. The paper continues with an overview of some technical challenges in the implementation of services. The final part of the paper provides an analysis of the digital twin technology (DTT) as a method of optimizing the adaptation of smart city services and the needs of PwD and as a way of influencing the development of services.

II. STRUCTURE AND NEEDS OF PWD POPULATION

There is no unique, universally accepted definition of a person with a disability. However, they all consider physical and/or mental impairment and interaction with the environment [1] [2]. In general, various legal acts, including national laws, observe PwD as a person who has a physical or mental impairment that has a significant and long-term adverse effect on his or her ability to perform daily routine activities. This ability strongly depends on adjustment of environment which can increase/decrease the level of person dependency.

According to a World Health Organization (WHO) report, in 2011., between 15-19 % of the world's population lives with some form of disability, of whom 2-4% have significant difficulties in day-to-day activities [3] [4]. The global prevalence of PwD is higher than previous WHO estimates from 1970s which presented with a figure of cca 10%. By a UN report [6] from 2008. cca 15% of

population live with some form of disability. Furthermore, unofficial report from June 2019 claims figure of 20% [2].

This global estimate for PwD on the rise due to population ageing and the rapid spread of chronic diseases, as well as improvements in the methodologies used to identify disability.

The definition of elderly population is people older of 65. Because of aging of world's population, the share of elderly population is consequently growing. (Figure 1.) [8]. The spread of elderly population is not equal in country regions because they are concentrated only in few areas. These areas consequently have specific social and economic issues such as government and private spending on education, pensions, health care, traffic, economic growth and welfare [9]. Furthermore, vast majority of them are not so familiar with modern technology.

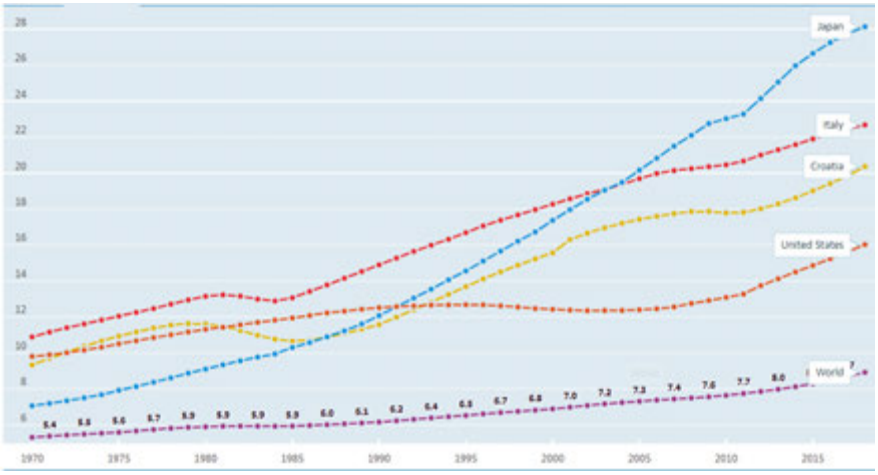


Figure 1. Elderly population share, 1970 – 2018

According to the nomenclature of Croatian Institute for Public Health (HZJZ), we can divide PwD into five basic groups according to the type of disabilities:

1. Mobility issues (MOI),
2. Mental and behavioral disorders (MBD),
3. Chronic illnesses (CRI),
4. Eyesight damage (EYD),
5. Hearing damage (HED).

Figure 2. shows the structure of PwD by type in the Republic of Croatia, as a percentage of the total number of persons with disabilities (N = 511 281) in 2020. in comparison to the 2019. It is about 12,6% of total population in 2020. in comparison with 12,4% in 2019. [10]. Each of the groups of PwD can use some of assistive technologies (ATs) to bridge communication problems [11]:

- Assistive listening devices (ALDs) amplify the sounds of oral speech,
- Augmentative and alternative communication (AAC) devices help to communicate in all forms exclude oral speech. It can be in a simple form of a picture board or in a complex form of computer program with synthesizing speech to text and via versa,
- Alerting devices which are connected to a doorbell, telephone or alarm, are emitting a loud sound or flashing light to warn people with hearing impairments,
- When computers and mobile devices are used there are set of different computer software and hardware, such as voice recognition programs, screen readers, and screen enlargement applications, that help people with motor and sensory impairments.

III. Smart City Applications and Persons with Disabilities

The "Smart City" concept is a comprehensive system of services and modern technology applications, which aims to improve the lives of its citizens [12]. Such a service system is based on IoT applications, encompassing all aspects of IoT applications:

- communication,
- the cloud technology as a repository of databases and platforms for algorithm of performing of services,
- information and management algorithms,
- the M2M addressing, and
- the 5G as a communication base.

A. Universal design

The planning of smart city applications should be guided by the principle of universal design [13] [14]. This means designing buildings, products or environments to make them accessible to all people, regardless of age, disability or other factors. Certain groups of PwD would still need supportive devices such as wheelchair regardless if the universal design concept was used.

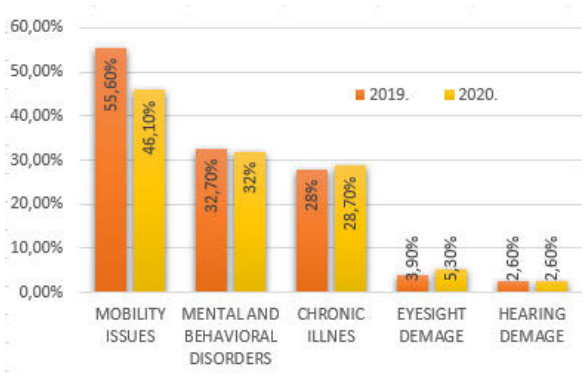


Figure 2. PwD structure and 2020-2019 comparison in Croatia

Universal design is achieved by:

- Designing products, services and applications that can be used by a large number of users without modification (accessible entrances, buildings without thresholds, remote control systems),
- Designing products that are easily adapted to other users (by adapting their user interfaces),
- Existence of standard interfaces that are compatible with specialized equipment.

From the definition of universal design, the design of products and services should take into account the easy adaptation to other users, ie, in this case, to the elderly population and PwD. Such services can be used directly, as they are, or with minor upgrades. In Article 9 of the Convention on the Rights of Persons with Disabilities (CRPD) [15] the economic, social and legal framework are defined. It emphasizes the obligation of Member States to provide access to information and communication technologies (ICT) in accordance with the physical environment and transportation. This gives importance and universality of ICT application in all areas of life in today's age of information.

From the PwD point we can consider two significant approaches of applications for PwD:

1. Applications designed for PwD (including standard applications that require significant customization for PwD),
2. Applications with integrated features for PwD (including applications that require minor customization for PwD).

First applications are rare because of the high price of development and relatively small market in comparison with market for consumers without disabilities. It can be designed universal or specialized for some of disability. On market exists diapason of different ATs that are used to support speech, hearing, visual, mental and motor impairment [16]. Table I. presents the types of ATs in correlation with an impairment type. We can see that there is no one universal AT that could suite everyone. Furthermore, within specific impairment AT usage depends of impairment level, e.g. for severe motor impairment only few sets of gesture are possible.

TABLE I. ASSISTIVE TECHNOLOGY FOR DIFFERENT IMPAIRMENT CONDITIONS

AT	IMPAIRMENT				
	Motor	Visual	Speech	Hearing	Mental
Voice recognition	YES	YES	NO	YES	YES
Voice to text/text to voice	YES/ YES	YES / YES	NO/ YES	YES /NO	YES/ YES
pictogram	YES	NO *	YES	YES	YES
Vibration/ alert light	YES/ YES	YES /NO	YES/ YES	YES/ YES	YES/ YES
Braille	NO	YES	NO	NO	NO
Gesture control	YES **	YES	YES	YES	YES

*depends of size of pictogram and of visual impairment level
**type of gesture depends of motor impairment level

Some inventive applications are integrated in wearable parts like insoles. The insoles are designed to work with a mobile device to help the user navigate a city without looking at a smartphone for directions. The insoles do this by vibrating to let the wearer know where to go, and to make recommendations for specific locations based on the wearer's learned behavior. This is a great invention for people with visual or hearing impairments as well as for people with orientation problems [17].

Second group is far easier to develop because it needs only few adjustments for PwD like Smart parking that includes information about parking slots for PwD. But, it has to be developed obeying universal design principles as we mentioned before. Some applications don't need any adaptation because they can be used by all users, like smart bulb which changes the color of light depend of programmed situation. It could help people with cognitive impairments navigate through the house or reminding them about things that they still need to do. Deaf people could use this device by setting the bulbs to shine in colors when there are certain noises like phone notifications or changes in weather.

Today, there are more then 5,5 million applications available on application repositories (iOS, Google play, Amazon and Microsoft combined) [18]. Very few of them are dedicated for PwD and only some of them have integrated features for PwD. The study that was conducted between 2015. and 2016. by the Rehabilitation Engineering Research Center for Wireless Technologies Shepherd Center, showed that 91% of PwD were using smartphones or tablets [19]. So, it is important to have in mind the needs and abilities of this population during the planning IoT apps.

B. Model of defining applications for PwD in Smart City

According to mentioned considerations we have to extend the set of IoT applications for PwD needs in smart cities to applications relevant to life and work in the city, structure them according to their needs into four basic groups and view them as an integrated set of services related to each other and to the environment.

In next step we will arrange the applications into 4 basic groups:

1. Mobility (MO),
2. Heath (HE),
3. Smart Home (SH),
4. Others (OT).

Such structuring of applications ensures their clear positioning within the smart city concept, clear relationships and connections with other stakeholders of smart city (Figure 2.). It also enables easier implementation of universal design approach and more appropriate analysis of the services regarding the needs of PwD.

Table II. shows links between previous considerations. It starts from the list of most often used smart city applications and connects them with the

TABLE II. DISTRIBUTION OF THE TYPE OF DIFFICULTY

application	difficulties by type					appl. group	appl. availability
	MOI	MBD	CRI	EYD	HED		
Smart Parking	H	L	M	M	M	MO	to improve
Smartphone Detection	M	M	M	H	H	MO	direct
Traffic Congestion	H	M	H	H	L	MO	direct
Smart Lighting	H	M	M	-	H	MO	direct
Perimeter Access Control	M	H	H	H	M	MO	direct
Smart Public Transportation	H	H	H	H	H	MO	to improve
Smart Stick	M	L	L	H	L	MO	to improve
Mobility band	L	M	L	H	H	MO	to improve
Smart Wheelchair	H	-	-	-	-	MO	to improve
Noise Urban Maps	M	H	M	L	-	HE	direct
Air Pollution	L	M	H	L	L	HE	direct
Weather condition	M	H	H	H	M	HE	direct
Potable water monitoring	L	L	H	H	M	HE	direct
Ozone Presence	L	L	H	L	L	HE	direct
Meteorological Station Network	M	M	H	H	H	HE	direct
Animal Tracking	M	M	M	M	M	HE	direct
Health control	M	H	H	M	M	HE	direct
Fall control	H	H	M	M	M	HE	direct
House guardian	M	H	M	M	M	SH	direct
Energy and Water Use	M	L	M	M	M	SH	direct
Remote Control Appliances	H	H	M	H	M	SH	direct
Smart light buble	H	H	M	L	H	SH	direct
Ring bell and door lock	H	H	M	H	H	SH	direct
Public safety	H	H	M	H	M	OT	direct
Intelligent Shopping Applications	L	H	H	H	H	OT	to improve

elements important for their usage for PwD. It shows the significance of applications according to the type of difficulties defined by Figure 1. and Table I. Significance is defined in three levels, H for high significance, M for medium and L for low. It also groups the according to Figure 2. Finally, it defined the usage possibility of application for PwD – without adaptation (direct) or with adaptation (to improve).

This model can be expanded by specifying certain region of country or city. In other words some applications will be of greater interest in some area compared to other, e.g. Air pollution application is relevant for regions with such problem like Slavonski Brod but not for Rovinj. This can be significant in making a decision which applications first to include in a Smart City design.

IV. PWD IOT APPLICATIONS AND INFRASTCTURE

IoT applications have a large and complex requirements toward communication and information infrastructure. These requirements arise from large amounts of expected applications, but also from specific and strict characteristics of applications itself.

Communication requirements are reflected in parameters such as the information transfer rate, capacity, and latency. These parameters relate to applications itself. Some PwD IoT applications require fast information processing, some are more demanding in terms of storage capacity. However, these parameters are connected also with users themselves, i.e. PwD, and their ability to the adaptation to a particular application. Furthermore, in the next step, the implementation of AI features, IoT applications independently will be able to manage the devices and make decisions.

Speed is determined by the need of users for quickly transmitting the information. Some applications, especially in the mobility and health groups, are looking for higher speeds for its very purpose. Examples could be a smart stick and fall detection. Some applications require a higher speed because the user due to their difficulties does not have the ability to wait a long time for a response. Examples could be people with mental disabilities. Latency, except in some vital situations like in the areas of health and mobility, has a much greater significance in industrial applications. Capacity has significance because of the amount of data that is

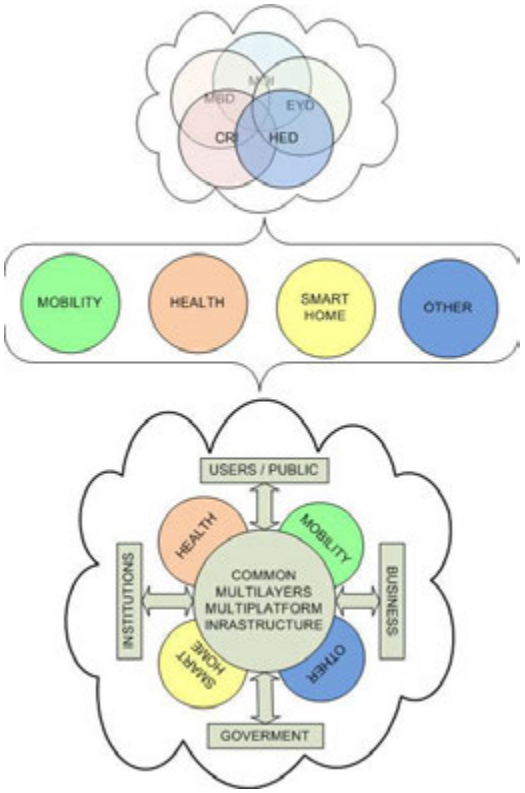


Figure 3. Application structure in Smart City Environment

transferred through the network, including application dedicated servers with cloud functionality. In fact, cloud technology is key to processing requests generated by PwD trough their IoT applications and for delivering responses that are more important to PwD than to ordinary users. Therefore, when implementing such applications, it is necessary to establish connections between the disability that the users have (MOI, CBD, CRI, EYD, HED), the functioning of applications (MO, HE, SH, OT), and the complex communication and information structure that ensures their implementation (Figure 3.).

The communication infrastructure of the entire 4IR, such as the PwD IoT, is mainly the 5G network as a lever that drives massive scale and is crucial to a huge increase in opportunities for people with disabilities. However, because of their specific purpose for persons with disabilities, it is necessary to ensure a slightly higher level of QoS for their use. In terms of capacity for most IoT PwD applications, NB-IoT is enough in principle. NB-IoT is a standardized radio technology for connecting a very large number of devices and services. The emphasis is on long-range, penetration into hard to reach places, low power / long battery life (> 10 years), low cost devices in large quantities that communicate through a single base station. The challenge, however, is the amount of traffic that can be expected. In some applications, such as Traffic Congestion, one request may be expected to carry more data and use more of the spectrum capacity. Latency and speed, on the other side, much more depend on the application itself and can significantly vary.

From the present point of view, the main challenge is spectrum capacity. This challenge can be significantly reduced by applying the principles of scalability and flexibility in spectrum use.

The principle of flexibility allows the same service to be realized with frequencies from different parts of the limited spectrum that are designated for 5G (700 MHz, 2.6 GHz, 26 GHz). This is achieved by covering the area with base stations for different parts of the frequency spectrum. In doing so, the same application, depending on the need and possibilities, can use any of these parts of the spectrum. This achieves an optimization effect on the use of the 5G spectrum given the needs of users and applications on the one hand and the capabilities of the network on the other.

IoT concept is basically user and service-oriented and as such is not dependent on the implementation of technology itself. However, PwD IoT applications themselves can also be technologically scalable, relying on a variety of wireless transmission technologies. The scalable use of wireless technology is an approach that considers the structure and amount of services provided as well as the structure and amount of service users. Therefore, the use of a heterogeneous set of wireless technologies is an optimizing way of delivering IoT services with respect to spectrum capabilities, and with respect to enhancing the capabilities of a portion of the spectrum that the legacy of historical technology. Scalability of addressing communication needs PwD IoT applications can be achieved by harmonizing various wireless technologies:

- Proximity (RFID, NFC),
-Wireless Personal Area Network, WPAN (Bluetooth),
- Wireless Local Area Network, WLAN (Wi-Fi),
- Wireless Wide Area Network, WWAN (LTE, 5G).

This approach of the flexible and scalable use of wireless technologies and associated parts of the frequency spectrum allows a minimal dependence of the application on the capabilities of the spectrum. On the other hand, this optimizes the spectrum usage, i.e. enables the applicability of spectrum to many applications.

V. DIGITAL TWIN TECHOLOGY IN PLANINING SMART CITY

Digital twin models are basically digital copies of their physical counterpart entities. This includes replicating the geometry, processes, and interactions of an object as well as tracking changes as they occur in the actual real-life system. The DTT concept is not a novelty but the development of 3D data capturing technologies along with cloud computing in conjunction with Artificial intelligence (AI) resurgence has provided a platform for cost-effective solutions. Thanks to their scalability, DTTs can be used to replicate areas of whole cities or even countries. Therefore, the technology is very suitable for IoT infrastructure providers like telecom networks who are core for Smart City and IoT applications. With DTT it would be possible to plan and optimize the network architecture in complex urban environments. As different structures are more and more cramped in a modern city it is challenging to utilize higher frequency bands like 5G needed for IoT implementation.

Consider the hypothetical example of a network that today has 10,000 base station sites serving 10 million human subscribers. Let us assume that in the next 5 years, this network will need 100,000 extra sites to support 100 million new connected IoT devices, and advanced 5G use cases requiring 10 x lower delay and 100 x times faster data speeds. The increasing complexity is going to be pushing the limits of traditional network planning methods. One way to overcome this is by utilizing advanced digital twin technology (DTT) possibilities with its ability to simulate the real world. In this simulation all of the available real-world data can be fed into the DTT model the physical environment, usage at different locations, mapping, buildings, vegetation lifecycles, historical weather conditions, and the location of street furniture.

When it comes to network planning and optimization, a DTT in combination with AI provides far more sophisticated software tool that opens a whole new world of possibilities. Unlike traditional techniques, it incorporates everything that affects the network including data about where and how people are using it, as well as trends gathered from social media [20].

The digital twin is like a bridge between the physical and the digital world. First, smart components that use sensors to collect real-time, operational, or positional data are integrated with physical items. The components are connected into a cloud-based system that receives and

processes all the data that the sensors monitor. This listing is analyzed based on business and other contextual information. Lessons are learned and opportunities in the virtual environment that can be applied to the physical world are discovered - ultimately to transform a specific IoT application to become PwD adaptive. Artificial intelligence and connectivity in the cloud opens the possibility.

The purpose of DTT is to create, test and build a process, product or service in a virtual environment. This pairing of the virtual and physical worlds enables data analysis and system monitoring to solve problems before they even occur, prevent downtime, develop new capabilities, and even plan with simulations. It could be also used to develop, validate and share concepts of universal design which would be in service for PwD. DTTs cover the entire life cycle of their real-life entity and form the basis for related products and services. Last, but not least, costs of creating, testing and building processes, products or services are scientifically reduced.

Connected products and services are the new paradigm of the business world, and in the case of PwD IoT services, they are exactly the model of development that is needed and that maximally incorporates their needs and capabilities.

The fastest and most effective way to develop and adapt IoT applications to specific needs such as PwD needs while minimizing the risk of PwD failure to accept the application as a user is via DTT. Since these are software simulations for which physical limitations are becoming less significant. On the other hand, educations that are becoming more significant are becoming more and more accessible in a virtual digital environment and with fewer and fewer obstacles.

VI. CONLUSION

IoT applications from the Smart City application group go beyond their original framework, in several directions. The variety, quantity, and applicability of such applications are increasing. Besides the most common targeted market of users, the population of PwD and the elderly in cities is significant and is constantly growing and it is expected to be a trend in the future too. Therefore, investing in the development and use of IoT applications in a smart city for PwD is an investment in them themselves, but also in social and economic development in general. This paper describes needs and specificities od PwD as well as IoT applications for PwD. It described our model of defining applications for PwD in Smart City. According to this model, the architecture of the network of Smart City has to be prepared for easy inclusion of application for PwD. That would be the main initiator for developers to develop applications using principles of universal design. As the main technological lever for the development of the applications is that they would be designed for all but especially for the PwD and the elderly, DTT has great potential. It makes it possible to decrease the cost of the development process of IoT applications in general and increase the quality of planning a Smart City in smart and quick way.

As a further step, several, for instance 5, trial applications should be selected based on PwD user needs analysis and infrastructure capabilities. Selected applications should cover all four described groups of applications (MO, HE, SH, OT) and should cover all disability groups in accordance with the division given in the work (MOI, MBD, CRI, EYD, HED). It would be very good if the selected applications cover more than one disability group at the same time, i.e. to be oriented to the type of impairment (Table I.). These applications would be piloted with two objectives. The first goal is to find a development model for PwD IoT Smart City applications using universal design. This development model would include the active participation of the PwD population and the use of simulation methods such as DTT. Another goal is to determine the viability of such applications in the IoT market. The inclusion of the PwD population and DTT technology can significantly contribute to this second goal. Even more, the inclusion of the PwD population has to be the future obligation of all participations obeying, as their moto actually says, “Nothing about us without us”.

REFERENCES

[1] https://www.disabled-world.com/definitions/disability-definitions.php ; May, 2020.

[2] https://www.disabled-world.com/disability/ ; May, 2020.

[3] https://www.who.int/disabilities/world_report/2011/report.pdf ; May, 2020.

[4] https://en.unesco.org/partnerships/partnering/icts-persons-living-disabilities ; May, 2020.

[5] https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS ; May, 2020.

[6] https://www.un.org/disabilities/documents/Disability_and_Urban_development.pdf ; May, 2020.

[7] https://unstats.un.org/unsd/demographic-social/sconcerns/disability/statistics/#/countries ; May, 2020.

[8] https://data.oecd.org/pop/elderly-population.htm; May, 2020.

[9] [https://data.oecd.org/pop/elderly-population.htm] ; May, 2020.

[10] https://www.dzs.hr/Hrv/publication/StatisticsInLine.htm; May, 2020.

[11] https://www.nichd.nih.gov/health/topics/rehabtech/conditioninfo/device; May, 2020.

[12] Deakin, Mark and Al Waer, Husam, "From Intelligent to Smart Cities", Journal of Intelligent Buildings International: From Intelligent Cities to Smart Cities, vol 3 (3), 2011, pp 140–152.

[13] http://cdrnys.org/blog/advocacy/ronald-mace-and-his-impact-on-universal-design/; May, 2020.

[14] http://universaldesign.ie/What-is-Universal-Design/The-7-Principles/; May, 2020.

[15] https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html; May 2020.

[16] Ghassan Kbar, Mustufa Haider Abidi, Syed Hammad Mian, Ahmad A. Al-Daraiseh and Wathiq Mansoor, “University-Based Smart and Context Aware Solution for People with Disabilities (USCAS-PwD)”, MDPI, Computers, Volume 5, Issue 3 2016, article 2, https://doi.org/10.3390/computers5030018; May 2020.

[17] https://medium.com/@imn/5-promising-examples-of-iot-and-wearable-devices-that-enable-people-with-disabilities-f50df601e046; June, 2020.

[18] https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/ ; May, 2020.

[19] https://www.inclusivecitymaker.com/infographic-smartphone-use-disabled-people/ ; May, 2020.

[20] https://www.nokia.com/blog/digital-twin-technology-double-meet-challenges-iot-5g/; May, 2020.

COMPUTING IN TELECOMMUNICATIONS

Cloud Computing Virtualization: A Comprehensive Survey

Amarildo Rista *, Jaumin Ajdari *, Xhemal Zenuni*
* SEE University/Faculty of Contemporary Sciences and Technologies, Tetovo, North Macedonia
ar29102@seeu.edu.mk; j.ajdari@seeu.edu.mk; xh.zenuni@seeu.edu.mk;

Abstract - **Cloud computing technology nowadays is a leading IT technology and introduces the next generation of distributed systems with a great acceptance and impact in the businesses. It influences the businesses on structural and operational processes with potential to reduce total costs and increase revenues. This technology is mainly service oriented and focuses on cost reduction, like hardware reduction, reduction on maintenance and total cost of ownership, and with different pricing models such as pay-per-use. Virtualization is the foundation process and technology in Cloud computing, as it hides the complexity of underlying hardware and software resources. It is the objective of virtualization to achieve an improvement in data security aspects, to provide flexible operations and data transfer, elimination (masking) of system failures and to ensure economic benefits. The aim of this research is through a systematic literature review of the existing work to provide a comprehensive understanding of the role and significance of virtualization within the domain of Cloud Computing**.

Keywords - Virtualization, Cloud Computing, Network Virtualization, Software Virtualization, Server Virtualization, Storage Virtualization

I. INTRODUCTION

Cloud computing technology is considered as a new generation in the field of IT. It is based on a Service Oriented Architecture [1] and Service Level Agreement [2]. Different definitions are given for cloud computing [3], considering it as internet-based computing whereby shared resources, software, storage and information are provided to client on demand. Three service models that are offered by Cloud Computing are: Infrastructure as a Service [4], Platform as a Service [5] and Software as a Service [6]. One of the most important topics in this domain is virtualization [7]. Virtualization is considered as the key technology in providing virtual platform of servers, operating systems and hardware devices such as storage. This platform helps the user by providing multiple machines at the same time. Also, it allows sharing a single physical instance of resource or an application to multiple users. Through cloud virtualizations is managed the workload and make it more scalable, economical and efficient. The environment in which virtualization is built can be private or public [8]. One of the features of virtualization is that it allows sharing of resources to multiple clients and companies. Thus, the client can maximize the resources and reduces the physical systems. Types of virtualization in Cloud Computing [9] are: Network Virtualization that combines the resources of network and its functionality in a single virtual network; Server Virtualization, where a single physical server is separated into multiple servers on demand; Software Virtualization, where a single physical server runs one or more instances of the same software and usually is implemented by use of virtual machines; Storage Virtualization which is considered as a sharing of physical storage into multiple storage devices. In this paper, a systematic literature review of the existing literature is presented with the aim of understanding of the role, the significance and the importance of virtualization in Cloud Computing. The rest of the paper is structured as follows: Section 2 presents a brief discussion on the related works. Section 3 presents the Methodology. Section 4 is focused on the literature review. Section 5 presents the discussion related to the overall results observed in the reviewed literature. And some conclusions are listed in the section 6.

II.RELATED WORK

Virtualization of resources is considered as a key element in Cloud Computing technology. It is the objective of virtualization to achieve an improvement in data security aspects, to provide flexible operations and data transfer, elimination (masking) of system failures and to ensure economic benefits. This section presents different virtualization technologies. One of the efforts in this domain is the network virtualization. The NetVM framework [10] is a tool of network virtualization by enabling high bandwidth network functions to operate at near line speed, while taking advantage of the flexibility and customization of low-cost commodity servers. GLANF (Glasgow Network Functions) framework [11] is another technology tool in network virtualization based on container which creates, deploy sand manages virtual network functions (NFs) to achieve low overhead, fast deployment and high reusability. Many studies are focused on the development of programmable networks through Software Defined Networking (SDN) [12]. OpenStack contains virtual network framework [13] based on SDN architecture in providing and managing the network environment and virtual network infrastructure. The improvement of the SDN is continuously proposed [14], as to minimize the response

time, a dynamic controller assignment schema is proposed. In [15], the authors with the use of OpenFlow architecture into SDN have achieved a significant improvement in terms of SFC link throughput and packet loss. The full-stack SDN framework including all protocol layers of the network stack is considered as the next SDN generation, and in the [16] this idea is explained. Andromena is another network virtualization platform designed to improve agility, availability, isolation, and scalability of network [17]. This is a technology developed for Google Cloud Platform (GCP). Recently a new paradigm of Coud Computing, namely the serverless computing, is supported by Network function virtualization (NFV) technology and the research of this is found in papers [18], [19]. To scale cloud resources, it should be forecast the NFV workload, and a method called Wavelet-GMDH-ELM (WGE) for NFV workload forecasting in the paper [20] is proposed and the method predicts and ensembles workload in different time-frequency scales. Another important resource virtualization in Cloud Computing is storage virtualization and basic introduction of this is found in [21]. One way of the storage virtualization, is the so called elastic memory management, described in [22] where this method enables dynamic memory allocation over provisioning and controlling application performance and penalty. This adjustment is beneficial for the resource provider, since the host can take advantage of that memory reduction in order to dedicate it for other processes. In [23], a Heterogeneous Cloud Storage Platform is proposed with a variety of storage technologies and uniform distribution of data stored. This platform is supported by SDN architecture, where the files and parts of file are distributed on different storage servers. In [24] the MDev-NVMe as a new virtualization implementation for NVMe (non-volatile memory express) storage device is proposed and where high throughput, low latency performance and a good device scalability could be achieved. In order to implement huge data storage with high performance, NVMe over remote processor messaging (NVMe-over-RPMsg) is applied to heterogeneous Multi-Core SoCs. The software framework emulates remote storage system as local NVMe device without requiring VMs [25]. The increase of "big data" on cloud computing is another issue [26] and different technologies of storing big data are explained [27]. In [28] the authors have used a network storage device such as a NAS or a SAN to configure it to a large-capacity storage device, in a network environment. NDAS storage systems as well are used to configure a large-capacity storage device in a network environment [29]. In this work, it is shown that the increase of the number of disks into the NDAS system proportionally increases the performance. Virtualization of servers [30] and OS has a very important role in cloud computing. Through server virtualization the resources are more efficiently used, and from user point of view it eliminates the cost of hardware. In Cloud Computing, by use of this virtualization, the workload is balanced and is divided into multiple servers where sometimes virtual servers perform a dedicated task. In [31], the performance analysis of the server virtualization is done and as conclusion the authors have noted that a key element in system performance is the virtualization technology used, notably in case of big data applications. There are two virtualization technologies [32], virtualization technologies based on hypervisor and virtualization technologies based on container. In [33], it is referred that the Docker and LXC virtualization technologies based on container technology introduce lower power consumption compare to KVM and Xen, based on hypervisor virtualization. Also, in the Linux operating system it turns out that Docker is faster than KVM compared by the analysis performed in [34]. Another very important property which needs to be satisfied is the elasticity. Elasticity is one of the main cloud performance and measuring parameter. It measures the ability of a cloud to give fast response to consumer according to their need and demand. In [35], an optimize elasticity framework, based on container technology is proposed. The integration of Docker with OpenStack is another framework that implements a container management service called Yun [36]. Yun improves the container deployment and throughput as well as the system performance by optimizing the message transmission architecture between internal components. Another virtualization framework is the Docker swarm and Kubernetes technology [37]. This technology distributes the work on different nodes by avoid overloaded. In [38], an agent, called Elixir, is proposed. It supports the Docker Swarm to balance the workload among swarm nodes and automatic scaling-up of worker nodes. The issue of the overhead is analyzed in the paper [39] and the authors have proposed an adoption of the OverlayFS to reduce the overhead.

III. METHODOLOGY

This section describes the methodology we follow to design and conduct this research paper. We first describe the research procedure and the research focus. Then we present the metrics on which will be based the analysis and comparison of existing technologies.

I. Research Focus and Procedure

This paper provides a systematic literature review procedure as described by Kitchenham [40] and Brereton [41] on approaches and techniques of resource virtualization in cloud computing. The main goal of this study is to understand the role and significance of resource virtualization in Cloud Computing by considering its benefits. This study introduces the most significant and relevant techniques about resource virtualization that may provide some direction of how the architecture of Cloud environments should be designed. Because of the investigated subject is very extensive, it is impossible to describe all relevant topics. Hence, some related subjects will be briefly mentioned. Based on the searches obtained through Google Scholar, IEEE Xplore, ACM, Springer and Elsevier database we have selected

the most relevant and cited articles that describe technologies about virtualization of resources. Referring to keywords, title and content of articles, we have divided them into three section as follows:

A. Different technologies in network virtualization
B. Different technologies in server and OS virtualization
C. Different technologies in storage virtualization

I. Metrics

The metrics on which will be conducted the analysis and comparison of existing technologies, are:
- Performance of the framework
- Flexibility of management operations
- Cost-Effective Strategy
- Smother IT Operations
- Security and system failure

Referring to the advantages and disadvantages of each technology we give some recommendations that which techniques are most relevant to design the architecture of Cloud environments in the future.

IV. VIRTUALIZATION ON CLOUD COMPUTING

A. Different technologies in network virtualization

There are a lot of publications that study network virtualization in cloud computing. One of the most cited publication in this field is "Network Virtualization and Software Defined Networking for Cloud Computing: A Survey" performed by Raj Jain and Subharthi Paul [12]. They have introduced the virtualization of layer 2 and layer 3 of network components by presented a software defined networking (SDN) which makes a special contribution to network virtualization. The most important innovation of the SDN is programmable control plan, which splits the network into several virtual networks with different policies between them and with share of a common hardware infrastructure. As shown in Fig.1, the SDN consists of a centralized control plan with an API for communication with hardware and network applications.

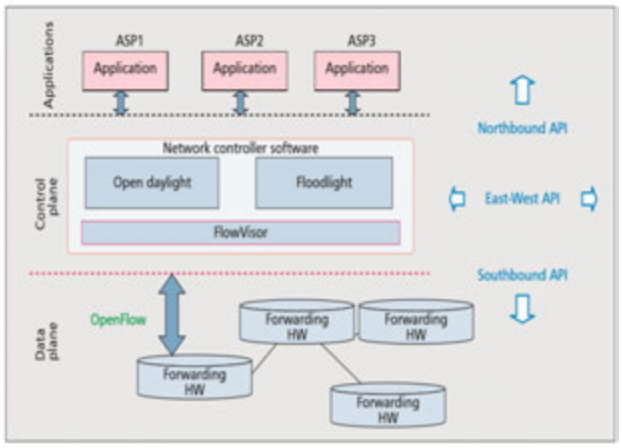


Figure 1. Software defined networking API [12].

Through SDN the networks can be made programmable, partitionable and virtualizable. A critical component for virtualizing SDN networks is hypervisor that abstracts the underlying physical SDN network into multiple logically isolated virtual SDN networks. In [42], a comprehensive understanding of network virtualization hypervisors for SDN is presented. The hypervisor is responsible to ensure that each user is independent from the other users operating a vSDN on the same underlying physical SDN network. In [43] an SDN controller platform called Meridian that supports a service-level model for application networking in clouds is presented. The Meridian platform architecture design is organized in three main logical layers: network model and APIs, network orchestration, and interfaces to underlying network devices as is shown in Fig.2.

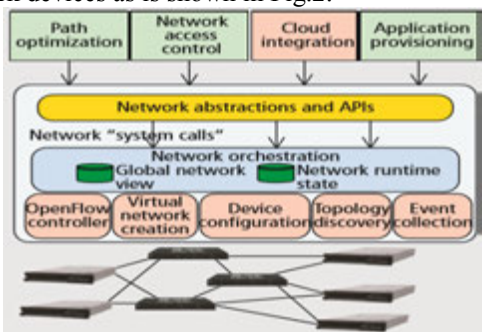


Figure 2. Meridian SDN cloud networking platform architecture [43]

The goal of the API layer is to present applications with a network model and associated APIs that expose only the information needed to interact with the network. The network orchestration layer plays three important roles in the SDN architecture. First, it performs a logical-to-physical translation of commands issued through the abstraction layer above. Second, it provides a set of network-wide services to applications. Third, it provides coordination and arbitration between network requests issued by applications, and mapping of those requests onto the network. And the lowest layer enables the controller to interface with various network technologies or tools. Through this platform can be dynamically update virtual network entities efficiently, and to scalable orchestration of network control as well as configuration tasks. Dmitry Drutskoy , Eric Keller&Jennifer Rexford in [44], introduce a FlowN framework to scalable network virtualization in SDN. This framework is based in two key concepts. First, it lets users to write arbitrary controller software that has full control over the address space and can target an arbitrary virtual topology. It is used a shared controller platform rather than running a separate controller for each user in analogy to container based virtualization. Second, it is used a modern database technology to perform the mapping between the virtual and physical address space. This framework gives each user the own address space, topology, controller, and leverages database technology to efficiently store and manipulate mappings between virtual networks and physical switches. In [45] a SDN/NFV management and orchestration framework is presented. This framework provides a multi-domain network orchestration and virtualization mechanism to offer dynamic and flexible end-to end connectivity and virtual network provisioning services across multi-domain and multi-technology

networks. Also, it performs an integrated orchestration of distributed cloud resources and network resources to dynamically deploy virtual machines and VNF instances and provide the required network connectivity between DCs and end-points across heterogeneous multi-domain transport networks. In [46] a SDN integrated with NFV architecture is presented. It consists a control module, forwarding devices and NFV platform at the edge of the network as is shown in Fig.3.

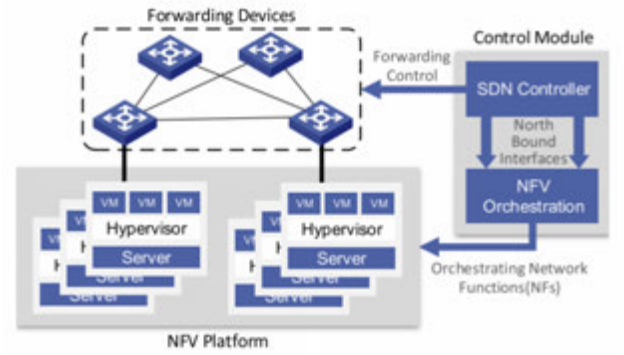


Figure 3. Software-defined NFV system [46]

The packet forwarding is done by the SDN controller according to logical forwarding tables. The Openflow protocol can be used as standardized interfaces in communicating between the SDN controller and distributed forwarding devices. Hypervisors run on the servers to support the VMs that implement the NFs. The SDN controller and the NFV compose the logical control module. The NFV is responsible for virtualized network functions and is controlled by the SDN controller through standard interfaces. While control module computes the optimal function assignments and chooses optimized routing paths. This framework promises agile traffic steering and joint optimization of network functions and resources. In [47], a network virtualization platform (NVP) is proposed in order to provide connectivity on demand. The NVP's architecture is shown in Fig.4.

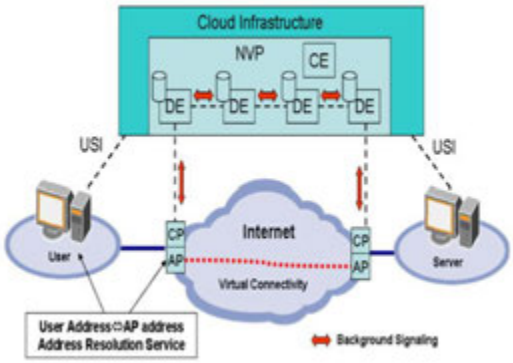


Figure.4. Network Virtualization Platform (NVP) architecture [47].

The NVP's design is based on requirements and service oriented network architecture model [48, 49]. As shown in Fig. 4, the NVP is composed by a set of distributed entities (DEs) and a centralized entity (CE). Clients request NaaS by sending messages to the NVP via the user - service interface (USI). After this, the DE maps it in the parameters characterizing the CP connectivity. The CE is a database with the client' data and relevant service level agreement (SLA) for the end user authentication and relevant network service authorization. In this way NVP is placed to the cloud virtualization platform and it accepts on-demand NaaS requests from the client through a dedicated interface. Through this platform cloud user can request on-demand connectivity without any knowledge of the complexity and the network technology. A highly cited study in this domain, is conducted by Omar Sefraoui, Mohammed Aissaoui&Mohsine Eleuldj [50]. In this study an OpenStack solution framework is proposed. OpenStack architecture is built based on three main components: OpenStack Compute, Image and Object as is shown in Fig.5.

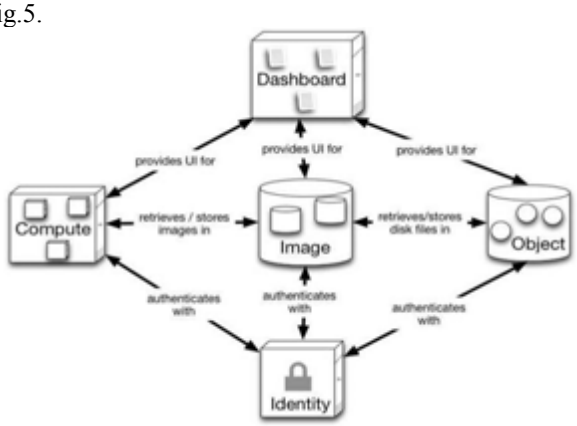


Figure 5. OpenStackArchitecture [50]

OpenStack Compute is a management platform that controls the infrastructure to control IaaS clouds. It allows managing large networks of virtual machines and redundant and scalable architectures. It provides an administrative interface and an API needed for the orchestration of the Cloud. Also, it includes instances management for servers, networks and access control. Imaging Service provides storage services, recording and distributing the images to virtual machine disks. While Object Storage is used to create a storage space redundant and scalable for storing multiple data. This framework promises a scalable, compatible, flexible and open source IaaS. Another study in this domain conducted by Franco Callegati [13]. In this study, a network virtualization architecture based on OpenStack is proposed. The OpenStack [51] provides cloud managers to control a cluster of hosting servers and to manage the virtual network infrastructures. The OpenStack cluster is composed by: a controller node that manages the cloud platform; a network node that host cloud networking services; a number of compute nodes that execute VMs and a number of storage nodes for data and VM images. To manage service networks, a separate component called Neutron which provides a flexible interface for virtual network management is used. An OpenStack user can create a new VM instance specifying the subnet that is assigned via DHCP. In Fig.6 is presented a virtual network infrastructure implemented by OpenStack. As shown in fig.6, a network node and a compute node are running two VMs

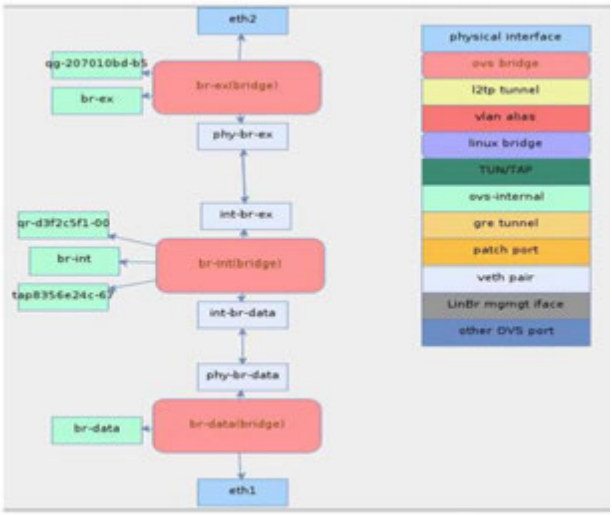


Figure 6. Network elements within an OpenStack network node [13].

The network node includes br-data for the data network and br-ex for the external network. While the compute node includes only br-data connected to the data network. Layer 2 virtualization is implemented by VLANs or Virtual eXtensible LAN (VXLAN) that is supported by the programmable features available in OVS through the insertion of appropriate OpenFlow mapping rules in br-int and br-data. When a new VM instance is created, the OpenStack chooses the best compute node that will host it. In [52], a cloud management platform based on OpenStack is proposed to optimize power consumption, host resources and networking. The main issue discussed in this paper is the problem of VM consolidation in cloud scenarios. We summarize briefly the simple use case of a new VM instantiation in order to understand how the different OpenStack services interact. In the first step each compute service periodically sends updates to a SQL server that stores load in formations about provisioned VMs and physical servers. When a new VM is requested the API Service sends a request to the scheduler service which chooses the compute service in which the VM should be launched. After, it sends a new VM instantiation request to the chosen compute node that requests network configuration parameters for the new VM to the network service. Based on experimental results is concluded that VM consolidation based on OpenStack is an extremely feasible solution to reduce power consumption but it has to be carefully guided to prevent excessive performance degradation. Another open source solution for virtual infrastructure management in cloud is OpenNebula [53]. OpenNebula is an open source virtual infrastructure manager that can be used to deploy virtualized services on both a local pool of resources and on external IaaS clouds. To manage a VM, the OpenNebula Core orchestrates three different management areas, namely: image and storage technologies to prepare disk images for VMs; the network fabric to provide the VMs with a virtual network environment; and the underlying hypervisors to create and control VMs. In [54], a comparison between OpenStack and OpenNebula is presented. It is concluded that both frameworks have their advantages and disadvantages depending on the application.

B. Different technologies in server and OS virtualization.

There are three types of server virtualization in cloud computing; para-virtualization, hypervisor and full-virtualization. In [55], is compared and evaluated the system performance of three techniques. The Xen-PV is used for para virtualization, Open VZ for container virtualization and XenServer for full virtualization. Based on various tests conducted by the benchmarking tools, the Xen Server used for full virtualization has comparatively higher system performance in terms of file copy, pipe based context switching, process creation, shell scripts and floating point operations than the Xen-PV and Open VZ. One of the most cited publication in this domain is the paper "Container-based Operating System Virtualization: A Scalable, High-performance Alternative to Hypervisors" [56]. In this paper a Container-based OS virtualization framework is proposed. This framework provides a shared, virtualized OS image consisting of a root file system, a set of system libraries and executables. Resources are assigned to each VM when it is created, but can be dynamically varied at run time.Fig.7 shows the design architecture.

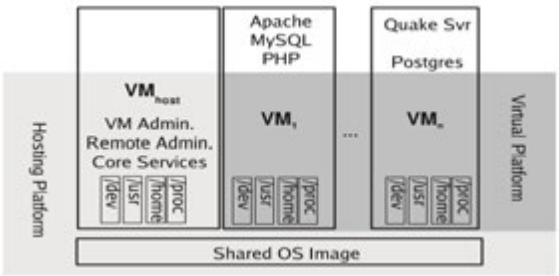


Figure.7 Container –based architecture [56]

As shown in the fig.7, there are three basic components. The hosting platform consists essentially of the shared OS image and a privileged host VM. The virtual platform is part of system seen by the guest VMs. Applications running in the guest VMs work as a corresponding non-container-based OS image. At this point there is little difference between a container and hypervisor based system. However, they differ fundamentally in the techniques they use to implement isolation between VMs. In this study, Linux-VServer technology was considered, comparing it with other technologies based on hypervisor. Based on experiments results container-based systems provide high performance compare to hypervisor based systems for server-type workloads and scale further while preserving performance. Also container based systems such as VServer include more features and compete strongly against hypervisor systems as Xen for various scenarios.
Another highly cited publication in this domain is "An introduction to Docker for reproducible research" [57]. In this study Docker technology is presented which combines several areas such as operating system virtualization, cross-platform portability, modular re-

usable elements, versioning, and a 'DevOps' philosophy. The Docker architecture [58] consists of four main elements; Docker client and server, Docker Images, registries and Docker containers as is shown in Fig.8.

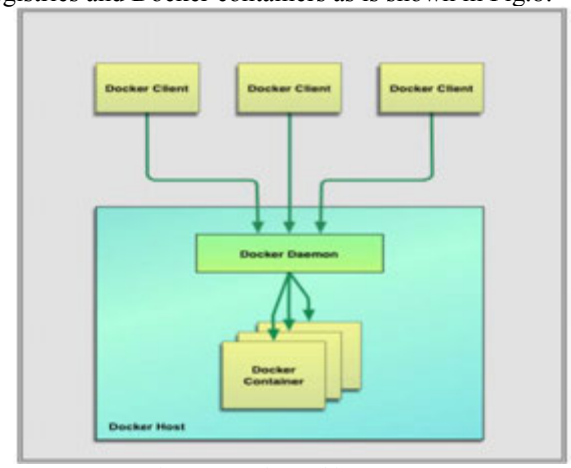


Figure 8. Docker architecture [58]

Docker is as a client-server application. The Docker server (daemon) and client can be run on the same host or local Docker client through a remote daemon run on another host. Images are the building blocks of the Docker. The images can be considered as the "source code" for containers. They are highly portable and can be shared, stored, and updated. Registers serves to store the Docker images. While a Docker container is considered as an image format, a set of standard operations and an execution environment. In [34], four different virtualization technologies are compared: KVM and Xen, which are based on hypervisor virtualization and Docker and LXC which are based on container virtualization. Hypervisor-based virtualization technology operates at the hardware level, while container virtualization operates at the OS level. Fig.9 (a) present the hypervisor-based virtualization architecture and fig.9(b) present the container - based virtualization architecture.

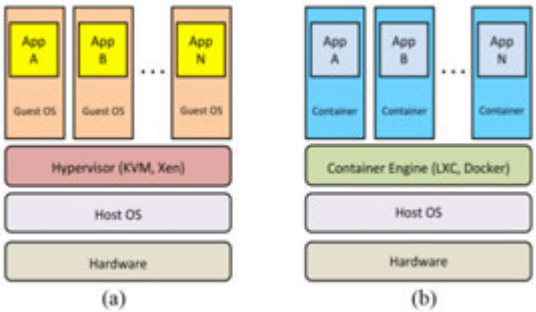


Figure 9. Virtualization Architecture: (a) hypervisor-based; (b) container-based [34].

The container technology allows virtualization of the larger number of instances and the disk image is smaller. Two paradigms of container - based, Docker and LXC are introduced. The performance of each virtualization technology is evaluated in terms of idle state performance, CPU, memory performance and network performance. Based on experiments, the performance of idle state for both technologies, hypervisor and container employ an optimized use of the Linux power saving system. Performance of CPU and Memory show that there is not significant change in power consumption regardless heavy CPU loads and the number of virtual runs. While the network performance shows different results through the different case studies. This is mainly due to the fact that the network packets have to be processed by more layers in a hypervisor environment compare to a container environment. An important study in this direction is presented in [32]. In this paper, the performance of different Apache Spark applications, using both Virtual Machines (VM) and Docker containers comparing for a big data enterprise cloud environment are investigated. Containers and VMs have similar resource isolation benefits but different architecture approach and resource management. Containers perform shared resource management to enable more flexible sharing of resources, while VMs perform distributed resource management where each virtual machine runs with its specific assigned set of resources. The differences between them are summarized in Table 1.

TABLE I. DIFFERENCE BETWEEN VM HYPERVISOR AND DOCKER CONTAINER

Specs	Virtual Machine	Container
Products	VMware, Xen, KVM, etc.	Dockers, rkt, etc.
Guest OS	Included	Not included
virtualization Controller	VM hypervisor	Docker engine
Resource Management	Distributed	Shared
Machine Emulation	Complete (isolated OS kernel)	Partial (shared OS kernel)
Spatial Size (Bytes)	Large	Small
Launch Time	Long	Short
Migration	May require image conversion	Easy

To compare those two technologies, a number of evaluation are conducted. One of the comparison of VM and Docker is by use of the Spark and different implementations run on those, based on different benchmarks. First, it is evaluated the total execution time for both implementations taking into consideration different number of iterations and different number of clusters. And at the end, the utilization of different system resources such as the CPU, disk and memory is investigated. Based on the results, it is concluded that the use of Spark within Docker is better for map and calculation intensive applications. Another highly cited publication in this domain is "Containers and Cloud: From LXC to Docker to Kubernetes" [59]. In this paper is presented an overcome of container-based technology from LXC to Docker to Kubernetes. The innovation of this study is movement around Kubernetes. Kubernetes is considered as the decoupling of application containers from the details of the systems on which they run. This decoupling simplifies application development also simplifies data center operations. In Kubernetes each group of containers deserves its own unique IP address that is reachable from any other group in the cluster, whether they are located on the same physical machine or not. In [60], performance evaluation of leading micro-hosting virtualization approaches are presented with specific focus on Docker and Flockport and their comparison with native platforms. Docker compared to other includes several other capabilities such as NAT, which helps to reduce some of the difficulties of Docker container utilization. On the other hands this directly impact input on throughput quality. As a results Docker

can be no quicker compared to Flockport in some cases. In this way the design of one IP address for each Docker container as suggested by the Kubernetes may support quality assurance as well as throughput. These results may provide some direction of how the architecture of cloud environments should be designed.

C. Different technologies in storage virtualization.

About storage virtualization in cloud computing various technologies are proposed. Each of them has its own advantages and disadvantages. In [61], a Software Defined Storage (SDStorage) framework is proposed. SDStorage is designed to handle all the challenges facing traditional storage systems. In large data storage as data centers, that store a huge amount of data it exploits virtualization to deploy the system. The virtualization is used to enhance the utilization of the system resource, also increase the management overhead across the multiple layers produced when the system is virtualized. SDStorage manage huge data in storage systems by isolating the data control layer from the data storage layer. The control layer refers to the software component that manages and controls the storage resources, whereas the data layer refers to the underlying infrastructure of the storage assets. Fig.10 shows the architecture of the SDStorage system.

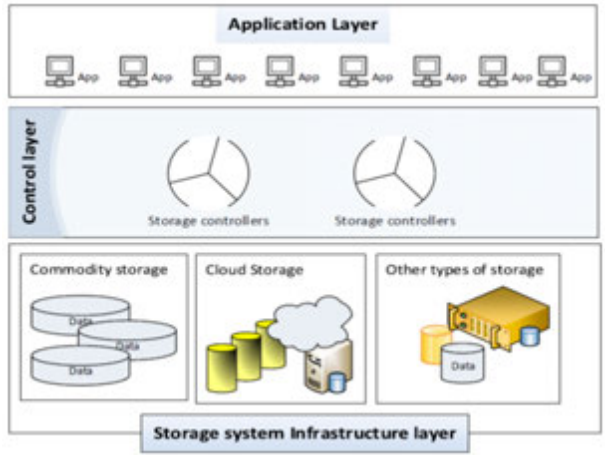


Figure 10. The Software Defined Storage system architecture [61]

The architecture is composed by three layers: infrastructure layer which combines various storage devices storing the raw data, control layer which converts different policies to different instructions inside the system and application layer that allows the end user to interact with storage devices as is shown in fig.10.Another framework in this domain is presented by German Molto [23]. He presents the impact of vertical elasticity for applications with dynamic memory requirements when running on a virtualized environment through the proposal of an architecture that dynamically adapt the memory size of the VM. Vertical elasticity has the ability to rapidly modify the capabilities of the single VMs, in terms of CPU and RAM. Fig.11 shows the architecture of this framework.

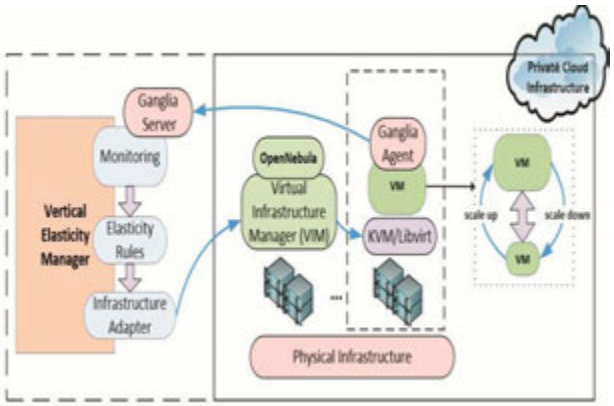


Figure 11. Proposed implementation to enable elastic memory management [23]

This architecture is supported by the KVM [62], which is basically a kernel driver inside the guest OS that behaves as a not-swappable process and can be expand or shrink its memory usage, controlled by the hypervisor on the host OS. If the memory size of a running VM expands the physical memory available in the VM is reduced, that compels the guest OS to reduce the memory of other processes where insufficient free memory is detected. Then, the memory allocated by the memory size of a running VM in the guest OS can be reclaimed by the host OS, which can be possibly used by other VMs. The VMs are deployed via OpenNebula tool [63] on a private cloud infrastructure that are based on a Virtual Machine Image with Ubuntu 12.04 LTS, which kernel supports memory KVM techniques when run as a guest OS. By use of Ganglia Monitoring System [64] the memory usage, in each VM is periodically monitored, in order to report back to the Ganglia' server the periodic state of the VM memory used. The VEM (Vertical Elasticity Manager) includes the implementation of the elasticity rules, to maintaining a user-defined percentage of free memory on the VM. The elasticity rules enable the system to decide when to scale up or scale down, delegating on the Infrastructure Adapter which interacts with the VIM (Virtual Infrastructure Manager) to effectively modify the VM memory size. The VEM can be executed as part of the private Cloud infrastructure or as an external component to the Cloud, which require only inbound connectivity from the Ganglia Agent and outbound connectivity to the VIM. The VEM can simultaneously operate with several VMs to independently scale up and down. To evaluate these techniques, a case study on a private Cloud infrastructure is executed [23]. Based on the results of some tests, in different benchmarks, it is possible to adapt the VM memory size in order to reduce the memory consumption of the VM while maintaining the level of performance of the running application. The vertical elasticity, in the shape of dynamic memory management, enables to mitigate memory over provisioning with controlled application performance penalty. Another study in this domain is presented by Li Zhang & Bing Tang [65]. In this paper, a platform called GrandStore which offers free storage space for individual users, while the free space is quite limited is presented. Through this system each user can manage all their accounts in a unique access interface integrating a plenty

of free accounts to get a unified large-scale free personal cloud storage. This platform is designed and implemented based on the principle of OAuth protocol [66] and open API which manage and control the account applied by user where authentication credentials are stored in back-end database of GrandStore, which realizes easily enlarging personal free storage space and managing all storage spaces in a unique access entry. GrandStore is a scalable and open platform that adds dynamically new accounts without disturbing it. Also, it can store a plenty of accounts to obtain large space. Fig.12 shows the architecture of this platform.

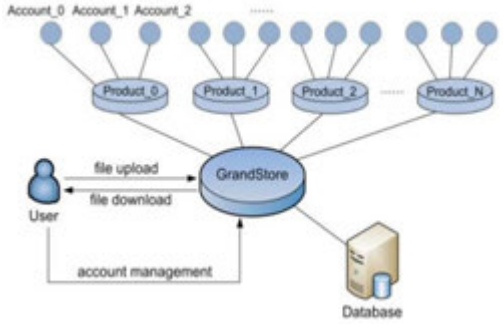


Figure 12. The architecture of GrandStore system [65].

Users firstly open a few accounts, integrate them by GrandStore and then utilize the unique storage space through it. Also, they can manage, integrate, and maintain all their own accounts through GrandStore. This system consists of four components:

- Product. It means free personal cloud storage provider, which provides SDK for developers, such as Google Drive, Dropbox and Kuaipan.
- Account. Each user can register many accounts for one product. Each account has a limited storage space.
- User. User is responsible for adding dynamically new products and new accounts and has the right of all kinds of file operations.
- Database. It is used to store account authentication credential and user's file information.

This study promises a system with a great practical value to get a unified large scale free personal cloud storage.

Another study presented by Amol Jaikar [29], show two technologies, NAS and SAN that can be used for decentralized storage system. Fig.13 shows the Network Attached Storage (NAS) architecture. NAS is composed by a head computer and storage which is connected to a network. This storage contains one or more hard disks with less featured operating systems such as FreeNAS, FreeBSD and others. To increase the performance, multiple copies of files are stored on multiple locations.

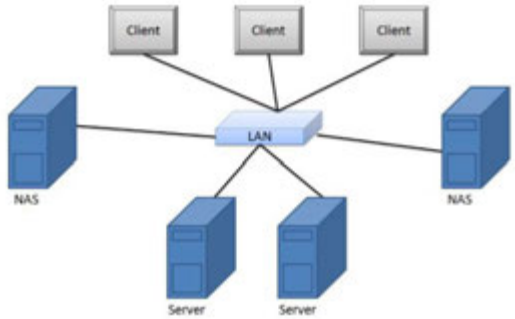


Figure 13. NAS Architecture [29].

In Fig.14 the Storage Area Network (SAN) architecture is presented. The SAN provides storage system with block-level storage. Users also can install file system on the top of SAN to provide file-level access. SAN is normally connected with fiber channel to improve the performance and speed. File sharing in SAN depends on the operating system. The data is accessed by block which is replicated to multiple locations to increase the availability.

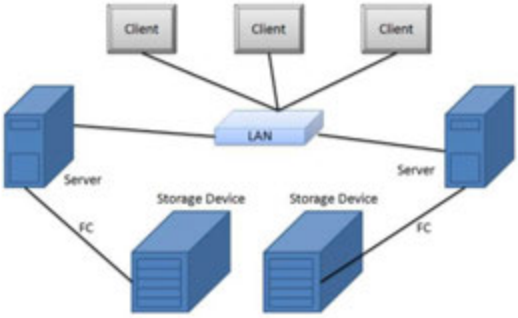


Figure 14. SAN Architecture [29].

Another framework is presented by J. Saravanan & Saravanan. P. in [67]. Direct Attached Storage (DAS) in which storage drives are directly attached to server machine. Fig.15 shows the DAS architecture. One of the benefits of DAS is high performance because of network connectivity and congestion problems that have no immediate effect. Also, DAS is easy to set up and configure as well as has a low price.

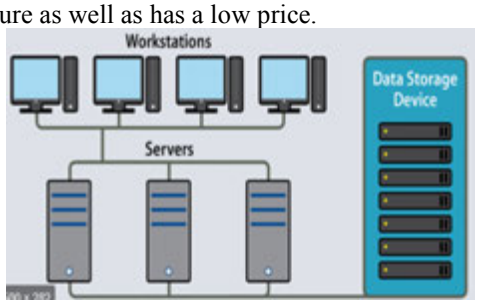


Figure 15. DAS Architecture [67].

V. DISCUSSION

This paper presents a comprehensive understanding about resources virtualization in cloud computing. It is mainly focused on the analysis of the recent technologies about this area. The analysis is focused on the metrics of each technology highlighting the advantages and disadvantages of them. Also, it is analyzed the behavior of different technologies in different applications highlighting the strengths and weaknesses for each application.

A. Different technologies in network virtualization

Network virtualization improves scalability, data transfer rates, security, versatility and reliability in a network. The proposed approach by Fabio Baroncelli [47] presents an abstraction of NaaS that enables cloud users to request on-demand connectivity without any knowledge of network details, just through Network Virtualization

Platform (NVP). Based on the approach, proposed by Raj Jain [12], through SDN the network can be done programmable, partitionable and virtualizable. The network virtualization architecture based on OpenStack [13], provides a cloud manager to control a cluster of hosting servers and to manage the virtual network infrastructures. The performance of network virtualization in this case increases significantly. By analyzing and comparing different technologies about network virtualization we can say that SDN is considered the next generation in this domain. Integration of OpenFlow and OpenStack architecture in SDN and consideration of the full-stack SDN framework that include all protocol layers in the network stack, the network performance can be improved in all aspects. Integrating these frameworks into a single framework are considered the current and future challenge about network virtualization

B. Different technologies in server and OS virtualization

Server virtualization is considered as a separation of physical server into multiple isolated virtual servers. Operating system virtualization is considered as a type of server virtualization. The paper presented by Janki Bhimani & Zhengyu Yang [31] investigates the performance of big data processing applications running on Spark and with different virtualization frameworks. They conclude that the use of Spark under Docker, in case of map and intensive calculation applications, performs better in comparison with Spark under VM. This is due that Docker provides lightweight operation, copy-on-write and intermediate storage drivers that assist Spark application to perform better. Spark used on Docker reaches speed-up of 10 times higher compared to VM. Another study is presented by Roberto Morabito [33], he compares the performance in terms of power consumption of four different virtualization technologies: KVM, Xen, Docker and LXC. Based on the experimental results, is confirmed a similar power consumption, when systems are challenged with a heavy CPU and memory workloads. As far as power consumption is concerned, both container-based solutions introduce lower power consumption. In [59], is presented the Kubernetes framework that is considered as the decoupling of application containers from the details of the systems on which they run in order to simplify application development also simplifies data center operations. By analyzing and comparing different technologies about server virtualization, based on the advantages and disadvantages of them we conclude that Docker is the more efficient technology in server virtualization, and it is considered as the future of them. The Docker improve the performance, reduce overhead, is faster than other technologies by balancing of the workload.

C. Different technologies in storage virtualization

Storage virtualization in Cloud Computing is considered as sharing of physical storage into multiple logical storage devices and such logical device looks like a single physical storage device. A technique for large-scale free personal cloud storage is presented by Li Zhang in his paper [49]. He has proposed a system called GrandStore which offers free storage space for individual users. Users apply for several accounts, and then the account authentication's credentials are stored in the GrandStore back-end database, and the system allows managing of all storage spaces in a unique access entry. Another result is presented by German Molt & Miguel Caballer in their paper [22]. They used an approach of adapting the VM memory size in order to reduce the memory consumption while maintaining the level of running application performances. The vertical elasticity, in the shape of dynamic memory management, enables to mitigate memory over provisioning with controlled application performance penalty. This adjustment is beneficial for the resource provider, since the host can take advantage of that memory reduction in order to dedicate it for other processes. In domain of storage virtualization we found out three types of data storage:

1. The NAS, described in [27] is used for file sharing, backup store memory space, and sharing device among the different machines. To increase the performance, multiple copies of files are stored on multiple locations.
2. The SAN, described into paper [28], provides storage system with block-level storage. The data is accessed by block which is replicated to multiple locations to increase the availability.
3. While in the DAS, described into paper [29] storage drives are directly attached to server machine. One of the benefits of DAS is high performance because of network connectivity and congestion problems have no immediate effect. Regarding the above mentioned, about the storage virtualization we conclude that each technology has its advantages and disadvantages and depends on their use. Integration of the advantages of NAS, SAN, DAC and SDN technologies in a heterogeneous Cloud Storage Platform remains one of the challenges of the future in this domain.

VI. CONCLUSION

In this paper, a systematic literature review of existing work to provide a comprehensive understanding of the role and significance of virtualization within the domain of Cloud Computing is presented. The scope of this paper is targeted towards: network virtualization, server virtualization, OS virtualization and storage virtualization. Network virtualization promises cloud users to request on-demand connectivity without any knowledge of the complexity and the network technology details. In this way, flexibility of management operations is increased, and can be created different virtual networks isolated from each other, allowing the share of a common physical infrastructure among multiple customers and reduce the operational costs. Storage virtualization promise sharing of physical storage into multiple storage devices which further appears to be as a single storage device. Through storage virtualization is achieved an easily data movement, data management improvement,

security increase and system failure avoidance. Server virtualization promises a partition of physical servers into multiple isolated virtual servers, where each virtual server runs its own operating system and applications. While OS virtualization is considered as a part of server virtualization which promises to integrate server hardware by moving services on separate servers as well as different computer programs can be run at the same time. In this way, by divided a single server into multiple virtual servers the physical hardware cost and maintenance cost are reduced and at the same time the productivity, security and system failure avoidance, are improved. By analyzing and comparing different technologies about resource virtualization in cloud computing is recommended that integration of OpenFlow and OpenStack architecture in SDN and consideration of the full-stack SDN framework is seen as the best solution in network virtualization and is considered for the next generation of this domain. Regarding to server and operating system virtualization is recommended that the use of Docker technology is seen as the best solution and it is considered for the future use. And regarding to storage virtualization is recommended integration of advantages of NAS, SAN, DAC and SDN technologies in a heterogeneous Cloud Storage Platform and it is considered as one of the big challenges of the future in this domain. However, these technologies remain to be studied in the future by analyzing their performance in different applications and big data as well as in terms of data security.

REFERENCES

[1] Wei-Tek Tsai, Xin Sun, Janaka Balasooriya , “Service-Oriented Cloud Computing Architecture”,Seventh International Conference on Information Technology, IEEE 2010.

[2] Pankesh Patel, Ajith Ranabahu, Amit Sheth,” Service Level Agreement in Cloud Computing” Digital Commons Network,USA 2009.

[3] Luis M. Vaquero, Luis Rodero-Merino, Juan Caceres, Maik Lindner,” A Break in the Clouds: Towards a Cloud Definition” ,ACM Volume 39, Number 1, January 2009.

[4] Sushil Bhardwaj , Leena Jain , Sandeep Jain,” CLOUD COMPUTING: A STUDY OF INFRASTRUCTURE AS A SERVICE (IAAS)”, 2010 International Journal of Engineering and Information Technology.

[5] Ling Qian,Zhiguro Luo,Yujian Du,Leita Guo,” Cloud Computin: An Overview”, IEEE & Springer 2009.

[6] Manish Godse ,Shrikant Mulik,” An Approach for Selecting Software-as-a-Service (SaaS) Product “,2009 IEEE International Conference on Cloud Computing.

[7] Rakesh Kumar , Shilpi Charu,” An Importance of Using Virtualization Technology in Cloud Computing”, Global Journal of Computers & Technology Vol. 1, No. 2, February 25, 2015.

[8] Swarnpreet Singh and Tarun Jangwal,” Cost breakdown of Public Cloud Computing and Private Cloud Computing and Security Issues”, International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 2, April 2012.

[9] Aaqib Rashid, Amit Chaturvedi, “Virtualization and its Role in Cloud Computing Environment”, International Journal of Computer Sciences and Engineering Vol.7(4), Apr 2019.

[10] Jinho Hwang ,K. K. Ramakrishnan, Timothy Wood, “NetVM: High Performance and Flexible Networking using Virtualization on Commodity Platforms”, IEEE Transactions on Network and Service Management (Volume: 12 , Issue: 1 , March 2015).

[11] Richard Cziva, Simon Jouet, Kyle J. S. White and Dimitrios P. Pezaros,” Container-based Network Function Virtualization for Software-Defined Networks”, 2015 IEEE Symposium on Computers and Communication (ISCC).

[12] Raj Jain and Subharthi Paul, “Network Virtualization and Software Defined Networking for Cloud Computing: A Survey”, IEEE Communications Magazine November 2013.

[13] Franco Callegati, Walter Cerroni, Chiara Contoli ,Giuliano Santandrea, “ Performance of Network Virtualization in Cloud Computing Infrastructures: The OpenStack Case”, 2014 IEEE 3rd International Conference on Cloud Networking (CloudNet).

[14] Samaresh Bera ; Sudip Misra ; Niloy Saha,"Traffic-aware Dynamic Controller Assignment in SDN" ,IEEE 2020

[15] Hiep Binh Nguyen ; Ngoc-Thanh Dinh ; Jaewook Oh ; Younghan Kim "An Openflow-based Scheme for Service Chaining’s High Availability in Cloud Network",2019 International Conference on Information and Communication Technology Convergence (ICTC),IEEE.

[16] Gianni Antichi,Gábor Rétvári,"Full-stack SDN: The Next Big Challenge?" ,Proceedings of the Symposium on SDN , Pages 48–54, ACM 2020

[17] Michael Dalton, David Schultz, Jacob Adriaens, Ahsan Arefin, Anshuman Gupta, Brian Fahs, Dima Rubinstein, Enrique Cauich Zermeno, Erik Rubow, James Alexander Docauer, Jesse Alpert, Jing Ai, Jon Olson, Kevin DeCabooter, Marc de Kruijf, Nan Hua, Nathan Lewis, Nikhil Kasinadhuni, Riccardo Crepaldi, Srinivas Krishnan, Subbaiah Venkata, Yossi Richter, Uday Naik, and Amin Vahda,"Andromeda: performance, isolation, and velocity at scale in cloud network virtualization",15th USENIX Conference on Networked Systems Design and Implementation,ACM 2018

[18] Paarijaat Aditya,Istemi Ekin,|Andre Beck,Ruichuann Chen,Volker Hilt,vica Rimac,Klaus Satzkeand and Manuel stein “Will Serverless Computing Revolutionize NFV?”, Proceedings of the IEEE (Volume: 107 , Issue: 4 , April 2019

[19] Deze ZengLin GuShengli PanSong Guo "Software Defined Systems" 2020 – Springer

[20] Sima Jeddi,Saeed Sharifia ,"A hybrid wavelet decomposer and GMDH-ELM ensemble model or Network function virtualization workload forecasting in cloud computing ",Elsevier 2019

[21] Wenying Zeng,Yuelong Zhao ,Kairi Ou,Wei Song ,"Research on cloud storage architecture and key technologies", 2009 ACM.

[22] German Molt,Miguel Caballer, Eloy Romero, Carlos de Alfonso ,“Elastic Memory Management of Virtualized Infrastructures for Applications with Dynamic Memory Requirements”, International Conference on Computational Science, ICCS 2013,Elsevier.

[23] Chao-Tung Yang ; Shuo-Tsung Chen ; Wei-Hsun Cheng ; Yu-Wei Chan ; Endah Kristiani,"A Heterogeneous Cloud Storage Platform With Uniform Data Distribution by Software-Defined Storage Technologies",11 October 2019 IEEE

[24] Bo Peng,Haozhong Zhang,Jianguo Yao,Yaozu Dong,Yu Xu,Haibing Guan "MDev-NVMe: a NVMe storage virtualization solution with mediated pass-through",Proceedings of the 2018 USENIX Conference on Usenix Annual Technical Conference,July 2018 ,Pages 665–676,ACM

[25] Xiaohao Zhang ; Yunjie Li ; Gang Chen,"NVMe-over-RPMsg: A Virtual Storage Device Model Applied to Heterogeneous Multi-Core SoCs", 2020 10th Annual Computing and Communication Workshop and Conference (CCWC),IEEE

[26] Ibrahim Abaker Targio Hashem , Ibrar Yaqoob , Nor Badrul Anuar,Salimah Mokhtar ,Abdullah Gani, Samee Ullah Khan,"The rise of “big data” on cloud computing: Review and open research issues",Elsevier 2014

[27] Chellammal Surianarayanan,Pethuru Raj Chelliah "Storage Fundamentals and Cloud Storage",Springer 2019

[28] Amol Jaikar, Syed Asif Raza Shah, Seo-Young Noh and Sangwook Bae.” Performance Analysis of NAS and SAN Storage for Scientific Workflow”, 2016 International Conference on Platform Technology and Service (PlatCon) IEEE.

[29] Jung Kyu Parka,Jaeho Kim "Big data storage configuration and performance evaluation utilizing NDAS storage systems",Elsevier 2018

[30] Malgorzata Steinder , Ian Whalley , David Carrera , Ilona Gaweda,David Chess,Thomas J. Watson, “Server virtualization in autonomic management of heterogeneous workloads”, : 2007 10th IFIP/IEEE International Symposium on Integrated Network Management.

[31] Janki Bhimani, Zhengyu Yang, Miriam Leeser, and Ningfang Mi ,“Accelerating Big Data Applications Using Lightweight Virtualization Framework on Enterprise Cloud”, 2017 IEEE High Performance Extreme Computing Conference (HPEC).

[32] Zheng Li , Maria Kihl , Qinghua Lu and Jens A. Andersson, “Performance Overhead Comparison between Hypervisor and Container based Virtualization”, 2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA).

[33] Roberto Morabito, “Power Consumption of Virtualization Technologies: An Empirical Investigation” 2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC).

[34] MinSu Chae, HwaMin Lee & Kiyeol Lee "A performance comparison of linux containers and virtual machines using Docker and KVM",Published 20 December 2017 Springer

[35] Noor e Sahir, Muhammad Amir Shahzad, Muhammad Sohaib Aslam,Waseem Sajjad, Muhammad Imran "Optimize Elasticity in Cloud Computing using Container Based Virtualization",International Journal of Innovations in Science & Technology 2020

[36] Sijie Yang, Xiaofeng Wang, Xiaoxue Wang, Lun An & Guizhu Zhang ,"High-performance docker integration scheme based on OpenStack",Springer 2020

[37] Nikhil Marathe ; Ankita Gandhi ; Jaimeel M Shah,"Docker Swarm and Kubernetes in Cloud Computing Environment",2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI),IEEE

[38] Michail S. AlexiouEuripides G. M. Petrakis, "Elixir: An Agent for Supporting Elasticity in Docker Swarm",Springer 2020

[39] Naoki Mizusawa ,Yuya Seki,Jian Tao,Saneyasu Yamaguchi "A Study on I/O Performance in Highly Consolidated Container-Based Virtualized Environment on OverlayFS with Optimized Synchronization",2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM),IEEE

[40] Barbara Kitchenham , O. Pearl Brereton , David Budgen, Mark Turn,er,John Bailey, Stephen Linkman“Systematic literature reviews in software engineering – A systematic literature review” Elsevier 2009.

[41] Pearl Brereton, ,MarkTurner[a] Mohamed Khalil,“ Lessons from applying the systematic literature review process within the software engineering domain”,Elsevier 2007.

[42] Andreas Blenk ; Arsany Basta ; Martin Reisslein ; Wolfgang Kellerer, "Survey on Network Virtualization Hypervisors for Software Defined Networkin",IEEE Communications Surveys & Tutorials (Volume: 18 , Issue: 1 , Firstquarter 2016)

[43] Mohammad Banikazemi ; David Olshefski ; Anees Shaikh ; John Tracey ; Guohui Wang, "Meridian: An SDN Platform for Cloud Network Services", IEEE Communications Magazine (Volume: 51 , Issue: 2 , February 2013)

[44] Dmitry Drutskoy ; Eric Keller ; Jennifer Rexford, "Scalable Network Virtualization in SoftwareDefined Networks",IEEE Internet Computing (Volume: 17 , Issue: 2 , March-April 2013)

[45] Raul Munoz ; Ricard Vilalta ; Ramon Casellas ; Ricardo Martinez ; Thomas Szyrkowiec ; Achim Autenrieth, "Integrated SDN/NFV management and orchestration architecture for dynamic deployment of virtual SDN control instances for virtual tenant networks", IEEE/OSA Journal of Optical Communications and Networking (Volume: 7 , Issue: 11 , November 1 2015)

[46] Yong Li ; Min Chen, "Software-Defined Network Function Virtualization: A Survey", IEEE Access (Volume: 3 ,December 2015)

[47] Fabio Baroncelli & Barbara Martini & Piero Castoldi,” Network virtualization for cloud computing”,Springer 2010.

[48] N. M. Mosharaf Kabir Chowdhury and Raouf Boutaba,” Network Virtualization: State of the Art and Research Challenges”, IEEE Communications Magazine (Volume: 47 , Issue: 7 , July 2009).

[49] Barbara Martini, Valerio Martini, Fabio Baroncelli, Karim Torkman, and Piero Castoldi,” Application-Driven Control of Network Resources in Multiservice Optical Networks”, Journal of Optical Communications and Networking 2009.

[50] Omar SEFRAOUI,Mohammed AISSAOUI, Mohsine ELEULDJ, "OpenStack: Toward an Open-Source Solution for Cloud Computing", International Journal of Computer Applications (0975 - 8887) Volume 55 - No. 03, October 2012

[51] OpenStack project website. Available: http://www.openstack.org,(accessed 12.01.2020)

[52] Antonio Corradi, Mario Fanelli, Luca Foschini, "VM consolidation: A real case based on OpenStack Cloud", Future Generation Computer Systems,Volume 32, March 2014, Pages 118-127 ,Elsevier

[53] Borja Sotomayor, Ruben S. Montero, Ignacio M. Llorente, and Ian Foster, "An Open Source Solution for Virtual Infrastructure Management in Private and Hybrid Clouds", IEEE Internet Computing ,Special Issue on Cloud Computing 2009

[54] Xiaolong Wen ; Genqiang Gu ; Qingchun Li ; Yun Gao ; Xuejie Zhang, "Comparison of open-source cloud management platforms: OpenStack and OpenNebula", 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery,IEEE

[55] Anish Babu S. ; Hareesh M.J. ; John Paul Martin ; Sijo Cherian Rajagiri ; Yedhu Sastri, "System Performance evaluation of Para virtualization, Container virtualization and Full virtualization using Xen, OpenVZ and XenServer", 2014 Fourth International Conference on Advances in Computing and Communications,IEEE

[56] Stephen Soltesz, Herbert Potzl ,Marc E. Fiuczynski, Andy Bavier, and Larry Peterson, "Container-based Operating System Virtualization: A Scalable, High-performance Alternative to Hypervisors", ACM 2007

[57] Carl Boettiger , "An introduction to Docker for reproducible research", ACM SIGOPS Operating Systems ,January 2015

[58] James Turnbull, "The Docker Book: Containerization is the new virtualization", Google book 2014

[59] David Bernstein, "Containers and Cloud: From LXC to Docker to Kubernetes", IEEE Cloud Computing (Volume: 1 , Issue: 3 , Sept. 2014)

[60] Zhanibek Kozhirbayeva ,Richard O.Sinnott, "A performance comparison of container-based technologies for the Cloud",Future Generation Computer Systems Volume 68, March 2017, Pages 175-182,Elsevier

[61] MahmoudAl-Ayyouba, Ala’Darabseha ,Elhadj Benkhelifab, Mladen Voukc, Andy Rindo, "Software defined cloud: Survey, system and evaluation",Future Generation Computer Systems Volume 58, May 2016, Pages 56-74,Elsevier

[62] KVM,KVM,CPU,Hotplug.Available:http://www.linux-kvm.org/page/CPUHotPlug,(accessed 14.01.2020)

[63] Dejan Milojičić Hewlett-Packard Laboratories; Ignacio M. Llorente ; Ruben S. Montero,”OpenNebula: A Cloud Management Tool” , IEEE Internet Computing (Volume: 15 , Issue: 2 , March-April 2011).

[64] Matthew L. Massie , Brent N. Chun , David E. Culler,” The ganglia distributed monitoring system: design, implementation, and experience” ,ELSEVIER Parallel Computing Volume 30, Issue 7, July 2004, Pages 817-840.

[65] Li Zhang, Bing Tang, “GrandStore: Towards Large-Scale Free Personal Cloud Storage”, 2017 International Conference on Computer Network, Electronic and Automation (ICCNEA) & IEEE.

[66] Barry Leiba, “OAuth Web Authorization Protocol”, IEEE Internet Computing (Volume: 16 , Issue: 1 , Jan.-Feb. 2012).

[67] J.Saravanan & Saravanan,” The Storage Process in Virtualization for Cloud Computing”, JAC : A Journal Of Composition Theory, Volume XII Issue XII December 2019.

Algorithms for Computing in Fog Systems: Principles, Algorithms, and Challenges

Nikheel Soni[1,2], Reza Malekian[2,3*], and Dijana Capeska Bogatinoska[4]
[1]Amazon Web Services, Cape Town, 8001, South Africa
[2]Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria, 0002, South Africa
[3]Department of Computer Science and Media Technology, Malmö University, Malmö, 20506, Sweden/ Internet of Things and People Research Center, Malmö University, Malmö, 20506, Sweden
[4]University of Information Science and Technology "St. Paul the Apostle", Ohrid, Macedonia
nikheel@amazon.com, reza.malekian@ieee.org (corresponding author), dijana.c.bogatinoska@uist.edu.mk

***Abstract* - Fog computing is an architecture that is used to distribute resources such as computing, storage, and memory closer to end-user to improve applications and service deployment. The idea behind fog computing is to improve cloud computing and IoT infrastructures by reducing compute power, network bandwidth, and latency as well as storage requirements. This paper presents an overview of what fog computing is, related concepts, algorithms that are present to improve fog computing infrastructure as well as challenges that exist. This paper shows that there is a great advantage of using fog computing to support cloud and IoT systems.**

Keywords *– algorithms, big data, cloud computing, fog computing IoT, latency*

I. INTRODUCTION

Fog computing is an architecture that is used to distribute resources closer to users, at the edge of the network, to improve the overall performance of applications and services [1], [2]. Fog computing is not designed to replace cloud computing, but to complement it by distributing computing, communication, control, and storage elements closer to the end-users [1]. The use of fog computing can be seen in various applications and services that do not work well in the cloud paradigm which include applications that are latency dependent such as gaming, geo-distributed such as pipeline monitoring, fast mobile applications such as connected vehicles and large-scale distribution such as smart grid [2]. Other applications where fog computing can prove to be highly effective include augmented reality, content delivery and caching and analytics on big data [1], [2]. As latency is a topic that is widely discussed, latency reduction does not occur instantly and depends on where application components are placed and is not critical for every application [3]. Environments in which fog computing architecture is implemented need to be thoroughly evaluated as an incorrect assessment can result in an increase in factors such as costs, security risks and processing time.

For instance, if an Internet of Things (IoT) environment is constantly changing and developing, implementation of fog can prove to be beneficial, as it will enable easy integration for any changes made. If an IoT environment does not change at all, the implementation of fog could be redundant. Implementation of fog should be considered when many devices are generating data, there is a high priority in obtaining a result from data that is generated and in areas where there is intermittent access to the cloud [9], Fig.1. Making use of fog computing architecture has numerous advantages. Advantages include: security, cognition, agility, latency, and efficiency [1]. Security is improved as data needs to travel a smaller distance which makes it less susceptible to attacks, fog nodes acting as proxies used to deploy certain security functions [1], keeping sensitive data is contained inside the network [4], act as the first line of access control and traffic encryption and provide data integrity to mention a few [1]. Cognition is created as fog devices are closer to the user, therefore, resulting in a better understanding of the user's requirements for an application [1]. The architecture is agile as it allows for instantaneous scaling concerning both hardware and software instances [1]. By reducing latency, time-constrained applications can be performed much faster to obtain results. Lastly, efficiency is improved through all the mentioned advantages by reducing network traffic to the cloud by only uploading what is required [4] and performing computing tasks closer to the user. As fog nodes are closer to the end-user, users obtain a better-tailored application that suits their needs.

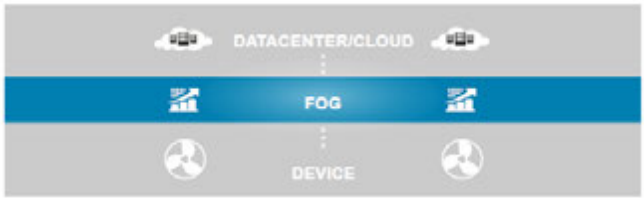


Figure 1. Fog lies between the Cloud and End-User (Taken from [4])

The structure of the paper is as follows. Section 2 describes some of the related concepts to fog computing, section 3 discusses a few of the computational algorithms that are currently in use, section 4 contains the discussion of the algorithms and the paper is concluded in section 5.

II. Related Concepts

The concept of fog computing is not a new one as there have been different relatable technologies that bring resources closer to the end user to improve processing, storage, and speed of applications. These include cyber foraging, cloudlets, and multi-access edge computing.

A. Cyber Foraging

One of the first concepts used for edge computing was cyber foraging [3]. Making mobile devices small and light in weight while being able to perform desktop-level applications was a major challenge that needed to be addressed [5]. By creating a dynamic host of computing power and data staging servers nearby to a user, the needs of a wireless mobile computer can use these resources to perform applications that are limited to the mobile device [5]. Systems that are in place, that help mobile computers in computational and storage tasks, are called surrogates; where these systems are not administrated by an authority and are considered untrusted. For example, a mobile device will detect a surrogate in their operating environment and negotiate with the surrogate for their resources through short-range wireless communication technology and when the mobile device needs to process a relatively large amount of data, the mobile device directs this computation required to the surrogate [5]. For applications such as facial recognition, the mobile device will capture the images, pass it to the surrogate where it will perform the relevant processing algorithms using a database and deliver the result to the mobile device [3]. Information on the surrogate used for the mobile device is removed when the mobile device does not require the surrogate; or is disconnected from the surrogate. Data can also be staged by predicting what the user's needs to reduce cache miss service times [5]. As this is an effective solution to reduce latency and bandwidth constraints for a given user, challenges such as management, trust level and load balancing of surrogate devices need to be addressed.

B. Cloudlet

Improving the concept of cyber foraging, cloudlets are trusted devices that are connected to the Internet which provide a large array of resources available to nearby mobile devices [6]. These are decentralized and dispersed across a large area for mobile computers to use. Cloudlets can be viewed as cloud-based systems on the edge of the network as they reuse cloud-based techniques such as virtual machines and are within a single-hop proximity of mobile devices [3]. Cloudlets act and perform as a scaled-down data center in a box that allows a mobile user to connect via Wi-Fi to create and initiate virtual machine instances [7]. If there are no cloudlet services around that a mobile device requires, the device can resort to using conventional cloud-based systems or its own dedicated hardware and software [6]. Looking again at facial recognition, processing will be performed on virtual machines instead of real machines which allow cloudlets to dynamically change depending on the intensity of processing required [3]. As cloudlets can represent full cloud infrastructures in certain scenarios, they can exist in a standalone environment due to virtual machines deployed without the resources of a conventional cloud [3].

C. Multi-Access Edge Computing (MEC)

Also referred to as mobile edge computing, MEC focuses on mobile networks and virtual machines [3]. Working with real-time Radio Access Network (RAN) information, developers can create applications to suit a user's needs by obtaining data such as the user's location [8]. This information can also be useful to businesses wanted to sell a product or service by learning about a user's interests based on their location. By allowing resources to be located at the edge of the network, within the range of the RAN, the user's latency and bandwidth consumption decrease [8].

D. Similarities and Differences with Fog Computing

All the proposed technologies have the same characteristics in common: to reduce latency, network constraints and provide the user with a product or service instantaneously. Cyber foraging is succeeded by cloudlets in terms of better security and management and can operate as a standalone device or connected to the cloud. Fog computing can operate in both standalone or connect to the cloud to help assist in applications and services. MEC has a fall back here in the case that it does not work with cloud-based systems and only in a standalone environment [3]. The main difference between the technologies is application usage. Cloudlets work by reducing processing and storage limits to a user whereas MEC works with applications that are best suited for edge environment operation [3]. Fog computing can take both characteristics by allowing compute and storage resources to lie both at the edge in a fog node, in the cloud or spanning across the two depending on the application requirements [3].

III. Computing Algorithms

A general outline of how algorithms designed for fog systems have been outlined in [3] discussing points of heterogeneity, QoS management, scalability, mobility, federation, and interoperability. Heterogeneity is how fog components can be diverse in its application such as perform several different functions regardless of the environment placed in with respect to computation and storage needs. QoS management is how fog systems provide improved user experience by reducing latency and processing time. Scalability is how fog computing can be implemented on a great horizontal and vertical spectrum with regards to the number of users to the number of applications. This factor should be dynamic as there may not always be a need for a large number of fog components in low demand situations. Mobility is how fog components can be moved and relocated if necessary while in operating conditions to perform a function. This can be seen with vehicle-to-vehicle communication [9] where vehicles [10] can act as fog nodes used to collect data onboard a vehicle [11] while moving at different speeds and through different operating environments. Federation ties together with heterogeneity in the case that fog systems need to be diverse and be able to be managed by different operators with different requirements

regardless of the fog components in place. Lastly, fog components need to be interoperable with other devices from other infrastructures such as cloud and IoT systems as fog helps improve these.

Fog infrastructure can be viewed from both algorithmic and architectural views. Algorithms for fog can be sub-divided into computing, content storage, and distribution, energy consumption and application-specific algorithms where architecture can be subdivided into application-agnostic architectures and application-specific architectures. As the field of fog computing is a very large one, this paper covers what has been done in the algorithmic environment with respect to computing in fog systems.

By analyzing and pairing nodes in a fog network, system efficiency can be improved by node pairing in the same fog domain [12] by using a utility-based scheme. This is done by refining Irving's matching algorithm and creating a model as a one-sided stable matching game with a quota [12]. A utility-based list is created that considers factors that include transmission power, transmission distance, and cost. A pairing of nodes is performed through the utilities by the creation of preference lists in each node to other nodes in the fog domain. By performing this matching algorithm for fog nodes in the same network, an improvement in performance in the form of full potential was seen [12]. By making use of a utility-based node matching algorithm, efficiency is improved over a greedy algorithm where a node would pair with a neighboring node without evaluation of the situation. By making use of short-range communication methods, local resources can be leveraged to reduce communication and computational loads.

By allowing for radio access clustering in fog

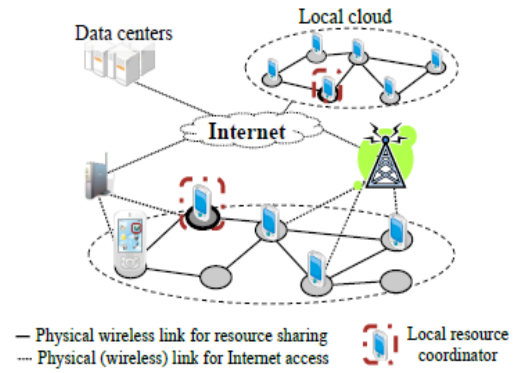


Figure 2. System model used for local resource coordinator (Taken from [14])

computing applications, improved user experience is obtained by reducing power consumption or latency for each user [13]. By creating a small cluster of cells, the computation can be shared in a cluster to improve the efficiency of fog components and reduce network traffic. This is done by distributing the computational load across adjacent radio access points with computational capacities. Two approaches to creating clusters included forming a cluster to perform a task with the lowest latency and the other to reduce the amount of power usage [13]. Clusters contain inside communication and computational resource allocation and are scalable. Trade-offs can be seen by tuning the latency performance of a cluster by scaling the size of the cluster accordingly. This implementation shows an improvement over the algorithm used in [12] as this accommodates heterogeneity as well as QoS by reducing latency.

By mapping particular parameters to time resources, a mathematical framework for the sharing of resources is implemented [14]. The main issue is how to distribute resources between nodes in a fog system. A simple approach is to perform numerical calculations or downloading of data, however, this does not factor in the service that is being delivered. This can result in an increase in factors such as latency for latency intensive applications such as augmented reality. Therefore, by performing task-orientated sharing, the processing is minimized. This is done by creating a unified framework to map resources that include power, bandwidth, and latency to distribute resources. From these parameters, optimization problems are created and solved through convex optimization approaches.

From Fig. 2, the system model is shown where a local resource coordinator handles the resource sharing of the fog domain. Nodes in the fog domain communicate with each other through various short-range communication methods that include Wi-Fi and/or Bluetooth.

The local resource coordinator is selected by the nodes in the domain through factors that include connectivity, CPU performance, and energy usage. The coordinator receives messages from the connected nodes in the domain about the amount of resources it has and the tasks it needs to perform. The coordinator then assigns tasks to each node to maximize the resource usage of each node. This proves to reduce latencies as well as improve energy efficiency. Increasing the number of nodes in a fog domain enables the system to perform better due to more resources being available, however, the communication range between nodes should be increased which will result in an increase in energy usage [15, 16]. Therefore, as a constraint, resource sharing can improve both latency and energy usage to a certain degree.

Using a mathematical model, [17] makes use of defined policies that are constructed around latency used to determine where tasks are performed in a fog infrastructure. The problem definition is to create a low latency network that consists of a group of users in the same domain, acting as one source node and a fog server as another node. Three different policies are created that are used to determine which fog node is to perform a task. The first policy assigns a task to a randomly selected node from a uniform distribution. The second policy selects the fog node with the lowest latency given the current state of the fog system. The third policy makes use of the fog node with the largest amount of resources available to perform a task. The system determines which fog nodes are to be used in the candidate list for redistributing the workload based on the workload generated by each source node. Based on the policies, a fog node will process the workload. By adjusting the number of nodes available and threshold values, it was proven that the lowest latency policy was the most effective.

By looking at a service orientated resource management model, [18] makes use of a user's behavior towards the usage of a resource in order to allocate resources accordingly. This makes use of a customer's historical use of the resources over time. Usage of resources can become more efficient when there is a predictive model in place to determine what resource needs to be used. With new users, default probabilities are used which results in the system. This can prove to be an issue if a new user takes advantage of resources over a user that has a good history for use of resources. An issue with this approach is that scheduling of tasks does not consider the heterogeneity of fog components which can degrade the performance of the fog system. Considering the type of service, overall service relinquishes probability and service-oriented relinquish probability, this implementation helps improve resource management that prevents wasting resources.

Making use of edge intelligence in fog computing, [19] makes use of two implementations that will help with application offloading and storage expansion for mobile devices [20, 21]. With application offloading, resources that are viewed include network latency, bandwidth, size of the overhead data and dynamic allocation of the partitions used for the loading. In storage expansion, making use of the user's devices that are near is viewed to improve latency, security, and privacy. Linear regression models usually result in static decisions since they do not account for the dynamic environment of resources being used. By monitoring resources constantly, intelligence in fog computing can encourage effective offloading decisions in a fog network. A test bench of 5 different environments is created with three representing fog-based systems and two representing cloud-based systems. The fog systems contain short-range communication methods

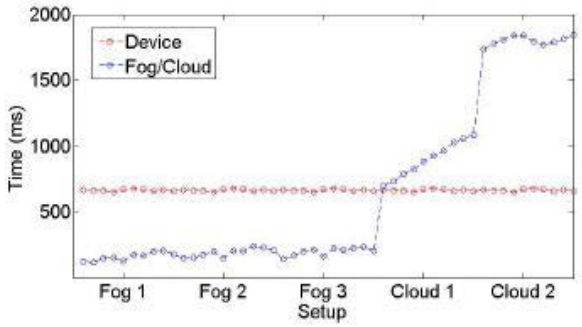


Figure 4. Fog vs Cloud execution time (Taken from [14])

such as WLAN or Wi-Fi which help reduce latency, opposed to the cloud systems that make use of 3G or 4G connectivity while cloud systems contain better CPU performance over the fog systems.

Making use of the face recognition method of Picaso, it can be viewed in Fig. 3 that the fog systems reduce the time drastically to process data. From this result, a model can be created to predict the performance of various tasks to offload them efficiently based on parameters such as bandwidth, latency, CPU usage, and memory. For the offloading of storage, the goal is to achieve a reduction in cost and latency for the user. This is done through a geographically distributed location policy that creates a cluster of files based on the data type, operation, access and access in the future [19].

Using Cloud Atomization Technology, fog nodes on different levels can be used to create virtual machine nodes [22]. This is performed using graph partitioning to improve load balancing techniques in a fog system. By doing this, a dynamic load balancing method is effectively implemented to handle resources and reduce the consumption of node migration that is introduced by system changes [22]. This implementation is focused on a balancing algorithm that is proportional to the number of resources available. Improvement of this algorithm can be done by analyzing the change in the system environment

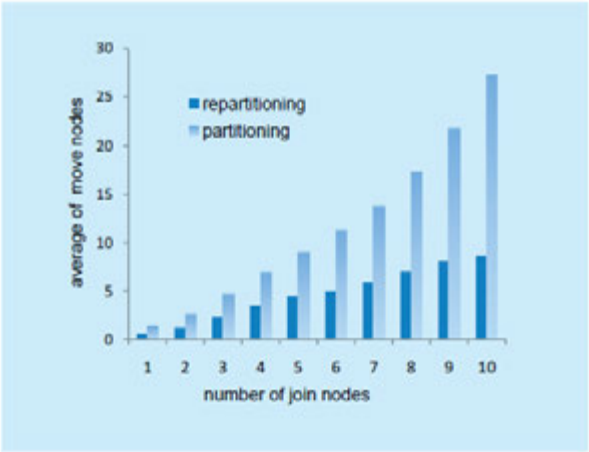
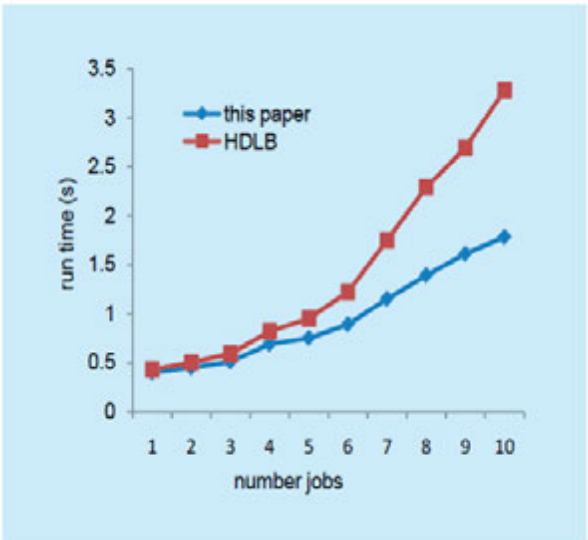


Figure 3. Cloud atomization using graph partitioning scheme (Taken from [22])

and improving the performance of the elastic load balancing method.

Looking at Fig. 4, the number of nodes that are moved is considerably less when repartitioning the fog system for load distribution. Making use of a static balancing algorithm will need to assess the load of each node which can result in partitioning occurring more often than required to balance the workload. By repartitioning, there



is a reduction in the number of resources used and waiting time.

Figure 5. Cloud atomization repartitioning scheme run time (Taken from [22])

Fig. 5 shows a reduced run time of the repartitioning scheme compared to the classical hybrid strategy. Mentioned earlier, the workload will need to be computed each time before a partition can take place which will increase the runtime of the system.

Reducing redundant elements of computing and storage elements in a fog system can reduce the bandwidth consumption and latency for users [23]. Making use of coding concepts that include minimum bandwidth codes and minimum latency codes. Minimum bandwidth codes create a trade-off between computational and communication loads by making use of computing resources that are not being used to their potential to reduce communication loads. Minimum latency codes are the opposite of minimum bandwidth codes in the sense that applications can be completed much faster by incorporating more computational resources. This can be further improved by making use of a unified coding framework that decreases the communicational and computational load of a system.

IV. DISCUSSION

All the discussed computing algorithms in section III provide a better understanding as to what can be achieved from an algorithmic point of view for improving fog systems. The system proposed in [9] created an efficient way to utilize the maximum potential of resources in a short-range communication spectrum. This gives the heterogeneity of quality to the system. This system is improved by adding for QoS through dynamic sizing and management of resources of the computation clusters [10]. This is also the same case with [11], [13], [16] where the fog system is heterogeneous and QoS is achieved. QoS is achieved only with [12], [14] and [15]. What is the challenge in these implementations is the scalability and mobility factors. Considering the characteristics of what fog systems should contain, these algorithms lack federation and mobility. This is where fog algorithms can be further developed, by incorporating all the factors of heterogeneity, QoS, scalability, mobility, and federation to create a tradeoff system to ultimately improve cloud and fog systems.

V. CONLUSION

This paper described what fog computing is, the challenges that exist in this infrastructure and what can be obtained through algorithms that are designed for fog systems. These solutions can also be applied to cloud-based systems where applicable to improve computing, network bandwidth, and latency as well as storage constraints.

REFERENCES

[1] M. Chiang, and T. Zhang, “Fog and IoT: An Overview of Research Opportunities,” IEEE Internet Things J., vol. 3, no. 6, 2016, pp. 854–864.

[2] F. Bonomi, R. Milito, P. Natarajan, and J. Zhu, “Fog Computing: A Platform for Internet of Things and Analytics,” vol. 546, 2014, pp. 169–186.

[3] C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow, and A. Polakos, “A Comprehensive Survey on Fog Computing : State-of-the-art and Research Challenges,” IEEE Communications Surveys & Tutorials, vol. 20, no. 1, 2017, pp. 1–49.

[4] Cisco Systems, “Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are,” https://www.cisco.com/c/dam/en_us/solutions/trends/iot/docs/computing-overview.pdf, 2016.

[5] R. Balan, J. Flinn, M. Satyanarayanan, S. Sinnamohideen, and H. I. Yang, “The case for cyber foraging,” Proceedings of the 10th workshop on ACM SIGOPS European workshop, 2002, pp. 87-92.

[6] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, “The Case for VM-Base Cloudlets in Mobile Computing,” IEEE Pervasive Computing, vol. 8, no. 4, 2009, pp. 14–23.

[7] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, “On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration,” IEEE Communications Surveys & Tutorials, vol. 19, no. 3, 2017, pp. 1657–1681.

[8] A. Ahmed, and E. Ahmed, “A Survey on Mobile Edge Computing,” 10th IEEE International Conference on Intelligent Systems and Control, (ISCO 2016), 2016.

[9] A. Thakur, M. Reza, "Fog Computing for Detecting Vehicular Congestion, an Internet of Vehicles Based Approach: A Review", IEEE Intelligent Transportation Systems Magazine, Vol.11, no.2, pp.8-16, 2019.

[10] N. Ye, Z. Wang, R. Malekian, Y. Zhang, R. Wang "A Method of Vehicle Route Prediction Based on Social Network Analysis," Journal of Sensors, vol. 15, pp.1-10, 2015.

[11] R. Malekian, Alain Kavishe, B. Maharaj, P. Gupta, G. Singh, H. Waschefort, “Smart vehicle navigation system using Hidden Markov Model and RFID Technology“, Wireless Personal Communications, Springer, Vol. 90, issue, 4, pp. 1717-1742, 2016.

[12] S. F. Abedin, M. G. R. Alam, N. H. Tran, and C. S. Hong, “A Fog based system model for cooperative IoT node pairing using matching theory,” Network Operations and Management Symposium (APNOMS), 17th Asia-Pacific, 2015, pp. 309–314.

[13] J. Oueis, E. C. Strinati, S. Sardellitti, and S. Barbarossa, “Small Cell Clustering for Efficient Distributed Fog Computing: A Multi-User Case,” Vehicular Technology Conference (VTC Fall), IEEE 82nd, 2015, pp. 1–5.

[14] T. Nishio, R. Shinkuma, T. Takahashi, and N. B. Mandayam, “Service-oriented heterogeneous resource sharing for optimizing service latency in mobile cloud,” Proceedings of the first international workshop on Mobile cloud computing & networking, MobileCloud '13, Bangalore, India, 2013, pp. 19-26.

[15] C. Shao, Y. Sun, R. Malekian, "Research on Cost-Balanced Mobile Energy Replenishment Strategy for Wireless Rechargeable Sensor Networks", IEEE Transactions on Vehicular Technology, Vol.69, no.3, pp.3135-3150 , 2020.

[16] C. Sha, C. Ren, M. Reza, M. Wu, H. Huang, N. Ye, "A Type of Virtual Force-Based Energy-Hole Mitigation Strategy for Sensor Networks", IEEE Sensors Journal, Vol. 20, No. 2, pp.1105 - 1119, 2020.

[17] K. Intharawijitr, K. Iida, and H. Koga, “Analysis of fog model considering computing and communication latency in 5G cellular networks,” Pervasive Computing and Communication Workshops (PerCom Workshops), 2016 IEEE International Conference, 2016, pp. 5–8.

[18] M. Aazam, and E. N. Huh, “Dynamic resource provisioning through Fog micro datacenter,” IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), 2015, pp. 105–110.

[19] M. A. Hassan, M. Xiao, Q. Wei, and S. Chen, “Help your mobile applications with fog computing,” Sensing, Communication, and Networking - Workshops (SECON Workshops), 12th Annual IEEE International Conference, 2015, pp. 49–54.

[20] W. Liu, Z. Li, R. Malekian, M. Sotelo, Z. Ma, W. Li, "A Novel Multi-Feature based On-Site Calibration Method for LiDAR-IMU Systems", IEEE Transactions on Industrial Electronics, Vol.67, 2020.

[21] G. Shao, Y. Ma, R. Malekian, X. Yan, Z. Li, "A Novel Cooperative Platform Design for Coupled USV–UAV Systems", IEEE Transactions on Industrial Informatics, Vol. 15, no.9, pp.4913-4922, 2019.

[22] S. Ningning, G. Chao, A. Xingshuo, and Z. Qiang, "Fog computing dynamic load balancing mechanism based on graph repartitioning," China Communications, vol. 13, no. 3, 2016, pp. 156–164.

[23] S. Li, M. A. Maddah-Ali, and A. Salman Avestimehr, "Coding for Distributed Fog Computing," IEEE Communications Magazine, Issue on Fog Computing and Networking, vol. 55, no. 4, 2017, pp. 34–40.

DNN Placement and Inference in Edge Computing

Mounir Bensalem, Jasenka Dizdarević and Admela Jukan
Technische Universität Braunschweig, Germany
{mounir.bensalem, j.dizdarevic, a.jukan}@tu-bs.de

***Abstract*—The deployment of deep neural network (DNN) models in software applications is increasing rapidly with the exponential growth of artificial intelligence. Currently, such models are deployed manually by developers in the cloud considering several user requirements, while the decision of model selection and user assignment is difficult to take. With the rise of edge computing paradigm, companies tend to deploy applications as close as possible to the user. Considering this system, the problem of DNN model selection and the inference serving becomes harder due to the introduction of communication latency between nodes. We present an automatic method for DNN placement and inference in edge computing; a mathematical formulation to the DNN Model Variant Selection and Placement (MVSP) problem is presented, it considers the inference latency of different model-variants, communication latency between nodes, and utilization cost of edge computing nodes. Furthermore, we propose a general heuristic algorithm to solve the MVSP problem. We provide an analysis of the effects of hardware sharing on inference latency, on an example of GPU edge computing nodes shared between different DNN model-variants. We evaluate our model numerically, and show the potentials of GPU sharing, with decreased average latency by 33% of millisecond-scale per request for low load, and by 21% for high load. We study the tradeoff between latency and cost and show the pareto optimal curves. Finally, we compare the optimal solution with the proposed heuristic and showed that the average latency per request increased by more than 60%. This can be improved using more efficient placement algorithms.**

I. Introduction

The rapid development of DL techniques, as a result of both the hardware and software advancements, as well as the increasing demand for data processing at the network edge has led to the new research direction, which sees these two trends combined [1]. But, the confluence of DL and edge computing raises mulitple challenges, as it assumes capabilities of performing computationally demanding DL tasks on resource constrained edge computing devices [2]. Among multiple open research issue, of special interest are modelling machine learning (ML) inference latency and ML model selection optimization in edge computing, along with related task placement and processor sharing. This is due to the fact that such models [3] developed for cloud computing cannot be directly applied in edge computing. The DNN placement problem in the edge needs to consider in particular the communication delay between nodes and the hardware heterogeneity of devices. To the best of our knowledge there has been no study of the DNN application selection, placement and inference serving problem in consideration of edge computing.

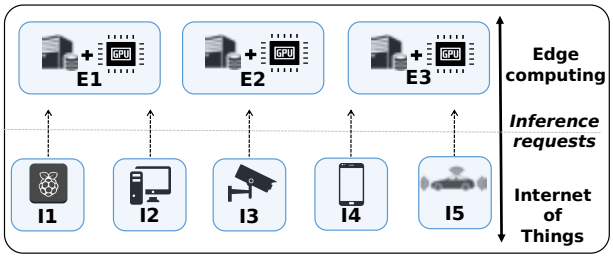


Fig. 1: A reference inference service system

In this paper, we extend the Deep Neural Network (DNN) Placement and Inference model in edge computing proposed in [4]. We provide a mathematical formulation of the problems of ML placement and inference service, considering inference latency of different model-variants, communication latency between nodes and utilization cost of edge computing nodes (resources). Our model also includes an analysis of the effects of hardware sharing, with GPU edge computing nodes shared between different model-variants, on inference latency. As a result, the model can be used to define different scenarios of GPU sharing to approximate the inference latency of DNN models in various DNN placements in edge computing.

The illustration of the DNN application placement problem is presented in Figure 1 [4] with the arrival of inference requests from IoT nodes to the edge computing layer. IoT nodes are assumed to be devices with processing and sensing capabilities, but not enough to run DNN models. In this system abstraction, edge computing layer, consisting of edge nodes with GPUs for running ML models, serves as an inference service system to the requests from IoT nodes. For the illustrated system we focus on designing a placement strategy of ML models, taking into account different possibilities of model-variants and how to forward requests coming from IoT nodes.

The remainder of this paper is organized as follows. Section II gives the necessary background in the context of state-of-the-are and presents the reference scenario to assessing the inference latency. Section III introduces a mathematical model for the MVSP problem. Section V numerically evaluates the proposed model. We conclude the paper in VI.

II. ML Inference Scenarios

A. Background

While there are many existing ML inference serving systems, such as Google Cloud ML, Azure ML, and AWS Sage Maker, which are centralized in the cloud, the implementation

and evaluation of ML inference serving systems in edge computing networks is still an open issue. One of the most obvious expected benefits of edge computing is the improved latency of inference, which in turn is allowing for faster application responses to users' requests. In addition, deep learning applications in edge computing have benefited from the improvements related to algorithms and software, as well as the development of novel hardware architectures for edge devices [5].

To serve ML inference queries GPU accelerators play an important role. Google Research has adapted DNN inference to run on top of mobile GPU [6]. Similarly, Amazon Web Services proposed a solution to run inference models on integrated GPUs at the edge [7]. But for this kind of applications, besides running inference models on GPU accelerators, it is necessary to consider GPU sharing as well, in order to allocate resources efficiently, opening another area of research. This approach is set to improve on the low utilization and scaling performances of unshared access to a GPU. That idea of GPU sharing can be promising as seen in [8], where authors studied the performance of temporal and spacial GPU sharing and [9], which presented a GPU cluster manager enabling GPU sharing for DL jobs.

B. Impact of GPU Sharing

For inference serving systems which deploy ML models, devices like GPU, TPU, dedicated accelerators are used due to their high performances and as of now most of them work exclusively for one ML model at a time [3]. We mentioned previously that GPU sharing is emerging as a promising direction in terms of further resource allocation optimization, and its effects on inference analysis that we want to explore in this section. With the rising efforts in the deployment of demanding ML algorithms on edge computing devices more resources are being required, adding complexity to the task of resource allocation. This has led to the utilization of new container-based lightweight virtualization technologies for dealing with limited resources and enabling software execution on heteregenous hardware architectures, as is often the case in the edge. This would allow for separating the model instances that would run in parallel in the same machine. In general this means that resource management systems can scale-up and scale-down allocated resources based on the load variation using these new virtualization technologies. However, sharing those devices accross ML models needs to be studied in terms of performance and scalability due to the additional latency that may be generated while deploying two models or more in the same machine.

Studying the impact of GPU sharing on the performance of ML models is important task because such analysis can highly affect the design of a management system especially how to scale-up and scale-down resources and how to choose the best model-variant. Due to the complexity of the GPU analysis, which requires a detailed study of numerous existing benchmarks with different ML models, different batch sizes, and GPU memory limitations for our application interference model, we will use a simplified analysis of the effects that replication and co-locating of model-variant can have on the inference latency. For this analysis, we propose two scenarios for calculating inference latency. The proposed inference latency given here will later be added to our Edge computing network model in Section III.

In order to model the inference latency of model-variants we use the following scenario assumptions:

- Scenario 1: No GPU sharing among ML models
  In the first scenario, we propose a common approach of unshared access to a GPU, used in well known machine learning cloud based services such as Amazon AWS SageMaker and Google CloudML Engine. We observe the inference latency when each model-variant workload has been placed in an exclusive GPU.
- Scenario 2: It is possible to share the acces to the same GPU to at most K model-variants
  In the second scenario, we assume parallel usage of hardware in terms of GPU sharing. This type of approach is still in its early phases in the context of edge computing. To address this problem, we analyze two potential behaviors of contribution of inference latency in presence of multiple model-variants: constant and linear increase.
  - Scenario 2.1: The inference latency of a model-variant is constant in presence of other model variants $mv1$ to mvK
    $$L_{mv1}^{mv2,...,mvK} = L_{mv1} \tag{1}$$
  - Scenario 2.2: The inference latency of a model-variant increases linearly in terms of the latency of co-located model-variants
    $$L_{mv1}^{mv2,...,mvK} = L_{mv1} + \sum_{k=2}^{K} \alpha_k L_{mvk} \tag{2}$$

where $L_{mv1}^{mv2,...,mvK}$ is the inference latency of model-variant $mv1$ in presence of $mv2$,...,mvK and L_{mv1}, L_{mv2} are the inference latency of $mv1$ and mvk running exclusively in the device, respectively. α_k is a coefficient that we call *interference coefficient* of model variant $mv1$ in presence of model variant mvk. This coefficient is introduced to estimate the latency of co-located models in terms of the latency of models running exclusively in the hardware. Latency implications of co-locating a different model on the same node, presented in this section, will be added to the total latency of our system model in Section III.

III. System Model

A. Reference Edge Computing Network Model

In order to analyze MVSP problem in edge computing network we define a system model which will consider inference latency of different model-variant with shared and unshared access to GPUs, node communication latency and utilization cost. The considered system consists of N_I IoT nodes, e.g. smart phone, security camera and smart car cameras, and N_E edge nodes, e.g. access points. Let $\mathcal{N}_\mathcal{I} = \{1, ..., N_I\}$ and $\mathcal{N}_\mathcal{E} = \{1, ..., N_E\}$ denote the set of indexes of IoT nodes and

edge nodes, respectively. Edge nodes are able to host various ML applications designated to serve the inference requests coming from IoT nodes. Every edge node $e \in \mathcal{N_E}$ has a computing unit specific for inference serving tasks e.g. CPU, GPU and TPU as well as memory capacity C^e. We assume that we have M different ML models that can be used for different tasks such as face recognition and object detection. Each ML model m can have V_m variants with different sizes and inference latencies per request and can be deployed via a VM or a container. We denote by $\mathcal{M} = \{1, ..., M\}$ the set of ML models and $\mathcal{V}_m = \{1, ..., V_m\}$ the set of variants of model m. Each model variant (m, v) has a minimum memory requirement R_{mv} to be loaded and can process at most $Load_{mv}$ with a stable performance. Each IoT node i can define its own latency requirement $Lmax^i_m$ for each infered model m as well as the number of inference requests r^i_m.

The system is designed to decide the placement strategy of ML models in edge nodes considering different possibilities of model-variants and the forwarding decision of requests coming from IoT nodes.
We introduce a binary variable x^{ie}_{mv} to indicate the forwarding decision of requests of model-variant (m, v) from IoT node $i \in \mathcal{N_I}$ to edge node $e \in \mathcal{N_E}$. The placement decision of model-variant (m, v) in an edge node e is defined by an integer variable n^e_{mv}, which indicates the number of deployed instances.

B. Latency Model

We consider two types of latencies: communication latency between IoT and edge nodes, and inference latency of model-variants in edge nodes. We denote by CL^{ie} the communication latency between IoT node i and edge node e. The inference latency of a model-variant (m, v) running exclusively on edge node e is denoted by L^e_{mv}. We define IL^e_{mv} as the inference latency of a model-variant (m, v) running on edge node e. For mathematical model of this latency we include the effects that sharing with other model-variants can have, as well as observing the case with unshared access to GPU. With that in mind, we assume in our formulation that an edge node can be shared by at most K model-variants. The average latency per request is given by:

$$L = \frac{1}{\sum_{i\in\mathcal{N_I}}\sum_{m\in\mathcal{M}} r^i_m} \sum_{i\in\mathcal{N_I}}\sum_{e\in\mathcal{N_E}}\sum_{m\in\mathcal{M}}\sum_{v\in\mathcal{V}_m} r^i_m (CL^e_i + IL^e_{mv}) x^{ie}_{mv} \tag{3}$$

The communication latency between node $i \in \mathcal{N_I}$ and node $e \in \mathcal{N_E}$ is considered as the sum of the delay on each link in the shortest path in both directions (sending request and receiving response). The delay d_l on each link l is assumed to have a random value with an average μ, including all possible existing delays in the link i.e. transmission, queuing, propoagation and processing. We denote by P_{ie} the set of links in the shortest path between node i and e.

$$CL^e_i = \sum_{l\in P_{ie}} d_l \tag{4}$$

We model the inference latency of a model-variant (m, v) running on edge node $e \in \mathcal{N_E}$ for each scenarios presented in Section II as follows:
Scenarios 1 & 2.1:

$$IL^e_{mv} = L^e_{mv} \tag{5}$$

Scenarios 2.2:

$$IL^e_{mv} = L^e_{mv} + \alpha_{mv} L^e_{mv}(n^e_{mv} - 1) + \sum_{m'\in M}\sum_{\substack{v'\in\mathcal{V}_m \\ v'\neq v}} \alpha_{m'v'} L^e_{m'v'} n^e_{m'v'} \tag{6}$$

where the inference latency is the sum of the inference latency of a model-variant running exclusively on an edge node (L^e_{mv}), the additional latency created by replication and the additional latency created by co-locating a different model on the same node.
IoT nodes (users) are assumed to express their latency requirements for the inference of a model $m \in \mathcal{M}$ with a latency requirement constraint given by:

$$\sum_{v\in\mathcal{V}_m} (CL^e_i + IL^e_{mv}) x^{ie}_{mv} \leq Lmax^i_m \quad \forall i \in \mathcal{N_I},\ \forall e \in \mathcal{N_E},\ \forall m \in \mathcal{M} \tag{7}$$

This constraint assumes that the round trip time (RTT) cannot exceed a maximum value of latency given by the user as a requirement $Lmax^i_m$. In our case, RTT is the sum of the communication delay (cumulative delay among the path) and the processing delay of inference request in the edge node.

C. Utilization Cost Model

The utilization cost model is an abstract formulation of all the costs induced from the utilization of edge resources, assuming that such a cost would increase with the increase of resource utilization. As an example, the utilization cost can represent the power consumption and energy efficiency measurements in a unity of power (Watt), considering different hardware components such as CPU, GPU, memory, and I/O. [10] shows some inference benchmarks of several DNN model-variant using Jetson AGX Xavier, Nvidia GPU.

For the sake of generality, we define a continuous variable z^e denoting the utilization cost of a node $e \in \mathcal{N_E}$. The average utilization cost of all edge nodes is given as:

$$C = \frac{1}{N_E} \sum_{e\in\mathcal{N_E}} z^e \tag{8}$$

The cost of every edge node e is related to its memory utilization u_e. Similarly to [11], the utilization cost follows an exponential function $\phi(.)$ of the utilization. We denote by $\mathcal{Y}$ a set of linear functions tangent to $\phi(.)$. Using the set of linear functions $\mathcal{Y}$, we approximate the utilization cost as follows:

$$\phi(x) = \max_{y\in\mathcal{Y}} y(x), \forall x \in \mathbb{R} \tag{9}$$

The following constraints insure that the variable z^e gets a value approximately equal to $\phi(u_e)$.

$$z^e \geq y(u_e), \forall e \in \mathcal{N_E},\ \forall y \in \mathcal{Y} \tag{10}$$

$$u_e = \frac{1}{C^e} \sum_{m \in \mathcal{M}} \sum_{v \in \mathcal{V}_m} R_{mv} n^e_{mv} \leq 1, \ \forall e \in \mathcal{N}_\mathcal{E} \tag{11}$$

Constraint (11) shows the definition of an edge node utilization u_e, as the sum of utilization of all possible model-variants mv i.e. in terms of required memory per loaded model. Then we divide the obtained sum by the memory capacity of a node (mainly GPU memory).

D. Loading and Scaling Model

A model-variant can be loaded in a specific node to serve requests coming from users (IoT nodes). When the requests load increases the deployed model-variant may not be able to serve users. Considering this scenario, the system may replicate the model instance on top of a new VM or container which scale up the throughput (based on container technologies) or use a different model-variant with a bigger batch size (i.e has higher fps).

The following constraints can insure that the load on a specific model variant on a specific node cannot exceed the maximum load $Load_{mv}$. Moreover, by conserving the maximum load, this constraint can scale-up the number of model-variant replicas (called *variant replication* in [3]), or choose to use another model variant that has less inference latency (called *variant upgrading* in [3]) regarding the minimization of the average latency (eq. 3).

$$\sum_{i \in \mathcal{N}_\mathcal{I}} r^i_m x^{ie}_{mv} \leq Load_{mv} n^e_{mv}, \ \ \forall e \in \mathcal{N}_\mathcal{E}, \forall m \in \mathcal{M}, \forall v \in \mathcal{V}_m \tag{12}$$

The following constraints assure that the number of model-variant instances is bigger than 1 only if at least one node is sending inference requests.

$$n^e_{mv} \geq x^{ie}_{mv} \ \ \forall i \in \mathcal{N}_\mathcal{I}, \forall e \in \mathcal{N}_\mathcal{E}, \forall m \in \mathcal{M}, \forall v \in \mathcal{V}_m \tag{13}$$

The number of instances shared by a node can be at most K:

$$n^e_{mv} \leq K \ \ \forall e \in \mathcal{N}_\mathcal{E}, \forall m \in \mathcal{M}, \forall v \in \mathcal{V}_m \tag{14}$$

The memory capacity constraints can be defined as follows:

$$\sum_{m \in \mathcal{M}} \sum_{v \in \mathcal{V}_m} R_{mv} n^e_{mv} \leq C^e, \ \ \forall e \in \mathcal{N}_\mathcal{E} \tag{15}$$

E. Problem Formulation

The Model Variant Selection and Placement (MVSP) problem aims to minimize both the average latency per request (eq. 3) and the average utilization cost of edge nodes (eq. 8):

$$\begin{aligned} &\mathbf{P:} \ \min \alpha L + (1-\alpha) C \\ &s.t. \sum_{v \in \mathcal{V}_m} \sum_{e \in \mathcal{N}_\mathcal{E}} x^{ie}_{mv} = 1, \ \ \forall i \in \mathcal{N}_\mathcal{I}, \ \ \forall m \in \mathcal{M} \\ &\text{Constraints: } (7), (10), (11), (12), (13), (14), (15) \end{aligned} \tag{16}$$

where α denotes weight of the average latency in the objective function. The first constraint in the problem (16) insures that a request from a specific IoT node can be processed only by one edge node. Constraint (7) assures that RTT cannot exceed the maximum tolerated latency. Constraint (10) and (11) are used to compute utilization cost per edge node. Constraint (12) insures that the load assigned to a specific model-variant deployed in a specific edge node does not exceed its maximum processing capacity. Constraints (13) and (14) defines the values of binary and integer variables used in the model. Constraint (15) assures the satisfaction of memory capacity per edge node.

IV. Heuristic Solution for MVSP Problem

In order to solve MVSP problem for large scale networks, we propose a general heuristic solution **Algorithm. 1** as well as two algorithms for placement **Algorithm. 2** and for assignment **Algorithm. 3**. The main idea behind the general heuristic is to place models one by one using a placement algorithm. After placing a model m, we check all IoT nodes in the system that send inference requests to m. Then, we assign the sending node to an edge node that satisfy its latency requirements considering the processing capacity of edge nodes, and for that we use an assignment algorithm. At the end of the execution, we compute the total cost of the solution and the average latency per request, using the same equations from the previous utilization cost model, and latency model, respectively. The RPA **Algorithm. 2** consists on randomly selecting a model-variant $v \in \mathcal{V}_m$, and an edge node $e \in \mathcal{N}_\mathcal{E}$. Afterwards, it computes the utilization of the edge nodes and checks if the placement is possible to take the allocation decision. The IAA **Algorithm. 3** is triggered by the GH whenever it tries to assign an IoT node, requesting a model, to an edge nodes deploying the a variant of the same model. To satisfy this demand, the IAA check all existing edge nodes sequencially, and on each one it checks all possible model-variant related to the demanded model, if the model exists, and have enough resources to satisfy new requests, then IAA takes the assignment decision, otherwise it tries with the next node.

Algorithm 1 General Heuristic (GH)

1: **Initialize**: $\overrightarrow{n} \leftarrow \overrightarrow{0}, \overrightarrow{x} \leftarrow \overrightarrow{0}$
2: **for** $m \in \mathcal{M}$ **do**
3: $\quad \overrightarrow{n_m} \leftarrow$ Place(m, $\mathcal{N}_\mathcal{E}$)
4: $\quad$ **for** $i \in \mathcal{N}_\mathcal{I}$ **do**
5: $\quad\quad \overrightarrow{x^i_m} \leftarrow$ Assign(i, $\overrightarrow{n_m}$, $\mathcal{N}_\mathcal{E}$)
6: $\quad$ **end for**
7: **end for**
8: $C \leftarrow$ Cost($\overrightarrow{n}$)
9: $L \leftarrow$ Latency($\overrightarrow{n}$, $\overrightarrow{x}$)
10: **return** C, L

V. Numerical Analysis

In this section, we evaluate our optimization model using two problem instances P_1 and P_2, based on MANIAC mobile ad hoc network. Table I shows the topology of each studied problem. For each problem, we choose the DNN models randomly from a pre-defined list. The communication latency is obtained from [12], which was estimated to have a random value with an average 12.23 ms per link. The inference latency of each model-variant was measured on a GPU GTX 1050 Ti using tensorflow framework.

Algorithm 2 Random Placement Algorithm (RPA)

function PLACE(m, $\mathcal{N}_\mathcal{E}$)
  Select randomly v from $\mathcal{V}_m$ and e from $\mathcal{N}_\mathcal{E}$
  Compute u_e using (11)
  if $u_e < 1$ **then**
    Place model-variant (m, v) in node e
  end if
  return $\overrightarrow{n_m}$
end function

Algorithm 3 Inference Assignment Algorithm (IAA)

function ASSIGN(i, n_m, $\mathcal{N}_\mathcal{E}$)
  for $e \in \mathcal{N}_\mathcal{E}$ **do**
    for $v \in \mathcal{V}_m$ **do**
      if $n^e_{mv} \neq 0$ and $Load^e_{mv} < Load_{mv}$ **then**
        Assign requests on m from i to e
        return $\overrightarrow{x^i_m}$
      end if
    end for
  end for
  $\overrightarrow{n_m} \leftarrow$ Place(m, $\mathcal{N}_\mathcal{E}$)
  $\overrightarrow{x^i_m} \leftarrow$ Assign(i, $\overrightarrow{n_m}$, $\mathcal{N}_\mathcal{E}$)
end function

In addition to analysis provided in this section, in [4] we have analysed the impact of the inference load and the effects that replication and co-locating of model-variant can have on the inference latency. In cases of high load, the optimization tended to satisfy inference requests by allocating new model-variants in distant edge nodes that have enough capacity to host the instances. The obtained results also showed the potentials of GPU sharing, with decreased average latency by 33% of millisecond-scale per request for low load, and by 21% for high load.

A. Impact of Application Interference

We consider three different scenarios as described in II. In the first scenario, an edge node can run only one DNN model-variant. In the second and third scenarios with GPU sharing in consideration, we set the maximum number of co-located models to 2. We set weight α equal to 0.1, node memory capacity $C^e = 0.1, \quad \forall e \in \mathcal{N}_\mathcal{E}$ and interference coefficient $\alpha_{mv} = 0.1, \quad \forall m \in \mathcal{M}, \ \forall v \in \mathcal{V}_m$. We test our optimization model using different request rates, which correspond to the average of requests per node (e.g $E(r) = 5.5$ where r denotes the random variable of request rates). Table II shows the optimization results of latency and cost for the two problems. For the first scenario, the model can satisfy requests only for low request rates, which can be explained by the fact that in this scenario, it is not allowed to share the GPU by more than one DNN model while in order to satisfy the increased load thy system should be scaled up. For scenario 2.1, we considered the possibility of GPU sharing without considering the effect of application interference (i.e. the additional latency caused by co-located instances). The average latency and the average cost increase with the increase of the request rate. This result means that the optimization is placing new instances in more distant edge nodes to IoT nodes (on average), causing extra communication latency. For scenario 2.2, we considered both the possibility of GPU sharing and the effect of application interference. In this case, as expected, the average latency is higher than the corresponding average latency for scenario 2.1 because the only difference between the two cases is the additional latency of application interference. However, the average cost is also higher in this case, which means that the placement decision is not the same for the two scenarios because the average cost increases in terms of node utilization. We remark in the placement decision of the optimization that bigger model-variants have been used to satisfy inference requests.

TABLE I: Topologies

Problem	N_I	N_E	M	V_m
P1	10	5	3	8
P2	11	4	4	8

TABLE II: Average latency and cost for different scenarios

$P1$	($\alpha = 0.1$, $\alpha_{mv} = 0.1$, $C^e = 8$, $K = 2$)					
	$E(r) = 5.5$		$E(r) = 22$		$E(r) = 33$	
Scenario	C	L	C	L	C	L
1	0.027	20.83	-	-	-	-
2.1	0.051	17.60	0.073	18.41	0.099	20.96
2.2	0.055	17.66	0.083	18.65	0.112	37.10
$P2$	($\alpha = 0.1$, $\alpha_{mv} = 0.1$, $C^e = 8$, $K = 2$)					
1	0.016	21.27	-	-	-	-
2.1	0.050	13.36	0.050	13.41	0.058	14.08
2.2	0.051	13.38	0.051	13.43	0.06	14.10

B. Trade-off between the Average Latency and Cost Function

We set the configuration parameters similar to the previous experiment, setting the average request load to 27.5, and we alter the weight of the latency and cost in the objective function to evaluate the trade-off between the two objectives. Figure 2 shows the opposite behavior of the average latency and the average utilization cost. When we consider only the cost ($\alpha = 0$), the optimization tends to allocate the least possible number of instances of each model that can satisfy inference requests. This results in a high value of latency due to the assignment of IoT nodes to distant edge nodes: average latency is 31 ms, average utilization is 67% and average cost is 0.23. When we increase the value of α, we consider more the latency in the decision making. An increase in α cause an increase in the value of the cost and a reduction in the average latency, until a maximum value in which the two objectives converge ($\alpha = 0.2$). The intersection of Pareto optimal curves, i.e. curves of the two goals: average latency and average cost, happens when α is equal to 0.04 with latency equal to 22.7 ms, cost equal to 0.3 and utilization equal to 70%. Setting the configuration at the intersection point of goals, decreases the cost by 40% compared to a configuration that set a slightly

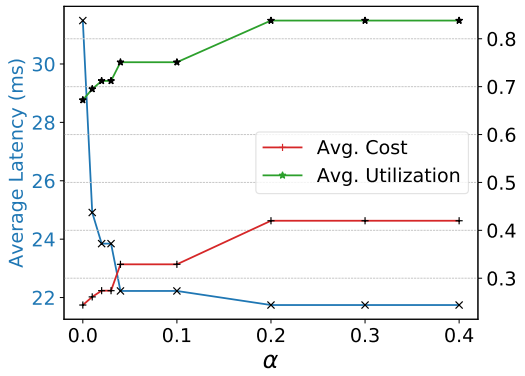


Fig. 2: Trade-off between the latency and cost in MANIAC network.

higher value of α. It is worth it to mention that we show the average utilization curve in the figure because it represents a significant metric while it cannot replace the average cost as the intersection between goals is different than the intersection between latency and utilization. The optimization tends to allocate multiple instances from the same model on different edge nodes (possibly with different variant types in each edge node depending on the possible capacity).

C. Optimal vs Heuristic Solution

We set the configuration parameters similar to the experiment of P_2 in V, and we vary the average load per node to evaluate the optimal model and the proposed heuristic.

Table III shows the optimization results of latency and cost, obtained from the optimal modeling and from the proposed heuristic solution for the problem $P2$. For all inference request load, we see that the average latency per request increased by more than 60%, when we used a heuristic solution based on random placement. For low load, the random placement based solution gave a better utilization cost than the optimal solution (decreased by 13%) and a worst latency per requests (increased by 71%). For higher load, heuristic solution give dramatically worst results in both cost and latency: for a load equal to 33, the cost increased by 220% and latency by 62%. Thus, we conclude that a random placement based solution is not efficient for our problem and we need a better heuristic solutions for online inference systems.

We have to mention that in this short analysis, we tested our heuristic on the same network topology used for the optimal modeling which is a small topology only to compare their performance. Furthermore, we couldn't show results for larger networks because the current optimal modeling has a very high computing complexity for larger network. As a future work, we plan to design more robust and efficient heuristics to outperform the random placement based solution and to deal with large networks and online inference serving systems.

VI. Conclusion

In this paper, we studied the DNN Model Variant Selection and Placement (MVSP) in edge computing networks. A mathematical model was proposed to formulate the problem considering the inference latency of different model-variants, the communication latency between nodes, and the utilization cost of edge computing nodes (resources) as well as proposing a heuristic solution. Also, we analyzed the effects of hardware sharing on inference latency regarding GPU edge computing nodes shared between different model-variants. We studied the placement results of the optimization and its effect on the average latency and cost. We showed that GPU sharing is a valuable approach to handle the increase of inference request rate. We plan to further work on extending our model to consider more parameters like multiple frameworks, multiple hardware devices, and stochastic demand on DNN applications as well as implementing more sophisticated online solutions.

TABLE III: Average latency and cost for different solutions

$P2$	$(\alpha = 0.1, \alpha_{mv} = 0.1, C^e = 8, K = 2)$					
	$E(r) = 5.5$		$E(r) = 22$		$E(r) = 33$	
Model	C	L	C	L	C	L
Optimal	0.051	13.38	0.051	13.43	0.06	14.10
Heuristic	0.045	22.89	0.093	23.03	0.192	22.88

References

[1] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, Aug 2019.

[2] C.-J. Wu, D. Brooks, K. Chen, D. Chen, S. Choudhury, M. Dukhan, K. Hazelwood, E. Isaac, Y. Jia, B. Jia *et al.*, "Machine learning at facebook: Understanding inference at the edge," in *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2019, pp. 331–344.

[3] F. Romero, Q. Li, N. J. Yadwadkar, and C. Kozyrakis, "Infaas: Managed & model-less inference serving," *arXiv:1905.13348*, 2019.

[4] M. Bensalem, J. Dizdarević, and A. Jukan, "Modeling of deep neural network (dnn) placement and inference in edge computing," 2020.

[5] G. Plastiras, C. Kyrkou, and T. Theocharides, "Edgenet: Balancing accuracy and performance for edge-based convolutional neural network object detectors," in *Proceedings of the 13th International Conference on Distributed Smart Cameras*, ser. ICDSC 2019. New York, NY, USA: ACM, 2019, pp. 8:1–8:6. [Online]. Available: http://doi.acm.org/10.1145/3349801.3349809

[6] J. Lee, N. Chirkov, E. Ignasheva, Y. Pisarchyk, M. Shieh, F. Riccardi, R. Sarokin, A. Kulik, and M. Grundmann, "On-device neural net inference with mobile gpus," *arXiv preprint arXiv:1907.01989*, 2019.

[7] L. Wang, Z. Chen, Y. Liu, Y. Wang, L. Zheng, M. Li, and Y. Wang, "A unified optimization approach for cnn model inference on integrated gpus," *arXiv preprint arXiv:1907.02154*, 2019.

[8] P. Jain, X. Mo, A. Jain, H. Subbaraj, R. S. Durrani, A. Tumanov, J. Gonzalez, and I. Stoica, "Dynamic space-time scheduling for gpu inference," *arXiv preprint arXiv:1901.00041*, 2018.

[9] J. Gu, M. Chowdhury, K. G. Shin, Y. Zhu, M. Jeon, J. Qian, H. Liu, and C. Guo, "Tiresias: A {GPU} cluster manager for distributed deep learning," in *16th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 19)*, 2019, pp. 485–500.

[10] "Jetson agx xavier: Deep learning inference benchmarks," https://developer.nvidia.com/embedded/jetson-agx-xavier-dl-inference-benchmarks, accessed: 2020-01-15.

[11] F. Carpio, A. Jukan, and R. Pries, "Balancing the migration of virtual network functions with replications in data centers," in *NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2018, pp. 1–8.

[12] Q. Qin, K. Poularakis, G. Iosifidis, S. Kompella, and L. Tassiulas, "Sdn controller placement with delay-overhead balancing in wireless edge networks," *IEEE Transactions on Network and Service Management*, vol. 15, no. 4, pp. 1446–1459, 2018.

MACHINE LEARNING APPLICATIONS

Predicting and Analyzing Absenteeism at Workplace Using Machine Learning Algorithms

Amarildo Rista *, Jaumin Ajdari *, Xhemal Zenuni*
* SEE University/Faculty of Contemporary Sciences and Technologies, Tetovo, North Macedonia
ar29102@seeu.edu.mk; j.ajdari@seeu.edu.mk; xh.zenuni@seeu.edu.mk;

Abstract – **Absenteeism is the usual or recurrent absence from work is continuously causing disruption in the smooth running of business, affecting the organizational performance and productivity and impacting on the employees' morale. The Oil Refinery in Albania (ARMO), employing 1200 employees is facing high rate of absences. If necessary measures are not being serious dealt with, the issue of absenteeism may jeopardize the operation and production. Prediction of absenteeism is too complex influenced by many factors. Usage of data mining and machine learning algorithms is a good solution to predict and analyze it. The aim of this paper is to identify and evaluate the appropriate ML algorithms to predict and analyses absenteeism at workplace. The dataset taken into account consists of some attributes such as: age, education, employment category, day, month, length of service ect, and 125000 records are considered. Analysis and comparison of various algorithms in terms of accuracy, precision and sensitivity are done in Weka tool.**

Keywords - data mining, absenteeism, machine learning algorithms, prediction.

I.INTRODUCTION

The Oil Refinery in Albania (ARMO) have developed different ways and means in improving employee resource management practices particularly with the objective in reducing absenteeism and to control employee turnover. Employee engagement, motivation and other means such as employee welfare, job performance evaluation and continuous training have been implemented with the objective in reducing employee absenteeism and to improve the organizational performance. Despite of the awareness regarding the adverse effect of absenteeism on the organizational efficiency, the level of absenteeism at workplace still remains high. This paper will focus on prediction and analyses of absenteeism at ARMO, using data mining (DM) and machine learning (ML) algorithms. The ARMO is one of the largest refineries in Balkan that processes oil and its sub-products. The ARMO employing 1200 employees and recently is facing high rate of absenteeism at work. Using DM and ML methods for analyzing and prediction of absenteeism is a good solution. DM is defined as a process used to extract important and useful information from large sets of data [1]. While ML is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed [2]. The first step to analyze data through DM is model designed [3]. A model is an algorithm or set of rules that connect a set of inputs to a specific output. There are two ways to adapt data with design model [4]: 1- Predictive methods; which are used to extract the models that describe important data classes or that predict future data trends. These methods use the known values to predict the unknown values.2- Descriptive methods: represents the recognition of connections that are hidden in the data and provides different results. Based on the features of the dataset and the aim of this paper we have chosen classification methods to analyze our data. These methods use the known values to predict the unknown values. The dataset contains some categories of the reason that provide to be effective in detecting the absenteeism at work such as: age, education, employment category, length of service, residence, civil status, distance from workplace, suffering from any disease, job satisfaction and leadership style. This paper can make a significant contribution in human resource department of an organization. The management can get a sense of understanding of employee activities and behavior which can eventually help in crucial decision making on both monitoring employees at work and recruiting potential employees. The rest of the paper is structured as follows: Section 2 presents a brief discussion on the related works. Section 3 is focused on the data mining classification techniques. Section 4 presents the Methodology. Section 5 presents the result and analysis. And some conclusions are listed in the section 6.

II. RELATED WORK

Application of DM and ML methods in human resource management is a new research field. Predicting and analyzing the absenteeism at workplace play a crucial role in demonstrating the productive and profitable capacity of a company. Some classification techniques are used in [5] to predict absenteeism of employees at workplace. In this paper are analyzed and compared 4 ML algorithms namely Decision Tree, Gradient Boosted Tree, Random Forest and Tree Ensemble. Based on experiments the Gradient Boosted Tree produced the best result with an accuracy rate of 82% compared to other algorithms. In [6] others classification algorithms are

analyzed to find out the probability of an employee being absentee at work for one or more days in a future period of seven days. The authors have analyzed the results of Random Forest, Multilayer Perceptron, Support Vector Machine, Naive Bayes, XGBoost and Long Short Term Memory algorithms concluding that XGBoost was reached the best results being 72% accuracy. Neural Networks and Deep Learning algorithms play an important role to predict the behavior of employees towards punctuality at workplace. In [7], is presented a model designed based on neural network to study the behavior and predict absenteeism of employees at workplace. This model is design based on Deep Neural Network that are composed by multiple hidden layers instead of a single layer of a Shallow Neural Network. The dataset used contains 20 different features and 740 samples which reflect human behaviors. The accuracy results 90.6% compared to 73.3% performance in a single-layer Neural Network and 82% performance in Decision Tree, SVM, and Random Fore. Another similar model based on neural network is presented in [8]. This model is trained with a dataset that has 38 attributes and 2243 records. This model tends to reduce the number of attributes using Rough Sets obtaining good results even in this form. In [9] are implemented three neural network models; Backpropagation, Radial Basis Function and Long-Short Term Memory to solve prediction problem of absenteeism. Based on the experimental results is concluded that Long-Short Term Memory neural network has a prediction rates as 99.9% compared to other. Another ML algorithm is applied in [10]. In this paper are used linear regression and SVR to get predictive absenteeism model. The linear regression is performed with all attributes, while for Support Vector Regression (SVR), are considered two parameters age and seasons of the year. Referring to the results of both algorithms can be used to predict absenteeism at workplace with good accuracy.

III. DATA MINING CLASSIFICATION TECHNIQUES

Classification is considered as the process of finding a model that predicts data classes. Classification techniques [11], are part of predictive methods and are useful to analyze large amounts of data. The main purpose of classification algorithms is to maximize the predictive accuracy obtained by the classification model. There are presented various classification techniques.

A. Decision Tree

The Decision Tree [12] can be used for solving regression and classification problems. The Decision Tree create a training model that can be used to predict the class or value of the target variable by learning decision rules inferred from prior data. This model consists of nodes and a root. The root represents the sample and this further gets divided into two or more sets. The nods are divided in to categories; Decision Node (nodes that are split into further sub-nodes) and Leaf Node (nodes that do not split further). Decision trees classify the instance by sorting them down the tree from the root to some leaf node, with providing the classification of the instances. Each node in the tree acts as a case for some attribute, and each edge descending from the node corresponds to the possible answers to the case. This process is recursive and is repeated for every subtree rooted at the new node. Fig.1 represent the Decision Tree Classifiers.

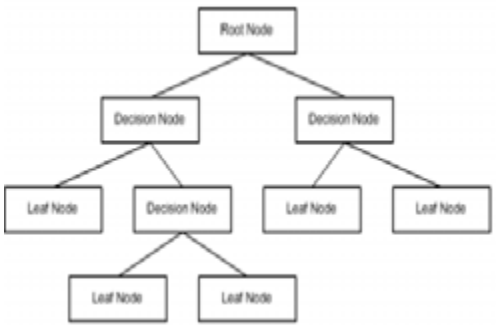


Figure 1. Decision Tree Classifier

Some based decision tree algorithms are: J48, Hoffeding Tree, Random Tree and REPTTree. The J48 classifies the class attribute based on the input attributes and often is referred to as a statistical classifier. The J48 is based on the C4.5 algorithm [13]. The C4.5 builds decision trees from a set of training data using the concept of information entropy. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is based on difference entropy. The attribute with the highest difference of entropy is chosen to make the decision. The Hoeffding Tree [14] algorithm can learn from large data streams assuming that the distribution generating examples does not change over time. This algorithm works well with small samples, as it uses the Hoeffding bound which computes the number of observations that are necessary to estimate statistics values within a prescribed precision. Random Tree [15] is a supervised classifier. It employs a bagging idea to produce a random set of data for constructing a decision tree. In standard tree each node is split using the best split among all variables. In a random tree, each node is split using the best among the subset of predicators randomly chosen at that node. This algorithm can be used both classification and regression problems. In the case of use as a classifier, the random trees take the input feature vector, classifies it with every tree in the forest, and outputs the largest class. While in case of a regression, the classifier response is the average of the responses over all the trees in the forest. Random Forests improve the performance of decision trees considerably. It creates a balanced tree where one global setting for the ridge value works across all leaves, thus simplifying the

optimization procedure.REPTree algorithm [16], uses the regression tree logic to creates multiple trees in different iterations. To build the trees it uses the information gain as the splitting criterion, and prunes it using reduced error pruning. REPTTree can be consider as a fast decision tree learner and it is based on C4.5.

B. Support Vector Machines (SVM)

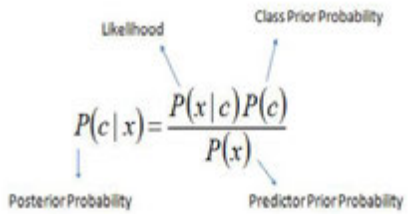
SVM [17] is a linear model for regression, classification and pattern recognition. It can solve linear and non-linear problems and work well for various practical problems. The idea of SVM is simple: The algorithm creates a hyperplane which separates the data into classes. This algorithm takes the data as an input and outputs a line that separates those classes if possible. The SVM make a decision in such a way that the separation between the classes is as wide as possible. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high

C. Logistic Regression

Logistic regression is a generalization of linear regression [18]. It is a statistical model that uses a logistic function to model a binary dependent variable, or multi-class dependent variables although many more complex extensions exist. It is used to classify the low dimensional data having no linear boundaries. It also provides the difference in the percentage of dependent variable and provides the rank of individual variable according to its importance. So, the main aim of Logistic Regression is to determine the result of each variable correctly.

D. Naive Bayes and BayesNet Classifiers

NB is considered as a statistic classifier [19]. It can predict class membership based on probabilities. In statistics everything revolves about hypotheses. First is made a hypothesis and then are collected evidence to test that hypothesis. To calculate the probability of a hypothesis the formula as follow is used:



$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Where:

- P(c/x) is the posterior probability of class c given predictor.
- P(c) is the probability of class.
- P(x|c) is the like lihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor

NB can be used in real time prediction also is known for multi class prediction feature. It can predict the probability of multiple classes of target variable. Whereas, the Bayes Net is a probabilistic graphical model that represents a set of variables and their conditional dependencies through a directed acyclic graph [20]. The Bayes Net' graphical model is comprised of nodes and links, where each node represents a random variable and each link describes the probabilistic dependency between two variables.

IV. METHODOLOGY

This section defines the analytical methods to achieve the set objectives by providing clarification and rationalization of the research design, sampling, data collection and analysis.

A. Research design

This research comprises a quantitative data collection where numerical and standardized data were collected to relationships and trends. The dataset that is analyzed consists of 125000 instances and 10 attribute such as: age, education, employment category, length of service, residence, civil status, distance from workplace, suffering from any disease, job satisfaction and leadership style. Based on this dataset are analyzed various ML algorithms. The environment where the experiments were performed is Weka tool.

B. Target Population and sampling

The study population is mainly targeted to staff separated in three operational level where different jobs category where been selected on a balance of percentage. The target population was organized in three groups namely; Operator, Supervisor, Technical and Engineer. The categories and group number of participants (N) are indicated below. Table 1 represent the characteristics of target population.

TABLE 1. CHARASTERISTICS OF TARGET POPULATION

Group	Category	N	Group
1	Level 1	166	Technical and Engineer
2	Level 2	167	Supervisor
3	Level 3	167	Operator
	Total	500	

A stratified random sampling technique has been chosen to increase the representation of the population. Respondents were randomly selected from each department/section and each respondent has got the equal chance to participate in the selection process.

C. Data Collection Methods

Both secondary and primary data have been considered as useful for this research. Secondary data was used for background understanding of the types and reasons of absenteeism, whereas primary data was collected through survey questionnaire for analyzing attributes that can be affect the absenteeism.

I. Secondary Data

As secondary data we have classified the following documentation; unauthorised absences report, medical survey report, absenteeism follow-up record and absence report. Table 2 summarizes the collection of secondary data.

TABLE 2. SECONDARY DATA COLLECTED

Sources	Purpose
Unauthorised absences report	Assess the unauthorised absences
Medical survey report	Type of sick leave
Absenteeism follow-up record	To analyses the period of sick leaves
Summary absenteeism & missing clocking report	To analyses the percentage of sick leave/clocking report
Absences report	Type of absences

II. Primary data

Survey questionnaire has been designed to collect first-hand data in order to design the attributes of dataset to meet the objectives of this paper. The survey questionnaires were divided into two parts; the first part assessed the demographic data such as age, educational background and the basic working conditions data and the second part addressed the level of job satisfaction and leadership style.

V. RESULT AND ANALYSIS

In this section are presented the experimental results. The experiments are performed in Weka tool and it is used the 10-fold cross-validation procedure [21], which means the dataset is separated into 10 subsets known as folds. Cross-validation is a technique for evaluating machine learning (ML) models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. The DM algorithm uses the 19 subset as the training set and uses 1 subset as a test set to measure the performance of the algorithm. This process is repeated 20 times and as a result the average is obtained. Depending on the size of our dataset k=20 is selected. The dataset is composed by 125000 instances that includes working hours for 500 employees during a year.

A. Assessment Criteria

The evaluation and comparison of algorithms is based in terms of accuracy, recall and precision as in [22].

I. Precision –It is the number of correct positive results divided by the number of positive results predicted by the classifier. Mathematically, it can be expressed as:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (1)$$

II. Recall-It is the number of correct positive results divided by the number of all relevant samples that should have been identified as positive. Mathematically it can be expressed as:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (2)$$

III. F-Measure-It is the harmonic mean between precision and recall. It means how precise is the classifier, how many instances it classifies correctly, as well as how robust it is. It tries to find the balance between precision and recall. The greater the F-measure, the better is the performance of model. Mathematically it can be expressed as:

$$\text{Fmeasure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (3)$$

- True positive (TP): correct positive prediction
- False positive (FP): incorrect positive prediction
- True negative (TN): correct negative prediction
- False negative (FN): incorrect negative prediction

Data preprocessing is the first step in the data mining process. It includes preparation, cleaning, normalization and transformation of data [23]. Fig.2 shows the visualization results from the preprocessing phase and specifically represents the distribution of absenteeism according to the "education" attribute.

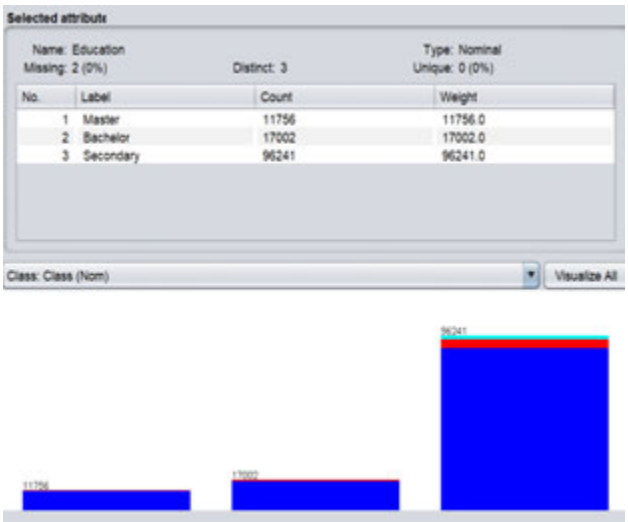


Figure 2. Visualize results of dataset in preprocessing phase.

During this phase the distribution of the absenteeism according to the dataset attributes occurs. After preprocessing phase is built the model and the results for each analyzed algorithm are announced. The algorithms we have analyzed are: BayesNet, Naïve Bayes, Logistic, J48, Random Tree, REPTTree and Hoeffding Tree. Table 3 presents the results of these algorithms.

TABLE 3. RESULTS OF ALGORITHMS

Algorithms	Correctly Classified Instances	Incorrectly Classified Instances	Recall	Precision	F-measure
Bayes Net	99.81%	0.09%	0.999	0.999	0.999
Naïve Bayes	86.08%	13.91%	0.861	0.82	0.839
Logistic	88.94%	11.05%	0.889	0.868	0.874
J48	99.97%	0.02%	1	1	1
Random Tree	99.61%	0.38%	0.996	0.996	0.996
REPTTree	99.96%	0.03%	1	1	1
HoeffdingTree	93.46%	6.53%	0.935	0.933	0.93

As can be shown from the results in table 3, all analyzed algorithms give acceptable results. Can be selected the Bayes Net, J48, RandomTree and REPTTTree algorithms that have very high accuracy and classify more than 99% of instances. The Bayes Net algorithm compared to other algorithms it is simpler to present results. Fig 3, shows the visualize graph of Bayes Net algorithm. This algorithm has classified the absenteeism occurred based on all dataset attributes: age, level of education, day, month, length of service, civil status, distance from workplace, job satisfaction, employment category and leadership style.

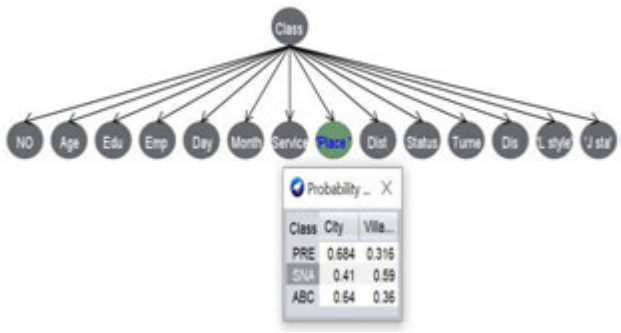


Figure 3. Visualize graph of Bayes Net Algorithm

After building the model and evaluating the data, referring to the Bayes Net algorithm results is concluded that: employees who are between 20 -30 and 60-70 years old tend to have more absences at work. Regarding to the level of education, employees with secondary school have a higher number of absences. Also, employees who live in the village, married, have a distance of 20 km from the workplace, work with turn, suffer from a disease tend to absence more at work. Regarding to the length of service, employees over 40 years of experience tend to absence more at work. While regarding to leadership style and job satisfaction, employees that are expressed "neutral" tend to absence more at work. Regarding the type of absences, it is noticed that employees that have secondary education are over 50 years old, live in city, work with turn, are part of level 1 employ category tend to make more unauthorized absences.

VI. CONCLUSION

Absenteeism at workplace is continuously causing disruption in the smooth running of business, affecting the organizational performance and productivity and impacting on the employees' morale. The focus of this research was to identify and evaluate the appropriate ML algorithms to predict and analyses absenteeism at workplace. The analysis is focused on data obtained from the ARMO company operating in Albania. The attributes that have been studied are: age, education, employment category, length of service, residence, civil status, distance from workplace, diseases, job satisfaction and leadership style. By analyzing and comparing various ML classifier algorithms in terms of precision, sensitivity and accuracy, it is concluded that Bayes Net, J48, Random Tree and REPTTTree algorithms have very high accuracy and classify more than 99% of instances. The Bayes Net algorithms compared to other algorithms it is simpler to present results. Usage of DM and ML techniques s is a good solution to predict and analyze the absenteeism at workplace. These techniques give a significant contribution in human resource department of an organization. The management can get a sense of understanding of employee activities and behavior which can eventually help in crucial decision making on both monitoring employees at work and recruiting potential employees. This research could be used as a relevant instrument for further investigations in the future at ARMO or other companies.

REFERENCES

[1] David L Olson, Dursun Delen,"Advanced Data Mining Techniques" Springer 2015.

[2] W. James Murdocha,Chandan Singhb,Karl Kumbiera, Reza Abbasi-Aslband Bin Yua "Interpretable machine learning: definitions,methods, and applications",PNAS 2019

[3] Bhatnagar V. and Gupta S.K. Modeling the KDD Process. Encyclopedia of Data Warehousing and Mining, Second Edition. Information Science Reference, Hershey, New York, f.1337 – 1344, 2008

[4] Usama, M.Fayyad, et al., Advances in Knowledge Discovery and Data Mining Cambridge, Mass.MIT Press, 1996

[5] Zaman Wahid ,Zaman Wahid,A K M Zaidi Satter ,A. K. M. Zaidi Satter,Abdullah Al Imran,Abdullah Al Imran,Touhid Bhuiyan,Touhid Bhuiyan "Predicting Absenteeism at Work Using Tree-Based Learners",Proceedings of the 3rd International Conference on Machine Learning and Soft Computing,ACM 2019

[6] Evandro Lopes de OliveiraJosé M. Torre,"Absenteeism Prediction in Call Center Using Machine Learning Algorithms", Springer 2019

[7] Syed Atif Ali Shah, Irfan Uddin , Furqan Aziz , Shafiq Ahmad ,Mahmoud Ahmad Al-Khasawneh,and Mohamed Sharaf "An Enhanced Deep Neural Network for Predicting Workplace Absenteeism",Hindawi ComplexityVolume 2020, Article ID 5843932, 12 pages

[8] Ricardo Pinto Ferreira., Andréa Martiniano., Domingos Napolitano, Edquel Bueno Prado Farias and Renato José Sassi"A rtificial Neural Network and Their Application in the Perdiction of Absenteeism at Work ",International Journal of Recent Scientific Research Vol. 9, Issue, 1(G), pp. 23332-23334, January, 2018

[9] Kagan Dogruyol,Boran Sekeroglu, "Absenteeism Prediction: A Comparative Study Using Machine Learning Models",10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions 2020

[10] Krittika Tewari,Shriya Vandita,Shruti Jai,"Predictive Analysis of Absenteeism in MNCS Using Machine Learning Algorithm" Springer 2020

[11] Thair Nu Phyu,"Survey of Classification Techniques in Data Mining",Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol IMECS 2009, March 18 - 20, 2009, Hong Kong

[12] Himani Sharma,Sunil Kumar "A Survey on Decision Tree Algorithms ofClassification in Data Mining",International Journal of Science and Research (IJSR)

[13] Quinlan, J. R.: C4. 5: programs for machine learning. Elsevier (2014).

[14] Hamoud, A. K.: Selection of Best Decision Tree Algorithm for Prediction and Classification of Students Action. American International Journal of Research in Science, Technology, Engineering and Mathematics 16(1), 26–32 (2016).

[15] Ian H. Witten, Eibe Frank & Mark A. Hall., "Data Mining Practical Machine Learning Tools and Techniques, Third Edition." Morgan Kaufmann Publishers is an imprint of Elsevier.

[16] [Ian H. Witten, Eibe Frank & Mark A. Hall., "Data Mining Practical Machine Learning Tools and Techniques, Third Edition." Morgan Kaufmann Publishers is an imprint of Elsevier

[17] Janmenjoy Nayak, Bighnaraj Naik and H. S. Behera,"A Comprehensive Survey on Support Vector Machine in DataMining Tasks: Applications & Challenges",International Journal of Database Theory and Application Vol.8, No.1 (2015), pp.169-186

[18] Charu C. Aggarwal, "Data classification: algorithms and applications,ACM July 2014

[19] Raj Kumar,Rajesh Verma,"Classification Algorithms for Data Mining:A Survey",International Journal of Innovations in Engineering and Technology (IJIET)

[20] Remco R. Bouckaert "Bayesian Network Classifiers in Weka",University of Waikato, 2004

[21] Tzu-TsungWong,"Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation", Elsevier 2015

[22] https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234 ,accessed 27/04/2020

[23] García, Salvador, Luengo, Julian, Herrera, Francisco "Data Preprocessing in DataMining" Springer 2015

Example of Using Algorithms for Switching Hybrid FSO/RF Systems

R. Haluška*, Ľ. Ovseník* and P. Šuľaj*
* Technical University of Košice Dept. of Electronics and Multimedia Communications, Košice, Slovakia
renat.haluska@tuke.sk, lubos.ovsenik@tuke.sk, peter.sulaj@tuke.sk

***Abstract* - This article discusses the problem of optical communication with Free Space Optics (FSO), its use, and the conditions that affect this link. The article deals with the design and construction of a monitoring system. This system is designed to collect and process data characterizing the nature of the conditions along the transmission path. Because of the vulnerability of the FSO transmission channel to weather conditions, it is necessary to predict the strength of the received signal and switch to the back-up line by machine learning based on decision trees.**

Keywords - component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Demand for high bandwidth in metropolitan networks with short access times is increasing. In addition, the requirements of flexibility and cost-effectiveness of service provision cause imbalances. This issue is often referred to as the "last kilometer obstacle".

Optical fiber is the first most visible way of addressing the lack of bandwidth. Fiber is undoubtedly the most reliable means of optical communication, but digging, fiber storage costs and time to market are the most serious disadvantages of optical fiber. The second option is communication via radio frequency (RF) [1]. This technology is advanced and often deployed today. Radio-based networks require huge investments to obtain spectrum licenses, but they cannot be compared to fiber-optic transmission capacity. The third alternative is Free Space Optics (FSO) communication. FSO is the optimal solution in terms of technology, bandwidth scalability, deployment speed and cost-effectiveness [2].

The main requirement for the operation of fiber-free optical systems is the seamless mutual visibility between two communication points, as FSO systems use light for communication and this cannot pass through solid obstacles such as walls, buildings, trees and others [3]. The communication scheme of the FSO system is shown in Fig. 1.



Fig. 1 FSO communication

Free Space Optics (FSO) communication involves the transmission, absorption and scattering of light through the Earth's atmosphere. The atmosphere interacts with light due to the composition of the atmosphere, which normally consists of molecules of various gases and small suspended particles called aerosols [4]. This interaction delivers a wide range of optical phenomena:

- Selective mitigation of radiation that is propagated in the atmosphere.
- Absorption at specific optical wavelengths due to molecules.
- Generation by distraction (blue sky, red sunset, etc.)
- Radiation emission of the optical beam.
- Scintillation action due to air refractive index change [5].

Hybrid Wireless Free Space Communication (FSO) and Radio Frequency Communication (RF) (Fig. 2) is a way to ensure reliable communication for critical real-time outdoor traffic as weather like fog affects FSO much more than RF connections [6]. The main condition for using the FSO is that the receiver must be in line of sight with the transmitter [7].

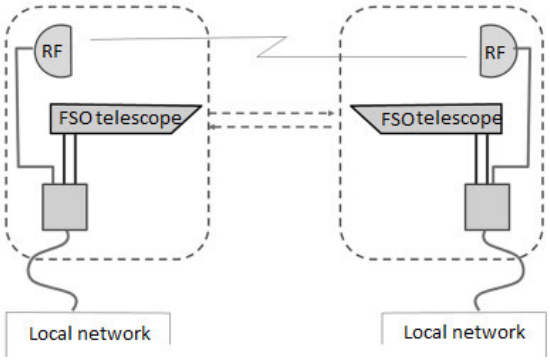


Fig. 2 Hybrid FSO/RF system

Hybrid FSO/RF systems are capable of providing high availability even under adverse weather conditions [8]. The cause of the RF signal degradation in the high-frequency RF transmission is rain, as the carrier wavelength is almost identical to the rain droplet size, and FSO communication is greatly affected by fog. The availability of the whole system can be improved by using a low data rate connection as a back-up line when the FSO line cannot be used [9].

II. COLLECTION OF WHEATHER DATA FOR FSO/RF LINE SWITCHNIG

An important task for the operation of a hybrid FSO /RF system is to monitor weather conditions along the atmospheric channel using sensors such as a temperature, humidity, wind speed sensor and more [10]. The measurement results are used as inputs for machine learning methods, for processing and prediction of received optical power, which is the implementation of Received Signal Strength Indicator (RSSI) parameter in hybrid FSO/RF line switching (Fig. 3).

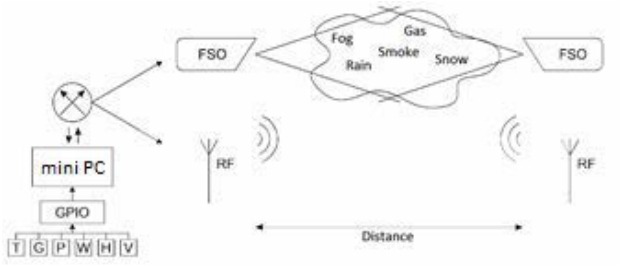


Fig. 3 Scheme of the Hybrid FSO/RF System

Computational performance and open operating system are suitable features of a minicomputer to be implemented in an experimental system to analyze the availability and reliability of FSO systems [11]. A substantial part of the selected minicomputer is the GPIO bus, which is used to connect analog and digital sensors.

The first head from the used steam FSO line located in the area of the Technical University of Košice (TUKE) is located on the roof of the building marked PK13. The second FSO warhead is located on the roof of the main building TUKE (Letna 9) shown in fig. Distance between FSO heads is around 230 meters.



Fig. 4 Localization of FSO beams

The main components of the software system to increase the availability and reliability of hybrid FSO / RF systems include, in addition to control scripts and MySQL, as well as the Grafana web server, which provides the ability to interact with instantaneous physical values (Fig. 4). A central database was created to back up the measured values. The replication of records between the Master and Slave databases takes place in real time, ensuring redundancy of the data obtained.

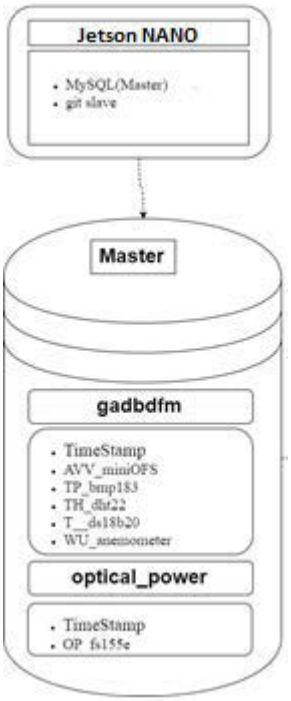


Fig. 5 Organization of database system

Periodic execution of individual operations is controlled by the Linux command crontab. The selected device is characterized by very low power consumption and has compact dimensions, which corresponds to the requirement of miniaturization experiment monitoring station (Fig. 6).



Fig. 6 Monitoring station

III. PREDICTION OF RSSI PARAMETER

Machine learning techniques are widely used today for many different tasks. Different data types require different methods. The task of machine learning in a hybrid FSO/RF system is to predict the strength of the received optical signal, which is affected by weather conditions [12]. A form of machine learning to the gradient boosting on the decision tree, which works by gradually training more complex models, is designed to maximize the accuracy of predictions. Gradient enhancement is particularly useful for predictive models that analyze organized data and categorical data. Decision trees are used for RSSI classification prediction.

A decision tree is a classifier expressed as a recursive part of an instance space. A decision tree consists of nodes that form a rooted tree, which means that it is a directional tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves [13]. In the decision tree, each inner node divides the instance space into two or more subspaces according to a certain discrete function of the input attribute values [14]. In the simplest and most common case, each test considers one attribute, so the instance space is divided by the value of the attributes. For numeric attributes, the status refers to a range.

In the classification problem, use different metrics as a criterion for tree splitting. One option is an index that expresses the measure of total variance between classes *K*. When creating a decision tree, the input and training set is divided into smaller subsets, which gradually characterize the values of the output variables. The recursive division process is performed until the termination condition is met. In the process of atmospheric channel analysis for the FSO/RF system, a relatively extensive system of cases of input variables *X* with the corresponding mute output stages was designed. variables *y*. The output variable y is RSSI in this case. The general training set of cases has the following structure.

$$P = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} & y_1 \\ x_{21} & x_{22} & \dots & x_{2k} & y_2 \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} & y_n \end{bmatrix} \quad (1)$$

The input and training set is split to create a decision tree smaller subsets that gradually characterize the values of the output variables. The training matrix of the input variables for the FSO/RF system has the following structure.

$$P = \begin{bmatrix} P_1 & T_1 & G_1 & V_n & H_n & \cdots & W_1 \\ P_2 & T_2 & G_2 & V_n & H_n & \cdots & W_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ P_n & T_n & G_n & V_n & H_n & \cdots & W_n \end{bmatrix} \quad (2)$$

where P is barometric pressure (hPa), H represents ambient humidity (%), T is air temperature (°C), wind speed W (m/s), G represents concentration Airborne particulate matter (mg/m3) and visibility (m). The matrix of the output variable *y* (target), which represents the received optical power of RSSI, is interpreted as follows:

$$y = [y_{RSSI.1} \quad y_{RSSI.2} \quad \cdots \quad y_{RSSI.n}]^T \quad (3)$$

Training set is used to create a description that can be used to predict previously unseen examples. Attributes (sometimes called fields, variables, or functions) are typically one of two types: nominal (values are members of an unordered set) or numeric (values are real numbers). Attribute *a*, it is useful to mark its domain values as:

$$dom(a_i) = \{v_{i,1}, v_{i,2}, \dots, v_{i,|dom(a_i|}\} \quad (4)$$

Instantial space (a set of all possible examples) is defined as the Cartesian product of all input attribute domains:

$$X = dom(a_1) \times dom(a_2) \times \dots \times dom(a_n) \quad (5)$$

Universal instance space *U* is defined as the Cartesian product of all inbound attribute domains and target attribute domains:

$$U = X \times dom(y) \quad (6)$$

A training set is an instance consisting of a set of m-tics. Formally, the training file is marked as $S(B) = (\langle x_1, y_1 \rangle, \dots, \langle x_m, y_m \rangle)$ where $x_q \in X$ and $y_q \in dom(y)$ [15]. It is usually assumed that the training sets are generated randomly and independently according to some fixed and unknown common probability distributions of *D* above *U*. Generally it is a generalization of a deterministic case when the supervisor classifies the set using the function *y = f (x)* [7].

To create a prediction *w* using a node in a tree that contains a set of instances, the weight of the instance in the current tree is predicted:

$$MSE(a, X) = \frac{1}{l}\sum_{i=1}^{l}(a(x)_i - (y_i)^2 \quad (7)$$

Overwritten in terms of residues:

$$MSE(a, X) = \frac{1}{l}\sum_{i=1}^{l}(r_i^2 - 2r_i w + w^2) \quad (8)$$

After simplifying the substitution for $\sum_{i=1}^{l} r \ as \ R$:

$$MSE(a, X) = \frac{1}{l}\sum_{i=1}^{l} r_i^2 - Rw + \frac{1}{l}nw^2 \quad (9)$$

By substituting for the loss function for the simultaneous iteration and simplifying, we get the effect of predicting *w** in the given leaf:

$$MSE(a, X) = \frac{1}{2}\sum_{i=1}^{l} r_i^2 - \frac{1}{l}\frac{R^2}{n} \quad (10)$$

As a basic algorithm it can be usually choose the simplest algorithm, for example short decision tree:

$$b_1(x) = argmin_b \frac{1}{l}\sum_{i=1}^{l}(b(x_i) - y_i)^2 \qquad (11)$$

The second algorithm must be designed to minimize system error:

$$b_2(x) = argmin_b \frac{1}{l}\sum_{i=1}^{l}(b(x_i) - (y_i - b_1(x))^2 \quad (12)$$

Then it is possible to derive the formula $b_N(x)$:

$$b_N(x) = argmin_b \frac{1}{l}\sum_{i=1}^{l}\left[b(x_i) - \left(y_i - \sum_{n=1}^{N-1} b_n(x_i)\right)\right]^2 \quad (13)$$

One of the most popular opensource libraries for machine learning in the Python programming language is Scikit-learn. The scikit-learn [17] library contains a *Decision-Tree-Classifier* class that can train a binary decision tree with entropic impurities. The prediction of RSSI parameter that can be used to switch the FSO line and backup RF line is implemented by Algorithm 1.

Algorithm 1 RSSI prediction

Input: Data in a simple format
Input: Start year
Input: End year
Input: Number of iteration

Output: Score of prediction.

Format the data for use by scikit-learn library
classifier ← *decision tree classifier*
trainings ← *training data file*
tests ← *tests data file*

for each in *iteration* **do**
  expected ← Clasification(test)
  actual ← Prediction(*classifier, trainings, tests*)
  score ← Getscore(*expected, actual*)
end for

Print *score*
Show *data plot*

Determination coefficient or score determines the probability of how well the learning model will predict future samples of the output variable *y*. The best achievable score value is 1. The max depth parameter is selected based on optimum MSE and score values. For the selected algorithm, the max depth parameter was gradually tested from value 5 with step 1 to value 300. Fig. 7 shows the development of the MSE and score depending on the max depth. The cyclic training model showed that the optimal value of the parameter max depth for MSE = 0.94.

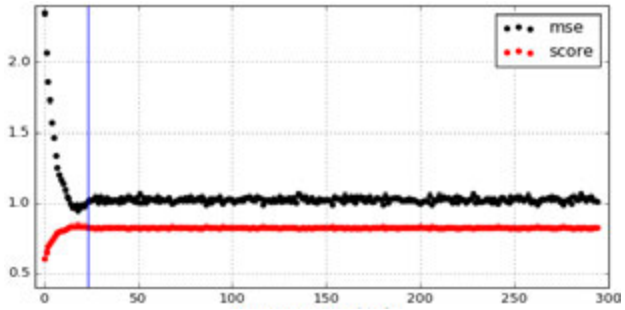


Fig. 7 Max depth parameter

The comparison of predicted and real RSSI values is shown in Fig. 8.

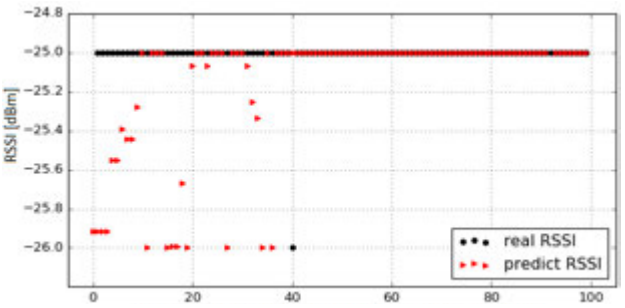


Fig. 8 measured and predicted data from the input matrix

IV. CONLUSION

Due to the high impact of weather on FSO transmission, a backup RF line is required, which is less prone to, but much slower. For this reason, a hybrid FSO/RF line is required, which can be switched to the hard or soft switching principle. To create such a system, it is necessary to analyze the conditions and thus obtain the characteristics of the variables that affect the transmission channel. Appropriate weather prediction, using machine learning methods, makes it possible to quickly switch over communications over the RF line and back to a faster FSO line. In this way, it is possible to increase the efficiency of communication via the hybrid FSO/RF system.

For efficiency, the tree can also be pruned using the max depth parameter. However, it is not always easy to find out which value is best to do a grid search or cross validation. Of course, it is extremely important to avoid over-specialization in the training set, which could lead to a model learning, so it is often necessary to carve a tree at a certain level. In this way, it is easier to find a good compromise between training and accuracy. If the model is not too large, it is possible to manually test different values and select the optimum value.

ACKNOWLEDGMENT

This work was supported by the Slovak Research and Development Agency under the contract no. "APVV-17-0208 - Resilient mobile networks for content delivery".

REFERENCES

[1] V. Brazda, O. Fiser, and J. Svoboda, "FSO and radio link attenuation: meteorological models verified by experiment," in *Free-Space and Atmospheric Laser Communications XI*, A. K. Majumdar and C. C. Davis, Eds. SPIE-Intl Soc Optical Eng, sep 2011.

[2] P. Singal, S. Rai, R. Punia, and D. Kashyap, "Comparison of different transmitters using 1550nm and 10000nm in FSO communication systems," *International Journal of Computer Science and Information Technology*, vol. 7, no. 3, pp. 107–113, jun 2015

[3] S. A. Al-Gailani, A. B. Mohammad, R. Q. Shaddad, and M. Y. Jamaludin, "Single and multiple transceiver simulation modules for free-space optical channel in tropical malaysian weather," in 2013 *IEEE Business Engineering and Industrial Applications Colloquium (BEIAC)*, April 2013, pp. 613–616.

[4] J. Tóth, "High availability and reliability in wireless optics using data analytics techniques" - 2017.In: *SCYR 2017*. - Košice : TU, 2017 S. 118-119. - ISBN 978-80-553-3162-1

[5] M. L. B. Riediger, R. Schober, and L. Lampe, "Blind detection of on-off keying for free-space optical communications," in *2008 Canadian Conference on Electrical and Computer Engineering*, May 2008, pp. 001 361–001 364

[6] F. Nadeem, V. Kvicera, M. Awan, E. Leitgeb, S. Muhammad, and G. Kandus, "Weather effects on hybrid FSO/RF communication link, " *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 9, pp. 1687–1697, dec 2009 [Online]. Available: https://doi.org/10.1109%2Fjsac.2009.091218

[7] T. Hastie and J. M. Chambers. "Statistical models." *Statistical Models in S*. Routledge, 2017. 13-44.

[8] H. Gavin. *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd, 2017.

[9] M. Brzoska, "Modellierung zwischen „overfitting' und „underfitting'," *Rationale Entscheidungen unter Unsicherheit*, pp. 152–156, 2018.

[10] G. Mittag and S. Möller, "Detecting Packet-Loss Concealment Using Formant Features and Decision Tree Learning," *Interspeech 2018*, Feb. 2018.

[11] X. Zhong and W. Li, "A Decision-Tree-Based Multi-objective Estimation of Distribution Algorithm," *2007 International Conference on Computational Intelligence and Security (CIS 2007)*, 2007.

[12] L. Rokach and O. Maimon, "Top-Down Induction of Decision Trees Classifiers—A Survey," *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 4, pp. 476–487, 2005.

[13] O. Maimon and L. Rokach, "Introduction to Supervised Methods," *Data Mining and Knowledge Discovery Handbook*, pp. 149–164.

[14] A. S.bhatlavande and A. A. Phatak, "Data Aggregation Techniques in Wireless Sensor Networks: Literature Survey," *International Journal of Computer Applications*, vol. 115, no. 10, pp. 21–24

[15] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," *Biomedical Signal Processing and Control*, vol. 52, pp. 456–462, 2019.

[16] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, "Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm," *Procedia Technology*, vol. 25, pp. 326–332, 2016.

[17] F. J. Brims, T. M. Meniawy, I. Duffus, D. D. Fonseka, A. Segal, J. Creaney, N. Maskell, R. A. Lake, N. D. Klerk, and A. K. Nowak, "A Novel Clinical Prediction Model for Prognosis in Malignant Pleural Mesothelioma Using Decision Tree Analysis," *Journal of Thoracic Oncology*, vol. 11, no. 4, pp. 573–582, 2016.

Crime Analysis and Prediction using Machine Learning

Olta Llaha

South East European University/Faculty of Contemporary Sciences and Technologies, Tetovo, North Macedonia
E-mail: ol29064@seeu.edu.mk

***Abstract* - Data mining and machine learning have become a vital part of crime detection and prevention. The purpose of this paper is to evaluate data mining methods and their performances that can be used for analyzing the collected data about the past crimes. I identified the most appropriate data mining methods to analyze the collected data from sources specialized in crime prevention by comparing them theoretically and practically. Some attributes of this dataset are, gender, age, employment status, crime place. Methods are applied on these data to determine their effectiveness in analyzing and preventing crime. Evaluations on the data showed that the method with a higher performance is "Decision Tree". This was achieved by some performance measures, such as the number of instances correctly classified, accuracy or precision and recall, that has brought better results compared to other methods. I come to the conclusion that the data mining methods contribute to the predictions on the possibility of occurrence of the crime and as a result in its prevention.**

Keywords - Machine Learning, Prediction, Crime Analysis, Data Mining

I. INTRODUCTION

The increase in crime data recording coupled with data analytics resulted in the growth of research approaches aimed at extracting knowledge from crime records to better understand criminal behavior and ultimately prevent future crimes.

Crime is a complex social phenomenon that has grown due to major changes in society. Law enforcement agencies need to learn the factors that lead to an increase in crime tendency. To curb this, there is always a need for strategies and policies to prevent crime. As a result of technology development, science and information, data mining and artificial intelligence tools are increasingly prevalent in the law enforcement community.

Law enforcement agencies face a large volume of data that needs to be processed and turned into useful information, and data mining can improve crime analysis by helping to predict and prevent it. By processing criminal data, law enforcement agencies can use models that may be important in the crime prevention process.

The use of data mining accelerates data analysis, and analysts can examine existing data to identify patterns and trends of crime. This paper is structured as follows: Section. 2 describes the relationship that exists between data mining, machine learning and criminology. The methodology and description of the dataset are described in Section. 3. Sections. 4 and 5, represent a theoretical description of the methods and algorithms that will be applied practically to our data. Section 6 presents the results of the application of algorithms and an explanation for the algorithm with the best results. In sect. 7 the conclusions and future work are discussed.

II. USING DATA MINING AND MACHINE LEARNING IN CRIMINOLOGY

Criminology is an area where the scientific study of crime and criminal behavior focuses. This is one of the most important areas when applying data mining techniques that can produce significant results [1].

Crime analysis, as part of criminology, is tasked with exploring and discovering crime and its relationship with criminals. Law enforcement is a process that aims to identify the characteristics of crime. Identifying crime characteristics is the first step in developing further analysis. The high volume of crime data and the complexity of the relationships between them have made criminology an appropriate field for applying data mining techniques [2].

Data mining can be used to examine many large datasets involving a large set of variables beyond what a single analyst, or even an analytical team or task force, can consider correct, whereas machine learning uses neural networks, predictive model and automated algorithms to make the decisions. Like any other problem solving method, the task of data mining begins with a problem definition. The identification of the data mining problem enables the determination of the data mining process and

the modeling technique. Machine learning is a subfield of data science that deals with algorithms able to learn from data and make accurate predictions [3]. Data mining gives law enforcement agencies the opportunity to learn about crime trends, how and why crimes are committed. Using data mining methods and machine learning improves crime analysis and help reduce and prevent crime.

III. DATA AND METHODOLOGY

I compare theoretically and practically data mining methods to discover the most appropriate method for our data. The methods were compared by applying machine learning algorithms to concrete data in the WEKA "Waikato Environment for Knowledge Analysis" [4] environment. The implemented algorithms are: Simple Logistic, Logistic, Multilayer Perceptron, Naive Bayes, Bayes Net, SMO, C4.5.In data collection step I am collecting data from law enforcement agencies. The collected data is stored into database for further process. They relate to the areas where crime and perpetrator data occur.

The dataset is made up of 100 records or instances.

The name of the dataset	Number of examples	Number of input attributes	Number of possible classes	Total number of attributes	Values that are missing
Crime Data	100	6	2	7	0

Table 1. Dataset details

The variables or attributes of this dataset are: age (from 17 to 55 years old), gender, education (middle school. high school, university) employment status (whether employed or not), civil status (whether married, single, or divorced), the area where the crime occurred (urban or rural) and whether the person who committed the crime was previously convicted or not. Crime dataset is in CSV format.

IV. CLASSIFICATION METHODS

Classification is a data mining technique that categorizes data in order to assist in more accurate predictions and analyzes [5, 6]. It is one of the data mining methods that aims to analyze very large datasets. It is used to derive patterns that accurately define the important data classes within the data set. Classification consists in predicting a given result based on a given input [6]. Classification algorithms attempt to detect relationships between attributes that would make it possible to predict the result. They analyze the input and produce a prediction.

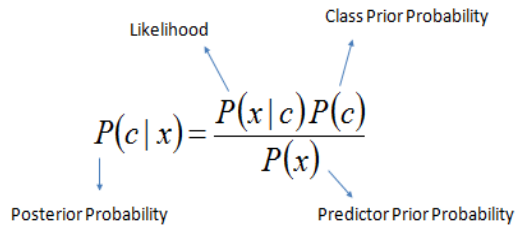
A. Artificial Neural Networks

Neural networks are an area of Artificial Intelligence (AI) based on the inspiration from the human brain. I use them to find data structures and algorithms for learning and classifying data. By applying neural network techniques, a program can learn from the examples and create an internal set of rules for classifying different inputs. Artificial Neural Networks (ANNs) are capable of predicting new observations from existing observations. A neural network consists of interconnected processing elements also called units, nodes, or neurons [5].

All processes of a neural network are performed by this group of neurons or units. Each neuron is a separate communication device, making its operation relatively simple. The function of one unit is simply to receive data from other units, as a function of the inputs it receives to calculate an output value, which it sends to other units. In artificial neural networks, neurons are organized in layers which process information using dynamic state responses to external inputs [6]. The Multilayer Perceptron (MLP) is a feed-forward artificial neural network model that maps sets of input data to a set of appropriate outputs [7]. In a feed-forward neural network, the input signal traverses the neural network in a forward direction from the input layer to the output layer through the hidden layers.

B. Naive Bayes Classifier

Bayesian classification represents a supervised learning method as well as a statistical classification method. It assumes a high-probability underlying model, which allows us to determine in principle the uncertainties for the model, thus determining the probability of the results. The Naive Bayes Classifier technique is based on the Bayesian theorem and is used especially when the dimensionality of the inputs is high [5, 8]. Naive Bayes classifier is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. Bayesian classification provides practical learning algorithms and prior knowledge, here the observed data can be combined. Bayesian classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates the apparent hypothetical probability. The algorithm works as follows. Bayes' theorem offers a way to calculate the probability of a hypothesis based on our prior knowledge.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ is the prior probability of *class*.
- $(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

Class (c) is independent of the values of other predictors. Naïve Bayes Classifier can be trained effectively in supervised learning [8]. After calculating the conditional probability for a different number of hypotheses, I can solve the hypothesis (class) with the highest probability. An advantage of the Naive Bayes classifier is that it requires a small amount of training data to calculate the parameters (mean and variance of the variables) needed for the classification [8]. Because the independent variables are assumed, then only the discrepancies of the variables for each class need to be determined and not the full matrix distribution. The Naive Bayesian classifier is fast and incremental can deal with discrete and continuous attributes, has excellent performance and can explain its decisions.

C. *Support Vector Machine*

Support Vector Machines are based on the concept of decision making plans that set the boundaries of decisions. A decision plan is one that divides a group of objects that have different class memberships. Classification tasks that are based on the dividing lines between different class membership objects are known as hyper-plane Classifiers [9]. SVMs are a set of related supervised learning methods used for classification and regression. Support Vector Machine (SVM) is primarily a classification method that performs classification tasks by constructing hyper-plane in a multidimensional space. The SVM uses statistical learning theory to search for a regularized hypothesis that fits the available data well without over-fitting. SVM also supports regression and classification techniques and can handle multiple continuous and categorical variables [9].

The efficiency of SVM-based classification is not directly dependent on the dimension of the classified entities. SVM can also be extended to learn nonlinear decision functions by first projecting the input data into a high dimensional space using kernel functions and formulating a linear classification problem in that space. SMO (Sequential Minimal Optimization) implements John C. Platt's sequential minimal optimization algorithm for training a Support Vector classifier using polynomial or RBF(Radial Basis Function) kernels [9].This implementation globally replaces all lost values and transforms nominal attributes into binary ones. It can be seen that the choice of kernel function and best value of parameters for particular kernel is critical for a given amount of data. It also normalizes all attributes by default.

D. *The decision tree*

The decision tree is a method in which data is presented in a tree structure based on the values of their attributes. It splits the data in the database into subsets based on the values of one or more fields. This process will be repeated for each subgroup recursively until all instances are a node in a single class. The result of the decision tree is a tree-shaped structure that describes a series of decisions given at each step [5, 6]. These decisions are then considered as rules for the classification task. The algorithms commonly used to construct decision trees are; ID3 and C4.5.

The ID3 (Iterative Dichotomiser 3) algorithm [10] induces classification models, or decision trees, from data. It is a supervised learning algorithm that is trained by examples for different classes. After being trained, the algorithm should be able to predict the class of a new item. ID3 identifies attributes that differentiate one class from another. All attributes must be known in advance, and must also be either continuous or selected from a set of known values. For instance, temperature (continuous), and country of citizenship (set of known values) are valid attributes. To determine which attributes are the most important, ID3 uses the statistical property of entropy [10].

The C4.5 algorithm [11] overcomes this problem by using another statistical property known as information gain. Information gain measures how well a given attribute separates the training sets into the output classes. This algorithm has input in the form of training samples and samples. Training samples in the form of sample data that will be used to build a tree that has been substantiated. C4.5 algorithms are algorithms result of the development of the algorithm ID3 [11]. C4.5 algorithm works by grouping several training sample data that will result in a decision tree based on the facts on the training data.

V. ASSOCIATION RULES AND REGRESSION

Association Rule is one of the most important canonical tasks in data mining and probably one of the most studied techniques for pattern discovery. Association rules are if/then statements that help to uncover relationships between unrelated data in a database, relational database or other information repository [12]. Association rules are used to find the relationships between the objects which are frequently used together [12]. Association Rules identify the arguments found together with a given, event or record: "the presence of one set of arguments brings the presence of another set". This is how rules of type are identified: "if argument A is part of an event, then for a certain probability argument B is also part of the event" [13]. The objective of the association rule was to discover interesting association or correlation relationships among a large set of data items. Support and confidence are the most known measures for the evaluation of association rule interestingness.

While classification provides categorical, discrete labels, regression has continuous function values. So regression is used mainly to predict missing numeric data values rather than discrete class labels. Regression analysis is a statistical methodology often used for numerical prediction, although there are other methods for doing this [14]. Regression also involves identifying the distribution of trends based on available data. For this purpose regression trees can be used as well as decision trees whose nodes have numerical values instead of categorical values. Linear regression is a mathematical technique that can be used to make a numerical data set by creating a mathematical equation [14]. On the other hand logical regression estimates the probability of verifying an event under certain circumstances, using the factors observed together with the occurrence of the event [14].

VI. EXPERIMENTAL RESULTS

To conduct this study I used WEKA [4] software based on the approach and familiarity with its use. WEKA is a collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, association rules, and visualization. It can be used to detect the various hidden patterns in our dataset and find the most determining data factors.

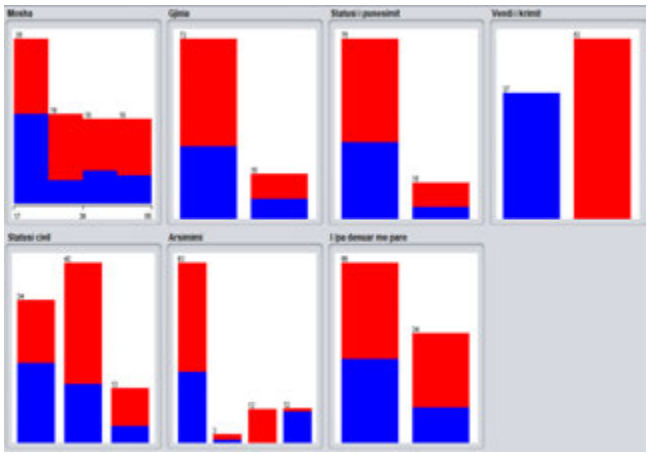


Figure. 1. Pre-processed data visualization

Experiments are done by using cross-validation on default option folds = 10. Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. The process is repeated 10 times for each fold. Performance indicators are given on the following Table 2.

Method	Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Recall	Precision	F-Measure
Regression	Simple Logistic	68 (68%)	32 (32 %)	0.680	0.680	0.665
	Logistic	71 (71%)	29 (29 %)	0.710	0.707	0.707
Bayes Classifier	Naive Bayes	73 (73%)	27 (27%)	0.730	0.737	0.732
	Bayes Net	72 (72%)	28 (28%)	0.720	0.725	0.721
SVM	SMO	67(67%)	33 (33 %)	0.670	0.666	0.666
Decision tree	C4.5	76 (76%)	24 (24 %)	0.760	0.762	0.761
Artificial Neural Network	Multilayer Perceptron	63 (63%)	37 (37%)	0.630	0.637	0.632

Table 2: Comparison of the results of the algorithms applied in WEKA

In this paper I used some algorithms (Table 2) and among them is C4.5 algorithm, which is a Decision Tree algorithm. This algorithm is clear and easy when I used it to interpret the results. The model construction is done by modifying the parameter values and this algorithm classifies crime data with a higher accuracy than other

algorithms of data mining methods. I converted our data to format. The C4.5 algorithm was implemented in this data.

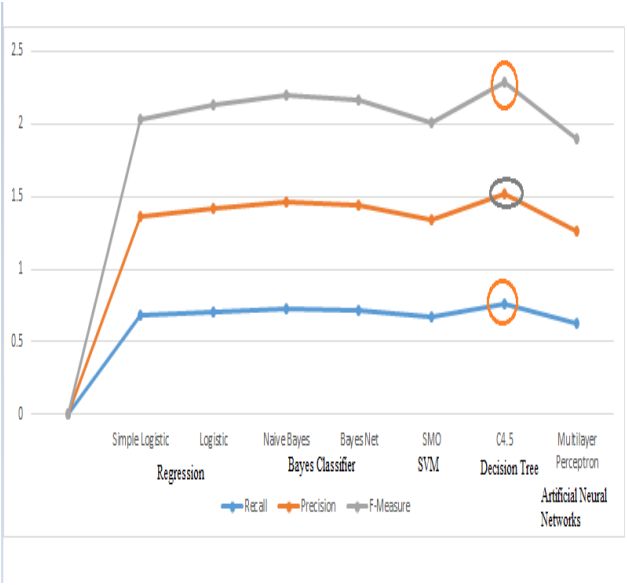


Figure 2: Performance of algorithms

The C4. 5 algorithm for building decision trees is implemented in WEKA as a classifier called J48. J48 has the full name weka.classifiers.trees.J48. What came out of this algorithm: the visualization and the decision tree are presented in Figure 3 and Figure 4.

```
=== Summary ===

Correctly Classified Instances          76               76      %
Incorrectly Classified Instances        24               24      %
Kappa statistic                          0.5106
Mean absolute error                      0.3194
Root mean squared error                  0.3997
Relative absolute error                 65.5331 %
Root relative squared error             80.973  %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.738    0.224      0.705     0.738    0.721      0.511    0.829     0.740     urbane
                 0.776    0.262      0.804     0.776    0.789      0.511    0.829     0.842     rurale
Weighted Avg.    0.760    0.246      0.762     0.760    0.761      0.511    0.829     0.799

=== Confusion Matrix ===

  a  b   <-- classified as
 31 11 |  a = urbane
 13 45 |  b = rurale
```

Figure 3: C4.5 (J48) Classifier

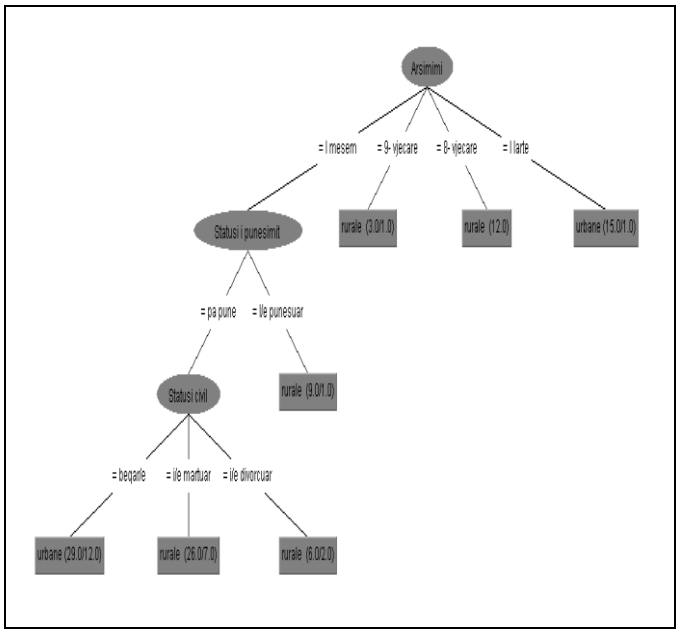


Figure 4: Decision Tree

Figure 3 shows the result of implementing the C4.5 algorithm. It shows that the number of correctly classified instances is 76 with a percentage of 76% and the number of incorrectly classified instances is 24, so 24%.

F-measure is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score: precision is the number of correct positive results divided by the number of all positive results returned by the classifier, and recall is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

The results of this algorithm for recall and precision values are respectively 0.760 (recall) and 0.762 (precision).

$$\text{F-Measure} = \frac{2 * Recall * Precision}{Recall + Precision}$$

- True positive (TP): correct positive prediction
- False positive (FP): incorrect positive prediction
- True negative (TN): correct negative prediction
- False negative (FN): incorrect negative prediction

F-measure after the application of the algorithm has the value 0.761.

The implementation of this algorithm has classified the crime data based on the dataset attributes as e.g. the place where the crime occurred (urban areas, rural areas) where: the number of correctly, classified instances, the accuracy or precision and recall have the highest values compared to other algorithms of data mining methods.

Figure 4 shows the visualization of the decision tree which is generated by the implementation of the C4.5 algorithm. Through the decision tree generated I understand in which areas more crimes occur, as well as the characteristics of the people who committed the crimes. Having this information helps law enforcement agencies to create policies or make decisions about areas where the crime rate is higher.

VII. CONCLUSION AND FUTURE WORK

The purpose of this study is to examine crime analysis through the applicability of data mining methods in the process of crime prediction and prevention. The results of experiments conducted in this research by implementing algorithms of data mining methods have revealed that these methods are applicable in the process of crime prediction. The decision tree as a data mining classification method has classified crime data at an accuracy rate of 76%. This method has shown promising results for the problem of crime prediction as the accuracy rate is high in the experiments performed. Furthermore, the decision tree seems more viable due to the fact that in contrast to other algorithms, it expresses the rules explicitly. These rules can be expressed in human language so that anyone can understand them. The use of machine learning and data mining in crime analysis is important because data mining methods can be used in the decision making process. Decision making is very important in crime prevention in order to decide accurate actions and law enforcement strategies. Through our data analysis law enforcement agencies can create strategies, operating in areas where most crimes occur. In the future extension of this study some models will be created for predicting the crime hot-spots that would help the deployment of police to places of crimes. Algorithms' behavior changes will be looked at when more data is added. I also plan to look into developing social link networks of criminals, suspects and gangs. I also intend to implement this study to an integrated enterprise software that will be created.

REFERENCES

[1] K. Zakir Hussain, M. Durairaj and G. R. J. Farzana, "Criminal behavior analysis by using data mining techniques," IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012), Nagapattinam, Tamil Nadu, 2012, pp. 656-658.

[2] Keyvanpour, Mohammad & Javideh, Mostafa & Ebrahimi, Mohammadreza. (2011). Detecting and investigating crime by means of data mining: A general crime matching framework. Procedia CS. 3. 872-880. 10.1016/j.procs.2010.12.143.

[3] Ioannis Kavakiotis OlgaTsave Athanasios Salifoglou Nicos Maglaveras Ioannis Vlahavas Ioanna Chouvarda, Machine Learning and Data Mining Methods in Diabetes Research, Computational and Structural Biotechnology Journal Volume 15, 2017, Pages 104-116

[4] Frank, Eibe & Hall, Mark & Holmes, Geoffrey & Kirkby, Richard & Pfahringer, Bernhard & Witten, Ian & Trigg, Len. (2010). Weka-A Machine Learning Workbench for Data Mining. 10.1007/978-0-387-09823-4_66.

[5] Pang-Ning Tan; Michael Steinbach; Anuj Karpatne; Vipin Kuma Introduction to Data Mining 2nd ed, Publisher: Pearson, 2019, Print ISBN: 9780133128901, 0133128903 eText ISBN: 9780134080284, 013408028

[6] M. Kantardzic, Data Mining Concepts, Models, Methods, and Algorithms, 2nd ed, John Wiley & Sons, Inc., Hoboken, New Jersey 2011, ISBN 978-0-470-89045-5 , oBook ISBN: 978-1-118-02914-5, ePDF ISBN: 978-1-118-02912-1, ePub ISBN: 978-1-118-02913-8

[7] Ahishakiye, Emmanuel & Opiyo, Elisha & Wario, Ruth & Niyonzima, Ivan. (2017). A Performance Analysis of Business Intelligence Techniques on Crime Prediction. International Journal of Computer and Information Technology. 06. 84 - 90.

[8] Marlina, Leni & Muslim, Muslim & Siahaan, Andysah Putera Utama. (2016). Data Mining Classification Comparison (Naïve Bayes and C4.5 Algorithms). International Journal of Emerging Trends & Technology in Computer Science. 38. 380-383. 10.14445/22315381/IJETT-V38P268.

[9] Himani Bhavsar, Mahesh H. Panchal, (2012). A Review on Support Vector Machine for Data Classification, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 1, Issue 10, December 2012, ISSN: 2278 – 1323.

[10] Xiaohu, Wang & Lele, Wang & Nianfeng, Li. (2012). An Application of Decision Tree Based on ID3. Physics Procedia. 25. 1017-1021. 10.1016/j.phpro.2012.03.193.

[11] Hssina, Badr & MERBOUHA, Abdelkarim & Ezzikouri, Hanane & Erritali, Mohammed. (2014). A comparative study of decision tree ID3 and C4.5. (IJACSA) International Journal of Advanced Computer Science and Applications. Special Issue on Advances in Vehicular Ad Hoc Networking and Applications. 10.14569/SpecialIssue.2014.040203.

[12] Kumbhare, Trupti A. and Santosh V. Chobe. "An Overview of Association Rule Mining Algorithms." (2014).

[13] Chengqi Zhang Shichao Zhang, Association Rule Mining Models and Algorithms, ISSN 0302-9743, ISBN 3-540-43533-6 Springer-Verlag Berlin Heidelberg New York, Springer, 2002

[14] Larose, Daniel T. Data mining methods and models, Published by John Wiley & Sons, Inc., Hoboken, New Jersey, 2006, ISBN-13 978-0-471-66656-1 ISBN-10 0-471-66656-4

A Machine Learning Approach for Analysis of Spectrum Availability in Kosovo based on Experimental Measurements

Zana Limani Fazliu[1], Hëna Maloku[1], Mimoza Ibrani[1], Myzafere Limani[1,2] , Blediona Gashi[1]
[1]Faculty of Electrical and Computer Engineering, University of Prishtina, Prishtina, Kosovo
[2] Academy of Sciences and Arts of Kosovo, Prishtina, Kosovo

Abstract - **Establishing the utilization level of spectrum resources is an essential step in evaluating the true availability of the frequency spectrum which could potentially become available for opportunistic use. Such assessments typically require extensive measurement campaigns, in order take into account various factors which may affect the availability of spectrum, such as frequency allocation at national level, and the particulars of the environment and terrain. In this study, we focus on the bands usually reserved for TV broadcasters, since studies show that these bands tend to be least efficiently utilized. In addition, we take advantage of a large set of data obtained from an extensive spectrum measurement campaign conducted over the entire territory of Kosovo, which enables us to attain a comprehensive picture of spectrum availability in TV bands. Spectrum sensing techniques using energy detectors are applied in order to determine whether the spectrum is utilized. However, due to the significant amount of measurement data collected, machine learning techniques can be leveraged to improve the accuracy of the approach, by improving the detection and false alarm rates. As an end result, a highly accurate spectrum availability map is obtained which can be in turn used by cognitive radio network to opportunistically access the spectrum.**

Keywords – TV bands, wireless technologies, spectrum measurements, machine learning

I. INTRODUCTION

Under-utilization of licensed spectrum, especially in the frequency bands traditionally allocated for TV broadcasting, is a fact that has been well established in the literature [1]. Furthermore, with the transition from analog to digital broadcasting, substantial amount of spectrum in TV bands is expected to become further available. These bands, often referred to as TV White Space (TVWS), are of great interest due to the superior signal propagation characteristics [2]. While, in most countries, these frequencies remain pre-allocated for TV broadcasting, the opportunistic use by secondary devices has been considered, and allowed, most notably in USA [3] and UK [4].

Determining the utilization level of TVWS is an essential step in evaluating the true availability of the frequency spectrum which could potentially become available for opportunistic use. Such assessments typically require extensive measurement campaigns, in order take into account various factors which may affect the availability of spectrum, such as frequency allocation at national level, and the particulars of the environment and terrain.

In this paper we present the results from spectrum measurements over TVWS conducted over a wide area of the territory of Kosovo. To our knowledge, this is the first study of this scale to be conducted in Kosovo, and one of the very few in the Western Balkans region. Similar studies have been performed in Bosnia and Herzegovina [5], Hungary [6], and Macedonia [7]. Preliminary results of this project, covering only the urban area of the capital, Prishtina, have been presented in our previous work [8, 9].

Applying comparable methodologies to other similar measurement campaigns performed in other countries [10, 11] we perform a spectrum occupancy analysis over the Ultra High Frequency (UHF) bands. The detection of activity in a specific channel was evaluated using energy detection techniques based on measured values of received power. To improve the accuracy of the detection process, an unsupervised machine learning techniques based on *k-means* clustering were applied, an endeavor which was made possible by the significant amount of measurement data collected.

Finally, we would like to note that the analysis presented in this work gives an overview of the current spectrum availability, which may change in the future as TV broadcasting in Kosovo has not yet fully transitioned from analog to digital broadcasting. Our goal, however, is to quantify the current levels of utilization of the TV spectrum band. The results of our measurement campaign can be used to assess the current status of the spectrum use and the availability of the spectrum for other, opportunistic, users. In addition, the outcomes of our extensive analysis can give valuable information to the national regulators about the efficiency of the current spectrum allocations. Hopefully, in the future, this data can be used for maintaining spectrum usage data and facilitate spectrum sharing in an operational environment.

The rest of the paper is organized as follows: Section II gives a brief description of the relevant literature. Section III describes the details of measurement campaign protocol. The spectrum availability analysis and the learning machine approach we apply is described in section IV. Finally, numerical results are provided in Section V.

II. RELATED WORK

The opportunistic utilization of spectrum through the use of cognitive radio (CR) devices, is an idea that has been studied intensively in the last 15 years [7]. A comprehensive review on performing measurement campaigns for effective spectrum occupancy monitoring is presented in [10]. Another survey focusing on

characterizing and modelling the availability of white spaces is provided in [2]. An important aspect when assessing activity in the surveyed spectrum band, is the detection method applied. In particular, when applying energy based techniques which rely on a pre-determined threshold to decide whether a channel is active or not, the assessment of the noise level becomes critical. In general, most studies [5-7], for sake of simplicity apply fixed thresholds against which the energy of sampled observations is compared to. In our own previous work [8], we also addressed the availability of TVWS within Prishtina city limits, using a single fixed threshold. However, as we argued in [9], a fixed predetermined threshold is not flexible enough to provide an accurate picture of the spectrum availability. To this end, more intricate methods have been studied and presented in literature. The estimation of the detection threshold without prior knowledge about the signal and noise characteristics is addressed in [11]. The use of iterative algorithms based on impulse suppression principles for spectrum sensing purposes is proposed in [12, 13]. The authors in [14] propose instead the application of adaptive double thresholds to enhance the performance of sensing process, which was also the approach taken in [9]. Finally, machine learning techniques have also found potential application in this area. The authors in [15, 16] propose the use of unsupervised and supervised machine learning techniques on observed samples to determine the activity within a specific frequency channel. A cooperative technique based on Hidden Markov Models is instead proposed [17].

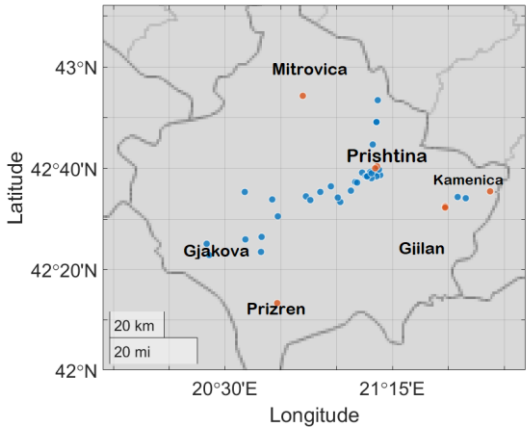


Figure 1 Measurement locations (blue) and TV transmitter (orange) positions

In this work, we adopt a simple unsupervised learning technique and apply it to obtain an accurate and realistic overview of the spectrum availability in the UHF bands, at national level, within the territory of Kosovo.

III. Measurement Campaign Protocol

The analog terrestrial TV broadcasting in Kosovo is allocated in the following spectrum bands: VHF 174-230 MHz (8 channels, each 7MHz wide) and UHF 470 – 860 MHz (48 channels, each 8MHz wide), according to data obtained from Kosovo Independent Media Commission [18]. A list of TV transmitters, their coordinates, antenna height and transmit power is also provided in their website [18].

The frequency bands under assessment included UHF channels over the entire $470 - 854$ MHz band. The measurement campaign was conducted at 36 different outside locations across the territory of Kosovo. Out of these locations 14 (around 40%) were situated in urban/suburban areas, while 18 were situated in rural areas. The measurements were performed using the NARDA Selective Radiation Meter SRM-3006, in the spectrum analysis mode, with a frequency resolution of 100 kHz, and sweep time of ~200 ms. At each location, 10 individual measurements were performed, each lasting 1 minute. For each measurement, the minimum, maximum and average values of the power received were recorded.

In particular, 7 active transmitters were identified during the measurement campaign. The measurement locations as well as the locations of the 7 TV broadcasters are shown in Fig. 1. The locations were chosen so as to obtain a sufficient distance resolution from the transmitters. The choice of the locations was also constrained by the availability of roads and accessibility. Within the city center of Prishtina measurements were taken in several different locations within the city center as well as in the surrounding areas.

IV. Spectrum Availability Analysis

A. Theoretical coverage

The theoretical coverage of the TV broadcasters can be established from the official list of licensed transmitters [18], which includes detailed information on antenna locations and characteristics. In particular, we were able to confirm that 7 transmitters from the official list were currently active, in 10 TV channels. Previous studies [19], show that the propagation model that best fits Kosovo environment is the Ericsson path loss model [20]. Calculating the respective distance between each measurement location and transmitting antenna, and applying the Ericsson path loss equation, we can derive the estimated received power at each location from each transmitter.

To decide whether a location is under the theoretical coverage of a TV transmitter or not, a coverage threshold can be applied based on the expected receiver sensitivity. The value of the threshold can impact significantly the theoretical values of occupancy, however there are tools available on how to establish this threshold. In this paper we have used the FCC's *FM and TV Propagation Curves* online tool, which is based on its own established service contours [21]. The tool can be used to find the distance to a service or interfering contour, or the corresponding field strength at a given contour distance. Applying the Ericsson propagation model, and the values obtained from the tool, we derive that a reasonable threshold value would be in the range of -75 to -70 dBm.

B. The spectrum detection problem

In general, spectrum detection is presented as a binary hypothesis-testing problem:

$$H_0: \text{ Primary user is not present}$$
$$H_1 \quad \text{Primary user is present}$$

In order to test the hypothesis, the cognitive device listens, i.e., senses, the channel and makes a decision based

on the observed samples. Usually, once the sensing is performed the energy of sensed samples is calculated and used as a test statistic. Since the equipment used in this measurement campaign, directly measured the received power level, the latter was used as a test statistic instead. In the following, we shall formulate the spectrum detection problem for our spectrum availability analysis.

Let $\mathcal{T}$ be the set of the TV transmitters, i.e., primary users, $\mathcal{L}$ be the set of measurement locations and $\mathcal{C}$ the set of frequency channels under assessment. For each transmitter $t \in \mathcal{T}$, we denote its transmitting power on channel c as $P_t(t,c)$. We denote the power received at location l, on channel c at time s as $P_r(l,c,s)$. We define the binary variable $\gamma(t,c,s)$ to indicate whether transmitter t is active on channel c at measurement time s. Our test statistic becomes:

$$P_r(l,c,s) = \begin{cases} N_0 & H_0 \\ \sum_{t\in\mathcal{T}} P_t(t,c)\alpha(t,l,c)\gamma(t,c,s) + N_0 & H_1 \end{cases}$$

where $\alpha(t,c)$ represents the pathloss exponent between transmitter t and location l on channel c, and N_0 is the white noise power level.

In the traditional energy detection technique, the energy, i.e., the power, of the sensed samples is compared to a fixed pre-defined threshold, in order to reach a decision. In that case, the decision on whether a channel c is active at time k, at a location l, can be expressed as:

$$A(l,c,s) = \begin{cases} 0 & \text{if } P_r(l,c,s) \le \delta \\ 1 & \text{if } P_r(l,c,s) > \delta \end{cases}$$

where δ is the predefined threshold.

Two key metrics to be considered in these kinds of problems are the probability of misdetection defined as $\Pr\{A=0|H_1\}$, and the probability of false alarm defined as $\Pr\{A=1|H_0\}$. While the energy detection technique is particularly attractive due to its simplicity and the fact that it requires no prior knowledge about the signal, its drawback is that it is very vulnerable to noise, which can lead to high rates of *false* or high number of *misdetections* depending on the threshold. On the one hand, if the threshold is set too low, the detector will be biased towards detecting the presence of a signal, even when there is only noise, leading to a false alarm. A high rate of false alarms can lead to significant underestimation of the availability of the spectrum, rendering the process useless. On the other hand, a high threshold, can underestimate the presence of the primary signal, i.e. TV signal, therefore increasing the risk of causing harmful interference to the primary user.

C. A machine learning approach

In this paper we adopt an unsupervised machine learning approach to detect the presence of a primary signal. In this approach, we again use the measured levels of received power to detect the signal, however instead of comparing them to a pre-determined threshold, we use them to construct *feature vectors*, which are fed to a classifier. The classifier in turn labels each vector with an inactive class, A_0, or an active class, A_1. Applying machine learning to the spectrum detection problem is attractive, because it is less susceptible to noise, compared to the simple threshold technique. A similar technique was applied by the authors in [15], albeit in a cooperative context, by constructing feature vectors from samples taken by several CR devices at different locations. In this work, we do not consider cooperation between devices to reach a decision, i.e., at each location we only use the measured power levels collected there to evaluate the occupancy of the spectrum.

As already mentioned, during our measurement campaign received power levels at each location were measured. The minimum, maximum and average values of the power were recorded. Therefore, we build our feature vector, referred to as power vector, for each frequency channel, as follows:

$$\mathbf{P}_{\mathrm{l,c,s}} = \left[P_r^{\min}(l,c,s)\ P_r^{\mathrm{avg}}(l,c,s)\ P_r^{\max}(l,c,s)\right]$$

However, as the name implies, in order to enable the classifier to accurately label the received power vectors, it must first undergo a *learning* process. Therefore, the process is separated into two phases, the *learning* phase and the *evaluation* phase, as described below.

1) Learning phase

The goal of the learning phase is to build a classifier which is able to accurately label power vectors as inactive or active. In this phase we apply an unsupervised learning technique based on *k-means* clustering. In order to do so, we must feed the learning algorithm with samples of feature vectors. As these samples must be different from those used during the evaluation phase, for each location we split the data collected during the measurement campaign into two sets: the training set, $\mathbf{P}_{\mathrm{l}}^T$, and the evaluation set, $\mathbf{P}_{\mathrm{l}}^E$. The samples from the training set are used to feed the k-means clustering learning algorithm described in Alg. 1. The k-means clustering algorithm [22], is attractive because of its simplicity, and also because it does not require the training vectors to be labeled *a priori*. However, the number of clusters needs to be pre-specified, which in our case is appropriate, because we already know that we are interested in two clusters only, the active and the inactive cluster.

Algorithm 1: *k-means* clustering for spectrum learning

Input: $\mathbf{P}_{\mathrm{l}}^T$, $k = 2$
1: Choose initial cluster centroids $\mathbf{C}_{A_0}, \mathbf{C}_{A_1}$
1: **for all** $\mathbf{P} \in \mathbf{P}_{\mathrm{l}}^T$
2: Calculate distance $d(\mathbf{P}, \mathbf{C}_{A_i}) = \|\mathbf{P} - \mathbf{C}_{A_i}\|$ from all centroids
3: Assign $\mathbf{P}$ to cluster A_i, where $\min_i d(\mathbf{P}, \mathbf{C}_{A_i})$
4: **end for**
5: Recalculate cluster centroids $\mathbf{C}_{A_i} = \frac{1}{|\mathrm{A_i}|}\sum_{\mathbf{P}\in A_i}\mathbf{P}$
6: **for all** $\mathbf{P} \in \mathbf{P}_{\mathrm{l}}^T$
7: Reassign $\mathbf{P}$ from cluster i to cluster j, if $\sum_i \sum_{\mathbf{P}\in A_i}\|\mathbf{P} - \mathbf{C}_{A_i}\|^2$ is reduced
8: Recalculate cluster centroids $\mathbf{C}_{A_i} = \frac{1}{|\mathrm{A_i}|}\sum_{\mathbf{P}\in A_i}\mathbf{P}$
9: **end for**
10: Repeat steps 6-8 until cluster assignments do not change, or the maximum number of iterations is reached.
11. Return $\mathbf{C}_{A_0}, \mathbf{C}_{A_1}$

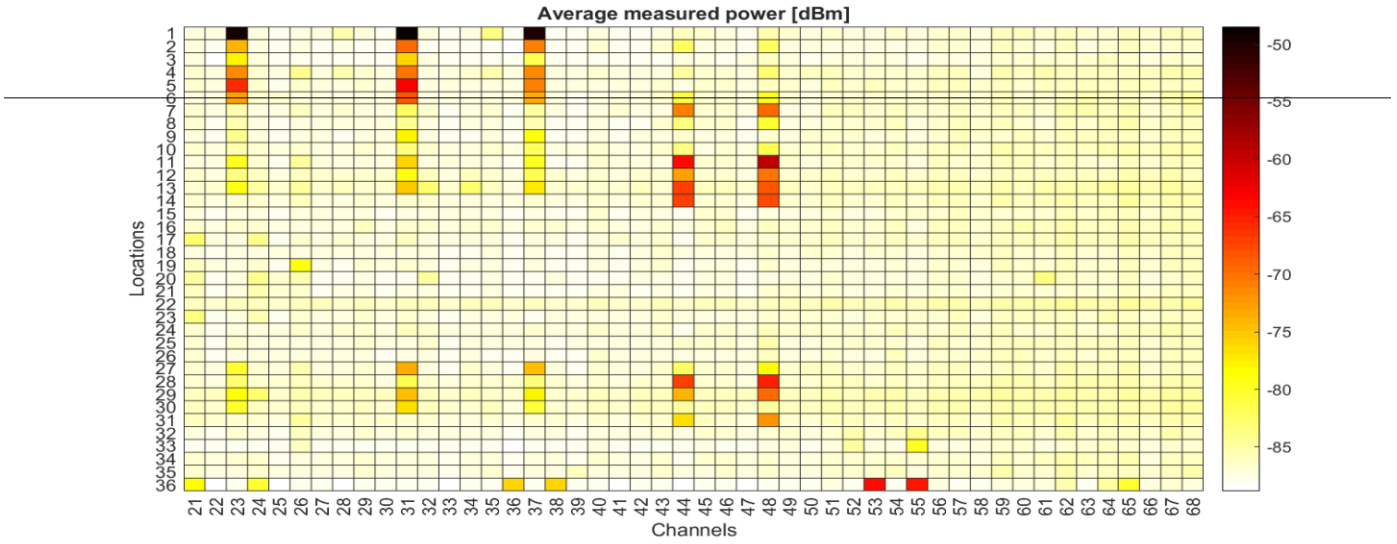


Figure 2. Heatmap of average measured power level in dBm

The input to the algorithm are the training vectors for location l, and the number of clusters which in our case is 2. The algorithm will initially choose the cluster centroids, by selecting two random power vectors from the training set. In the first stage, it will assign each power vector to the cluster with the closest centroid. Once this stage is done, the centroids are recalculated based on the assignment. The second stage is an iterative procedure whereby the algorithm goes through each power vector, and evaluates whether changing its cluster assignment would reduce the within-cluster sum-of-squares point-to-cluster-centroid distances:

$$\sum_{i} \sum_{\mathbf{P} \in A_i} \left\| \mathbf{P} - \mathbf{C}_{A_i} \right\|^2$$

The second stage is repeated until the clusters converge, or a maximum number of iterations is reached. At the end, the algorithm returns the centroids for the two clusters. It should be noted that the algorithm is performed independently for each location. Therefore for each location, we will have two centroid vectors denoted as $\mathbf{C}_{A_i}^{l}$.

2) Evaluation phase

The second phase of the spectrum detection process is the actual evaluation of whether a primary user is active over a channel, at a specific location and time instant, or not. For each location, the power vectors from the evaluation set $\mathbf{P}_{l,c,s} \in \mathbf{P}_l^E$, are fed into the classifier. The classifier evaluates each power vector, by calculating its distance from the location cluster centroids:

$$d\left(\mathbf{P}_{l,c,s}, \mathbf{C}_{A_i}^{l}\right) = \left\| \mathbf{P} - \mathbf{C}_{A_i}^{l} \right\|, \quad i = 0,1$$

The classifier will assign to the vector, the label which corresponds to the smallest distance from the cluster centroid. It should be noted that the performance of the algorithm depends heavily on the quality of the training data. In certain locations, all training vectors may belong to the same class, which makes it difficult for the classifier to distinguish between the two clusters. A reliability indicator can be defined, by calculating the distance between the cluster centroids:

$$R(l) = \left\| \mathbf{C}_{A_0}^{l} - \mathbf{C}_{A_1}^{l} \right\|$$

If this reliability indicator is below a certain threshold, additional means to ensure low probability of false alarm can be taken as proposed in [15]. Namely, in cases where $R(l) \leq \rho$, the classifier will label a power vector as inactive, A_0, only if the following criteria is met:

$$\frac{\left\| \mathbf{P} - \mathbf{C}_{A_0} \right\|}{\min_i \left\| \mathbf{P} - \mathbf{C}_{A_i} \right\|} \geq \beta$$

Where ρ and β are threshold parameter to control the tradeoff between the misdetection and the false alarm probabilities.

V. Spectrum Detection and Analysis Results

In this section we analyze the measurement data collected during the campaign described in Section II.

At every location, the received power level was measured, and the minimum, average and maximum values of the power, for the measurement duration, were recorded. The received power was measured in dBm over the entire $470 - 854$ MHz band, in bins of 100 kHz. The power levels detected in the individual bins within each channel band were then summed to obtain the received power level at the specific channel:

$$P_{l,c,s} = 10\log 10 \sum_{b=1}^{B_c} 10^{\frac{P_{l,b,s}}{10}}$$

Where $P_{l,b,s}$ is the power measured at location l, on frequency bin b, at measurement time s. B_c is the number of bins in channel c, which depends on the channel bandwidth and can vary from channel to channel.

Using the list of active TV transmitters, the distance between each transmitter - location pair was calculated. This distance was used to establish whether a location falls under the theoretical coverage area of a transmitter as described in Sec. IV-A.

The availability of the spectrum was assessed through processing of measurement data with MATLAB software. First, in Fig. 2, we show the average measured power at every location and channel. The max, min and average values of the measured power for each location, channel and measurement time, were stored, and separated into the training set and evaluation set, as described in Sec. IV-C. The training set was fed to the k-means clustering algorithm to obtain the cluster centroids for each location. The validation set was then submitted for classification, and an occupancy value of 0, or 1, was assigned for each (l, c, s) triple.

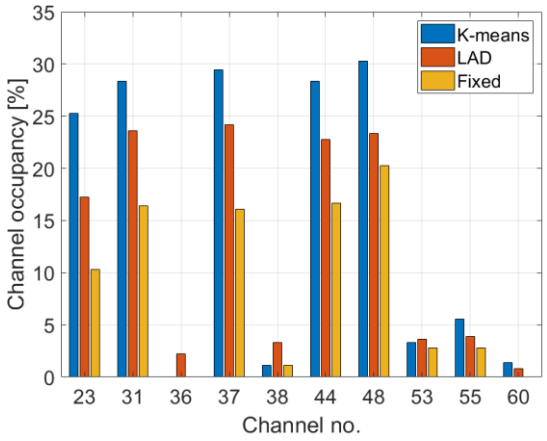


Figure 3. Averaged occupancy over all locations and measurement times for known active channels.

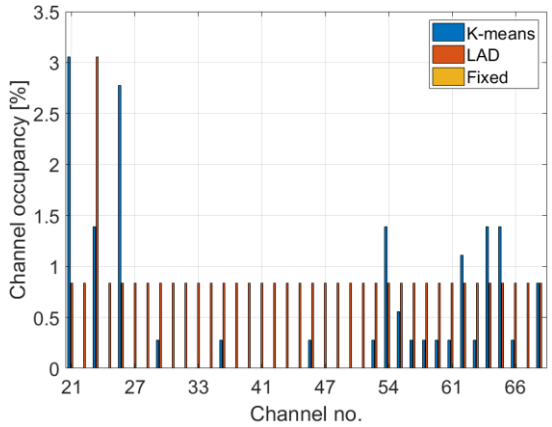


Figure 4 Averaged occupancy over all locations and measurement times for known other channels with unknown activity status.

In addition, for comparative purposes we also performed the detection using a fixed threshold, and double adaptive thresholds (LAD) as described in [9]. This analysis produced availability picture for the measurement locations and frequency band under study. The average occupancy of known active channels over all measurement locations and measurement times, is shown in Fig. 3.

The results show that even known active channels appear un-occupied in more the 70% of the locations. K-means is able to detect the presence of the primary users more often that fixed and LAD methods in most locations. However, LAD is able to detect primary user presence in some locations that k-means does not. It should be noted that LAD is an adaptive method which tunes its threshold based on the noise level, therefore it could perform better in noisy locations at higher frequencies, since noise is known to increase with frequency [11].

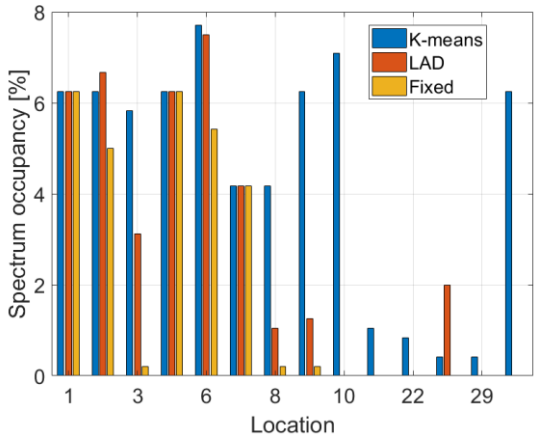


Figure 5. Spectrum occupancy percentage over the UHF channels, in suburban locations.

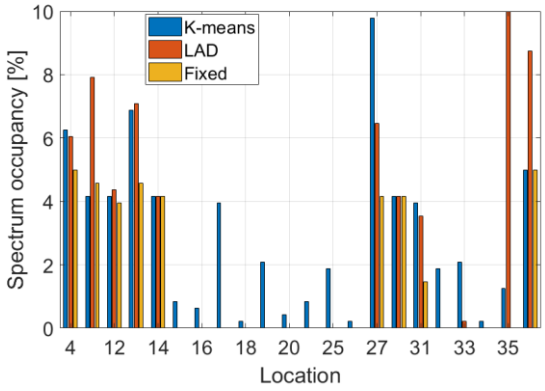


Figure 6 Spectrum occupancy percentage over the UHF channels, in rural locations.

For other channels, for which we have no confirmation about their activity, occupancy percentage varies between 0-3%, as shown in Fig. 4. LAD exhibits a near-constant occupancy rate of 0.8%, which is connected to the fact that LAD thresholds are determined based on fixed target probability of false alarm values. The fixed threshold method does not detect any activity in any of the channels, because the threshold is set too high. K-means displays variable occupancy in these channels, although the occupancy increases with frequency. This could indicate an increase in false alarms at higher frequencies due to the increase in noise levels.

Finally, we look at the spectrum occupancy at the different locations. At suburban locations the spectrum occupancy over the entire 470-860 MHz band, varies between 0-8%, as shown in Fig. 5, while in rural areas between 0.02-10%, as shown in Fig. 6. A map view of the occupancy levels is shown in Fig. 7.

CONCLUSION

In this paper we address the utilization level of spectrum resources in the TV bands over the territory of Kosovo. To do so, we leverage a large set of data obtained from an extensive spectrum measurement campaign which enabled us to attain a comprehensive picture of spectrum availability in TV bands. The significant amount of measurement data collected, enabled us to apply machine learning techniques to improve the accuracy the spectrum detection process. As an end result, we obtained a highly accurate spectrum availability map which can be in turn used by cognitive radio network to opportunistically access the spectrum. Our analysis show that occupancy of spectrum all over the territory of Kosovo, does not surpass 10%, in any of the locations. Furthermore, we show that even known active channels are un-occupied in more than 70% percent of the measurement locations. This indicates that the under-utilization of these bands is significant. Results of our analysis will be shared with regulatory authorities in order to initiate a discussion over the liberalization of these frequency bands for potential opportunistic use.

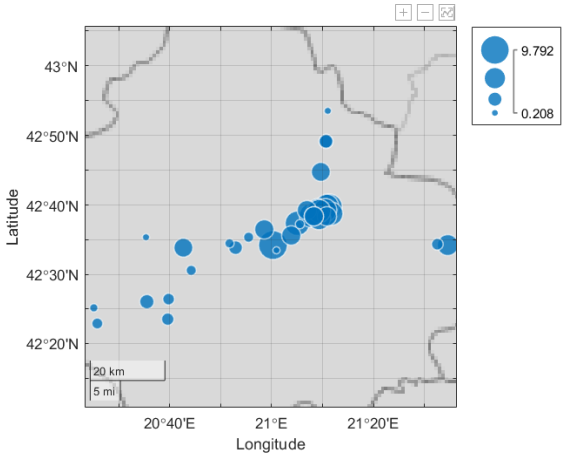


Figure 7 A map view of the spectrum occupancy levels. Larger circles represent higher occupancy ratio

ACKNOWLEDGMENT

This work was supported by research project "Research on the Reusability Possibilities of New Frequency Bands UHF, VHF and Millimeter Waves for Wireless Communication Networks in territory of Kosovo" funded by the Kosovo Academy of Sciences and Arts.

REFERENCES

[1] J. J. Lehtomaki, R. Vuohtoniemi, K. Umebayashi, “On the Measurement of Duty Cycle and Channel Occupancy Rate”, IEEE Journal on Selected Areas in Communications, May 2013.

[2] B. Fette, Cognitive Radio Technology, Elsevier, 2008.

[3] Federal Communications Commission, " Notice of proposed rulemaking and order: Facilitating opportunities for flexible, efficient, and reliable spectrum use employing cognitive radio technologies", ET Docket No. 03-108, Feb.2005

[4] OFCOM Consultation: Digital Dividend: Cognitive Access, Feb 16, 2009. http://www.OFCOM.org.uk/consult/condocs/cognitive/summary]

[5] M. Hadzialic, J. Muovic, M. Hamza, M. Milisic, K. Huseinovic, J. Dizdarevic and M. Dulic, "TV White Space: Solution for bridging the gap between user's demand and the network capacities", 22nd Telecommunications forum TELFOR Belgrade, 2014.

[6] L. Csurgai-Horvath, I. Rieger, and J. Kertesz, "A Survey of the DVB-T Spectrum: Opportunities for Cognitive Mobile Users", Mobile Information Systems, 2016.

[7] B. Jankuloska, M. Pavloski, M. 1 Zahariev, V. Atanasovski and L. Gavrilovska, "Efficecient Spectrum Utilization: A Cognitive Approach", IEEE ICC 2010, Cape Town, South Africa.

[8] H. Maloku, Z. L. Fazliu, M. Ibrani, A. Mekuli, E. Sela and M. Rajarajan, "Measurement of Frequency Occupancy Levels in TV Bands in Urban Environment in Kosovo," 2018 18th Mediterranean Microwave Symposium (MMS), Istanbul, Turkey, 2018, pp. 268-271.

[9] Z. Limani Fazliu, H. Maloku, A. Mekuli, M. Ibrani, "Detection Threshold for TVWS Spectrum Occupancy Determination in Urban Environments in Kosovo", 42nd MIPRO Convention, Opatija, Croatia, May, 2019.

[10] M. Höyhtyä et al., "Spectrum Occupancy Measurements: A Survey and Use of Interference Maps," in IEEE Communications Surveys & Tutorials, vol. 18, no. 4, pp. 2386-2414, 2016.

[11] M. Lopez-Benitez and F. Casadevall, “Methodological aspects of spectrum occupancy evaluation in the context of cognitive radio,” European Transactions on Telecommunications, vol. 21, no. 8, pp. 680–693, Dec. 2010.

[12] H. Saarnisaari, P. Henttu, and M. Juntti, “Iterative multidimensional impulse detectors for communications based on the classical diagnostic methods,” IEEE Trans. Commun., vol. 53, no. 3, March 2005, pp. 395–398.

[13] J. Vartiainen and R. Vuohtoniemi,"False Alarm Rate Analysis of the FCME Algorithm in Cognitive Radio Applications", AICT 2015, The Eleventh Advanced International Conference on Telecommunications, June 2015.

[14] J. Vartiainen, H. Sarvanko, J. Lehtomaki, M. Juntti, and M. Latva-aho, “Spectrum sensing with LAD based methods,” in Proc. IEEE PIMRC, Athens, Greece, Aug. 2007.

[15] K. M. Thilina, K. W. Choi, N. Saquib and E. Hossain, "Machine Learning Techniques for Cooperative Spectrum Sensing in Cognitive Radio Networks," in IEEE Journal on Selected Areas in Communications, vol. 31, no. 11, pp. 2209-2221, November 2013.

[16] Haozhou Xue and Feifei Gao, "A machine learning based spectrum-sensing algorithm using sample covariance matrix," 2015 10th International Conference on Communications and Networking in China (ChinaCom), Shanghai, 2015, pp. 476-480.

[17] H. Eltom, S. Kandeepan, Y. Liang and R. J. Evans, "Cooperative Soft Fusion for HMM-Based Spectrum Occupancy Prediction," in IEEE Communications Letters, vol. 22, no. 10, pp. 2144-2147, Oct. 2018.

[18] Kosovo Independent Media Commission. “Gjendja reale dhe e planifikuar e planit frekuencor televiziv të Republikës së Kosovës”, https://www.kpm-ks.org/assets/cms/uploads/files/Gjendja%20reale%20dhe%20e%20planifikuar%20planit%20frekuencor%20dhe%20televiziv%20t%C3%AB%20Kosov%C3%ABs.doc

[19] H. Maloku, Z. Limani Fazliu, E. Sela, M. Ibrani, "Path loss model fitting for TV bands based on experimental measurements for urban environments in Kosovo" 42nd MIPRO Convention, Opatija, Croatia, May, 2019.

[20] Zakaria, Yahia, Jiri Hosek, and Jiri Misurec. "Path Loss Measurements for Wireless Communication in Urban and Rural Environments." American Journal of Engineering and Applied Sciences 8.1 (2015): 94. Radio Wave Propagation Mechanisms And Empirical Models For Fixed Wireless Access Systems.

[21] https://www.fcc.gov/media/radio/fm-and-tv-propagation-curves

[22] Lloyd, Stuart P. “Least Squares Quantization in PCM.” IEEE Transactions on Information Theory. Vol. 28, 1982, pp. 129–137.

[23] Arthur, David, and Sergi Vassilvitskii. “K-means++: The Advantages of Careful Seeding.” SODA ‘07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. 2007, pp. 1027–1035.

SIGNAL PROCESSING

The Influence of Multipath Propagation of the Signal on the Accuracy of the GNSS Receiver

Teodor B. Iliev*, Ivaylo S. Stoyanov**, Strahil A. Sokolov***, Ivan H. Beloev****
* University of Ruse/Department of Telecommunication, Ruse, Bulgaria
** University of Ruse, Department of Electrical and Electronic Engineering, Ruse, Bulgaria
*** University of Telecommunications and Post/Department of Information technologies, Sofia, Bulgaria
****University of Ruse, Department of Transport, Ruse, Bulgaria
e-mail address: tiliev@uni-ruse.bg, stoyanov@uni-ruse.bg, strahil.sokolov@gmail.com, ibeloev@uni-ruse.bg

***Abstract* - The GNSS system allows on a planetary scale to fulfill navigation functions, that is, to determine the current location of consumers of navigation information and their speed and to carry out accurate time on the Earth's surface and in near-Earth space. There are several sources of errors that distort the code and the carrier of the navigation signal. The noise floor of the receiver error and the errors from multipath propagation of the signal are independent for each receiver. Multipath propagation of the signal is the main source of errors for high-precision navigation. In this paper we propose a methods for estimation of these errors, which are based on the properties of the time correlation of the multipath propagation of the navigation signal.**

Keywords - GNSS, multipath propagation, user equipment

I. INTRODUCTION

The propagation of radio waves in far-field is characterized by reflections off the surfaces of natural or artificial objects. Diffraction from peaks, hills and buildings and the refraction from the lower layers of the atmosphere also have an effect. [1, 2] The signal is subjected to two main types of attenuation:

- attenuation due to prolonged fading, propagation and shading losses;
- attenuation due to short-term fading, multipath signal propagation and Doppler signal propagation.

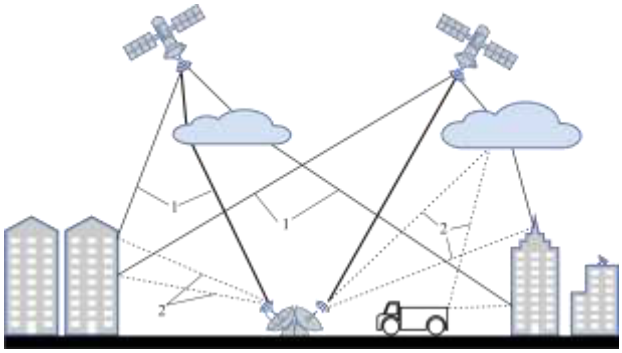


Figure 1. Multipaht propagation of the signal:
1 – direct signal; 2 – reflected signal.

When the signal is reflected, the path it traverses is extended, and thus the measured distance to the satellites changes (Fig. 1). In this way the reflected signal is mixed with the direct ones leading to inaccuracy in determining the values of the distances and errors are received while calculating the position. Multipath errors, together with ionospheric delays, have the greatest impact on position accuracy. [3]

II. METHODS FOR ELIMINATING ERRORS CAUSED BY MULTIPATH PROPAGATION

There are a number of methods for eliminating errors caused by multipath propagation, which underlie the application of temporal or spatial selection. One possible attempt to reduce the influence of ground multipath is the use of ground planes or choke ring antennas.

Choke rings are a rather simple means to mitigate multipath. They are ground planes consisting of several concentric rings, three to five concentric ring structures that are located around the GNSS antenna (Fig.2). The resulting grooves are able to cancel out the primary and secondary wave of a reflected signal. This can be achieved if the groove's depth is approximately equal to a quarter of the carrier wavelength. This is why choke rings are normally designed for a dedicated carrier frequency. However, so-called dual-depth choke rings have also been developed in the past. [4]

The Fig. 2 helps to describe how the innovative technology responds to unwanted multipath signals. The desirable signals "1" a shown with continuous line and undesirable signals are shown with dashed line "2".



Figure 2. Choken ring weakens multiparh signals:
1 – direct signal; 2 – reflected signal.

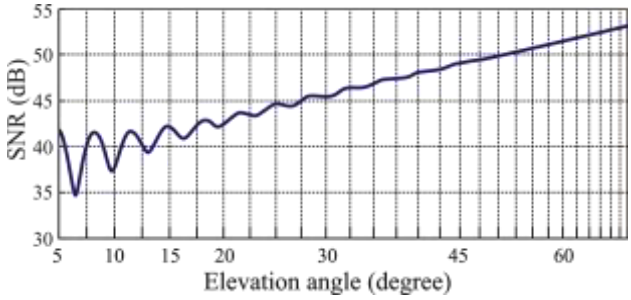


Figure 3. Signal-to-noise ratio in GPS L2C signal

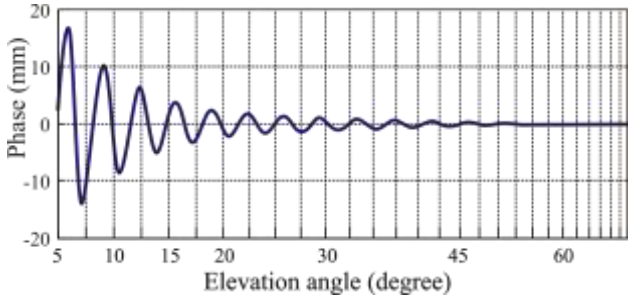


Figure 4. Carrier phase multipath error in GPS L2C signal

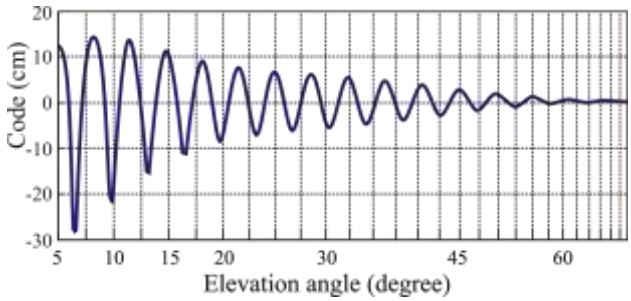


Figure 5. Multipath error in GPS L2C signal

Spatial methods use antennas with special directional diagrams of the choke-ring type, suppression of the signal arriving from the lower hemisphere, but this method is ineffective at passing the reflected signal coming from the upper hemisphere. [5]

On Fig. 3 - 5 are shown the model results for GPS L2C signal identifiable, like: signal/noise ratio (*SNR*), carrier phase multipath error and multipath error, where error means the difference with respect to multipath-free conditions. The results are received for horizontal ground surface, made up of soil with medium-level moisture and negligible roughness, observed with a choke-ring antenna installed upright on a 1.5 m tall monument.

We should note that the reflecting surface of the receiver antenna is horizontal, made up of soil with medium level moisture and negligible roughness. The installed antenna is a choke-ring on a 1.5 m high monument.

Other methods are based on the use of temporary selection of a delayed reflected signal. Due to the fact that the navigation signal is continuous, the methods used to operate the pulse signals in this situation are not acceptable.

In order to limit the multipath errors, so-called strobe pulse method is used. [6]. This method is based on the creation of a special type of discriminatory characteristic that generates an error signal in the chain of the modulation code delay tracking system. The discriminatory characteristic means the dependence of the signal on the output of the discriminator. The discriminatory characteristic defines many important properties of the delay tracking system, such as the range of operation (when entering trace mode), tracking errors with adaptive interference, and the delay time at which the reflected signal is suppressed. The necessary form of discriminatory characteristic is achieved by selecting the reference signal to be given to the discriminator. In this method, the signal consists of a sequence of short pulses called strobe pulses.

An elementary (simple) strobe impulse is a single impulse of rectangular shape, and a complex one consists of a group of impulses with different polarity. After determining the form of the strobe pulse, it is necessary to determine its duration based on the trade-off between multipath errors and noise.

A. Theoretical calculation of multipath errors

There are theoretical methods for calculating multipath errors for given types of discriminators and pseudorandom sequence. In this regard, it is necessary to choose an appropriate model that describes the behavior of the reflected signals. The delay tracking system discrimination algorithms are determined by the relationship between the output value of the discriminator Z_d and the offset time ε of the reference signal against the modulating pseudorandom sequence.

The error in signal phase tracking (φ) in systems with phase automatic tuning and the delay tracking system error (ε) can be represented by the correlations of the correlation signals as follows:

$$\begin{aligned} I(\varepsilon) &= R_0(\varepsilon)\cos(\varphi); \quad Q(\varepsilon) = R_0(\varepsilon)\sin(\varphi); \\ dI(\varepsilon) &= \Delta R_0(\varepsilon)\cos(\varphi); \quad dQ(\varepsilon) = \Delta R_0(\varepsilon)\sin(\varphi); \end{aligned} \tag{1}$$

where:

$R_0(\varepsilon)$ is the reciprocal correlation function of the pseudorandom sequence of the high-frequency output of the receiver and the reference signal, which is a copy of the modulating pseudorandom code;

$\Delta R(\varepsilon)$ is the mutual correlation function of the pseudorandom sequence of the high frequency output of the receiver and the reference signal in the channel of the delay tracking system in the form of a sequence of strobe pulses. In [7], two types of discriminators are considered: coherent and non-coherent, often used in professional navigation receivers. In this case, the non-coherent discriminator is described by the expression:

$$\overline{Z}_{di}(\varepsilon) = I(\varepsilon).dI(\varepsilon) + Q(\varepsilon).dQ(\varepsilon) \tag{2}$$

Out of (2) it is to see that the non-coherent discriminator does not depend on the error in signal phase tracking.

For the coherent discriminator, the following is calculated:

$$\overline{Z}_{dc}(\varepsilon) = I(\varepsilon).dI(\varepsilon) \tag{3}$$

If the considered phase automatic frequency tuning has a small tracking error φ=0, then under the direct signal the discriminant characteristics will be the same and have the following form:

$$\overline{Z}_{di}(\varepsilon) = \overline{Z}_{dc}(\varepsilon) = R_0(\varepsilon).\Delta R(\varepsilon) \tag{4}$$

In the case of a single reflected signal of a certain amplitude, the signal of the incoherent discriminator can be recorded as:

$$\overline{Z}_{di}(\varepsilon) = \overline{Z}_{di}(\varepsilon) + \alpha\cos\theta\left[R_0(\varepsilon-\delta).\Delta R(\varepsilon) + R_0(\varepsilon).\Delta R(\varepsilon)\right] + \alpha^2 R_0(\varepsilon-\delta).\Delta R(\varepsilon-\delta) \tag{5}$$

where α – the relative amplitude of the reflected signal, δ - delay of the reflected signal relative to the direct signal, θ - phase shift of the carrier of the reflected signal relative to the direct signal.

For a coherent signal, the effect of multipath errors on the phase automatic frequency tuning needs to be taken into account. Then the output of the coherent discriminator will be:

$$\overline{Z'}_{dc}(\varepsilon) = Z_1(\varepsilon).Z_2(\varepsilon).Z_3(\varepsilon), \tag{6}$$

where:

$$Z_1(\varepsilon) = \frac{1}{1+\left[\frac{\alpha\sin\theta.R_0(\varepsilon-\delta)}{R_0(\varepsilon)+\alpha\cos\theta.R_0(\varepsilon-\delta)}\right]^2};$$

$$Z_2(\varepsilon) = \Delta R(\varepsilon) + \alpha\cos\theta.R(\varepsilon-\delta) + (\alpha\sin\theta)^2\frac{\Delta R(\varepsilon-\delta).R_0(\varepsilon-\delta)}{R_0(\varepsilon)+\alpha\cos\theta.R_0(\varepsilon-\delta)};$$

$$Z_3(\varepsilon) = R_0(\varepsilon) + \alpha\cos\theta.\Delta R(\varepsilon-\delta) + (\alpha\sin\theta)^2\frac{R_0{}^2(\varepsilon-\delta)}{R_0(\varepsilon)+\alpha\cos\theta.R_0(\varepsilon-\delta)}.$$

Then multipath errors can be defined as the root of the equation $\overline{Z'}_{dc}(\varepsilon)=0$ and $\overline{Z'}_{di}(\varepsilon)=0$. The time delay and the phase change are related to the expression:

$$\Theta = 2\pi.f_c.\delta, \tag{7}$$

where f_c is the carrier frequency.

Therefore, the reflected signal is determined by the amplitude and delay relative to the direct signal. It is also important that with a slight fluctuation in the delay time, the phase shift changes very strongly and multipath errors will change their sign. For this reason, many researchers are examining the multipath error envelope that is most relevant to errors.

III. ANALYSIS OF ERRORS CAUSED BY MULTIPATH PROPAGATION

The Phase Multipath Mitigation Window (*PMMW*) proposed in [8] is used to estimate multipath errors in phase measurements. Instead of combining the received in-phase and anti-phase parameters of the signals at any given time (about 40 GHz), only samples immediately preceding and following the code transition are used. (Fig.6). The PMMW technology is said to allow a practical evaluation of the Doppler shift in the presence of multipath propagation of the signal.

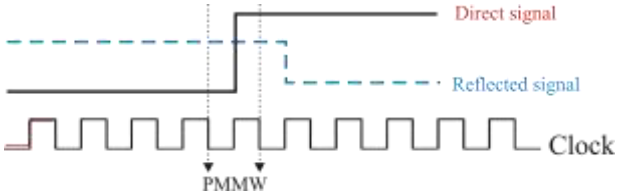


Figure 6. Phase multipath mitigation window (PMMW)

To reduce one of the major limitations of the Phase Multipath Mitigation Window (*PMMW*), an algorithm can be used to recover errors caused by multipath that are invariant to the distance between the antenna and the reflector. The multipath can be described by the following parameters:

- Functional dependence of signal - to - noise ratio change (indicated by SNR or C/N_0) from the relative position angle of the navigation satellite;
- Code error obtained by accounting for deviations from the average ionosphere parameters for frequencies L1 and L2.

$$C_1 - \left(1 + \frac{2}{(f_1/f_2)^2 - 1}\right)\Phi_1 + \left(\frac{2}{(f_1/f_2)^2 - 1}\right)\Phi_2 \tag{8}$$

- PMMW correction, which is a phase error estimate Φ.

The effect of multipath can be described as follows:

$$C/N_{0_{multipath}} \sim K\sum_i \alpha_i \cos\Theta_i \tag{9}$$

$$\tau \sim K\sum_i \alpha_i \cos\Theta_i * d_i \tag{10}$$

$$\Phi \sim \sum_i \alpha_i \cos\Theta_i \qquad (11)$$

for n reflected signals ($i = 1 \ldots n$),

where: $\alpha_i = R(d_i + \tau)/R(\tau)$; R is the autocorrelation function of the code; d_i is the delay of the code of the ith reflected signal; Θ_i is the phase delay of the ith reflected signal; α_i is the ratio of the amplitude of the direct and reflected i^{th} signal; K is a constant independent of the direct and reflected i^{th} signal. The value of the coefficient K determines the characteristic of the particular type of GPS receiver in terms of signal/noise (C/N_0).

The above relations are based on the rules of geometric optics [9, 10, 11]. Phase error corrections, called multipath errors, are based on the signal-to-noise ratio [11]. This method is based on the harmonic nature of the change in signal-to-noise ratio (10) close to its nominal value when multitrack arises. The determination of the amplitude A_i and the argument h_i is possible by combining different classical signal processing algorithms: first, adaptive notch filters (*ANF*) for frequency and amplitude identification and second adaptive least squares (*ALS*). The general dependency has the form:

$$C/N_{0_{multipath}} = \sum_i \hat{A}_i \cos\hat{h}_i \qquad (12)$$

This identification is based on the phase correction method. The phase correction of multipath is given in (11), where the same amplitude α_i and the argument i are given as the multipath deviation of the signal/noise ratio. Therefore, A_i and h_i, treated as signal-to-noise variables, can be substituted in the phase correction expression (12):

$$\Phi \sim \frac{1}{K} \sum_i \pm \hat{A}_i \sin\hat{h}_i \qquad (13)$$

In spite of this, calculation by changing the signal-to-noise ratio (C/N_0) does not allow the sign of the argument to be obtained, since this ratio allows you to restore only the cosine of the argument and not the sine, which is why the sign does not change.

IV. CONLUSION

The article provides classification of the factors that affect the accuracy of real-time GNSS positioning systems. The types of disturbing effects and mathematical models of errors caused by multipath propagation were analyzed.

The presented methodology refers to the theoretical description of errors due to multipath. The complexity of the presented method in this article is primarily related to the adequate and accurate description of the reflected signal. For the precise calculation, it is necessary to specify any model, while keeping in mind that this model may be far from specific real conditions.

The considered theoretical approach for the calculation of multipath errors can be used in the design of a navigation receiver, in order to determine as accurately as possible the frequency range of the receiver, the strobe pulse sequences for suppressing errors from the multipath signal and the filters in the receiver.

ACKNOWLEDGMENT

This work is also partially supported by the Science Fund of University of Ruse "Angel Kanchev".

REFERENCES

[1] B. Parkinson and J. Spilker (ed.). Global Positioning System: Theory & Applications. Washington, DC: American Institute of Aeronautics and Astronautics, Inc. 1996.

[2] I. Rumora, N. Sikirica and R. Filjar, An Experimental Identification of Multipath Effect in GPS Positioning Error. TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation., 2018, 12. 29-32. 10.12716/1001.12.01.02.

[3] M. Filić and R. Filjar, Smartphone GNSS Positioning Performance Improvements through Utilisation of Google Location API, 2018 41th International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO 2018, Opatia, Croatia, pp. 507 – 510.

[4] V. Filippov, I. Sutiagin, J. Ashjaee, "Measured Characteristics of Dual Depth Dual Frequency Choke Ring for Multipath Rejection in GPS Receivers", Proceedings of the 12th International Technical Meeting of the Satellite Division of the Institute of Navigation, September 14-17, 1999, Nashville, Tennessee

[5] http://trl.trimble.com/docushare/dsweb/Get/Document-410894/022543-429G_Antenna_BRO_USL_0817_LR.pdf

[6] F. Sousa, F. Nunes and J. Leitao. (2006). "Strobe pulse design for multipath mitigation in BOC GNSS receivers". Proceedings of IEEE/ION PLANS 2006, San Diego, CA, April 2006, pp. 348-355.

[7] F. Nievinski and K. Larson, "Forward modeling of GPS multipath for near-surface reflectometry and positioning applications". GPS Solutions. Vol. 18, Issue 2, 2014, pp. 309-322.

[8] D. Bétaille, J. Maenpa, H.J.Euler and P.Cross, "A New Approach to GPS Phase Multipath Mitigation", Proceedings of the 2003 National Technical Meeting of The Institute of Navigation, Anaheim, CA, January 2003, pp. 243-253.

[9] J. K.Ray, M. E. Cannon, and P. Fenton, „Code Range and Carrier Phase Multipath Mitigation Using SNR, Range and Phase Measurements in a Multi-Antenna System". Proceedings of ION GPS-99. Nashville. September, 1999.

[10] Y. Georgiadou, „Kleusberg On carrier signal multipath effects in relative GPS positioning" // Manuscripta Geodaetica. Vol. 13. 1988.

[11] C. J. Comp and P. Axelrad, „An adaptive SNR-based carrier phase multipath mitigation technique". Proceedings of ION GPS-96. The Institute of Navigation. Kansas City. 1996.

Error Correction with Systematic RLNC in Multi-Channel THz Communication Systems

Cao Vien Phung, Anna Engelmann and Admela Jukan
Technische Universität Braunschweig, Germany
Email: {c.phung, a.engelmann, and a.jukan}@tu-bs.de

***Abstract*—The terahertz (THz) frequency band (0.3-10THz) has the advantage of large available bandwidth and is a candidate to satisfy the ever increasing mobile traffic in wireless communications. However, the THz channels are often absorbed by molecules in the atmosphere, which can decrease the signal quality resulting in high bit error rate of received data. In this paper, we study the usage of systematic random linear network coding (sRLNC) for error correction in generic THz systems with with $2N$ parallel channels, whereby N main high-bitrate channels are used in parallel with N auxiliary channels with lower bit rate. The idea behind this approach is to use low-bit rate channels to carry coding redundancy generated based on data sent over high-bit rate channels. The coding redundancy can be used at destination to compensate errors occurred during THz transmission. The presented analysis evaluates different settings of THz transmission system in term of the amount of coding redundancy, a code rate, required bit rate of auxiliary channels, the modulation format and transmission distance as required system configurations for a fault tolerant THz transmission.**

I. Introduction

The terahertz (THz) frequency band ranging from 0.3 THz to 10 THz has evolved as a prime candidate to satisfy the exponentially growing data volumes in wireless networks. However, the quality of received THz signal, e.g., Signal-to-Noise ratio (SNR), is always a challenge resulting in a high bit error rate (BER) with a increasing signal modulation level and THz transmission distance. Also, increase in BER is often caused by molecule absorption in the atmosphere [1], [2]. To this end, the research community gives novel methods to overcome the problem of high BER of THz signal when sending data over longer transmission distances with high modulation format, and under atmospheric effects.

So far, many authors studied the key components of high-speed THz communication with compensated path attenuation. The authors in [3]–[5] researched on fast and efficient amplitude and phase modulators. The authors in [6]–[10] studied high gain and massive multiple-input multiple-output (MIMO) antennas and the efficient THz beamforming. In [2], WiFi and THz channel are combined to utilize the distance estimation of the THz receiver and to measure the relative air humidity. The research on an adequate channel coding method is however still in its infancy. Currently, the state-of-the-art Forward Error Correction (FEC) Codes such as Reed Solomon, Low Density Parity Check (LDPC) and Polar code are under discussion [11], [12], and their implementations without reducing the THz channel throughput is an open challenge.

In this paper, we use the idea of deploying generic $2N$ parallel channels to enhancing the overall throughput of THz transmission. We propose to use sRLNC in addition to a generic low complexity FEC, as proposed by the standard [13], and in parallel we send the coding redundancy generated during encoding process for erasure error correction at the receiver. To this end, out of $2N$ channels we deploy N main channels and N auxiliary (coding redundancy) channels. Our system configures the source data to be transferred over N THz main channels with high bit rate, while the coding redundancy is dispatched over N auxiliary channels with lower bit rate. We analyze and compare the scenarios of THz system with different channel configurations, modulation formats and transmission distances in term of the amount of sRLNC redundancy, the related code rate and the transmission rate of auxiliary channels required for certain reliability of THz transmission. The results show that the main THz channels can use a higher level modulation format and configure the longer transmission distance to achieve a higher overall throughput in the case of auxiliary channels with low bit rate supported.

The paper is organized as follows. Sec. II designs fault tolerant multi-channel THz transmission system with sRLNC. In Sec. III we analyze the amount of coding redundancy, code rate and bit rate of redundant channels required. Sec. IV present numerical results. Sec. V concludes the paper.

II. System design

This section designs a fault tolerant multi-channel THz transmission system with $2N$ channels and sRLNC. This proposal extends the 2-channel THz transmission system from our previous work [14], [15]. The notations of the paper are summarized in Table I.

As shown in Fig. 1, source data are distributed and simultaneously transmitted over $2N$ parallel THz channels, where N channels use for original data and N channels are for redundant data. Encoding with sRLNC and traffic parallelization are performed in the source prior to THz transmission.

At the source, the input native data are a bit stream structured into a sequence of symbols, i.e. K symbols, where each symbol has a fixed size of s bits. All native symbols will be temporarily stored for the encoding process at the encoding buffer. For the encoding process, we define a generation as a set of symbols used to code and decode together at the source and receiver, respectively, where each generation is divided from the input data stored in the encoding buffer. In

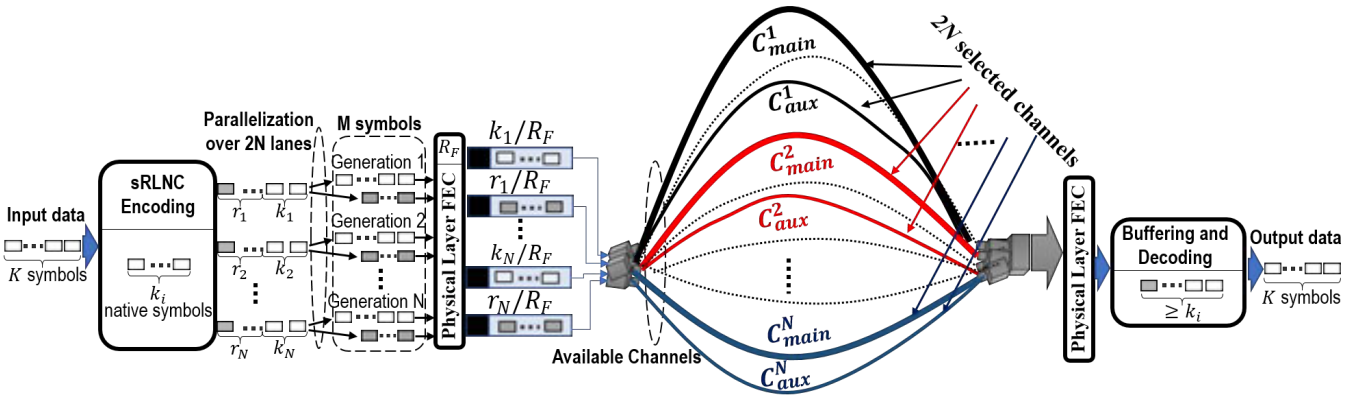


Figure 1. Multi-Channel THz transmission system with systematic random linear network coding (sRLNC).

our example of Fig .1, a generation contains k native symbols encoded together by sRLNC, where the sRLNC is performed over a finite field $\mathbb{F}_{2^s}$, coding coefficients is randomly chosen and generation size is constant. For simplicity, in Fig. 1 we only show N generations, where any i^{th} generation has k_i native symbols. During encoding process, r additional symbols are generated with linear combinations of a generation with k native symbols along with coding coefficients, however, the native symbols keep unchanged and their size as well as the size of any redundant symbols keeps s bits. Thus, r_i redundant symbols of any i^{th} generation are generated by forming linear combinations from k_i symbols. The number of native symbols of any generation i is fixed, i.e., $k_i = k_j = k$, whereby the amount of redundant traffic depends on the expected BER on its related main channel, i.e., $r_i \neq r_j$.

After sRLNC encoding process, $M = K + R$ symbols, where $K = \sum_{i=1}^{N} k_i$ is the total size of generation data and $R = \sum_{i=1}^{N} r_i$ is total amount of redundancy, need to be parallelized and distributed over $2N$ channels, i.e., N high bit rate main THz channels and N low bit rate auxiliary channels. The parallelization process includes two steps. In the first step, each generation is distributed over N lanes, while, in the second step, the redundant symbols are separated from the native symbols requiring two lanes per encoded generation and $2N$ lanes in total. After the parallelization of any N generations over $2N$ lanes, FEC per lane is required. In this paper, we do not consider any specific FEC in the physical layer, while sRLNC is applicable in the higher layer (e.g., Ethernet) to complement any FEC mechanisms with code rate R_F, when FEC either fails or yields insufficient performance. As a result, before the THz signal transmission, any k_i native symbols and their related redundancy are packetized and protected by FEC with code rate R_F, leading to the size of each generation i and its redundancy increased as $\frac{s \cdot k_i}{R_F}$ bits and $\frac{s \cdot r_i}{R_F}$ bits, respectively.

Finally, any generation i with k_i native symbols is sent on the ith main channel and its related redundancy r_i is transferred on the corresponding i^{th} auxiliary channel. All generations and their redundancy are dispatched at the same time on the transmission channels. For our proposed THz system, the high-level modulation format is used for main channels to achieve a high bit rate, while the related auxiliary channels with lower bit rate are for reliably transferring additional coded symbols to correct error symbols happened on the main channels. Some parameters, such as modulation level, transmission distance, signal power, etc., are necessary to be reasonably tuned on the main and related auxiliary THz channels to address symbol errors occurred during Thz transmission.

At the receiver, the error symbols will be deleted through error detector of FEC mechanism, when these erroneous bits could not be corrected. Therefore, after FEC those lost symbols can be replaced by the sRLNC redundancy from the same coded generation transmitted on the auxiliary channels. The decoding process for recovery of each generation performed by Gauss-Jordan elimination is only successful when at least k_i out of $k_i + r_i$ symbols from any coded generation i arrive at decoder. Finally N decoded generations are serialized. Thus, in order to temporarily store symbols until a generation is complete and decoding process can start, the buffering at the receiver is required. However, it should be noted that the buffering in high speed systems is a big engineering challenge.

III. Analysis

Next, an analysis of code rate and minimal transmission rate of auxiliary channels is presented. Without loss of generality, we assume that the minimal Hamming distance of a generic FEC code with a code rate R_F for the main channels and auxiliary channels are $\Delta^k_{min} = \frac{k_i \cdot s}{R_F} - (k_i \cdot s)$ and $\Delta^r_{min} = \frac{r_i \cdot s}{R_F} - (r_i \cdot s)$, respectively. Additional assumption is that any bit errors of redundant data from the auxiliary channels can be corrected during FEC operation resulting in no residual BER. Thus, the Hamming distance is determined as

$$\Delta_{min} = \frac{k_i \cdot s}{R_F} - (k_i \cdot s) \equiv \Delta^k_{min} \tag{1}$$

Based on Eq. (1), for each main channel, the number of erroneous bits that can be detected is given

$$t_e = \Delta_{min} - 1 \tag{2}$$

Table I
LIST OF NOTATIONS.

Notation	Meaning
P_i^b	Residual bit error rate on the ith main channel.
P_i^s	Residual symbol error rate on the ith main channel.
p_i^e	Expected BER on the ith main channel.
s	Symbol size in bits.
R_F	Code rate of FEC code.
R_L	Code rate of sRLNC.
K	Total size of generation data in symbols.
k_i	Size of the ith generation in symbols.
R	Total amount of redundancy in symbols.
r_i	Redundancy of the ith generation in symbols.
M	Total size of coded generation in symbols.
N	Total number of main and auxiliary channels.
C_{main}^i	Transmission rate of the ith main channel.
C_{aux}^i	Transmission rate of the auxiliary i^{th} channel.
τ_{main}^i	Propagation delay on the ith main channel.
τ_{aux}^i	Propagation delay on the ith auxiliary channel.
t_{main}^i	Transmission delay on the ith main channel.
t_{aux}^i	Transmission delay on the ith auxiliary channel.
T_{main}^i	Total delay on the ith main channel.
T_{aux}^i	Total delay on the ith auxiliary channel.
d_{main}^i	Transmission distance of the ith main channel.
d_{aux}^i	Transmission distance of the ith auxiliary channel.
c_p	Propagation speed of light, $c_p = 3 \cdot 10^8$m/s.

, and the number of erroneous bits that can be corrected is expressed

$$\begin{cases} t_k = \frac{\Delta_{min}-2}{2} & if\ \Delta_{min}\ is\ even \\ t_k = \frac{\Delta_{min}-1}{2} & if\ \Delta_{min}\ is\ odd \end{cases} \tag{3}$$

, then the expected residual BER P_i^b on any i^{th} main transmission channel after FEC process of correcting bit errors with expected BER p_e^i at receiver can be calculated as

$$P_i^b = \frac{(k_i \cdot s \cdot p_e^i) - (R_F \cdot t_k)}{k_i \cdot s} \tag{4}$$

, where all bit errors of the ith main channel are totally corrected by FEC in the case of $P_i^b \leq 0$. Using Eq. (4), the residual symbol error rate on any i^{th} main channel caused by some of bit errors that cannot be corrected by FEC process is determined as

$$P_i^s = 1 - (1 - P_i^b)^s \tag{5}$$

As a result, if the expected number of symbol errors per any generation i is $P_i^s \cdot k_i$, then to successfully decode source data, receiver needs at least $P_i^s \cdot k_i$ redundant coded symbols sent over the corresponding i^{th} auxiliary channel, i.e.,

$$\forall i \in N:\ r_i \geq P_i^s \cdot k_i \tag{6}$$

, where the number of additional coded symbols r_i sent on the ith auxiliary channel is considered as a function of the residual symbol error rate of the ith main channel, while any generation size k_i is constant.

Thus, the total size of native data symbols sent over all main THz channels can be calculated as.

$$K = \sum_{i=1}^{N} k_i \tag{7}$$

and the total number of redundant symbols sent over all additional auxiliary channels is determined with Eq. (6) as.

$$R = \sum_{i=1}^{N} r_i = \sum_{i=1}^{N} (P_i^s \cdot k_i) \tag{8}$$

, where $R \geq 0$. Thus, the total number of the native and redundant symbols sent over all $2N$ THz channels can be determined as

$$M = K + R = \sum_{i=1}^{N} (k_i + r_i) \tag{9}$$

, where $k_i \leqslant k_i + r_i$, $r_i \geq 0$ and $K \leqslant M$. Note, if $N = 1$, there are in total 2 THz channels utilized, where the total number of redundant symbols transferred on the auxiliary channel is $R = r_1 = P_1^s \cdot k_1 = P_1^s \cdot K$. While the code rate R_F of the generic FEC is constant, the overall code rate R_L of sRLNC is function of channels' configuration and the number of additional symbols sent to protect all N generations.

$$R_L = \frac{\sum_{i=1}^{N} k_i}{\sum_{i=1}^{N} k_i + r_i} = \frac{K}{M} \tag{10}$$

For any main channel i, k_i native symbols of generation i arriving at the receiver after time interval are

$$T_{main}^i = t_{main}^i + \tau_{main}^i = \frac{k_i \cdot s}{R_F \cdot C_{main}^i} + \frac{d_{main}^i}{c_p} \tag{11}$$

, where t_{main}^i is the transmission delay of k_i native symbols of generation i, τ_{main}^i stands for the propagation delay of the channel i, C_{main}^i denotes a transmission rate of the main channel i, d_{main}^i is the transmission distance between sending and receiving THz antenna of the main channel i and c_p is the propagation speed of light.

For any auxiliary channel i, r_i redundant coded symbols of generation i arriving at the receiver after time interval are

$$T_{aux}^i = t_{aux}^i + \tau_{aux}^i = \frac{r_i \cdot s}{R_F \cdot C_{aux}^i} + \frac{d_{aux}^i}{c_p} \tag{12}$$

, where t_{aux}^i is the transmission delay of r_i additional symbols of generation i, τ_{aux}^i represents the propagation delay of auxiliary channel i, d_{aux}^i is the transmission distance between sending and receiving THz antenna of an auxiliary channel i and C_{aux}^i stands for a bit rate of auxiliary channel i.

The THz system should be configured so that the arriving time interval of k_i native symbols at the receiver with Eq. (11) is equal to that of r_i additional symbols with Eq. (12) to avoid a very large buffer, i.e., $T_{main}^i = T_{aux}^i$, or

$$\frac{k_i \cdot s}{R_F \cdot C_{main}^i} + \frac{d_{main}^i}{c_p} = \frac{r_i \cdot s}{R_F \cdot C_{aux}^i} + \frac{d_{aux}^i}{c_p} \tag{13}$$

. Therefore, the bit rate of any auxiliary channel i is a function of redundancy and transmission distance of main and auxiliary channels and can be determined from Eq. (13) as

$$C^i_{aux} = \frac{r_i \cdot s \cdot c_p \cdot C^i_{main}}{R_F \cdot C^i_{main} \cdot (d^i_{main} - d^i_{aux}) + c_p \cdot k_i \cdot s} \quad (14)$$

Generally, $C^i_{aux} \geq 0$, whereby $C^i_{aux} = 0$ means that it is unnecessary to create an auxiliary channel i for additional symbols. Based on Eq. (14), we can derive a constraint $R_F \cdot C^i_{main}(d^i_{main} - d^i_{aux}) + c_p \cdot k_i \cdot s > 0$, then the transmission distance of any i^{th} auxiliary channel has to be limited as

$$d^i_{aux} < \frac{k_i \cdot s \cdot c_p}{R_F \cdot C^i_{main}} + d^i_{main} \quad (15)$$

The purpose for the transmission distance limit of auxiliary channels is to maximally reduce the buffer size at the receiver.

Assume that we configure the transmission rate of all main channels is the same, i.e., $C^i_{main} = C^j_{main} = C_{main}$, all generations have the same size, i.e. $k_i = k_j = k$, the transmission distance of all main channels are the same, i.e., $d^i_{main} = d^j_{main} = d_{main}$ and the transmission distance of all auxiliary channels are the same, i.e., $d^i_{aux} = d^j_{aux} = d_{aux}$. Then, based on Eq. (14), the total transmission rate C^t_{aux} of all auxiliary channels is given by

$$C^t_{aux} = \sum_{i=1}^{N} C^i_{aux} = \frac{s \cdot c_p \cdot C_{main} \cdot \sum_{i=1}^{N} r_i}{R_F \cdot C_{main} \cdot (d_{main} - d_{aux}) + c_p \cdot k \cdot s} \quad (16)$$

IV. Numerical results

In this section, we analyze the results of THz system scenarios with $2N = 2$ channels and $2N = 4$ channels for the code rate of sRLNC R_L and for the transmission rate of auxiliary channels calculated with Eqs. (10), (6), (4) and (14), (16), respectively, whereby two channel types, i.e., Channel B ($660 - 695$ GHz) and channel C ($855 - 890$ GHz), are considered and simulated by the raised cosine filter with roll-off factor 0.4 and omni-directional antennas in [2]. The transmission rate of any main channel i used is $C^i_{main} = C_{main} = 25$ GBd/s$= 2 \cdot 10^{11} \cdot L$ bps, where L denotes the number of bits per symbol for a certain modulation level. The expected BER p^i_e of any main channel i collected from the measurements in [2] is a function of THz channel type, distance and modulation format. Assume that any auxiliary channel i in our evaluation is configured in term of transmission distance, modulation format, channel type, etc., so that the residual BER and co-channel interference on that one is approximately equal to 0. The symbol size is set to $s = 8$ bits. The total number of symbols is set to $K = 100$ symbols. The code rate $R_F = 0.73$ of generic FEC is used for all channels. The transmission distance of main channels between sending and receiving antennas is set up in interval $[200, 2000]cm$ and the modulation formats investigated are 16PSK, 8PSK, QPSK and BPSK. The transmission distance of all main channels (B and C channel) as well as all auxiliary channels is configured the same, i.e., $d^i_{main} = d^j_{main} = d_{main}$ and $d^i_{aux} = d^j_{aux} = d_{aux}$, respectively. All of main channels use the same modulation format.

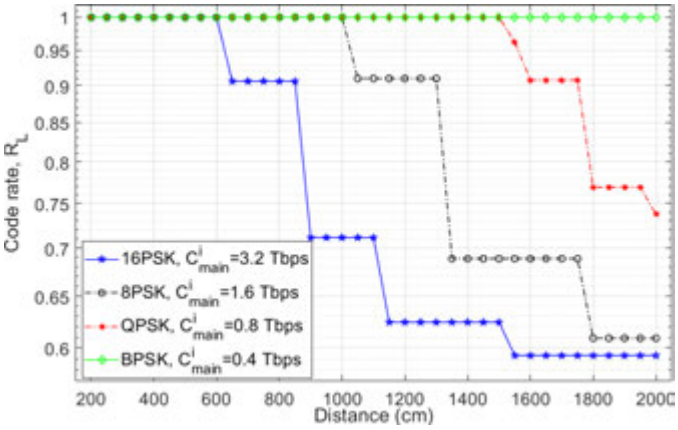


Figure 2. Code rate vs. transmission distance d_{main} in configuration with two main channels (B and C), i.e., $N = 2$.

For a THz system configuration of $2N = 2$ channels, the Hamming distance calculated with Eq. (1) is $\Delta_{min} = \frac{100 \cdot 8}{0.73} - (100 \cdot 8) = 296$ bits for any generation of $K = k_i = 100$ symbols. Then, the number of erroneous bits that can be corrected by FEC and calculated with Eq. (3) is $t_k = \frac{296-2}{2} = 147$ bits.

For a THz system configuration of $2N = 4$ channels, the Hamming distance of any channel i defined in Eq. (1) is set to $\Delta_{min} = \frac{50 \cdot 8}{0.73} - (50 \cdot 8) = 148$ bits, where $k_i = k = 50$, and the number of erroneous bits on each of two main channels that can be corrected by FEC and calculated with Eq. (3) is $t_k = \frac{148-2}{2} = 73$, i.e., 146 bits in total.

Figure 2 presents the code rate for the fault tolerant THz transmission system with $2N = 4$ channels. Thus, there are two main channels (B and C channel) and any two auxiliary channels with lower transmission rate. For BPSK, no redundancy is required and the code rate is always $R_L = 1$ because all bit errors on the two main channels (B and C) can be corrected by FEC. In the case of QPSK, the additional redundancy needs to support the main channels, when transmission distance $d_{main} \geq 1550cm$. That results in the code rate, which varies from 1 to 0.74. The minimal code rate of 0.61 in case of 8PSK is reached when transmission distance d_{main} is larger than or equal to $1800cm$. With 16PSK, the code rate changes from 1 to 0.59, when transmission distance d_{main} increases from $600cm$ to $1550cm$. We observe that with an increasing distance d_{main} and an increasing modulation level the code rate R_L reduces. When we use the modulation format on the two main channels, the code rate R_L can be independent and achieved a maximal value, i.e., $R_L \approx 1$. In addition, it is possible to use modulation format up to 16PSK with a code rate $R_L \leq 0.91$.

Figure 3 shows a comparison of the code rates in the scenarios of $2N = 2$ channels with main channel B, $2N = 2$ channels with main channel C and $2N = 4$ channels with the two main channels B and C. The presented results are for the modulation format 16PSK only, which enable to

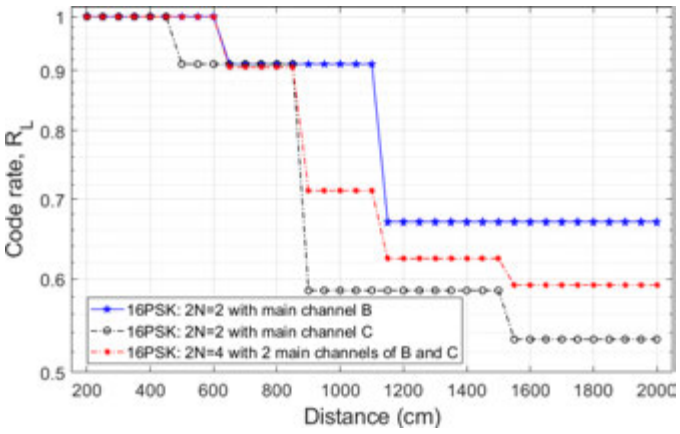


Figure 3. Comparison of code rates as a function of transmission distance d_{main} in configuration with one main channel B; one main channel C and two main channels (B and C), i.e., $N = 2$.

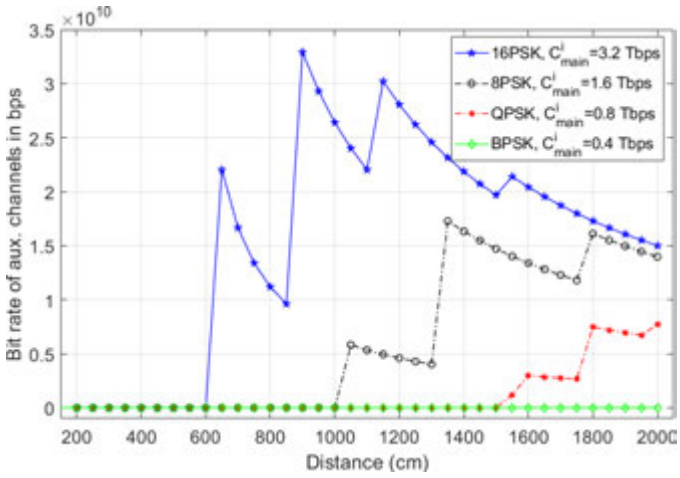


Figure 4. Total transmission rate of all auxiliary channels in configuration with $N = 2$ vs. transmission distance d_{main} of main channels (B and C), where $d^1_{main} = d^2_{main}$ and $d^1_{aux} = d^2_{aux} = 500cm$.

achieve a high transmission rate. The highest and lowest code rate can be considered for the case of $2N = 2$ channels with C main channel and B main channel, respectively. The configuration with $2N = 4$ channels, where two main channels B and C are utilized simultaneously, provides a trade off. Generally, using channel B as a main channel with 16PSK and a code rate of 0.9, the transmission distance can reach up to $d_{main} = 1100cm$ in configuration with $2N = 2$ channels.

Next, we evaluate the transmission rate of auxiliary channels required for a reliable THz system with configuration of $2N = 4$ channels in Fig. 4, whereby channel B and channel C are utilized as main channels. The transmission rate of any auxiliary channel i calculated with Eq. (14) is a function of a number of redundant symbols r_i and a transmission distance of the main and auxiliary channels, d^i_{main} and d^i_{aux}. For simplicity, we show the total transmission rate C^t_{aux} over all auxiliary channels using Eq. (16). The mean transmission rate of an auxiliary channel can be calculated as $C^i_{aux} = \frac{C^t_{aux}}{N}$, where N is the number of auxiliary channels utilized. The transmission distance of any auxiliary channel i is fixed as $d^i_{aux} = d_{aux} = 500cm = 5m$. Next, we investigate the total transmission rate C^t_{aux} considering different transmission distances and modulation formats of main THz channels. For BPSK, $C^t_{aux} = 0$ for all transmission distances of main THz channels, therefore we do not need to use any auxiliary channels because FEC code can correct all error bits from all main THz channels. Observe Fig .4, we see that the behavior of 16PSK, 8PSK and QPSK is quite similar, whereby 16PSK, 8PSK and QPSK have a fluctuation in interval $[600, 2000]cm$, $[1000, 2000cm]$ and $[1500, 2000cm]$, respectively. We take an example of 16PSK to explain the behavior. C^t_{aux} gradually decreases from $2.201 \cdot 10^{10}$ bps to $0.962 \cdot 10^{10}$ bps when the transmission distance d_{main} increases because the total amount of redundancy R sent, which is constant, i.e., 11 symbols, in transmission distance interval d_{main} $[650, 850]cm$ and also at the same time the increasing transmission distance d_{main} of any main channel i leads to a decreasing total transmission rate C^t_{aux} of all auxiliary channels, which have a fixed transmission distance $d^i_{aux} = d_{aux} = 5m << d^i_{main} = d_{main}$, and then achieves a peak at $3.290 \cdot 10^{10}$ bps because the increasing number of redundant symbols required, i.e., from 11 additional symbols at $d_{main} = 850cm$ to 41 additional symbols at $d_{main} = 900cm$. After the transmission distance of $d_{main} = 900cm$, the behavior can be similarly explained as in interval $[650, 900]cm$.

Fig. 4 shows that the total transmission rate C^t_{aux} of auxiliary channels increases when the modulation level on the main THz channels increases. Additionally, the results confirm our statement in Eq. (15) that when $d_{aux} << d_{main}$, the total transmission rate C^t_{aux} of auxiliary channels can reach the minimal value, whereby the symbol data of the main channels and the auxiliary channels will simultaneously arrive to maximally reduce the buffering overhead at receiver.

Next, we consider and compare the total transmission rate of auxiliary channels for the THz transmission system with $2N = 2$ channels configured main channel either as channel B or as channel C and with $2N = 4$ channels configured two main channels including channel B and channel C, when the transmission distance of main and auxiliary channels are the same, i.e., $d_{main} = d_{aux}$. For analyzed results with high transmission rate of the main THz channel, Fig. 5 only investigate the modulation format of 16PSK and is a function of the transmission rate d_{main} and main channel configuration. Based on Eq. (16), we see that the total transmission rate C^t_{aux} of auxiliary channel only depends on the number of additional symbols when $d_{main} = d_{aux}$. As BER increases, the increasing number of coding redundancy needs a high transmission rate of auxiliary channels. Also, if the transmission distance of main channels d_{main} insignificantly increases, then the number of additional symbols required as well as the total transmission rate of auxiliary channels C^t_{aux} will keep unchanged. For instance of the scenario with $2N = 4$ channels, the number of redundant symbols keeps unchanged with 11 symbols in interval $[650, 850]cm$ requiring the constant total transmission rate $C^t_{aux} = 0.665 \cdot 10^{12}$ bps.

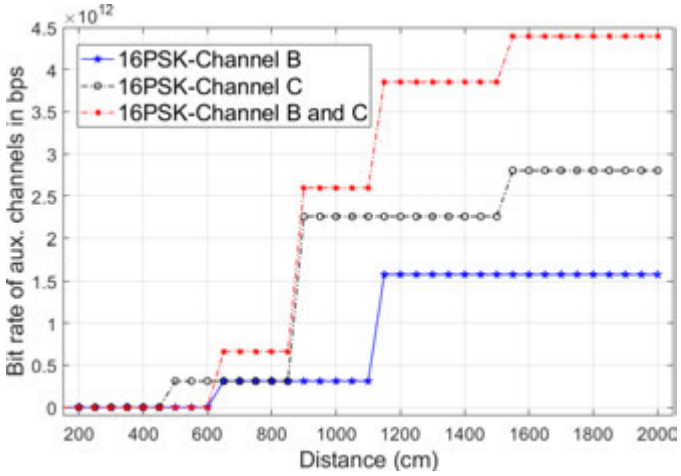


Figure 5. Total transmission rate of all auxiliary channels in configuration with one B main channel, one C main channel and with $N = 2$ (B and C channels) vs. transmission distance d_{main} of main channels, where $d^1_{aux} = d^2_{aux} = d^1_{main} = d^2_{main}$.

We observe that the transmission rate of auxiliary channel in the THz system $2N = 2$ channels with the main channel configured as channel C is always higher than that in the case of the THz system $2N = 2$ channels with main channel B because the symbols sent on the channel C often occur more error bits requiring more additional coded symbols for erroneous symbol recovery at receiver. On the other hand, although the configuration with $2N = 4$ channels provides a trade off, its total transmission rate is the highest because its generation size with $k = 50$ symbols is less than $K = 100$ symbols of the system with $2N = 2$ channels as derived from Eq. (16). For example, with the same transmission distance of $d_{main} = d_{aux} = 2000cm$, the THz system of $2N = 2$ channels with main channel B, $2N = 2$ channels with main channel C and $2N = 4$ channels need auxiliary channels with the total transmission rate $C^t_{aux} = 1.577 \cdot 10^{12}$ bps, $2.800 \cdot 10^{12}$ bps and up to $4.394 \cdot 10^{12}$ bps.

V. Conclusion

In this paper, we propose to use $2N$ channels with systematic RLNC to improve the error correction in THz communications. Our idea is to complement a generic low complexity FEC code by a low complexity sRLNC, whereby multiple THz channels can be simultaneously operated. For our design, we send the native data over N main channels and coding redundancy over N auxiliary channels in parallel. There are other wireless or wireline technologies that can use for auxiliary channels, such as the Radio frequency (RF) or Free Space Optics (FSO) required in Mbps area, but these ones also suffer from atmospheric, such as fog, rain, etc, or even re-using THz frequency band. The auxiliary channels should be configured the distance and modulation type and channel type so that their residual BER and co-channel interference approximately achieve 0 for a certain reliability. Our analysis evaluates and compares the scenarios of the THz system with the different number of channels in term of code rate of sRLNC, coding redundancy and the transmission rate of auxiliary channels required for a certain reliability transmission. The results show that the main channels can use a high-level modulation format and send data over longer distances, when the auxiliary channels can send additional symbols with lower transmission rate for the erasure error correction at the receiver. In future works, we will do practical implementation for our design to better understand the impact of network coding on the performance of multiple-channel THz systems.

Acknowledgment

This work was partially supported by the DFG Project Nr. JU2757/12-1, "Meteracom: Metrology for parallel THz communication channels."

References

[1] C. Han, A. O. Bicen, and I. F. Akyildiz, "Multi-wideband waveform design for distance-adaptive wireless communications in the terahertz band," *IEEE Transactions on Signal Processing*, vol. 64, no. 4, pp. 910–922, Feb 2016.

[2] F. Moshir and S. Singh, "Rate adaptation for terahertz communications," in *2016 13th IEEE Annual Consumer Communications Networking Conference (CCNC)*, Jan 2016, pp. 816–819.

[3] Z. Zhou, S. Wang, Y. Yu, Y. Chen, and L. Feng, "High performance metamaterials-high electron mobility transistors integrated terahertz modulator," *Opt. Express*, vol. 25, no. 15, pp. 17 832–17 840, Jul 2017. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-25-15-17832

[4] S. Rout and S. R. Sonkusale, "A low-voltage high-speed terahertz spatial light modulator using active metamaterial," *APL Photonics*, vol. 1, no. 8, p. 086102, Nov 2016.

[5] M. Mittendorff, S. Li, and T. E. Murphy, "Graphene-based waveguide-integrated terahertz modulator," *ACS Photonics*, vol. 4, no. 2, pp. 316–321, 2017.

[6] N. Khalid and O. B. Akan, "Experimental throughput analysis of low-thz mimo communication channel in 5g wireless networks," *IEEE Wireless Communications Letters*, vol. 5, no. 6, pp. 616–619, Dec 2016.

[7] S. Ullah, C. Ruan, and T. U. Haq, "Z-shaped dual band circular polarized microstrip antenna for thz communication," in *2019 44th International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz)*, Sep. 2019, pp. 1–2.

[8] C. T. Jiang, X. C. Tu, P. Xiao, S. M. Zhai, X. L. Guo, X. Q. Jia, L. Kang, J. Chen, and P. H. Wu, "The design of a bowtie antenna for 0.65 thz detection," in *2018 43rd International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz)*, Sep. 2018, pp. 1–2.

[9] C. Lin and G. Y. Li, "Adaptive beamforming with resource allocation for distance-aware multi-user indoor terahertz communications," *IEEE Transactions on Communications*, vol. 63, no. 8, pp. 2985–2995, Aug 2015.

[10] L. Yan, C. Han, and Q. Ding, "Hybrid beamforming architectures of terahertz communications," in *2019 44th International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz)*, Sep. 2019, pp. 1–2.

[11] *Next-Generation Channel Coding Towards Terabit/s Wireless Communications*. Zenodo, Aug. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1346686

[12] N. Wehn, "Channel coding for tb/s communications," Aug. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3360520

[13] M. Wang and B. Li, "How practical is network coding?" in *200614th IEEE International Workshop on Quality of Service*, June 2006, pp. 274–278.

[14] A. J. Thomas Kuerner, Xiaomin Chen, "Verfahren zur drahtlosen datenbertragung ber zwei drahtlos-datenbertragungskanle sowie datenbertragungseinrichtung und computerprogramm," Europisches Patentamt (EPA) WO/2016/156109, PCT/EP2016/056241, Oct. 2016.

[15] C. V. Phung, A. Engelmann, T. Kürner, and A. Jukan, "Improving THz Quality-of-Transmission with systematic RLNC and auxiliary channels," in *IEEE ICC 2020 Workshop on Terahertz Communications (IEEE ICC'20 Workshop - TeraCom)*, Dublin, Ireland, Jun. 2020.

Measurement-based optimized propagation model for urban, suburban and rural environments for UHF bands in Kosovo

Hëna Maloku[1], Zana Limani Fazliu[1], Mimoza Ibrani[1], Myzafere Limani[1,2] , Blediona Gashi[1]
[1]Faculty of Electrical and Computer Engineering, University of Prishtina, Prishtina, Kosovo
[2] Academy of Sciences and Arts of Kosovo, Prishtina, Kosovo

Abstract- **Severe spectrum shortage in conventional bands used for wireless communications has encouraged an increased interest in researching alternative frequency bands, which could potentially accommodate the exponential growth of various wireless technologies. The ultra-high frequency (UHF) band, traditionally reserved for TV broadcasters, in particular, has garnered interest, due to the fact that studies have consistently shown that the spectrum in this band is heavily under-utilized. However, in order to study the true availability of the spectrum and analyze potential scenarios for opportunistic use, it is necessary to have an accurate propagation model for the channel. This paper, derives such a model by using country-wide experimental spectrum measurements conducted in the territory of Kosovo, to optimize the parameters of known propagation models shown to fit the geographical terrain of Kosovo. The model is extended to encompass urban, suburban and rural environments. The model is verified against an additional set of experimental measurement data and shows a high level of accuracy.**

Keywords - ***TV bands, propagation model, wireless technologies, parameter optimization***

I. Introduction

The development of various wireless technologies and the spectrum shortage in conventional bands has encouraged research community to explore alternative frequency bands. Research has shown that TV band (UHF frequency band) is severely underutilized [1, 2]. This spectrum band is expected to further increase with the transition to digital transmission by TV broadcasters that operate in this band. The path-loss model plays the most important role in designing a network and in the parameters of the quality of the communication links. However, in order to assess the exact availability in this band for opportunistic usage, the propagation model that is appropriate for the country terrain has to be determined first. Previous study [3] has shown that the path-loss model that best fits the urban environment in our country is Hata model for short distances between transmitter and receivers of a network and Ericsson model for larger respective distances Since path-loss depends on many factors including the terrain conditions and environment, country-wide spectrum measurements were conducted to derive a model that fits the geographical terrain of the entire country. In previous works, empirical propagation models that are based in measurements have been used to determine the availability of UHF band in the country. However, since empirical models are very dependent on the location terrain, no single model could be best-fit for the entire country terrain profile. Thus the propagation model coefficients need to be optimized [4,5,6].

In this work, we have chosen three widely used propagation models [7, 8, 9, 10] to be considered in the optimization task using numerical analysis and simulations in Matlab. The propagation models used in this work are: Hata, Ericcson and COST 231 model. The experimental data was collected using NARDA Selective Radiation Meter SRM-3006 using Spectrum analysis mode of the device to measure the received power levels emitted by TV broadcasters in the UHF band (470-860) MHz [11] and compare them with simulation results. Optimization of the parameters of the propagation models shown to fit the geographical terrain of the country will provide a framework for the design of future heterogeneous wireless networks that are envisioned to operate in the TV band.

The rest of the paper is organized as follows: Section II describes the propagation models chosen to be used for analytical analysis. The measurement data collections are presented in Section III. The optimization process is described in section IV. The comparative analysis and result discussion is given in section V with conclusions drawn is section VI.

II. Propagation models description

The empirical propagation models used for comparative analysis, are described briefly below:

A. Hata Model

The Hata model is derived from Okumura-Hata model but is extended for distances 20-100km and it covers frequency range from 15-1500MHz. Hata's basic formulation is given for suburban and rural environments [12]:

Suburban area:

$$PL(dB) = a_1 + a_2 log f_c - a_3 log h_t + (a_4 - a_5 log h_t) log R - 2\left[\left(log\left(\frac{f_c}{28}\right)\right)^2 + 5.4\right] \quad (1)$$

Rural area:

$$PL(dB) = a_1 + a_2 log f_c - a_3 log h_t + (a_4 - a_5 log h_t) log R - 4.78(log f_c)^2 + 18.33 log f_c + 40.94 \quad (2)$$

where, f_c is the carrier frequency in MHz, h_t base station antenna height in meters, h_r is the receiver height and $R = r * 10^{-3}$ is the distance between transmitter and receiver in [km], whereas r is the respective distance in [m]. The constants a_1, a_2, a_3, a_4 *and* a_5 are model parameters that can be varied according to the environment. Typical values suburban and rural environments are: 69.55, 26.16, 13.82, 44.9 and 6.55 respectively.

B. Ericsson

The Ericsson model is an extension of Okumura-Hata model but it allows to adjust the parameters based on environment. The path loss in this model is calculated with the following formula :

$$PL_{Ericsson} = a_0 + a_1 \log(d) + a_2 \log(h_t) + a_3 \log(h_t) \log(d) - 2(\log(11.75 h_r))^2 + \mathrm{g(f)} \quad (3)$$

where, f is the carrier frequency in MHz, h_t base station antenna height in meters, h_r is the receiver height and d is the distance between transmitter and receiver in km. The g(f) parameter is defined as:

$$g(f) = 44.9 \log(f) - 4.78(\log(f))^2 \quad (4)$$

The values of the constants a_0, a_1, a_2, a_3 for different environments are: Suburban 43.20, 68.63, 12, 0.1 and Rural 45.95, 100.6, 12, 0.1 respectively [13].

C. COST 231

The COST 231 model is also an extension of Hata model and its designed to be used in frequency range from 500-2000 MHz. The expression for the path loss is given by [14]:

$$PL = a_1 + a_2 log f - a_3 log h_t - a(h_r) + (a_4 - a_5 log h_t) log d + c_m \quad (5)$$

where, d is the distance in meters, h_t and h_r are base station and receiver antenna height in meters and f is frequency in MHz. Parameter c_m is defined as 0 dB for suburban and rural environment. For suburban and rural environment, the parameter $a(h_r)$ is defined as:

$$a(h_r) = (1.1 log f - 0.7) h_r - (1.56 log f - 0.8) \quad (6)$$

The values for a_1, a_2, a_3, a_4 *and* a_5 are: 46.3, 33.9, 13.82, 44.9 and 6.55 respectively. Although the frequency range of this model is outside of the measurements frequency range, the opportunity for factor correctness made it widely used in the bands of interest [15].

III. Measurement data collection

Experimental data was obtained from measurements all over the country. The level of power received from 7 TV transmitters (3 of which are in the capital city of Prishtina from which 2 are collocated) was recorded in more than 36 locations with 10 individual measurements performed at the same location and lasting for 10 minutes. The power received in 8 MHz channel was calculated as average of all power levels detected in 80, 100 kHz wide bins of the same channel. The locations were chosen such as to represent different types of environments for different distances from each transmitter starting from 10-100 km with an increment of 10km. The choice of the locations was also constrained by the availability of roads and accessibility. The transmitter and measurement positions shown are shown in Fig. 1. Transmitter locations are marked in red, while measurement locations are marked in yellow color.

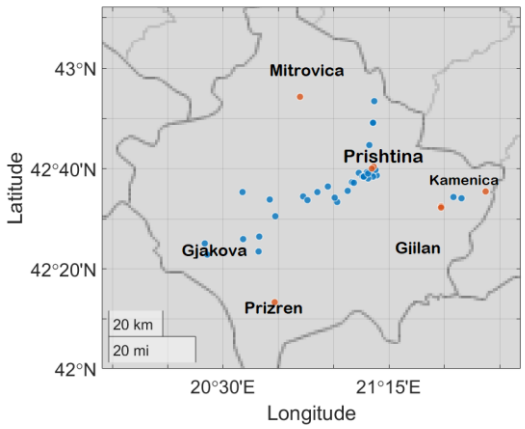


Fig.1. Measurement locations

Since we know the TV transmitter positions [11], as well as their antenna characteristics and transmit power, we are able to calculate the distance between each transmitter and measurement location, and estimate the measured path loss.

IV. Optimization

The optimization process aims to minimize the root means square error (RMSE) between the measured and predicted path-loss values [16]. The measured path values, for each transmitter, location are obtained by calculating the difference between transmitted power from transmitter t, and the measured power at location l :

$$PL_m^{t,l} = P_{tx}^t - P_{rx}^l \tag{7}$$

where P_{tx}^t is the transmitted power by transmitter t, and P_{rx}^l is the power receive at location l, in dBm. The models described in Sec. III, are applied to obtain the predicted path loss values, denoted as $PL_p^{t,l}$.

The optimization problem is therefore formulated as a minimization of the RMSE values, as follows:

$$\min_{a_i} \sqrt{\frac{\sum(PL_m^{t,l} - PL_p^{t,l})^2}{N}} \tag{8}$$

Where N is the number of samples and a_i are the models parameters that are being optimized.

The performance of standard and optimized propagation models was analyzed based on RMSE [17, 18] and error probability:

$$\Delta\bar{x} = \frac{1}{N}\sum_{i=1}^{N} \Delta x_i \tag{9}$$

where, $\Delta x_i = |PL_{predicted} - PL_{measured}|$ (10)

V. Results

The path-loss values measured and estimated with three aforementioned propagation models are plotted with respect to distance for suburban and rural environments separately. In order to compare the results and see the scale of correctness in path-loss values, the optimized model was also plotted in the same graph. The measured values of path loss are shown in Fig. 2, with black squares whereas in different colors (red, yellow, purple, and green) are shown estimated path-loss when using different standard propagation models and the optimized model respectively.

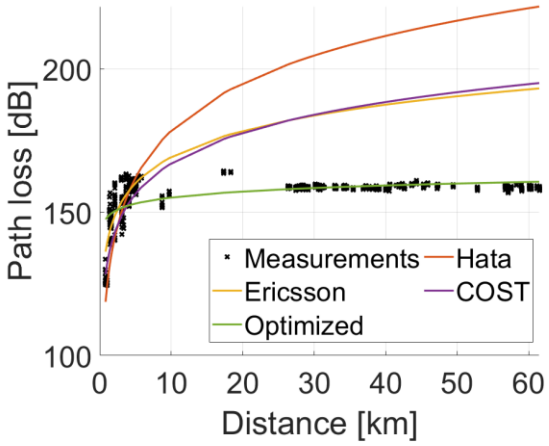


Fig. 2. Measured, estimated and optimized path-loss for suburban environment

As we can see from Fig. 2, the estimated path-loss values when applying standard models, are higher than the measured values because we don't account for losses due to shadowing and multipath effects. In the other hand, the optimized path-loss models fit almost perfectly the measured data. The optimized values for each model are shown in Table 1.

The results of model performance for the standard propagation models and optimized ones are shown in Table 2.

Table 1: Optimized propagation coefficients for suburban environments

	a_0	a_1	a_2	a_3	a_4	a_5
Hata	/	82.95	30.38	14.35	12.52	7.63
Ericsson	43.30	6.80	13.47	0.13	/	/
Cost231		52.72	37.59	13.99	17.89	7.61

Table 2: Comparison of performance of propagation models for suburban environments

	Hata	Ericsson	COST	Optimized
RMSE	44.9059	24.5237	25.4261	4.6031
$\Delta\bar{x}$	39.6186	21.7405	22.7789	2.5816

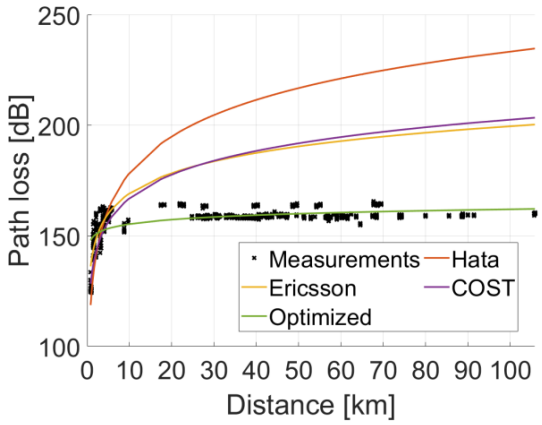


Fig. 3. Measured, estimated and optimized path-loss for rural environment

For the rural environment as well, the optimized path-loss models fit almost perfectly the measured data, as shown in Fig. 3. The optimized values for each model are shown in Table 3 below.

Table 3: Optimized propagation coefficients for rural environments

	a_0	a_1	a_2	a_3	a_4	a_5
Hata	/	81.54	30.96	14.12	12.73	8.06
Ericsson	44.38	6.49	13.10	0.12	/	/
Cost231		49.82	40.06	16.43	18.70	8.39

The results of model performance for propagation models and optimized ones in rural environments are shown in Table 4.

Table 4: Comparison of performance of propagation models for rural environments

	Hata	Ericsson	COST	Optimized
RMSE	50.0749	27.0817	28.2790	4.0598
$\Delta\bar{x}$	45.9762	24.9013	26.0998	2.4860

In Fig. 4, we have also plotted the RMSE values with respect to distance for all propagation models including the optimized ones.

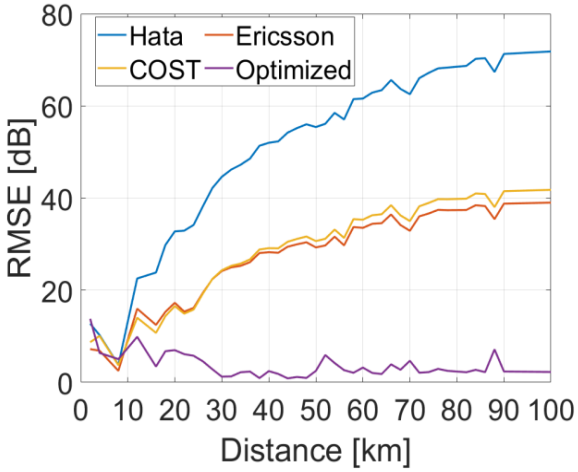


Fig. 4. RMSE of path-loss for all environments

It is evident from Fig. 4 that for both types of environments with respect to distance, the Hata model overestimates the path-loss while Ericsson and COST 231 have similar behavior. For short distances, the values of RMSE for all models are in the range of acceptable values (10-15 dB) [19], whereas for larger distances the path-loss values exceed the acceptable values. In the other hand, the optimized models show a great performance especially for larger distances.

VI. Conclusions

The true availability of spectrum is difficult to determine without the proper propagation model that best fits the type of environment. In the other hand, the propagation models are very dependent on the type of environment and terrain making them difficult to use for different locations. Therefore, the optimized version of some of the standard propagation models applied in literature for similar types of environments, was developed. The measured and estimated data of path loss values are compared to the path-loss values generated from optimized models. From the analysis it is clear that that the optimized propagation models have the highest accuracy on calculating the path-loss in terms of error probability and RMSE values for suburban and rural environments. The findings from this study will be shared with the national spectrum regulatory authorities for planning on future opportunistic usage of these TV bands.

ACKNOWLEDGMENT

This work was supported by research project "Research on the Reusability Possibilities of New Frequency Bands UHF, VHF and Millimeter Waves for Wireless Communication Networks in territory of Kosovo" funded by the Kosovo Academy of Sciences and Arts.

REFERENCES

[1] H.Maloku, Z.Limani Fazliu, M.Ibrani, A.Mekuli, E. Sela, M. Rajarajan, "Measurement of frequency occupancy levels in TV bands in urban environment in Kosovo", 2018 18th Mediterranean Microwave Symposium (MMS)

[2] Z.Limani Fazliu, H.Maloku, A.Mekuli, M.Ibrani, "Detection Threshold for TVWS Spectrum Occupancy Determination in Urban Environments in Kosovo", 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2019

[3] H.Maloku, Z.Limani Fazliu, E. Sela M.Ibrani, "Path loss model fitting for TV bands based on experimental measurements for urban environments in Kosovo", 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2019

[4] R Mardeni, TS Priya, " Optimised COST-321 Hata models for WiMAX path loss preditction in suburban and open urban enviornments", Modern Applied Science, Vol4, No.9, 2010

[5] Segun I. Popoola, et.al, "Calibrating the Standard Path Loss Model for Urban Environments using Field Measurements and Geospatial Data", Proceedings of the World Congress on Engineering 2017 Vol I WCE 2017, July 5-7, 2017, London, U.K.

[6] Faruk, Nasir; Adediran, Yinusa. A.; Ayeni, Adeseko A., "Optimization Of Davidson Model Based On Rf Measurement Conducted In Uhf/Vhf Bands", ICIT the 6th Conference on Information Technology, 2013

[7] P. Prajesh, and R.K. Singh, "Investigation of outdoor path loss models for wireless communication in Bhuj." 171-178.

[8] J. Chebil, A.K. Lawas, and M. D. Islam, "Comparison between measured and predicted path loss for mobile communication in Malaysia." World Applied Sciences Journal 21 (2013): 123-128.

[9] Bolli Sridhar and Mohammed Zafar Ali Khan, "RMSE comparison of path loss models for UHF/VHF bands in India", 10th IEEE Region Symposium,2014, Malaysia

[10] F.S. Silva, L.J. Matos, F.A.C. Peres, and G.L. Siqueira, "Coverage prediction models fitted to the signal measurements of digital TV in Brazilian cities." In Microwave & Optoelectronics Conference (IMOC), 2013 SBMO/IEEE MTT-S International, pp. 1-5. IEEE, 2013

[11] Kosovo Independent Media Commission "Gjendja reale dhe e planifikuar e planit frekuencor televiziv të Republikës së Kosovës", http://www.kpmks.org/index.php?mod=materiale&id=830

[12] R. Mardeni and K. F. Kwan, "Optimization Of Hata Propagation Prediction Model In Suburban Area In Malaysia", Progress In Electromagnetics Research C, Vol. 13, 91–106, 2010

[13] Zakaria, Yahia, Jiri Hosek, and Jiri Misurec. "Path Loss Measurements for Wireless Communication in Urban and Rural Environments." *American Journal of Engineering and Applied Sciences* 8.1 (2015): 94. Radio Wave Propagation Mechanisms And Empirical Models For Fixed Wireless Access Systems

[14] Lusilao Zodi, Milanović Josip, Snježana Rimac-Drlje, and Ivo Majerski. "Radio wave propagation mechanisms and empirical models for fixed Wireless access systems." *Tehnički vjesnik* 17.1 (2010): 43-53.

[15] V.S.Abhayawardhana , I.J.Wassell, D.Crosby, M. P. Sellars and M.G. Brown" Comparison of empirical propagation path loss models for fixed wireless access systems", IEEE 61st Vehicular Technology Conference, Sweden,2005.

[16] N. Faruk, A. Ayeni, and Y.A. Adediran, "On the study of empirical path loss models for accurate prediction of TV signal for secondary users." Progress In Electromagnetics Research B 49 (2013): 155-176.

[17] Pooja Prajesh and Dr. R.K.Singh, "Investigation Of Outdoor Path Loss Models For Wireless Communication In Bhuj", Internatinal Journal of Electronics and Communication Engineering & Technology, Volume 3, Issue 2, pp.171-178,2012

[18] Segun I. Popoola, Et.al, "Standard Propagation Model Tuning for Path Loss Predictions in Built-Up Environments", Springer International Publishing, ICCSA 2017, Part VI, LNCS 10409, pp.363-375

[19] Blaunstein, N., D. Censor, D. Katz, A. Freedman, and I. Matityahu, "Radio propagation in rural residential areas with vegetation," Progress In Electromagnetics Research, Vol. 40, 131– 153, 2003.

BUSINESS PROCESSES & DATABASES

Transforming Product Catalogue Relational into Graph Database: a Performance Comparison

Josip Lorincz*, Vlatka Huljić**, Dinko Begušić*
* Faculty of electrical engineering, mechanical engineering and naval architecture (FESB)/Department of electronics and computing, University of Split, Split, Croatia, e-mail: josip.lorincz@fesb.hr; dinko.begusic@fesb.hr
** Kron d.o.o., Split, Croatia, e-mail: vlatka.huljic@kron.hr

Abstract* - Due to the continuous expansion of telecom operators (TOs) services and subscriber's base, a number of objects in TO product catalogue (PC) traditionally developed employing some relational database (RD) rapidly increase. This can result in degraded PC database performance in terms of reduced object visibility or slow query response time, which mandates a search for new types of databases that can cope with such challenges. A possible solution to these challenges is the transformation of RD into some graph database (GD), with visualisation and organisation capabilities that can cope with the increasing complexity of PC RDs. Hence, in this paper, a novel implementation concerning the transformation of a set of PC data and functions of the TO RD into ArangoDB GD is proposed. The proposed approach allows the upgrading of PC RD into a higher presentation layer through the implementation of the GD, which enables advanced data analysis and visualization. The developed solution shows that GD with appropriately structured queries, data and links between them, improves the visibility and simplifies the entire search process. Obtained results show that GD outperforms RD in terms of search query response time for different functions executed on an equal amount of database objects.

***Keywords* - Product Catalogue, eTOM, SID, REST API, Graph, Relational, Database, ArangoDB, NoSQL, BSS/OSS**

I. INTRODUCTION

The telecom operator's (TOs) product catalogue (PC) containing all details of the products (services) to costumers is a very important module for their business. The PC structure defined with Enhanced Telecom Operations Map (eTOM) standards consists of entities and relationships based on which, packages and telecommunication service offers to the costumers are formed. eTOM is a part of Frameworx, developed by the TeleManagement Forum (TM Forum) [1]. The TM Forum is an industry association focused on transforming business processes, operations, and systems for managing and monetizing online information, communication, and entertainment services [2]. Hence, Frameworx is a suite of best practices and standards that provide the blueprint for effective and efficient business operations of enterprises such as TOs.

TM Forum Frameworx contains the following frames: application framework, business process framework, information framework and integration framework. The business process framework or eTOM describes different levels of enterprise (TO) processes, according to their significance and priority for the business [3]. Shared Information Data (SID) model or the information framework describes the information entities over which the processes operate, with their characteristics and relationships. SID provides definitions for all the information that passes through the enterprise among business partners and corresponding service providers. SID domain corresponds to the lowest level of the eTOM model and is central to the TOs next generation operations system support (NGOSS). This model represents a key principle for enhancing operations support systems (OSS) and business support systems (BSS) processes interoperability of TOs [4].

Therefore, the SID model offers a "common language" for software developers and integrators for describing information management of TOs processes. It enables simpler and more efficient integration of BSS/OSS software applications deployed by different vendors into the TO infrastructure. SID model takes into account the principles and concepts needed for defining shared information. This model also precisely prescribes many of the business elements (known as "*entities*") of interest to TOs and corresponding *attributes* that specify these entities (known as Aggregate Business Entities (ABSs)).

SID as model covers a vast area of TO activities and many companies in the world are working on the development and implementation of TO systems covering SID related process. Currently, the SID model is based on seven domains (market and sales, product, customer, service, resource, engaged party and common business entities domain) [5]. Product domain presented in Figure 1. contains different ABEs, which deals with product, product specification, product offering, strategic product portfolio plans, product performance, product usage, product configuration, product test and loyalty [6]. In fact, SID product domain elements are those elements which are fully covered and specified within the PC of most of the TOs.

The product domain as part of the SID is one of the TO system domains which has been analysed in this paper and PC as the main component of the SID product domain has been developed utilizing graph database (GD) concept. Since the number, lifecycle and contract operations related to TO products' can be huge, especially for large TOs, PCs eventually become very complex and hard to manage by means of relational databases (RDs) which are commonly used so far for PC management. Thus, in this paper, several functions of the real PC used by a prominent European TO are extracted from an implemented RD and transformed into the newly developed multifunctional GD. The capabilities and functions of a GD implemented on a real model of PC are explored and compared with that equivalent in the RD.

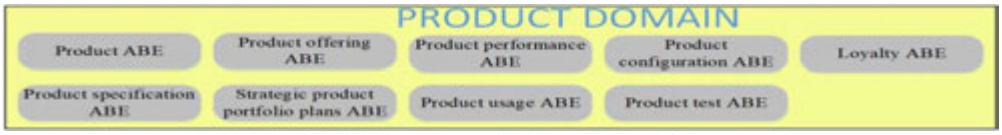


Figure 1: Product domain aggregate business entities of the SID model [6]

The comparison of the two PC database implementations shows that the proposed GD implementation outperforms the RD implementation.

The remaining structure of this paper is organised as follows: related work presenting an overview of relevant literature dedicated to the transformation of RD into GD is discussed in Section II. Section III explains research motivation and describes the analysed PC of TO. In Section IV, the transformation of the RD into GD is presented. In the next Section V, the functionality of PC as GD is described. A comparison of developed multifunctional GD and RD in terms of performance are shown in Section VI. Finally, some concluding remarks are given in Section VII.

II. Related Work

There has been some extent of work in the area of transforming an RD into a GD. Soussi in [7] analyses the possibility that an RD could be transformed into some other model that can further be used for developing the GD. Some of the models proposed in the literature are based on eXtensible Markup Language (XML), Entity-Relationship (ER) or Resource Description Framework (RDF) [8-10]. The basic drawback of these approaches is in the lack of possibility to detect interactions among generated nodes without the appropriate ontology which is needed for defining tables and data being built beforehand. Hence, efficient algorithms for converting RD into GD by building a graph through some other in-between model are yet to be proposed.

The other approach to the transformation of RD into GD is based on tools that enable feedback from an RD in the form of a graph, through posing a query in a format of search keywords that could be found somewhere in the database tuples. Such tools proposed in literature differ in modelling the tuples as the graph nodes which are connected by edges representing foreign keys or transitive tables [11-14]. Recently proposed tools for generating a GD from an RD enable graph mining techniques on relational data, however, the drawback of these tools is that they rely on the user which must strictly define the data being converted [15-16]. Another approach in transforming RD to GD proposed in [17] and [18] are based on conceptual graphs or relation-of-relations graphs, respectively. The first approach is not suitable for frequent subgraph mining from the perspective of entities, while the second approach, although having appropriate conversion duration, lacks testing on larger databases. Approaches for converting the entire RD to a GD are proposed in [19 -21]. The approach presented in [19] is based on transforming dependency graphs for the entities in the system into a hypergraph model for the RD, which can be used to develop the domain relationship model that can be converted into a GD model. In [20], an approach to automatically migrate data and queries from RD to GD is presented. Algorithms proposed in [21] are independent of the data semantics and provide the conversion between RD and GD without data loss.

Presented overview of solutions for transferring RD into GD confirms that there is no ideal tool that can merge automatic transfer of any RD into (any) GD. Due to differences in structure and data organization of versatile RDs and GDs, the conversion tools must be adopted for specific RD and GD, in order to read the relational data of RD and load it into the GD in as few steps as possible. Hence in this paper, another approach to transferring RD characteristic for PC of TOs into contemporary GD is presented.

There are a number of GDs currently available, such as Sparksee (formerly known as DEX), AllegroGraph, Neo4j, HyperGraphDB, Infinite Graph, ArangoDB, OrientDB, etc. and comparison of these and some other GDs are given in [22-23]. According to [23], ArangoDB and Neo4j GD obtain the highest scores in terms of comparison based on features such as flexibility, query language, sharing, backups, multimodality, multi-architecture capabilities, scalability and cloud readiness. This motivates the selection of ArangoDB as one of the best candidates for analyses presented in this paper. Although ArangoDB and Neo4j outperform other currently available GDs in terms of mentioned features, results of performance comparison among RDs and Neo4j GD have been already reported in [24-25]. However, to the best of our knowledge, a performance comparison between the RD and ArangoDB GD on the concrete application has not been reported in the literature. To fill this gap, this paper presents performance comparison among ArangoDB GD and an RD for PC as the specific TO use case.

III. Research Motivation for Product Catalogue Transformation

To achieve the functionality of the PC, modules that contain product information such as the basic entities (the general names for any service), attributes, prices, product offerings, etc. are required [26]. The existence of the relationship between these entities enables the creation of complex products grouped in packages offered to the users by the TO. The structure and number of elements of each TO PC can differ in complexity and size.

The data model of the PC module analysed in this work contains different basic entities (attributes, products, product relations, attribute mapping, products business rules, product offerings, prices, price lists and sales context) which are presented in Figure 2. Links between entities of the PC model emphasize the complex structure of the PC.

In the presented analyses, part of the real PC model used by one European middle-size TO was originally

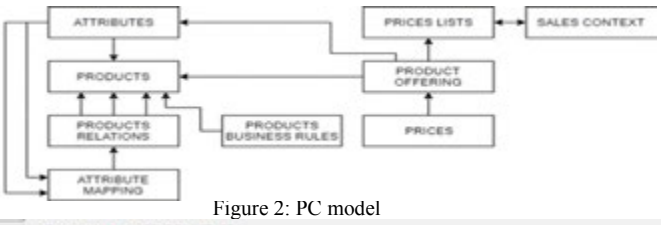


Figure 2: PC model

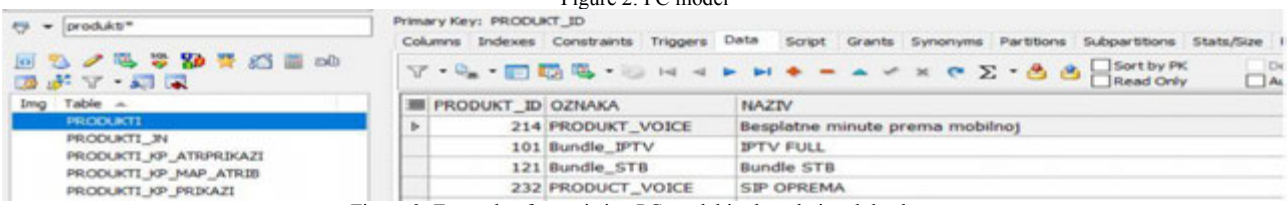


Figure 3: Example of an existing PC model in the relational database

implemented and used in the form of the RD. An example presenting a listing of the PC elements in the form of the table, with main products structured in the RD are shown in Figure 3. A *Toad* software tool was used to work with RDs (Figure 3). This tool allows developers, database administrators and database analysts to manage RDs and non-relational databases by Structured Query Language (SQL) queries.

Since it is challenging to cope with the multitude of data and have a good overview of all database elements, administrators of such RDs start to have the problem of managing the PC. The question that arises was can some other PC data structure offer faster PC database search and better visualisation of PC elements, especially in light of necessity for the continuous expenditure of TOs PC with new products and corresponding elements. The transformation of the existing RD into new datasets with an upgrade to the GD was a valuable option to the TO. Hence, the research presented in this paper was motivated with the goal related to the transformation of data from an RD into the GD format. This can gain a new visualisation dimension of PC and additional information that will serve for better analysis of the PC data. The idea was to use the graph of the database to attribute a PC in order to get information about: where the services are used and under what conditions, what are the geographic locations of services used by users and how used services are represented. All this information cannot be easily visualized in existing TO RD. However, such information is important for TO marketing, which must optimize offering of the service in the geographical area covered by TO and also for the technical implementation and maintenance of the TO equipment including fault recovery.

IV. Implementing Product Catalogue by Means of Graph Database

According to presented in the previous section, PC database management can be a difficult task when it comes to large amounts of data, and the problem can be compounded when multiple data formats are involved. For the development of GD, in this work, the open-source ArangoDB as not only SQL (NoSQL) database type having GD capabilities were selected for administration of PC data transformed from Oracle v12c RD. This database type is specifically designed to manage multiple data formats in a single instance. The ArangoDB provides GD capabilities, database document management and key-value storage [27].

Unlike RDs that have a strictly defined data schema, ArangoDB has a dynamic data schema [28]. Data in ArangoDB are stored as JavaScript Object Notation (JSON) documents. The documents are grouped into "collections" that correspond to a table in RDs. The documents in the collection have a unique *primary document key* that automatically stores and encodes the identity of the document, thus storing the document in the key/value container. The created collection consists of documents stored as vertices and edges in the form: *_from* the source point value *_to* the value of the target vertex [29]. Combining vertices and edge collections produce a graph. The proposed transformation of data from the RD into GD is performed through the following phases:

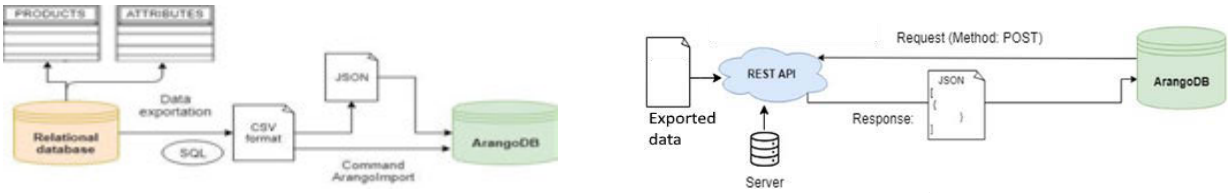
A. *The first phase: export data from a relational database*

The first phase in the process of transforming PC from the RD into GD structure is related to the export of selected data from an RD into a GD, more precisely into the ArangoDB database (Figure 4). These data operations are performed using SQL queries (Figure 4a).

The first two queries extract the data that will be entered into the vertices collections of GD, and the third query extract data which will be extrapolated into the edges collection of GD. The first phase ends when data from the RD are extracted in a .xls file format that is stored in Comma Separated Values (CSV) data format (Figure 4a).

B. *The second phase: import data into ArangoDB*

The second phase of transferring RD into GD is dedicated to the import of exported data form RD into the GD (Figure 4). The ArangoDB contains a command-line client tool called *Arangoimport* (Figure 4a). This client command tool can be used to import JSON, CSV and tab-separated values (TSV) format data into ArangoDB. In our analyses, JSON and CSV as common data formats are used (Figure 4a). *Arangoimport* is called from the command line every time when each file (JSON or CSV) wants to be imported into the GD (Figure 4a). Hence, the



a) b)

Figure 4. The proposed concept for transforming relational into GD based on relational data export and: a) direct SQL queries, b) REST API methods

CSV data format can be imported directly (Figure 4a), or it can be formatted into JSON format using format converter and imported in GD (Figure 4a). The data importing results with the transfer of PC data objects consisted of attribute-value pairs and array data types in GD. When data import is done, it is possible to continue working with the same data model previously used.

Another way of importing data from an RD is to expose the ArangoDB application programming interface (API) over the hypertext transfer protocol (HTTP) server, making the server easily accessible for different clients and tools (Figure 4b). Data exported in the first phase are reachable through exposed API and converted into JSON format for purpose of transmission into ArangoDB. According to Figure 4b, the JSON data is sent and received over a well-known HTTP server protocol which provides a RESTful API (REST API).

The REST API methods (GET, POST, PUT, DELETE, PATCH, etc.) represent the architectural style used for creating a web service for the transfer of resource state views (Figure 4b). Usage of these methods is popular in practice due to their simplicity and the fact that they rely on existing HTTP protocol and corresponding features, while HTTP defines the format and method of messaging [30].

C. *The third phase: creating collections in ArangoDB and graph formations*

In ArangoDB, documents are stored in collections. Collections have names, which describe what kind of information the documents contain. Each document requires a key that uniquely identifies it within a collection. It can be assigned by the user upon creation, or ArangoDB will generate one. Document keys and document identifiers (IDs) are always indexed. The index on the *_key attribute* is called a primary index. It exists for each collection and can't be removed. There are two types of collections in ArangoDB: vertex and edge collections (Figure 5). Documents in an edge collection have two additional attributes, *_from* and *_to*. Both must have document IDs to link documents together. The document that links them is called an edge and the linked documents are called vertices in the graph model (Figure 5). When creating a graph model based on an RD model, it is important to be aware that the row in the RD is equivalent to a node in the GD, and that the table name is equivalent to the node label in the GD model.

The ArangoDB uses ArangoDB query language (AQL) for executing procedures over graphs, as well as over individual collections. Although some keywords overlap, AQL syntax is different from RD SQL syntax. The most notable difference is in the concept of loops in AQL, which makes AQL look like programming language [31]. AQL has a language construct for looping through data and the most common form is the FOR loop used with a collection of documents (similar to SELECT in SQL). Complex object and array manipulations can be done using AQL queries (INSERT, UPDATE, REPLACE, DELETE, etc.) and operators (FILTER, SORT, LIMIT, etc.), what makes AQL very powerful query language that remains easy to write and read.

V. THE FUNCTIONALITY OF PRODUCT CATALOGUE AS GRAPH DATABASE

As previously indicated, the PC presents description and characteristics of TO services offered to customers and contains all necessary information related to the specific product. An example of graphical representation obtained through developed GD for specific TO product is presented in Figure 5. The graphical representation obtained through implementation of GD shows the relationship between templates (e.g. user mail addresses ("MAIL"), contracts ("OTT ugovor"), etc. in Figure 5) of PC and their attributes which can be characteristic for the specific template (e.g. additional service ("Dodatna usluga"), service and service packet ("Usluga i paket usluge"), etc. in Figure 5) or shared among different templates (e.g. service name ("Naziv usluge"), name and surname of the owner ("Ime i prezime vlasnika"), etc. in Figure 5).

The template is prepared textual document used when generating a new document instance and its attributes are used to describe an entity or to extend the meaning of an entity. This means that if there is a link between templates and their attributes, internal logic generates a document with its attributes representing dynamic fields in it. The view is filtered by the name contained in the template or by attribute description, what generates query results as those presented in Figure 5. This approach gives a credible view of all related and unrelated elements in the graph of GD. In this way, an error can easily be detected if there must be a connection between PC elements (templates and attributes), but no link exists in the graph model. Unlike RDs where data search is a long process, GD improves the entire search process by simple visualization of the data and the links between them (Figure 5).

Services offered by the TO are tracked through the concept of the customer's assets. They represent the state of services that the customers use at any given time. For TOs, it is important to have a correlation between used services and corresponding products and locations where they are installed. As an example, for some selected dataset, Figure 6 presents a relationship between products

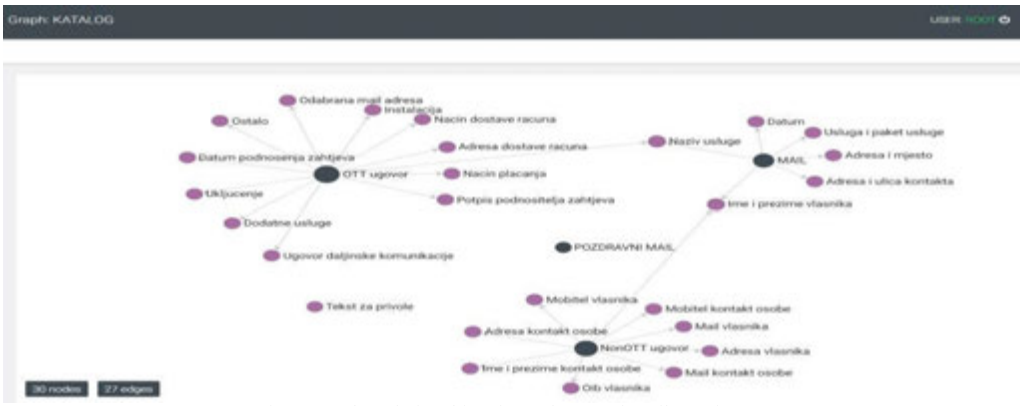


Figure 5. The relationship of templates and attributes in GD

and locations of product usage, where templates (e.g. "SESVETE", "ZAGREB", etc. in Figure 6) represent locations of used products, while attributes (e.g. S service packet no. 19 ("S paket 19"), L service packet no. 19 ("L paket 19"), etc. in Figure 6) identify a specific product. The selected dataset realistically shows the relationship of products sold to a customer in a particular place of residence (Figure 6). Displaying the density of services and corresponding product distribution by location (place), gives the TO an insight into the actual state of use of the specific service and corresponding product. This information is useful for making decisions about the need to change the current state, i.e. create new services and corresponding products or use existing ones in the areas not covered with specific TO service.

Hence, developed ArangoDB GD ensures one of the main advantages of GDs, which is performing simple queries on a diverse set of data, what results with a comprehensive and detailed view of the requested information.

VI. Performance Analyses

Performance analysis is based on a comparison of developed multifunctional GD (ArangoDB) and existing RD (Oracle). Analyses are based on the measurement of SQL and AQL query execution response time. For realistic comparison, the same amount of data related to the PC of TO was used for comparing query response time in RD and GD. This measurement is performed for the three queries (Q1, Q2, and Q3) executed in the RD and GD (Figure 7). Queries differ in terms of scope and complexity of the search, which is directly related to the number of analyzed database elements.

The Q1 query in the RD search for a custom set of data (e.g. places, products, link type, etc.) and 90 database elements are the result of the search. The Q1 query execution time in the RD is 109 ms, while in the ArangoDB GD is significantly faster and equals to 0.695 ms (Figure 7).

The Q2 query provides information about which attributes are attached to the selected template and gives 9 elements as a result. The speed of query execution differs significantly in this example, because the GD has a great advantage in terms of execution time which is equal to 0.960 ms and the 43 ms for the RD (Figure 7).

To have reliable analyses, a larger amount of PC data was taken in the next example with the Q3 query, where the query results in 1000 database elements. The Q3 query search for all the attributes contained in the products table (collection) of the PC (e.g. group ("GRUPA"), type ("VRSTA"), model ("TIP"), line ("LINIJA")). In this case, query execution time in the RD is 226 ms (Figure 7), while in the GD equals to 5.355 ms.

Based on the measured execution time of individual queries, it can be concluded that query execution in the GD is significantly faster compared to the RD (Figure 7). For some query types, GD outperforms RD with more than 42 times faster response time. Also, in Figure 7 can be seen that better performance is evident for GD where the query execution time does not increase significantly with the increase in the number of search data objects, which is not the case with RD. Hence in developed GD, queries are executed faster and outcomes of queries give more comprehensive results in terms of their visualization. The faster execution time of GD is a direct consequence of the better structural organization of ArangoDB GD, what confirms that this type of GD is suitable for transforming the PC of TO from a conventional RD.

VII. Conclusion

The PC structure based on eTOM standards contains entities and relationships according to which packages of telecommunication services are formed and offered to users. PCs are traditionally developed in the form of an RD. Due to the constant increase in a number of services and subscriber base, the complexity of relational PC database management became challenging for TO database administrators. This requires new concepts in the development and management of the relational PC databases of TOs. This paper gives a contribution to devising such new concepts through the presentation of a new approach for transformation of TO relational PC database into GD. The process phases of transforming RD (Oracle) into GD (ArangoDB) are described, with the presentation of developed GD structure and performance comparison among analyzed databases.

It is shown that GD with advanced queries supported through AQL, graph data structure and links between GD elements, simplify the entire search process and it is applicable for use when a larger amount of different data structures must be managed. A comparison of analysed

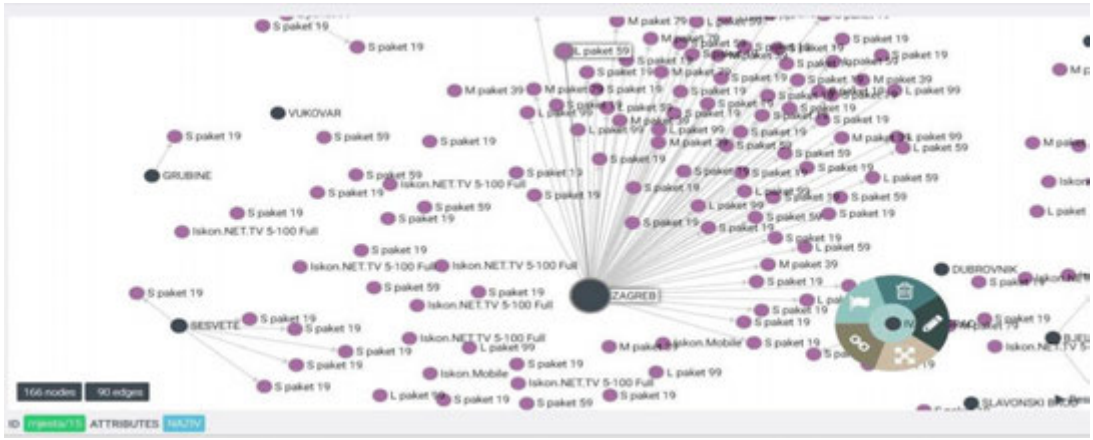


Figure 6. Relationship between product and location of product usage presented in GD

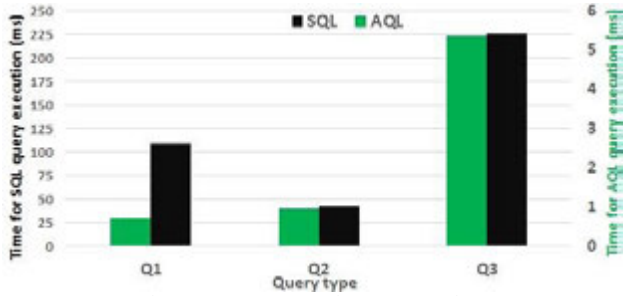


Figure 7. Query execution response time

databases show that GD offers a higher presentation level which enables more comprehensive data analysis and visualization. The greatest advantage of developed GD over the RD is reflected in the speed of query execution, for which GD significantly outperforms RD. Due to a very important feature of managing and storing large amounts of versatile data structures, the ArangoDB is expected to be widely used in practice, and it is particularly suitable for the development of TO PCs. Our future research activities will be focused on the automatization of data transfer form RDs into GDs.

REFERENCES

[1] Internet, August 2019, https://www.tmforum.org/

[2] Internet, August 2019, https://www.cisco.com/c/en/us/products/collateral/services/high-availability/white_paper_c11-541448.html

[3] Internet, August 2019, http://www.ilsa.kz/etom/main/tamdomain5.htm

[4] Internet, August 2019, http://www.sebconsulting.ie/understanding_ngoss_sid.html

[5] Internet, August 2019, https://www.tmforum.org/Browsable_HTML_SID_R18.5/content/3E3F0EC000E9_root.html

[6] Internet, August 2019, https://www.tmforum.org/tm-forum-frameworx-2/

[7] R. Soussi, "Querying and extracting heterogeneous graphs from structured data and unstrutured content", PhD thesis, Ecole Centrale Paris, 2012.

[8] M. Hert, G. Reif, H. C. Gall, "A comparison of rdb-to-rdf mapping languages", In Proc. of the I-SEMANTICS '11, 2011. pp: 25 – 32.

[9] W. Hu, Y. Qu, "Discovering simple mappings between relational database schemas and ontologies", In Proc. of the ISWC/ASWC '07, 2007., pp. 225-238.

[10] J. Sequeda, M. Arenas, D. P. Miranker, "On directly mapping relational databases to rdf and owl", In Proc. of the WWW '12, 2012., pp. 649-658.

[11] V. Hristidis, Y. Papakonstantinou, "Discover: Keyword search in relational databases," in Proc. of the 28th VLDB '02, 2002. pp. 670–681.

[12] S. Agrawal, S. Chaudhuri, G. Das, "Dbxplorer: enabling keyword search over relational databases," in Proc. of the ACM SIGMOD '02, 2002, pp. 627–627.

[13] V. Kacholia, S. Pandit, S. Chakrabarti, S. Sudarshan, R. Desai, H. Karambelkar, "Bidirectional expansion for keyword search on graph databases", in Proc. of the VLDB '05, 2005, pp. 505–516.

[14] H. He, H. Wang, J. Yang, P. S. Yu, "Blinks: ranked keyword searches on graphs," in Proc. of the ACM SIGMOD '07, 2007., pp. 305–316.

[15] K. Xirogiannopoulos, U. Khurana, A. Deshpande, "Graphgen: Exploring interesting graphs in relational data", Proceedings of the VLDB Endowment, vol. 8, no. 12, 2015., pp. 1-4.

[16] D. Simmen, K. Schnaitter, J. Davis, Y. He, S. Lohariwala, A. Mysore, V. Shenoi, M. Tan, Y. Xiao, "Large-scale graph analytics in aster 6: bringing context to big dana discovery," Proceedings of the VLDB Endowment, vol. 7, no. 13, 2014., pp. 1405–1416.

[17] S. Palod, "Transformation of relational database domain into graphs based domain for graph based data mining," Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, SAD, 2004.

[18] S. Pradhan, S. Chakravarthy, A. Telang, "Modeling relational data as graphs for mining" in Proc. of the COMAD '09, 2009., pp. 1-6.

[19] S. Bordoloi, B. Kalita, "Designing Graph Database Models from Existing Relational Databases", International Journal of Computer Applications, Volume 74, No.1, 2013, p.p.: 25-31.

[20] R. De Virgilio, A. Maccioni, R. Torlone, "Converting relational to graph databases," in Proc. of the GRADES '13, 2013, pp. 1-6.

[21] O. Orel, S. Zakošek, M. Baranović, Property Oriented Relational-To-Graph Database Conversion, Automatika, Vol. 57(2016), 3, 2016, p.p.: 836–845.

[22] R. Angles, "A comparison of current graph database models", in Proc. of the ICDEW '12, 2012, pp. 171–177.

[23] D. Fernandes1, J. Bernardino, "Graph Databases Comparison: AllegroGraph, ArangoDB, InfiniteGraph, Neo4J, and OrientDB", In Proc. of the DATA 2018,2018., pp. 373-380.

[24] O. Est, "Comparative analysis of relational and graph database behavior on the concrete web application", Master thesis, Tallin University of Technology, 2016, p.p.: 1-68.

[25] Yaowen Chen, "Comparison of Graph Databases and Relational Databases When Handling Large-Scale Social Data", Master thesis, University of Saskatchewan, pp: 1-91

[26] Kron Company, "Kron - Product Catalog - Assets – Order", Internal Document, 2019.

[27] Internet, August 2019, https://www.arangodb.com/arangodb-graph-course/

[28] Internet, August 2019, https://repozitorij.pmf.unizg.hr/islandora/object/pmf%3A3234

[29] J. Stücke: „From Zero to Advanced Graph Query Knowledge with ArangoDB" Internet, August 2019, https://www.arangodb.com/2017/05/zero-advanced-graph-query-knowledge-arangodb/

[30] Internet, August 2019, http://searchmicroservices.techtarget.com/definition/REST-representational-state-transfer

[31] Internet, August 2019, https://www.arangodb.com/docs/3.4/getting-started-coming-from

Značajka analitičkog pristupa u dizajnu, optimizaciji i održavanju poslovnih procesa

Darko Markulin, Kristina Musa, Ana Mršić
HT d.d.
Draškovićeva 26, 10000 Zagreb, Hrvatska
Telefon: (385) 1-4912 721
E-mail: darko.markulin@t.ht.hr, kristina.musa@t.ht.hr, ana.mrsic@t.ht.hr

***Sažetak* – Za uspješan dizajn novih poslovnih procesa ili optimizaciju već postojećih, potrebna je primjena različitih metoda pristupa u razvoju poslovnih procesa. Dobro dizajnirani poslovni procesi na području telekomunikacija preduvjet su za uspješnu realizaciju usluga, a isto tako i osnova za uspješnu analizu i optimizaciju postojećih poslovnih procesa. Različite metode pristupa koriste i različite alate, kako za dizajn, tako i za detaljnu analizu kvalitete dizajna samih procesa. Za dizajn procesa mogu se koristiti različiti alati kao što su Visio, UML ili BPMN. Neki od alata koji se koriste za analizu poslovnih procesa mogu biti Excel, pivot tablice ili PBI (power business inteligence) alati. Korištenje tih alata preduvjet su za uspješnu analizu i definiranje uskih grla u procesu. Jedan od važnijih segmenata novih alata za analitički pristup je vizualizacija koja znatno pospješuje njihovo uspješno korištenje u dizajnu, optimizaciji i održavanju složenih poslovnih procesa kao što su telekomunikacijski poslovni procesi.**

Ključne riječi – dizajn, analitički pristup, analitički alati, poslovni procesi, telekomunikacije

I. UVOD

Uspješno korištenje poslovnih procesa tijekom čitavog njihovog životnog ciklusa zahtijeva analitički pristup i primjenu različitih metoda i alata. Naročito se to odnosi na složene poslovne procese kao što su procesi u telekomunikacijama za aktivaciju i održavanje telekomunikacijskih usluga. Dobro dizajnirani poslovni procesi, prikazani kroz odabrani model procesa, jedan su od osnovnih preduvjeta za uspješnu primjenu i održavanje poslovnih procesa. Važna svojstva odabranog modela poslovnog procesa su točnost i odgovarajući način prikaza (vizualizacije) poslovnog procesa. Dobro modelirani poslovni procesi olakšavaju njihovu implementaciju u sustavima za podršku operacijama koji se koriste prilikom aktivacije ili održavanje telekomunikacijskih usluga. Isto tako, preduvjet su za uspješnu analizu i optimizaciju postojećih poslovnih procesa. Analizu postojećih poslovnih procesa dodatno olakšava primjena različitih analitičkih alata koji na vizualan način prikazuju njihovo ponašanje. Vizualizacija pomaže i u boljem razumijevanju procesa i uočavanju mogućih nepravilnosti te potencijalnih uzroka problema unutar raznih procesnih scenarija. To omogućava proaktivno djelovanje i smanjenje mogućih poteškoća u realizaciji telekomunikacijskih usluga.

U radu je opisana primjena različitih alata koji se u svrhu dizajna i analize poslovnih procesa koriste u Hrvatskom telekomu (HT). Pojašnjeni su razlozi za odabir pojedinih alata i način njihovog korištenja.

II. DIZAJN POSLOVNIH PROCESA

Specificiranje poslovnih procesa znatno je lakše korištenjem modela poslovnog procesa, koji je zapravo grafički prikaz poslovnog procesa putem dijagrama. Prilikom odabira modela pomoću kojeg se dizajnira poslovni proces bitno je definirati njegovu osnovnu namjenu. Modelom poslovnog procesa mogu se prikazivati osnovni koraci poslovnog procesa uz definiranje detalja kojima se oni detaljnije opisuju (kao što su akteri poslovnog procesa, ulazni i izlazni podaci) i uvjeti za daljnji tijek odvijanja procesa. Osnovna namjena takvog modela je opis poslovnog procesa na razumljiv način kako bi bilo jasno koje je akcija potrebno poduzeti kako bi se postigao osnovni cilj poslovnog procesa. Na primjer, aktivacija usluge od strane telekomunikacijskog operatora. S druge strane, kroz model poslovnog procesa mogu se dati takvi detalji koji na jasan način definiraju sve što je potrebno kako bi se u sustavima za podršku operacijama uspješno implementirala programska podrška za upravljanje poslovnim procesima. Primjeri takvih sustava za podršku operacijama su OSS (Operational Support Systems) sustavi koji se u telekomunikacijama koriste za upravljanjem poslovnim procesima potrebnim za aktivaciju ili održavanje telekomunikacijskih usluga. Model daje apstraktnu (pojednostavljenu) sliku sustava i njime se, između ostalog, može definirati[2]:

A. Kontekst

Određuje se granica između sustava i njegove okoline te sučelja

B. Ponašanje

Promatraju se transformacije podataka, reakcija sustava na događaje i promjenja njegovih stanja

C. Struktura

Definira se arhitektura samog sustava ili građa podataka koje sustav obrađuje.

O namjeni modela poslovnog procesa uvelike ovisi i o odabir alata za modeliranje. Ukoliko se modelom poslovnog procesa želi samo specificirati poslovni proces i kroz njegov grafički prikaz pospješiti komunikaciju s IT-om zaduženim za njegovu implementaciju u sustave za podršku operacijama, primjer alata koji se uspješno može koristiti je Visio. Visio je alat u kojem se jednostavno mogu kreirati dijagrami i u kojem se može koristiti široki spektar

template-ova s različitim notacijama (elementima) dijagrama.

Ukoliko se želi kreirati model poslovnog procesa s visokom točnošću i kojim se omogućuje simulacija poslovnog procesa, potrebno je koristiti programske alate namijenjene dizajnu poslovnih procesa i softverskih rješenja kao što su UML ili BPMN. UML (Unified Modelling Language) je standardizirani jezik koji koristi grafičku reprezentaciju za vizualno modeliranje softverskih rješenja. Mogu se koristiti različiti dijagrami kojima se sustavi modeliraju iz različitih aspekata[1]. Koristi se više vrsta dijagrama za modeliranje statičke strukture sustava i dinamičkog ponašanja sustava. Dijagram kojim se može dobro modelirati poslovni proces je UML dijagram aktivnosti. BPMN (Business Process Modelling Notation) je grafički jezik koji ima široku uporabu u modeliranju poslovnih procesa i kroz njegovu notaciju moguće je modelirati poslovni proces na precizan i razumljiv način kako poslovnim analitičarima, tako i IT stručnjacima odgovornim za implementaciju u sustave. Dijagramima se daje jasan prikaz organizacijskih struktura i odgovornosti.

Model poslovnog procesa važna je komponenta validacije poslovnog zahtjeva za sustavom za podršku operacijama. Validacija zahtjeva utvrđuje odgovara li sustav koji gradimo korisničkoj specifikaciji[2]. Validacija implementirane programske podrške provodi se njezinim testiranjem. Validacija se provodi kroz testne slučajeve, a model je vrlo koristan kod definiranja kriterija za uspješno provođenje testnog slučaja.

U HT-u se modeliranje poslovnih procesa provodi na temelju slijedećih principa:

A. Visio kao alat za izradu dijagrama

Visio se koristi za modeliranje poslovnih procesa jer je relativno jednostavan za korištenje, ima široku uporabu i jednostavno je dobivene dijagrame korisiti u sklopu specifikacija za potrebnu programsku podršku. Troškovi korištenja ovog alata nisu visoki, a korištenje drugih alata specijaliziranih za modeliranje poslovnih procesa zahtijeva dodatno specijalizirano znanje.

B. Notifikacija UML2 dijagrama aktivnosti

UML2 dijagram aktivnosti se koristi jer je njegova notifikacija pogodna za modeliranje poslovnih procesa, jezik je pogodan za specifikaciju softverskih rješenja (u ovom slučaju programske podrške) i njegovim korištenjem dobiva se na točnosti specifikacije te stvaraju preduvjeti za modeliranje složenih poslovnih procesa i eventualno korištenje modela u svrhu simulacije poslovnih procesa. U [5], [6] i [7] opisan je i definiran način korištenja modela za modeliranje složenih poslovnih procesa i notifikacija (elementi) modela . U [1] i [4] opisano je rješenje kojim je moguće kroz takav model i uz pomoć korištenja formalnih metoda postići simulaciju odvijanja procesa.

C. Modeliranje kroz više razina složenosti

Kako su procesi koji se modeliraju složeni, model poslovnih procesa podijeljen je u više razina složenosti (module)[1]. Moduli su predstavljani podprocesima. Modulima na višoj razini apstraktnosti (manja složenost) definirane su osnovne faze procesa uz njihov dodatni opis. Ovi moduli se koriste za osnovni opis poslovnog procesa i razumljivi su širem krugu sudionika u razvoju programske podrške. Slikom 1 prikazan je takav modul. Modulima na nižoj razini apstraktnoosti (viša razina složenosti) modeliraju se detalji bitni za implementaciju programske podrške. Ti detalji nisu razumljivi širem krugu sudionika u razvoju programske podrške i prvenstveno su namijenjeni za sudionike koji aktivno sudjeluju u implementaciji programske podrške u sustave. Slikom 2 prikazan je modul na toj razini složenosti.

III. ANALIZA I OPTIMIZACIJA POSTOJEĆIH POSLOVNIH PROCESA

Da bi se pratilo ponašanje i odvijanje poslovnih procesa te se na temelju toga definirali postupci za optimizaciju i poboljšanje postojećih procesa, potrebno je definirati način korištenja postojećih alata za analizu poslovnih procesa. Alati koji se vrlo često koriste za potrebe analitičkog razmatranja u različite svrhe su Excel i pivot tablice. Rezultati dobiveni korištenjem tih alata često su dobra osnova za daljnji postupak optimizacije poslovnih procesa. No, nedostatak vizualizacije i potreba za specifičnim znanjima kako bi se očitali dobiveni rezultati osnovni su nedostaci korištenja tih alata u svrhu analize i optimizacije poslovnih procesa. Power BI alati nadomještaju spomenute nedostatke vizualnim prikazom definiranih pokazatelja kroz koje se prati ponašanje poslovnih procesa. Time ti alati postaju snažno sredstvu za analizu poslovnih procesa i pomoć za definiranje slijedećih koraka u sklopu optimizacije poslovnih procesa.

Slikama 3 i 4 prikazan je način kako se u HT-u koriste analitički alati u sklopu praćenja ponašanja poslovnih procesa kroz jedan od poslovnih segmenata HT-a. Kroz definirane parametre prikupljaju se potrebni podaci na temelju kojih se prati utjecaj poslovnih procesa na uspješnost korištenja usluga od strane korisnika. Kroz Power BI alat omogućen je vizualni prikaz (dashboard) ponašanja i utjecaja poslovnih procesa na kvalitetu postojećih i novih korisničkih usluga. Primjer takvog prikaza dat je Slikom 3. Na slici 4 može se uočiti utjecaj određenog procesa na aktivnu uslugu kod korisnika nakon što se otklonila greška koja se pojavljivala tijekom odvijanja procesa. Analizom kroz postojeći model poslovnog procesa utvrđeno je da se ispravno odvijaju svi procesni koraci definirani tim modelom, ali dodatnom analizom utvrđeno je da se greška događala u samom postupku aktivacije jednog od mrežnih elemenata. Nakon ispravka spomenute greške, može se uočiti trend smanjenja utjecaja tog poslovnog procesa na broj pojavljivanja smetnji na aktivnim uslugama kod korisnika.

Kako bi se olakšala analiza procesa za koji se zaključilo da je potrebna njihova optimizacija u cilju smanjenja njihovog negativnog utjecaja na uspješnost korištenja usluga od strane korisnika, aktivno se koriste modeli poslovnih procesa nastali tijekom njihovog dizajna. Kroz model se lako može pratiti tijek odvijanja aktivnosti te utvrditi jesu li odrađene sve aktivnosti i je li ispravan njihov tijek. Naravno, često je potrebna dodatna analiza kao što je analiza aktivnosti unutar različitih sustava za podršku operacijama koji se koriste za upravljanje poslovnim procesima. Takav primjer je sustav za aktivaciju mrežnih elemenata koji pokreće svoju aktivnost na temelju akcija



Naziv procesa: **HALO uključenje (MSAN)**

Proces se pokreće ukoliko korisnik uključuje samo Halo uslugu.

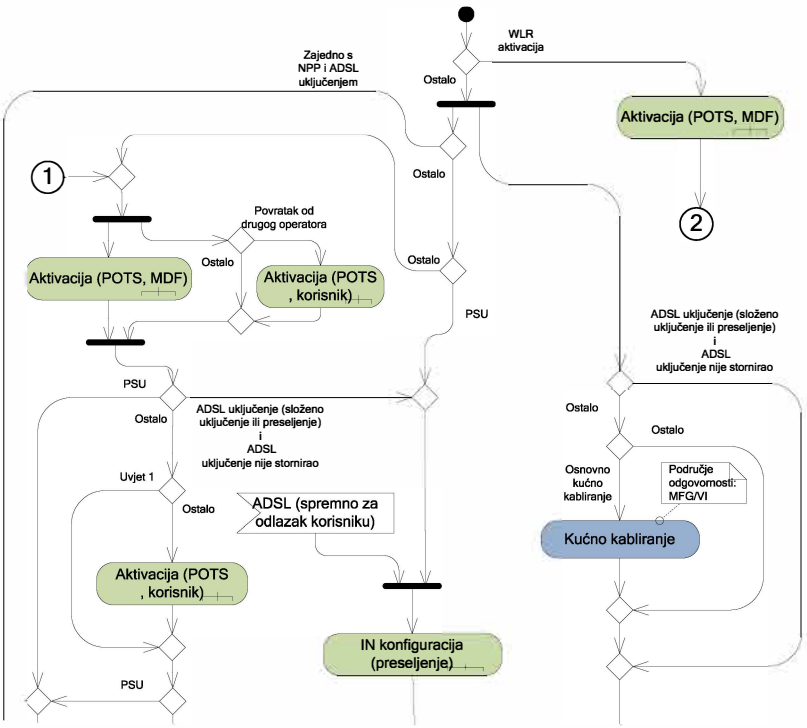
Dolazni podaci:
- Podaci za Halo (govornu) uslugu

Konfiguracija na centrali:
- Konfiguracija mrežnih kapaciteta (paralelne akcije):
  - automatska konfiguracija (uključenje) MSAN porta (new; UNDO)
  - automatska konfiguracija (uključenje) broja na IMS-u (new; UNDO)
- Odlazni pozivi su blokirani

Aktivacija na terenu:
- Realizacija na terenu (slijedne akcije):
  - ručni zadatak - POTS spajanje linije (MDF)
  - ručni zadatak - POTS - Spajanje linije na strani korisnika
- Prije zatvaranja zadatka potrebno je pokrenuti automatsku konfiguraciju (deblokiranje odlaznih poziva) broja na IMS-u

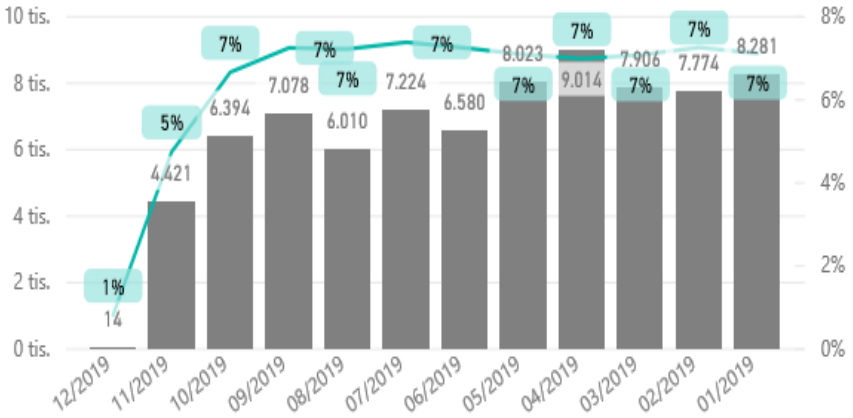
Odlazni podaci:
- PHONE_NUM (broj telefona)
- Due date (vrijeme instalacije)

Slika 1: Segment modela poslovnog procesa više razina apstraktnosti

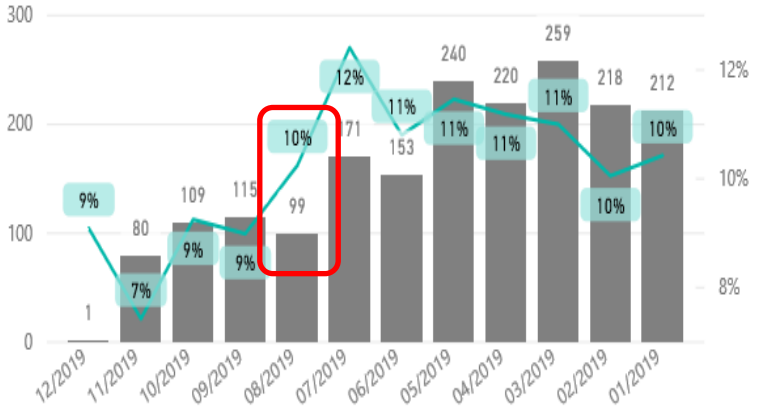


Uvjet 1	Povratak s VOIPa ili povratak od drugog operatora ili (APBX i nema vodeći broj i nema Main PBX asset)

Slika 2: Segment modela poslovnog procesa niže razine apstraktnosti (aktivacija na terenu - dio modula)



Slika 3: Praćenje određenog segmenta ponašanja postojećih poslovnih procesa



Slika 4: Primjer praćenja trenda u ponašanju određenog poslovnog procesa

pokrenutih od strane workflow sustava kojim se upravlja tijekom akcija u sklopu poslovnog procesa.

Osim navedenog primjera primjene analitičkog praćenje ponašanja poslovnih procesa, kroz analitičke alate mogu se pratiti i ostali segmenti ponašanja poslovnih procesa kao što je vrijeme odvijanja cijelog poslovnog procesa ili njegovih segmenata, uspješnost automatske konfiguracije mrežnih elemenata, ponašanje poslovnih procesa nakon implementacije promjena u sustavima za podršku operacijama i tako dalje.

IV. Zaključak

Važno je naglasiti da model poslovnog procesa ne može pojednostaviti složenost poslovnog procesa, ali ga može prikazati na odgovarajući način kako bi svim zainteresiranim stranama tijekom životnog tijeka procesa bilo što jasniji tijek i namjena poslovnog procesa. Također, model osigurava lakše promjene i unaprjeđenja postojećih procesa, kao i simulacije prije donošenja odluka o promjenama.

U odabranom modelu poslovnog procesa za dizajn i specifikaciju potrebnog sustava za podršku operacijama, poslovni proces se prikazuje kroz više aspekata i time se nastoji postići višestruka namjena modela. S jedne strane, modelom se opisuju osnovni procesni koraci kako bi svakome tko razmatra model bilo jasno što se s procesom želi postići. S druge strane, modelom se definiraju svi potrebni detalji za implementaciju programske podrške u sustavima za podršku operacijama kojom se upravlja odvijanjem poslovnih procesa kao i područja odgovornosti za pojedine procesne korake.

Opisani model poslovnih procesa koristi se i u svrhu analize postojećih poslovnih procesa gdje se, u kombinaciji sa specijaliziranim analitičkim alatima kao što je Power BI, želi postići snažan analitički mehanizam kojim se olakšava otkrivanje uskih grla u poslovnim procesima i definiranje koraka za potrebnu optimizaciju i poboljšanje poslovnih procesa. Definiranje podataka, prikaza i trendova unutar PBI-a temelji se na modelu procesa i karakterističnim pokazateljima vezano za tok procesa, faze i povezivanje podataka. Sve zajedno omogućava lako razumljive prikaze ponašanja poslovnog procesa i utjecaja određenih aktivnosti tijekom procesa na ponašanje usluge. Time se olakšava donošenje zaključaka i odluka što je potrebno mijenjati kako bi se omogućila bolja kvaliteta realizacije i promjene na uslugama, kao i što je potrebno promijeniti

unutar samog procesa za osiguranje efikasnije i brže isporuke usluga krajnjim korisnicima.

LITERATURA

[1] Markulin D., Kunštić M., "Modelling Business Processes in the Complex Telecommunication Environment", *ConTEL 2013, Zagreb, 2013.*

[2] Magner R., "Softversko inženjerstvo", Sveučilište u Zagrebu, Prirodoslovno Matematički Fakultet, skipta, 2005.

[3] Jović A., Horvat M., Ivošević D., Frid N., „Procesi programskog inženjerstva", Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, skripta, 2015.

[4] Markulin D., Kunštić M., "Result of using formal methods in modeling telecommunication business management", *Mipro 2012, HT d.d.*, 2012.

[5] Musa K., Markulin D., "Proposal for business process modelling and development in condition of complex telecommunication market", *Mipro 2012, HT d.d.*, 2012.

[6] Markulin D., Musa K., "Uloga kvalitetnog dizajna poslovnih procesa u složenom okruženju telekomunikacijskog tržišta", *Mipro* 2015, HT d.d., 2015.

[7] Markulin D., Musa K., Pisačić Lj., Mršić A., "Uloga standarda u postupku redizajna poslovnih procesa telekom operatora", Mipro 2017, HT d.d., 2017.

Komparativna analiza upravljanja obrtnim kapitalom u odjeljcima područja djelatnosti Informacije i komunikacije u Republici Hrvatskoj

Dina Korent*
* Fakultet organizacije i informatike/Katedra za gospodarstvo, Varaždin, Hrvatska
dkorent@foi.unizg.hr

***Sažetak* - Upravljanje zalihama, distribucija trgovačkih kredita kupcima i sposobnost poduzeća da prolongira plaćanje obveza prema dobavljačima, pa u skladu s time i potrebe za obrtnim kapitalom, razlikuju se među poduzećima iz različitih djelatnosti. Razlog tome su tradicija i priroda poslovanja koje nerijetko određuju potrebe za obrtnim kapitalom i ciklus konverzije novca u pojedinoj djelatnosti. Posljedično, može se provocirati hipoteza da ulaganja u obrtni kapital značajno ovise o djelatnosti. Nastavljajući se na prethodno, vjerojatnim se čini da će poduzeća iz različitih djelatnosti u nastojanju ostvarenja ciljeva slijediti i različite strategije upravljanja obrtnim kapitalom. Upravo iz tog razloga, svrha ovog rada jest analizirati, ispitati značajnost i utvrditi postojanje sličnosti odnosno razlika u upravljanju obrtnim kapitalom u odjeljcima područja djelatnosti Informacije i komunikacije u Republici Hrvatskoj. Upravljanje obrtnim kapitalom u navedenima analizira se u kontekstu pokazatelja istog. Kao izvor podataka za izračun pokazatelja koriste se baze podataka Financijske agencije za promatrano razdoblje od 2008.-2015. godine. Pored deskriptivne analize, komparativna analiza upravljanja obrtnim kapitalom uključuje i testiranje signifikantnosti razlika među pokazateljima upravljanja obrtnim kapitalom predmetnih odjeljaka djelatnosti. Rezultati analize ukazuju da su spomenute razlike signifikantne. Potonje se može objasniti inherentnim strukturno-poslovnim obilježjima odjeljaka područja djelatnosti Informacije i komunikacije.**

Ključne riječi - upravljanje obrtnim kapitalom, odjeljci područja djelatnosti Informacije i komunikacije, Republika Hrvatska, komparativna analiza

I. UVOD

Upravljanje obrtnim kapitalom je dio upravljanja financijama poduzeća koje je, u pravilu, fokusirano na kratkoročna ulaganja i kratkoročno financiranje poduzeća [1], [2]. Vodeći se distinkcijom koncepata obrtnog kapitala, upravljanje obrtnim kapitalom se može definirati na dva načina, kao upravljanje bruto obrtnim kapitalom te kao upravljanje neto obrtnim kapitalom. Upravljanje bruto obrtnim kapitalom predstavlja koncept upravljanja investicijama u ukupnu tekuću imovinu i financiranjem ukupne tekuće imovine. Međutim, premda se u literaturi najčešće eksplicitno ne definira o kojem je konceptu upravljanja riječ, pod terminom upravljanja obrtnim kapitalom većinom se misli na kratkoročno upravljanje financijama poduzeća, odnosno upravljanje neto obrtnim kapitalom [3]. Predmetno obuhvaća upravljanje tekućom imovinom, odnosno planiranje i kontrolu iznosa i sastava tekuće imovine, te upravljanje tekućim obvezama poduzeća, odnosno korištenje kratkoročnim ili tekućim obvezama za financiranje investicija u tekuću imovinu [1], [3]–[12]. Nastavi li se sužavanje koncepta obrtnog kapitala može se problematizirati upravljanje neto operativnim obrtnim kapitalom kao i upravljanje neto operativnim obrtnim kapitalom u užem smislu. Potonje se može definirati kao integrativno upravljanje zalihama, potraživanjima od kupaca i obvezama prema dobavljačima, odnosno kao upravljanje ulaganjima poduzeća u neto operativni obrtni kapital u užem smislu [13].

Neovisno o konceptu od kojeg se polazi u definiranju upravljanja obrtnim kapitalom, navedeno bi trebalo težiti postizanju uravnoteženih razina komponenata obrtnog kapitala [8], odnosno optimalne veličine, strukture i cirkulacije obrtnog kapitala koje najviše pridonose ostvarenju fundamentalnog cilja poslovanja poduzeća, dakle, maksimalizaciji njegove sadašnje vrijednosti [5], [9]. Uzme li se pritom u obzir konfliktnost maksimalizacije profitabilnosti i minimalizacije rizika nelikvidnosti impliciranih navedenim ciljem [9], [14]–[17], cilj upravljanja obrtnim kapitalom relaksira se i svodi na uspostavu i održavanje optimalne likvidnosti poslovanja pojmljeno kroz prizmu najpovoljnijeg kompromisa između profitabilnosti i rizika, a u kontekstu utjecaja na stvaranje vrijednosti poduzeća [13].

U nastojanju postizanja navedenog cilja, poduzeća na temelju razmatranja međuovisnosti rizika i profitabilnosti te sukladno preferencijama prema riziku definiraju i provode strategiju upravljanja obrtnim kapitalom. Pritom je moguće razlikovati agresivnu, konzervativnu te umjerenu strategiju upravljanja obrtnim kapitalom [4], [7], [18]–[20]. Agresivna strategija implicira relativno mala ulaganja u tekuću imovinu i relativno visoke iznose kratkoročnog financiranja [7]. Oprečno, konzervativna strategija podrazumijeva relativno velika ulaganja u tekuću imovinu i relativno male iznose kratkoročnog financiranja, te rezultira visokim neto obrtnim kapitalom i tekućim omjerom [1], [7]. Umjerena strategija nalazi se između agresivne i konzervativne strategije [18], te uključuje umjerena ulaganja u tekuću imovinu i umjeren iznos kratkoročnog duga [7].

Nastavljajući se na prethodno, premda je cilj upravljanja obrtnim kapitalom istovjetan za sva poduzeća neovisno o djelatnosti, veličini i drugim karakteristikama, načini njegova ostvarenja nerijetko se razlikuju u zavisnosti o navedenima. S aspekta djelatnosti, tome je tako jer razlike u prirodi poslovanja i poslovnim običajima

među različitim djelatnostima, u pravilu, determiniraju i razlike u upravljanju obrtnim kapitalom u istima [6]. Naime, s jedne strane, poduzeća iz nekih djelatnosti sklonija su minimalizaciji zaliha i/ili potraživanja od kupaca, dok, s druge strane, poduzeća iz drugih djelatnosti tendiraju maksimalizirati obveze prema dobavljačima [21]. Primjera radi, Dinku [22] konstatira da su, u cilju povećanja novčanih tokova, poduzeća u manje kompetitivnim djelatnostima sklonija smanjiti obujam danih trgovačkih kredita te dane vezivanja potraživanja od kupaca, dok su poduzeća u djelatnostima s velikom bazom dobavljača sirovina i materijala usmjerena povećanju broja dana vezivanja obveza prema dobavljačima.

II. PREGLED LITERATURE I HIPOTEZE

Uvažavajući teorijsku pretpostavku da se upravljanje obrtnim kapitalom, odnosno upravljanje zalihama, distribucija trgovačkih kredita kupcima i sposobnost poduzeća da prolongira plaćanje obveza prema dobavljačima, pa u skladu s time i potrebe za obrtnim kapitalom, razlikuju među poduzećima iz različitih djelatnosti, istraživanja područja upravljanja obrtnim kapitalom usmjerena su ispitivanju značajnosti navedenih razlika. Literatura područja obuhvaća istraživanja koja uključuju analizu upravljanja obrtnim kapitalom u pojedinim djelatnostima [21], [23]–[28], istraživanja odrednica upravljanja obrtnim kapitalom koja kao jednu od odrednica uključuju utjecaje djelatnosti, u pravilu, kao indikator varijable za djelatnosti [29]–[36], a u slučaju [31] i u pogledu poduzorkovanja po djelatnostima (proizvodnja, usluge), potom istraživanja koja kao jednu od nezavisnih varijabli u modelima uvrštavaju industrijski prilagođenu zavisnu varijablu upravljanja obrtnim kapitalom [37], te kombinacije različitih tipova istraživanja [13], [16], [38]–[52].

U fokusu razmatranja spomenutih istraživanja su različite djelatnosti. Analizom sastava uzoraka istih dijagnosticirano je da su najprisutnije proizvodne djelatnosti, odnosno prerađivačka industrija, zatim maloprodaja i veleprodaja, poljoprivreda, šumarstvo i ribarstvo, rudarstvo i vađenje, te građevinarstvo. Nerijetko su iz uzoraka istraživanja isključena poduzeća iz financijske djelatnosti i djelatnosti osiguranja [6], [10], [23], [47], [49], [51], [53]–[55]. Potonje je argumentirano specifičnom prirodom poslovanja, drugačijim računovodstvenim zahtjevima te razlikama u definiranju koncepta obrtnog kapitala. Pored navedenih djelatnosti, u nekim su istraživanjima, iz uzoraka izostavljena i ostala uslužna poduzeća, odnosno sva uslužna poduzeća [5], [9], [24], [42], [55]–[57].

Rezultati većine istraživanja [16], [21], [23]–[32], [37]–[53], [58], [59], provedenih u razvijenim, ali i zemljama u razvoju, otkrivaju da poduzeća iz različitih djelatnosti imaju različite potrebe za obrtnim kapitalom te da se zbog inherentnih karakteristika djelatnosti upravljanje obrtnim kapitalom razlikuju među istima. Potonje sugerira da je djelatnost signifikantna odrednica upravljanja obrtnim kapitalom [13].

Kongruentno potonjem, Mongrut et al. [37], za sve države u uzorku, uočavaju signifikantno pozitivan utjecaj prosječnog industrijskog ciklusa konverzije novca na ciklus konverzije novca poduzeća, obrazlažući navedeno nastojanjem poduzeća da svoje upravljanje obrtnim kapitalom usklade s uobičajenim praksama u djelatnosti. Nastavno, osim na postojanje razlika u relativnoj razini agresivnosti ili konzervativnosti strategija ulaganja i financiranja obrtnog kapitala između različitih djelatnosti, Afza i Nazir [23] te Weinraub i Visscher [25] u svojim istraživanjima detektiraju i da su te razlike veće i značajnije za strategije ulaganja u obrtni kapital no za strategije financiranja istih. Pored navedenog, identificirano je i to da relacija strategija upravljanja obrtnim kapitalom različitih djelatnosti perzistira kroz vrijeme [16], [21], [23], [25]. Konačno, Aktas i Petmezas [51] ističu kako u tridesetogodišnjem razdoblju obuhvaćenom njihovim istraživanjem dolazi do smanjenja heterogenosti u domeni upravljanja obrtnim kapitalom u većini promatranih djelatnosti [13].

Usprkos teorijskim indikacijama o utjecajima djelatnosti na upravljanje obrtnim kapitalom i zamjetnog broja emiprijskih istraživanja koja govore u prilog navedenoj tezi, određena istraživanja, premda rijetka, ne pronalaze empirijsku potvrdu egzistencije signifikantnog utjecaja djelatnosti na upravljanje obrtnim kapitalom [33]–[35].

Iako pronalaze robusnost upravljanja obrtnim kapitalom u ovisnosti o efektima djelatnosti, istraživanja koja su proveli Hill et al. [42] te Naser et al. [41] svojim analizama upravljanja obrtnim kapitalom poduzeća po djelatnostima nedvojbeno sugeriraju njihovu egzistenciju.

Sukladno teorijskim pretpostavkama i brojčano znatno zastupljenijim empirijskim istraživanjima koja problematiziraju i ispituju utjecaj djelatnosti na upravljanje obrtnim kapitalom poduzeća, provocira se hipoteza da upravljanje obrtnim kapitalom značajno ovisi o djelatnosti, odnosno, specifično, da postoje signifikantne razlike u upravljanju obrtnim kapitalom među odjeljcima područja djelatnosti Informacije i komunikacije u Republici Hrvatskoj.

III. METODOLOŠKI OKVIR ISTRAŽIVANJA

A. Uzorak

Uzorak istraživanja čine 7.718 poduzeće-godina opažanja, odnosno podaci o poduzećima koja su registrirana u Republici Hrvatskoj i najveći dio prihoda ostvaruju u području djelatnosti Informacije i komunikacije (J prema Nacionalnoj klasifikaciji djelatnosti (NKD) 2007), odnosno potonje je njihova pretežita djelatnost, u razdoblju od 2008.-2015. godine, a zavedena su u bazi podataka Financijske agencije (FINA-e). Navedeno je područje djelatnosti odabrano prije svega zbog njegove važnosti u gospodasrtvu Republike Hrvatske. Naime, analiza financijskih rezultata poduzeća registriranih u Republici Hrvatskoj, bez banaka i osiguravajućih društava, po djelatnostima pokazuje da je područje djelatnosti Informacije i komunikacije u promatranom razdoblju, izuzev 2013. i 2014. godine, jedno od tri najznačajnije djelatnosti u gospodarstvu Republike Hrvatske [60].

Iz početnog uzorka od 7.718 poduzeća i 36.929 poduzeće-godina opažanja, korištenjem programskog

paketa Stata 14, izolirana su poduzeće-godina opažanja bez zaposlenih (na bazi sati rada), bez pozitivnih vrijednosti, odnosno s nepozitivnim vrijednostima prihoda od prodaje, poslovnih rashoda, aktive ili kapitala i rezervi. Nakon navedenog koraka u uzorku je ostalo 3.388 poduzeća i 16.918 poduzeće-godina opažanja. Prema podacima Financijske agencije za 2008. godinu, struktura poduzeća odnosno poduzeće-godina opažanja u uzorku prema odjeljcima djelatnosti je kako slijedi: J.58 Izdavačke djelatnosti: 568 poduzeća i 3.114 poduzeće-godina opažanja, J.59 Proizvodnja filmova, videofilmova i televizijskog programa, djelatnosti snimanja zvučnih zapisa i izdavanja glazbenih zapisa: 287 poduzeća i 1.316 poduzeće-godina opažanja, J.60 Emitiranje programa: 159 poduzeća i 1.064 poduzeće-godina opažanja, J.61 Telekomunikacije: 216 poduzeća i 949 poduzeće-godina opažanja, J.62 Računalno programiranje, savjetovanje i djelatnosti povezane s njima: 1.923 poduzeća i 9.152 poduzeće-godina opažanja, J.63 Informacijske uslužne djelatnosti: 298 poduzeća i 1.323 poduzeće-godina opažanja, a prema veličini poduzeća sljedeća: 3.350 malih poduzeća (16.554 poduzeće-godina opažanja), 53 srednje velika poduzeća (239 poduzeće-godina opažanja) te 23 velika poduzeća (125 poduzeće-godina opažanja).

B. Podaci i varijable

Kako bi se testirala istraživačka hipoteza, u radu su prije svega na temelju navedenih podataka, primjenom programskog paketa MS Excel, izračunati pokazatelji varijable upravljanje obrtnim kapitalom. Korišteni su pokazatelji spomenute varijable korišteni u drugim sličnim istraživanjima, a njihov je odabir determiniran teorijskom utemeljenošću, zastupljenošću i signifikantnošću istih u prethodnim istraživanjima te dostupnošću podataka potrebnih za njihov izračun. Podaci o pretežitom odjeljku djelatnosti poduzeća tretirani su i korišteni kao indikator varijable. Navedeni su pokazatelji definirani za svako poduzeće-godinu. Konačno, kao strukturno-poslovni pokazatelji izračunati su Herfindahl-Hirschmanov indeks (HHI) i koeficijent radne intenzivnosti. Nazivi i oznake te načini mjerenja svih korištenih (pokazatelja) varijabli dani su u Tablici I.

C. Metode

Pored deskriptivne statistike, usmjerene analizi obilježja upravljanja obrtnim kapitalom poduzeća iz uzorka, kako bi se testirala istraživačka hipoteza, u radu su korišteni Kruskal-Wallisov test jednakosti populacija i Dunnov post hoc test.

Primjenom Kruskal-Wallisova testa jednakosti populacija i Dunnova post hoc testa se, uslijed neudovoljavanja pretpostavkama za primjenu odgovarajućih parametarskih testova, testira postojanje signifikantnih razlika između srednjih vrijednosti, konkretno medijana, pokazatelja upravljanja obrtnim kapitalom za promatrane odjeljke područja djelatnosti Informacije i komunikacije u Republici Hrvatskoj. Navedeno omogućava ispitivanje utjecaja djelatnosti na upravljanje obrtnim kapitalom poduzeća.

TABLICA I. NAZIVI, OZNAKE I NAČINI MJERENJA POKAZATELJA

Naziv i oznaka pokazatelja	Način mjerenja pokazatelja
Omjer tekuće i ukupne imovine (TIUI)	tekuća imovina/ukupna imovina
Omjer zaliha i ukupne imovine (ZUI)	zalihe/ukupna imovina
Omjer potraživanja od kupaca i ukupne imovine (PKUI)	potraživanja od kupaca/ukupna imovina
Omjer tekućih obveza i ukupne imovine (TOUI)	tekuće obveze/ukupna imovina
Omjer obveza prema dobavljačima i ukupne imovine (ODUI)	obveze prema dobavljačima/ukupna imovina
Omjer neto operativnog obrtnog kapitala u užem smislu i ukupne imovine (NOOKUI)	neto operativni obrtni kapital u užem smislu/ukupna imovina
Dani vezivanja zaliha (DVZpr)	(prosječne zalihe/poslovni rashodi)*365
Dani vezivanja zaliha (DVZpp)	(prosječne zalihe/prihodi od prodaje)*365
Dani vezivanja potraživanja od kupaca (DVK)	(prosječna potraživanja od kupaca/prihodi od prodaje)*365
Dani vezivanja obveza prema dobavljačima (DVDpr)	(prosječne obveze prema dobavljačima/poslovni rashodi) * 365
Dani vezivanja obveza prema dobavljačima (DVDpp)	(prosječne obveze prema dobavljačima/poslovni rashodi) * 365
Ciklus konverzije novca (CKN)	DVZpr + DVK - DVDpr
Neto trgovinski ciklus (NTC)	DVZpp + DVK - DVDpp
Djelatnost	Indikator varijabla za odjeljke područja djelatnosti J Informacije i komunikacije (šest modaliteta tj. šest odjeljaka): J.58, J.59, J.60, J.61, J.62 i J.63.
Herfindahl-Hirschman indeks (HHI) odjeljka	Suma kvadrata tržišnih udjela (ukupni prihodi poduzeća/ukupni prihodi odjeljka) svih poduzeća u podjedinom odjeljku.
Koeficijent radne intenzivnosti odjeljka	((broj zaposlenih u odjeljku)/broj zaposlenih gospodarstva)/)/dugotrajna imovina odjeljka/dugotrajna imovina gospodarstva))

IV. EMPIRIJSKI REZULTATI

A. Deskriptivna statistika

Rezultati deskriptivne statistike pokazatelja upravljanja obrtnim kapitalom za sva poduzeća iz uzorka, tj. za sva poduzeća područja djelatnosti Informacije i komunikacije obuhvaćena istim odnosno za pojedine odjeljke spomenutog područja djelatnosti [1] generalno demonstriraju postojanje zamjetnih raspona varijacije, standardnih devijacija te koeficijenata varijacije promatranih pokazatelja. Potonje ukazuje na velike varijacije između vrijednosti te na njihovu veliku disperziju oko odnosnih prosječnih vrijednosti. S obzirom na to da koeficijenti varijacije svih pokazatelja prelaze graničnih 30, a nerijetko i 100 i više %, spomenuti, kao pokazatelji reprezentativnosti prosjeka, sugeriraju manju ili veću nereprezentativnost istih. Posljedično, prosjek, odnosno aritmetička sredina nije relevantna mjera vrijednosti predmetnih pokazatatelja za odnosna poduzeća, te se kao prikladnija mjera deskriptivne statistike u analizi i donošenju zaključaka koristi medijan.

Medijalne vrijednosti omjera tekuće i ukupne imovine (0,84), omjera potraživanja od kupaca i ukupne imovine

[1] Rezultati deskriptivne statistike dostupni su na zahtjev.

(0,25), omjera tekućih obveza i ukupne imovine (0,46) te omjera neto operativnog obrtnog kapitala u užem smislu i ukupne imovine (0,18) za sva poduzeća iz uzorka, no isto tako i za pojedine poduzorke poduzeća, upućuju na važnost upravljanja obrtnim kapitalom u promatranim poduzećima. Promatrano po odjeljcima djelatnosti, medijalna vrijednost omjera tekuće i ukupne imovine pritom je iznad one za cijeli uzorak za odjeljke J.58, J.62 i J.63 odnosno ispod one za cijeli uzorak za odjeljke J.59, J.60 i J.61. Medijalna vrijednost omjera potraživanja od kupaca i ukupne imovine u odjeljcima J.60 i J.62 veća je od medijana istog za cijeli uzorak, dok je u preostalim odjeljcima od navedenog manja. Nadalje, medijani omjera tekućih obveza i ukupne imovine promatranih odjeljaka, izuzev u odjeljcima J.62 i J.63, premašuju medijan spomenutog pokazatelja za cijeli uzorak poduzeća. Konačno, medijalne vrijednosti omjera neto operativnog obrtnog kapitala u užem smislu i ukupne imovine, osim za odjeljke J.58 i J.62, ne prelaze vrijednost medijana potonjeg za uzorak svih poduzeća.

Oprečno elaboriranim pokazateljima upravljanja obrtnim kapitalom koji signaliziraju na njegovu važnost, medijalne vrijednosti omjera zaliha i ukupne imovine (0,00) te omjera obveza prema dobavljačima i ukupne imovine (0,09) čitavog uzorka, kao i poduzoraka poduzeća po odjeljcima područja djelatnosti, ukazuju na suprotno. U tom kontekstu, izuzev u odjeljcima J.58 i J.61, i medijalne vrijednosti omjera zaliha i ukupne imovine u preostalim odjeljcima iznose 0. Ovako niske medijalne vrijednosti omjera zaliha i ukupne imovine u području djelatnosti Informacije i komunikacije jasne su uzme li se u obzir neopipljivost kao inherentna karakteristika njihovih proizvoda, odnosno usluga [13]. Pretežiti dio promatranih poduzeća ne proizvodi opipljiva dobra te shodno tome, u pravilu, ne nabavlja zalihe sirovina i materijala za angažiranje u proizvodnji, njihovu transformaciju u gotove proizvode te naposljetku prodaju tih proizvoda na tržištu [13]. Štoviše, mnoga poduzeća u području djelatnosti Informacije i komunikacije uopće nemaju zalihe ili navedene čine samo mali dio njihove ukupne imovine. Uz zalihe, kao što je spomenuto, i obveze prema dobavljačima inferioriziraju značaj upravljanja obrtnim kapitalom u području djelatnosti Informacije i komunikacije. S aspekta odjeljaka, medijani omjera obveza prema dobavljačima i ukupne imovine odjeljaka, u odnosu na sva poduzeća područja, ukazuju na inferirorniju važnost u odjeljcima J.62 i J.63 te na superiorniju važnost u svim preostalim odjeljcima.

Uzimajući u obzir važnost upravljanja obrtnim kapitalom, ciklus konverzije novca i neto trgovinski ciklus, kao i dani vezivanja istih pobliže otkrivaju specifičnosti upravljanja obrtnim kapitalom u odjeljcima područja djelatnosti Informacije i komunikacije. Na razini cijelog uzorka, medijani ciklusa konverzije novca i neto trgovinskog ciklusa iznose 173,93 odnosno 172,72 dana. Shodno tome, polovica vrijednosti ciklusa konverzije novca te neto trgovinskog ciklusa poduzeća iz uzorka jednaka je ili manja, dok je druga polovica istih jednaka ili veća od 173,93 odnosno 172,72 dana. Medijalni neto vremenski interval između novčanih izdataka za kupovine proizvodnih resursa i novčanih primitaka od prodaje proizvoda poduzeća u odjeljcima J.58, J.59 i J.60 dulji je, a u odjeljcima J.61, J.62 i J.63 kraći od onog čitavog uzorka. Relativno dugački medijalni ciklusi rezultat su dugih ciklusa konverzije zaliha (DVZpr: 187,20; DVZpp: 179,10) i vremena naplate potraživanja od kupaca (DVK: 186,84), koji nisu kompenzirani, premda dugačkim, vremenima plaćanja obveza prema dobavljačima (DVDpr: 201,51; DVDpp: 190,85). Nastavno, analizom je identificirano da su navedeni općenito kraći u odjeljcima J.61, J.62 i J.63 odnosno dulji u odjeljcima J.58, J.59 i J.60.

Posljednje implicira da je relativan odnos ulaganja u zalihe te potraživanja od kupaca i financiranja obveza prema dobavljačima u odjeljcima J.61, J.62 i J.63 na strani posljednjeg, a u odjeljcima J.58, J.59 i J.60 na strani prvotnih. Predmetno ukazuje na postojanje sličnosti odnosno razlika u upravljanju obrtnim kapitalom među pojedinim odjeljcima promatranog područja djelatnosti.

B. Rezultati testiranja razlika u upravljanju obrtnim kapitalom

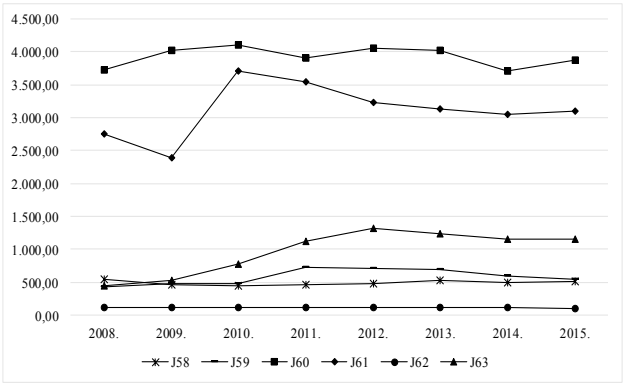
Daljnja analiza (Kruskal-Wallisov test jednakosti populacija i Dunnov post hoc test)[2] otkriva postojanje statistički signifikantnih razlika u ciklusu konverzije novca te neto trgovinskom ciklusu između svih odjeljaka područja djelatnosti Informacije i komunikacije, izuzev između parova odjeljaka J.61, J.62 i J.63, naročito u kontekstu neto trgovinskog ciklusa. Na uobičajenim razinama signifikantnosti, statistički signifikantna razlika u ciklusu konverzije novca nije detektirana samo između odjeljaka J.62 i J.63.

U svrhu uvida u izvore razlika u trajanju ciklusa konverzije novca i neto trgovinskog ciklusa među poduzećima iz različitih odjeljaka, provedeno je i testiranje razlika u upravljanju komponentama ciklusa konverzije novca odnosno neto trgovinskog ciklusa. Vezano za upravljanje zalihama, rezultati pokazuju da je navedeno signifikantno različito, izuzev između odjeljaka J.59 i J.62, J.60 i J.61, J.60 i J.63 te J.61 i J.63 u kontekstu ciklusa konverzije odnosno između J.58 i J.60 te J.61 i J.62 u kontekstu neto trgovinskog ciklusa. Upravljanje potraživanjima od kupaca signifikantno je različito među svim odjeljcima osim između odjeljaka J.62 i J.63. Istovjetan je rezultat uočen i za upravljanje obvezama prema dobavljačima promatrano u odnosu na prihode od prodaje, dok su u relaciji s poslovnim rashodima razlike nešto uže i ne egzistiraju među odjeljcima J.58 i J.60, J.61 i J.62 te J.61 i J.63.

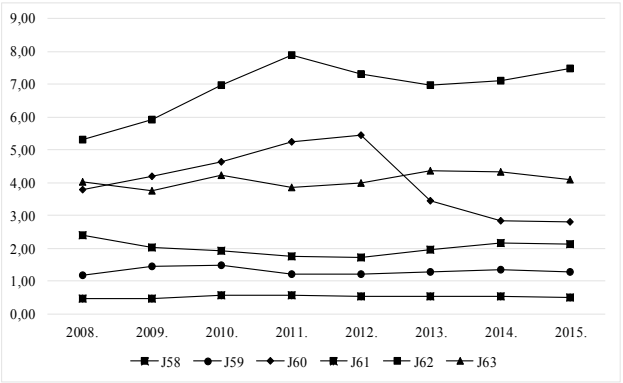
Na temelju izloženih rezultata može se, uz određene iznimke, napose za odjeljke J.61, J.62 i J.63, a sporadično po komponentama upravljanja obrtnim kapitalom i za neke druge odjeljke, prihvatiti hipoteza da postoje signifikantne razlike u upravljanju obrtnim kapitalom među odjeljcima područja djelatnosti Informacije i komunikacije u Republici Hrvatskoj. Te su razlike šire i signifikantnije za upravljanje potraživanjima od kupaca i obvezama prema dobavljačima, no za upravljanje zalihama, značaj kojeg je u domeni upravljanja obrtnim kapitalom analiziranih poduzeća najinferiorniji. Dodatno, neovisno o širini i signifikatnosti spomenutih razlika, integrativno i parcijalno upravljanje obrtnim kapitalom u

[2] Rezultati analize dostupni su na zahtjev.

odjeljcima J.61, J.62 i J.63 signifikantno je agresivnije nego u preostalim odjeljcima područja djelatnosti Informacije i komunikacije. Dobiveni rezultati, kao i nalaz testiranja hipoteze mogu se obrazložiti fundamentalnim razlikama u proizvodima i tržištima, odnosno strukturno-poslovnim obilježjima promatranih odjeljaka, naročito koncentracijom i radnom intenzivnošću istih. Naime, izuzev za odjeljak 62 u kontekstu HHI-a odnosno za odjeljak 61 u kontekstu radne intenzivnosti, vrijedi da generalno odjeljci s većim HHI-om, konzistentno s [22], i većom radnom intenzivnošću imaju kraće cikluse upravljanja obrtnim kapitalom. Kretanja vrijednosti HHI-a i koeficijenta radne intenzivnosti po odjeljcima područja djelatnosti Informacije i komunikacije kroz godine ilustrirana su grafičkim prikazima na Slici 1. odnosno Slici 2.



Slika 1. HHI po odjeljcima područja djelatnosti Informacije i komunikacije



Slika 2. Radna intenzivnost po odjeljcima područja djelatnosti Informacije i komunikacije

V. ZAKLJUČAK

Upravljanje obrtnim kapitalom predstavlja kratkoročno upravljanje financijama koje uključuje upravljanje tekućom imovinom i tekućim obvezama poduzeća. Cilj upravljanja obrtnim kapitalom je maksimalizirati sadašnju vrijednost poduzeća. Navedeni cilj se zbog razlika u prirodi poslovanja i poslovnim običajima među djelatnostima, u istima nerijetko postiže odnosno nastoji postići implementacijom različitih strategija upravljanja obrtnim kapitalom.

U skladu s teorijskim presumpcijama i dominantnim emiprijskim istraživanjima, u ovom je radu provocirana hipoteza da upravljanje obrtnim kapitalom značajno ovisi o djelatnosti, odnosno, specifično, da postoje signifikantne razlike u upravljanju obrtnim kapitalom među odjeljcima područja djelatnosti Informacije i komunikacije u Republici Hrvatskoj. Rezultati istraživanja i nalazi testiranja istraživačke hipoteze omogućuju njeno djelomično prihvaćanje, odnosno prihvaćanje uz određene iznimke. Integrativno, ali i parcijalno upravljanje obrtnim kapitalom, izuzev za odjeljke J.61, J.62 i J.63, a sporadično po komponentama upravljanja obrtnim kapitalom i za neke druge odjeljke, signifikantno se razlikuje. Navedeno se može argumentirati inherentnim strukturno-poslovnim obilježjima odjeljaka područja djelatnosti Informacije i komunikacije.

LITERATURA

[1] S. A. Ross, R. W. Westerfield, and B. D. Jordan, *Essentials of corporate finance*, 6th ed. New York: McGraw-Hill/Irwin, 2008.

[2] A. K. Sharma and S. Kumar, "Effect of Working Capital Management on Firm Profitability: Empirical Evidence from India," *Glob. Bus. Rev.*, vol. 12, no. 1, pp. 159–173, 2011.

[3] S. Orsag, *Poslovne financije*. Zagreb: Avantis: Hrvatska udruga financijskih analitičara, 2015.

[4] J. C. Van Horne and J. M. Wachowicz, *Fundamentals of Financial Management*, 13th ed. Harlow: Prentice Hall Inc, 2008.

[5] M. Deloof, "Does working capital management affect profitability of Belgian firms?," *J. Bus. Financ. Account.*, vol. 30, no. 3–4, pp. 573–587, 2003.

[6] A. M. A. Eljelly, "Liquidity - profitability tradeoff: An empirical investigation in an emerging market," *Int. J. Commer. Manag.*, vol. 14, no. 2, pp. 48–61, 2004.

[7] R. C. Moyer, J. R. McGuigan, and W. J. Kretlow, *Contemporary Financial Management*, 10th ed. Mason, OH: South-Western, Thomson Corporation, 2006.

[8] V. Ganesan, "An analysis of working capital management efficiency in telecommunication equipment industry," *Rivier Acad. J.*, vol. 3, no. 2, pp. 1–10, 2007.

[9] A. Raheman and M. Nasr, "Working Capital Management And Profitability – Case Of Pakistani Firms," *Int. Rev. Bus. Res. Pap.*, vol. 3, no. 1, pp. 279–300, 2007.

[10] J. Enqvist, M. Graham, and J. Nikkinen, "The impact of working capital management on firm profitability in different business cycles: Evidence from Finland," *Res. Int. Bus. Financ.*, vol. 32, no. August, pp. 36–49, 2012.

[11] R. A. Brealey, S. C. Myers, and F. Allen, *Principles of Corporate Finance*, 10th ed. New York: McGraw-Hill/Irwin, 2011.

[12] S. Besley and E. F. Brigham, *Essentials of Managerial Finance*, 14th ed. Mason, OH: Thomson South-Western, 2007.

[13] D. Korent, "Identifikacija odrednica upravljanja obrtnim kapitalom u funkciji povećanja profitabilnosti poduzeća u odabranim djelatnostima u Republici Hrvatskoj," Doktorski rad, Zagreb: Ekonomski fakultet Zagreb, 2018.

[14] M. Singhania, N. Sharma, and J. Y. Rohit, "Working capital management and profitability: evidence from Indian manufacturing companies," *Decision*, vol. 41, no. 3, pp. 313–326, 2014.

[15] L. J. Gitman and C. J. Zutter, *Principles of Managerial Finance*, 13th ed. New Jersey: Prentice Hall Inc, 2011.

[16] S. Baños-Caballero, P. J. García-Teruel, and P. Martínez-Solano, "How does working capital management affect the profitability of Spanish SMEs?," *Small Bus. Econ.*, vol. 39, no. 2, pp. 517–529, 2012.

[17] A. Bhunia and A. Das, "Affiliation between Working Capital Management and Profitability," *Interdiscip. J. Contemp. Res. Bus.*, vol. 3, no. 9, pp. 957–968, 2012.

[18] E. F. Brigham and P. R. Daves, *Intermediate financial management*, 9th ed., vol. 21, no. 3. Mason, OH: Thomson South-Western, 2007.

[19] O. Nwankwo and G. S. Osho, "An empirical analysis of corporate survival and growth: Evidence from efficient

working capital management," *Int. J. Sch. Acad. Intellect. Divers.*, vol. 12, no. 1, pp. 1–13, 2010.
[20] E. F. Brigham and M. C. Ehrhardt, *Financial Management: Theory and Practice*, 13th ed. Mason, OH: South-Western Cengage Learning, 2011.
[21] G. Filbeck and T. M. Krueger, "An Analysis of Working Capital Management Results Across Industries," *Am. J. Bus.*, vol. 20, no. 2, pp. 11–20, 2005.
[22] T. Dinku, "Impact of Working Capital Management on Profitability of Micro and Small Enterprises in Ethiopia: The Case of Bahir Dar City Administration," *Int. J. Account. Tax.*, vol. 1, no. 1, pp. 15–24, 2013.
[23] T. Afza and M. S. Nazir, "Working Capital Approaches and Firm's Returns in Pakistan," *Soc. Sci.*, vol. 1, pp. 25–36, 2008.
[24] A. Uyar, "The Relationship of Cash Conversion Cycle with Firm Size and Profitability: An Empirical Investigation in Turkey," *Int. Res. J. Financ. Econ.*, vol. 24, no. 24, pp. 187–193, 2009.
[25] J. H. Weinraub and S. Visscher, "Industry practice relating to aggressive conservative working capital policies," *J. Financ. Strateg. Decis.*, vol. 11, no. 2, pp. 11–18, 1998.
[26] G. Hawawini, C. Viallet, and A. Vora, "Industry influence on corporate working capital decisions," *Sloan Manage. Rev.*, vol. 27, no. 4, pp. 15–24, 1986.
[27] G. Filbeck, T. Krueger, and D. Preece, "CFO Magazine's 'Working Capital Survey': Do Selected Firms Work for Shareholders?," *Q. J. Bus. Econ.*, vol. 46, no. 2, pp. 3–22, 2007.
[28] Ž. Barać Aljinović, T. Vuko, and T. Vučak, "Effects of working capital management on profitability: evidence from Croatia," in *9th International conference Economic integrations, competition and cooperation: accession of the Western Balkan countries to the European Union*, 2013, pp. 578–589.
[29] R. Kieschnick, M. LaPlante, and R. Moussawi, "Corporate working capital management: determinants and consequences," *Int. J. Manag. Financ.*, vol. 3, no. 2, pp. 164-177., 2006.
[30] M. S. Nazir and T. Afza, "Working Capital Requirements and the Determining Factors in Pakistan," *IUP J. Appl. Financ.*, vol. 15, no. 4, pp. 28–38, 2009.
[31] A. Gill, "Factors that influence working capital requirements in Canada," *Econ. Financ. Rev.*, vol. 1, no. 3, pp. 30–40, 2011.
[32] A. A. Onaolapo and S. O. Kayjola, "What are the Determinants of Working Capital Requirements of Nigerian Firms?," *Res. J. Financ. Account.*, vol. 6, no. 6, pp. 118–128, 2015.
[33] J.-R. Chiou, L. Cheng, and H.-W. Wu, "The Determinants of Working Capital Management," *J. Am. Acad. Bus.*, vol. 10, no. 1, pp. 149–156, 2006.
[34] E. Manoori and D. Muhammad, "Determinants of working capital management: Case of Singapore firms," *Res. J. Financ. Account.*, vol. 3, no. 11, pp. 15–24, 2012.
[35] A. N. Saarani and F. Shahadan, "The Determinant Factors of Working Capital Requirements for Enterprise 50 (E50) Firms in Malaysia: Analysis Using Structural Equation Modelling," *Soc. Sci. Sci. Stud.*, vol. 5, no. 2, pp. 52–66, 2012.
[36] G. A. Afrifa and K. Padachi, "Working capital level influence on SME profitability," *J. Small Bus. Enterp. Dev.*, vol. 23, no. 1, pp. 44–63, 2016.
[37] S. Mongrut, D. F. O´Shee, C. C. Zavaleta, and J. C. Zavaleta, "Determinants of Working Capital Management in Latin American Companies," *Innovar*, vol. 24, no. 51, pp. 5–18, 2014.
[38] M. L. Jose, C. Lancaster, and J. L. Stevens, "Corporate Return and Cash Conversion Cycle," *J. Econ. Financ.*, vol. 20, no. 1, pp. 33–46, 1996.
[39] S. Baños-Caballero, P. J. García-Teruel, and P. Martínez-Solano, "How do market imperfections affect working capital management?," *Work. Pap. Doc. Trab. Ser. EC (Instituto Valencia. Investig. Económicas)*, vol. 14, no. 1, pp. 3–25, 2009.
[40] A. Rimo and P. Panbunyuen, "The effect of company characteristics on working capital management A quantitative study of Swedish listed companies," Master thesis, Umeå University, Umeå School of Business, Economics and Statistics, 2010.
[41] K. Naser, R. Nuseibeh, and A. Al-Hadeya, "Factors Influencing Corporate Working Capital Management: Evidence from an Emerging Economy," *J. Contemp. Issues Bus. Res.*, vol. 2, no. 1, pp. 11–30, 2013.
[42] M. D. Hill, G. W. Kelly, and M. J. Highfield, "Net Operating Working Capital Behavior: A First Look," *Financ. Manag.*, vol. 39, no. 2, pp. 783–805, 2010.
[43] J. Koralun-Bereźnicka, "On the Relative Importance of Corporate Working Capital Determinants: Findings from the EU Countries," *Contemp. Econ.*, vol. 8, no. 4, pp. 415–434, 2014.
[44] J. F. T. B. Russo, "The working capital management: the determinants and the effect on profitability: Evidence from Portuguese SMEs," Master Thesis, ISCTE – Instituto Universitário de Lisboa, 2013.
[45] K. Padachi, "Trends in Working Capital Management and its Impact on Firms' Performance: An Analysis of Mauritian Small Manufacturing Firms," *Int. Rev. Bus. Res. Pap.*, vol. 2, no. 2, pp. 45–58, 2006.
[46] P. J. García-Teruel and P. Martínez-Solano, "Effects of working capital management on SME profitability," *Int. J. Manag. Financ.*, vol. 3, no. 2, pp. 164–177, 2007.
[47] M. A. Zariyawati, M. N. Annuar, H. Taufiq, and A. S. A. Rahim, "Working capital management and corporate performance: Case of Malaysia," *J. Mod. Account. Audit.*, vol. 5, no. 11, pp. 47–54, 2009.
[48] S. T. Quayyum, "Relationship between Working Capital Management and Profitability in Context of Manufacturing Industries in Bangladesh," *Int. J. Bus. Manag.*, vol. 7, no. 1, pp. 58–70, 2012.
[49] G. A. Afrifa, "Net Working Capital, Cash Flow and Performance of UK SMEs," *Rev. Account. Financ.*, vol. 15, no. 1, pp. 21–44, 2016.
[50] D. Yazdanfar and P. Öhman, "The impact of cash conversion cycle on firm profitability," *Int. J. Manag. Financ.*, vol. 10, no. 4, pp. 442–452, 2014.
[51] N. Aktas, E. Croci, and D. Petmezas, "Is working capital management value-enhancing? Evidence from firm performance and investments," *J. Corp. Financ.*, vol. 30, no. February 2015, pp. 98–113, 2015.
[52] S. Lončar and M. Ćurak, "Efficiency of working capital management and its influence on the profitablity - the case of Croatian companies," *J. Int. Res. Publ. Econ. Bus.*, vol. 2, no. 2008, pp. 246–265, 2008.
[53] M. S. Nazir and T. Afza, "Impact of Aggressive Working Capital Management Policy on Firms' Profitability," *J. Appl. Financ.*, vol. 15, no. 8, pp. 19–31, 2009.
[54] H. Nobanee and M. AlHajjar, "A Note on Working Capital Management and Corporate Profitability of Japanese Firms." p. 9, 2009.
[55] D. M. Mathuva, "The Influence of Working Capital Management Components on Corporate Profitability: A Survey on Kenyan Listed Firms," *Research Journal of Business Management*, vol. 4, no. 1. pp. 1–11, 2010.
[56] H. Shin and L. Soenen, "Efficiency of working capital management and corporate profitability," *Financ. Pract. Educ.*, vol. 8, no. 2, pp. 37–45, 1998.
[57] A. Gill, N. Biger, and N. Mathur, "The Relationship Between Working Capital Management And Profitability: Evidence From The United States," *Bus. Econ. J.*, vol. 10, no. 1, pp. 1–9, 2010.
[58] A. B. Czyzewski and D. W. Hicks, "Hold Onto Your Cash," *Manag. Account.*, vol. 73, no. 9, pp. 27–30, 1992.
[59] S. Baños-Caballero, P. Martínez-Solano, and P. J. García-Teruel, "Working capital management in SMEs," *Account. Financ.*, vol. 50, no. 457, pp. 511–527, 2010.
[60] FINA, "Analiza 2008: Financijski rezultati poduzetnika Republike Hrvatske," FINA, Zagreb, 2009.

Pregled performansi djelatnosti J61 Telekomunikacije u Republici Hrvatskoj

M. Bubanić*
* Fakultet organizacije i informatike/Katedra za gospodarstvo, Varaždin, Hrvatska
mabubanic@foi.hr

***Sažetak* - Djelatnost telekomunikacija smatra se jednim od značajnijih segmenata razvoja svakog gospodarstva, pa tako i hrvatskog. Cilj rada jest utvrditi performanse poduzeća klasificiranih prema Nacionalnoj klasifikaciji djelatnosti 2007. u područje J61 Telekomunikacije u Republici Hrvatskoj. U radu su analizirani strukturno-poslovni pokazatelji područja djelatnosti J61 Telekomunikacije, izračunati temeljem podataka iz baze podataka Financijske agencije (Financijski rezultati poslovanja poduzetnika Republike Hrvatske obveznika poreza na dobit) za razdoblje od 2008. do 2017. godine. Analiza uključuje analizu stanja i kretanja broja poduzeća s osvrtom na kategorije veličine poduzeća, kao i analizu stope preživljavanja poduzeća. Kretanja unutar populacije poduzetnika telekomunikacijskih usluga ukazuju na pozitivan demografski prirast mjeren neto ulaznim stopama. Osim poslovne demografije, analiza uključuje analizu likvidnosti i profitabilnosti, produktivnosti rada, te analizu koncentracije. Provedena analiza performansi, predstavlja temelj za ocjenu tržišne strukture djelatnosti telekomunikacija u Hrvatskoj.**

Ključne riječi - telekomunikacije; performmanse; poslovna demografija

I. UVOD

Autori Šain, Krajnović i Predovan [1], te Kovačević, Krajnović i Čičin-Šain [2] ocjenjuju tržište telekomunikacija u Republici Hrvatskoj kao oligopolističko. Na takvom tržištu poduzeća mogu zaračunavati cijene iznad graničnog troška što omogućava ostvarivanje iznadprosječnih profita. Svrha predmetnog rada jest analizirati profitabilnost poduzeća koja djeluju u djelatnosti telekomunikacija u Republici Hrvatskoj kao i ostale performanse poduzetnika iz promatrane djelatnosti.

U Hrvatskoj analiza performansi provedena je za različite djelatnosti, a koje uključuju varijable: profitabilnost, likvidnost, kapitalna intenzivnost, broj i distribucija poduzeća, stupanj koncentracije itd. Primjerice, Kovačević i Vuković [3] i [4] proveli su analizu performansa djelatnosti informacijsko-komunikacijske tehnologije gdje je analizirana distribucija poduzeća s obzirom na njihov broj i veličinu, analizirani su ukupni prihodi, broj zaposlenih, produktivnost, stopa preživljavanja i kapitalna intenzivnost. Korent [5] u analizu performansi djelatnosti za prerađivačku industriju i djelatnost trgovine na veliko i malo uključuje minimalnu efikasnu veličinu, koeficijent kapitalne intenzivnosti, broj i veličinu poduzeća, zaposlenost, ukupne prihode i stupanj koncentracije. Za hrvatsku mljekarsku industriju ocjenjena je profitabilnost te je ista komparirana sa rentabilnošću mljekarske industrije u Srbiji i Sloveniji [6]. Analiza značajki softverske industrije u Hrvatskoj [7] obuhvaća broj i relativnu veličinu poduzeća, tehnološke i troškovne uvjete, uvjete potražnje, broj i relativnu snagu kupaca, te stupanj jednostavnosti ulaska i izlaska iz nje. Analiza broja zaposlenih, ukupnih prihoda, dobiti i gubitka, te vanjskotrgovinskog poslovanja analizirana je za tekstilnu industriju u Hrvatskoj [8]. Temeljem nabrojenih analiza, odabrane su najčešće korištene varijable koje predstavljaju temelj za provedbu pregleda performansi djelatnosti telekomunikacija u Hrvatskoj, a koja obuhvaća analizu profitabilnosti, analizu broja i distribuciju poduzeća, stopu preživljavanja, stupanj koncentracije, likvidnost i produktivnost.

II. METODOLOŠKI OKVIR ISTRAŽIVANJA

U analizi su korišteni podaci Financijske agencije (FINA-e). Podaci obuhvaćaju registrirane poduzetnike u Republici Hrvatskoj čija je pretežita djelatnost prema Nacionalnoj klasifikaciji djelatnosti (2007) odjeljak J61 Telekomunikacije. Vremensko razdoblje obuhvaća podatke od 2008. do 2017. godine, te skup podataka obuhvaća samo poduzetnike koji su predali financijske izvještaje za referentnu godinu.

Odjeljak J61 Telekomunikacije obuhvaća četiri razreda: J6110 Djelatnosti žičane telekomunikacije, J6120 Djelatnosti bežične telekomunikacije, J6130 Djelatnosti satelitske telekomunikacije, te J6190 Ostale telekomunikacijske djelatnosti.

Za izračun pokazatelja korišteni su podaci iz temeljnih financijskih izvještaja (bilance i računa dobiti i gubitka) koje su poduzetnici predali FINA-i, a koji obuhvaćaju: broj zaposlenih, ukupne prihode, ukupne rashode, dobit prije oporezivanja, dobit nakon oporezivanja, ukupnu imovinu (aktiva), kapital i rezerve, kratkotrajnu imovinu i kratkoročne obveze. Podaci su obrađeni primjenom programskog alata MS Excel. Korištena je metoda deskriptivne statistike te su dobiveni rezultati prikazani grafički i tabelarno, u svrhu interpretacije određenih rezultata korištene su mjere centralne tendencije.

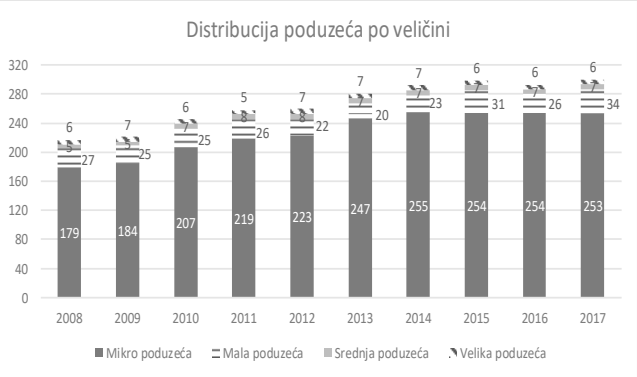
III. PERFORMANSE PODUZETNIKA IZ DJELATNOSTI TELEKOMUNIKACIJE U REPUBLICI HRVATSKOJ

A. Broj i distribucija poduzeća po veličini, stope preživljavanja

Slika 1. prikazuje distribuciju poduzeća odjeljka J61 Telekomunikacija s obzirom na veličinu poduzeća. Poduzeća su klasificirana na mikro, mala, srednja i velika

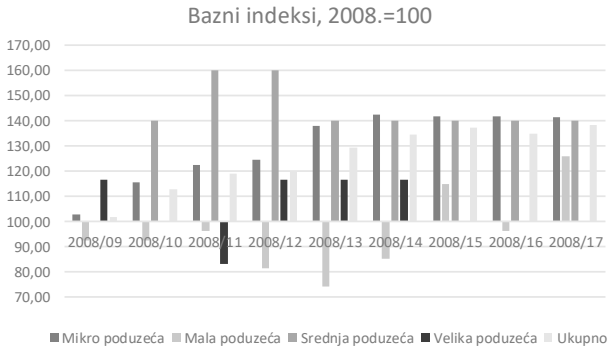
poduzeća, prema kriterijima iz članka 5. Razvrstavanje poduzetnika Zakona o računovodstvu [9]. Kriteriji o razvrstavanju poduzetnika mijenjali su se tokom promatranog razdoblja (od 2008. do 2017. godine). Kako bi se izbjegla metodološka pogreška izazvana tim promjenama, te omogućila komparacija poduzeća s obzirom na veličinu kroz promatrano razdoblje, za njihovu klasifikaciju su korišteni jedinstveni Zakonski kriteriji [9].

Distribucija poduzeća ukazuje kako promatrana djelatnost u većini godina, izuzev 2016., bilježi pozitivan prirast broja poduzeća. Mikro poduzeća u prosjeku čine 85,24% djelatnosti, mala poduzeća 9,81%, srednja 2,56%, a velika 2,40%. Kontinuirani rast broja poduzeća upućuje na privlačnost ove djelatnosti.



Slika 1. Distribucija poduzeća djelatnosti J61 Telekomunikacije s obzirom na velčinu od 2008. do 2017 godine

Bazni indeksi (2008.=100) prikazani na Slici 2. pokazuju da je ukupan broj poduzeća u 2017. godini u odnosu na 2008. porastao za 38,25% (u apsolutnom iznosu 83 poduzeća). Najveći prirast od početka do kraja promatranog razdoblja ostvarila su mikro poduzeća za 41,34% (74 poduzeća), zatim srednja poduzeća sa 40% (u apsolutnom iznosu to su tek dva poduzeća), zatim mala poduzeća sa 25,93%, dok je broj velikih poduzeća u 2017. godini u odnosu na početak promatranog razdoblja ostao nepromijenjen (6 poduzeća).



Slika 2. Bazni indeksi kretanja broja poduzeća s obzirom na veličinu poduzeća, te ukupan broj poduzeća (2008.=100)

S obzirom na to da neto ulasci/izlasci predstavalju razliku između broja poduzeća u razdoblju t (promatrana godina) i broja poduzeća t-1 (prethodna godina), oni „prikivaju“ stvarni broj ulazaka i izlazaka poduzeća. S ciljem boljeg uvida u demografiju i mobilnost poduzeća valja promatrati bruto ulaske i izlaske. Tablica I. prikazuje bruto izlaske (I) i bruto ulaske (U) u odjeljak J61 Telekomunikacije. Bruto izlasci predstavljaju stvarni broj izlazaka poduzeća iz djelatnosti u promatranom vremenskom razdoblju, dok bruto ulasci predstavljaju stvarni broj ulazaka poduzeća u djelatnost u promatranom vremenskom razdoblju, u konkretnom slučaju u razdoblju od 2008. do 2017. godine. Pri izračunu bruto ulazaka i izlazaka, mobilnost poduzeća koja podrazumijeva prelazak poduzeća iz jedne kategorije veličine u drugu kategoriju, u ovom se slučaju tretira kao bruto ulazak, odnosno izlazak. Iako mobilnost ne predstavlja stvarni ulazak/izlazak iz djelatnosti, u ovom slučaju se ona tretira tako jer se nastoji naglasiti dinamika kretanja poduzeća. Bruto ulasci i izlasci poduzeća pokazuju da je najveća dinamika prisutna kod mikro poduzeća u svim promatranim godinama. Iako bazni indeksi pokazuju da u 2017. godini naspram 2008. nije bilo promjena u broju velikih poduzeća, bruto ulasci i izlasci pokazuju da je tijekom promatranog razdoblja došlo do i/ili ulazaka/izlazaka i/ili migracije poduzeća.

Stopa preživljavanja računata je po uzoru na metodu koju su primijenili autori Kovačević i Vuković [3] koristeći sljedeći izraz (1):

$$\%\ preživljavanja = \frac{broj\ preživjelih\ poduzeća\ iz\ odjeljka\ J61}{broj\ poduzeća\ odjeljka\ J61} \quad (1)$$

pri čemu broj poduzeća odjeljka J61 čine sva poduzeća koja su se zatekla u navedenom odjeljku 2008. godine, a

TABLICA I. BRUTO ULAZI I IZLAZI PODUZEĆA OD 2008. DO 2017. GODINE S OBZIROM NA VELIČINU

Godine	**2009/08**		**2010/09**		**2011/10**		**2012/11**		**2013/12**		**2014/13**		**2015/14**		**2016/15**		**2017/16**	
Veličina poduzeća	**I**	**U**	**I**	**U**	**I**	**U**	**I**	**U**	**I**	**U**	**I**	**U**	**I**	**U**	**I**	**U**	**I**	**U**
Mikro poduzeća	34	39	27	50	25	37	27	31	27	51	34	42	40	39	38	38	34	33
Mala poduzeća	8	6	6	6	4	5	7	3	7	5	5	8	3	11	9	4	5	13
Srednja poduzeća	1	1	1	3	2	3	2	2	1	0	1	1	1	1	0	0	1	1
Velika poduzeća	0	1	1	0	1	0	0	2	0	0	1	1	1	0	0	0	0	0

broj preživjelih poduzeća iz odjeljka J61 čine poduzeća koja su kontinuirano perzistirala u promatranom odjeljku od 2008. pa do godine za koju se izračunava stopa preživljavanja. Iz Tablice II. vidljivo je kako se u 2008. godini u odjeljku J61 našlo 217 poduzeća, te je prikazano koliko je iz godine u godinu broj preživjelih poduzeća od početnih 217 u 2008. godini. Stopa preživljavanja pokazuje kako je od 2008. godine do 2017. godine u odjeljku J61 Telekomunikacija preživjelo 41,47% poduzeća.

TABLICA II. BROJ PREŽIVJELIH I STOPA PREŽIVLJAVANJA PODUZEĆA OD 2008. DO 2017. GODINE

Godina	Broj preživjelih poduzeća	% preživljavanja
2008	217	100,00%
2009	184	84,79%
2010	166	76,50%
2011	155	71,43%
2012	141	64,98%
2013	128	58,99%
2014	122	56,22%
2015	112	51,61%
2016	96	44,24%
2017	90	41,47%

Prema metodologiji koju koristi Državni zavod za statistiku, stopa preživljavanja novorođenih poduzeća definirana je kao udio rođenih poduzeća u ukupnoj populaciji rođenih, a koja su nastavila neprekinutu aktivnost nakon godine rođenja [10]. Broj novorođenih poduzeća u 2009. godini iznosi 36[1]. Od ukupnog broja novorođenih, promatra se broj poduzeća koja su preživjela određeni broj godina. Tablica III. prikazuje broj preživjelih novorođenih poduzeća i stopu preživjelih novorođenih poduzeća, gdje je vidljivo kako je u 2017. godini od ukupnog broja novorođenih u 2009. preživjelo 58,33% poduzeća.

TABLICA III. BROJ PREŽIVJELIH NOVOROĐENIH I STOPA PREŽIVLJAVANJA NOVOROĐENIH PODUZEĆA OD 2009. DO 2017. GODINE

Godina	Preživjela novorođena poduzeća	Stopa preživljavanja novorođenih poduzeća
2009	36	100,00%
2010	29	80,56%
2011	27	75,00%
2012	26	72,22%
2013	25	69,44%
2014	24	66,67%
2015	23	63,89%
2016	22	61,11%
2017	21	58,33%

[1] Broj novorođenih poduzeća izračunat je za 2009. godinu u odnosu na 2008. godinu, jer autoru nisu dostupni podaci za 2007. godinu, te stoga nije bilo moguće izračunati broj novorođenih poduzeća u 2008. godini. Broj novorođenih poduzeća u 2009. godini nije identičan izračunatom broju bruto ulazaka, jer bruto ulasci uključuju i mobilnost iz različitih kategorija veličine poduzeća što se ne smatra novorođenim poduzećem.

B. Stupanj koncentracije

Stupanj koncentracije predstavlja mjeru tržišne moći [11]. Kako bi se ocijenio stupanj koncentracije djelatnosti telekomunikacija korištene su dvije kvantitativne mjere industrijske koncentracije: (a) koeficijent koncentracije reda n (C_n) i (b) Herfindahl-Hirschmannov indeks (HHI). Izračun obiju pokazatelja temelji se na udjelu ukupnih prihoda pojedinog poduzeća iz djelatnosti u ukupnim prihodima ostvarenim u odjeljku J61 Telekomunikacije, pomnožen sa 100 (s_n). Koeficijent koncentracije reda *n* (C_n) mjeri koliki je zbroj relativnih udjela prvih *n* najvećih poduzeća u djelatnosti. Koeficijent koncentracije C_4 sumira tržišne udjele četiri najveća poduzeća, dok C_8 uključuje sumu tržišnih udjela osam najvećih poduzeća. Stoga je za izračun navedenog pokazatelja potrebno najprije poduzeća rangirati s obzirom na veličinu udjela (od najvećeg do najmanjeg), te se zatim zbrajaju udjeli za onoliki broj poduzeća (*n)* koliko ih je predviđeno mjerenjem [11]. Za ocjenu stupnja koncentracije djelatnosti telekomunikacija u Hrvatskoj korišten je koeficijent koncentracije C_4 i C_8. Rezultati su prikazani u Tablici IV.

TABLICA IV. KOEFICIJENT KONCENTRACIJE REDA ČETIRI I REDA OSAM ODJELJKA J61 TELEKOMUNIKACIJE OD 2008. DO 2017. GODINE

Godina	**C4**	**C8**
2008	89,18	94,99
2009	84,82	93,34
2010	85,31	92,14
2011	84,52	91,70
2012	82,76	91,28
2013	83,04	91,50
2014	83,05	92,19
2015	83,42	92,01
2016	84,66	92,44
2017	83,19	91,85
Prosjek	84,39	92,34

Rezultati izneseni u Tablici IV. pokazuju kako se radi o tržištu na kojem kontinuirano dominiraju četiri vodeća i ostala manja poduzeća [12]. Od 2008. do 2009. godine vodeća poduzeća u odjeljku J61 Telekomunikacija su Hrvatski telekom d.d., T-Mobile Hrvatska d.o.o., Vipnet d.o.o. i Tele2 d.o.o. koja zajedno drže 89,19% tržišta telekomunikacija u Hrvatskoj. Godine 2010. spajanjem Hrvatskog telekoma d.d. i T-Mobile Hrvatske d.o.o. u Hrvatski telekom d.d. dolazi do povećane koncentracije te navedeno poduzeće u referentnoj godini drži 53,68% tržišta. Te iste godine četvrto mjesto vodećih poduzeća zauzima OT-Optima Telekom d.d., te navedena vodeća

poduzeća zadržavaju svoje pozicije sve do 2017. godine. Od 2011. godine udio Hrvatskog telekoma d.d. pokazuje tendenciju postupnog smanjenja (s 53,68% u 2010. godini, na 46,03% u 2017. godini), dok Tele2 d.o.o. pokazuje tendenciju postupnog povećanja tržišnog udjela (s 3,55% u 2008. godini, na 10,19% u 2017. godini).

Herfindahl-Hirschmannov indeks (HHI) predstavlja sumu kvadrata tržišnih udjela svih *n* poduzeća u određenoj djelatnosti [13] odnosno jednak je izrazu (2). Izračunava se na način da se izračunati tržišni udjeli za svako pojedino poduzeće kvadriraju, te se potonje kvadrirani tržišni udjeli sumiraju. Vrijednost HHI indeksa kreće se od nula do 10.000. Prilikom ocjene tržišne koncentracije, prednost se daje HHI-u, jer HHI naspram koeficijenta koncentracije (a) uzima u obzir sva poduzeća koja se nalaze u promatranoj djelatnosti, a ne samo četiri ili osam najvećih poduzeća, te (b) kvadriranjem tržišnih udjela pridaje veća važnost poduzećima s većim tržišnim udjelom.

$$HHI = \sum s_n^2 \quad (2)$$

U Tablici V. prikazani su rezultati Herfindahl-Hirschmannovog indeksa (HHI), te pokazuju kako je u prosjeku djelatnost telekomunikacija visoko koncentrirano tržište [14]. Ovakvi rezultati komplementarni su s dobivenim koeficijentima koncentracije. U 2010. naspram 2009. godini vidljiv je značajni porast HHI indeksa što je u skladu s analizom da je te godine došlo do spajanja Hrvatskog telekoma d.d. i T-Mobile Hrvatske d.o.o. u Hrvatski telekom d.d. čime su zauzeli više od polovice tržišnog udjela. Postupno smanjenje tržišne moći Hrvatskog telekoma d.d. vidljivo je i u postupnom smanjenju HHI indeksa.

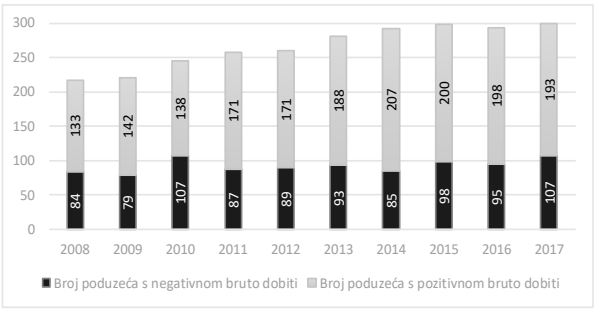
TABLICA V. HERFINDAHL-HIRSCHMANNOV INDEKS (HHI) ZA DJELATNOSTI J61 TELEKOMUNIKACIJA OD 2008. DO 2017. GODINE

Godina	HHI
2008	2604,92
2009	2224,09
2010	3444,17
2011	3328,10
2012	3047,35
2013	2955,76
2014	2867,98
2015	2886,42
2016	2926,95
2017	2805,67
Prosjek	2909,14

C. Profitabilnost

Rentabilnost kao jedan od temeljnih ciljeva poslovanja [15] predstavlja ujedno i odrednicu atraktivnosti za potencijalne ulaznike u određenu djelatnost. Mogućnost postavljanja cijene iznad graničnog troška, omogućava ostvarivanje iznadprosječnih profita, što privlači nova poduzeća na tržište. Slika 3. prikazuje distribuciju broja poduzeća u djelatnosti telekomunikacija u Hrvatskoj koja su ostvarila pozitivnu odnosno negativnu dobit prije oporezivanja[2] od 2008. do 2017. godine. Kroz promatrano razdoblje, u prosjeku 65% od ukupnog broja poduzeća ostvaruje pozitivnu dobit prije oporezivanja.

Profitabilnost djelatnosti mjerena je s dva pokazatelja: (a) rentabilnost ukupne imovine (ROI) i (b) rentabilnost vlastitog kapitala (ROE). Rentabilnost ukupne imovine predstavlja odnos između bruto dobit ili neto dobiti te ukupne imovine, shodno čemu se razlikuje bruto i neto rentabilnost imovine [16]. Slika 4. prikazuje bruto i neto rentabilnost ukupne imovine djelatnosti telekomunikacija u Hrvatskoj[3]. Rentabilnost ukupne imovine pokazuje



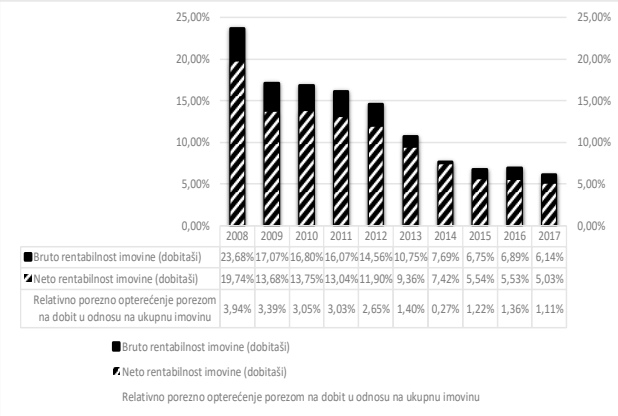
Slika 3. Broj poduzeća u odjeljku J61 Telekomunikacije koji ostvruju pozitivnu i negativnu bruto dobit od 2008. do 2017. godine

tendenciju smanjenja kako kroz godine raste broj poduzeća u djelatnosti, a što je u skladu s teorijskim postulatima da visoka razina profitabilnosti privlači nove ulaznike, što dovodi do iscrpljivanja profitabilnosti, te čini djelatnost za nove ulaznike manje atraktivnom [17].

Razlika između pokazatelja bruto i neto profitabilnosti ukupne imovine pokazuje relativno porezno opterećenje porezom na dobit u odnosu na ukupnu imovinu [16], a što je vidljivo na Slici 4. Relativno porezno opterećenje porezom na dobit u odnosu na ukupnu imovinu pokazuje tendenciju smanjenja kroz promatrano razdoblje. Porez na dobit poduzeća iz odjeljka J61 Telekomunikacija se od 2008. do 2014. godine kontinuirano smanjivao, dok je ukupna imovina u promatranom razdoblju u prosjeku pokazivala relativno povećanje. Veće relativno smanjenje poreza na dobit u odnosu na relativno povećanje ukupne imovine dovelo je do kontinuiranog smanjenja relativnog poreznog opterećenja porezom na dobit u odnosu na ukupnu imovinu u promatranom razdoblju. U 2014. godini naspram 2013. došlo je do smanjenja poreza na dobit za 70,72% te istovremeno do relativnog povećanja ukupne imovine za 6,69% što je dovelo do relativnog poreznog opterećenja porezom na dobit u odnosu na ukupnu imovinu od 0,27%. U 2015. godini došlo je do relativnog povećanja porezom na dobit u odnosu na 2014. godinu, dok je ukupna imovina zabilježila relativno smanjenje, što je dovelo do rasta poreznog opterećenja.

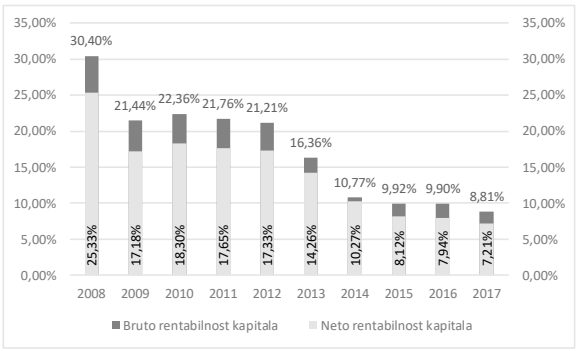
[2] Poduzeća koja su ostvarila dobit prije oporezivanja jednaku nuli, odnosno ukupni prihodi jednaki ukupnim rashodima, u analizi su obuhvaćena u skupu s poduzećima koja su ostvarila pozitivnu bruto dobit.

[3] Bruto rentabilnosti ukupne imovine odjeljka J61 Telekomunikacija izračunata je na način da je u brojnik uključena suma bruto dobiti poduzeća koja su u promatranoj godini ostvarila dobit prije poreza veću ili jednaku od nule, a u nazivnik je uključena suma ukupne imovine samo onih poduzeća koja su ostvarila bruto dobit veću ili jednaku nuli. Na sličan način izračunata je i neto rentabilnost ukupne imovine, osim što su u obzir uzeta poduzeća koja su ostvarivala neto dobit veću ili jednaku nuli.



Slika 4. Bruto i neto rentabilnost imovine odjeljka J61 Telekomunikacija od 2008. do 2017. godine

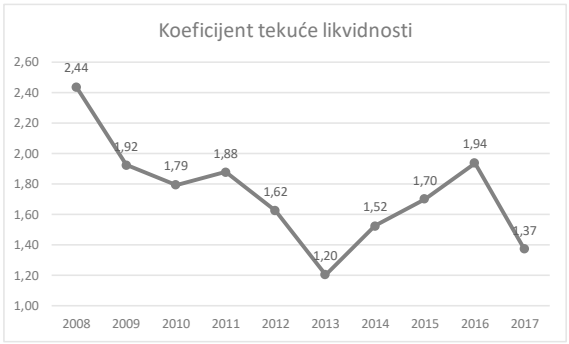
Rentabilnost vlastitog kapitala izračunava se kao odnos između bruto ili neto dobiti i vlastitog kapitala, izražen u postocima, te ovisno o korištenoj varijabli razlikujemo bruto i neto rentabilnost vlastitog kapitala [16]. Slika 5. prikazuje bruto i neto rentabilnost vlastitog kapitala djelatnosti telekomunikacija u Hrvatskoj [4]. Kao i rentabilnost ukupne imovine, i rentabilnost vlastitog kapitala pokazuje tendenciju smanjenja kako kroz vrijeme dolazi do povećanja broja poduzeća u promatranoj djelatnosti. Kapital i rezerve do 2013. godine pokazuju relativno smanjenje iz godine u godinu, te od 2014. godine pa do kraja promatranog razdoblja pokazuju kontinuirano relativno povećanje, ali što i dalje nije bilo dovoljno da postignu početnu veličinu iz 2008. godine. Kapital i rezerve su u 2017. u odnosu na 2008. godinu smanjene za 13,70%.



Slika 5. Bruto i neto rentabilnost vlastitog kapitala odjeljka J61 Telekomunikacija od 2008. do 2017. godine

D. Likvidnost

Likvidnost je pojam koji se veže uz cirkulaciju imovine u poslovnim transakcijama [18]. Pokazatelji likvidnosti mjere sposobnost podmirivanja dospjelih obveza kroz odnose kratkotrajne imovine i kratkoročnih obveza te su izravno vezani uz solventnost poslovanja [15]. Za ocjenu likvidnosti djelatnosti telekomunikacija u Hrvatskoj korišten je koeficijent tekuće likvidnosti izračunat kao odnos sume kratkotrajne imovine svih poduzeća koja su se našla u odjeljku J61 Telekomunikacija u određenoj godini, te sume kratkoročnih obveza tih poduzeća. Slika 6 prikazuje kretanje koeficijenta tekuće likvidnosti odjeljka J61 Telekomunikacija. U svim godinama iz kratkotrajne imovine moguće je podmiriti kratkoročne obveze. Koeficijent tekuće likvidnosti se u 2013. godini u odnosu na 2008. smanjio za 50,81%. U istom razdoblju kratkotrajna imovina se smanjila za 24,84%, dok su se kratkoročne obveze povećale za 52,30%. U 2016. godini u odnosu na 2013. kratkotrajna imovina se smanjila za 0,18%, dok su se kratkoročne obveze smanjile za 38,02%. Relativno veće smanjenje kratkoročnih obveza u odnosu na relativno smanjenje kratkotrajne imovine dovelo je do



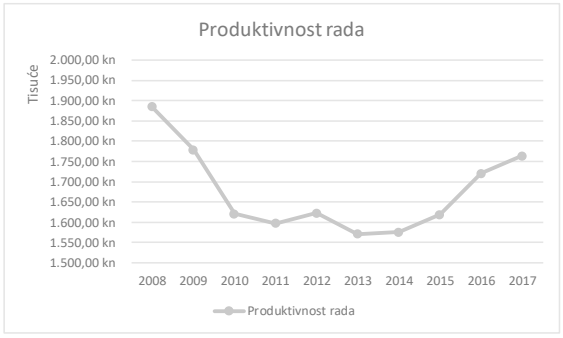
Slika 6. Koeficijent tekuće likvidnosti odjeljka J61 Telekomunikacija od 2008. do 2017. godine

povećanja koeficijenta tekuće likvidnosti u istom razdoblju za 61,67%. U 2017. godini u odnos na 2016. smanjenje koeficijenta tekuće likvidnosti posljedica je relativno većeg povećanja kratkoročnih obveza za 42,34% u odnosu na relativno povećanje kratkotrajne imovine za 0,75%.

E. Produktivnost

Proizvodnost rada općenito se definira kao odnos između količine učinaka i količine rada utrošenog za stvaranje tih učinaka [19]. Veća produktivnost doprinosi većoj učinkovitosti djelatnosti, a time i cjelokupnog gospodarstva. Produktivnost rada odjeljka J61 Telekomunikacija prikazana je na Slici 7. Produktivnost je mjerena kao odnos sume ukupnih prihoda i ukupnog broja zaposlenih u odjeljku J61. U promatranom odjeljku prosječno od 2008. do 2017. godine je po zaposleniku ostvareno 1.675.242,96 kn ukupnih prihoda. Produktivnost rada u 2013. u odnosu na 2008. godinu je smanjena za 16,67%, što je posljedica relativno većeg smanjenja ukupnih prihoda za 24,51%, ali i relativnog smanjenja zaposlenih za 9,41%. U 2017. godini u odnosu na 2013. došlo je do rasta produktivnosti za 12,29% što je posljedica povećanja ukupnih prihoda za 0,83%, te smanjenja zaposlenih za 10,20%.

[4] Bruto rentabilnosti vlastitog kapitala odjeljka J61 Telekomunikacija izračunata je na način da je u brojnik uključena suma bruto dobiti poduzeća koja su u promatranoj godini ostvarila dobit prije poreza veću ili jednaku od nule, te čije su kapital i rezerve bile veće ili jednake nuli, a u nazivnik je uključena suma kapitala i rezervi samo onih poduzeća koja su ostvarila bruto dobit veću ili jednaku nuli, te čije su kapital i rezerve bile veće ili jednake nuli. Na sličan način izračunata je i neto rentabilnost vlastitog kapitala, osim što su u obzir uzeta poduzeća koja su ostvarivala neto dobit veću ili jednaku nuli i kapital i rezerve veće ili jednake nuli.



Slika 7. Produktivnost rada odjeljka J61 Telekomunikacija od 2008. do 2017. godine (u 000 HRK)

IV. ZAKLJUČAK

Stopa preživljavanja poduzeća devetogodišnjeg razdoblja od 41,47% i stopa preživljavanja novorođenih poduzeća osmogodišnjeg razdoblja od 58,33% govore u prilog tome da kad poduzeća uspiju ući u djelatnost telekomunikacija, velik dio njih preživi te ostvaruje rast tržišnog udjela što dovodi do smanjenja tržišne koncentracije djelatnosti. Usprkos navedenom, stupanj koncentracije telekomunikacijske djelatnosti je i dalje visok, što potvrđuje rezultate prethodnog istraživanja [1] te upućuje na postojanje četiri dominantna poduzeća i većeg broja manjih poduzeća. Iako se stupanj koncentracije tokom godina smanjuje, djelatnost telekomunikacija je i dalje daleko od fragmentiranog tržišta.

Djelatnost telekomunikacija u Hrvatskoj iskazuje na početku promatranog razdoblja visoke stope profitabilnosti imovine i vlastitog kapitala što se može protumačiti kao odrednica atraktivnosti za potencijalne ulaznike. Iz promatranog se vidi korelacija rasta broja poduzeća kroz promatrano vremensko razdoblje i topljenja profitabilnosti.

Djelatnost telekomunikacija u promatranom razdoblju iz tekuće imovine može podmiriti dospjele tekuće obveze. Iako je rast kratkoročnih obveza doveo do smanjenja likvidnosti, financijska stabilnost djelatnosti je i dalje neupitna. Smanjenje broja zaposlenih posljedično se očituje u promjenama produktivnosti rada. Iako je na smanjenje produktivnosti do 2013. godine utjecalo i smanjenje ukupnih prihoda i smanjenje broja zaposlenih, na rast produktivnosti nakon 2013. godine uglavnom utječe daljnji pad ljudskih potencijala.

Temeljem iznesenih rezultata, nerealno je očekivati značajnije promjene u skoroj budućnosti u djelatnosti J61 Telekomunikacije u Hrvatskoj. Tržišni udjeli dvaju najvećih „diva“ u razdoblju od 2008. do 2017. godine nisu se značajnije mijenjali. Ulazak novih giganta, preuzimanjem postojećih, ne dovodi do veće konkurentnosti (a čemu bi trebali težiti), već samo zauzimanju postojećeg mjesta novim imenom. Pozitivne neto ulazne stope ukazuju na postepeno labavljenje tržišta, ali ono je sporo.

Telekomunikacije omogućuju razmjenu informacija između različitih korisnika, bez vremenskog i prostornog ograničenja, te stoga predstavljaju važnu gospodarsku djelatnost. Iz navedenog proizlazi uska povezanost poduzeća iz djelatnosti telekomunikacija i poduzeća iz djelatnosti informacijske tehnologije. Njihova međuovisnost predstavlja temelj razvoja inovacija, a što dovodi do rasta obiju djelatnosti i gospodarstva u cjelini [20].

Osim analiziranih performansi, za buduća istraživanja valjalo bi uključiti i analizu minimalne efikasne veličine, analizu kapitalne i/ili radne intenzivnosti, analizu zaposlenosti, plaća, ekonomičnosti. Predmetna analiza provedena je temeljem agregiranih podataka, koji djelom „zamagljuju stvarnu sliku“ poslovanja malih poduzeća bez značajnijeg utjecaja na djelatnost. Za podobnije zaključke o poslovanju pojedinih poduzeća u promatranoj djelatnosti autorica ima namjeru provesti i analizu temeljenu na mikroekonomskim podacima poduzeća registriranih u djelatnosti J61 Telekomunikacija koja bi navedene nedostatke dijelom otklonila. Dobivene rezultate budućih istraživanja valjalo bi usporediti s rezultatima analiza ostalih djelatnosti, a što će omogućiti rangiranje djelatnosti s obzirom na njihovu važnost u cjelokupnom gospodarstvu Republike Hrvatske.

Bez obzira na trendove, poduzeća koja u djelatnosti telekomunikacija slove kao dominantni starosjedioci i dalje uživaju prednosti ovakve tržišne strukture, te nemaju razloga za njenim labavljenjem. Međutim, postavlja se pitanje utjecaja na blagostanje i nedostataka koje donosi ovakva tržišna struktura za gospodarstvo u cjelini.

LITERATURA

[1] D. Č. Šain, A. Krajnović, and M. Predovan, “Oligopol na tržištu telekomunikacijskih usluga u Republici Hrvatskoj,” Oeconomica Jadertina, vol. 1, no. 1, pp. 57–75, 2011.

[2] D. Kovačević, A. Krajnović i D. Čičin-Šain, "MARKET ANALYSIS OF THE TELECOMMUNICATIONS MARKET – THE CASE OF CROATIA", DIEM, vol.3, br. 1, str. 161-175, 2017. [Online]. Dostupno na: https://hrcak.srce.hr/187376. [Citirano: 09.04.2020.]

[3] Z. Kovačević i K. Vuković, "Performanse poduzeća u hrvatskom sektoru informacijsko-komunikacijske tehnologije (ICT)", Ekonomska misao i praksa, vol., br. 2, str. 217-240, 2006. [Online]. Dostupno na: https://hrcak.srce.hr/10680. [Citirano: 28.02.2020.]

[4] Z. Kovačević i K. Vuković, "INDUSTRIJA INFORMACIJSKO-KOMUNIKACIJSKE TEHNOLOGIJE (ICT) U HRVATSKOM GOSPODARSTVU", Poslovna izvrsnost, vol.1, br. 1, str. 97-112, 2007. [Online]. Dostupno na: https://hrcak.srce.hr/38525. [Citirano: 28.02.2020.]

[5] D. Korent, “Identifikacija odrednica upravljanja obrtnim kapitalom u funkciji povećanja profitabilnosti poduzeća u odabranim djelatnostima u Republici Hrvatskoj,” Doktorski rad, Zagreb: Ekonomski fakultet Zagreb, 2018.

[6] S. Muminović i V. Pavlović, "Profitability of Dairy Industry in Slovenia, Croatia and Serbia", Mljekarstvo, vol.62, br. 2, str. 96-110, 2012. [Online]. Dostupno na: https://hrcak.srce.hr/83326. [Citirano: 28.02.2020.]

[7] N. Rupčić i M. Kurjaković, "Značajke tržišnoga natjecanja u softverskoj industriji Republike Hrvatske", Praktični menadžment, vol.5, br. 1, str. 29-40, 2014. [Online]. Dostupno na: https://hrcak.srce.hr/134945. [Citirano: 28.02.2020.]

[8] Đ. Borozan i Lj. Dragišić, "Hrvatska tekstilna industrija ne prekretnici: od preživljavanja do rasta", Ekonomski vjesnik, vol.XVIII, br. 1-2, str. 29-42, 2005. [Online]. Dostupno na: https://hrcak.srce.hr/199722. [Citirano: 28.02.2020.]

[9] Zakon o računovodstvu. (2015 – 2018). NN 78/15, 134/15, 120/16, 116/18. Dostupno: https://www.zakon.hr/z/118/Zakon-o-ra%C4%8Dunovodstvu. [Citirano: 24.02.2020.].

[10] Državni zavod za statistiku, „Osnovni pokazatelji poslovne demografije u 2017.“, 2018. [Online]. Dostupno na: https://www.dzs.hr/Hrv_Eng/publication/2018/15-02-01_01_2018.htm. [Citirano: 28.02.2020.]

[11] D. Tipurić, M. Kolaković and K. Dumičić, "Koncentracijske promjene hrvatske bankarske industrije u desetogodišnjem razdoblju (1993.-2002.)", Zbornik Ekonomskog fakulteta u Zagrebu, vol.1, br. 1, str. 1-22, 2003. [Online]. Dostupno na: https://hrcak.srce.hr/26164. [Citirano: 27.02.2020.]

[12] M. Naldi and M. Flamini, „The CR4 Index and the Interval Estimation of the Herfindahl-Hirschman Index: An Empirical Comparison,“ SSRN Electronic Journal, 10.2139/ssrn.2448656, 2014.

[13] J. R. Church and R. Ware, „Industrial organization: a strategic approach,“ Irwin McGraw Hill, 2000.

[14] M. Kostić, L. Maksimović and B. Stojanović, "The limitations of competition in the insurance markets of Slovenia, Croatia and Serbia," Economic Research-Ekonomska Istraživanja, vol. 29, (1), pp. 395-418, 2016.

[15] S. Orsag, „Vrijednosni papiri: Investicije i instrumenti financiranja,“ Revicon, 2011.

[16] I. Mamić Sačer, K. Žager, L. Žager, A. Ježovita and S. Sever Mališ, „Analiza financijskih izvještaja, 3. izmijenjeno i dopunjeno izdanje,“ Hrvatska zajednica računovođa i financijskih djelatnika, Zagreb, 2017.

[17] L. Skuflic and M. Turuk, "Barijere ulaska malih i srednjih poduzeća u prerađivačku industriju u Hrvatskoj," Business Excellence, vol. 7, (2), pp. 23, 2013.

[18] M. Mikić, S. Orsag, N. P. Vokić and S. Švaljek, (Eds.), „Ekonomski leksikon,“ Leksikografski zavod Miroslav Krleža, 2011.

[19] S. Dvorski and V. Kovšca, „Ekonomija za poduzetnike,“ TIVA-Tiskara, Varaždin, 2011.

[20] I. Lovrek, „Telekomunikacijska tehnologija i specifičnosti telekomunikacijskog tržišta“, Element, Zagreb, 2007.

IMPLEMENTATIONS OF STANDARDS & BEST PRACTICES

Implementation of NETCONF Standard by Major Customers in Croatia

D. Valencic
University of Applied Sciences Velika Gorica, Velika Gorica, Croatia
davorin_valencic@yahoo.com

Abstract - **The traditional way of managing and configuring network devices is by using Command Line Interface (CLI) method with Secure Shell (SSH) protocol and Simple Network Management Protocol (SNMP). CLI has proven to be a very complex and ineffective solution, especially with fast development and more complexity in the computer networking area. To solve these issues, IETF (Internet Engineering Task Force) standard organization created in 2006 the NETCONF (Network Configuration Protocol) standard. Using the NETCONF standard should enable standardized, simpler, programmatic and more efficient configuration of R/S (router/switch) devices. Today, more than ten years after the creation of the NETCONF standard there is the question about the actual acceptance of the NETCONF protocol in current customers' networks and about the actual maturity of the NETCONF standard. The scope of this paper is to show the results of the scientific research (qualitative or quantitative) made with major customers (partners and end customers) in Croatia with R/S network devices: in which networks and on which type of devices NETCONF is implemented, what are the reasons for NETCONF (non)implementation, their knowledge about the NETCONF standard, their view of availability of NETCONF training and documentation, etc.**

Keywords - **NETCONF, standard, computer networks, router, switch, implementation, customer**

I. INTRODUCTION

NETCONF (Network Configuration Protocol) is an important component of SDN architecture (as defined in Request for Comments (RFC) 7426) [1], and software-defined networking (SDN) is being increasingly implemented. The users who use SDN networks, sometimes implement also the NETCONF protocol (as a component of the SDN product), not even being aware of this. NETCONF performs its functions in the background and it is hidden by the SDN user application.

This paper is one in a series that covers NETCONF and its implementation that is based on the doctoral thesis *Increasing the availability of network devices by implementing the command configuration standard* (author D. Valenčić) [2]. The topic of this doctoral dissertation deals with the following:

- defining of advantages of the NETCONF standard in order to obtain higher availability of network devices;
- the current state of implementation of the NETCONF standard by world vendors of networking equipment is defined and recommendations for the increase and improvement of the implementation are proposed.
- Also, the level of implementation by the network equipment users is determined and the operation model (activities) with the users is proposed that could assure an increase in the usage of the NETCONF standard with all its advantages.

This paper covers the last abovementioned topic of the doctoral thesis (Implementation of NETCONF by the network equipment users in Croatia).

As part of the research the hypothesis was tested which says: "the users of network equipment do not implement the NETCONF standard on their network devices" that is the main topic of this article.

II. RESEARCH RESULTS: IMPLEMENTATION OF NETCONF STANDARD ON R/S DEVICES BY MAJOR CUSTOMERS IN CROATIA

A. Data processing method

This section presents the use of tabular presentation method of the structure of responses to:

- Interview questions by the employees of the vendors' partners, and
- Survey questions by the major users in Croatia.

Figure 1 shows an overview of information about the selection of examinees and their companies for the research [2].

Testing the differences in relation to the theoretical value is performed by the use of Wilcoxon test for one independent sample. The analysis was performed in the statistical software SPSS 25 and the conclusions have been made at the border significance of 5%.

University of Applied Sciences Velika Gorica, Velika Gorica, Croatia

Method of research / type of data	Interview vendors	Interview partners	Questionnaire users
Company of the respondent:	The biggest world R/S device vendors.	The biggest partners of R/S device vendors in Croatia.	Companies in Croatia that have or maintain larger and major computer networks.
Condition that has to be met by the selected companies (required value in the research plan):	Coverage of R/S market: min 80%	Coverage of R/S market: min 60%	Min 50 end users who have availability $\geq$ 99.9% (downtime annually $\leq$ 8.76 hours).
Condition that has to be met by the selected companies (met value):	Coverage of R/S market: 82%	Coverage of R/S market: 64%	53 end users
Description of the selected companies:	Four world vendors of highest presence in Croatia	Seven out of eleven biggest partners in Croatia	73 companies out of which twenty are partners (system integrators) and 53 end users.
Number of obtained interviews / questionnaires:	5	7	73
Condition met by the respondents:	Technical specialists who have the highest level of knowledge and experience for R/S devices and technologies in their companies for Croatia and the region.	Technical specialists who have the highest level of knowledge and experience for R/S devices and technologies in their companies.	Work on technical activities related to computer networks.
Knowledge (certification) of respondents in the field of computer networks:	3 advanced (CCIE) level, 2 intermediate (CCNP) level.	6 advanced (CCIE) level, 1 intermediate (CCNP) level.	5 advanced (CCIE) level, 27 intermediate (CCNP) level, 26 basic (CCNA) level, 14 no passed certificates.

Figure 1. Overview of examinee selections for research [2]

B. *Results of interviews with the R/S device vendors*

1) *Implementation of the NETCONF standard on R/S devices*

Responding to question "In which types of networks do you use NETCONF on R/S network devices and in which percentage by your users?" the examinees expressed their opinion in the class values from 0 to 5, where zero denoted 0% whereas five denoted 100%.

Table I. shows the obtained results ordered from the highest to the lowest average value. The examinees use most NETCONF on R/S network devices in the networks of service providers (average value of response 3.33 (66.6%); std.dev. 1.15). The examinees, however, do not use NETCONF in IoT networks to which no examinee gave an answer. The total level of examinees' responses marked an average value of 0.94 with the average deviation from the arithmetic mean 0.84 which shows very low average implementation (18.8%) of the NETCONF standard in different types of networks.

TABLE I. IMPLEMENTATION OF THE NETCONF STANDARD ON R/S DEVICES BY USERS (INTERVIEW PARTNERS) [2]

	N	Average	Std. Dev.	Z	P*
Internet service provider – ISP. (Includes all types of ISP networks, e.g. broadband, wireless, core, access)	3	3.33	1.15	0.58	0.564
Data Center – DC networks	6	1.50	1.05	2.04	0.041
WAN networks of major enterprises (*Enterprise Networks*)	7	0.57	0.53	2.43	0.015
LAN networks of small and medium enterprises (*Small Medium Business* – SMB)	7	0.14	0.38	2.53	0.011
Internet of Things (IoT)	4	0.00	0.00	2.00	0.046
LAN networks of major enterprises (*Enterprise Networks*)	7	0.71	0.49	2.46	0.014
Total	7	0.94	0.84	2.37	0.018

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

It is important to note that the engineers are very often specialized in one of two big areas of implementing R/S devices: ISP or enterprise. The examinees who work in the ISP area have more knowledge and experience about the NETCONF standard, whereas the examinees in enterprise networks have much lower knowledge (due to less frequent implementation and lesser needs of the users).

Additional observations of the examinees are that the most important advantage of the NETCONF protocol is automation, i.e. NETCONF is used mainly on larger networks in order to simplify the configuration. NETCONF is best applicable in SP and the so-called cloud environments, i.e. in SDN solutions (e.g. SD-WAN, SD-Access).

2) *The implementation of R/S network devices per vendors*

Responding to question "From which vendors do the R/S network devices come and in which percentage they are installed in the networks of your customers?" the examinees expressed their opinion in class values from 0 to 5, where zero denoted 0%, whereas five denoted 100%.

Table II. shows the obtained results ordered from the highest to the lowest average value. R/S devices that have been installed in the networks of user examinees most often come from the vendor Cisco (average value of response 4.86 (97.2%); std.dev. 0.38), whereas the rarest are the R/S devices from the vendor Arista (average response value 0.43 (8.6%), with standard deviation 0.53).

Additional observation of the examinees is the additional type of devices owned by the users, that is Nokia (former Alcatel). Based on this information the survey questionnaire has been updated for the users.

3) Reasons for implementing the NETCONF standard on R/S devices of the users

Responding to question "What are the main reasons for implementing the NETCONF standard on R/S network devices of your users?" the examinees expressed their opinion through class values from 1 to 5, where one denoted a not important reason and five as extremely important reason.

Table III. shows the obtained results ordered from the highest to the lowest average value. The total average level of reasons for implementing the NETCONF standard on R/S devices by the users is 3.79 with average deviation from the arithmetic mean 1.33 and this refers to a high level of increase (Z= 2.20; p=0.028).

During the interview one of the examinees indicated the possibility of deleting the configuration (*change management*) on R/S devices as a significant property of the NETCONF standard. After the service stops working, the configurations on the devices need to be deleted or brought to the default state. This is an enormous current operative problem of SP, and the solution is possible with SDN technology (and the NETCONF protocol).

4) Reasons of non-implementing the NETCONF standard on the users' R/S devices

Responding to question "What are the main reasons of non-implementing the NETCONF standard on R/S network devices of your users?" the examinees expressed their opinion in class values from 1 to 5, where one denoted an unimportant reason, and five an extremely important reason.

TABLE II. IMPLEMENTATION OF R/S NETWORK DEVICES PER VENDORS (INTERVIEW PARTNERS) [2]

	N	Average	Std. Dev.	Z	P*
Cisco	7	4.86	0.38	2.53	0.011
F5	7	1.86	1.57	1.63	0.102
Hewlett Packard Enterprise	7	1.43	0.98	2.23	0.026
Huawei	7	1.00	1.00	2.27	0.023
Juniper Networks	7	0.86	1.07	2.25	0.024
Extreme Networks	7	0.57	0.98	2.46	0.14
Arista	7	0.43	0.53	2.43	0.015

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

TABLE III. REASONS FOR IMPLEMENTING THE NETCONF STANDARD ON USERS' R/S DEVICES (INTERVIEW PARTNERS) [2]

	N	Average	Std. Dev.	Z	P*
NETCONF as a standard since it enables programmability (automation) which significantly reduces the human error frequency.	7	4.14	1.46	2.24	0.025
"Backup & restore" possibility offered by the NETCONF standard.	7	4.14	1.46	2.24	0.025
Possibility of simultaneous and coordinated configuration of all or several devices in the network	7	4.14	1.46	2.24	0.025
Possibility of checking configuration before implementation in the NETCONF method of device configuration.	7	3.71	1.38	2.14	0.033
Possibility of locking configuration file provided by NETCONF.	7	3.71	1.50	2.06	0.040
NETCONF as a standard way of configuration via remote access to IP address of the device since the structure of commands and data is the same for all vendors.	7	3.57	1.40	2.05	0.040
NETCONF standard uses a clear difference in the structure between configuration and operating data.	7	3.43	1.51	1.90	0.057
Possibility of partial configuration of devices using the NETCONF standard.	7	3.43	1.27	2.06	0.040
Total	7	3.79	1.33	2.20	0.028

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

Table IV. shows the obtained results ordered from the highest to the lowest average value. The main reasons for non-implementing the NETCONF standard on the users' devices are:

- ignorance of programming languages by the user (average response value 4.43; std.dev. 0.53),
- unavailability on older devices and operating systems (average value 4.29; std.dev. 0.76),
- users being used to vendors' CLI and OS (average value 4.00; std.dev. 1.00).

The least important reason of non-implementing the NETCONF standard on the users' devices is that the end users use RESTCONF as a better standard solution than NETCONF (average response value 1.71; std.dev. 0.76).

The selected additional observations of the examinees are as follows:

TABLE IV. REASONS FOR NON-IMPLEMENTING THE NETCONF STANDARD ON R/S DEVICES OF THE USERS (INTERVIEW PARTNERS) [2]

	N	Average	Std. Dev.	Z	P*
Users' ignorance of programming languages	7	4.43	0.53	2.43	0.015
NETCONF is not available on older devices and older OS that do not support it	7	4.29	0.76	2.40	0.016
NETCONF is not needed since the end users are used to OS and CLI of the vendor	7	4.00	1.00	2.40	0.016
Vendors do not support all the capabilities, e.g. "Candidate", "Startup", "Validate", "Rollback-on-Error" and others, so that a majority of the advantages of the standard cannot be used.	7	3.71	1.25	2.07	0.038
Fear of new technologies	7	3.71	1.11	2.22	0.026
Required additional tools (e.g. NETCONF browser or application in Python)	7	3.00	1.29	1.84	0.066
Vendor's CLI is simpler than the NETCONF standard	7	2.57	1.27	1.13	0.257
NETCONF uses XML syntax which is redundant and extensive which may tire and confuse the user.	7	2.29	1.60	0.41	0.680
End users use RESTCONF which is a better standard solution than NETCONF	7	1.71	0.76	1.00	0.317

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

- the users are used to CLI, as well as enterprise environments and do not require as much the automation.
- A very frequent reason of non-implementation is the fear from automation, although automation (if qualitatively implemented) is more secure (than when a human does the configuration) because of the reduction of human error. The fear of automation occurs because the users have a sense that they have no control over the devices.
- Very often the users are not familiar with the advantages of the NETCONF standard, i.e. the users understand CLI, but do not yet understand NETCONF.

C. Results of the survey questionnaire of R/S device users in Croatia

1) Implementation of the NETCONF standard in the network of the examinee's company

Responding to the question of the survey questionnaire: "Do you use the NETCONF standard on R/S network devices of your company?" the examinees responded with "*Yes*" or "*No*".

The great majority of companies do not use the NETCONF standard on their R/S devices (response "*No*": 65 companies; 89.04%). NETCONF is used by only a small number of examinees' companies (response "*Yes*": 8 companies; 10.96%). Table X. shows the obtained results.

Only the examinees who gave an affirmative answer to this question (who implement NETCONF in the networks of their companies), answered the following three questions while the others continued with the fifth question in this section.

2) Implementation of the NETCONF standard per types of networks

Responding to question "In which types of networks of your company do you use NETCONF on R/S network devices and in which percentage?" the examinees expressed their opinion in class values from 0 to 5, where zero denoted 0%, whereas five marked 100%. This question was answered only by the examinees who responded affirmatively to the first question in this section.

Table V. shows the obtained results ordered from the highest to the lowest average value. All the examinees do not use all types of networks in implementing the NETCONF standard. The number of examinees who do not use certain types of networks in this question can be calculated by the difference between the total number of examinees who use NETCONF (eight examinees) and the number of the obtained answers N in every row of the table. Other responses (implementation of the NETCONF standard per certain types of networks) have been processed and presented by descriptive statistics.

The companies usually use the NETCONF standard in the networks of the service providers (average response value 3.63 (72.6%); std.dev. 1.77), whereas it is least used in the Internet of Things networks (average response value 1.33 (26.6%); std.dev.1.97).

TABLE V. IMPLEMENTATION OF THE NETCONF STANDARD PER TYPES OF NETWORKS (USER QUESTIONNAIRE) [2]

	N	Average	Std. Dev.	Z	P*
Internet Service Provider – ISP	8	3.63	1.77	0.90	0.366
LAN network of major enterprises (Enterprise Networks)	7	3.29	1.98	0.61	0.607
Data Center (DC)	7	3.14	1.95	0.11	0.915
WAN network of major enterprises (Enterprise Networks)	7	3.00	1.83	0.92	0.915
LAN network of small and medium enterprises (Small Medium Business – SMB)	6	2.50	2.26	0.68	0.496
Internet of Things networks (IoT)	6	1.33	1.97	1.70	0.089

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

3) Implementation of the NETCONF standard per vendors

Responding to question "On R/S devices of which vendors do you use NETCONF and in what percentage in your company?" the examinees expressed their opinion with the class values from 0 to 5, where zero meant 0%, and five meant 100%. This question was answered only by the examinees who gave an affirmative answer to the first question in this section.

Table VI. shows the obtained results ordered from the highest to the lowest average value. All the examinees do not use the devices of all the vendors in their networks. The number of examinees who do not use the devices of certain manufacturers in this question can be calculated by the difference between the total number of examinees who use NETCONF (eight examinees) and the number of the obtained N responses in every row of the table. Other responses (implementation of the NETCONF standard for certain vendors) were processed and presented through descriptive statistics.

The largest number of R/S network devices with the NETCONF standard comes from the vendor Juniper Networks (average response value 3.83 (76.6%); std.dev. 1.47), whereas R/S devices in the fewest cases come from the vendor Huawei (average response value 1.67 (33.4%); std.dev. 1.53).

4) Reasons for implementing the NETCONF standard on R/S devices

Responding to question "What are the main reasons for implementing the NETCONF standard on R/S network devices of your company?" the examinees expressed their opinion in the class values from 1 to 5, where one meant an unimportant reason and five meant an extremely important reason. This question was answered only by the examinees who gave an affirmative answer to the first question in this section.

Table VII. shows the obtained results ordered from the highest to the lowest average value. The most important reason of implementing the NETCONF standard on network devices of the companies is the "Backup & restore" possibility (average response value 4.14; std.dev. 0.69), whereas the least important reason for the implementation is the clear difference in the structure between the configuration and operational data of NETCONF (average response value 3.25; std.dev. 1.75).

TABLE VI. IMPLEMENTATION OF THE NETCONF STANDARD PER VENDORS (USER QUESTIONNAIRE) [2]

	N	Average	Std. Dev.	Z	P*
Juniper Networks	6	3.83	1.47	1.39	0.163
Arista	4	3.75	1.26	1.13	0.257
Cisco	8	3.63	1.41	1.30	0.194
F5	4	3.25	1.50	0.38	0.705
Hewlett Packard Enterprise (HPE)	6	3.17	1.94	0.27	0.785
Extreme Networks	3	2.67	2.08	0.27	0.785
Nokia (Alcatel-Lucent)	1	2.00	N/a	1.00	0.317
Huawei	3	1.67	1.53	1.34	0.180

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

TABLE VII. REASONS FOR IMPLEMENTING THE NETCONF STANDARD ON R/S DEVICES (USER QUESTIONNAIRE) [2]

	N	Average	Std. Dev.	Z	P*
"Backup & restore" possibility of NETCONF.	7	4.14	0.69	2.41	0.016
Possibility of simultaneous and coordinated configuration of all or several devices in the network.	8	3.88	1.13	2.39	0.017
Programmability (automation) of NETCONF	7	3.86	1.21	2.23	0.026
Possibility of checking configuration before implementation.	8	3.75	0.89	2.56	0.011
NETCONF is a standard which has the same structure of commands and data for all device vendors.	8	3.63	1.41	2.21	0.027
Possibility of partial configuration of devices.	8	3.63	1.41	2.21	0.027
Possibility of locking configuration file.	8	3.63	0.74	2.57	0.010
Clear difference in the structure between NETCONF configuration and operational data.	8	3.25	1.75	1.71	0.086

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

5) Reasons of non-implementing the NETCONF standard on R/S devices

Responding to question "What are the main reasons of non-implementing the NETCONF standard on R/S network devices of your company?" the examinees expressed their opinion in class values from 1 to 5, where one denoted an unimportant reason, and five denoted an extremely important reason. This question was answered by all the examinees.

Table VIII. shows the number and percentage of the answers "Don't know" to this question. A high percentage of responses "Don't know" in this question shows that a significant number of examinees (users) is not familiar with the NETCONF standard and its implementation by individual vendors. Other responses are shown by descriptive statistics in Table IX. ordered from the highest to the lowest average value.

The main reason of non-implementing the NETCONF standard on the company devices are the following statements:

- *NETCONF is not needed because of the users being used to the vendors' OS and CLI* (average response value 3.13; std.dev. 1.33), and

TABLE VIII. REPLIES "DON'T KNOW" ABOUT REASONS FOR NON-IMPLEMENTING THE NETCONF (USER QUESTIONNAIRE) [2]

	Reply: "Don't know"		Reply: 1-5		Total	
	N	**%**	**N**	**%**	**N**	**%**
Vendor's CLI is simpler than the NETCONF standard	*25*	*34.72*	47	65.28	72	100
NETCONF is not needed because of being used to OS and CLI of the vendor	*18*	*25*	54	75	72	100
NETCONF uses XML syntax which is redundant and extensive which may be tiring and confusing.	*34*	*47.22*	38	52.78	72	100
RESTCONF is used, which is a better standard solution than NETCONF	*33*	*45.83*	39	54.17	72	100
NETCONF is not available on older devices and older OS which do not support it	*26*	*36.11*	46	63.89	72	100
Insufficient knowledge of programming languages.	*21*	*29.17*	51	70.83	72	100
Vendors do not support all the capabilities, e.g. "Candidate", "Startup", "Validate", "Rollback-on-Error" and others, so that the majority of the advantages of the standard cannot be used.	*34*	*47.22*	38	52.78	72	100
Additional tools are necessary (e.g. NETCONF browser or the application in Python)	*30*	*41.67*	42	58.33	72	100
Fear of new technologies.	*20*	*27.78*	52	72.22	72	100

TABLE IX. REASONS FOR NON-IMPLEMENTING THE NETCONF STANDARD ON R/S DEVICES (USER QUESTIONNAIRE) [2]

	N	Average	Std. Dev.	Z	P*
NETCONF is not needed because of being used to OS and CLI of the vendor	54	3.13	1.33	4.84	<0.001
NETCONF is not available on older devices and older OS which do not support it.	46	3.11	1.45	2.27	<0.001
Insufficient knowledge of programming languages.	51	2.82	1.21	4.09	<0.001
Vendor's CLI is simpler than the NETCONF standard	47	2.72	1.23	3.53	<0.001
The vendors do not support all the capabilities, e.g. "Candidate", "Startup", "Validate", "Rollback-on-Error" and others, so that the majority of advantages of the standard cannot be used.	38	2.68	1.32	2.87	0.004
Additional tools are necessary (e.g. NETCONF browser or application in Python)	42	2.57	1.36	2.56	0.010
NETCONF uses XML syntax which is redundant and extensive and this may be tiring or confusing.	38	2.45	1.13	2.45	0.014
Fear of new technologies.	52	2.23	1.25	1.46	0.143
RESTCONF is used, which is a better standard solution than NETCONF.	39	1.77	0.93	1.80	0.072

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

- *NETCONF is not available on older devices and older OS that do not support it* (average response value 3.11; std.dev. 1.45).

The least important reason for non-implementing the NETCONF standard on the users' devices is that RESTCONF is used, which is a better standard solution than NETCONF (average response value 1.77; std.dev. 0.93).

D. Research results: Testing of hypothesis

As part of the research the hypothesis was tested which says: "the users of network equipment do not implement the NETCONF standard on their network devices".

This hypothesis is tested based on the users' opinions collected through the survey questionnaire. The question was: "Do you use the NETCONF standard on R/S (*Router/Switch*) network devices of your company?". Table X. shows the responses of the users to this question.

The hypothesis is tested by the use of Chi square (χ^2) test which checks the difference in the presence of the responses regarding an equal distribution (50% of answers *No* and 50% of answers *Yes*). In case that the empirical p value is less than 5% (0.05), it is determined that there prevails the answer in the population of users that has a larger percentage of presence among the users. If the empirical p value is greater than 0.05 the conclusion is made that the distribution of the answers *No* and *Yes* in the population does not deviate statistically significantly from the ratio 50% to 50%.

Table XI. shows the Chi square test (χ^2) of this hypothesis. It can be determined from the table that the number of examinees who answered negatively is by 57

TABLE X. IMPLEMENTATION OF THE NETCONF STANDARD BY THE USERS (USER QUESTIONNAIRE) [2]

	N	**%**
YES	8	10.96
NO	65	89.04
Total	73	100.0

TABLE XI. TESTING OF HYPOTHESIS [2]

Do you use the NETCONF standard on R/S network devices of your company?			χ^2	p*
No	65.00	89.04%	44,51	<0,001
Yes	8.00	10.96%		
Total	73	100.00%		

* χ^2 TEST

larger responding to question whether the NETCONF standard is used on R/S network devices in their companies.

The testing showed that there is a difference in the presence of the enterprises that use and those that do not use the NETCONF standard on R/S network devices; (χ^2=44.51; empirical p value <0.001), i.e. the agreement with the provided statement will prevail in the population.

Following the results of the performed test it may be concluded that the hypothesis which assumes *that the users of the network equipment do not implement the NETCONF standard on their network devices* **is accepted as true**.

III. CONCLUSION

This paper shows the level of the implementation of the NETCONF standard by the major users of network equipment in Croatia and gives proposals that will enable an increase in the usage of the NETCONF standard with all its advantages.

Following the results of the scientific research presented in this paper it can be concluded that the hypothesis which assumes *that the users of the network equipment do not implement the NETCONF standard on their network devices* is accepted as true.

The NETCONF standard is an extremely robust and high-quality conceived and developed protocol [2-3]. The basic properties of the NETCONF protocol identified by the examinees in this research as the main reasons for its implementation are as follows:

- "Backup & restore" possibility of NETCONF.
- Possibility of simultaneous and coordinated configuration of all or several devices in the network.
- Programmability (automation) using NETCONF.

NETCONF is a transaction-based protocol, which offers a great advantage over the CLI [2-4]. In other words, NETCONF is a stateful protocol and brings the system into a defined (desired) state or returns it to the previous state. Therefore, this prevents the occurrence of partly done configuration of the R/S devices. NETCONF allows automation thus simplifying the configuration of R/S devices on major and more complex networks [2-3].

The users' implementation of the NETCONF standard is currently very low. The users implement the NETCONF standard very rarely on their devices of different vendors. The results of the survey questionnaire show that the majority of companies (of a total of 73 companies of the examinees) do not use the NETCONF standard on their R/S devices (65 companies; 89.04%). NETCONF is used only by eight companies of the examinees (10.96%). Therefore, there is a significant possibility of extending the implementation of the NETCONF standard among the users.

According to the examinees' responses the causes of non-implementation of the NETCONF standard are as follows:

- NETCONF is not needed because the end users are used to vendors' OS and CLI.
- NETCONF is not available for older devices and older OS that do not support it.
- insufficient knowledge of the programming languages by the users.

It may be observed from the abovementioned causes that the majority of them regarding non-implementation is related to the habits of the network administrators and the time necessary to accept changes, so that they need not be considered as the actual drawbacks of the NETCONF standard.

The users have expressed their opinion about the ways of increasing the implementation of the NETCONF standard and they think that the vendors should (a) increase the promotion (education and information) about the NETCONF standard and related technologies (automation, SDN networks, programming) and (b) implement the NETCONF standard on all types of R/S devices (especially on SMB devices and ISP-managed services).

The current status of the NETCONF RFC 6410 standard has been "proposed", and in order to pass to the highest level (status of "Internet standard", as the one of SNMP) there should be a wide usage of this standard in practice by the users, i.e. RFC should be accepted in everyday implementation [3,5]. The users (network administrators, system engineers, network system architects, and others) determine by the selection and installation of a protocol in their networks the significance and maturity of RFC standard, i.e. of this protocol. There is a lot of room for extending the implementation of the NETCONF standard among the users, but also a threat in case of no extension of the implementation, that there is a possibility that instead of NETCONF some other solution would be implemented.

REFERENCES

[1] IETF RFC 7426 Software-Defined Networking (SDN): Layers and Architecture Terminology (2015) https://tools.ietf.org/html/rfc7426 (21.1.2020).

[2] Valencic D.: Increasing the availability of network devices by implementing the command configuration standard, doctoral thesis, FFZG, University of Zagreb, Croatia, 2020

[3] Valencic D., Mateljan V.: Implementation of NETCONF Protocol, Mipro 2019.

[4] IETF RFC 6241 Network Configuration Protocol (NETCONF) (2011) https://tools.ietf.org/html/rfc6241 (21.1.2020)

[5] IETF RFC 6410 Reducing the Standards Track to Two Maturity Levels (2011) https://tools.ietf.org/html/rfc6410 (21.1.2020).

Vendors' Implementation of NETCONF Standard on Routers and Switches

D. Valencic
University of Applied Sciences Velika Gorica, Velika Gorica, Croatia
davorin_valencic@yahoo.com

Abstract - **Internet Engineering Task Force (IETF) standard organization created in 2006 the NETCONF (Network Configuration Protocol) in order to define the standard solution for command-oriented CLI (Command Line Interface) user interface for the configuration of network devices (routers and switches) of different vendors. Using the NETCONF standard should enable standardized, simpler, programmatic and more efficient configuration of R/S (router/switch) devices. Today, more than ten years after the creation of the NETCONF standard there is the question about the actual acceptance of the NETCONF standard by vendors on their R/S devices and about the actual maturity of the NETCONF standard. The scope of this paper is to show results of the scientific research (qualitative with interviews) made with major worldwide vendors of network devices (routers and switches): on which type of devices the NETCONF is implemented, what are the reasons for NETCONF (non)implementation, the availability of the NETCONF training and documentation, etc. Also, it is shown (quantitative research with questionnaire) how major customers in Croatia see the vendors' implementation of the NETCONF standard and how it could be improved.**

Keywords - **NETCONF standard, computer networks, router, switch, management, implementation, vendor, software-defined networking, SDN**

I. INTRODUCTION

NETCONF (Network Configuration Protocol) is a standard (created by IETF in 2006 in Request for Comments (RFC) 4741) for the configuration of network devices that could solve the drawbacks of the non-standard CLI (Command Line Interface) and SNMP (Simple Network Management Protocol) configuration methods.

After more than ten years since the emergence of the NETCONF standard a question has been asked as to what the real acceptance and application of this standard is by vendors in their network devices [1-2].

That is the topic of this paper that is based on research published in the doctoral thesis *Increasing the availability of network devices by implementing the command configuration standard* (author D. Valenčić) [2].

University of Applied Sciences Velika Gorica, Velika Gorica, Croatia

As part of research the hypothesis has been tested which says: "The vendors do not implement the NETCONF standard on their devices" that is the main topic of this article.

II. RESULTS OF RESEARCH: VENDORS' IMPLEMENTATION OF NETCONF STANDARD ON ROUTERS AND SWITCHES

During the performed research the following scientific methods have been used [2]:

1. Semi-structured interview with the selected world vendors (qualitative research);
2. Semi-structured interview with the selected partners of vendors in Croatia (qualitative research); and
3. Structured questionnaire with users (partners and end users) in Croatia (quantitative research).

Qualitative research (interviews with vendors and partners) served also as a preparation for quantitative research (questionnaire).

A. Data processing method

By using the method of tabular presentation, this section presents the structure of responses to:

- interview questions from the vendor's employees,
- interview questions from the employees of the vendor's partners, and
- survey questions from the users in Croatia.

Using the method of descriptive statistics the mean values (arithmetic mean) are presented, as well as the dispersion around the mean values (standard deviation as the average deviation from the arithmetic mean).

Testing of differences in relation to the theoretical value is performed using the Wilcoxon test for one independent sample. The justification for the use of the Wilcoxon test results from:

- small sample sizes, and the rank of characteristics of numerical values (for interviews);

- rank of characteristics of numerical values that do not have a normal distribution (for the questionnaire),

due to which the conditions for conducting the parametric test versions have not been met.

The analysis was performed in the statistical software SPSS 25 and the conclusions have been made at border signification of 5%.

In [2] are presented

- by means of descriptive statistics and mean value the median and mod as well as the inter-quartal range (IQR) as an indicator of dispersion around the mean value of the median,
- as well as a more detailed processing of examinees' responses,

that could not be presented in this paper because of the space limitation.

B. Results of interviews with the R/S device vendors

1) Implementation of the NETCONF standard on R/S devices of the vendors

All the interviewed vendors' employees have confirmed that their companies implement the NETCONF standard defined in RFC standards, on their devices.

2) Implementation of the NETCONF standard on R/S devices per types of networks

Responding to question "On which types of networks has your company implemented NETCONF on R/S devices and in which percentage?" the examinees expressed their opinion through class values from 0 to 5, zero denoting 0% and five denoting 100%.

Table I. presents the obtained results ordered from the highest to the lowest average value. One may observe that the total level of the NETCONF standard implementation on R/S devices by the companies marks a high average value of 3.75 (75%) with average deviation from the arithmetic mean of 1.50.

TABLE I. IMPLEMENTATION OF NETCONF STANDARD ON R/S DEVICES PER TYPES OF NETWORKS (INTERVIEW VENDORS) [2]

	N	Average	Std.Dev.	Z	P*
Internet Service Provider	4	4.75	0.50	1.89	0.059
Data Center	5	4.60	0.89	2.00	0.046
WAN Enterprise Networks	5	4.20	1.30	1.66	0.098
LAN Small Medium Business (SMB)	4	3.50	2.38	0.37	0.713
LAN Enterprise Networks	5	3.20	2.05	0.14	0.888
Internet of Things	4	2.25	2.06	0.74	0.461
Total	5	3.75	1.50	0.94	0.345

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

NETCONF is part of the Juniper operating system (JUNOS) and has therefore been implemented on all R/S products of this manufacturer.

3) Reasons for implementing the NETCONF standard on R/S devices

Responding to question "What are the main reasons for implementing the NETCONF standard on R/S network devices of your company?" the examinees expressed their opinion by class values from 1 to 5, one meaning not important, whereas five meaning of extreme importance.

Table II. shows the obtained results ordered from the highest to the lowest average value. All the mentioned properties have obtained extremely high values of importance.

4) Reasons for non-implementation of the NETCONF standard on R/S devices

TABLE II. REASONS FOR IMPLEMENTING NETCONF STANDARD ON R/S DEVICES (INTERVIEW VENDORS) [2]

	N	Average	Std. Dev.	Z	P*
Possibility of partial configuration* of a network device that is offered by NETCONF. (*NETCONF achieves by using XML language and "XML subtree filtering" function)	5	5.00	0.00	2,24	0.025
Possibility of validating a configuration* prior to implementation that is offered by NETCONF. (*NETCONF achieves validation of the configuration datastore by using the capability <validate>)	5	5.00	0.00	2,24	0.025
NETCONF as a standard-based solution for configuration management over remote IP access to R/S device because the structure of commands and data is the same for all vendors.	5	4.80	0.45	2,21	0.034
NETCONF as a standard-based solution for configuration management because of providing the automated and programmable network operations which replaces the process of manual configuration and so significantly decreases the possibility of human error.	5	4.80	0.45	2,21	0.034
NETCONF because of making a clear distinction between configuration and operational data.	5	4.80	0.45	2,21	0.034
"Backup & Restore"* possibility offered by NETCONF. (*NETCONF achieves by (1) manipulation of different configuration datastores (<running>, <candidate>, <startup> and backup) and (2) by using the "Rollback-on-Error" capability.)	5	4.80	0.45	2,21	0.034
Possibility of network-wide configuration* offered by NETCONF. (*NETCONF achieves by using different configuration datastores (<running>, <candidate>, <startup> and backup) and with operations "lock" and "unlock".)	5	4.80	0.45	2,21	0.034
Possibility of locking a configuration datastore* offered by NETCONF. (*NETCONF operations "lock" and "unlock" protect from simultaneous change of configuration datastore from several sources (for example by access of other user with NETCONF session, CLI or SNMP protocol).	5	4.60	0.89	5,00	5.00

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

Responding to question "What are the main reasons of non-implementing the NETCONF standard on R/S network devices of your company?" the examinees expressed their opinion with the class values from 1 to 5, one denoting not important, whereas five denoting extremely important.

Table III. presents the obtained results ordered from the highest to the lowest average value.

It is interesting that the selected main reason for non-implementing was: "NETCONF has not been accepted by the users and it is rarely used by the users".

C. Results of interviews with partners of R/S device vendors

1) The implementation of the NETCONF standard on the users' R/S devices

Responding to question "On R/S (router/switch) devices of which manufacturers do you use NETCONF and in what percentage?" the examinees expressed their opinions about the class values from 0 to 5, where zero marked 0% and five meant 100%.

Table IV. shows the obtained results ordered from the highest to the lowest average value. NETCONF is usually used by the users on R/S devices of the vendor Cisco (average value of the responses 1.00, usage in 20% of users; std.dev. 0.58) and devices of the vendor Huawei (average value of responses 1.00 (20%); std.dev. 1.73). The total level of examinees' responses marked the average value of 0.61 with average deviation from the arithmetic mean 0.51, which shows a very low average implementation (12.2%) of the NETCONF standard on R/S devices of vendors.

D. Results of the survey questionnaire of R/S device users in Croatia

TABLE III. REASONS FOR NON-IMPLEMENTATION OF NETCONF STANDARD ON R/S DEVICES (INTERVIEW VENDORS) [2]

	N	Average	Std. Dev.	Z	P*
NETCONF has not been accepted by users and it is rarely used by users.	5	3.20	1.48	1.51	0.131
NETCONF is not necessary as our clients are used to our OS and CLI.	5	2.80	1.64	1.07	0.285
NETCONF uses XML syntax that is redundant and lengthy which could fatigue and confuse the user.	5	2.20	1.30	0.38	0.705
Our CLI is more user-friendly than NETCONF standard.	5	1.80	1.30	0.38	0.725
NETCONF does not yield sufficient benefits to be implemented.	5	1.60	1.34	0.71	0.480
We use RESTCONF which is a better standard solution than NETCONF.	5	1.40	0.55	1.73	0.083

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

TABLE IV. IMPLEMENTATION OF NETCONF STANDARD ON R/S DEVICES OF USERS (INTERVIEW PARTNERS) [2]

	N	Average	Std.Dev.	Z	P*
Cisco	7	1.00	0.58	2.45	0.014
Huawei	5	1.00	1.73	1.79	0.074
F5	4	0.75	0.96	1.84	0.066
Juniper Networks	4	0.75	0.96	1.84	0.066
Arista	3	0.33	0.58	1.63	0.102
Hewlett Packard Enterprise	6	0.17	0.41	2.33	0.020
Extreme Networks	2	0.00	0.00	1.41	0.157
Total	7	0.61	0.51	2.37	0.018

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

1) Implementation of the NETCONF standard on R/S devices of different vendors

Responding to question "According to your opinion in what percentage are various vendors implementing the NETCONF standard on their R/S devices?" the examinees expressed the opinions by class values from 0 to 5, where zero represented 0%, whereas five marked 100%.

Table V. presents the number and percentage of responses "Don't know" to this question. High percentage of the answer "Don't know" to this question shows that a significant number of examinees (users) are not familiar with the NETCONF standard and its implementation by certain vendors. Other responses are presented by descriptive statistics in Table VI. ordered from the highest to the lowest average value.

According to the opinions of examinees the vendor Cisco implements the NETCONF standard on their R/S devices in the highest percentage (46.2%) (average value of responses 2.31; std.dev.1.78), whereas the lowest percentage (24.4%) is achieved by the vendor Extreme Networks implementing the NETCONF standard on their R/S devices (average value of responses 1.22; std.dev.1.59).

TABLE V. REPLIES "DON'T KNOW" ABOUT NETCONF IMPLEMENTATION OF VARIOUS VENDORS (USER QUESTIONNAIRE) [2]

	Total number of replies	Replies "Don't know"		Number of replies for analysis
	N	**%**	**N**	**N**
Arista	*64*	67.19	*43*	21
Cisco	*72*	37.50	*27*	45
Extreme Networks	*65*	64.62	*42*	23
F5	*65*	60.00	*39*	26
Hewlett Packard Enterprise	*65*	61.54	*40*	25
Huawei	*65*	64.62	*42*	23
Juniper Networks	*66*	57.58	*38*	28
Nokia (Alcatel-Lucent)	*64*	70.31	*45*	19

TABLE VI. IMPLEMENTATION OF NETCONF STANDARD ON R/S DEVICES OF VARIOUS VENDORS (USER QUESTIONNAIRE) [2]

	N	Average	Std.Dev.	Z	P*
Cisco	45	2.31	1.78	2.49	0.013
Juniper Networks	28	2.07	1.80	2.61	0.009
Huawei	23	1.74	1.74	2.99	0.003
Arista	21	1.48	1.83	3.08	0.002
F5	26	1.42	1.55	3.68	<0.001
Hewlett Packard Enterprise (HPE)	25	1.36	1.50	3.77	<0.001
Extreme Networks	23	1.22	1.59	3.63	<0.001
Nokia (Alcatel-Lucent)	19	1.05	1.22	3.64	<0.001

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

2) Supporting the implementation of the NETCONF standard by the vendors

Responding to question "Do the vendors support a larger implementation of the NETCONF standard at the users and in what way?" the examinees expressed agreement with the values from 1 to 5, where one marked disagreement and five complete agreement with the statement.

Table VII. shows the obtained results ordered from the highest to the lowest average values.

The number of responses "Don't know" to this question can be calculated by the difference between the total number of examinees 73 and the number of obtained answers N in every table row, and on the average it amounts to 50.68% (37 out of 73 examinees). A high number of answers "Don't know" to this question shows that a significant number of examinees (users) are not familiar with the NETCONF standard and its support by the vendors.

The total grade of the examinees on the support of implementing the NETCONF standard by the vendors is rather low and amounts to an average value of 2.55 with average deviation from the arithmetic mean of 0.90.

3) Ways of increasing the implementation of the NETCONF standard by the vendors

Responding to question "In your opinion, in what way could the implementation of the NETCONF standard be increased by the vendors?" the examinees expressed their opinion using class values from 1 to 5, where one marked an insignificant reason, and five an extremely important reason.

The number of responses "Don't know" to this question can be calculated through the difference between the total number of examinees 73 and the number of the obtained responses N in every table row.

Table VIII. presents the obtained results ordered from the highest to the lowest average value.

E. Research results: Testing of hypothesis

As part of research the hypothesis has been tested which says: "The vendors do not implement the NETCONF standard on their devices".

This hypothesis is tested based on the users' opinion collected by means of the survey questionnaire. The question was: "Do you agree with the following statement: The vendors implement the NETCONF standard on their devices?" Table IX. shows the responses of the users to this question.

The examinees' responses have been grouped in two groups of agreement (*No* and *Yes*), whereas the responses of examinees who did not have any opinion were left out of the testing (the responses *Don't know* or *Neither agree nor disagree*).

The hypothesis is tested using Chi square (χ^2) test which checks the difference in the share of responses in relation to equal re-distribution (50% responses *No* and 50% responses *Yes*). In case the empirical p value is lower than 5% (0.05), it is determined that the response in the user population that has higher percentage of share among the examinees prevails. If the empirical p value is greater than 0.05 it is concluded that the distribution of responses *No* and *Yes* in the population does not deviate statistically significantly from the 50% to 50% ratio.

TABLE VII. SUPPORT OF IMPLEMENTING NETCONF STANDARD BY THE VENDOR (USER QUESTIONNAIRE) [2]

	N	Average	Std. Dev.	Z	P*
There is detailed documentation of the vendor which explains the operation and implementation of the NETCONF standard on network devices: instructions (*manuals*), recommendations for use (*white papers*), etc.	36	2.58	0.87	2.62	0.00
There are vendors' courses for self-learning (CBT, Webinar, etc.) about the NETCONF protocol.	35	2.57	0.81	2.78	0.00
There are organized instructor-led courses by vendors about the NETCONF protocol.	33	2.52	1.06	2.40	0.01
There are certification exams (and accompanying courses) of vendors that train in detail the users about the use and implementation of the NETCONF standard.	31	2.39	0.84	3.23	0.00
The vendors organize presentations and workshops on the advantages of implementation methods of the NETCONF protocol.	32	2.38	1.13	2.54	0.01
Total	36	2.55	0.90	2.78	0.00

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

TABLE VIII. METHODS OF INCREASING THE IMPLEMENTATION OF NETCONF STANDARD BY VENDORS (USER QUESTIONNAIRE) [2]

	N	Average	Std. Dev.	Z	P*
Increase education and informing about automation and programming of computer networks (*Automation tools and DevOps*)	57	3.86	0.90	6.34	<0.001
Increase education and informing about Software Defined Networks – SDN	57	3.81	0.90	6.43	<0.001
Increase informing about practical implementation of the NETCONF standard – e.g. publication of recommendations for use (white papers), etc.	56	3.80	0.92	6.36	<0.001
Increase informing about implementation of the NETCONF standard to INCREASE AVAILABILITY of network systems and R/S devices – e.g. publication of recommendations for use (white papers), etc.	54	3.78	0.86	6.27	<0.001
Implement NETCONF on all types of R/S devices – e.g. from small SMB devices to large ISP devices.	55	3.75	0.91	6.31	<0.001
Increase informing about the advantages of NETCONF standard – e.g. through organized vendors' courses, workshops, CBT, Webinar, etc.	56	3.70	0.97	6.18	<0.001
Increase education about programming and programming languages	58	3.69	1.03	6.39	<0.001

* WILCOXON TEST FOR ONE INDEPENDENT SAMPLE

TABLE IX. IMPLEMENTATION OF NETCONF STANDARD BY THE VENDORS (USER QUESTIONNAIRE) [2]

	Don't know		-2 Extremely NOT agree		-1 DO NOT agree		0 Neither agree nor disagree		1 Agree		2 Extremely agree		Total	
	%	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**
The vendors implement the NETCONF standard on their devices.	12.33	*9*	1.37	*1*	10.96	*8*	39.73	*29*	28.77	*21*	6.85	*5*	100	*73*

Table X. shows Chi square test (χ^2) of this hypothesis.

From the table it can be determined that the number of examinees who expressed agreement with the statement *that the vendors implement the NETCONF standard on their devices* is by 17 larger than the number of examinees who did not agree with the statement. The examinees who did not have an expressed opinion were left out of the testing. The test showed that there is a difference in the presence of agreement with the offered statement in relation to the number of examinees who did not express agreement (χ^2=8,26; empirical p value <0.001), i.e. in the population the agreement with the offered statement will prevail.

Following the results of the test, it can be concluded that the hypothesis which assumes *that the vendors do not implement the NETCONF standard on their devices* **is rejected as untrue**.

The explanation of the obtained result is as follows: the hypothesis was set two to three years before, and in the last few years there have been significant changes in the field of computer network management and NETCONF (as part of software-defined networking) is increasingly implemented on R/S devices by the vendors.

TABLE X. TESTING OF HYPOTHESIS [2]

Vendors implement the NETCONF standard on their devices.			χ^2	p*
No	9	25.71%	8.26	<0.001
Yes	26	74.29%		
Total	35	100.00%		

* χ^2 TEST

III. CONCLUSION

The NETCONF standard has significant advantages in relation to previous non-standard methods of command configuration as presented in [1-2].

This paper presents the degree of implementing the NETCONF standard by the vendors, and proposes also the recommendations to increase the implementation.

The majority of vendors do not support the NETCONF standard on all their devices and OS versions. The support of a device for NETCONF depends on the parameters of this device (type of device, model of device, type of operating system, version of operating system, groups of supported functionalities of the operating system). Because of frequent changes of these parameters, it is very complex to determine which network devices support NETCONF and in which version of the software, and the results change almost daily.

The extent of implementing the NETCONF standard by the R/S device vendors in this paper is determined through studying the information obtained by the vendors (through interviews) and users (through interviews and survey questionnaire).

The NETCONF standard had been implemented until now only on R/S devices of some types of networks and therefore, there is a large space for expanding the implementation. The vendors have implemented the NETCONF standard on the majority of their devices for internet service provider (ISP), data centre (DC) and WAN (wide area networks) networks. There is a significant possibility of increasing the implementation for other devices, and especially for SMB (small medium business) devices (e.g. for ISP-managed services).

The vendors believe that there is sufficient and available documentation and education on the NETCONF standard, and the partners and the users agree (but to a much lesser extent).

The users have expressed their opinion about the ways of increasing the implementation of the NETCONF standard and consider that the vendors should (a) increase promotion (education and information) about the NETCONF standard and related technologies (automation, software defined networking, programming) and (b) implement the NETCONF standard on all types of R/S devices (especially SMB devices and ISP-controlled services).

It should be also noted that at the moment of the request for the NETCONF standard (RFC 3535 in June 2002) and at the moment of its inception (RFC 4741 in December 2006), the virtualization of computer devices and cloud computing were at their very infancy. Although conceived initially as replacement for non-standard CLI approach, today NETCONF is mainly used for automation and programmability of computer network management and as a component of software defined networking (SDN) solution [3]. The area of automation and programming of computer networks is extremely active and has been developing fast over the last year. So, e.g. Cisco established a DevNet team that provides detailed free information and training about SDN and other related technologies, and also the equipment is available on which laboratory testing can be done [4]. Alternative and similar solutions are being developed, such as NETCONF (like e.g. RESTCONF standard defined in [5] and gRPC remote procedure call framework defined in [6]) and their implementation by the R/S device vendors can be the topic of future research.

Following the results of the scientific research made in [2] and presented in this paper it can be concluded that the hypothesis which assumes *that the vendors do not implement the NETCONF standard on their devices* is rejected as untrue. The explanation of the obtained result is that in the last period have been significant changes in the field of computer network management and NETCONF (as part of software-defined networking) is increasingly implemented on R/S devices by the vendors.

REFERENCES

[1] Valencic D., Mateljan V.: Implementation of NETCONF Protocol, Mipro 2019.

[2] Valencic D.: Increasing the availability of network devices by implementing the command configuration standard, doctoral thesis, FFZG, University of Zagreb, Croatia 2020

[3] IETF: RFC 7426 Software-Defined Networking (SDN): Layers and Architecture Terminology (2015) https://tools.ietf.org/html/rfc7426 (21.1.2020).

[4] Cisco DevNet: Cisco APIs, SDKs, Sandbox and Community for Cisco Developers:, https://developer.cisco.com/ (20.1.2020)

[5] IETF RFC 8040 RESTCONF Protocol (2017) https://tools.ietf.org/html/rfc8040 (15.2.2020)

[6] GRPC, https://www.grpc.io (15.2.2020)

Dohvat novih ključeva za mobilnu aplikaciju na bankomatu prema ITIL najboljoj praksi

Denis Muratović, Renato Barišić
Visoko učilište Algebra, Zagreb, Republika Hrvatska
denis.muratovic@racunarstvo.hr, renato.barisic@racunarstvo.hr

***Sažetak* - Dohvat ključeva za mobilno bankarstvo na bankomatu spoj je tradicionalnog kanala i moderniziranja tehnologije IT sektora. Kako je bankarski sustav pod raznim regulatornim obvezama, zbog osjetljivosti podataka koje pruža, ne postoji širok odabir novog načina za dohvat ključeva. Obzirom da usluga postoji u internetskom bankarstvu, postoji ekvivalent tokenu koji postojeći korisnici koriste - kartica tekućeg računa. Kartica je, kao i token, zaštićena PIN-om što korisnicima jamči sigurnost u pružanju usluge. Dohvat ključeva za mobilno bankarstvo na bankomatu oslanja se na ITIL najbolju praksu. Velika prednost ITIL-a je što nije standard te je jednostavan za implementaciju. Rad govori o implementaciji nove usluge, suradnji više bankarskih sektora koji rezultiraju kontinuiranim poboljšanjem usluge. Upravo kontinuirano poboljšanje usluge prati korisničko iskustvo i modeliranje prema korisnikovim potrebama.**

Ključne riječi – ključevi, bankomat, banka, ITIL, kartica

I. UVOD

Ključevi za aktivaciju mobilnog bankarstva sastoje se od dva važna ključa: identifikacijskog i aktivacijskog. Identifikacijski, odnosno aktivacijski ključ, skup je od osam brojeva kojima se aplikacija mobilnog bankarstva može aktivirati. Ključevi su potrebni za aktivaciju novoinstalirane aplikacije mobilnog bankarstva na mobilni uređaj.

Za projekt dohvata ključeva za mobilno bankarstvo na bankomatu potrebna je velika količina znanja, iskustva i organizacije. S obzirom na pedantnost sustava donesena je odluka za korištenje ITIL najbolje prakse. ITIL ima sve važne preduvjete za implementaciju nove usluge: stabilan je, ne traži ispunjenje svakog zahtjeva, nije standard, međunarodno je prihvaćen i preporučuje se kod uvođenja novih usluga. Skup je to najboljih praksi iz poslovanja, nije usmjeren na pojedinačno područje pa tako postaje najvrjedniji alat za svaku organizaciju.

Dohvat ključeva za mobilno bankarstvo na bankomatu susret je tradicionalnog kanala isporuke, bankomata i trenda digitalizacije usluga svakog područja poslovanja. Može se reći da se savršeno spajaju svi aspekti potrebni za uvođenje nove usluge. Mobilno bankarstvo relativno je mlada, ali napredna usluga koja broji preko pola milijuna korisnika. Oni često nadograđuju i brišu aplikacije te im je vrlo važna brza i jednostavna reaktivacija usluge. Iako postoji digitalna alternativa za dostavu ključeva, pokazala se zastarjelom za većinu korisnika.

II. DIGITALNO BANKARSTVO

Digitalno bankarstvo podrazumijeva dvije vrste usluga: elektroničko bankarstvo putem računala i mobilno bankarstvo putem mobitela. Mobilno bankarstvo novija je vrsta bankarstva, idealna za život u pokretu za koji su potrebna dva ispunjena uvjeta: pametni telefon i internetska veza. S obzirom na brzinu života, užurbani rast i napredak tehnologije, mobilno bankarstvo godinama bilježi porast korisnika.

Razvitak informacijsko-komunikacijske tehnologije postupno dozvoljava i razvoj digitalnog bankarstva. Elektroničko bankarstvo ili skraćeno e-bankarstvo sustav je koji se oslanja na informacijsku i komunikacijsku tehnologiju kako bi krajnjem korisniku pružio uslugu. To je sustav koji u realnom vremenu prikazuje stanje i promet transakcijskog računa, omogućava prijenos novca drugim korisnicima, omogućava praćenje limita po karticama itd. Usluga e-bankarstva u Sjedinjenim Američkim Državama pojavljuje se 1994. godine, a u Hrvatskoj 1999. godine. Prvo je uvodi Međimurska banka, a slijede ju današnji lideri na tržištu: Zagrebačka banka i Privredna banka Zagreb.

Mobilno bankarstvo u Hrvatskoj pojavilo se 2005. godine od strane Erste banke te 2008. godine od strane Zagrebačke banke. Često ga po godinama uspoređuju s prvim Appleovim pametnim telefonima iPhone ili Amazonovim uređajem za čitanje knjiga – Kindle. Prema istraživanju mrežne stranice wearesocial.com broj korisnika mobilnih uređaja u konstantnom je porastu, a 2018. godine brojio je više od 5 milijardi korisnika.

III. ITIL

ITIL (eng. Information technology infrastructure library) je skup najboljih praksi koje nude organizacijama preporuke za upravljanje IT uslugama. Osnovni cilj ITIL-a je dokumentiranje svakog koraka u izradi izmijenjene ili nove usluge koje rezultira kontinuiranim poboljšanjem usluge. S obzirom na to da nije standard, primjenjiv je u svim veličinama organizacije i nije ograničen na vrstu poslovanja kojom se ona bavi.

Sastoji se od pet faza koje su prikazane na slici 1. Strategija usluge nalazi se u sredini okružena dizajnom, tranzicijom i rukovanjem usluge. Kontinuirano poboljšanje usluge okružuje sve faze i pregledava svaku fazu za promjenu.

- Strategija usluge

- Dizajn usluge
- Tranzicija usluge
- Rukovanje uslugom
- Kontinuirano poboljšanje usluge



Slika 1. Faze ITIL-a

Strategija usluge (engl. Service strategy) prema ITIL najboljoj praksi početak je u nastanku nove ili izmijenjene usluge. Nalazi se u sredini kotača gdje je okružuju dizajn, tranzicija i rukovanje uslugom. Centralno pozicioniranje govori o podređenosti drugih ciklusa strategiji jer predstavlja srž životnog ciklusa usluge. Strategija usluge daje pružatelju usluge (engl. service provider) smjernice i prioritete kod ulaganja u nove ili izmijenjene usluge te je usmjerena na potrebe korisnika.

Dizajn usluge druga je faza životnog ciklusa prema ITIL najboljoj praksi. U fazi dizajna fokus je usmjeren na ulazne podatke iz strategije usluge kako bi se stvorili izlazni podaci za tranziciju usluge. Glavni ulaz u dizajn usluge su, prema ITIL Service Design knjizi, uvjeti za novu ili izmijenjenu uslugu. Fokusirat će se na procese, politike i dokumentaciju za novu ili izmijenjenu uslugu. Usluga dohvata ključeva na bankomatu pripada uvođenju nove usluge. Nova usluga se oslanja na postojeću infrastrukturu i transakciju dohvata ključeva putem elektroničkog bankarstva.

Tranzicija usluge (engl. Service transition) treća je faza u ITIL najboljoj praksi za modeliranje usluge. Ujedno je i zadnja faza koju klijent ne vidi. Cilj tranzicije usluge je da usluga dohvata ključeva na bankomatu bude u skladu s poslom koji je dokumentiran u prethodne dvije faze.

Cilj tranzicije usluge je:

- Planiranje i vođenje promjena prema efikasnosti i efektivnosti
- Upravljanje rizikom koji je povezan s uslugom dohvata ključeva
- Uvesti uslugu u podržano okruženje
- Izraditi bazu znanja o usluzi i komponentama

Rukovanje uslugom (engl. service operation) zadnja je faza u ITIL najboljoj praksi. Jedina je faza koju korisnik vidi, a cilj je održati zadovoljstvo i povjerenje u informacijsku tehnologiju. Održavanje zadovoljstva i povjerenja može se postići kroz efikasnu i efektivnu isporuku i podršku IT servisima. Usluga dohvata ključeva postaje dostupna korisnicima i treba minimizirati ispade na dnevnoj razini. Važno je osigurati da usluga bude isporučena samo ciljanim korisnicima. Povezivanjem sa strategijom usluge isporučuje se samo korisnicima s ugovorenom uslugom mobilnog bankarstva. Ovime se žele izbjeći neplanirani ispadi i troškovi koji nastaju identifikacijom srži problema (engl. root cause). U slučaju ispada usluge dohvata ključeva, važna je pravovremena reakcija kako bi se smanjilo trajanje i učestalost ispada. Rukovanje uslugom omogućava praćenje pristupa podacima. Pristup podacima o ugovorenim uslugama trebaju dobiti samo ovlaštene osobe.

Kako se vidi na slici 1, faza kontinuiranog poboljšanja usluge okreće se oko svih drugih faza. Kada je usluga implementirana i korištena od korisnika radi se na poboljšanju usluge. Svrha kontinuiranog poboljšanja usluge za novu uslugu dohvata ključeva na bankomatu je identifikacija i provođenje poboljšanja.

IV. STRATEGIJA USLUGE

Usluga je, prema knjizi Introduction to the ITIL Service Lifecycle [1], sredstvo za isporuku vrijednosti korisnicima kojima će se olakšati rad bez preuzimanja dodatnih troškova i rizika.

Prethodnim bilježenjem prijedloga i prigovora korisnika pružatelj usluge uvidio je nedostatak u mogućnostima dohvata novih ključeva za mobilno bankarstvo. Trenutni načini dostave ključeva su osobni dolazak u poslovnicu ili putem elektroničkog bankarstva uz uvjet da korisnik posjeduje token uređaj. Pozivanjem klijenata i ispitivanjem mnijenja zaključeno je da je potreban kanal dostave koji ne zahtijeva čekanje ili posjedovanje dodatnog uređaja. Novi kanal za dostavu ključeva je bankomat (engl. *Automated Teller Machine*, skraćeno ATM), 24-satna zona putem kojeg će korisnici uz karticu tekućeg računa jednostavno i brzo moći odraditi transakciju. Odabirom bankomata kao 24-satne zone može se procijeniti utjecaj usluge na daljnje poslovanje i razvoj. Korisnici su svi oni koji imaju ugovorenu uslugu mobilnog bankarstva, ali i novi korisnici koje će banka, kao pružatelj usluge, u budućnosti steći. Korisnici žele neovisnost u obavljanju transakcija zbog užurbanog životnog tempa i manjka vremena. Žele stalnu dostupnost usluge i jednostavnost poslovanja. Uvođenjem usluge dostave ključeva za mobilno bankarstvo na bankomatu nudi se jedinstvena usluga čija primjena još uvijek nije viđena na domaćem tržištu čime se želi povećati broj postojećih i novih korisnika. U portfelju postojećih usluga nudi se tekući račun uz korištenje kartice čime se povećavaju prilike za aktivaciju usluge. Sve usluge koje pružatelj nudi vidljive su svim korisnicima na mrežnim stranicama, a većina ih je podržana i u mobilnom bankarstvu.

Usluga dohvata ključeva na bankomatu u početku definira postojeće korisnike usluge mobilnog bankarstva kao ciljano tržište. Postojeći korisnici bili su okidač za

[1] Introduction to the ITIL Service Lifecycle, 2011. Stranica 15

uvođenje i izradu usluge. Strategijom se želi poboljšati odnos s postojećim korisnicima, a kasnije uspostaviti poslovni odnos s novim korisnicima čime se želi povećati konkurentnost i zadovoljstvo.

Uvjet korištenja usluge za postojeće korisnike je ugovoren transakcijski račun i mobilno bankarstvo. Kako bi se ispravno provodile transakcije na bankomatu potrebno je implementirati provjeru na SQL bazi kojom će transakcija biti vidljiva ovom tipu korisnika. Pružatelj usluge definira dvije vrste korisnika, nove i postojeće, kako bi strategiju mogao povezati s dizajnom usluge. Strategija definira poslovne rizike i izdvaja dva najčešća: ispad i održavanje bankomata te prigovore novih korisnika. Ispad bankomata može se dogoditi uslijed neispravnog funkcioniranja, nestanka mrežne veze ili kvara prouzročenog nemogućnošću isplate gotovine. Zbog poboljšanja korisničkog iskustva potrebno je u dizajnu napraviti transakciju kojom se omogućava nesmetani rad bezgotovinske transakcije. Bezgotovinske transakcije odnose se na sve one transakcije koje ne zahtijevaju isplatu gotovine, poput dohvata ključeva na bankomatu, kupnje bona za mobilni uređaj, pregleda stanja računa. Osim ispada bankomata prigovori novih korisnika smatraju se rizičnijima za daljnje poslovanje. Usluga u početku neće biti dostupna novim korisnicima čime se otvaraju mogućnosti za prigovore, ali i prijedloge novih korisnika.

V. DIZAJN USLUGE

Dizajn usluge doprinosi:

- Poboljšanju kvalitete servisa
- Poboljšanju dosljednosti usluge
- Poboljšanju korektivnih mjera
- Poboljšanju izvođenja usluge
- Poboljšanju informacija i donošenja odluka

Poboljšanje kvalitete servisa u dizajnu usluge može se postići dizajnom usluge koja udovoljava korisnikovim potrebama. Kod nove usluge dizajnirat će se nova transakcija čime se postiže jednostavnost u dohvatu novih ključeva. Poboljšana kvaliteta servisa odražava se u brzini odziva transakcija za dohvat ključeva. Dizajn nove usluge treba pratiti korporativnu strategiju, arhitekturu i ograničenja. Budući da se slična usluga nudi putem elektroničkog bankarstva, strategija se odlikuje u uhodanim aktivnostima. Ograničenje nove usluge je nemogućnost ugovaranja za nove korisnike, čime se prati mjera sigurnosti za korisnika. Poslovna potreba za uslugom dohvata ključeva očituje se u konstantnom praćenju raspoloživosti usluge, analizi i donošenju korektivnih mjera za ispravak.

Dizajn usluge slijedi procese iz strategije usluge čime se dizajnira kapacitet za novu uslugu i zahtjeve. Zahtjevi su temeljeni na već ponuđenoj usluzi putem elektroničkog bankarstva na temelju koje se rade analize. Usluga dohvata ključeva na bankomatu dizajnira izgled transakcije i definira mjerenje dostupnosti. S obzirom na različite vrste bankomata definirat će se pozicija transakcije. Pozicija transakcije osigurat će identičnu poziciju na svim vrstama bankomata zbog jednostavnosti snalaženja korisniku usluge. Plan oporavka definira nastavak korištenja usluge u slučaju nemogućnosti rada. Uspostavljanje metrika pomoći će u boljem donošenju odluka u poboljšanju usluge i korisnikovom iskustvu.

Razumijevanje IT servisa za uslugu dohvata ključeva na bankomatu omogućava pružatelju usluge korištenje kataloga usluga za oglašavanje usluge. Oglašavanje nove usluge važan je aspekt za njezino korištenje i generiranje zahtjeva za uslugom. Odnos između korisnika i pružatelja usluge može se uspostaviti jednostavnošću u oglašavanju usluge. Korištenje reklamnog materijala treba jasno i nedvosmisleno postaviti upute za korištenje nove usluge. Pružatelj usluge se fokusira na korisnika njegovim profiliranjem i obradom podataka. Imovina pružatelja usluge služi kao ishod profiliranja i ponude. Ponudom usluge dohvata ključeva na bankomatu korišteni su interni servisi, prigovori i pohvale korisnika. Prigovori i pohvale su mehanizam kojima je pružatelj usluge dao zahtjev za izradu nove usluge.

Mehanizmi za poboljšanje efikasnosti i efektivnosti koriste informacije iz svih izvora kojima korisnik pristupa. Tako se definiraju:

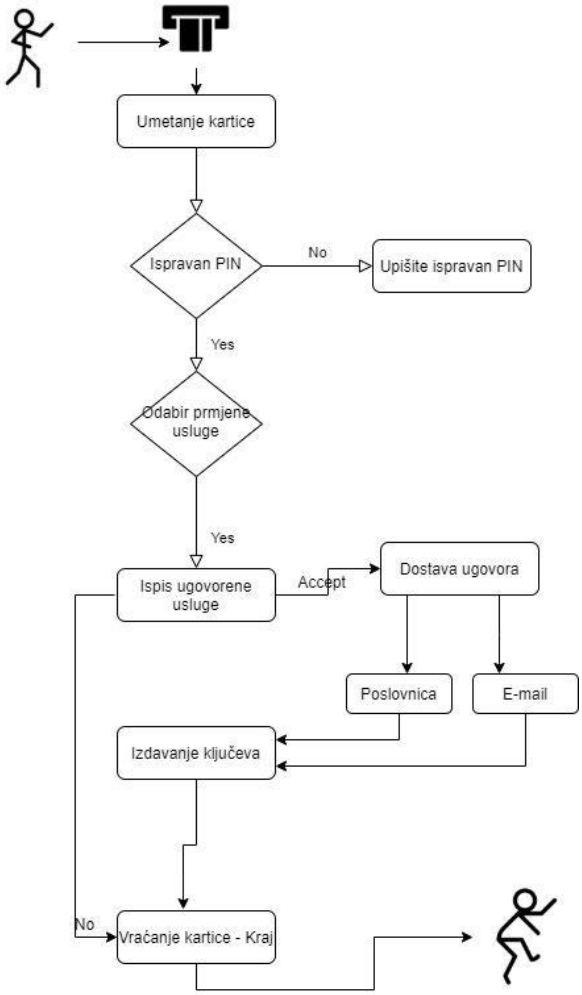
- Usluge prema korisnicima
- Prateće usluge

Usluge prema korisnicima (engl. customer facing services) su sve usluge koje korisnik koristi i koje mu olakšavaju rad. Na temelju ponuđenih usluga iz kataloga usluga korisniku se omogućava rad s novim uslugama. Svi korisnici kartica primarno koriste usluge podizanja gotovine i kupnju elektroničkih bonova. IT servisi pomažu korisniku u korištenju usluga, a nova usluga oslonit će se na iste IT servise i omogućiti dohvat novih ključeva.

Prateće usluge (engl. supporting services) su svi IT servisi koji korisniku pomažu ostvariti ishod. Ove servise korisnik obično ne vidi, ali igraju važnu ulogu u isporuci usluge. Nova usluga oslanja se na elektroničko bankarstvo, baze podataka, servere i mrežu koji omogućavaju uslugu dohvata novih ključeva.

Dizajnom usluge je zamišljeno da korisnik dolaskom na bankomat umetne karticu tekućeg računa za koji je usluga mobilnog bankarstva vezana. Umetanjem kartice i upisom PIN-a bankomat prikazuje moguće opcije od kojih je nova, Ponovna aktivacija usluge mobilnog bankarstva, smještena u donjem desnom kutu. Prije nego se transakcija prikaže vrši se provjera na bazi postoji li ugovorena usluga za korisnika. Baza podataka je ključna jer sadržava podatke o svakom korisniku (npr. stanje računa, isplata, ugovorene usluge, vezani računi, opunomoćenici, itd.) čime se dobiju podaci za validaciju same transakcije. Bankomat je umrežen i kontinuirano komunicira sa serverom dok s bazom samo u izvršavanju transakcije. Sinkronizacija bankomata i servera je na principu servisa koji je automatski pokrenut na serveru. U slučaju upisa neispravnog PIN-a bankomat korisnika usmjerava na ponovni upis ili Kraj transakcije. Odabirom promjene po usluzi prikazuje se ekran ugovorene usluge gdje podatke nije moguće promijeniti. Prihvaćanjem ispisanih podataka prikazuje se ekran za dostavu ugovora. Prilikom svake promjene po usluzi potrebno je korisniku

izdati novi ugovor o usluzi i opće uvjete za njezino ugovaranje tj. promjenu. Potpisivanje ugovora korisnik odrađuje samo prilikom prvog ugovaranje usluge dok se novi izdaje zbog promjena u uvjetima korištenja usluge i bilježenja datuma promjene po usluzi. Korisnik dobiva mogućnost za dostavu ugovora putem elektroničke pošte ili poslovnice. Ovaj korak je važan jer bez slaganja s uvjetima promjene transakciju nije moguće provesti. Zbog sigurnosti identifikacijski ključ štampa se na papiriću dok se aktivacijski zaprima putem SMS poruke na ugovoreni broj. Mobilnu aplikaciju nije moguće aktivirati samo jednim ključem te se ovim postiže dodatna sigurnost i povjerenje korisnika u proceduru za ponovno izdavanje ključeva. U slučaju da se neki od podataka ne slažu korisnik može završiti transakciju i ispraviti podatke u poslovnici. Završetkom transakcije korisniku se vraća kartica. Dijagram tijeka opisanog postupka može se vidjeti na slici 2.



Slika 2. Dijagram tijeka usluge

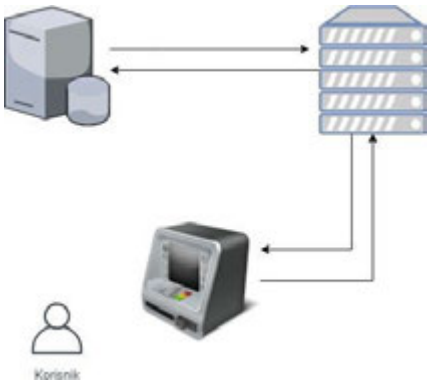
VI. TRANZICIJA USLUGE

Područje tranzicije usluge su upute i prebacivanje nove usluge u podržano okruženje. Usluga dohvata ključeva na bankomatu objedinjuje planiranje i puštanje u produkciju. Planiranje predstavlja naseljavanje prve verzije na testno okruženje. Testno okruženje uključuje implementaciju transakcije na bankomatu gdje se testiraju sve vezane transakcije za dohvat ključeva. Planiranjem i testiranjem provjerava se programsko rješenje definirano u dizajnu usluge. Nova usluga podliježe doradama, testiranju, evaluaciji i naseljavanju. Dorade uključuju promjene u performansama ili izgledu transakcije nakon čega se pristupa testiranju i evaluaciji programa. Na testnom bankomatu prolaze se svi mogući scenariji kako bi se dokazala neopterećenost infrastrukture. Nakon evaluacije dolazi naseljavanje na produkciju gdje usluga postaje dostupna za korisnike.

Vrijednost tranzicije počiva u:

- Procjeni troškova, vremena, resursa i rizika za tranziciju usluge
- Dijeljenju i korištenju resursa među različitim projektima
- Smanjenju zastoja ako se resursi koriste u drugim projektima
- Povećanju sigurnosti da će se usluga isporučiti bez utjecaja na druge usluge

Planiranje i podrška tranziciji daju pregled za planiranje tranzicije usluge i koordiniranje zahtjeva za resurse. Koordinacija zahtjeva za resurse dohvata ključeva na bankomatu uključuje infrastrukturne servise poput mreže, baze podataka, servera, radne memorije bankomata, električne energije i dr. Potrebno je s dobavljačima i timovima uspostaviti proceduru za rješavanje najčešćih problema. Najčešći problem je uvoz novih licenci za dostavu ključeva za omogućavanje neprekidnog rada korisnika. Kao strojni uređaj, bankomat se sastoji od više transakcija koje se nadziru. S time se uspostavlja nadzor bankomata u potpunosti. Koordinacija dizajna usluga uspostavit će, sa svim zainteresiranim stranama dizajn usluge i dokumentirati u paket dizajna usluge. Na slici 3 prikazana je komunikacija korisnika, bankomata i servera. Prva točka je korisnikova interakcija s bankomatom koji upit dalje prosljeđuje prema serveru i bazi podataka.



Slika 3. Komunikacija korisnika, bankomata i servera

VII. RUKOVANJE USLUGOM

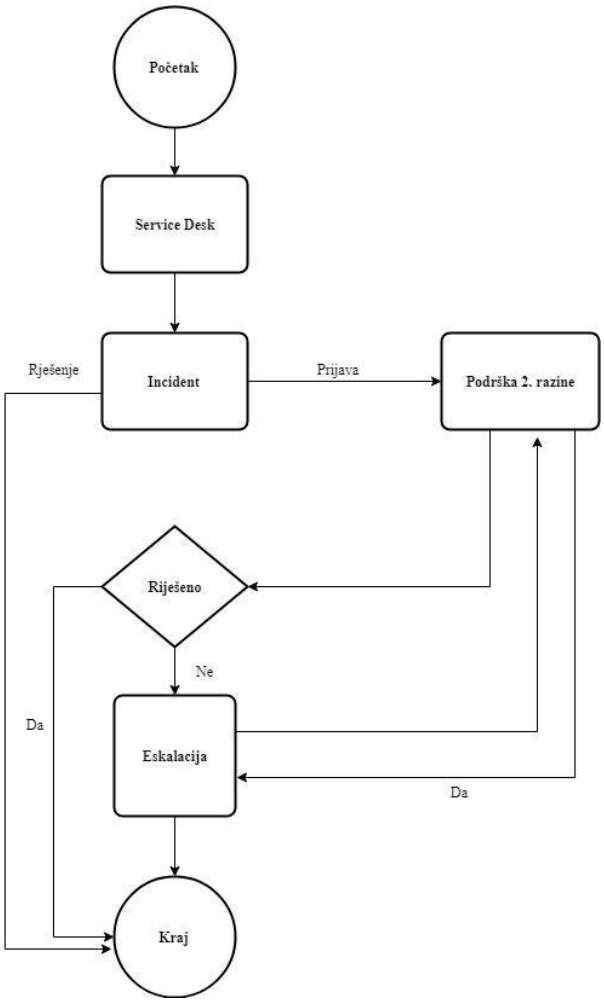
Za uslugu dohvata ključeva na bankomatu uspostavljen je 24-satni nadzor bankomata i poslužitelja baza podataka. Za ovu vrstu je već uspostavljen nadzor i detekcija događaja u upravljačkim alatima. Svrha nadzora ističe se u ranom otkrivanju incidenata, njihovom sprečavanju i nadzorom u realnom vremenu. Događaji se

kategoriziraju na temelju značaja: informativni, upozoravajući i izuzetak. Informativni događaj je obavijest o trenutnom stanju: broj prijavljenih korisnika na bankomatu ili broj korisnika koji je koristio uslugu dohvata ključeva u jednom danu. Upozorenje je obavijest da se isporuka usluge približava dogovorenom pragu isporuke (raspoloživost 99.5%). Upozorenje služi kao poduzimanje mjera prije nemogućnosti isporuke. Dolazi za nedostatak licenci, papira za ispis ključeva ili nemogućnost slanja elektroničke pošte za sklapanje ugovora. Obavijesti o izuzecima pojavljuju se kada usluga radi, ali s djelomičnim performansama. Za novu uslugu ovaj tip obavijesti se implementira kad se ne može poslati elektronička pošta ili isporučiti ključevi, s usluga je i dalje operativna.

Cilj upravljanja događajima je:

- Detekcija svih promjena stanja u isporuci usluge
- Uspostavljanje akcija i obavijesti svim zainteresiranim stranama
- Usporediti trenutno stanje i SLA ugovor

Incident je događaj koji dovodi do gubitka ili prekida rada usluge. Proces upravljanja incidentima opisuje aktivnosti koje se poduzimaju za povratak usluge u operativno stanje. Slika 4 prikazuje dijagram tijeka incidenta.



Slika 4. Tijek upravljanja incidentima

Za uslugu dohvata ključeva na bankomatu potencijalni incident se pokušava riješiti odmah u Service desku koji je točka kontakta između korisnika i pružatelja usluge gdje se rješavaju poteškoće na koje korisnici nailaze tijekom korištenja usluge.

VIII. KONTINUIRANO POBOLJŠAVANJE USLUGE

Za uspješno upravljanje kontinuiranim poboljšanjem usluge potrebno je uključiti tri izjave:

- Ne može se upravljati onime što se ne može kontrolirati
- Ne može se kontrolirati ono što se ne može mjeriti
- Ne može se mjeriti ono što nije definirano

Vođeni ovim izjavama u kontinuiranom poboljšanju usluge, usluga dohvata ključeva na bankomatu definira se kao nešto čime pružatelj usluge upravlja i kontrolira. Nova usluga dohvata ključeva na bankomatu je mjerljiva, njezina raspoloživost definirana je na mjesečnoj razini na 99,50 %. Definirana raspoloživost usluge može se poboljšati u kontinuiranom poboljšanju usluge usporedbom njezinog korištenja. Nova bi se raspoloživost mogla prema procjeni podići na 99,80 %.

Ciljevi poboljšanja usluge su:

- Revizija, analiza, prioritizacija i preporuke
- Revizija razine usluge
- Poboljšanje financija u isporuci bez narušavanja korisničkog zadovoljstva
- Shvaćanje procesa i razloga mjerenja

Kontinuirano poboljšanje usluge (engl. continual service improvement) oslanja se na registar poboljšanja usluge (engl. Continual service improvement registar; CSI register), zapise o mogućim poboljšanjima usluge. Poboljšanja mogu biti mala, srednja i velika. Važno je razumijevanje poboljšanja, kada su ona ostvariva i koje su dobiti za usluge. Vrijeme ostvarenja može varirati od brzog, sporog ili najsporijeg (engl. Long term). U poboljšanje se ide ovisno o financijskim mogućnostima, ali i dobiti.

Najvažniji aspekt kontinuiranog poboljšanja usluge dohvata ključeva na bankomatu je razumijevanje mjerenja i metrika. Kako bi se ispravno provela faza potrebno je odrediti:

- Metrike tehnologije
- Procesne metrike
- Metrike servisa

Mjerenjem dostupnosti i njegovim povećanjem definirana je formula za izračun. Formula se odnosi na praćenje prije poboljšanja usluge, a uzima u obzir ispad rada samog bankomata.

$$DOSTUPNOST = \frac{\text{Operativno vrijeme} - \text{Ispad bankomata}}{\text{Operativno vrijeme}}$$

U kontinuiranom poboljšanju za uslugu dohvata ključeva na bankomatu mjerenje se treba poboljšati. Za točniji izračun usmjerenost se prebacuje na samu transakciju dohvata ključeva koja je implementirana. S obzirom na broj bankomata pregledavat će se prijavljeni ispad transakcije dohvata ključeva na bankomatu. Nova formula za izračun vidljiva je u nastavku.

$$DOSTUPNOST = \frac{\text{Operativno vrijeme} - \text{Ispad transakcije}}{\text{Operativno vrijeme}}$$

Operativno vrijeme iz formule definira se kao vrijeme rada transakcije dohvata ključeva na jednom bankomatu, 24 sata svaki dan. Ispad transakcije i ispad bankomata odnosi se na prijavu ispada. Redovitim analizama dostupnosti svih bankomata, ali i poslužiteljskih sustava koji sudjeluju u procesiranju transakcija, doći će se do informacija o ukupnoj dostupnosti sustava bankomata.

Bez obzira na prijave ispada od strane krajnjih korisnika ili zaposlenika banke, kontinuirana automatizirana mjerenja i nadzori sustava na svim razinama (bankomat, bankomatska aplikacija, računalna i telekomunikacijska mreža, poslužitelji, baze podataka, ...) ključni su izvori podataka za proaktivna i reaktivna poboljšanja usluge i sustava.

Iako se proces kontinuiranog poboljšanja usluge stalno vrti oko svih faza, njegov kraj definira se jasnim metrikama. Nova dostupnost usluge može se povećati na 99,80 % praćenjem procesa i infrastrukture. Dostupnost usluge dohvata ključeva ne mjeri se samo ispadom transakcije i postaje kompleksnije uvođenjem novih praćenja poput infrastrukture i bankomata. Kada se sagledaju svi aspekti i kada usluga dosegne više od 99,80% znat će se odgovor na zadnje pitanje; „Jesmo li tamo?“. Procjena za uspjeh poboljšanja je jedna godina, mjesečnim mjerenjem i definiranjem trendova.

IX. ZAKLJUČAK

Ovaj rad prikazuje primjenu ITIL najbolje prakse u atipičnom sektoru - bankarstvu. Pomaže izgraditi uslugu iz njezinog temelja postavljajući pitanje „Zašto?“. Izgradnjom nove usluge pokušavaju se predvidjeti sve situacije koje bi mogle narušiti operativni rad. Pružatelj usluge, banka, sagledava sve perspektive. S obzirom na atipično okruženje u kojem se pružatelj usluge nalazi, pokušava se nositi s digitalizacijom, sigurnošću i tradicionalizmom. Spajanjem tri glavna temelja ostvarila se usluga dohvata novih ključeva za mobilno bankarstvo. Digitalizacija predstavlja brzinu i zadovoljstvo korisnika, a bankomat predstavlja tradicionalizam, što se manifestira i u sigurnosti usluge - bankovnoj kartici koja se koristi više od 20 godina.

Iako korisnik nije svjestan svih faza u uvođenju usluge, one su glavni ključ u isporuci usluge te održavanju zadovoljstva i vjernosti. Svaka aktivnost i proces su mjerljivi, monitorirani i podložni izmjenama. Mjerljivost i nadzor definirani su u prvim fazama, dok se izmjene definiraju u kontinuiranom poboljšanju usluge. Tako se postiže uspješno vođenje uslugom. Digitalizacija i tradicionalizam nove usluge predstavljaju napredak u poslovanju, koji je vidljiv u pomno određenim dijagramima tijeka u svakoj fazi usluge. Baza znanja i visoka dostupnost usluge uhodanim kanalom stvaraju dodatnu vrijednost koja prestiže korisnikovu potrebu.

ITIL najbolja praksa daje uvid u izgradnju nove ili izmijenjene usluge i ne traži da se slijepo slijede sve njegove točke. Naprotiv, daje okvir i prilagodljivost na organizacijsku strukturu. Implementacijom i podrškom usluge unificira se mjesto događaja svih procesa. Definiranjem mjesta događaja pospješuje se upravljanje uslugom, njezinu efikasnost.

Usluga dohvata ključeva na bankomatu prikaz je velike mogućnosti ITIL najbolje prakse čiji je jedini cilj zadovoljstvo korisnika. Zadovoljan korisnik je cilj uvođenja svake usluge, a nadzor i visoka dostupnost nove usluge samo su dodatak koji produbljuje njegovu vjernost.

LITERATURA

[1] P. Hearsum, Introduction to the ITIL Service Lifecycle. London: TSO, 2011.

[2] M. Nieves, ITIL Service Strategy. London: TSO, 2011.

[3] V. Lloyd, ITIL Service Design. London: TSO, 2007.

[4] L. Taylor, ITIL Service Transition. London: TSO, 2007.

[5] D. Cannon, ITIL Service Operation. London: TSO, 2007.

[6] G. Case, ITIL Continual Service Improvement. London: TSO, 2007.

[7] R. Barišić, „IT Service Management Systems (ISO/IEC 20000-1:2011 and ITIL 2011)“, Proceedings of the 15th Croatian Conference on Quality and 6th Scientific Symposium of the Croatian Society for Quality (International Scientific and Professional Conference), Pages: 27-35, ISSN: 1848-8633, Croatian Society for Quality, Zagreb, Croatia, 2015.

[8] R. Barišić, D. Kerić, T. Žitko „ITIL and ISO/IEC 20000 in Croatian IT Companies“, Proceedings of the 16th International Symposium on Quality, Pages: 499-510, ISBN: 978-953-57036-9-3, Croatian Quality Managers Society, Zagreb, Croatia, 2015.

Uvođenje usluge bežičnog pristupa internetu pomoću ITIL smjernica

Krešimir Brodnjak, Renato Barišić
Visoko učilište Algebra, Zagreb, Republika Hrvatska
kresimir.brodnjak@racunarstvo.hr, renato.barisic@racunarstvo.hr

***Sažetak* - Cilj ovog rada je prikazati kako smisleno uz preporučene smjernice, opet ne standarda, uvesti novu uslugu u rad, kako upravljati njome i poboljšavati je. Potrebno je prilikom toga proći kroz sve faze uvođenja usluge, od same ideje za uvođenje usluge do konačnog puštanja u rad, ispravljanja nedostataka do poboljšavanja rada usluge uz moguća proširenja i opsega same usluge u budućnosti. Kao usluga je navedena usluga bežičnog pristupa internetu, ali sami koncept rada je primjenjiv na bilo koju uslugu. Za smjernice je odabran ITIL kao temelj zbog svoje opsežnosti praćenja usluge od ideje do korištenja i daljnjeg unapređenja. Kroz rad je prikazan pristup uvođenju usluge kroz sve njezine faze.**

Ključne riječi – usluga, bežični pristup, ITIL, ITSM, CSI

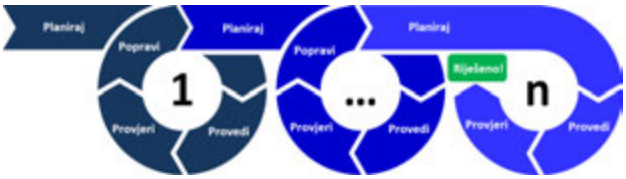
I. UVOD

Kroz ovaj rad će se prikazati primjer uvođenje nove usluge u rad unutar trenutnog poslovanja Kliničkog bolničkog centra Sestre milosrdnice. Samo uvođenje će biti prikazano po ITIL (Information Technology Infrastructure Library) v3 izdanje 2011 smjernicama za uvođenje usluge u rad po odvojenim fazama od začetka ideje same usluge pa kroz faze strategije usluge, dizajn usluge, tranzicije usluge, rukovanja uslugom i stalnog poboljšavanja usluge. Sama ideja uvođenja usluge bežičnog pristupa internetu za ležeće pacijente je došla od strane Uprave KBC Sestre milosrdnice zbog sve učestalijih upita od strane pacijenata za mogućnost pristupa internetu za vrijeme višednevnog boravka u ustanovi putem pametnih telefona, prijenosnih računala i tablet računala.

ITIL opisuje procese, postupke, zadatke i popise za provjeru koji nisu specifični za određenu organizaciju ili tehnologiju, ali ih organizacija može primijeniti za uspostavljanje integracije sa strategijom organizacije, isporuku vrijednosti i održavanje minimalne razine kompetencija. Ona omogućuje organizaciji da uspostavi osnovicu iz koje može planirati, provoditi i mjeriti. Koristi se za dokazivanje usklađenosti i mjerenje poboljšanja [1].

ITIL (engl. Information Technology Infrastructure Library) je podloga za upravljanje IT uslugama koja teži predvidivim i održivim uslugama usklađenima s potrebama društva i/ili organizacije. Osmišljen je kako bi se standardizirao izbor, planiranje, dizajn, isporuka i podrška za IT usluge u poslovanju. [2]

Povijest ITIL-a počinje davne 1980-te godine kada Britanska agencija CCTA (Central Computer and Telecommunications Agency) objavljuje preporuke jer su prepoznali da bez standardnih praksi vladine agencije i privatni sektor razvija neovisne načine upravljanja informatičkim procesima. Prvotno su preporuke bile okupljene u kolekciju knjiga, od kojih je svaka pokrivala određeno područje unutar upravljanja informatičkim procesima. Preporuke su bile bazirane na ponavljajućem procesnom modelu kontrole i upravljanja procesima, takozvanom PDCA (Plan-Do-Check-Act) modelu (slika 1).



Slika 1. Plan-Do-Check-Act model (prilagođeno od [3])

Za razliku od standarda, poput ISO/IEC 20000-1:2011, ITIL, kao skup dobrih praksi u području upravljanja informacijskim uslugama, daje veću fleksibilnost pri uvođenju i daljnjem održavanju i poboljšavanju usluge [4].

U hrvatskom gospodarstvu, ITIL je u značajnijoj mjeri prepoznat uglavnom u IT tvrtkama [5] iako se njegovi principi i preporuke koriste i puno šire.

ITIL dijelimo na pet faza, a to su:

Strategija usluge

U strategiji usluge definiramo kako ćemo, kao pružatelj usluge, uslugu isporučiti korisnicima sukladno njihovim očekivanju i definiramo strategiju kako ćemo upravljati tom uslugom.

Dizajn usluge

U dizajnu usluge istu dizajniramo na način da kreiramo pravo rješenje koje je u skladu sa poslovnim zahtjevima. Donosimo odluku kako ćemo isporučiti najoptimalniju uslugu.

Tranzicija usluge

Kroz fazu tranzicije usluge stvaramo uslugu koju ćemo isporučiti korisnicima. U njoj ispunjavamo zahtjeve koji su postavljeni u fazi strategije usluge, a do kraja razvijeni u fazi dizajna usluge.

Rukovanje uslugom

Namjena faze rukovanja uslugom je vršenje aktivnosti i procesa potrebnih za isporuku usluge korisnicima, nadzor isporuke usluge i pomoć korisnicima te za upravljanje tehnologijom koja se koristi za isporuku same usluge.

Stalno poboljšavanje usluge

Glavni ciljevi faze stalnog poboljšavanja usluge su analiza postojeće usluge koju pružamo i na temelju tih analiza donošenje preporuka za poboljšavanja kvalitete usluge i za nadogradnje iste.

II. STRATEGIJA USLUGE

U strategiji se definira primarno postavljanje samog cilja usluge, definiranje očekivanja od pružanja iste i koje vrijednosti se očekuju od uvođenja same usluge u katalog usluga. Kod postavljanja strategije usluge potrebno je odgovoriti na slijedeća pitanja [6]:

a. Iz koje perspektive krećemo – moramo znati što želimo raditi, koja nam je vizija same usluge?

Želimo uvesti uslugu mogućnosti bežičnog pristupa internetskim resursima za pacijente i zaposlenike Kliničkog bolničkog centra Sestre milosrdnice. Kada pacijenti dođu u prostor bolnice, da mogu za vrijeme boravka, bilo jednodnevnog ili višednevnog, kratiti vrijeme pristupom internetskim servisima. Trenutno takva usluga ne postoji, te bi donijela novost u poslovanju Bolnice, sa fokusom na zadovoljstvo korisnika a to su pacijenti, zaposlenici i tvrtke vanjski partneri.

b. Koji stav i poziciju ćemo tom uslugom zauzeti na tržištu?

Pozicija koju želimo zauzeti je da bolnica dođe na glas kao moderna Ustanova koja je otvorena za nove tehnologije u svrhu olakšanja boravka. Očekujemo da bi tom uslugom povećali broj pacijenata kao korisnika usluga same Bolnice, što bi dovelo do povećanja broja pretraga i pregleda. Modernizacija poslovanja same bolnice pozitivno bi utjecala i na poslovne procese, ubrzavanjem rada zaposlenika i mogućnosti upotrebe novih medicinskih uređaja čime bi otvorili mogućnost i novih medicinskih dijagnostičkih usluga.

c. Koji su nam planovi za postići željeno stanje, kojim redoslijedom?

Takvo stanje namjeravamo postići postavljenjem dovoljnog broja pristupnih točaka u Kliničkom bolničkom centru za pokrivanje Bolnice na cijelom području, kontrolom kvalitete usluge kako bi ista bila adekvatno pružena korisnicima, stalnim nadzorom rada usluge da i ukoliko dođe do prekida pružanja usluge, sami prekid bude učinkovito uklonjen u razumnom vremenu. Poboljšavanjem i nadogradnjom usluge konstantno bi podizali kvalitetu pružene usluge.

d. Kojim skupovima aktivnosti to namjeravamo postići?

To ćemo postići pravilnim planiranjem broja pristupnih točaka, zamjenom neadekvatne opreme, vršenjem nadzora putem centralnog mjesta upravljanja, i redovitog izvještavanja. Temeljem izvještaja kvaliteta usluge se analizira i traže se mjesta za poboljšanje i nadogradnju usluge.

Istovremeno je potrebno identificirati i definirati koje usluge ćemo pružati kojim korisnicima. Korisnici koje prepoznajemo su slijedeći:

Pacijenti – Pacijenti su krajnji korisnici zbog kojih se i uvodi bežični pristup internetu. Usluga koja se njima pruža mora biti dostupna stalno i adekvatna za korištenje.

Zaposlenici – Zaposlenici će koristiti uslugu bežičnog pristupa internetu, točnije bolničkim mrežnim resursima, radiološkom sustavu, laboratorijskom sustavu i bolničkom informacijskom sustavu.

Vanjske tvrtke partneri – Vanjske tvrtke partneri koji će koristiti mogućnost bežičnog pristupa bolničkim mrežnim resursima će u stvari povezivati medicinske uređaje na bolničke resurse putem bežičnog pristupa.

Iz vrsta korisnika vidimo da su nam potrebne dvije vrste usluge:

Usluga pristupa internetskim resursima

Usluga koja se pruža korisnicima pacijentima potrebno im je omogućiti pristup legalnim servisima, ali isto tako onemogućiti pristup bolničkim resursima iz sigurnosnih razloga. Iz toga proizlazi da taj segment bežične mreže mora biti logički izdvojen od ostatka bolničke mreže i infrastrukture.

Usluga pristupa bolničkim resursima

Korisnicima koji koriste bežičnu mrežu za pristup bolničkim resursima, a to su Bolnički informacijski sustav, Laboratorijski informacijski sustav i Radiološki informacijski sustav, potrebno je omogućiti nesmetan i siguran rad na tom mrežnom segmentu koji mora biti zaseban. U te korisnike spadaju zaposlenici KBC Sestre milosrdnice i tvrtke vanjski partneri.

III. DIZAJN USLUGE

Dizajn usluge pokazuje kako planirano rješenje uklopiti u postojeće poslovno i tehničko okruženje, koja je tehnologija i arhitektura infrastrukture potrebna za podršku samoj usluzi. Ova faza je mjesto povratka informacija iz ostalih faza, kada se uoči neki nedostatak u radu usluge, na temelju te informacije mijenjamo dizajn usluge kako bi postigli što bolji nivo kvalitete [7].

Za uspostavljanje usluge bežičnog pristupa internetu potrebno je :

- Strateški definirati broj pristupnih točci zbog pokrivenosti bežične mreže
- Proporcionalno prosječnom broju ležećih pacijenata definirati tehničku specifikaciju pristupne točke po njenim karakteristikama
- Definirati logički odvojen mrežni segment koji će se koristiti za dodjeljivanje adresa pacijentima
- Definirati model pristupa usluzi od strane korisnika
- Definirati proizvođača opreme i izvođača radova

- Definirati način unaprjeđenja usluge sukladno povratnim informacijama iz faze tranzicije i rukovanjem
- Definirati način proširenja usluge

Kada radimo dizajn usluge, moramo postići namjeravanu razinu usluge uz prihvatljivu razinu rizika i cijene pružanja usluga (cijene samog pružanja, ne cijene za krajnjeg korisnika usluge). Smanjenje troškova pružanja usluge možemo postići na više načina, od kojih su neki:

Način isporučivanja usluge

Ovisno o tome da li će isporuku usluge vršiti trenutni zaposleni KBC Sestre milosrdnice ili će se isporuka prepustiti nekoj vanjskoj tvrtki što bi dovelo do povećanja troškova jer je isto potrebno platiti i uklopiti u dugoročni financijski plan. Istovremeno, prepuštanje isporuke i upravljanja trenutnim zaposlenicima ne mora biti financijski jeftinije jer ovisi o kapacitetu IT odjela i znanju zaposlenika.

Cijena dostupnosti usluge

Dostupnost usluge je vrlo bitna stavka, jer ako usluga nije dostupna, nije ni pružena. Kada planiramo razinu dostupnosti usluge, moramo voditi računa da povećanje dostupnosti podiže i cijenu usluge. Ako ćemo koristiti najnižu razinu dostupne opreme i tehnologije cijena će biti niža nego da koristimo zadnju tehnologiju na tržištu čime osiguravamo najbolju dostupnost. Samim time treba pronaći ravnotežu u omjeru cijene i dostupnosti, jer neka vrhunska rješenja sa svojim mogućnostima možda i nisu potrebna jer se ista neće koristiti, a time smo financijska sredstva potrošili uzalud umjesto da smo ih preusmjerili na nešto drugo.

Jedna od preporuka je i sklopiti OLA (Organizational Level Agreement). Iako, možemo reći da postoji nepisana OLA između uprave KBC Sestre milosrdnice i Službe informatike (koja je konačni pružatelj usluge) jer je ista dužna isporučivati IT usluge krajnjim korisnicima. Potpisivanjem OLA-e možemo očekivati odličnu razinu usluge jer obvezuje pružatelja usluge na kvalitetu pružanja. Njome definiramo dužnosti Službe informatike koja se obvezuje na adekvatno pružanje usluge ali u skladu sa tehnologijom koju ima pravo koristiti kao i u skladu sa financijskim mogućnostima. Kako o financijskom dijelu omogućuje uprava KBC-a, ugovorom se definira razina pružanja usluge sukladno mogućnostima. Prilikom sklapanja, Služba informatike treba definirati koje su financijske potrebe (računajući sve resurse potrebne, ljudstvo, tehnologije, alate, uređaje) za pružanje usluge na zadovoljavajućem nivou, te sukladno tome upravlja uslugom. Također, ukoliko Uprava KBC-a nije u mogućnosti u nekom trenu zadovoljiti uvjete koji su definirani, Služba informatike nije odgovorna za pad kvalitete usluge jer od strane Uprave nije poštovana OLA, primjerice nisu omogućena tekuća sredstva za zamjenu neispravnih pristupnih točaka pa kod kvara nema mogućnosti zamjene.

Uz ugovor na razini organizacije potrebno je napisati SLA (Service Level Agreement). SLA se provodi između pružatelja usluge i korisnika, putem kojem se korisnici slažu sa uvjetima korištenja usluge, to jest pružatelj usluge se obvezuje pružiti uslugu pod određenim uvjetima. SLA će biti napravljen kao uvjeti korištenja, sa kojima se korisnici usluge moraju složiti ukoliko žele koristiti uslugu. Prilikom spajanja na bežičnu mrežu, korisnicima će putem korisničkog portala biti omogućen uvid u uvjete korištenja, koje moraju označiti kao pročitane da bi mogli koristiti bežičnu mrežu.

A. Snimka postojećeg stanja

Klinički bolnički centar sagrađen je kao paviljonski sustav koji se sastoji od 39 zgrada, s time da pojedine zgrade imaju do pet katova. Do takve gradnje došlo se godinama, od prvotne jedne do trenutnog stanja. Centralna usmjernička točka nalazi se u poslužiteljskoj sobi u prizemlju zgrade broj 1. (slika 2). Komunikacijski ormari za smještaj aktivne mrežne opreme nalaze su zgradama unutar Ustanove i povezani su sa centralnom točkom putem optičkih kablova, dok su veze između komunikacijskih ormara unutar zgrada gdje postoje, napravljene pomoću UTP (Unshielded Twisted Pair) Cat5e kabela. Postoje dvadeset i dvije optičke veze između centralne pristupne točke i komunikacijskih ormara koji su smješteni po zgradama Ustanove, od kojih je jedanaest brzine 10 gigabita u sekundi, a jedanaest brzine 1 gigabit u sekundi.



Slika 2. Plan KBC Sestre milosrdnice

Ukupno postoji 44 komunikacijska ormara u kojima je smješteno ukupno 57 komada aktivne mrežne opreme, proizvođača Cisco Systems.

U ormarima za smještaj aktivne komunikacijske opreme ne postoje uređaji za neprekidno napajanje i nema mjesta za smještaj istih u ormare. Kod svakog ormara, to jest ispod svakog postoji zasebni ormarić namijenjen za smještaj uređaja za neprekidno napajanje, klasičnih dimenzija. Svaki ormar ima svoj dovod električne energije na koji je spojen strujni razdjelnik sa pred naponskom zaštitom i samo su pojedini ormari spojeni na generator električne energije koji koristi cijela Ustanova.

Ukupno postoje 32 VLAN-a (Virtual LAN, Virtual Local Area Network) koji se koriste za pristup računalne

opreme na samu mrežu KBC Sestre milosrdnice i resursima na mreži, te će biti potrebno u fazi tranzicije napraviti rekonfiguraciju mrežne infrastrukture.

B. Planiranje broja pristupnih točaka

Kako je Ustanova koncipirana kao skup nepovezanih zgrada i pacijenti i korisnici se nalaze u zatvorenim prostorima, iste je potrebno pokriti signalom bežične mreže. Zgrade koje se koriste kao servisne, to jest u kojima se nalaza servisi kao primjerice praonica rublja i tehnička radiona također je potrebno uzeti u planiranje zbog budućih projekata koji bi mogli zahtijevati pristup bežičnoj mreži. Površina koju je potrebno pokriti adekvatnim signalom je ukupno 44.317 kvadratnih metara, koje možemo podijeliti prema slijedećem popisu:

- Bolesničke sobe – 7.073,24 m2
- Intenzivne jedinice – 1.230.05 m2
- Operacijske dvorane – 2.217,45 m2
- Ambulante/dijagnostike/labosi – 7.287,39 m2
- Sterilizacija – 365,00 m2
- Prehrana/čajne kuhinje/boravak – 628,68 m2
- Sanitarni čvorovi – 1.921,52 m2
- Hodnici/čekaonice – 9.584,28 m2
- Uredski prostori – 13.685,57 m2

Također, kod planiranja broja pristupnih točaka treba voditi računa o broju korisnika koji se spajaju na pristupne točke. U Kliničkom bolničkom centru postoji 827 bolesnički ležaj, a dnevno kroz Ustanovu prođe i do 3.500 prolaznih pacijenata.

IV. TRANZICIJA USLUGE

Faza tranzicije opisuje kako iz postojećeg (ukoliko postoji) stanja usluge preći u novo stanje usluge, ili kako uvesti novu uslugu u svakodnevno poslovanje. Tranzicija usluge je faza koju krajnji korisnik ne vidi, za razliku od slijedećih faza. U ovoj fazi, usluga koja se uvodi kao takva još ne postoji, ovo je posljednja faza prije uspostavljanja usluge prema krajnjem korisniku. Po završetku faze tranzicije usluge, ona postaje dostupna korisnicima. Objedinjava se sve navedeno u prethodne dvije faze i usluga se pušta u rad [8].

U fazi tranzicije za navedeni primjer potrebno je:

- Kupiti uređaje za bežični pristup internetu
- Kupiti i zamijeniti neadekvatnu aktivnu mrežnu opremu
- Izvesti potrebna kabliranja na odjelima od kabelskih ormara do mjesta gdje će biti postavljeni uređaji i postaviti uređaje
- Kreirati novi mrežni segment za pristupne uređaje i odvojiti ga od ostatka mreže
- Kreirati novi mrežni segment za pacijente, odvojiti ga logički od ostatka mreže
- Provesti mjerenja pokrivenosti i brzine na odjelima gdje su postavljene pristupne točke
- Provesti testiranje sigurnosti ostatka mreže nakon uspostave novog mrežnog segmenta
- Provesti testiranje ponovne uspostave usluge ukoliko dođe do prekida iste – simulacija prekida usluge zbog mehaničkog kvara ili softverskog kvara, nestanka struje.

Ukoliko se u fazi tranzicije otkrije nedostatak, primjerice nedovoljan broj pristupnih točci ili izračunat premali mrežni segment za pacijente, potrebno je pružiti povratnu informaciju o tome u fazu dizajna usluge da se naprave potrebne preinake u dizajnu. Vrlo je bitno što više uočenih mana ispraviti prije faze rukovanja uslugom, prije nego sama usluga postane dostupna i korisnici primijete mane i nedostatke same usluge.

Kao što smo vidjeli u fazi dizajna usluge, veći broj postojećih mrežnih uređaja koji se nalaze u upotrebi nisu adekvatno upotrebljivi, te je potrebno iste zamijeniti novim modelima koji će svojim specifikacijama zadovoljiti potrebe brzine i kapaciteta potrebne za uslugu bežičnog pristupa internetu i bolničkim resursima. Potrebno je raspisati natječaj za nabavu mrežnih preklopnika i pristupnih točaka prema odgovarajućim specifikacijama.

Kako se trenutna mrežna infrastruktura sastoji od preklopnika marke Cisco i ne mijenjaju se svi uređaji zbog potreba uvođenja bežične mreže, preklopnici koji se nabavljaju moraju biti isto tako od proizvođača Cisco Systems kako ne bi došlo do postojanja dva različita proizvođača mrežne opreme u produkciji što bi uvelike otežalo eventualnu dijagnostiku problema u komunikaciji zbog različitosti podržanih protokola [9].

Zbog specifičnosti okruženja u kojem se uvodi bežični pristup, potrebno je isplanirati zamjenu postojeće mrežne opreme to jest preklopnika sa što manjim prekidima u radu korisnika. 25 preklopnika koje je potrebno zamijeniti nalaze se u 24 mrežna ormara. Sama izmjena preklopnika uz prethodnu konfiguraciju novih traje okvirno do dva sata po ormaru. Najveća frekvencija pacijenata a samim time i potreba za korištenjem mrežnih resursa unutar KBC Sestre milosrdnice se dešava u vremenu od 7h ujutro do 15h popodne radnim danom, dok je vikendom smanjen opseg pacijenata te su najveća opterećenja u prostorima hitnih službi. Vodeći se time, većina opreme će se mijenjati subotom od 9h ujutro, nakon jutarnjih vizita i očitanja nalaza. Mrežni preklopnici koji se nalaze u zgradama Uprave, Stare vile i tehničke radione, njih ukupno osam, će se zamijeniti radnim danom u 15h.

V. RUKOVANJE USLUGOM

Rukovanje uslugom je potrebno kako bismo osigurali upravljanje i isporuku usluge na razini. U tu svrhu je potrebno organizirati izvođenje aktivnosti i procesa čijim bi se provođenjem održavala ugovorena usluga. Samim održavanjem usluge zadržali bi i povećali zadovoljstvo korisnika, utjecaj prekida usluge bio bi minimalan sa minimalnim utjecajem na procese u KBC Sestre milosrdnice. Kako bi mogli kvalitetno rukovati uslugom,

moramo definirati što kreira uslugu, koji su procesi pomoću kojih upravljamo uslugom, koje tehnologije su nam potrebne za isporuku usluge, tko će rukovati uslugom, to jest raspodijeliti rukovanje uslugom da djelatnike koji rukuju istom.

Kako bi kvalitetno upravljali uslugom, istu moramo nadzirati, uspoređivati kvalitetu iste sa definiranim normama, i kontrolirati dobivene rezultate. Rukovanje uslugom dijelimo na pet osnovnih procesa [10]:

- upravljanje događajima
- upravljanje incidentima
- ispunjavanje zahtjeva
- upravljanje problemima
- upravljanje pristupom

Upravljanje događajima nam se nameće kao temelj za nadzor usluge i kontrolu, te nam kroz životni ciklus događaja omogućuje otkrivanje, razumijevanje i definiranje odgovarajućih aktivnosti kao odgovor na događaj. Upravljanje događajima omogućuje automatizaciju i samim time povećanje efikasnosti jer se automatizacijom nekih aktivnosti može postići oslobađanje ljudskih resursa za druge aktivnosti. Također automatizacijom možemo postići brži odgovor na neki događaj, a samim time povećavamo i zadovoljstvo korisnika usluge.

Upravljanje incidentima dolazi do izražaja kod brzine uspostave usluge u slučaju prekida. Svaki prekid ili smetnja u korištenju usluge smatra se incidentom i potrebno je kod pojave istoga što prije vratiti sustav u normalno funkcionalno stanje. Incidenti se mogu prijavljivati automatski kroz automatizirane procese, primjerice slanje maila tehničkom osoblju koji vode brigu o sustavu prilikom prestanka rada pristupne točke.

Ispunjavanje zahtjeva služi kako bi zadovoljstvo korisnika usluge bilo na što višoj razini dok istu koriste. U tu svrhu potrebno je imati jasan komunikacijski kanal kojim se predaju zahtjevi, korisnicima treba dati (i davati) informacije o dostupnosti usluge, primjerice na svim mjestima gdje je dostupan bežični pristup internetu treba korisnicima to naglasiti i omogućiti davanje informacija o usluzi, dostupnosti i načinu korištenja.

Upravljanje problemima je potrebno zbog što efikasnijeg uklanjanja istog, sa time da treba naglasiti da svaki problem nije incident, dok je svaki incident problem. Potrebno je dokumentirati svaki problem, pronaći temeljni uzrok, primjerice na odjelu gdje je dostupna usluga korisnicima se daje kriva pristupna zaporka. Za svaki poznati problem potrebno je voditi evidenciju o trajnom rješenju i o zaobilaznom rješenju, te evidenciju kako bi ponavljajuće probleme spriječili promjenom procesa zbog kojeg dolazi do tog problema. Aktivnosti upravljanja problemima se svode na pitanja na koja je potrebno odgovoriti kako bi istog riješili:

Upravljanje pristupom nam je potrebno kako bi znali tko je spojen i šta smije raditi. Korisnici moraju biti autorizirani dok se neautoriziranima sprječava pristup na bežičnu mrežu i korištenje usluge. Time se podiže sigurnost bolničke infrastrukture i smanjuje broj eventualnih sigurnosnih incidenata. Pravo pristupa na bežičnu mrežu za pacijente imaju svi pacijenti koji to usmeno zatraže od djelatnika odjela na kojem borave koji će im po upitu pružiti informacije o spajanju na bežičnu mrežu. Zaposlenici Kliničke bolnice Sestre milosrdnice se moraju javiti u Službu informatike s uređajem kojim će se spajati na bežičnu mrežu, kako bi se isti evidentirali i povezali s korisnikom. Prilikom prestanka radnog odnosa, djelatnici ostaju bez prava spajanja na bolnički bežičnu mrežu. To će se regulirati sukladno izvješću kadrovske službe koja će po odlasku djelatnika obavještavati Službu informatike da se ukine pravo pristupa.

VI. STALNO POBOLJŠAVANJE USLUGE

Uslugu kao takvu, potrebno je držati na zadovoljavajućoj razini [11]. Međutim, potrebno je voditi računa i o poboljšavanju usluge i to neprestano. Svaki uočeni prostor za poboljšanje potrebno je iskoristiti. Svrha poboljšavanja usluge je usklađivanje sa promjenama u poslovanju, u navedenom slučaju primjerice pojačavanje signala, (bez obzira na zadovoljavajuće trenutno stanje), ili povećanje pokrivenosti uslijed širenja odjela u novu zgradu.

Potrebno je stalno tražiti način poboljšanja usluge, u vidu povećanja efektivnosti usluge, potrebno je definirati što efektnije procese u radu usluge i po mogućnosti smanjivati troškove usluge, recimo ako se primijeti da je negdje postavljeno previše pristupnih točaka, iste je potrebno izmjestiti na područje gdje nedostaju a ne da se kupuju nove pristupne točke.

Stalno poboljšavanje usluge možemo bazirati na analitičkim izvještajima od strane Service deska i na temelju mjerenja performansi usluge.

Iz toga vidimo da je potrebno imati kontrolno mjesto pristupnih točaka, mjeriti pokrivenost signalom, eliminirati ponavljajuće probleme i incidente smanjiti na minimum ili ih ukloniti. Da bi se to moglo ispravno i pravilno provoditi, potrebno je kreirati registar stalnog poboljšavanja, koji mora sadržavati popis svih uočenih prilika za poboljšanje, koje unutar njega treba kategorizirati prema opsegu angažmana za aktivnosti koje su potrebne da se poboljšanje napravi, uz popisane očekivane rezultate provedenih aktivnosti, to jest njihove dobrobiti.

Usluga kao takva mora biti dostupna, pouzdana i adekvatna. Kako bi to osigurali, potrebno je redovito provoditi mjerenja dostupnosti usluge, pouzdanosti usluge i performansi usluge. Za provođenje tih mjerenja, potrebno je ustanoviti takozvani „base line“ (engl. Osnovne vrijednosti) navedenih karakteristika, koji nam služi kao polazna točka za usporedbu dobivenih rezultata.

A. Praćenje rada usluge

Nakon uspostave usluge, istu je potrebno pratiti da bi se mogli uočiti nedostaci, točnije rečeno da se uoči prostor za poboljšanje usluge. Praćenje rada usluge može se vršiti automatskim nadzorom prekida u radu, dok kvalitetu same usluge možemo temeljiti na anketama provedenim sa

korisnicima. U tu svrhu potrebno je napraviti dvije vrste ankete: jednu za pacijente i drugu za zaposlenike KBC Sestre milosrdnice. Moraju biti anonimne i lako dostupne. Prilikom otpusta iz Bolnice, pacijenti će biti zamoljeni da ispune anonimnu anketu o korištenju pristupa Internetu putem bežične mreže, dok će anketa namijenjena korisnicima biti dostupna na intranet stranicama bolnice, a sami korisnici će biti zamoljeni da je ispune svaka tri mjeseca tokom prve godine, a nakon toga svakih šest mjeseci.

B. Automatsko praćenje dostupnosti usluge

Da bi pružatelj usluge mogao istu pružiti na zadovoljavajućoj razini i uočio mjesto za poboljšanje, potrebno je istu pratiti i nadzirati njezin rad. To možemo raditi na nekoliko načina. Jedan od načina koji uvelike to pojednostavljuje je praćenje rada usluge bežičnog pristupa putem softvera za nadzor koji ima automatsko obavještavanje pružatelja usluge o promjenama dostupnosti usluge.

C. Planiranje načina poboljšavanja usluge i poboljšavanje usluge

Temeljem rezultata prije navedenih anketa možemo planirati poboljšavanje usluge. Inicijalna anketa prema zaposlenicima će se provesti trideset dana po puštanju usluge u rad, dok će redoviti raspored ispunjavanja anketa biti svaka tri mjeseca tokom prve godine, a nakon toga svakih šest mjeseci. Dobivene rezultate potrebno je analizirati i na temelju dobivenih rezultata korigirati rad usluge. Rezultati dobiveni jednom mjesečno za prethodni mjesec koristiti će se za korekcije rada usluge dok će se na temelju kvartalnih i godišnjih rezultata anketa planirati veće izmjene u usluzi. Uočeni problemi kroz ankete su nam povratna informacija za fazu Dizajn usluge u kojoj će nam poslužiti kao osnova za redizajn usluge i poboljšavanje.

Informacije dobivene anketama, čiju analizu radi Voditelj service deska, koristiti će se za poboljšavanje usluge. Kao primjer, ako se na temelju rezultata anketa za pacijente uoči da je pokrivenost signalom na nekom dijelu odjela slabija i zbog toga nije moguće koristiti bežičnu mrežu na zadovoljavajućem nivou, na tom dijelu odjela postaviti će se dodatna pristupna točka.

Uspostavom usluge otvaraju se nebrojene mogućnosti za daljnje poboljšanje kvalitete i učinkovitosti rada u KBC Sestre milosrdnice, primjerice uvođenje „digitalne vizite“, na kojoj će se vizite i uvid u dokumentaciju pacijenata za vrijeme iste raditi putem tableta ili prijenosnih računala, bez potrebe nošenja pacijentove medicinske dokumentacije. Otvara se i mogućnost unapređenja opremljenosti bolesničkih kreveta koji mogu biti povezani sa Bolničkim informacijskim sustavom, može se uvesti praćenje kretanja pacijenata putem pametnih narukvica, dok bi korištenje bežične telemetrije za praćenje trenutnog stanja zdravstvenih parametara pacijenata olakšalo rad zaposlenicima i boravak pacijentima.

VII. ZAKLJUČAK

Ovim radom je prikazano uvođenje usluge bežičnog pristupa internetu pomoću ITIL smjernica. Opisano je na primjeru koliko je bitno sustavno i određenim redoslijedom uvoditi novu uslugu u rad. Potrebno je voditi brigu o svim fazama uvođenja usluge, ali i tijekom cijelog njezinog životnog ciklusa. Faze su ponavljajuće te se korigiraju temeljem povratnih informacija iz drugih faza. Čak i dizajn usluge se može korigirati temeljem povratnih informacija iz drugih faza, posebno u fazi korištenja usluge sa ciljem stalnog poboljšavanja usluge. Životni vijek usluge traje sve dok se ona pruža i briga o usluzi ne prestaje u trenu njenog uvođenja, dapače tek onda počinje. Nakon uvođenja usluge u rad možemo uočiti mjesta za poboljšanje usluge i vidimo gdje smo možda nešto previdjeli pa na temelju toga šaljemo povratne informacije u fazu dizajna usluge i vršimo promjene u dizajnu.

Dobro planiranje usluge je temelj za bazne funkcionalnosti, a kasnije nakon što je usluga u radu, povratne informacije se koriste u svrhu poboljšavanja iste. Kroz tranziciju se vrši priprema za puštanje usluge u rad na temelju faze dizajna usluge. Rukovanje uslugom daje pogled na rad same usluge, te se nadgledanjem stanja dolazi do informacija koje se prosljeđuju radi stalnog poboljšavanja usluge. Kontinuirano poboljšavanje usluge otvara prilike za poboljšanje i nadogradnju temeljenu na informacijama iz faze rukovanja uslugom, praćenjem rada usluge i korisničkim anketama.

U fazi stalnog poboljšavanja usluge planira se kako nadograditi uslugu novim mogućnostima ili kako uvesti novitete u ostalim radnim procesima ustanove koje se temelje na korištenju usluge bežičnog pristupa internetu.

ITIL prati razvoj usluge od prvotne ideje pa kroz korištenje usluge nudi modele unapređenja i praćenja te se zbog toga pokazuje kao vrlo dobro rješenje za realizaciju usluge. Građen je na temelju višegodišnjeg iskustva u radu, to jest najboljih praksi koje su primjenjive u bilo kojem području.

LITERATURA

[1] P. Hearsum, Introduction to the ITIL Service Lifecycle. London: TSO, 2011.

[2] Cartlidge, C.; Rudd, M.; Smith, P.; Wigzel, S.; Rance, S.; Shaw, T., "An Introductory Overview of ITIL® 2011", The Stationery Office, 2012.

[3] PDCA (en.wikipedia.org/wiki/PDCA), pristup siječanj 2020.

[4] R. Barišić, „IT Service Management Systems (ISO/IEC 20000-1:2011 and ITIL 2011)“, Proceedings of the 15th Croatian Conference on Quality and 6th Scientific Symposium of the Croatian Society for Quality (International Scientific and Professional Conference), Pages: 27-35, ISSN: 1848-8633, Croatian Society for Quality, Zagreb, Croatia, 2015.

[5] R. Barišić, D. Kerić, T. Žitko „ITIL and ISO/IEC 20000 in Croatian IT Companies“, Proceedings of the 16th International Symposium on Quality, Pages: 499-510, ISBN: 978-953-57036-9-3, Croatian Quality Managers Society, Zagreb, Croatia, 2015.

[6] M. Nieves, ITIL Service Strategy. London: TSO, 2011.

[7] V. Lloyd, ITIL Service Design. London: TSO, 2007.

[8] L. Taylor, ITIL Service Transition. London: TSO, 2007.

[9] Cisco Press, Implementing Cisco Switched Networks (SWITCH) Foundation learning Guide, 2010.

[10] D. Cannon, ITIL Service Operation. London: TSO, 2007.

[11] G. Case, ITIL Continual Service Improvement. London: TSO, 2007.

CE

International Conference on

COMPUTERS IN EDUCATION

Steering Committee

Chairs: Marina Čičin-Šain, University of Rijeka, Croatia
Ivanka Sluganović, I. Technical School Tesla, Zagreb, Croatia
Jaak Henno, Tallinn University of Tehnology, Estonia
Hannu Jaakkola, Tampere University, Pori, Finland
Snježana Babić, Polytechnic of Rijeka, Croatia

Members: Michael E. Auer, Carinthia Tech Institute, Villach, Austria
Marjan Krašna, University of Maribor, Slovenia
Duško Lukač, Siegburg, Germany
Robert Repnik, University of Maribor, Slovenia
Slavomir Vukmirović, University of Rijeka, Croatia

Program Committee

Members: Lejla Abazi-Bexheti, SEE University, Tetovo, North Macedonia
Nikola Kadoić, University of Zagreb, Croatia
Ivan Kaštelan, University of Novi Sad, Serbia
Mario Konecki, University of Zagreb, Varaždin, Croatia
Božidar Kovačić, University of Rijeka, Croatia
Gorana Mudronja, University of Rijeka, Croatia
Dana Paľová, Technical University of Košice, Slovakia
Libuša Révészová, Technical University of Košice, Slovakia
Frano Škopljanac-Mačina, University of Zagreb, Croatia
Andreja Špernjak, University of Maribor, Slovenia

PAPERS

Parents' Attitudes toward Programming in Elementary Schools in City of Osijek

A., Papić*, I. Ćosić*
* Faculty of Humanities and Social Sciences/Department of information sciences, Osijek, Croatia
apapic@ffos.hr, icosic1@ffos.hr

***Abstract* - Programming is the process which enables a production of software, applications and web sites. Programming develops skills for logic and critical thinking and problem solving necessary for living in today information society. In the theoretical part is pointed out a lot of positive aspects of Micro: bit technology, the suitable programming tools for elementary school pupils such as Scratch are described as well as numerous projects of programming in city of Osijek oriented toward this particular population. In the empirical part the parents' attitudes toward programming in elementary schools in city of Osijek are explored. The results of the research show parents' familiarity with elementary concepts in regard to programming, the possible connections between parents' workplaces and their perceptions and support of programming within school or out of school activities.**

Keywords – elementary school; parents; programming;

I. INTRODUCTION

Programming can be defined as the process which enables a production of software, applications and web sites [5]. Programming is not just writing of computer programs but moreover programming is problem solving. To solve certain problem it is necessary to develop appropriate strategies which are not necessary inside the field of programming. The technology is children's natural surrounding thus technology and programming should be incorporated into educational process in schools. Computational thinking became skill integrated within everyday reading and writing at computer [3]. Programming language is defined as group of commands and other syntax for creating of software program. Program development involves several steps: problem definition, problem solution, program testing and program documentation [9]. Programming language enables to write the language of machine in readable format. Learning the programming languages is similar to learning ordinary languages thus children should learn programming languages as earlier as possible to accept them better. Nowadays a great attention is given to Micro: bit technology and its implementation across elementary schools in Croatia. Namely, Micro: bit is a little hand device specially designed to make programming more accessible for teaching and learning. It enables beginners to learn programming through 25 LED monitor, buttons and embedded sensors. It is designed by BBC, Microsoft, Samsung and Science Scope and it gives insight into introduction of coding for pupils around the world. Micro: bit is very easy for use, efficient and cost-effective hardware. It can be programmed through every web browser without additional installation of software. There are more than 200 different activities and resources which can be try out at Micro: bit, from simple experiments to challenges of creative coding. Micro: bit programming is compatible with programming languages Python and C++. Microsoft offers two programs for easier coding – Microsoft Touch Develop, so called textual programming and Microsoft Blocks, so called graphic programming language [1]. Micro: bit can be connected to another mother boards which enables transmission of signals to robots and motors through electrodes. Micro: bit is more and more employed into elementary schools because of easier and simplest knowledge transfer regarding programming. For visual programming in elementary school a very appropriate programming tool is Scratch. Scratch is a programming tool especially for children which enables research and experiments with concepts of computer programming through graphical interface which is easy to use [8]. Programming language Scratch is born at MIT University and it is based on idea of programming language Logo. Scratch enables creating games, stories and interactive presentations. Scratch interface consists of stage with spites and commands are grouped into pallets according to their purpose. Scratch is programming language with following main properties: (1) low floor – enables easy start, (2) high ceiling – gives opportunities to increase complexity and (3) wide walls – supports different kinds of projects to involve people with different preferences [12]. This paper aims to explore parents' attitudes toward programming in elementary schools in city of Osijek thus the research questions raised in this paper are the following:

(1) What kind of attitudes toward programming in general in elementary school parents have?

(2) What kind of attitudes toward programming with Micro: bit parents have?

(3) Do parents' attitudes toward programming and their support of programming within school or out of school activities depend upon parents' workplaces and their jobs?

After the introduction which gives theoretical insight into the theme of the paper follows the description of the method used in the research. Third section of the paper consists of the overview of gained research results and related discussion. The last concluding section of the paper brings the main findings of the research and points out its possible implications.

II. METHOD

In this paper the method used in this research is an online questionnaire distributed in September 2019 year toward random sample of parents of pupils in all elementary schools in city of Osijek. The online questionnaire was firstly distributed to teachers and than they distributed online questionnaire to parents of their pupils. Total of 102 fulfilled questionnaires were analyzed for the purpose of answering to the research questions.

III. RESULTS AND DISCUSSION

The total of 102 parents filled out the online survey regarding their attitudes toward programming of their children in elementary schools in city of Osijek. According to gender in research participated 46 (45.1 %) of male parents and 56 (54.9%) female parents. At Fig. 1. it can be seen parents' gender.

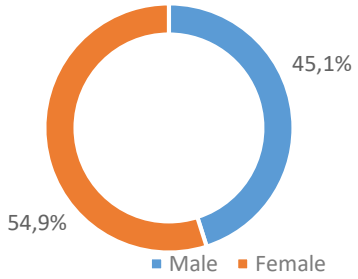


Figure 1. Parents' gender

According to the age in research participated 9 parents between 25 - 35 years old (8.8%), 50 parents between 35 - 45 years old (49.1%), 39 parents between 45 - 55 years old (38.2%) and 4 parents between 55 - 65 years old (3.9%). Almost half of parents had between 35 - 45 years old (49.1%) and for those parents and younger namely 9 parents between 25 - 35 years old (8.8%) it could be supposed that they learned about some main programming concepts during their formal or informal previous education. The parents were asked are they familiar with the main concepts of programming. 85 parents (83.3%) answered that they are in some way familiar with programming concepts but 17 parents (16.7%) answered that they did not heard for this terms at all. Fig. 2. shows the pecentage of parents regarding their familiarity with main programming concepts.

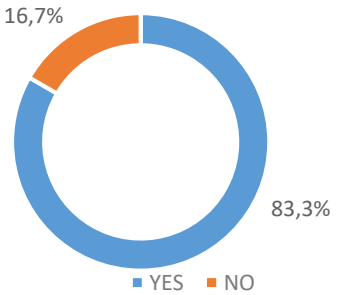


Figure 2. Parents' familiarity with main programming concepts

T

Parents were asked weather their workplace and job is in some kind of relation with programming skills. The total of 96 parents answered this question. 79 parents (82.3%) claimed that there does not exist any kind of relation of their job with programming. The minority of parents, 17 (17.7%), answered that their workplace and job is in relation to programming. Fig. 3. shows that the majority of parents' workplaces and jobs were not in any kind of relation to programming. Although, results shown at Fig. 2. and Fig. 3. seem contradictory it could be explained that parents find programming interesting and are informed about programming either during their own formal or informal previous education or they learned about programming together with their children or perhaps from their children. Although, parents claimed that they were familiar with main programming concepts for majority of them their workplaces and jobs were not in any kind of relation to programming.

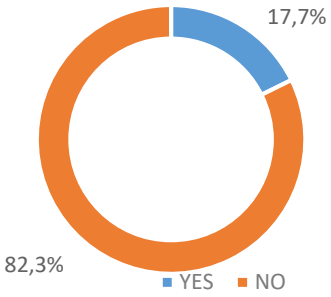


Figure 3. Parents' workplace and job in relation to programming

Parents were asked do they think that learning with Micro: bit is very interesting and just one parent completely disagreed with this. 9 parents (8.8%) disagreed with this, 24 parents (23.5%) neither agreed, neither disagreed with this. The majority of parents even 61 parents (59.8%) agreed with this and 7 parents (6.9%) parents completely agreed with this. Parents were asked do they think that Micro: bit within computer science in school should be more employed. 2 parents completely disagreed with this and 5 (4.9 %) of them disagreed with this. 26 parents (25.5%) neither agreed, neither disagreed with this. The majority of 58 parents (56.9%) agreed with this and 11 of them (10.8%) completely agreed with this. Parents do not completely agreed with claim that programming with Micro: bit is simplier. 11 parents (10.8%) disagreed with this claim. 34 parents (33.3%) neither agreed, neither disagreed with this.

Less than half of parents 45 parents (44.1%) agreed with this and 12 parents (11.8%) completely agreed with this. Parents were asked do they think that non-computer science teachers should also use Micro: bit and just one parent does not completely agreed with this. 6 parents (5.9%) disagreed with this and 47 of them (46.1%) neither agreed, neither disagreed with this. The third of parents 35 of them (34.3%) agreed with this and 13 parents (12.7%) completely agreed with this. Parents were asked do they think that Micro: bit enables easier entrance into the digital world for their children and 3 parents (2.9%) completely disagreed with this. 7 parents (6.9%) disagreed with this and even 27 parents (26.5%) neither agreed, neither disagreed with this. Less than half of parents 44 of them (43.1%) agreed with this and 21 parents (20.6%) completelly agreed with this. Fig. 4. shows parents' attitudes toward programming with Micro: bit.

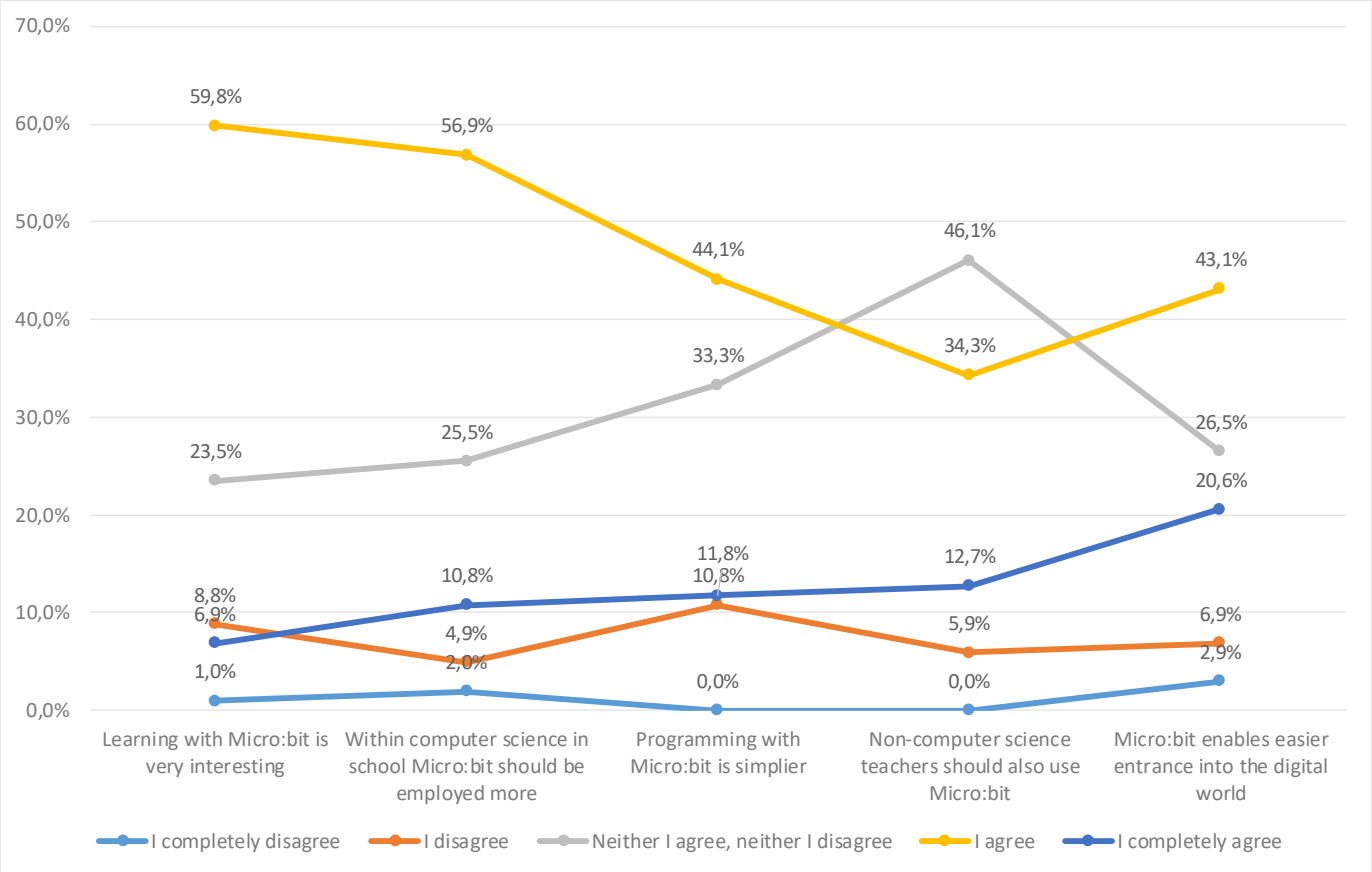


Figure 4. Parents' attitudes toward programming with Micro: bit

Parents were asked to grade the most important opportunities of Micro: bit technology. 35 parents (34.3%) agreed that speed of use is important characteristic. The majority of parents, 55 parents (53.9%), agreed that easiness of use of Micro: bit is the most important characteristic. 44 parents (43.1%) agreed that compatibility with other devices such as laptops, tablets, mobile devices etc. is very important characteristic.

The half of parents, 52 parents (51%) agreed that multidisciplinarity is the most important characteristic of Micro: bit. 26 parents (25.5%) agreed that Micro: bit enables education not just for pupils but also for teachers and that this is very important characteristic. Fig. 5. shows parents' attitudes toward different opportunities of Micro: bit.

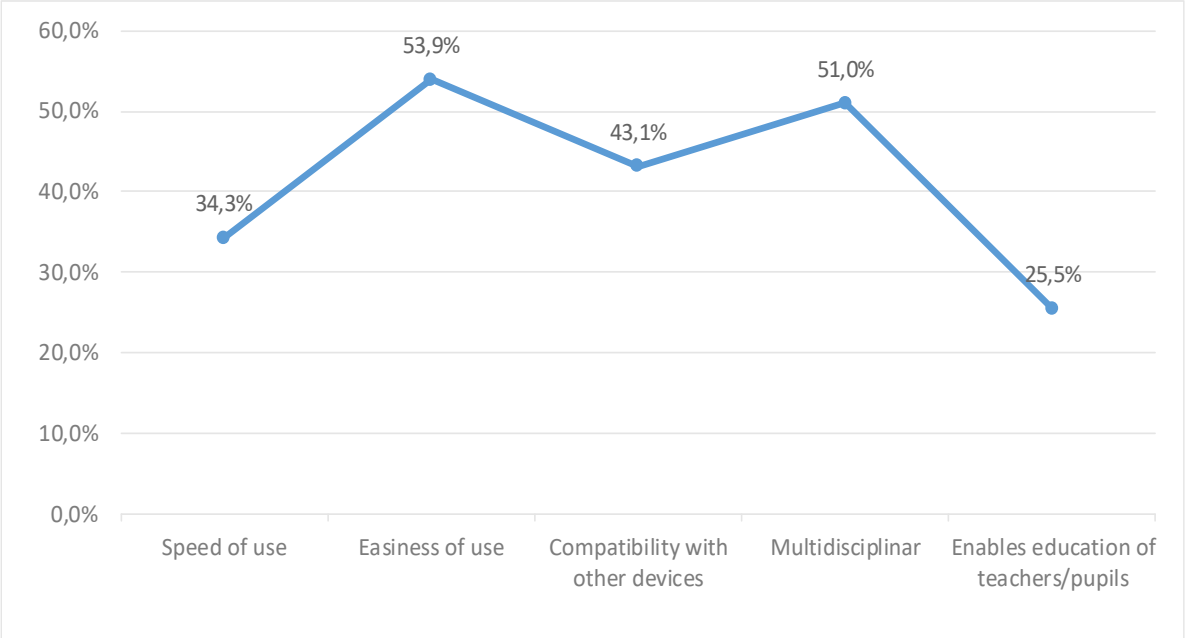


Figure 5. Parents' attitudes toward opportunities of Micro: bit

Parents were asked weather they are familiar with following programming tools namely weather their children learned about them in elementary school. The majority of parents, 55 parents (53.9%), answered that they are familiar with programming tool Scratch [11]. Less than half of parents, 42 parents (41.2%) answered that they are familiar with programming tool Code.org [4].

Almost third of parents, 27 parents (26.5%) answered that they are familiar with programming tool Blockly [2]. Just 13 parents (12.7%) were familiar with programming tool Hopscotch [7]. 20 parents (19.6%) were familiar with programming tool Pocket Code [10] and 14 parents (13.7%) were familiar with programming tool Gdevelop [6]. Fig. 6. shows parents' familiarity with different programming tools.

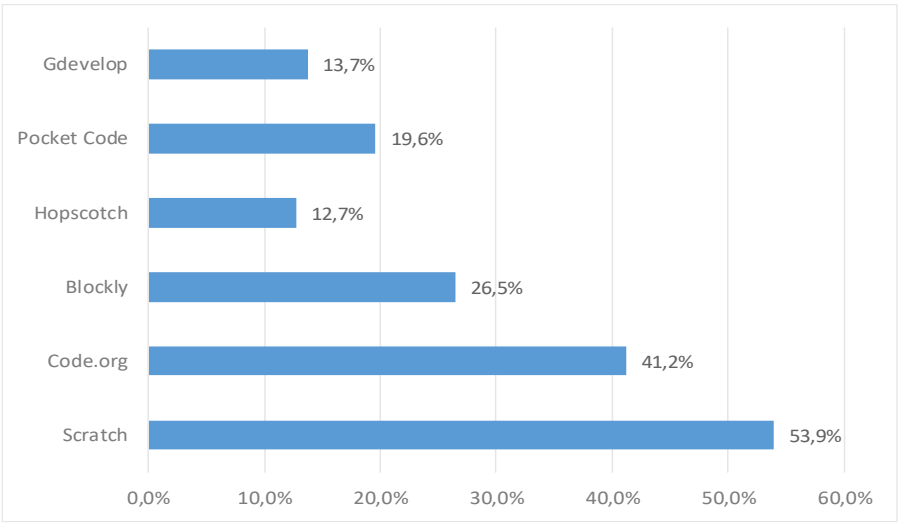


Figure 6. Parents' familiarity with programming tools

Parents were asked do they think that programming is necessary skill for their child for further life. 2 parents completely disagreed with this. 11 parents (10.8%) disagreed with this. 13 parents (12.7%) neither agreed, neither disagreed with this. The half of parents, 51 parents (50%) agreed with this. 25 parents (24.5%) completely agreed with this. Parents were asked do they think that programming should have priority at computer science classes. 2 parents (2%) completely disagreed with this. 8 parents (7.8%) disagreed with this. 21 parents (20.6%) neither agreed, neither disagreed with this. More than half of parents, 53 (52%) parents agreed that programming should have priority at computer science classes. 18 parents (17.6%) completely agreed with this. Parents were asked do they think that programming should be elective subject. Just one parent completely disagreed with this. 8 parents (7.8%) disagreed with this. 21 parents (20.6%) neither agreed, neither disagreed with this.

More than half of parents, 54 parents (52.9%), agreed that programming should be the elective subject. 18 parents (17.6%) completely agreed with this. Parents were asked do they think that programming develops logical reasoning of children. None of the parents completely disagreed with this. 6 parents (5.9%) disagreed with this. 20 parents (19.6%) neither agreed, neither disagreed with this. More than half of parents, 56 parents (54.9%), agreed that programming develops logical reasoning. 20 parents (19.6%) completely agreed with this. Parents were asked do they think that programming should be obligatory subject. 5 parents (4.9%) completely disagreed with this. 8 parents (7.8%) disagreed with this. 26 parents (25.5%) neither agreed, neither disagreed with this. Almost half of parents, 42 parents (41.2%), agreed with that programming should be obligatory subject. 21 parents (20.6%) completely agreed with this. Fig. 7. shows parents' attitudes toward programming in general in elementary schools.

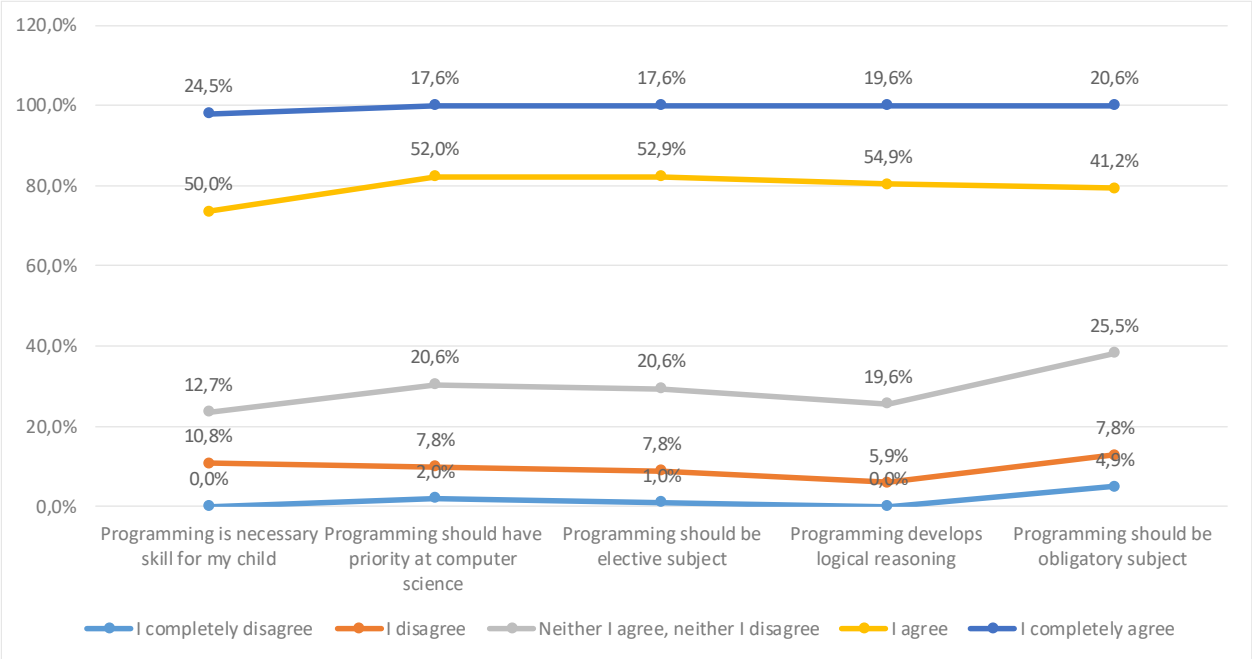


Figure 7. Parents' attitudes toward programming in general in elementary schools

IV. CONCLUSION

In this paper an online questionnaire in elementary schools in city of Osijek was conducted in purpose to explore parents' attitudes toward programming. According to the gained results it could be concluded that the more than half of parents agreed that programming is necessary skill for their child for further life. Furthermore, more than half of parents agreed that programming should have priority at computer science classes. Also, more than half of parents agreed that programming develops logical reasoning. Although, parents generally support programming more than half of them would like that programming be the elective subject and not to be obligatory. Regarding parents' attitudes toward programming with Micro: bit the majority of parents agreed that learning with Micro: bit is very interesting. Also, the majority of parents agreed that Micro: bit within computer science in school should be more employed. Parents graded the most important opportunities of Micro: bit technology in the following order: easiness of use, multidisciplinary, compatibility with other devices, speed of use and education for teachers and pupils. Among different programming tools the majority of parents are the most familiar with programming tool Scratch. Since the majority of parents' workplaces and jobs were not in relation to programming it could be concluded that their strong support toward programming within school or out of school activities does not depend upon their own professional orientation.

REFERENCES

[1] BBC Micro:Bit: Što je to i čemu služi?, https://pcchip.hr/ostalo/tech/bbc-microbit-sto-je-to-i-cemu-sluzi/, 17-1-2020

[2] Blockly, https://developers.google.com/blockly/guides/overview, 17-1-2020

[3] N. Bubica, M. Mladenović and I. Boljat. "Programiranje kao alat za razvoj apstraktnog mišljenja", https://bib.irb.hr/datoteka/702093.Programiranje_kao_alat_za_razvoj_apstraktnog_miljenja-CUC-zbornik.pdf, 17-1-2020

[4] Code.org, https://studio.code.org/courses, 17-1-2020

[5] Digital Technologies Hub. General-Purpose Programming/Coding, https://www.digitaltechnologieshub.edu.au/teachers/topics/general-purpose-programming-coding, 17-1-2020

[6] Gdevelop, https://gdevelop-app.com/, 17-1-2020

[7] Hopscotch, https://www.gethopscotch.com/, 17-1-2020

[8] Micro:bit, https://sciencescope.uk/microbit/,_17-1-2020

[9] Purpose of Programming Languages, https://www.freetext.org/Introduction_to_Computer_Programming/Introduction/Purpose_of_Programming_Languages/, 17-1-2020

[10] Pocket Code, https://share.catrob.at/app/, 17-1-2020

[11] Scratch, https://scratch.mit.edu/, 17-1-2020

[12] Scratch- Vizualni programski jezik za djecu, https://cuc.carnet.hr/2010/images/b1_52a14.pdf?dm_document_id=182&dm_%20dnl=1, 17-1-2020

Artificial Intelligence and Education

H. Jaakkola*, J. Henno **, A. Lahti***, J-P. Järvinen**** and J. Mäkelä****
* Tampere University, Pori, Finland
** Tallinn University of Technology, Tallinn, Estonia
*** University of Turku, Turku, Finland
**** University of Lapland, Rovaniemi, Finland
hannu.jaakkola@tuni.fi; jaak.henno@taltech.ee; {aleksi.lahti | jussi-pekka.jarvinen}@utu.fi; jumakela10@gmail.com

***Abstract* - Artificial Intelligence (AI) is one of the emerging technologies of today. It provides applications in addition to the traditional computing environments, and is also in affordable smart devices, making AI available everywhere. These affordable devices have built-in capabilities to handle complex computing tasks (edge computing), flexible access to fast network resources, access to cloud-based services to solve complex problems on a collaborative basis, and access to an enormous amount of open and closed data resources. In education, AI contributes in at least two ways: (1) the scope and content of education - what kind of education is needed; and (2) the process of education – support and changes to education and the teacher's work. In the scope of education, we must consider that AI (and related technologies) will replace some jobs (education will no longer be needed), some jobs will change dramatically (education content must be modified), and a lot of new jobs will be generated (new education must be established). In the process of education – the work itself - AI will have and has the role of reformer and enabler, which also changes the characteristics and division of duties. In this paper we analyze the opportunities and challenges caused by AI in education. Although the focus is on the role of AI, it is difficult to separate it from other technology-driven changes, especially in the discussion about work life.**

Keywords* - *Artificial Intelligence, Learning, Teaching, Machine Learning, Intelligent Systems, Computers in Education, Expert Systems

I. INTRODUCTION

A. Artificial Intelligence

Artificial Intelligence (AI) has grown in importance in the field of Information and Communication Technology (ICT) over recent decades. It is one of the "solid" application areas that has lasted from the 1950s until today, albeit manifested in varying forms. Our paper [6] separates *four sequential waves* in the progress of AI: AI in program code from the middle of the 1950s, AI in (rule based) Expert Systems from the late 1970s, AI based computer architectures from the mid-1980s, and *AI based on Machine Learning (ML)* from the middle of the 2000s. As is typical for new *radical technologies*, at the beginning they take the role of *change driver* but are slowly adopted widely by potential users and become part of "*the new normal*". Which is why the waves above have a start but no end – all of them still exist as part of everyday ICT. The programming languages Lisp and Prolog are used in special purpose applications and expert systems support decision-making in a wide variety of application areas (e.g., IBM's Watson in medicine), while specialized application-oriented computer architectures are now part of computing infrastructure.

AI (*fourth wave)* is undoubtedly one of the emerging technologies today. *Emerging technologies* can be described by the following properties: radical novelty, opportunities still largely unrealized, capable of changing the status quo, relatively fast growth in use. The *emergence phase* is somewhat uncertain and ambiguous from the user point of view. The first users of an innovation are *innovators* – users that are eager to test new ideas - *and early adopters*[1] - users that make their own decisions without waiting for the experiences of other users. These user groups are ready to carry and accept the risks of uncertainty so as to gain early benefits of the innovation. Technical progress finally convinces most of the potential users (followers: early majority, late majority, laggards), who accept new technologies in a more mature phase of their life cycle, based on the experiences of others (imitation).

In a way AI is already widespread in many application areas. In the *education* sector, we seem to be in the embryonic phase - a variety of trials exist, simple "commercial" applications are available, but it is not yet widely in general use. AI is somehow hidden and embedded in the more general trend, that of *computers and ICT in education*. The spread of AI to new application areas – also in the education sector - is slower than the progress of the technology (AI) itself, because the focus of the research and development (R&D) is on AI technologies, rather than its applications.

Traditionally, the term "AI" relates to *expert systems*, which are *rule based* applications that manage knowledge. AI based on ML handles and manages (properties of the) data using *learning algorithms*. In *supervised learning* it is *trained* to understand the basics of the problem area. During use it learns from mistakes and corrects its future behavior (*self-learning*). In supervised learning the results of the analysis are known in advance (by the trainer) and the learner (trainee, machine) makes the right or wrong decisions. What is critical for the learning results is the size of the *training data set.* Massive amounts of data are

[1] See E. Rogers, Diffusion of Innovations, 5th Edition, New York: Free Press, 2003.

needed to guarantee good learning results – e.g., masses of photos to train a system to find content elements in photos, a lot of live video material to identify people from the faces in it, etc. AI applying ML is based on learned content (handled by the algorithms of linear statistics), whereas a traditional expert system handles the data based on rules. A good example of the transfer from rule-based intelligence (expert system) to ML is the progress of Google Translator: the coded rule base of 0.5 MLOC (Million Lines of Code) was replaced by a learning algorithm of 500 LOC; the learning capability has improved the quality of translations dramatically, because the training data set is growing all the time along with use.

Learning can also be *unsupervised (autonomous)* [2], wherein no data set is available, and learning is fully based on algorithmic reasoning. The current AI is called *narrow / weak AI*, which means its applicability is in a *narrow specific application area.* The transition toward *general / strong AI* and unsupervised learning has started. The system understands facts and their relationships without the restrictions of a specific application area.

Although AI itself is not the topic of our paper, we wanted to provide this general overview of the progress and its basics to give the necessary background to understand the opportunities and challenges of AI, especially its limitations. More background is available in our last MIPRO paper [6], of which this paper is a continuation.

B. *Artificial intelligence in the society of today*

Thanks to the growing intelligence of ICT-based systems, the role of human work is changing. Automatization itself already has a long tradition: factories are partially robotized, computers do the boring routine work. Robotics is also being transferred to office work in the form of software robots (*Robotic Process Automation – RPA*)[3]. Although intelligent tools and applications support human activities in many ways, most of these are based on mechanical automation of (parts of) processes or replace human work in the mechanical stages with programmable machines and computer applications.

Advanced AI brings new opportunities for the use of ICT. Instead of mechanical automation it provides means for embedding *"human like" intelligence,* both in existing systems and in totally new application areas. The role of human work is changing: some tasks and jobs are *disappearing*, some are *changing* dramatically, and a lot of *new ones are being born*. Productivity is expected to rise with less pressure on humans. This will give rise to challenges in society – the education sector included.

C. *AI-supported education*

AI will change the education sector in two ways:

- First, *changes in work and the demand for it*: the education sector must be prepared for the accelerating changes in work life, concerning expertise in different professions and the demand for new work. Some jobs are disappearing, most will change, and a lot of new jobs will be generated.
- Second, changes *in education practices*: AI affects education itself. AI not only has the role of reformer and enabler but also a supportive role. It provides a means of re-dividing the workload. Examples of such activities in education are automation in the grading of examinations, analyzing (changing) students' needs in real-time, and pointing out places where courses need to be improved (based on student behavior).

AI can also support the transition to new learning paradigms – *flipped* learning, *time- and place-independent* learning and *blended (multiform)* learning. *Intelligent assistants* and *chatbots* are ready to help students independently of the time and this relieves the teacher from routine mentoring. Students' learning activities can be also supported by AI tutors, AI based feedback, etc.

The topic of *AI in education* is new. Rather than providing remarkable innovations in education practices, it is more a question of empirical trials and prototypes in *learning analytics*, *visualization* of data collected, and *statistical reasoning* based on it (Sedrakyan et al., [12]). Some learning platforms utilize machine learning but to a limited extent. The use of the term AI has become quite prevalent in the marketing speak of digital educational platforms, even though actual AI-based applications range from non-existent to very rare.

However, a lot of *ethical issues* must be considered in this context: the reliability of results, the validity of data to forecast the real progress of students, the ownership of the data collected and the right to use it, the context dependency of the algorithms used in the analysis, etc. Plenty of data is collected in the various systems that are integrated into daily teaching. One important question is what is the *teacher's ability (and will)* to utilize this data and to interpret it in the right way. Cultural differences play an important role – the difference of the goals and needs in the "culture of development" and the "culture of usage" might cause conflicts.

The use of AI at schools is a kind of by-product of general computerization and adoption of ICT-based applications / course materials / platforms. Its manifestation can be seen in the use of digital materials, which support learning-based adaptability in the students' progress. Data is collected, but teachers are not educated / trained to use this knowledge to improve their work. Whether or not they should be is another question. Of course, it should be the starting point of all efforts in this area.

D. *Research question and the structure of the paper*

The aim of this paper is to analyze the changes caused by AI in the current society, especially from the educational point of view. Above, we have introduced the topic and its background. We have a preconceived notion about the situation. The motivation to write this paper was to learn

[2] Good article to study on the role of ML in AI is: B. Marr, "Supervised -v- Unsupervised Machine Learning -- What's The Difference?". Retrieved from https://www.forbes.com/sites/bernardmarr/2017/03/16/supervised-v-unsupervised-machine-learning-whats-the-difference/#7e933555485d in February 12th, 2020.

[3] See e.g. https://en.wikipedia.org/wiki/Robotic_process_automation

more about AI and to assess the *state of the art*. We have set the following *research problem* for our paper:

What is the role of Artificial Intelligence in education now, and in the near future?

Two *sub-questions* can be derived from it:

1. What will the technology-driven changes mean for future jobs?
2. How does AI support and change teaching?

These questions also provide a structure to the paper. We return to the findings of our previous paper [6], which handled the progress of AI in the sequence of waves; the current paper is a continuum of it. The paper is structured in the following way. The findings discussed are based on a simple *literature review* conducted as a mapping study[4]. In *Section 2* we focus on the *technology driven changes* in *work*. *Section 3* handles *AI-driven changes in teaching*. *Section 4* concludes the paper and binds the different viewpoints together.

II. TECHNOLOGY-DRIVEN CHANGES IN WORK

In his books[5], *Yuval Noah Harari* has analyzed the changes of humankind from the historical, current, and future perspective. He separates the *physical* and *cognitive* side of a human being. The physical part can mainly be automatized and replaced by cognitive activities. Currently, cognitive tasks can also be automatized: humans solve problems by perceiving and reasoning; AI does the same faster by *handling big amounts of data* with statistical operations. Harari handles the changing role of the human being and the birth of societies. The latest book reviews the future (challenges) and the role of AI, robotics, and automatization.

The *World Economic Forum* published in 2018 contains an analysis of changes in work [21]. The report handles a wide set of viewpoints and aims to recognize changes, to be *proactive,* and to recognize the *driving forces* in the background (to be prepared). The report has a short-term focus on the changes – the floating five years, current (2018-2022) - but it also debates the *far-term* perspective until 2050. The report lists the *technologies* that companies have planned to adopt (and that will have an effect on job profiles): big data analytics, app- and web-enabled market, the Internet of Things, ML, cloud computing, digital trade, augmented and virtual reality, encryption, new materials, wearable electronics, distributed ledger (blockchain), 3D printing, autonomous transport, robots (stationary, non-humanoid land, humanoid, aerial and underwater), quantum computing, and biotechnology [21, p.7]. All these represent remarkable *opportunities to change* the characteristics of work. The list also provides a (narrow) vision of the *required technical skills* for planning education. In addition to the technology drivers behind the changes, the report lists "*soft skills*" for which there is a growing demand in work: analytical thinking and innovation, active learning, creativity, technology design and programming, critical thinking and analysis, complex problem-solving, leadership and social influence, emotional intelligence, reasoning, problem-solving and ideation, and systems analysis and evaluation. The report points out the changes in the job profile (over a period of five years): roles remain stable 48%, new roles 27%, redundant 21%, other 4% [21, p. 9]. These figures mean a lot of challenges to be solved in education: how to keep workers relevant in the changing job market. NB – the timespan is less than *five years*. There is a need not only to take into account the *needs of the change drivers* (of the time), but also to provide *permanent and reformable* skills, which are applicable in new contexts and situations (along with the changes in the drivers). These skills are often referred to as 21st century skills: these can be elaborated to be the most essential general skills needed in 21st century work life (Griffin & Care, [4]).

McKinsey & Company (Manyika et al.) report on the changes in the job market [8]: Close to 50% of current work activities can be automated by adopting currently demonstrated technologies; 60% of occupations have more than one third of activities that are automatable; only 5% of occupations consist of activities that can be fully automated. The results differ significantly by country and occupation, and also by demographic figures: developing countries have the highest demand for new workers and in highly developed countries the tendency is to automate. Professions in the "danger zone" (decrease / no growth) cover predictable and unpredictable physical work, office support work, and customer interaction. "Growth zone" professions are highly educated, for instance, care providers, managers and executives, teachers, and technology professionals. The report estimates that up to 800 million people could be displaced by automation worldwide and will need to find new jobs by 2030; 75 - 375 million of these workers may need to *switch occupational categories* and learn new skills. Despite the changes, the report is optimistic about the future: there will be enough work even in the future: the labor market will be able to adjust to the changes, as it has done throughout history. Automatization will decrease the amount of routine work, which means higher education of professionals. More time will be spent on activities that machines are less capable of doing, such as managing people, mentoring, applying expertise, and communicating with others. These tasks require more social and emotional skills and advanced cognitive capabilities, such as logical reasoning and creativity.

The Fast Future report [16] analyzed future jobs from 2010 to 2030. The "forecast" is based on an analysis of economic, political, socio-demographic, environmental, science and technology related changes in the period. Based on the *futurists'* analysis, it mentions twenty future professions, some of which are listed here: body part maker, old age wellness manager, memory augmentation surgeon, new science ethicist, vertical farmers, weather modification police, virtual lawyer, avatar manager, virtual teachers, waste data handler, social 'networking' worker,

[4] See e.g., K. Petersen, S. Vakkalanka, and L. Kuzniarz. "Guidelines for conducting systematic mapping studies in software engineering: An update". Information and Software Technology 64 (August 2015), 1–18. DOI:https://doi.org/10.1016/j.infsof.2015.03.007.

[5] Author: Y.N. Harari. Three books: Sapiens – A Brief History of Humankind; Homo Deus – A Brief History of Tomorrow; 21 Lessons for the 21st Century.

personal branders. In this context we will content ourselves with the list above; details can be found in the original report. Reality or not – we leave the conclusions up to the reader; at least some of these seem to exist already.

III. AI IN EDUCATION

A. Work Profile of Teachers

The study of McKinsey & Co (Bryant et al., [2]) takes a look at the teacher's work profile (in primary and secondary level) and analyzes the opportunities of AI to support teaching work. The demand for education professionals is growing (in the USA by 5-24% till 2030; in emergent societies the growth would be 100%). To fill the demand gap, the amount of teacher training should be increased and / or the productivity of existing training should be raised by automatization and focusing human work on essential activities instead of routines.

According to the study, 20 - 40 percent of teacher hours are spent on activities that could be automated. Teachers' weekly workload is approximated at 50 hours; less than half (49%; 24.5 h) of it is allocated to direct *interaction with students*. The rest is used for preparation (10.5 h), evaluation and feedback (6.5 h), professional development (3.0 h), and administration (5.0 h). The study concludes, based on the evaluation of existing technology and expert interviews, that technology could provide the potential to *reallocate 20-30% of the teacher's time* to activities that support students' learning. Further, the hours not related to direct student interaction could be halved.

B. AI-Driven Changes in Education

We started this study by making a simple mapping study using the keywords "*artificial intelligence in education*". The publication intensity proved to be high: Scobus over 3,000, Science Direct over 6,800 and Google Scholar over 123,000 references. A general Google Search provides 615,000 hits. We limited the search to the last 3+ years (2017-2020). Our aim was not to implement a systematic literature search, but to find enough well-focused texts (based on the expert analysis) to create a general overview about the state of the art in this topic. The sources for the following were selected on a *purpose basis* (as is typical for a mapping study).

Teachthought lists [18] ten roles of AI in education. The use of *AI in grading* is found to be a tool to decrease the amount of basic activities in education: traditional mechanical automated grading support (multiple choice, fill-in-the-blank) was not counted. Promising results are available in the use of *AI in text analysis*, although essay grading software is still seen to be in its infancy phase.

The current educational software provides a means to *adapt learning* to student needs. In *individualized learning,* the learning platform responds to the students' needs, puts focus on the topics needed by the student, repeats topics that are not yet mastered etc. Essentially the software differentiates the material at hand to meet the students' changing needs for increased challenges or need for support. Doing this with software requires a deep understanding of how and why learning happens. It presents a challenge for *technologically oriented* developers to bring *pedagogical expertise* to the development process. As a simple technical solution, learning platforms collect data. Using it needs a pedagogical approach – i.e., to recognize places where *courses must be improved* – places where masses of users fail. Do these two aspects (data collection, and its use) match each other easily or not?

Smart data collection (and analysis) – when done and utilized "in-large" – can be used in *development* and *evaluation* (comparison to others) by a school, and in *marketing* it to potential students. Collected data is also useful in planning study programs and allocating resources to activities.

Human tutoring work can be replaced by *AI tutors*. Some trials exist and good experience is available in the context of the fundamentals. AI-driven programs can give students and educators *feedback* – by combining students' and teachers' views of learning, based on the monitoring data collected from a longer period of the course or study program. The data also supports the improvement of the *user experience* of the system – how the users interact with it. This is similar to the use of many current commercial systems, which make recommendations based on the users' behavior and adapt it in their interaction process.

Teachthought sees AI as an opportunity for the teacher to focus on more essential work in teaching – AI (learning platform) provides the means for routines, and teachers will supplement lessons by providing students with hands-on support for students. AI is also seen as a "neutral" tool for *trial-and-error learning*. Students do not like to fail in front of human teachers, but in the use of a learning platform, the situation is quite different.

The survey of Acer Education technology [1] quotes the findings reported by the *Stanford University* study panel related to the future of AI [15]. The report covers the progress of AI widely and lists some key findings of the report. The transfer toward *online education* is becoming a reality in all levels of education and allows the facilitation of more *customizable* approaches to consider students' individual needs in learning. The innovations cover learning analytics, personalized education, and identification of the educational needs of the students; IoT and cloud technologies support peer-based and AI-tutor mentoring. Textbooks and course material can be customized and learning resources are accessible from multiple devices (cloud).

As the students learn, so will the online education systems. Learning analytics will accelerate the development of new tools for personalized education and provide teachers with better insight of their students' performance. Robot teachers (physical, software) are not alternatives to a human teacher, but would have an assisting role to give humans the opportunity to focus on the most essential tasks. Teaching and the work of a teacher will remain; social interaction still requires authentic human skills.

The Stanford report [15] mentioned above handles the topic on a more concrete level. It refers to the history of robots as educational devices (Lego Mindstorm Kit from the 1980s). The spread of AI technologies has been slow

due to the lack of resources (funding, skilled teachers), but also because of the missing evidence of their benefits in learning. *Intelligent Tutoring Systems (ITS)* are transitioning from laboratory experiments to real use. The report mentions *Carnegie Speech* (spoken language assessment and training software to improve voice and accent training) and *Duolingo* (foreign language training) using *Automatic Speech Recognition*. Other ITS products cover *Carnegie Cognitive Tutor* (used in teaching mathematics), *SHERLOCK82* (to be used to teach Air Force technicians to diagnose problems with electrical systems) and some avatar tools (training people to deal with people from different cultural backgrounds).

The fast-growing role of Massive Open Online Courses (MOOCs) provide a means for synchronous and asynchronous education and a platform to apply adaptive learning tools. New opportunities are provided especially in professional and life-long learning, with lesser importance placed on face-to-face interaction. *Learning analytics* plays an important role in adapting these platforms for the needs of students, teachers, and schools. It applies deep learning, natural language processing, and other AI techniques to analyze the student's engagement, behavior, and outcomes. MOOCs when used by masses of students provide enough monitoring data to make the analysis reliable.

The Stanford report lists some additional future opportunities: human teachers will be assisted by AI technologies; the use of Virtual Reality (VR) in teaching, MOOCs and other forms of online education will become part of learning at all levels; the transition from hard copy books to digital and audio media and texts will continue; digital reading devices will become 'smarter' and provide easy access to additional information about the subject matter; and Machine Translation (MT) technology will make it easier to translate educational material into different languages.

The Forbes article by Marr [10] highlights the future changes in the work of teachers: "Even though most experts believe the critical presence of teachers is irreplaceable, there will be many changes to a teacher's job and to educational best practices." Educational solutions of AI are starting to mature and are ready to support teachers' work even in practice. The same factors as discussed above are pointed out: personalization, adaptability, automatization of admin. tasks, personalization, the use of digital platforms (AI-supported learning, testing, and feedback to students). New tools may support learning for the (seeing, hearing) impaired and people speaking foreign languages. Universal access to the classroom supports remote participation in the classes.

The article by Schmelzer [13] (Forbes) deals with AI applications in education. It introduces a new term "*AI-Enabled Hyperpersonalization*" to emphasize the power of AI and ML to tailor and personalize the learning environment and opportunities. In the middle of the 2020s, close to half of the learning management tools will have embedded AI capabilities. Instead of creating a single curriculum for a group of students, educators will have intelligent assistance to provide a wide range of materials for the same core curriculum to meet individual needs. The business volume of such tools is expected to reach over $6 (US) billion by the middle of the 2020s. Studies will become paperless and the learning platform will provide a means for direct interactive feedback and follow-up. Augmented tutoring provided by autonomous agents will support learning and mentoring. Intelligent assistants will enhance adaptive learning so that each of the students can learn at their own pace or time frame.

In comparison to another field, *medicine*, there the expected business volume is expected to rise to 36.15 (US) billion, with a CAGR of 52% (Compound Annual Growth Rate) in the forecast period of 2018-2025 (MarketsandMarkets, [9]). The most important AI technologies are ML, natural language processing, context-aware computing, computer vision, and querying methods. The driving force for adopting AI in healthcare is the rising need for services due to the imbalance between the health care professionals and patients; the same – a lack of skilled teachers – applies to education, too. These frontier applications are the same as in the AI market[6] in general, in which the annual CAGR is calculated to be 36.6%. In the education sector, the expected CAGR is 44.2%[7].

The paper by Kulkarni [7] brings a couple of interesting viewpoints to the discussion. It introduces the terms *AI-Led* and *AI-Assisted* (*AI-Powered*) education and concludes – as do many others – that the future lies in *AI-Assisted education*. The main responsibility still falls on a human teacher. The paper discusses the current state and lists some successful implementations: Duolingo[8] (referring to the claim that "34 hours on Duolingo's app are equivalent to a full university semester of language education") and McGraw Hill's ALEKS[9] as a web-based, AI-powered assessment and learning system. In spite of the main development effort being in the USA, *China* is catching up (and overtaking). China's current investments in AI are already slightly higher than the USA's. China, having a population of 1.4 billion, fast growth of education needs, and a lack of skilled teachers, will be in the vanguard of the serious use of AI in education. The importance of education in Chinese culture has deep roots; Chinese families spend 20-26% of their income on their children's education. Companies like Squirrel AI (collaborating with CMU) and ALO7[10] are mentioned as the leading organizations.

The paper by Shroer [14] lists twelve American companies that have AI-based *products* for use in education. The application areas cover speech recognition and transcription for students having a limited writing ability (*Nuance*), adaptive learning technology (*Knewton*), virtual learning assistant based on conversational interaction (*Cognii*), customizable STEM (Science, Technology, Engineering, Mathematics) tutoring lessons (*Querium*), personalized learning plans and reduced

[6] https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-market-74851580.html

[7] https://marketersmedia.com/new-report-2019-artificial-intelligence-ai-in-education-market-to-grow-at-a-tremendous-cagr-of-44-2/511172

[8] https://www.duolingo.com/

[9] McGraw Hill, Assessment and LEarning in Knowledge Spaces, based on the Knowledge Space Theory: https://www.aleks.com/about_aleks

[10] http://squirrelai.com/; https://www.alo7.com/en/

workloads for instructors utilizing cognitive neuroscience and data analytics (*Century Tech*), a speech-to-text tool designed for small children (*Kidsense*), AI and machine learning to help students to develop a deeper conceptual understanding of math (*Carnegie Learning*), an adaptive learning platform to help educational institutions to collect data and to increase learner's engagement (*Kidaptive*), combined computer vision intelligence and augmented reality to enhance the way students learn in the classroom (*Blippar*), a customized tutoring program that tracks the work step-by-step (*Thinkster Math*), Knowledge Engine to synthesize course and quiz results and briefings to find knowledge gaps among employees in enterprises (*Volley*), and adaptive studying and learning tools (*Quizlet*). These confirm our earlier findings in the mainstream of the current AI technologies used in education – mainly for monitoring and adaptation purposes.

At the end of this review we make three references [11; 19; 20] to reports by UNESCO, without going into details. UNESCO has handled the *role of AI in education* from two different points of view. The first report deals with the role of AI in education from three perspectives: *AI for* education, *AI as* education, and *AI in* education and examines activities to implement the principles of inclusive and equitable participation in "AI and education". The two other reports focus on the role of AI in sustainable development.

To end this sub-section, we will give a brief overview of an ambitious Finnish project - *DigiTala*[11], as an example of a practical effort to use AI in education. *DigiTala* is an interdisciplinary (pedagogical, phonetics, behavioral, signal processing) project with the goal of developing a tool for computer-aided language tests. It will be used to grade the skills of spoken foreign language at the end of general upper secondary education. Most of these (written) tests are already computerized. Currently, the oral language part is not yet included in the set of tests. In other contexts, these are based on the evaluation of human assessors. The goal is to include a computerized spoken language test in the final examination and to provide an objective (human-assessor independent) spoken language evaluation.

IV. CONCLUSIONS

In this paper we have reviewed factors related to the role of AI in education. Partially we had to extend the approach to cover other technical innovations – "*technology driven changes"* in education, because AI cannot be separated from it. Two aspects were considered: the role of technology-driven changes in *work life*, and the role of AI (and related technologies) in *education*.

Work is undergoing the pressure of rapid changes. Demand for higher productivity and the opportunities provided by new technologies are driving industry and society towards automatization. This means radical changes in human work: some tasks will disappear; most are changing, and some new ones will be born. From a historical perspective, it is easy to see that society has always been able to adapt to change and (close to) full employment in society has been the status quo. What will happen in the future? Re-education, continuing education, and life-long learning are needed to adapt to the changes. The socio-economic structure is changing fast: high-welfare societies are aging and fewer people of working age have to take care of a larger number of the older generation, who are also living longer than earlier. Automatization and robotics as replacement and support of human work would be more than welcome in this structural change. In contrast, in lower welfare societies, which suffer low employment, the birth rate is high, and people tend to have a lower level of education. If these people are to be used to replace aging workers, a new kind of education is needed.

In education AI has a growing role, too. The mapping study of publications indicates that the topic is an important one, and a lot of work is going on in this area. However, research work on AI itself is being done in different silos than its applications. This means that the results are not being fully exploited in a variety of application areas – especially not in the education sector. AI application related research is concentrated in "*money-rich*" areas – education sector is not one, but medicine for instance seems to be. Our survey, covering some relevant articles, shows that future opportunities are envisioned, but real breakthrough applications are still waiting for their time to come. The article by Bryant et al. [2] reported the difficulties in adopting new technologies in teaching: the benefits of technology-rich, personalized, blended learning pilots are not realized on full scale; students who use tablets, laptops, and e-readers in the classroom are performing worse than those who do not. The explanation for this is that adoption of new technology transfers the focus *from the subject to learning to use the technology* itself and the teachers' time is spent acting as a facilitator and coach in using the new technology rather than on the learning interaction.

Are teachers ready to use these new technologies? Some are, but the majority not at all. According to the Deloitte study [7], teachers in European schools are still using ICT in classrooms to a diverse extent. On EU level, the range of ICT intensity in lessons by teachers is polarized. Roughly half of the teachers reported using ICT in classrooms in more than 50 % of the lessons. At the same time, nearly half of the teachers reported using ICT in the classroom rarely, in less than 50 % of the lessons. Teachers' perceptions of pedagogy-related obstacles to using ICT in teaching relate to insufficient pedagogical support and to a lack of pedagogical models on how to use ICT for learning, not to a lack of learning platforms or devices. One example of this complex situation can be given from the Finnish school system, which is known for its highly educated teachers. According to the report by Tanhua-Piiroinen et al. [17], over 50 % of Finnish teachers reported to have only "basic skills in ICT" and 10 % reported deficiencies in their ICT skills. The challenges in applying AI to education is not only about the algorithms and platforms but about the *capabilities possessed by the teachers*. Time and money for teachers' continuing education is required.

The reported AI-based tools mainly cover algorithmic intelligence, rather than the current AI with ML on a wide scale. Key applications seem to be based on algorithmic reasoning (like expert systems), data mining, big data type

[11] https://blogs.helsinki.fi/digitala-projekti/about-digitala/

analysis and visualization, not on a wide scale in system intelligence and its learning capability. The support provided to teachers mainly includes decision support, not automatic decision making. Grading, adaptation to the students' needs (often only adaptive user experience[12]), the use of electronic study materials, bidirectional feedback collection and analysis, and assistance "robots" are the main applications. The real use of ML is still minor: natural language processing (speech recognition, speech synthesis, natural language processing and analysis) has found its way into real applications but these are culture- and language-sensitive, not global; minority languages will still have to wait for a long time to convert these opportunities into practice. Progress in machine translation will help the situation (in the future) but does not solve the problem that education itself is culture-sensitive, having *different values and goals in different cultures*. The transition towards the use of cloud-based systems and universal learning platforms, MOOCs, especially in professional and continuing education as well as in regular education, seems to be clear. Because these systems are used by masses of people, even the follow-up data (because of the large amount) is useful on an ML basis to really make the platform AI-based.

What does the future look like through the eyes of experts in the field? We will close this paper by quoting (albeit loosely) the opinions of 17 experts, based on the interviews of Disruptor in [3]. These *expert opinions* reconfirm our findings: AI will assist instructors in performing tasks that are mostly mechanical (grading with helpful feedback, or by identifying students having difficulties). Students will receive increasingly complex support for their study problems. AI can help students to personalize their learning experience and it tracks the student's performance, enabling the students' needs to be addressed. AI will play the role of a teaching assistant (chatbots to address student questions between lectures). AI will be implemented to improve online and remote education. AI is a component in larger systems - combinations of AI with other technologies such as augmented reality. AI will make education accessible to all sectors of the population – especially students with disabilities. Education will be redefined as continuous learning instead of discrete curricula-based courses. All learning in basic areas will be managed by AI, with "teachers" replaced by facilitators and behavior managers. With the integration of IoT and 5G, classrooms will be replaced by virtual classrooms. Most students complete AI-supported flipped classroom and blended learning based programs virtually, and more in-depth studies on-site. AI is an aid for teachers in repetitive and data-driven tasks. Education is going to be AI-enhanced, but AI will not replace teachers completely, nor will all student learning become remote.

In an interview (Tekniikka&Talous 21.2.2020) *Professor Yutaka Matsuo* from Tokyo University pointed out that AI is not a threat to humans. Soon – within 5-10 years – AI will understand the semantics of words and moving robots will learn to manage their own movements. Simple tasks based on image processing (machine vision), speech recognition and natural language understanding will become possible. The current "intelligent assistants" – Siri, Alexa, Google assistant – just imitate something they have learned, without understanding it. Human thinking is based on two layers: a symbolic system and a system, which collaborate without our ability to explain it. Because we are not able to explain human behavior, it is difficult to transfer it to AI.

AI as a technology is already available and it is being adopted as its maturity reaches the expectations of the users. The future is already here to a certain extent, but there is more to come!

REFERENCES

[1] Acer for Education, "Is Artificial Intelligence transforming Education?", 2017. Retrieved from https://eu-acerforeducation.acer.com/innovative-technologies/is-artificial-intelligence-transforming-education/?gclid=CjwKCAjw4LfkBRBDEiwAc2DSlJ3KrOsrCemd5ZMS9oRydHgYtpKb4ym8kCvI1146dH7hhvK1056UwBoCR7kQAvD_BwE on February 19th, 2020.

[2] J. Bryant, C. Heitz, S. Sanghvi and D. Wagle, "How artificial intelligence will impact K-12 teachers", McKinsey & Company, 2017. Retrieved from https://www.mckinsey.com/industries/social-sector/our-insights/how-artificial-intelligence-will-impact-k-12-teachers?cid=soc-web&fbclid=IwAR0Qbb76ybBchgub_2wCxI5wD-_B9T-IB9Vg7dEbU4P2QGTfuqkhyoWsXWg on February 14th, 2020.

[3] Disruptor, What's The Future Of AI In Education? 17 Experts Share Their Insights. 2019, Retrieved from https://www.disruptordaily.com/future-of-ai-education/ on February 24th, 2020.

[4] P. Griffin, E. Care, Assessment and Teaching of 21st Century Skills: Methods and Approach. 2015. Springer. DOI: 0.1007/978-94-017-9395-7.

[5] Ipsos Mori,. 2nd survey of schools ICT in education: Objective 1: Benchmark progress in ICT in schools. European Commission Study by Deloitte, 2019. DOI: 10.2759/23401. Retrieved from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=57894 on March 6th, 2020

[6] H. Jaakkola, J. Henno, J. Mäkelä, and B. Thalheim, "Artificial intelligence yesterday, today and tomorrow," in 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2019 - Proceedings, K. Skala, Ed. Opatija, Croatia, May 2019, IEEE and MIPRO, 2019, pp. 985-992.

[7] A. Kulkarni, AI in Education: Where is It Now and What is the Future?, Lexalytics, 2019, Retrieved from https://www.lexalytics.com/lexablog/ai-in-education-present-future-ethics on February 24th, 2020.

[8] J. Manyika, S. Lund, M. Chui, J. Bughin, J. Woetzel, P. Batra, R. Ko and S. Sang, "Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages", McKinsey Global Institute. 2017. Retrieved from https://www.mckinsey.com/featured-insights/future-of-work/jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages on February 17th, 2020.

[9] MarketsandMarkets, Artificial Intelligence in Healthcare Market by Offering (Hardware, Software, Services), Technology (Machine Learning, NLP, Context-Aware Computing, Computer Vision), End-Use Application, End User, and Geography – Global Forecast to 2025. Retrieved from https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-healthcare-market-54679303.html on February 24th, 2020.

[10] B. Marr, "How Is AI Used In Education -- Real World Examples Of Today And A Peek Into The Future", Forbes, 2018. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/07/25/how-

[12] See e.g., Jaakkola, B. Thalheim, J. Henno, J. Mäkelä, and H. Keto, "Role of the User in Information Systems Development," in MIPRO 2018 - Proceedings of the 41st International Convention, K. Skala, Ed. GRAFIK, Rijeka: Mipro and IEEE, 2018, pp. 701-709.

is-ai-used-in-education-real-world-examples-of-today-and-a-peek-into-the-future/amp/ on February 19th, 2020.

[11] F. Pedró, M. Subosa, A. Rivas and P. Valverde, Artificial intelligence in education: challenges and opportunities for sustainable development, Document ED-2019/WS/8, Paris: UNESCO, 2019. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000366994?posInSet=1&queryId=2c58492c-931f-47d 2-967d-7194427f4062 on February 19th, 2020.

[12] G. Sedrakyan, J. Malmberg, K. Verbert, S. Järvelä and P.A. Kirschner, "Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation". Computers in Human Behavior. In Press, Corrected Proof. 2018. https://doi.org/10.1016/j.chb.2018.05.004.

[13] R. Schmelzer, "AI Applications In Education", Forbes, 2019. Retrieved from https://www.forbes.com/sites/cognitiveworld/2019/07/12/ai-applications-in-education/#4773a4a762a3 on February 19th, 2020.

[14] A. Schroer, 12 companies using AI in education to enhance the classroom, Builin, 2018. Retreived from https://builtin.com/artificial-intelligence/ai-in-education on February 24th, 2020.

[15] Standing Committee of the One Hundred Year Study of Artificial Intelligence, Artificial Inelligence and Life in 2030: One hundred year study of artificial intelligence, Report of the 2015 Study Panel, Stanford University, September 2016. Retrieved from https://ai100.stanford.edu/sites/g/files/sbiybj9861/f/ai_100_report_0906fnlc_single.pdf on February 19th, 2020.

[16] R. Talwar and T. Hancock, The shape of jobs to come - Possible New Careers Emerging from Advances in Science and Technology (2010 – 2030), Fast Future Research Report, 2010. Retrieved from https://fastfuture.com/shop/shape-of-jobs/ on February 17th, 2020.

[17] E. Tanhua-Piiroinen, S.S. Kaarakainen, M.T. Kaarakainen, J. Viteli, A. Syvänen, and A. Kivinen, Digiajan peruskoulu. Valtioneuvoston kanslia, Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 6/2019 , 2019. Retreived from http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161383/6-2019-Digiajan%20peruskoulu_.pdf on March 6th, 2020.

[18] Teachthought, "10 Roles For Artificial Intelligence In Education. 2018". Retrived from https://www.teachthought.com/the-future-of-learning/10-roles-for-artificial-intelligence-in-education/ on February 20th, 2020.

[19] UNESCO, Artificial intelligence in education, compendium of promising initiatives: Mobile Learning Week 2019, Document ED-2019/WS/38, Paris: UNESCO, 2019. Retrived from https://unesdoc.unesco.org/ark:/48223/pf0000370307 on February 19th, 2020.

[20] UNESCO, Artificial intelligence for Sustainable Development: synthesis report, Mobile Learning Week 2019. Document ED-2019/WS/39. Paris: UNESCO, 2019. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000370308?posInSet=2&queryId=2c58492c-931f-47d2-967d-7194427f4062 on February 19th, 2020.

[21] World Economic Forum, The Future of Jobs Report 2018, Geneva Switzerland 2018. Retrieved from http://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf on February 17th, 2020.

Teaching physics using programming of simulations

Robert Repnik[1,2,*], Peter Bernad[1] and Marjan Krašna[3]
[1] University of Maribor, Faculty of Natural Sciences and Mathematics, Physics Department, Koroska 160, SI-2000 Maribor, Slovenia
[2] Association for Technical Culture of Slovenia, Zaloska c. 65, 1000 Ljubljana, Slovenia
[3] University of Maribor, Faculty of Arts, Department of Pedagogy, Koroska 160, SI-2000 Maribor, Slovenia
* corresponding author: robert.repnik@um.si

***Abstract* - It has long been agreed that the best physics teaching is by observing the experiment. But in some cases, experiments are not feasible for different reasons (time of the experiment is too short or too long to observe, experiment can be dangerous or extremely expensive). For counted reasons we use simulations which are computer programs that simulate real environment with various degree of reality. Simulations are used to hide the underground complex physics from the less skilled physics students who can observe and analyse the natural phenomena even without knowing all physics background. But we wanted to try something else – teach physics through gradual building of more and more complex simulation – hence teaching physics through programming. In our article we are going to present the scenario how to use such approach, and the analysis of two programming languages Python and Visual Basic to see which is more appropriate for developing visual simulations for students.**

Keywords – education; simulation, physics; programming; cross-curricular approach

I INTRODUCTION

In our country, we see decline of students of STEM and particularly of physics for many years. The statistics also show the similar trends in different countries around the world [1]. It is interesting that we see such trends in the tertiary education hence the primary school students still like physics. It is hard to understand the reasons why physics becomes less popular after secondary education. In the same time, the IT is still highly valued among students. Could we do something to join the two disciplines and make the best out of them? A long time ago [2] the physics teachers have find out that mathematical models (even simplified) are not popular among students. Learning complex formulae for different physics phenomena did not help to the popularity of physics. Even some physics teachers may be blamed for contemporary decline of physics popularity. They were using complex mathematical tools beyond the knowledge of students and they shine in their superiority. Despite the fact that such teachers are nowadays rarity in the education, the horror of physics they seed have remained in the mind of generations of parents and grandparents and the influence is ever-present.

The concept of hiding complex calculations behind the simulations emerge at the turn of last century. Many small simulations were made (named physlets) that help students to understand the influence of parameters in the physics phenomena without the need to understand underlying mathematics. More or less complex physics are hidden in computer games that could be used as gamification of learning [3]. The advancement was visible almost in no time, the understanding of the fundaments of physics have grown but the low interest for studying physics have remained.

To raise the interest of physics education, especially in secondary school, students can be involved in different interdisciplinary projects [4]. The knowledge they gain from such projects is easily applied even outside of the school. Project based learning have multiple benefits: (1) the students achieve higher education with practical use and, (2) the teachers improve on their teaching methods as they adapt to new circumstances by mentoring in various projects. This didactics strategy requires a lot of effort, but the gain is not easily seen. The additional analysis of the entire TLS (teaching-studying-learning) processes, where we can grade the students' progress, is also needed. For the STEM students, especially physics and computer science, it is crucial to analyse the level of science literacy and the achieved digital competences through the process of learning. This cannot be achieved without clear guidelines [5]. The support to develop higher knowledge and skills is of the utmost importance. Teacher-to-teacher collaboration and the use of innovative learning methods continues also at the university [6]. Therefore, project based learning gives opportunities for the lifelong learning and the acquired knowledge usable outside of school, can improve students' quality of education and the life itself.

In the physics education, we often use simplified formulae to describe natural phenomena. The reality is much more complex. If we try to be more accurate we would need mathematical tools beyond the understanding of students. Therefore, the process of learning is to go from the simple formulae to more complex; and gradually ever closer to the reality. The formulae is therefore mathematical model that describes the natural phenomena up to certain degree of accuracy. Yet the formulae are just dead letters on the paper; complex reasoning is also required to understand what they represent. If we transform these formulae into the computer program, we

create simulation that transforms them into a meaningful tool. Possibility to change the parameters and run the simulation gives us the opportunity to observe parameters' effect in the simulation. Would the development of the simulation give us better results of understanding both, formulae and programming? Potential benefits are worth exploring this approach.

Cross-curricular assignments are extremely valuable approach to join otherwise sterile courses into life like processes. Such assignments give the purpose where and how to use knowledge and create highly prized competences. We cannot deny that the use of technology is crucial in scientific research and physics is no exception. Using digital sensors and electronic devices in experimental work is a common practice, yet the main interest in this article is the use of programming with the intent to learn the physics by creating simulation. The question is, how much programming experience do physics students get through their first years of university?

Comparing various physics programs of different universities we could see that students have enough computer science background to design and programme their simulations:

- The Karl-Franzens-University in Austria offers courses, where students acquire competences for database analysis, and are especially involved in computer sciences. One of the courses in their curriculum is dedicated to learning programming with the computer languages MATHEMATICA and C++ [7].
- The programming language C++ is used as the basic tool for coding at the courses Introduction to Programming and Computing for Scientists at the Department of physics at the Lund University in Sweden [8].
- The department of Physics at the University of Oxford includes a course on computer programming and numerical methods [9].
- Upon inspecting further universities in other countries in Europe (Hungary [10], Germany [11], Finland [12], Croatia [13]), each study program of physics includes a course, where students learn some basics of coding. It is no difference in Slovenia [14]; students of physics also have course of introduction of programming [15].

Many physicists have therefore learned some programming basics at some point of their education. Mostly in special courses or as a part of practical work. A debate could appear, if the use of programming inside of some individual courses is enough or if the use of programming should be present as a part of practical work throughout all semesters of the education. If programming would be used throughout the physics education students would have programming skills embedded for life.

In the approach presented in this article, we have prepared the scenario how to teach our physics teachers to develop simulations. Starting with simple simulation gradual increasing complexity we can improve the accuracy of the simulations in two programming languages: Python and Visual Basic (both freeware). Later we are going to have a workshop on how to develop physics simulations with the secondary and primary school students in the cross-curricular assignments.

This article is structured in five different but connected topics. The topic about "*Designing mathematical models for physics phenomena*" explains how we can start from a simple mathematical model of physics phenomena that explains the real-life phenomena in general. Later we add complexity that is suitable for more advanced analysis of the phenomena as learner's knowledge and mathematical skills progress through the educational levers. The topic "*Coding mathematical models*" gives us the background how to prepare computer simulation and how we can progress from simple to more complex simulations. The topic "*Comparison of the two programming languages in the view of the physics teacher*" compares two computer programming languages that are free and can be used in the primary and secondary level educational institutions. Python vs Visual Basic comparison is presented in this comparison on simple physics phenomena of a free fall and adding the horizontal wind parameter. "*Preliminary results*" topic covers students of educational physics using our approach to the physics simulations. Students of educational physics learn how to do the coding and learn physics but this assignment joins both topics together in shows how we can rise complexity level from simple to more complex. The topic "*Blueprint for the secondary schools test*" presents a scenario how to use this approach in secondary schools in the cross-curricular activities.

II DESIGNING MATHEMATICAL MODELS FOR PHYSICS PHENOMENA

In the physics education we all know that formulae that student need to learn are progressively more complex depending on the levels of education. In the beginning, we use simple formulae that roughly describes phenomena, but it does not correspond perfectly with the measurements. The differences are often described as measurement's errors, which is only one part of the truth. The mathematical models often use simplification and some assumptions that are not present in the real world.

In our case, we decide to start with the movement (projectile motion) because the mechanics is typically the first topic that students start learning in physics. Progressively this problem of (projectile motion) becomes more and more complex. We can start with the drop of the perfect ball to the ground and consider only the effect of gravity to the ball. To enhance the model we add the starting velocity (and angle). In reality we have wind therefore, we can add the wind speed (constant or changing) and wind's direction (velocity). Adding the air resistance (with constant or increasing air pressure with decreasing height) further complicate the mathematical model. Depending on location on the Earth, we need to add also the Coriolis force and so on. In the real world, we do not have perfect ball, any projectile have its own shape and other properties of the ball and the environment also influence the calculations.

We can therefore start with the simple beginning and add complexity to it as desired.

III CODING MATHEMATICAL MODELS

To make a computer simulation we can have two approach: fixed (constant) time interval or variable time interval. Since we wanted to show the movement of the object on the screen, we used fixed time interval simulation approach. Therefore we have to have all formulae time dependent and when we move through the time we calculate new position based on initial given parameters and previous calculations.

Therefore, we can summarize the simulation in the simple pseudocode (Figure 1).

```
FROM time=0 TO max_time
  x = calculate_x(time)
  y = calculate_y(time)
  draw_object(x, y)
NEXT time
```

Figure 1: Simulation's pseudocode

In the coding, we can use different approaches to model physics phenomena. We can use different mathematical formulae for each increase of complexity or we can use functional approach where we use simple functions for primary calculation and combine them in new function. We wanted to make program simple for understanding despite adding some complexity to the coding. Our idea is to make use of function and to achieve higher level of complexity of the model (Figure 2).

```
x = calculate_x(time)
x = add_wind(calculate_x(time),windparam, time)
x = add_resistance(calculate_x(time),resistanceparam, time)
x = add_resistance(add_wind(calculate_x(time),windparam,
time), resistanceparam, time)
```

Figure 2: Functional approach to the different complexity levels

In this approach, each function provides results that can be used as input to another function. In opposition of this approach, we would need different function for each combination. The results may be more structured from mathematical point of view, but it could confuse the learner.

For simple projectile motion, we have two formulae to work with (Figure 3).

$$x(t) = v*t$$
$$y(t) = g*t^2/2$$

Figure 3: Projectile motion basic functions

From the (Figure 3), we can calculate how long we need to perform the simulation by calculating when $y(t)$ equals maximum height. A bit later the simulation actually stops. We just need to decide how many steps we are going to need to finish the simulation. For the purpose of computer graphics, we need to calculate ratio between the simulation height and application window height. This ensures that the drawings stay inside the open window $y(t)$. We cannot assume the same for the $x(t)$ because we would need to restrict input value of the velocity in order not to get outside of the drawing area and this would be a little beyond the scope of our task.

IV COMPARISON OF THE TWO PROGRAMMING LANGUAGES IN THE VIEW OF THE PHYSICS TEACHER

A. Python

Since Python is a general-purpose programming language, several libraries are available that can contribute to creating a simple simulation application. With a variety of libraries, we need to choose which one would be the most beginner-friendly since the teaching of physics through programming is not aimed towards students with primary expertise in computer science.

The goal is to make the implementation of the physics behind the simulation application possible for a physics student with no advanced knowledge of coding required. Yet we can still achieve some basic algorithmic thinking without losing the focus on the physics background.

Upon inspecting some libraries/tools[1,2], the library turtle offers simple functions that can reveal how the projectile moves with time and trough space when it is influenced by initial speed, gravity and other factors later on. The functions are similar to the educational programming language Logo[3] that was developed in year 1966 and is aimed towards kids.

The previous mentioned development of the simulation application for projectile motion (Figure 4) consists of calculating the new position for each new frame/iteration that changes through time and converting the values into pixel values to display it correctly on the screen.

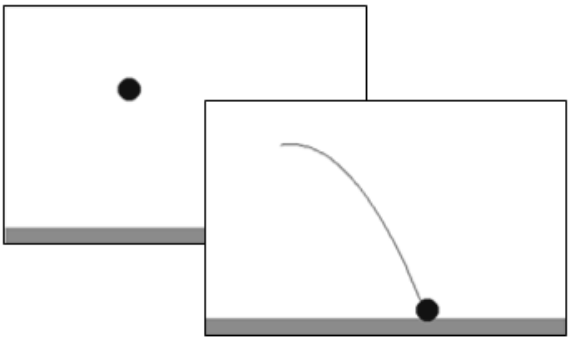


Figure 4: the blue circle represents the projectile marking the trajectory with a blue curve.

To understand the trajectory of the projectile motion, the movement in the horizontal direction has to be calculated independent from vertical. Functions are defined to return the new position in each axis on the screen (Figure 5). The number of functions might seem large, but the problem is divided into smaller ones. The main functions are calculate_Y() and calculate_X() which return new positions of the projectile for the current frame.

[1] https://vpython.org/, https://docs.python.org/2/library/tkinter.html
[2] https://docs.python.org/3.3/library/turtle.html?highlight=turtle
[3] https://en.wikipedia.org/wiki/Logo_(programming_language)

```
def return_pixel_value(value):
    return value - 290

def new_height(old_height, vy, dt, a):
    return old_height + vy * dt - 0.5 /
        / * a * dt * dt

def calculate_Y():
    global free_fall, start
    global current_height
    global current_Y_velocity
    global dt, g, t, t_previous
    global h_max, max_y_velocity

    if free_fall == True:

        h_new = new_height( /
        /current_height, /
        /current_Y_velocity,/
        dt, g)

        # the ball touches the ground
        if h_new < 0:
            t = t + dt
            free_fall = False
            t_previous = t + dt
            current_height = 0
        else:  # the ball is in the air
            t = t + dt
            current_Y_velocity = /
            /current_Y_velocity - g * dt
            current_height = h_new

    return return_pixel_value(current_height)

def calculate_X():
    global ball
    global velocity_X_start, dt, a_v, t, xball
```

Figure 5: The whole process of defining the functions is writing python code without any graphical interface.

In the function calculate_X(), we can recognize a basic equation for calculating the new position depending on the velocity and an acceleration (if there is any wind e.g.). The function calculate_Y() also includes some aspects, e.g. if the projectile is either rising or falling and if it touches the ground. Nevertheless, inside of the previously mentioned function, there is a call of the function new_height(), which also includes the basic function of the accelerated function that returns the new height of the projectile in the current frame.

Of course, we need to predefine initial values of the variables first for all those functions to work properly (Figure 6):

```
#the acceleration of the wind
a_v = 0
#gravity
g = 5
#time
t = 0
#time interval
dt = 0.03
t_previous = 0
#size of the projectile
size = 1
#the start coordinates of the ball
xball = -600
yball = 290
#conversion of the pixel coordinate
#to the height that we calculate
h_max = yball + 290
#newly calculated height in each iteration
h_new = 0
current_height = 0
#the maximum velocity in the vertical axis
max_y_velocity = math.sqrt(2 * h_max * g)
#stat velocity of the vertical axis
velocity_Y_start = 5
#current velocity in the vertical axis for
#calculating the new height in each iteration
current_Y_velocity = velocity_Y_start
#start velocity of the horizontal axis
velocity_X_start = 2
#the limit for the fall
floor_limit = 290
#to check if the ball is falling or rising
free_fall = True
```

Figure 6: All the variables have their description added as comments.

The user must call then one turtle function in the while loop to update the projectile position (Figure 7) in each new iteration (Figure 8).

```
#basic initialisation of the window
wn = turtle.Screen()
wn.bgpic('background.png')
wn.setup(1400,800)
wn.title("Projectile motion")
wn.update()

#definition of the projectile
ball = turtle.Turtle()
ball.shape("circle")
ball.color("blue")
ball.penup()
ball.turtlesize(size)
ball.speed(0)
ball.goto(xball,yball)
ball.dy = 0
```

Figure 7: In the window settings we define the width and the height, which is important to be set correctly, if we want to see adequate results. The functions for defining the window and the object are a part of the Turtle library.

```
while True:
    #set the position of the projectile
    ball.setpos(calculate_X(), calculate_Y())
```

Figure 8: The main while loop is readable due to the definition of functions outside of the loop.

If needed, the code can be upgraded by adding some other influences such as wind, drag and bounce. With small changes in the functions, the curvature of the projectile will change into shape including those influences. With that a more authentic situation can be simulated (Figure 9). In addition of the dissipative forces, the basics of energy conversion is introduced as the projectile can decrease the speed due to work of those forces, e.g. at bounce.

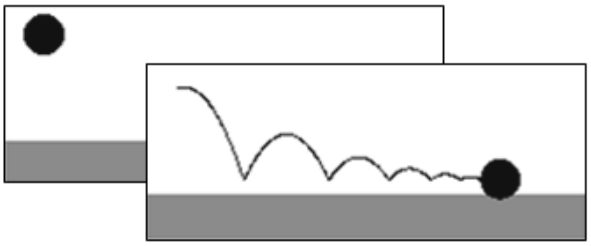


Figure 9: Compared to (Figure 4), the trajectory of the projectile shows the addition of the bounce.

In conclusion, even though the result of the programming is visual, the whole process of creating the functionalities behind that is pure code writing without any graphical interface. Yet with dividing the code into multiple parts (definition of the variables, definition of functions, definition of the elements of the simulation and the main while loop), the overview is simplified.

B. Visual Basic

The Visual Basic is event driven programming language. The main idea is that user is the controller and he/she initiate program's function to do predefined work. It can be used to start action immediately after loading but more often program loads and wait for user input to do different predetermined tasks.

Visual basic has strong support for windows forms and therefore programmer does not need to make much effort in the background of the design work for his/her program. To do the simulation drawing we can use multiple approach. We could calculate all trajectory of the moving object and store it into *x* and *y* array with the time as index. But we could use timer to control calculation and on each tick of a timer we calculate new *x* and *y* and draw new position of the object. We could use the combination too. The first approach require more processing power at the beginning and later we can draw the object forward and backward and analyse the behaviour through time. On the same time, this approach is memory capacity consuming since it store all positons of *x* and *y*. The second approach calculate the position of *x* and *y* position when needed. Its drawback is that on each movement, the calculation needs to be performed but we do not need much memory though. If we have enough processing power, the user should not perceive difference between the two approaches. Users with high environmental awareness could argue that first approach is less power consuming if we analyse movement of the object forward and backward.

The work in the Visual basic start as the new Windows Form Application project and continues with the form design. At the beginning, we start the simple model of free fall to test the formulae and graphics.

Free fall

For a simulation of a free fall of a ball, we need only gravity and height and do the math how much time it will take for a ball to fall to the ground. We simulate the fixed time step and decide to stop simulation in the 10 steps. We have set the topmost position of the ball to correspond to the height of the simulation and size of a window.

Therefore, the initial parameters for our simulations are set as shown in Figure 10 and are then set via the graphical interface of the simulation (Figure 11).

```
gravity = Val(tGravity.Text)
velocity = Val(tVelocity.Text)
ball_height = Val(tHeight.Text)
ratio = Me.Height / ball_height
'ration between screen pixel and height

time = Math.Sqrt(2 * ball_height / gravity)
'physic

time_tick = time / sim_steps
'simulation is stopped in 10 steps
```

Figure 10: Initialization of the parameters in the Visual Basic



Figure 11: Calculation of the ratio between pixels and height of the simulation

All the physics is hidden in two formulae (Figure 3 and Figure 12) which are known to anyone who learn physics. But to make them alive (Figure 11) we need to do additional computational steps to calculate ratio (Figure 10).

```
Private Function calculate_x(time As Double)
    Dim xx As Double
    xx = velocity * time
    Return xx
End Function

1 reference
Private Function calculate_y(time As Double)
    Dim yy As Double
    yy = gravity * time * time / 2
    Return yy
End Function
```

Figure 12: Functions that calculate physics motion of the ball

Simulation starts with the timer in visual basic that needs to stop in the predetermined condition. In our case predetermined number of steps (Figure 13). We could stop the simulation even when we hit the ground or if new location of the ball does not differentiate to the old one for certain amount of calculation cycles.

```
Private Sub Timer1_Tick(sender As Object, e As EventArgs) Handles Timer1.Tick
    Dim ball_new_x As Double
    Dim ball_new_y As Double
    If counter <= sim_steps Then
        ball_new_x = calculate_x(counter * time_tick)
        ball_new_y = calculate_y(counter * time_tick)
        'global counter that count sim_steps
        counter = counter + 1
        'new position of a ball in x axes
        ball.X = ball_new_x * ratio + bx
        'new position of a ball in y axes
        ball.Y = ball_new_y * ratio + by
        mygraphics.DrawEllipse(mypen, ball)
    Else
        Timer1.Enabled = False
        Button1.Text = "Start"
    End If
End Sub
```

Figure 13: Simulation for each step of the timer

The result is the consecutive drawing of the position of the ball in free fall (Figure 14).



Figure 14: The graphics outcome of the simulation

Free fall with the wind (add velocity)

The effect of the added horizontal wind (*x* direction) to the free fall is seen on Figure 15.

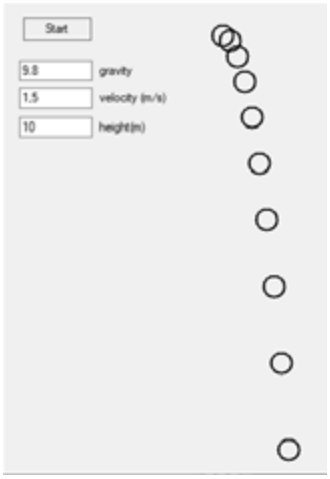


Figure 15: The graphics outcome with the velocity present

V PRELIMINARY RESULTS

It was our intention to try this approach with our students of educational physics study programs. But we have discovered that some constraints were a bit too great to overcome. We were not able to conclude them inside the time frame and topics of the courses. The evaluation process requires a little more than basic programming skills and could not be done in one half of the year.

The scenario to use this approach effectively is shown in the (Figure 16). Where preconditions are the basic knowledge of programming in one of the programming languages. If the preconditions are not achieved, the course of programming need to be finished first.

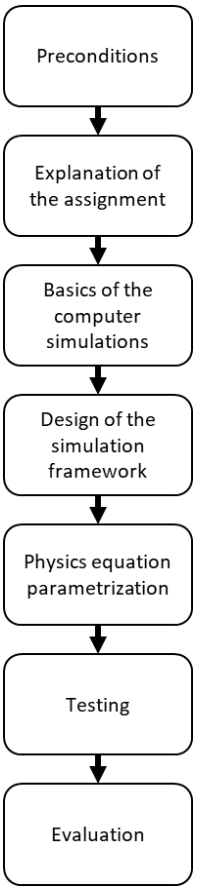


Figure 16: Scenario for the cross-curricular physics simulation programming in which we cannot proceed (visualized with the arrows between the steps) to the next step without meeting the preconditions

We should start the process with the explanation what is the desired outcome of our assignment and consider the constraints. The next step is to cover basics of computer simulations. We need to explain the (1) fixed time step simulation or (2) next event time simulation. In our case we can use fixed time since the phenomena is well known and we are capable to calculate the lifetime of the phenomena. Therefore, it is always possible to finish simulation in the predetermined numbers of step where time of the single step is calculated as the ratio: lifetime of the phenomena *over* number of steps of the simulation.

Later we point out the importance of the graphics in the design of the simulation framework. Depending on the selected computer language we explain what needs to be done for graphics initialization and how we draw shapes on the canvases.

In the next step we need to take physic formulae and convert them to such type that can be used in the physics simulation.

The testing phase is necessary to test the different input parameters and expected behaviour of the simulation. In this phase, most of the learning is achieved. Correcting the faults and errors is invaluable part of the process of learning.

The evaluation is the final step. It ends with the report and the understanding of the students that it is not just to learn something. Understanding that the process needs to be finished is something that they are going to need in all other assignments.

The described scenario can be applied in the cross-curriculum assignments where computer science and physics would cover mutually benefited topics. This would be beneficial for computer science too because it would empower the programming topics which are mostly not covered in our curriculum (in primary education and most of secondary education).

VI Blueprint for the Secondary School Test

The mathematical knowledge of primary school students is not developed enough to use full potential of cross curricular assignments in the projectile motion. Trigonometry is a topic of secondary education. Jet we do not need trigonometry to calculate the position of the ball and therefore the angles of a falling bodies based on non-zero horizontal velocity can be presented even in primary school. Therefore, a constant and accelerated movement is possible to be introduced in the primary schools.

The secondary school physics can use more complex mathematical tools. Students can use trigonometry even in 3D space. It is possible to cover 3D terrain and 3D wind. MSW Logo was the first programming language that managed to cover 3D graphics in a simple way. The other programming languages require elaborate calculation for the 3D to 2D transformation. Secondary school student are capable to use side view and top view. The 3-point perspective, which is the most authentic, is more complex to master. Most of the time it is part of the tertiary education for computer science.

For other calculations, we need integral equation. Integral equation was taught typically in the last year of the secondary school and therefore too late (too short time is this mathematical tool related knowledge available until the end of the secondary school education) to use it in the cross curricular assignments. If we wanted to experiment with the air resistance, we can simplify the mathematical model to the ideal ball and provide the formulae to the secondary school students. We cannot expect of them to understand the mathematics behind the air resistance but they would grasp the background concept of it.

In conclusion, using this kind of method at teaching physics can be beneficial for both, primary and secondary school (with some adjustments in primary school of course).

Although, the complexity must be narrowed down greatly in primary school since our country's curriculum does not include the necessary knowledge and skills for programming simulation of projectile motion in advanced complexities. The concept of learning a physics phenomenon through programming adds a fresh approach in teaching different than usually. We are aware that this approach is not achievable for primary schools in general because the computer science courses may not be mandatory in all countries.

VII Conclusion

In the last two decades, we see decline of the general knowledge of computer science in primary and secondary education. What is now called computer science in our country is actually application of computer in different areas of life and very far from term computer science. Our youth just not know how to code anymore. On the other hand, algorithmic thinking is needed in education. Some subjects are not popular between young populations. This is partly true because from the point of view of students it is so remote from their real-life experience and some of the teachers use old and cumbersome teaching approach from past century. It is true that they need to learn how to code algorithms; and it is true they need to learn basics of physic; and it is true that both area are for too many students boring. We have asked ourselves if we could do something to connect these two area and made them more appealing to the students. Hence, we developed and proposed our approach to teach physics by programming physics simulation and vice versa – teaching programming using physics.

We have been successful with the students (future teachers) of educational computer science and educational physics because of their background in both disciplines. The physics we used in the experiment is simple and students can understand it easily but the main benefit was in the computer science. Computer science part of experiment shows students the purpose of learning programming; potential benefit could be used in lower level of education too (secondary and primary education). It is important to point out that using programming as a method to teach various phenomenon in physics is not sufficient, but it can be a great addition with the combination of experiments. In that way, students can learn about the phenomenon in ideal circumstances and then test it in the real environment.

The method is more suitable for secondary school and for university undergraduate studies. The collaboration with computer sciences eases the approach into creating the simulation. Using the step-by-step approach in teaching students how to develop computer program, the gain is observed in both courses (programming and physics). This approach makes a solid foundation in the programming environment, students can create various simulations and with each new assignment, the knowledge in each expertise becomes greater and more efficient.

In the becoming years, we are going to test this approach in primary and secondary schools. We expect that it would raise the understanding why programming should be the part of the education. If students learn physics at the same time as they practice programming, we strongly believe we are on the right path.

Acknowledgment

This article is a result of research, co-financed by the Republic of Slovenia and the European Union under the European Social Fund within the project "Innovative

learning and teaching in higher education" [6]. The development of this article was also supported by Slovenian Research Agency grant number P1-0403 [16] and by the Republic of Slovenia and the European Social Fund within the projects PODVIG [4] and NAMA POTI [5].

REFERENCES

[1] H. Shapiro, S. Fuglsang Østergaard and K. Frøhlich Hougaard, "Does the EU need more STEM graduates? Final report," Publications Office of the European Union, Luxembourg, 2015.

[2] I. Gerlič, M. Brumen, M. Marhl, A. Šoštarič and M. Krašna, "Konceptualno poučevanje naravoslovnih znanosti : projekt CoLoS - Tempus : poročilo o nalogi za 1995 : [Tempus S_JEP-09641]," Faculty of Education, University of Maribor, Maribor, Slovenia, 1995.

[3] D. Robič and R. Repnik, "Computer game Angry Birds in teaching physics," in *MIPRO 2015*, Opatija, Croatia, 2015.

[4] National education institute Slovenia, "ZSRŠ projekt PODVIG," 2019. [Online]. Available: https://www.zrss.si/objava/projekt-podvig. [Accessed 01 03 2020].

[5] National education institute Slovenia, "Project NA-MA POTI, interdisciplinary authentic complex problem solving," National education institute Slovenia, 2017. [Online]. Available: https://www.zrss.si/objava/projekt-na-ma-poti. [Accessed 14 02 2020].

[6] INOVUP, "Inovative learning and teaching in higher education," 2019. [Online]. Available: http://www.inovup.si/en/. [Accessed 01 03 2020].

[7] Karl-Franzens-Universität Graz, "Curriculum für das Bachelorstadium Phsyik 2017," 2017. [Online]. Available: https://mibla-archiv.tugraz.at/16_17/Stk_12a/NAWI_Bachelor_Physik_Version_2017.pdf. [Accessed 29 01 2020].

[8] Lund University, "Courses - Basic Level," [Online]. Available: http://www.fysik.lu.se/index.php?id=131038. [Accessed 29 01 2020].

[9] University of Oxford, "3-year BA Physics," [Online]. Available: https://www2.physics.ox.ac.uk/study-here/undergraduates/the-courses/3-year-ba-physics. [Accessed 29 01 2020].

[10] University of Pécs, "BSc in Physics," [Online]. Available: https://international.pte.hu/study_programs/bsc_physics. [Accessed 29 01 2020].

[11] Jacobs University, "Bachelor of Science - Undergraduate Program," [Online]. Available: https://www.jacobs-university.de/study/undergraduate/programs/physics?gclid=CjwKCAiA66_xBRBhEiwAhrMuLdQDsT7X57OeQtfEeVu-RaAbf3qn6GBwKIhcXvfV4e5XUqaBfvwESRoCyPoQAvD_BwE. [Accessed 29 01 2020].

[12] University of Oulu, "BSc in Physics," [Online]. Available: https://www.oulu.fi/studiesarchive2013/5-6-year-degree-programmes-2013-2014/bsc-physics. [Accessed 29 01 2020].

[13] University of Zagreb, "Integral undergraduate and graduate study of physics," [Online]. Available: http://www.pmf.unizg.hr/phy/en/study. [Accessed 29 01 2020].

[14] M. Svetec, R. Repnik, R. Arcet in E. Klemenčič, „Educational technology at the study program of educational Physics at the University of Maribor in Slovenia," v *The role of technology in education*, F. Altlnay, Ured., London, IntechOpen, 2020, pp. 1-16.

[15] Faculty of Natural Sciences and Mathematics, "Univerzitetni študijski program prve stopnje Fizika," [Online]. Available: https://www.fnm.um.si/index.php/2010/10/04/univerzitetni-tudijski-program-prve-stopnje-fizika/. [Accessed 29 01 2020].

[16] Computationally intensive complex systems, Faculty of Natural Sciences and Mathematics, [Online]. Available: https://www.sicris.si/public/jqm/prg.aspx?lang=eng&opdescr=search&opt=2&subopt=700&code1=cmn&code2=auto&psize=1&hits=1&page=1&count=&search_term=perc%20matja%C5%BE&id=17624 . [Accessed 29 01 2020].

Improving the Quality of Entrepreneurial Education by ICT Education of HEI Pedagogical Staff

D. Paľová*, M. Vejačka*
* Department of Applied Mathematics and Business Informatics
Faculty of Economics, Technical University of Košice, Slovakia
dana.palova@tuke.sk, martin.vejacka@tuke.sk

Abstract – **In spite fact, that current unemployment rates reach a historical minimum since the last world economic crisis, the numbers connected to the unemployment of young people are still not encouraging. The rates of young people's unemployment are almost double in comparison with the unemployment of other groups of EU citizens. The Erasmus+ is EU funding scheme, which aims to support education, training, youth, and sport in Europe. Within this scheme, the project ReStart was designed. It aims to reinforce the entrepreneurial and digital skills of students and teachers to enhance the modernization of higher education in Moldova. Its' main objective is to innovate the education in the field of entrepreneurship in connection with European Union best practices, local market expectations with a better understanding of relevant entrepreneurship and digital learning needs of students, teachers, and local businesses. The paper will describe the present status of the education innovation process including the implementation of new pedagogical approaches based on using ICT. Also, the results of the ReStart project will be presented.**

Keywords – entrepreneurial education; LMS Moodle; interaction; teachers' education; ICT

I. INTRODUCTION

The consequences of the economic and financial crisis persist in the economy. However, unemployment rates slowly decrease each year. Moreover, the impacts of the present and continuing COVID-19 crisis of the economic environment of all countries around the world are more-less still unpredictable. But the fact is, that rates of unemployment in the EU will rise in comparison with status before March 2020. Following the strategic documents of the European Union (EU), the unemployment rate of the young citizens is almost double in comparison to the unemployment rate of the other EU citizens in productive age. This is the reason, why the employment of young people is one of the key challenges in Europe. European Commission performs several activities to improve the current situation (e.g. Youth Employment Initiative, European Qualification Framework, etc.). One of the most important tools that can be used to decrease youth unemployment is education. As the previous pandemic months showed us, education needs serious reform in way of using modern available technologies to enhance students' motivation to learn at their home and use the different on-line platform to share our knowledge and skills. Unfortunately, the situation showed, that there is a lot of work before all participants of the education process. But the positive of the situation is, that technologies got a chance to be used in different ways and individuals got the opportunity to achieve a relationship to lifelong learning at their homes.

However, to promote the skills of young people and to provide them with the knowledge they will need (especially at the labor market), it is necessary to transform education. One of the biggest challenges is to support higher education areas to promote entrepreneurship education. Students must gain practical skills and experience besides providing them with theoretical knowledge. This is the main objective of the international project "Reinforce entrepreneurial and digital skills of students and teachers to enhance the modernization of higher education in MOLDOVA" (ReStart; supported by EU program Erasmus+). The project aims to improve study programs, students' and teachers' entrepreneurial skills by employing digital education. The main purpose of the paper, as we are already in the midterm of the project lifetime, is to present the mentioned project and its already achieved and expected outcomes, without analysis of obtained results that are expected at the end of the year.

II. UNEMPLOYMENT OF YOUTH

Unemployment is one of the crucial economic measures for the country. The economic and financial crisis resulted in young people becoming one of the age groups most at risk of social exclusion. The unemployment of young citizens is usually higher when compared to the other groups of the citizens in productive age [1].

Typically, young new graduates face the reality of finding the job. The unemployment rate of young people aged 15-24 years in the EU reached a height of 24.0 % in February 2013, with peaks of 60.0 % in Greece, 56.2 %

in Spain, 49.8 % in Croatia, 44.1 % in Italy and 40.7 % in Portugal. Since then the youth unemployment rate decreased to 14.2% in 2019, thanks to the economic growth, but this number presents still a high level of unemployment [1].

The increase in employment, productivity, and social cohesion are pillars of the EU's smart growth. Intending to increase employment, especially the employment of the youth and recent graduates, it is important to support their ability to self-employment. The excessively theoretical approach in the educational process was identified amongst the major problems of education. The educational system should provide students the opportunity to acquire practical experience, which represents the advantage when applying for a job or developing their own business.

The Council addressed this situation and established the Youth Guarantee, a political commitment to ensure that all young people under the age of 25 years receive a good-quality offer of employment, continued education, an apprenticeship or a traineeship within a period of four months of becoming unemployed or leaving formal education. To support this policy EU created the Youth Employment Initiative (YEI). [2]

The Youth Employment Initiative (YEI) includes the following actions to promote the application of youth in the labor market:

- stepping up of early activation and intervention with the development of individual action plans,
- improvements to the quality of general education and training, including reforms in the vocational education and training system and in the apprenticeships system, better digital technologies, and foreign language learning, support for young people who have not completed secondary education to return to education or training,
- easier school-to-work transitions, for instance by helping young people secure their first work experience or traineeship, intending to increase their employability and labor market attachment, including in the public sector,
- acquisition of professional skills in new professions or professions where there are labor shortages, in training and on-the-job training,
- strengthening public employment services capacity, with the establishment of an integrated system of career guidance, help and guidance for job seekers, such as CV and cover letter, or job interview preparation,
- implementation of mentoring schemes as a part of labor market services, either when the young person is in training or on-the-job,
- activation support for self-employment, such as start-up support (including mentoring) for young entrepreneurs,
- incentives for employers, such as wage and recruitment subsidies and reductions of non-wage labor costs. [2]

All the above-mentioned actions are in close connection with strategic document Europe2020 targets, where entrepreneurship was identified as one of the key elements to promote innovation, competitiveness, and economic growth [3]. The entrepreneurship might be defined as a transversal competence, which applies to all spheres of the life: from the nurturing personal development to the actively participating in the society, to (re)entering the job market as an employee or as a self-employed person, and also to start-up ventures (cultural, social or commercial). Therefore, the entrepreneurial education of young generations is highly important.

In general, entrepreneurship represents an individual's ability to turn ideas into action. It includes creativity, innovation, and risk-taking, as well as the ability to plan and manage projects to achieve business objectives [4]. The development of the entrepreneurial skills at the multinational level is covered by the New Skills Agenda for Europe [5], [6] by its particular area - The Entrepreneurship Competence Framework (EntreComp) [7]. EntreComp conceptualizes entrepreneurship as competence needed in the XXI century. It covers the mixture of cognitive skills, social skills and relationship skills, technical skills, and managerial skills, also defined as important in other researches e.g. [8],[9].

A. Youth unemployment in the Republic of Moldova

The problem with youth unemployment is significant also in European countries outside the EU.

The latest global economic and financial crisis strongly influenced the economy of the Republic of Moldova, consequently, the investments in the economy have been reduced, the local currency depreciated, the remittances have been reduced and the purchasing power of the population diminished, etc. Despite these phenomena, Moldova's gross domestic product (GDP) grew by 3.4% in 2018, according to the World Bank data [10]. The most significant influence on GDP growth was the gross added value created in the wholesale and retail trade, transport and storage, hotels, and restaurants. Crediting the economy reflects the first signs of revival, the volume of new loans granted in the economy registered an increase of about 28% [11]. According to the Global Competitiveness Index 2017-2018 issued by the World Economic Forum, Moldova ranks 89th out of 137 included states. [12]

Following Figure 1 shows changes in youth unemployment rates in Moldova since its independence in 1991.

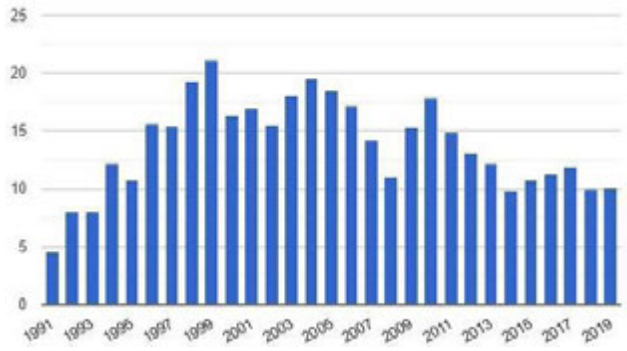


Figure 1. Youth unemployment rates in Moldova during years 1991 - 2019; source:[13]

The unemployment of the youth in Moldova increased during years of transformation with the maximum in 1999 at the level of 21.14%. During the 2000s the youth unemployment rate fluctuated until peaking in 2004 (19.51%). Since then it steadily decreased until the economic crisis with another increase peaking at 17.85% in 2010. Since then it decreased and now it is stagnating around 10% (10.03 in 2019). Still, it is a high level of unemployment and it must be addressed by economic measures[13]. The above-stated data show a complicated socio-economic situation in Moldova. The Republic of Moldova signed the Association Agreement with the European Union in 2014. Its agenda outlines the priorities for reform in Moldova. According to its recommendations, the Republic of Moldova should conduct multiple reforms also in the education area. One of the recommendations also includes the promotion of academic cooperation, capacity building, and student and staff mobility through the Erasmus+ program and researchers' mobility and training through Marie Skłodowska-Curie actions [14]. As the reaction to the economic situation of Moldova and in connection with that recalled necessary education reform, the project ReStart (in which also our university is participating) started under the Erasmus + program.

III. RESTART PROJECT

Project ReStart (Reinforce entrepreneurial and digital skills of students and teachers to enhance the modernization of higher education in Moldova) is co-funded by the program Erasmus+. The international consortium of the project is composed of universities from Moldova (State University of Moldova, Alecu Russo State University of Balti, Academy of Economic Studies of Moldova, State Agrarian University of Moldova and BP Hasdeu State University of Cahul), Bulgaria (University of Ruse "Angel Kanchev"), Romania (Politehnica University of Bucharest and The Agence Universitaire de la Francophonie - Central and Eastern Europe Office) and Slovakia (Technical University of Košice). Its primary objective is to contribute to the update of 2nd-degree study programs of Moldavian partner universities in the field of Business & Administration with the main aim of improving the entrepreneurial skills of students and teachers through digital education [15].

The most important benefits for the students of the project include:

- improving the level of knowledge in the field of entrepreneurship and business plan development,
- improving the level of digital skills,
- improving student preparedness for labor market needs.

The most important benefits from the teachers' point of view of the project include:

- development of 6 innovated curricula in the field that they can use in their university environment,
- the possibility of acquiring new information, knowledge, and skills on how to implement modern pedagogical approaches with the possibility of using tools and methods of digital education.

Since the start of the project, several activities were conducted that aim to achieve the main objective:

- assessing the needs of the current business in the Republic of Moldova,
- identifying the level of digital skills of participants in the educational process in Moldovan universities;
- integration of the "Best practices" into the education process and share of the experiences of the EU partner universities.

Best practices integration is realized through:

- organizing so-called "Training of Teachers", where the knowledge, skills, and experience are shared directly at the European universities,
- practical demonstrations of the teaching units carried out directly by the European partners in the environment of the Moldavian universities,
- creation of innovated curricula and courses in the LMS Moodle online environment,
- reviewing the curricula developed by the European partners with the aim of continuous improvement of created modern learning environment using innovative approaches and digital technologies.

Developed educational materials and online courses are directly tested as part of pilot testing by individual Moldavian universities. Students who have completed the courses then present the acquired knowledge within the so-called "Business Plan Competition", where students from all participating universities compete with each other. The students' business plans are evaluated by experts from European universities. The project outputs are communicated to the public in various ways, in all cases, it is not only a one-sided presentation of the project results, but the aim of the events is also to get feedback from the professional public (educators, entrepreneurs, etc.) operating in the Republic of Moldova. Most significant events include:

• Information meetings of target groups - aimed at students and educators of Moldavian universities, where the possibilities of education to support entrepreneurship and the possibilities of using innovative education are presented.

• Career days - meetings of students with entrepreneurs operating in Moldova to bring students closer to the reality of entrepreneurship and labor market needs.

• Multiplier events at Moldavian universities involving students, educators, entrepreneurs, but also professional associations, civic organizations, and various public administrations to raise awareness of the potential of business education contributing to a better business environment and a better understanding of new business approaches.

IV. IMPROVING THE QUALITY OF EDUCATION BY ICT PRACTICES

A. Advanced ICT Practices in LMS Moodle

When using LMS Moodle for education, teachers have multiple advanced ICT tools or practices suitable for improving the quality of education. Students might be navigated through the content of the Moodle course in the desired way using principles of the learning path. A learning path is a sequence or a structure of the content and activities that help students to get the right knowledge, in the right order, and without wasting time. To say it simply, a teacher could define, which content and activity are available and when, for example after completing a previous activity, quiz, etc. The learning paths could include intermediate quizzes, videos, presentations, assignments, and text.

Using this approach, the students' performance might be tracked and analyzed for two purposes:

- to improve the course in the following iteration,
- and to provide the feedback about achieved knowledge and skill to the student and the teacher.

Using the learning paths represents a tool for the student's higher engagement in the educational process, as it provides the student with feedback and pace of learning adapted to his / her knowledge and skills. Furthermore, the environment of LMS provides students the possibility to cooperatively work on practical problems and peer-asses their performance, which should be used as another tool to enhancement the students' involvement in the education process. Following examples of advanced activities in LMS Moodle that might be used to improve the education process:

Workshop - module enables the collection, review, and peer assessment of students' work. Students can submit any digital content or type text directly into a field. Submissions are assessed using a multi-criteria assessment form defined by the teacher. The process of peer assessment and understanding the assessment form can be practiced in advance with example submissions provided by the teacher, together with a reference assessment. [16]

Quiz - the quiz activity enables a teacher to create quizzes comprising questions of various types, (e.g. multiple-choice, matching, short-answer, etc.). The teacher can allow the quiz to be attempted multiple times, with the questions shuffled or randomly selected from the question bank. Quizzes may be used as course exams, mini-tests for reading assignments or topics, exam practice using questions from past exams, to deliver immediate feedback about performance and for self-assessment.[16]

Survey - module provides several verified survey instruments that have been found useful in assessing and stimulating learning in online environments. A teacher can use these to gather data from their students that will help them learn about their class and reflect on their teaching. [16]

B. Training-of-Teachers Event for Improving the Quality of Entrepreneurial Education

Training-of-teachers event was part of the Restart project and it was called "Transforming Teachers Mindsets Through Digital and Entrepreneurial Skills Development". The event was attended by 37 teachers from 5 different Moldavian universities and it took place in October 2019. The event contained the following modules:

- Module 1 – Innovative teaching and learning – which included analysis and training of soft skills training with a psychologist during the teaching session, and also practical advanced ICT-practices in blended learning courses using LMS Moodle.
- Module 2 – Gaming in education (games implemented in management and marketing courses).
- Module 3 – Entrepreneurship skills development for teachers and students case studies (lectures on topics of Entrepreneurship, Family business problems and prospects, family business case studies, etc.).
- Module 4 – Technological transfers – lecture on technology transfers and start-up incubators.
- Module 5 – Pedagogical aspects in counseling and mentoring of students (Key subjects in digital and entrepreneurial skills for teachers; Pedagogical aspects in counseling and mentoring of students).
- Module 6 – Professional insertion of MD students, international call for dual-learning projects, lessons learned, and best practices in dual-learning.

Within Module 1 of the Training-of-Teachers event, the innovative ICT tools and activities available in LMS Moodle were introduced by lecturers and practically tried by participating Moldavian teachers. The main idea was to improve the practical experience of teachers from Moldavian universities by working with LMS Moodle within the education process. However, to better understand the functioning of the entire education process, one of the goals was to demonstrate how students see the activities in the courses and on the other hand, how the teachers see them.

Before the Training-of-Teachers event, we reviewed the courses developed by the participants in LMS Moodle provided by their university. The result was that the courses in LMS Moodle used by participants did not use the potential of the given LMS to support the educational process as they were only at the basic level. Their courses used Moodle as a closed environment for publishing static content. Therefore, various advanced activities within practical exercises were presented to increase the attractiveness of participants' courses in LMS Moodle. The usage of advanced education tools in the LMS Moodle that offer to control the acquired knowledge and skills and to control the educational process of individual students was therefore demonstrated.

During the session design phase, the study material and demonstrations were prepared for participants, containing basic information about resources and activities, blended learning, learning paths, learning analytics, and their creation in LMS Moodle. The basic experience with LMS Moodle (to create static contents – uploading the texts, pdf files, titles, etc.) was assumed as the participants already prepared theoretical lectures during earlier training sessions aimed at basic use of LMS Moodle.

To ensure the efficient use of resources and time during a session, the participants were asked within the training proposals to prepare the following materials for advanced ICT practices training in advance:

- content of topic in their course (arbitrary depending on participant's specialization, research, etc.),
- resources – uploaded into the course by participants during the course content development,
- minimum of 5 different questions (and corresponding answers) for the test based on the content of the course topic,
- a task which other participants should solve while acting as students (short activity e.g. on-line text, essay, and detailed description of how the assignment would be evaluated by others).

During the realization phase, firstly, the participants were asked to answer the questionnaire focused on their experience with LMS Moodle and blended learning. It confirmed that all participants had basic experience with the usage of LMS Moodle. Almost 91% of them heard about learning competencies but only 54% ever set the learning competencies in their LMS. Just above 77% knew about the learning path but only 36% set the learning path in their course at least once before. Around 23% knew some of the tools for learning analytics in LMS Moodle and used them before. Similarly, around 23% ever involved students in the activity evaluation process. Following Table I contains participants' answers on questions regarding their experience with advanced tools in LMS Moodle.

TABLE I. ANSWERS ON QUESTIONS REGARDING PARTICIPANTS' EXPERIENCE WITH ADVANCED TOOLS IN LMS MOODLE

In which roles do you use LMS Moodle?	administrator	manager	teacher	a teacher without possibility to change course content	student	host
	9.09%	9.09%	100.00%	13.64%	4.55%	4.55%
Which resources in LMS Moodle do you know?	URL	Page	File	Folder	Book	Label
	81.82%	68.18%	86.36%	95.45%	59.09%	36.36%
Which activities in LMS Moodle do you know?	Assignment	Quiz	Choice	Workshop	Database	Glossary
	59.09%	81.82%	45.45%	22.73%	27.27%	50.00%
Which of the available resources do you use most often?	URL	Page	File	Folder	Book	Label
	9.09%	4.55%	40.91%	45.45%	0.00%	0.00%
Which of the activities do you use for your students most often?	Assignment	Quiz	Choice	Workshop	Wiki	Other
	45.45%	45.45%	4.55%	0.00%	0.00%	4.55%
Which of the advanced Moodle activities have you ever used before?	Workshop	Wiki	Database	Glossary	Survey	None of the mentioned
	13.64%	0.00%	4.55%	9.09%	40.91%	31.82%

own results

The main idea was to achieve experience in the course development as a teacher, but also gain insight into the education process from the student's side. From that point of view, practical training was divided into two parts. Firstly, when participants acted as teachers, they created the content of their course. Every content should have the character suitable for increasing student engagement in the course (in LMS Moodle: poll, forum, chat, dictionary, feedback, test, wiki, and workshop). Further, participants designed the quiz based on the content of the topic. Lecturers also described the possibilities of the workshop activity available in LMS Moodle and participants tried to prepare a workshop while acting as teachers. In the second part of the event, participants tried the roles of students in the course and tried out the content created by other participants (including tests and workshops). Switching the role between the student and the teacher was necessary (to manage their topics) until the end of the event. In LMS

Moodle only 32 topics were edited, while some of the teachers rather worked in pairs.

At the end of the practical training, participants gave feedback on this part of the training event. Almost 68% appreciated the professionalism of lecturers in the eLearning area and over 74% valued the possibility of consultation with the experts. More than 90% welcomed the chance to directly try out the advanced tools of eLearning. Almost 71% of participants appreciated acquiring new knowledge and 81% the skills trained during the event. Furthermore, over 29% of participants are planning to create new courses in their LMS Moodle and almost 71% will further develop and improve their current courses, from that 68% through increasing the interactivity of their courses. More than 93% of respondents will implement approaches of eLearning into education for increasing the interactivity and the motivation of students. Around 13 % will include quizzes and bank of questions into their courses (the most of respondents already used the quizzes).

All of the participants adduced that they learned something new in the area of advanced tools in LMS Moodle. They particularly highlighted achieved practical experience with quiz and workshop activities in LMS Moodle, which they are planning to implement into their courses. All of them agree that they can apply acquired skills and knowledge in their teaching praxis.

V. CONCLUSION

Youth unemployment is a very serious problem, which can be partially addressed by improving the quality of education. Especially, the entrepreneurial education can promote the self-employment of the youth. However, to improve entrepreneurial education, it is important to integrate the most up-to-date educational approaches and make use of available modern technologies. Innovation and improvement of entrepreneurial education in Moldova is the key aim of the ReStart project presented in the paper. Its outputs are the creation of innovated curricula and also the exchange of experience between educators from different universities, different states, and different business environments. The paper also indicates the results of the Training-of-Teachers event aimed to improve the quality of entrepreneurial education including practical try-out of advanced ICT practices available in LMS Moodle by teachers from Moldavian universities. This exchange of experiences drives the further development of the education process, which benefits not only the students themselves but also the labor market. The main aim of the paper was to present the share of knowledge and experience in the higher education innovation process at the Moldovan universities. The process is still in progress and the results of the "ReStart" project implementation are expected at the end of this year.

ACKNOWLEDGMENT

This contribution was made thanks to the participation of authors in an international project ReStart (Reinforce entrepreneurial and digital skills of students and teachers to enhance the modernization of higher education in Moldova) co-financed by the European Commission under the program Erasmus+, KA2 Program – Capacity Building in the Field of Higher Education.

REFERENCES

[1] Statista: "Youth unemployment rate in EU member states as of August 2019", [online], available at: https://www.statista.com/statistics/266228/youth-unemployment-rate-in-eu-countries/ [Accessed 20. January 2020]

[2] European Parliament: Youth Employment Initiative [online], available at: http://www.europarl.europa.eu/RegData/etudes/BRIE/2017/608688/EPRS_BRI%282017%29608688_EN.pdf [Accessed 21. January 2020]

[3] European Commission: "EUROPE2020: A European strategy for smart, sustainable and inclusive growth", [online], available at: http://ec.europa.eu/eu2020/pdf/COMPLET%20EN%20BARROSO%20%20%20007%20-%20Europe%202020%20-%20EN%20version.pdf, [Accessed 25. January 2020]

[4] World Bank: Doing Business 2019: Training for Reform – Moldova. [online], available at: http://documents.worldbank.org/curated/en/196231541152940523/Doing-Business-2019-Training-for-Reform-Moldova [Accessed 24. January 2020]

[5] European Commission: "Analytical Highlight Focus On Entrepreneurial skills", [online], available at: https://skillspanorama.cedefop.europa.eu/sites/default/files/EUSP_AH_Entrepreneurial_0.pdf, [Accessed 24. January 2020]

[6] European Commission: "EntreComp: The Entrepreneurship Competence Framework", [online], available at: http://publications.jrc.ec.europa.eu/repository/bitstream/JRC101581/lfna27939enn.pdf, [Accessed 7. January 2020].

[7] European Commission: "New Skills Agenda for Europe", [online], available at: https://ec.europa.eu/social/main.jsp?catId=1223, [Accessed 25. January 2020]

[8] E. Muzio, D. Fisher, E. R. Thomas, and V. Peters: "Soft Skills Quantification (SSQ) for Project Manager Competencies." In Project Management Journal, Pennsylvania, USA, vol. 38, no. 2, 2007, pp. 30–38.

[9] J. Wiklund and S. Dean: "Entrepreneurial Orientation and Small Business Performance: a Configurational Approach", in Journal of Business Venturing, vol. 20, 2015, pp. 71-91.

[10] World Bank: GDP growth (annual %). [online], available at: https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=MD

[11] National Association of Young Managers: Situation Of Social Entrepreneurship And Entrepreneurship Possibilities For Young People In Republic Of Moldova, National research in framework of the KA2 project "Entrepreneurial EAST (Entrepreneurship Activities – Strategic Training)", 2018

[12] World Economic Forum's Global Competitiveness Report 2017-2018, [online], available at: https://www.weforum.org/ [Accessed 23. January 2020]

[13] World bank: Moldova: Youth unemployment. [online], available at: https://www.theglobaleconomy.com/Moldova/youth_unemployment/ [Accessed 27. January 2020]

[14] Recommendation No 1/2017 of the EU-Republic of Moldova Association Council of 4 August 2017 on the EU-Republic of Moldova Association Agenda [2017/1489].

[15] ReStart project consortium: Project Description. [online], available at: http://www.restart-eu.upb.ro [Accessed 19. January 2020]

[16] Moodle docs: Add activities. [online], available at: https://docs.moodle.org/30/en/Main_page [Accessed 19. January 2020]

How to educate students for the future?

L. Révészová
Faculty of Economics Technical University of Košice
Department of Applied Mathematics and Business Informatics, Košice, Slovakia
libusa.reveszova@tuke.sk

***Abstract* - Technologies, especially ICT are the most pervasive force influencing human lives today. Computers and electronic devices are mostly considered ICTs, but technologies in general involve much more. Most of technologies nowadays consist of four main components: hardware, software, brainware and know – how. All of these components are equally important. The investment efficiency into the modern technologies depends especially on the interest, motivation and education of their users. We are educating students for their future career opportunities that require such kind of knowledge and skills, which we as teachers may only predict nowadays. There is a question to be answered: "What should every educator extract from wisdom of the past and knowledge of the present for education needed in the future?"**

The main goal of the article is to present and summarize our results of research of students' informatics knowledge and ICT competencies via questionnaire and problem solving. We also describe experience in teaching, the body of knowledge and methods which we have implemented into the subject Informatics II in the first year of Bachelor study at the Faculty of Economics, Technical University, Košice.

Keywords – education; computer science/informatics; information and communication technology;

I. INTRODUCTION

Nowadays, we live in the Industry 4.0 era – era of the fourth industrial revolution. All fields of human lives, science disciplines, economies and industries are influenced by new technologies, that combine the physical, digital and biological worlds. These technologies give us huge opportunity, for example connecting billions of people to the web. The efficiency of business and organizations is drastically changing and improving. Now it is possible to help regenerate the natural environment through better asset management, and repair the damage previous industrial revolutions have caused [10].

The pace of change and innovation has intensified markedly: product life cycles and time for development are being compressed, services are becoming a larger proportion of economic output, computer power and capacity unprecedentedly increase and hardware prices to fall, data transmissions costs are decreasing, and ICT is expanding. The recent transformations in the world show us that knowledge has become a primary factor of success in the economy competitiveness. Also in the tertiary education it is necessary to re-examine the policies and assumptions in a rapidly changing environment [26], [25].

The way we work transforms fundamentally. Human tasks and jobs are replaced by automation and 'thinking machines'. The skills that corporates, firms and organizations are looking for in their people are changing. All of us, and also businesses, organisations, governments etc. need to be prepared for a number of possible, even seemingly unlikely, outcomes. Some experts believe that artificial intelligence (AI) could create a new world where human abilities are amplified as machines help mankind process, analyse, and evaluate the abundance of data that creates today's world. In this new world humans will spend more time engaged in high-level thinking, creativity, and decision-making [24].

Across advanced economies the debate about technological unemployment among economists has intensified. The decline of employment in routine intensive occupations and the impact of computerization on labour market outcomes is well-established and documenting in the literature [5].

Problem-solving skills are becoming very productive. We can see the substantial employment growth in occupations involving cognitive tasks where skilled labour has a comparative advantage, as well as the persistent increase in returns to education [2], [5].

Throughout all human history, technological progress has vastly shifted the composition of employment and significantly influenced education. How technological progress in the twenty-first century will impact on labour market, and subsequently education we will see. But we should be prepared, perceiving changes and trends.

II. CURRENT CHALLENGES FOR EDUCATION AND FOR EDUCATORS

Many disciplines which are taught and the way how they are taught are both stalked by the threat of obsolescence. Nowadays, graduates need the kind of education that enables them to participate actively in all areas of their lives, to use different types of thinking and attitudes as global citizens as well as economic actors to support ethical, sustainable development of their societies. The key message from the Europe 2020 strategy is: *"we need more creative, flexible and entrepreneurial young people who are equipped for the challenges of today's ever changing work environment"* [4].

Human skills like creativity, thinking (logical, critical, computational), problem solving skills, leadership and empathy are now increasingly in demand. It is not just about science and technology. For educational institution it is necessary to identify the "knowledge and skills for the future" and start to concentrate on how to build them and how to use them alongside technology [24].

Education significantly develops human capital, and via human capital can support growth in two ways:

- first, by improving the capacity to absorb and adapt new technologies, which will affect short- to medium-term growth,
- second, by catalysing the technological advances that drive sustained long-term growth [17], [18].

As we can read in [25] tertiary (mostly university) education is more than the capstone of the traditional education pyramid. Universities are clearly a key part of all tertiary systems as a critical pillar of human development and a base for today´s lifelong-learning framework. Tertiary education provides not only the high-level skills necessary for every labour market but also the training essentials for teachers, doctors, nurses, civil servants, engineers, humanists, entrepreneurs, scientists, and myriad personnel. They are these trained individuals who develop the capacity and analytical skills that drive local economies, support civil society, teach children, lead effective governments, and make important decisions which affect entire societies.

The quality of education - teaching and learning - should be at the core of the needed higher education reform agenda. Based on our own experience it is necessary to focus on curricula that deliver relevant, up-to-date knowledge and skills, knowledge which is globally connected, which is useable in the labour market. Achieving this is not an easy task.

Undoubtedly, everyone agrees today, that higher education institutions should introduce and promote cross/trans and interdisciplinary approaches to teaching and learning. They are directly responsible for development of students' depth of understanding and entrepreneurial and innovative mind-sets. Curricula should be developed and monitored through dialogue and partnerships among teaching staff, students, graduates and labour market actors, drawing on new methods of teaching and learning, so that students acquire relevant skills that enhance their employability [4].

Skills that the 21st century teachers are supposed to have to promote high quality learning, according to [4], are the 'five Es': education, experience, enthusiasm, ease, and eccentricity. A teacher's knowledge base should not be restricted simply to his or her own subject, but must also include an understanding of learning theories and also how to incorporate them into practice. Teaching students well obviously implies that teachers produce up-to-date and good quality materials and supporting content for their lessons.

Students' interest and motivation are important factors as well. While certain elements may be affected by the teacher, motivation and learning styles are completely individual. To ensure efficient knowledge transfer, teachers must be adapted to conditions such as group size, timing and duration of the lectures.

Despite the benefits, which are reflected in the multitude of resources that facilitate course realization, the main challenge for teachers who teach computer science and introductory programming courses is how to teach problem solving and develop critical and computational thinking by using algorithms. Polya in [14] observed problem-solving process through four stages: understanding the problem, solution planning, its execution and reflection on the results

According to many authors personalization process in the modern educational systems has growing importance [12], [7].

III. INPORTANCE OF COMPUTATIONAL THINKING AND PROBLEM SOLVING SKILLS

Students of the 21st century start using computer and ICT even before they first attend formal education. They are considered as digital natives and computer and ICT are parts of their everyday lives and they are significantly accustomed to many of computer's functionalities because of playing computer games and using social media from a very early age [3].

Computational thinking is a new and much-needed part of literacy in the 21st century. It enables people to adjust computation to their needs.

Nowadays computer science and computational thinking is needed in all disciplines and professions beyond science and engineering. For example, computational economics and finance, computational social science, algorithmic medicine, computational archaeology, digital humanities, etc.

Computational thinking describes the mental activity in formulating a problem to admit a computational solution. It is a process involved in formulating problems and their solutions so that the solutions are represented in a form that can be effectively carried out by an information-processing agent, by a human or machine, or more generally, by combinations of humans and machines [21].

Everyone should learn a little computational thinking because it can be applied in daily life. Due to advances in technology, especially computer science and ICT, universities have to change curricula to promote computational thinking to all students, not just computer science majors.

Computational thinking for everyone means being able to:

- understand the limitations and power of computer - computational tools and techniques,
- understand what parts and aspects of a problem are amenable to computation,
- evaluate the match between computational tools and techniques and a problem,
- apply or adapt a computational tool or technique to a new use,
- recognize an opportunity to use computation in a new way,
- apply computational strategies such as division and conquer in any domain [21], [22], [23].

Computational thinking for professionals – scientists and engineers, means to:

- being able to apply new computational methods to solving their problems,
- reformulate problems to be amenable to computational strategies,
- discover new "science" through analysis of large data,
- ask new questions that were not thought of or dared to ask because of scale, easily addressed computationally,
- explain problems and solutions in computational terms [22].

Nowadays, immediate task ahead is to better explain to non-computer scientists what it means computational thinking and the benefits of being able to think computationally. The educational benefits are mostly in transferring to any domain by enhancing and reinforcing intellectual skills. Computer scientists see the value of thinking abstractly, at multiple levels, abstracting to manage complexity, abstracting to deal with scale, etc. [21], [22], [23].

Problem solving skills are considered to be an important part of teaching and learning according to many authors, and our experience, too. Many researchers concluded that via problem solving method the participants not only received knowledge, but actively participated in the educational process - at the same time they learned, associated and organized value to their acquired knowledge, built abstract knowledge, and adopted a belief system and a personal worldview.

Problem solving method enables students to create a personalized learning environment and a majority of students have agreed that this kind of educational activities enabled them to feel more confident and engaged into the learning process.

Many students have learned to solve problems in other disciplines by memorizing formulas and procedures. They often do this without an in-depth understanding of the essence of concepts and problems. Since the computer science and programming are oriented to critical and computational thinking, problem solving, but also gaining practical skills, students often mistakenly believe that it can be mastered mainly by studying textbooks without intensive practicing. This can be prevented by providing frequent opportunities during regular classes as well as in order to solve practical and real problems.

Problem solving requires a variety of skills that students often do not ad hoc possess sufficiently: understanding the problem, knowledge transfer, reflective thinking and perseverance. The teacher should ask the students to predict the results of certain activities, set new tasks to address some of the procedures used in previous solutions etc.

IV. Informatic Education at the Faculty of Economics TUKE

As teachers, we have been dealing with the problem of optimization and personalization our education in the subject Informatics II due to individual needs and knowledge base of different kinds of students to maximize development of their problem solving skills and computational thinking.

We are inspired by the social constructivist Vygotsky - [19] and constructionist Papert - [15] theoretical perception of the learning process. We are trying to apply pedagogical innovations in the forms of increased students' active role, in construction of learning resources, and learning environment. It is a good basis for consideration and deeper investigation of learning personalization process affordances enabling a learner to create his/her personal learning, as we can read in [7].

There are some important attitudes and characteristics of modern education (mentioned in previous parts) that we try to implement and develop:

- computational thinking,
- software engineering,
- personalization of learning,
- problem solving skills.

Everyone, not just those who study major in computer science, can benefit from thinking like a computer scientist. Computational thinking is used in the design and analysis of problems and their solutions. The most important and high-level thought process in computational thinking is the abstraction process. Abstraction is used in defining patterns, generalizing from instances, and parameterization [21].

Software engineering provides a number of processes, models, methods and techniques, as well as tools and notations. It is an area of computing that defines systematic, and quantifiable approaches for the development, disciplined operation and maintenance of software. [6].

The inclusion of software engineering practices can help students gain insight into some of the challenges in real software projects: user requirements development, design and development of software, user interface design, testing, configuration management, etc.

We agree with authors [8], [11], that educational/ learning process should be personalized. It is important to consider and follow the main characteristics/needs of the learners, i.e. prior knowledge, intellectual level, interests, goals, cognitive traits, learning behavioural type, and, finally, learning styles.

In determining both the content and the method of teaching Informatics II at our faculty, we consider an evaluation of the students' knowledge base in the field of Computer science as a very important first step. At the beginning of their study, a questionnaire as a research tool has been used. We search for both the extent and content of compulsory education in the field of Computer science/Informatics at secondary schools. We have been

carrying out our research since 2003, of course, with the necessary changes and modifications.

The sample of our research (in the mentioned period) consists of all secondary school graduates who entered their first year at Faculty of Economics Technical University of Košice. In more than 95 % of cases students in the studied sample are from the Eastern Slovakian Region.

There are several parts in our questionnaire which focus on different areas. Within the first part our students are asked the questions like: What type of secondary school did you attend? How many mandatory informatics lessons per week did you get over at secondary school in particular school years? In the second part of the questionnaire we focus on students´ knowledge in the field of basic terms of informatics, hardware architecture, operation systems, text editors etc. In the third part we want students to describe their experience in information systems (IS). The questionnaire evaluation results since 2003 were published in detail in e.g. [16]. Figure 1 presents evaluation of questionnaire in average percentage over the last three years.

One of the tasks of the questionnaire is: define and describe user requirements for a simple information system in a XY rental where we want to install new and effective electronic IS based on modern ICT instead of "old-styled/paper" one. We divided students´ solutions into categories as follows:

- A – acceptable; it was possible to recognize user requirements, there were structure and hierarchy between requirements, there were some graphical expressions.
- B – mostly continuous text with effort to design database tables or web site; partly incomplete.
- C – extensive continuous text with typical lack of exactness, structure, relationships; extent from ½ sheet to 2 sheets.
- D – text - few sentences, max. ½ of sheet; totally unsatisfactory.
- E – no answer.

Figure 2 represents evaluation of students´ user requirements in average percentage over the last three years.

Based on the questionnaire evaluation, we can say that the situation in the field of Informatics education at the Slovak secondary school is unfavorable. Teaching is flexible not enough for the requirements of the present, it absolutely does not meet the modern trends. Education of Computer science/Informatics at secondary schools is focused on handling different packages of applications´ programs. There is a lack of understanding of basic terms, of data processing, of modelling, of programing, algorithm, etc. which implies that computer thinking and problem solving skills are at a very low level. That is why during the second semester (via the subject Informatics II) students had to solve tasks involving basics of software engineering, computational thinking, problem solving.

As a dominant method we use problem-oriented project teaching, while we are using the functionality of LMS Moodle environment [13]. We try to develop the education with the potential to support our students by personalizing learning process due to the individual characteristics and needs. To achieve this, our system

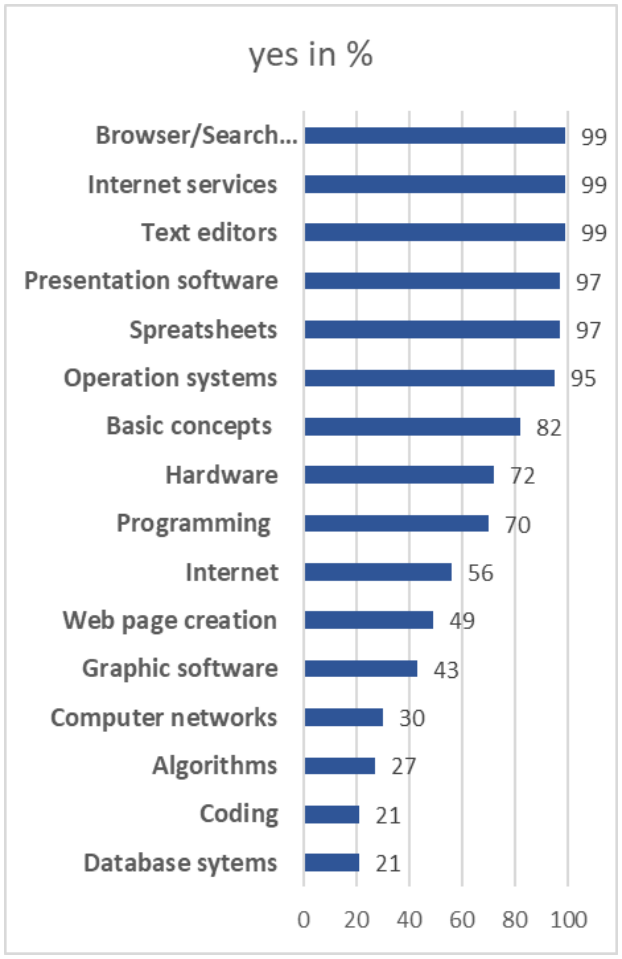


Figure 1. Percentage of students (in average) who worked with the terms, who worked in mentioned areas on secondary schools

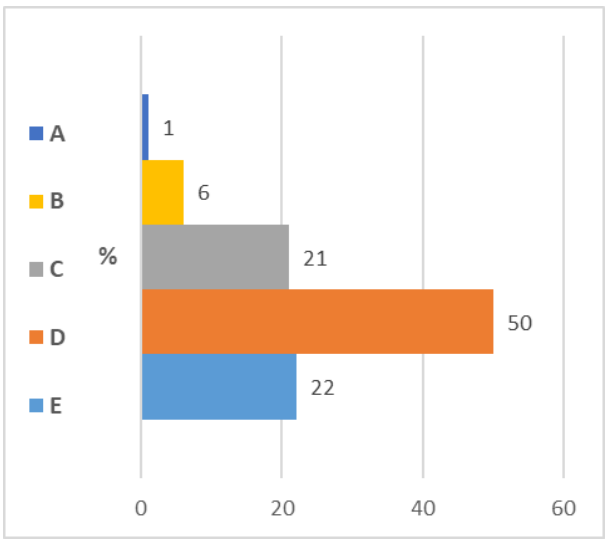


Figure 2. Percentage of students (in average) - evaluation of user requirements description

recommends learning materials, learning paths, consultations, teachers´ feedback and similar. The system presented mostly via recommended e-learning, activities and optional possibilities in a Moodle course provides personalization in the context of e-learning and learning by doing.

In this way of teaching/learning we are trying to transform the problem solving process into an active process of cognition. Our students are taking part in virtual projects of building IS in one of various types of virtual or real existing firms, companies or institutions to gain experience and skills via problem solving process. All our students find themselves in a position/role of IT executive such as Chief Information Officer (CIO) in ensuring continuous development and innovation of both IT applications and IT infrastructure. It offers us possibility for gamification of education that can motivate students to learn new concepts via game.

The recommended structure of the students' projects corresponds to ARIS methodology and consists of four parts focusing on computer science fundamentals including creation of proper models, algorithms, processes, analysis of user requirements, abstraction, division of problems into the parts, software engineering, computational thinking and systematic approach. We teach our students to take into account as well social aspects, their experience and creativity.

In the first part of project students specify the core business process by text and a graphic model. It represents analysis of the business and its processes, creating a basis for process management, strategic factors and goals, challenges, possibilities to support management of IS and ICT in the enterprise. In this part students have to think about the basic business model, the model the functional and organizational structure of the business, the product model of business processes and their parameters, and the structure of applications to support the business.

In the second part students create requirements on the IS in a written form. They describe the IS functionality in the third part using UML use case diagram. The basis for the IS data model is a proposal of classes and their attributes created as UML class diagram. The last part is a proposal of tables and their relations in MS Access environment, where students deal with logical structure of the database system.

Development of user requirements project parts mean creating a logical concept of information system in the chosen firm/organization, which includes the structure of the information processes, the organizational structure, the basic structure of the application, IS procedures, role of users, requirements of hardware and network infrastructure, security requirements, possibilities of IS/ICT further development, documentation and staff training requirements, operation and continuous improvement of processes.

To create new modern information system via project of specifying users' requirements, students have to think about how to differ from the others, what kind of innovations they should bring into their model of the main process of a chosen company. By using the method "learning-by-doing" they get knowledge of IS and its management together with experience on how to prepare the project proposals, how to communicate with others, how to present their ideas, design, suggestions. The ability of problem solving we consider as the most valuable skill.

V. DISCUSION

The European Commission Science Hub has been promoting computational thinking as an important 21st century skill or competence. Computational thinking could be added to the traditional three Rs: reading, writing and arithmetic as an additional basic skill needed especially by university students. They will be better prepared to choose a future career. However, it's impossible to predict exactly the skills that will be needed even five years from now. But educational institutions, their graduates as future employees/workers, firms and organizations need to be ready to adapt. It's important to have a clear view of what the future, driven by the megatrends and our own actions, will look like.

The graduate who has received high quality education is more adaptable, assured, innovative, entrepreneurial and employable in the broadest sense of the term. The graduate who has received poor education has wasted a lot of his/her time and money and in a competitive job market is at a disadvantage. In many cases, poor education contributes to the high dropout rates and lack of student success [4], [20].

Providing education is not enough. What is important, and what generates a real return on investment, is learning and acquiring skills. Schooling without learning is a terrible waste of precious resources and of human potential [26].

Personalization is a more and more studied issue. Its advantage in terms of time and motivations widely proved in the context of e-learning courses. Course personalization basically means to understand student's needs [11], [9].

VI. CONLUSION

We developed a way of teaching to support development of students´ computational thinking and problem solving skills. Our education is based on modelling, simulation and on a well-known opinion that "no kind of teaching can replace the personal experience". Students create interesting and relevant artefacts with the tools and techniques of computing and computer science, work with data using a variety of tools and techniques to better understand the many ways in which data is transformed into information and knowledge. Our education in the subject of Informatics II also includes examples of abstractions used in modelling the world, in managing complexity, and communication with people as well as with machines.

According to the Curriculum Guidelines for Undergraduate Degree Programs described in [1], our students are expected to demonstrate the following advanced user competencies:

- Professional knowledge - in formulating user requirements using the professional standards necessary for project managers.
- Technical knowledge - a basis for problem identification and analysis, verification, and documentation - an understanding of appropriate theories, models, and techniques.
- Teamwork - ability to work both individually and as part of a team to develop quality software artefacts.
- End-user Awareness - an understanding and appreciation of the importance of negotiation, effective work habits, leadership, and good communication with stakeholders.
- Design Solutions in Context - appropriate solutions in one or more application domains using approaches that integrate ethical, social, legal, and economic concerns.
- Perform Trade-Offs - finding acceptable compromises within the limitations of cost, time, knowledge, existing systems, and organizations.

To compete in the economy of the future, workers need strong basic skills and foundations for adaptability, creativity, and lifelong learning. Education raises human capital, productivity, incomes, employability, and economic growth [26]. But there are much more benefits of proper and modern education, which go far beyond monetary gains: education also makes people healthier and gives them more control over their lives. And it generates trust, boosts human and social capital, and brings prosperity.

ACKNOWLEDGMENT

The research was realized within the national project "Decision Support Systems and Business Intelligence within Network Economy" (Contract No. 1/0201/19) funded by Grant Agency for Science; Ministry of Education, Science, Research and Sport of the Slovak Republic.

REFERENCES

[1] ACM/IEEE, "Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering", 2014, [online] https://www.acm.org/binaries/content/assets/education/se2014.pdf

[2] D. Autor, and D. Dorn, "The growth of low skill service jobs and the polarization of the US labor market", American Economic Review 2013, 103(5): 1553–1597

[3] M. Akcaoglu, "Design and Implementation of the Game-Design and Learning Program", TechTrends, 2016, 60 (2), 114–123, https://doi.org/10.1007/s11528-016-0022-y

[4] European Commission, "Report to the European Commission on Improving the quality of teaching and learning in Europe's higher education institutions", Luxembourg: Publications Office of the European Union 2013, doi:10.2766/42468

[5] C. B. Frey and M. A. Osborne, "The future of employment: How susceptible are jobs to computerisation?" 2013, [online] https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

[6] IEEE CS. SWEBOK, "Guide to the Software Engineering body of knowledge" (3th ed.). Silver Spring,USA: IEEE, 2014.

[7] N. Ignatova, V. Dagiene, and S. Kubilinskiene, "ICT-based Learning Personalization Affordance in the Context of Implementation of Constructionist Learning Activities Informatics in Education", 2015, Vol. 14, No. 1, 51–65 © 2015 Vilnius University DOI: http://dx.doi.org/10.15388/infedu.2015.04

[8] T. Jevsikova, A. Berniukevičius, E. Kurilovas, E., "Application of Resource Description Framework to Personalise Learning: Systematic Review and Methodology", Informatics in Education, 16(1), 61–82., 2017

[9] C. Limongelli, F. Sciarrone, "Fuzzy Student Modeling for Personalization of e-Learning Courses." In: Proceedings of the 1st International Conference on Learning and Collaboration Technologies (LCT), Heraklion, Greece, JUN 22-27, 2014. Zaphiris, P. and Ioannou, A. (Eds) Learning and Collaboration Technologies: Designing and Developing Novel Learning Experience, PT I, Lecture Notes in Computer Science, 8523: 292–301.

[10] B. Marr, "Why Everyone Must Get Ready For 4th Industrial Revolution", 2019, [online] https://www.bernardmarr.com/default.asp?contentID=966

[11] J. Mamcenko, E. Kurilovas, and I. Krikun, "On Application of Case-Based Reasoning to Personalise Learning Informatics in Education",Vol. 18, No. 2, 345–358 © 2019 Vilnius University DOI: 10.15388/infedu.2019.16, 2019

[12] P. McRae, "The politics of personalization in the 21st century". ATA Magazine, 91 (October 1) 2010

[13] D. Paľová, "Experience with Usage of LMS Moodle not Only for the Educational Purposes at the Educational Institution" In: MIPRO 2016. - Rijeka p. 1006-1011.

[14] G. Polya, "How to Solve it", Princeton University Press, New York. 1957

[15] S. Papert, "The Children's Machine: Rethinking Schools in the Age of the Computer". Basic Books, New York, 1993

[16] L. Révészová, "Designing modern informatics education for future managers and advanced users according to their knowledge base," E + M:Ekonomie a Management,Vol. 19, no. 4, 2016, p. 186-201.

[17] P. M. Romer, "Endogenous Technological Change." Journal of Political Economy 98 (5): 71–102, 1990

[18] R. M. Solow, "A Contribution to the Theory of Economic Growth." Quarterly Journal of Economics 70 (1): 65–94, 1956

[19] L. S. Vygotsky, "Mind and Society: The Development of Higher Mental Processes", Harvard University Press, Cambridge, MA, 1978

[20] M. Vejačka, "Electronic commerce entrepreneurship education using virtual laboratory" In: MIPRO 2019, Rijeka, Croatian Society for Information and Communication Technology, Electronics and Microelectronics s. 670-675

[21] J. M. Wing, "Computational Thinking," Communications of the ACM, CACM vol. 49, no. 3, March 2006, pp. 33-35.

[22] J. M. Wing, "Computational Thinking: What and Why?" [online] www.cs.cmu.edu/~CompThink/resources/TheLinkWing.pdf

[23] J. M. Wing, "Computational thinking's influence on research and education for all". Italian Journal of Educational Technology, 2017, 25(2), 7-14. doi: 10.17471/2499-4324/922

[24] Workforce of the future The competing forces shaping 2030, 2018, [online] https://www.pwc.com/gx/en/services/people-organisation/publications/workforce-of-the-future.html

[25] The World Bank, "Constructing Knowledge Societies: New Challenges for Tertiery Education" Washington, DC: The World Bank, 2002.

[26] The World Bank, "World Development Report 2018: Learning to Realize Education's Promise", Washington, DC: World Bank, . 2018. doi:10.1596/978-1-4648-1096-1.

Game design based learning of programming for girls

M. Bevčič, J. Rugelj
University of Ljubljana
Mateja.Bevcic@pef.uni-lj.si

Abstract - We will present the use of game design as a method for teaching programming in primary and secondary schools. Students of the introductory course in computer programming design and program games. This represents a typical active learning approach that has proven to be the most effective way of learning programming. The act of building a game is itself a way to learn, regardless of whether the game is interesting to other people or not.

The described methodology and all corresponding learning materials for students and supporting materials for teachers are being developed within the Erasmus+ Coding4Girls project, which addresses equal opportunities for girls and boys in the highly rewarding computer science professions by highlighting links between programming and real life and highlighting the fact that such professions are rewarding regardless of gender or background. The results of the project will help to create the necessary background among the students that will enable them to succeed in their future endeavours either academic or professional. It will also address the development of teachers' competencies and the profile of the teaching profession by empowering educators to effectively build up the desired computer literacy skills among their learners.

Keywords - Game design based learning, Design thinking, Programming, Programming for girls

I. INTRODUCTION

Games are now widely used in the classroom, especially in the earliest stages of a child's development. However, in the later stages of student development, learning is still too often based on a teacher centred activities in which students play a passive role. On the other hand, learning theories developed over the last century encourage new student-centred approaches to learning and teaching that are problem based and motivating, directed to higher pedagogical and taxonomic levels and often supported by information and communication technologies (ICT) [1].

II. GAME-BASED LEARING

Serious games should be designed to facilitate learning goals [2], [3]. In most cases, the game is not an independent learning activity, but a part of the learning activity that can be carried out before, during or after the game.

- *Pre-game activities*
  We can use them as an introductory activity in which the teacher presents the rules and goals of the game as well as the context and learning goals. If students recognize rules and goals by themselves, these activities can be left out, and the player is also more active and involved in the game.
- *Activities during the game*
  They are used less frequently and are intended to support the player when problems in the game are too demanding or when additional feedback is needed.
- *Post-games activities*
  The players must have the opportunity to discuss and reflect with their teacher and to contribute their own experiences with the game. These activities are necessary for the students to see what they have learned in the game, what knowledge they have acquired and what could be improved to achieve their learning goals [1].

According to Kolb [4], the opportunity for learners to reflect on their own learning experiences is crucial to the learning process. In this way students reflect on their activities and explore what and why they do. For effective learning, [4] Kolb suggests in his book a Learning Cycle of four stages that the student must go through: 1) to have a *concrete experience*, 2) to review and reflect on the experience (*reflective observation*), 3) to infer and learn from the experience (*abstract conceptualization*), and 4) to plan and try out what has been learned (*active experimentation*).

All the above mentioned activities are collected in a so-called *pedagogical package*, which has to be prepared and offered to teachers who want to use game-based learning for their students [1].

Although serious games require one or more learning goals to be defined, they must also have all the characteristic of the game. Prensky in his book entitled *Digital Game-based Learning* [5] defines seven key elements of the game: 1) *Story* - the game must be based on the story, which provides a framework and connects parts of the game into a whole, 2) *Objectives and challenges* – the objectives must be related to the story and challenges, 3) *Additional challenges* – the game rules give structure to the game and some additional challenges are recommended, 4) *Interaction with other people* – the player has active interaction with other people or within the game environment, 5) *Conflict relationship* – the relationship with others is usually conflictual or competitive, 6) *Control over events* – when interacting, the player has a sense of control over the events and also over the possibilities of influencing the course of the

game, 7) *Feedback* – the game must provide feedback to the player so that he can check the appropriateness of his actions and the effectiveness of his progress. All these key elements motivate the player, help him stay focused and have fun.

In the game, the player is guided to the learning goals that are normally hidden and not obvious in the game. Therefore, the learning goals are achieved without the player being aware of the learning [6]. Nevertheless, game-based learning is not suitable for achieving learning goals in all subject areas. The appropriateness and usefulness of the game for learning can be strongly influenced by the following factors: subject area, learning theory and didactic method, taxonomic level, student's background and teacher's experience and expertise [7].

Serious games can offer students many learning benefits - risk taking, problem solving, interaction, situational meaning, research, teamwork, and critical thinking. Studies by Ibrahim and Jaafar [8] have shown that students are very interested in using educational games for learning. The results from Nor Azan and Wong [9] indicated that 92% of the students included in their survey had experience with computer games, that they play games for a variety of reasons, and suggested that games could be used as an alternative approach to teach subjects with lower student motivation. According to a number of studies [10], [11], [12], [13] games motivate learners, provide immediate feedback, support skills and influence behavior and attitudes.

III. CODING4GIRLS PROJECT

Coding4Girls is an ongoing international project that addresses the gap between girls and boys in computer science education and careers. According to the European Commission, only 2.9% of women with a Bachelor or other degree hold a degree in ICT (compared to 9.5% men) and only 0.4% women eventually work in ICT. The Coding4Girls project aims to make computer science attractive to all and focuses on the late years of primary education up to the early years of secondary education (10-16 years), when many girls lose interest in STEM (science, technology, engineering and mathematics), including computer science [14]. Coding4Girls is also concerned with the development of teaching competencies and the profile of the teaching profession. Teachers guide their students in the learning process by constructing solutions to the given problems and implementing them through game design [1].

Students participating in the project will: be prepared for computer science careers by building programming skills; be able to apply developed programming skills in larger contexts; build analytical and critical thinking; learn how ICT solutions can improve quality of life and address common needs; learn to think entrepreneurially to introduce solutions to real-world problems through design thinking mind sets [14].

IV. GAME DESIGN BASED LEARNING

Within the Coding4Girls project, the students will learn programming concepts by creating games. According to Kafai [15], students become even more active in game-based learning when they learn by developing games themselves. Also Zapušek and Rugelj [16] argue that learning by creating games can be more effective than traditional methods.

The act of building the game is itself a way of learning, regardless of whether the game is interesting to other people or not. Due to the expanding range of computer-based design and game design tools, the idea of "learning by design" has become widespread, based on the assumption that the best way to learn is to actively participate in the design and development process [17].

Between 2014 and 2015, extensive studies of published papers on the topic of children's learning while making games were conducted by Kafai and Burke [18]. Convincing evidence has been found that young children use different computer practises when developing games. In the project in which the fourth grade students designed the fraction game, they debugged, revised and tested their games again and again, especially after the evaluation sessions with younger students. They not only programmed the game, but explicitly focused on teaching fractions to younger students. Post-tests showed that the students who programmed the fraction game were much more familiar with programming in Logo (and also with understanding and representing fractions) compared to students who learned Logo in small, unrelated programming activities. These results have been confirmed in recent studies on the creation of mathematics games in Scratch by Ke [19], as students have been confronted with mathematical tasks on a daily basis and have improved their understanding. In another analysis, the students were asked to create different game designs. 30 college students in a semester-long course created 268 games and 33 middle school students in an 8-week class created 73 games. It turned out that both groups improved their computational thinking patterns, but the improvement for the college students was more substantial, as they started the course with more experience and also spent more time designing their games [18].

Serious games are also used to learn complex and difficult to understand topics in computer science education, such as data structures and algorithms. It turns out that students better understand the algorithms when they apply a new learning strategy for designing computer games that visualize algorithms [20]. For example, every computer game that teaches and visualizes algorithms has six common characteristics, which is also true for other serious games: 1) the game can be created by developing a new game from the beginning or by modifying an existing game, 2) the game must be simple enough so that the students do not lose concentration, but at the same time challenging enough so that they do not get bored, 3) the algorithm game must visualize the algorithm and its data structure and can be of any genre, 4) the game must be motivating and challenging, with clear goals, appropriate difficulty levels and feedback. The information in the game must be complex and unknown, which increases the curiosity of the players. The game must also increase competition and cooperation between the students, 5) the game graphics should reflect the characteristics of the

structure that the students are learning (e.g. using a group of boxes with a content to learn about the array data structure, or using an actual tree with leaves to learn about the tree data structure), 6) the algorithm game must simulate the behavior of the algorithm that the game visualizes.

V. INTERESTING GAMES FOR GIRLS

Before we prepared Snap! activities and wrote learning scenarios, we had to consider what kind of games would be interesting for girls.

Vermeulen et al. [21] stated that women play games more often, but less long than men, and that there are also different preferences for game genres between the genders. Female players like to play social games [22] with in-game rewards [23], educational and simulation games [24], cooperation and exploration games [25], virtual world games and party games [26], while men prefer action, strategy, role-playing and fighting games. An analysis of studies [27] reveals the following elements regarding the playing preferences of women:

- Explore: Women like role-paying and new identities, the exploration of new worlds and the discovery of new knowledge [28]. Studies have shown that women have a preference for exploratory play [23].
- Character customization: Women like to individualize their avatars, which includes customization of clothing, weapons and accessories. Role-playing games are known for these customization options [29].
- Storyline: The storyline is a main motivating element. It allows the player to interact with the characters and situations in the game [30] and women like to play games with a good story [29].
- Social interaction: Social interaction increases motivation and represents the element of collaboration and cooperation [31] and is a major factor for women to play games [32], [33].
- Collaboration: Collaboration is important for motivation and is one of the most popular forms of collaboration [30].
- Challenges: Challenge is a key component [34] and is one of the many motivating elements of the game for women [31].
- Fun: Fun and entertainment are a central motivation for women [34], [31].
- Control: Control also affects motivation [34] and the commitment and performance of players [35]. Women prefer to control the game, rather than the game controlling them [29].
- Feedback: Environments that provide feedback to the player improve engagement and performance [35].

In the No One Left Behind (NOLB) project, the students created 77 games (girls created 37 games, boys 34 and 8 games were created in mixed teams). Students aged 8 to 12 years developed games in the subject areas such as English, computer science and fine arts. The games made by girls were mainly adventure games - nature, winter wonderland, underwater world (33%), followed by text adventure games / storytelling - fantasy, nature, romance (24%), puzzle games - fantasy, crime story (17%), platform games - nature (15%), action games - nature, underwater world (9%) and action shooter / role-playing games - fantasy, realistic (4%). The games created by boys were also predominantly adventure games - space, nature, realistic (35%), followed by action shooter games – space, fantasy (22%), puzzle games - realistic, fantasy, fiction (19%), quiz - realistic (9%) and action games – nature, space (15%). The results show significant differences between gender and genre. Girls also used a nature-related theme most often (43%), while boys used a space theme most often (46%). The use of graphics for the main character shows even greater differences in gender. In games created by girls, the main characters were usually animals - dog, panda, butterfly, etc. (13 games), followed by fantasy - skeleton, fairy, Elsa, unicorn etc. (11), items (4), female characters (3), male characters (3), food (1) and no transport. In the games made by boys, the main character was usually transport - spaceship, UFO (12 games), followed by male character (6), animals - bird, dog (4), fantasy (3), food (2), items (2) and female characters (1) [36].

VI. LEARNING SHEETS FOR TEACHING PROGRAMMING

When writing learning scenarios, we took into account the goals of the Coding4Girls project and all the motivational elements and preferences of the girls in relation to the games. We have also tried to relate the activities to real-world problems.

The Coding4Girls approach encourages participation in programming activities through a "low-entry, high-ceiling approach", which initially places low demands on knowledge, but also includes complex challenges for more advanced students. The students solve partially completed solutions prepared in the programming language Snap! (similar to Scratch), adding missing blocks of code or creating their own solution [14].

Prepared learning scenarios contain all the information that will help teachers to integrate proposed serious games and design thinking learning methods into the learning process. The learning sheets contain all the following information:

- overall educational objective for each learning activity,
- concepts covered by the learning activity,
- specific learning objectives,
- expected learning outcomes,
- step-by-step use of the Coding4Girls game design-based learning approach,
- assessment methods for evaluating the developed knowledge,
- questions to stimulate discussion among learners [14].

More than 20 learning sheets for planned learning activities were prepared. Teachers will be able to use the scenarios and games in the suggested order, but they will be also able to customize the scenarios to suit their needs. The learning sheets cover the generic functionality of the proposed serious game, including user interaction processes, Snap! activities and half-baked scenarios with descriptions of all learning activities implemented in the proposed serious game [14].

The activities are divided into two parts. The first part contains basic activities covering one programming concept, followed by more advanced activities in several programming concepts. The duration of the basic activities is usually 45 minutes, while 90 minutes are provided for advanced activities. Students can design their games individually, in pairs or in groups.

In the initial activities, students familiarize themselves with the Snap! visual programming environment, followed by finding programming blocks, connecting these blocks into a sequence, moving a sprite, creating sprites that say something, and creating a meaningful sequence of blocks. Students learn how to change costumes and turn sprites by making a ballerina dance, how to add sounds by adding the voices of farm animals, and how to implement the chameleon's movement by using events and changing its colour according to the background by using conditionals. By helping the prince and princess find their animals in the maze, they learn how to draw a path, and by connecting dots on the board, they learn how to "write" a code to draw with loops. They meet with variables and duplicate sprites within a game where they pick up trash, clean up the park and count points. In the games where cats are fed and the number of cats in a shelter is guessed, random numbers and operations are introduced. There are more challenging activities such as catching healthy food, planting trees for a clean environment, inventing a story, catching the mouse, buying food for a picnic, calculator, recycling, playing a piano, testing, and Packman game.

The learning sheets will be published in the Coding4Girls web portal[1] and will be available in English and in all national languages of the project partners: Bulgarian, Croatian, Greek, Italian, Portuguese, Slovenian, and Turkish.

VII. CONCLUSION

The learning scenarios developed within the Coding4Girls project will be tested in different ways in primary and secondary schools. The test results will reveal our performance in composing activities, taking into account the girls' preferences in playing and creating games and their motivation in learning programming by creating games.

ACKNOWLEDGMENT

This research was partly funded by the Erasmus+ programme of the European Union project Coding4Girls (2018-1-SI01-KA201-047013).

[1] Coding4Girls web portal: https://www.coding4girls.eu/

REFERENCES

[1] J. Rugelj and M. Lapina, "Game design based learning of programming," Proceedings of SLET-2019 – International Scientic Conference Innovative Approaches to the Application of Digital Technologies in Education and Research, Stavropol – Dombay, Russia, 20-23 May 2019.

[2] M. Zapušek, Š. Cerar, and J. Rugelj, "Serious computer games as instructional technology," 2011 Proceedings of the 34th International Convention MIPRO, IEEE, 2011.

[3] M. Zapušek and J. Rugelj, "Learning programming with serious games", EAI Endorsed Transactions on Serious Games 1.1, 2013.

[4] D. A. Kolb, "Experiential learning: Experience as the source of learning and development," New Jersey: Prentice-Hall, 1984.

[5] M. Prensky, "Digital Game-based Learning", New York: McGraw-Hill, 2001.

[6] K. M. Kapp, "The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education", San Francisco: Pfeiffer, 2012.

[7] N. Hoić-Božić, M. Holenko Dlab, L. Načinović Prskalo, J. Rugelj, I. Nančovska Šerbec, "Games for learning algorithmic thinking - GLAT project", Proceedings [Colombo]: [Global Academic Research Institute], 2018.

[8] R. Ibrahim and A. Jaafar, "Using educational games in learning introductory programming: A pilot study on students' perceptions.", 2010 International Symposium on Information Technology. Vol. 1. IEEE, 2010.

[9] M. Nor Azan and S. Wong, "Game based learning model for history courseware: A preliminary analysis," 2008 International Symposium on Information Technology. Vol. 1. IEEE, 2008.

[10] D. A. Lieberman, "Management of chronic pediatric diseases with interactive health games: theory and research findings," The Journal of Ambulatory Care Management, 24.1, 2001, pp. 26-38.

[11] M. K. Virvao, G. Katsionis, and M. Konstantinos, "Combining software games with education: Evaluation of its educational effectiveness," Educational Technology & Society, 8.2, 2015, pp. 54-65.

[12] J. Hill and M. Y. Tollett, "Impacts of playing video games on learning in children," Literature Synthesis for Applying Research, M.Y. Tollett, Editor. 2006, The University of Georgia, Athens, Georgia.

[13] M. Sugimoto, "What can children learn through game-based learning systems," 2007 First IEEE International Workshop on Digital Game and Intelligent Toy Enhanced Learning (DIGITEL'07). IEEE, 2007.

[14] Coding4Girls, Project documentation, Erasmus+ Application form, 2018-1-SI01-KA201-047013, 2018.

[15] Y. B. Kafai, "Minds in play: Computer game design as a context for children's learning", Mahwah, NJ: Erlbaum, 1995.

[16] M. Zapušek and J. Rugelj, "Achieving teachers' competences in the serious game design process", 8th European Conference on Games Based Learning: Berlin, Germany 9-10 October 2014.

[17] J. Rugelj, "Serious games design as collaborative learning activity in teacher education," 9th European Conference on Games Based Learning: Steinkjer, Norway 8-9 October 2015.

[18] Y. B. Kafai and Q. Burke, "Constructionist gaming: Understanding the benefits of making games for learning," Educational Psychologist, 50.4, 2015, pp. 313–334.

[19] F. Ke, "An implementation of design-based learning through creating educational computer games: A case study on mathematics learning during design and computing," Computers & Education, 73, 2014, pp. 26–39.

[20] S. S. Shabanah, J. X. Chen, H. Wechsler, H., D. Carr, D., and E. Wegman, "Designing computer games to teach algorithms," 2010 Seventh International Conference on Information Technology: New Generations. IEEE, 2010.

[21] L. Vermeulen, J. Van Looy, F. De Grove, and C. Courtois, "You are what you play?: A quantitative study into game design preferences across gender and their Interaction with gaming habits," DiGRA 2011 Conf. Think Des. Play, 2011.

[22] S. A. Wallace, R. McCartney, and I. Russell, "Games and machine learning: a powerful combination in an artificial intelligence course," Computer Science Education 20.1, 2010, pp. 17-36.

[23] K. R. Hamlen, "Children's choices and strategies in video games," Computers in Human Behavior 27.1, 2011, pp. 532-539.

[24] M. H. Phan, J. R. Jardina, S. Hoyle, and B. S. Chaparro, "Examining the Role of Gender in Video Game Usage, Preference, and Behavior," Proceedings of the human factors and ergonomics society annual meeting. Vol. 56. No. 1. Sage CA: Los Angeles, CA: Sage Publications, 2012.

[25] L. Vermeulen and J. Van Looy, "'I Play So I Am?' A Gender Study into Stereotype Perception and Genre Choice of Digital Game Players," Journal of Broadcasting & Electronic Media 60.2, 2016, pp. 286-304.

[26] B. D. Homer, E. O. Hayward, J. Frye, and J. L. Plass, "Gender and player characteristics in video game play of preadolescents," Computers in Human Behavior 28.5, 2012, pp. 1782-1789.

[27] S. A. Alserri, N. A. M. Zin, and T. S. M. T. Wook, "Gender-based Engagement Model for Serious Games," International Journal on Advanced Science, Engineering and Information Technology 8.4, 2018, pp. 1350-1357.

[28] A. L. Krassmann, L. N. Paschoal, A. Falcade, and R. D. Medina, "Evaluation of game-based learning approaches through digital serious games in computer science higher education: a systematic mapping," 2015 14th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames). IEEE, 2015.

[29] E. Adams, "Fundamentals of Game Design," Design, Third Edit., New Riders, 2013, p. 576.

[30] A. I. Abdul Jabbar and P. Felicia, "Gameplay engagement and learning in game-based learning: A systematic review," Review of educational research 85.4, 2015, pp. 1–40.

[31] E. A. Van Reijmersdal, J. Jansz, O. Peters, and G. Van Noort, "Why girls go pink: Game character identification and game-players' motivations," Computers in Human Behavior 29.6, 2013, pp. 2640-2649.

[32] J. Rugelj, "Serious computer games in computer science education," EAI Endorsed Trans. Serious Games 2.6, 2015.

[33] S. A. Alserri, N. A. M. Zin, and T. S. M. T. Wook, "Gender-based engagement model for designing serious games," 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI). IEEE, 2017, p. 5.

[34] B. Pourabdollahian, M. Taisch, and E. Kerga, "Serious games in manufacturing education: Evaluation of learners' engagement," Procedia Computer Science 15, 2012, pp. 256-265.

[35] K. L. McClarty, A. Orr, P. M. Frey, R. P. Dolan, V. Vassileva, and A. McVay, "A literature review of gaming in education," Gaming in education, 2012, pp. 1-35.

[36] B. Spieler and W. Slany, "Game development-based learning experience: Gender differences in game design," Proceedings of the 12th European Conference on Games Based Learning, Sophia Antipolis, France, 03-05 October, 2018.

Gamification tools improving university students' involvement in the education process

D. Paľová*, M. Vejačka*
* Department of Applied Mathematics and Business Informatics
Faculty of Economics, Technical University of Košice, Slovakia
dana.palova@tuke.sk, martin.vejacka@tuke.sk

***Abstract* – Information and Communication Technologies (ICT) influence the field of education in a significant way. Technology makes education more efficient, scalable, accessible and interactive. Current students' generation, the Millennials, ask for different innovative pedagogical approaches. They like technology and it is the natural part of their everyday life as breathing. From the conducted surveys follows, that they are group-oriented and practiced users of digital technology. In order to impress and motivate them to learn, it is necessary to employ ICTs differently and interactively and it does not matter the age. The gamification of learning is a kind of educational approach, which uses game elements in a learning environment. The main aim is to enhance students' interest in learned topics and to inspire them to continue learning. At our faculty, education innovation is one part of the internal development project. The paper will present the short review of tools and methods usable for the gamification of the education process, followed by its implementation to the particular course Informatics II provided by LMS Moodle.**

Keywords – education process innovation; LMS Moodle; gamification; enhancement of students' motivation;

I. INTRODUCTION

The well-known fact is that at present, the world is highly dependent on Information and Communication Technologies (ICT). They influence all aspects of our life, not omitting the education. The current young generation is more "wired" and as ICT natives they naturally expect, that education will use all technologies known by them also in their education process [1]. The technology is part of their lifestyle and identity [2] and brings a huge amount of information by free. This dramatically changes the distribution of "power" between teachers and students. Nowadays, the teacher is not one and only representative of knowledge in the special topic, but his role was shifted to the role of a guide during the students' exploring the knowledge and achieving new skills. The main challenge of education seems to be teaching students how to find and navigate information and critically interpret it and if it is needed, use it by an innovative and creative approach.

According to [1], [3] the young generation spend on average 6.5 hours using ICT and other similar media (e.g. TV, DVD), retrieving new information, listen and record music, view, create and publish Internet content; play video games. It is difficult to compete with such an incentive, which, besides being of interest to the student, also brings content to him in a fetching manner. In connection with that fact, one of the hardest challenges of the teacher is to motivate students to be more involved in education within his field of the subject taught.

So, if we want to attract a student, we have to take several of their basic characteristics into account: they like to be in control, they like choice, they are group-oriented and social, they are inclusive, they are practiced users of digital technology, they think differently, they are more likely to take risks and they value time off because they look uncertain [3]. Only if we consider these aspects when creating the content and structure of education, do we have a chance to succeed and thus increase students' interest in education. In order to achieve the increase of students' involvement in the education process and the effectivity of learning, the teacher needs to:

- Set clear expectations of what needs to be reached before graduating the course and allow students to manage their education within the learning environment by using a mixture of different internal, external information resources, activities and interactions across the class [4]. It is important to provide analytics of students' activity tracking in order to achieve more detailed insight into the learners, their development needs, which presents the input for the continuous education improvement [5].
- Choose the right combination of modern technologies as LMSs, different kinds of games, Virtual Reality, Augmented Reality, 3D printing or chatbots, etc. The main challenge from that point of view is to use this kind of technology that is supported and available via smart devices. Implementing this, it offers the convenience to students to study anywhere and anytime, which results in their involvement in the education process not only at university but also outside of it. [6]

Implementing modern technologies into education goes hand in hand with the gamification of education. The gamification of education brings great potential for improving the performance, motivation, and engagement of students. The gamification of the education results in students' devotement more time and greater involvement to study what results in the improvement of the knowledge and skills of the students.

II. GAMIFICATION OF EDUCATION

The concept of education gamification is not so fresh. This approach was mentioned first already in the 80-ties of the previous century by the Malone [7], [8] and Bowman [9]. Chapman [10] presented the gamified learning process as a collection of activities taken in reality but supported by the ICT system to discover and validate activities, view progress and communicate and collaborate with other players. Gamification, in general, is presented as "*an integration of game elements and game thinking in activities that are not games*" [5], [11]. Games have some distinctive features which play a key role in gamification and are used to motivate and facilitate learning process: *narrative and users* (presented by all participants); *challenges/tasks* that users perform and progress towards defined objectives; *progress mechanism* that covers points that are accumulated as a result of executing tasks, levels which users pass after collecting predefined amount of points, badges which serve as rewards for completing actions (that can be repeated in order to reach the goal) and public ranking of users according to their achievements; *player control; immediate feedback; social connection and opportunities for collaborative problem-solving*. [5], [12]

The main idea of gamification education is to capture the learners' interest and inspire them to continue learning and help them to focus on meaningful learning tasks [13]. The appropriate composition of tasks helps students realistically estimate their chances of success and dividing complex tasks into shorter and simple subtasks help them to deal with the complexity of tasks. From that point of view, the structure of a course must be changed at different levels: change of roles of the students, teacher and learning environment, change of the language used in the course and finally yet importantly, change of the content of the course. By gamification of roles, a student might be adopting an avatar and the avatars might be grouped into guilds by the teacher (or some other rule), a teacher is responsible for design gamification, embedding game dynamics and providing rewards attractive to the students. [14] Changing the language of the course means, that instead of official terms, a teacher can describe the learning objectives and tasks as "quests", and process of teaching it as "completing a mission". At the same time, also results could be presented in knowledge points (KPs) experience points (XPs), health points (HPs) or badges, items, etc. All of this aims to motivate students to participate in the education process and to increase their interest in course content and activities and not pay their attention to the letters that they can achieve in the grade book for participating in the "game".

A. Gamification in LMS Moodle

LMS Moodle is one of the most popular learning platforms. Besides providing an interactive environment for communication and cooperation among all education process participants, it offers different features for gamification of the learning process. Based on the previously listed facts, it could be generalized that LMSs are suitable environments for education gamification. They provide tools for automatic tracking of students' results and progress. Using the discussions, forums, and blogs, taking part in creating the content of wiki pages, knowledge databases are students encouraged to take part actively in developing learning content.

Based on [15], the gamification of the learning process has the potential to make learning fun. According to [16], 80% of LMS users believe, that gamification would help them be more productive. Based on this survey, students presented a positive attitude towards gamified learning and that the LMS environment is more engaging, challenging and competitive.

LMS Moodle provides more in-build capabilities to gamify the learning process: user's profile contain a field for uploading a photo, so students can define them as an *avatar*; possibility to set *learning paths* – the multiple activity completion conditions/criteria which must be met by the students in order to access the activity; *the progress of the students in the course is visible* (teacher can it visible also by the progress bar like in games), that help them to understand that their actions are connected in a greater whole; *top ranking of the best students* in the quizzes or any set activities (what adds competitive nature of the learning); *instant feedback* on their tests, assignments or other activities help to motivate the students and encourage them in the action; learners could be *rewarded by badges* upon completion of several activities or for achieving a certain level of knowledge and competence. [5]

Besides above mentioned in-build possibilities of how to gamify the education in the LMS Moodle environment, a teacher has available different plugins and modules that allow creating a more gamified environment within a particular course. The most popular are the following [15]:

- *Level up!* is a customizable block that a teacher can add to a course to give experience points to students as they progress through a course. It displays their current level and progresses towards the next level.
- *Activities: Game* module makes use of questions, quizzes and glossaries to create offer a variety of interactive games hangman, crossword, cryptex, millionaire, Sudoku, Snakes, and Ladders, The hidden picture, Book with questions. The results of the games are part of the grade book. Using these games teachers can slightly force to study available study materials if they want to achieve the best results.
- *Quizventure* is a module that loads quiz questions from the course. The possible answers come down as space ships and you have to shoot the correct one. The quizventure is a more attractive way how to motivate the students to self-study also at their homes.
- *Stash* is a block that allows a teacher to create and then show items around a course. Students can then go and collect these items, which will then appear in their stash block. The stash block is a good way to encourage more interaction with activities and is invaluable for teachers looking to gamify their course.

B. *Gamification tools usable outside LMS Moodle*

Previously mentioned tools are the part of the LMS Moodle, which presents the on-line part of the education process. Nevertheless, what kind of tools could be used, while the teacher wants to motivate students to be more active during face-to-face learning? At the market, there are tools outside the LMS Moodle environment, but usable for education gamification. They aim to "wake up" students' attention during the lecture or lesson in different ways:

- Mentimeter is an on-line tool available at https://www.mentimeter.com/ by which you can build interactive presentations with the easy-to-use online editor. It provides the possibility to add questions, polls, quizzes, slides, images, gifs and more with the main objective - to create fun and engaging presentations. When a teacher presents the presentation, the audience (students) uses their smartphones to connect to the presentation where they can answer questions, give feedback and much more. Finally, the teacher can visualize their responses in real-time (Figure 1.). [17]



Figure 1. Example of the final visualization of the audience responses on the question "Why use a word cloud?"; source: [17]

- sli.do (available at https://www.sli.do/) is another online tool that empowers students to ask questions, vote in polls and is a part of the discussion by using a simple Q&A and polling tool in real-time (Figure 2.). Live polls are an easy way to kick-start the conversation, check knowledge or get instant feedback. A teacher can use it to warming up the audience's attention, to recap the content and to entertain or test the attendees. [18]

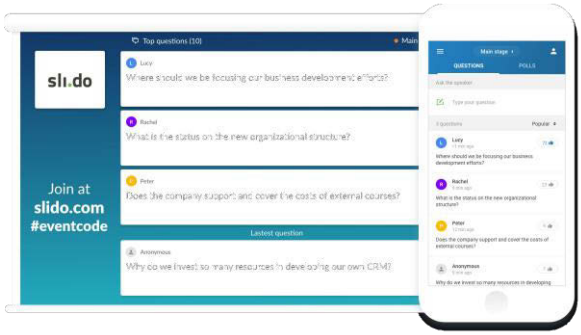


Figure 2. Questions and votes in Live polls in Slido application; source: [18]

- Kahoot! (available at https://kahoot.com/) is a game-based learning platform, used as educational technology in schools and other educational institutions. It provides learning games like multiple-choice quizzes trivia quizzes (Figure 3.). Kahoot! can be used to review students' knowledge, for formative assessment, or as a break from traditional classroom activities. [19]

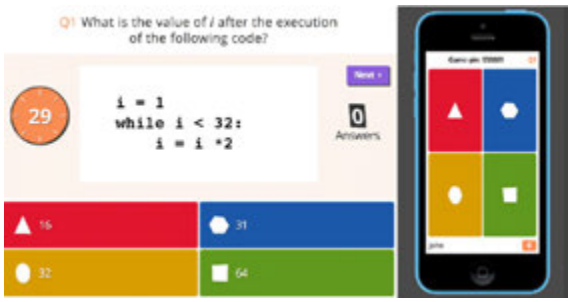


Figure 3. Example of quiz question designed in Kahoot!; source: [19]

III. EDUCATION INNOVATION OF INFORMATICS II USING GAMIFICATION

A. *Subject Informatics II at Faculty of Economics, Technical University of Košice*

The obligatory subject Informatics II is taught in the summer semester of the first year of bachelor study of the study branch Finance, Banking and Investments provided by Faculty of Economics, Technical University of Košice (FoE at TUKE). The main objective of the subject is to lead the students into the domains that they will meet in their practice on a daily basis, the domains of business informatics and business data analysis. The content of the subject is composed concerning students' previous knowledge and skill in the field of ICT [20]. The main covered topics are:

- *Fundamentals of business informatics* cover the history of business informatics, the introduction into different information systems used in enterprises like ERP (enterprise resource planning), CMS (content management systems), SCM (supply chain management), BI (business intelligence), etc.
- *Business process modeling (BPM)* is concerned with processes in the business itself and ways how and what kind of data management is needed to model and improve business processes.
- *Business data analysis* presents different types of data used in the business environment and for what purposes they are used.
- *Tools for business data processing* illustrate how ICT tools can be used for data processing and modeling. For that part of the subject, we examine and use a tool like ARIS, MS Excel, MS Access.
- *Economic aspects of business informatics,* where the students are introduced to how ICT could affect the financial part of the business.

- *The trend of business informatics* tries to show the upcoming technologies and their benefits for companies in the future.

The realization of the subject uses could be described in a way of blended learning. It is divided into two main parts (Figure 4.):

- F2F learning is realized by the lectures, where the students are provided by theoretical information further used during practical PC labs, where the students implement theoretical knowledge in solving real-world problems.
- Online learning part realized via LMS Moodle, serves mainly to publish study materials, news, tasks for students' self-study and preparation for practical exercises, for self-testing and for submitting semestral project files that are a part of the final evaluation of the particular student. It is also a space for online testing, chatting, feedback, etc. which ensures two-way communication between the participants in the education process.

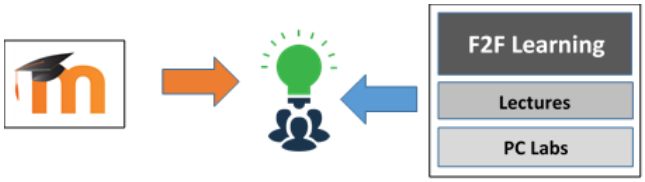


Figure 4. Model of Informatics II Education; Source: own contribution

Every time, after finishing the subject, there is realized the feedback on the subject. In an annual subject survey by the students, the subject is evaluated as a demanding but rewarding for their future life. In connection with the topic of the paper – possible gamification of the subject, we researched the habits of the students connected to the preferred way of study and the use of supportive LMS Moodle platform. From the conducted survey results, we are concerned with the following particular questions:

1. *What kind of study materials do you use during the study of the subject?*
2. *How often do you visit and use the course Informatics II in the LMS Moodle?*
3. *Is the information published in the LMS Moodle supportive to you during your study?*
4. *What kind of activity/study material did you miss during the study the Informatics II?*

As can be seen in the Figure 5. , the most widely used resources include online materials published in LMS Moodle (indicated by the 85% of respondents) and practical tasks and assignments, which are also published in LMS Moodle, but they are continuously solved during the F2F learning (indicated by the 92% of respondents).

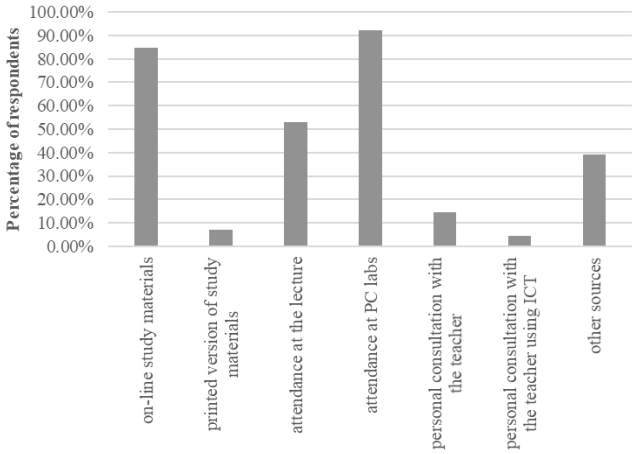


Figure 5. Preferred kind of study materials; Source: own contribution

In addition to the published materials, we were also interested in the frequency of visiting the course in LMS Moodle. The main reason for this interest was that in the past, we have noticed that students do not have an overview of what is available within the course in the system and they were not prepared for the practical exercises. One of the reasons is, that if the student does not visit the system regularly, he/she can easily lose the overview of the current course, which can affect his/her preparation for the final exam and his/her final evaluation. The survey shows that 58% of students attend the LMS Moodle course more than once a week, which is not a satisfactory number, especially if they have assignments and documents for preparation for F2F lessons just in LMS Moodle. At the same time, from that point of view, we realized also the alarming, that 11% of students visit the course in LMS Moodle even just once a week or less (Figure 6.).

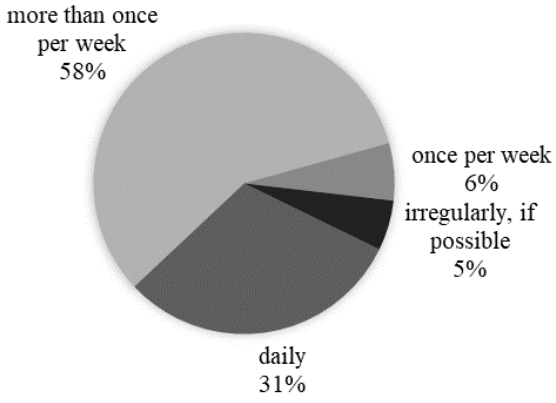


Figure 6. Frequency of students'visit of the course Informatics II in LMS Moodle; Source: own contribution

Within a more detailed analysis of individual students' logs in LMS, we found that students do not attend study materials during the semester, but only a list of tasks for the next practical PC lab. Their home preparation, therefore, consists only of identifying tasks that they are not able to work out independently. However, if they had previously studied the provided materials, this problem would not have arisen. However, students only view these

support materials (i.e., study materials, tutorials, etc.) before a credit or final exam.

100% of respondents agreed that study materials and tasks are supportive and useful during their study of the subject, but they miss the more interactive and motivating environment. 32% of respondents asked for more video-tutorials and 9% for more tasks for practicing, which will help them study better and in a more collaborative way.

B. Gamification possibilities of Informatics II

Based on the achieved results from the feedback on the Informatics II subject we try to find out how to motivate students to participate more actively in the educational process. As we found out, students use the published materials especially before the credits and final exams. What is a positive finding that students visit LMS Moodle more often, but they miss there an impulse, which would navigate them what to self-study. Therefore, we have decided to implement elements of gamification in order to improve the teaching of the subject Informatics II. We did not gamify the whole course, just the identified problematic parts by using some of the above-mentioned tools (Figure 7.).

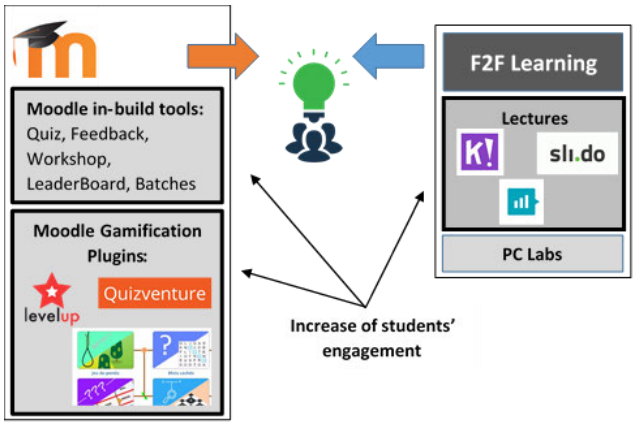


Figure 7. Gamified Model of Informatics II Gamification; Source: own contribution

At this stage, it is very important to realize the age of the target group and keep in mind the reasons for introducing a gamification approach to the course. Among the most critical parts of teaching the Informatics II, we can currently include the weak and passive attendance of students in the lectures. To solve this problem it is possible to use tools such as Mentimeter, Kahoot! and sli.do, which are accessible on-line and via smartphones, which the students have with them at the lecture. These tools are useful for "awakening" students' attention, especially in case of an explanation of the difficult topic when the students often lose their concentration while they just passively receive the information. These tools can be used to attract the students not only at the beginning of the lecture, but also to initiate the activity and participation of the students in the lecture's topic during the lecture, or to get instant feedback, resp. to discuss the issues that most interested the students within the explained topic at the close time of the lecture.

The second problem, what we expect to be solved by education gamification is to reach the continuous study of published study materials throughout the semester by the students. In that case, we can employ tools and plugins of the LMS Moodle that were designed for that purpose, all together with the appropriate setting of the learning paths in the Informatics II course, in order to raise the attractiveness of the learning environment.

At the higher education level, the students are not oriented only on the gaming aspects, but they take into consideration, how their participation affects their results in the subject evaluation. From that point of view, we are limited by the accreditation concerning the mandatory parts of the evaluation. Therefore, we decided to put the extra points to the gamified part that can be used to fill in the missing points in the final determination of the mark of the subject. Each student participating in the gamification part of the course could gain points for completing several simple tasks. The number of gained points depends on task completion (like attending classes) or actual student performance (like writing tests or completing the final project). If the student missed any test or failed to pass it, he was not given negative points as a penalty. He/she can choose to study in an obvious way. Of course, since this is a game, students who do not participate will not affect it in any way.

IV. CONCLUSION

Games shape the identity of the students as people and as learners. The gamification of the whole education process or just its particular parts presents a way to provide students with opportunities to act autonomously, to display competence, and to learn in relationship to others [21]. The research is done in [22] and [23] proved the positive impact on students' involvement after employing the gamification methods into the education process. Gamification, apart from the entertaining and more motivating form of education, help to build up and increase the skills and competencies that cannot be ignored in case of education for future jobs, especially: digital skills, digital navigation skills (what means finding information, prioritizing information and assessing the quality and reliability of information), followed by communication skills, literacy (e.g. media literacy, digital literacy, reading), critical thinking and judgment, problem-solving, teamwork, personal resilience and reflection [24].

Taking into account the previous paragraph, the gamification of education may look very attractive, but Fitz-Walter [25] and Decker [26] highlighted not only the potential benefits but also the challenges and potential conflicts when applying the gamification into Higher Education. There are no general rules or procedures on how to implement the gamification approach into education, it is highly dependent on the age of the attendance, learned subject and the purpose why the gamification of education needs to be implemented [12]. It is important teachers know the students, but take into consideration not only their interests but also the specific learning goals that hit on elements of knowledge from the curriculum.

The paper presented the main outcomes of the students' behavior towards course Informatics II in LMS Moodle and the expectations or proposals of the students on how we could enhance their engagement in the learning process. Based on the results of the conducted survey among our students and our research in the field of innovative teaching methods and tools, we decided to implement the learning gamification principle via different tools. During our research, we found out many various tools supporting the learning process gamification. After critical evaluation of them, we have to conclude, that the main aim is not to gamify the education for any price, but to find out the appropriate level and ways of gamification, which motivates the students but at the same time encourages them to study in entertaining way and not merely to concentrate on playing the games themselves.

There are a huge amount of tools supporting the process of gamification, but the topic of the subject and the level of education limit their use. Within this paper, we proposed an idea of how to gamify the courses taught at the higher education institution. The presented gamification proposal is implemented already in the course Informatics II and it will be tested during the next semester.

REFERENCES

[1] D.F. Roberts, U.G. Foehr, and V. Rideout, "Generation M: Media in the Lives of 8-18 Years-old," [online], available at: https://www.kff.org/wp-content/uploads/2013/01/generation-m-media-in-the-lives-of-8-18-year-olds-report.pdf [Accessed 21. January 2020]

[2] European Network of Education Councils, "Learning in the Digital Age," [online], available at: http://www.eunec.eu/sites/www.eunec.eu/files/event/attachments/report.pdf [Accessed 21. January 2020]

[3] International Education Advisory Board, "Learning in the 21st Century: Teaching Today's Students on Their," [online], available at: https://www.certiport.com/portal/common/documentlibrary/ieab_whitepaper040808.pdf [Accessed 20. January 2020]

[4] R. Chasse, "What is Digital Age Learning?" [online], available at: https://globalfocusmagazine.com/digital-age-learning-2/ [Accessed 18. January 2020]

[5] G. Kiryakova, N. Angelova, and L. Yordanova, "Gamification in education," in Proceedings of the 9th International Balkan Education and Science Conference, Edirne, Turkey, October 2014, retrieved from: https://www.sun.ac.za/english/learning-teaching/ctl/Documents/Gamification%20in%20education.pdf.

[6] M.S.F. Tham, "Smart Learning In The Digital Age," [online], available at: https://elearningindustry.com/smart-learning-digital-age [Accessed 20. January 2020]

[7] T. W. Malone, "What makes things fun to learn? Heuristics for designing instructional computer games," Proceedings of the 3rd ACM SIGSMALL symposium and the first SIGPC symposium on Small systems – SIGSMALL '80. ACM Press, New York, USA, pp. 162–169, 1980.

[8] T. W. Malone, " Heuristics for designing enjoyable user interfaces," Proceedings of the 1982 Conference on Human Factors in Computing Systems. ACM Press, 63-68, 1982.

[9] R. F. Bowman, "A Pac-Man theory of motivation. Tactical implications for classroom instruction," in Educational Technology,Vol. 22, No.9, pp. 14–17, 1982.

[10] J. R. Chapman and P. J. Rich, "Does educational gamification improve students' motivation? If so, which game elements work best?" in Journal of Education for Business, Vol. 93, No.7, pp. 315-322, doi: 10.1080/08832323.2018.1490687.

[11] K. Kapp, The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education. Pfeiffer, 2012.

[12] K. Werbach, and D. Hunter, For the Win: How Game Thinking Can Revolutionize Your Business. Philadelphia, PA: Wharton Digital Press, 2012.

[13] M. Borys, and M. Laskowski, "Implementing game elements into didactic process: A case study," Management, Knowledge and Learning International Conference proceedings: Zadar, Croatia, 2012. pp. 819–824.

[14] B. Kim, "Designing Gamification in the Right Way," Library Technology Reports, Vol. 51, No. 2, 2015.

[15] R. Gotarkar, "All About Gamification in eLearning and Why it's Still the Rage?" [online], available at: https://edwiser.org/blog/gamification-in-moodle-elearning/ [Accessed 12. January 2020]

[16] A. Apostolopoulos, "The 2019 Gamification at Work Survey," [online], available at: https://www.talentlms.com/blog/gamification-survey-results/ [Accessed 13. January 2020]

[17] Mentimeter, "What is mentimeter?" [online], available at: https://www.mentimeter.com/ [Accessed 20. January 2020]

[18] Slido, "Explore Slido" [online], available at: https://www.sli.do [Accessed 20. January 2020]

[19] Kahoot! "Kahoot! for schools - How it works" [online], available at: https://kahoot.com/schools/how-it-works/ [Accessed 20. January 2020]

[20] L. Révészová, "Designing modern informatics education for future managers and advanced users according to their knowledge base" in: Ekonomie a Management. Vol. 19, no. 4, pp. 186-201, 2016.

[21] R. M. Ryan, and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being," in American Psychologist, Vol. 55, pp. 68–78. 2000, doi:10.1037/0003-066x.55.1.68.

[22] O. Borras-Gene, M. Martinez-Nuñez, and L. Martín-Fernández, "Enhancing Fun through Gamification to ImproveEngagement in MOOC," in Informatics, Vol. 6, No. 28, 2019 doi: 10.3390/informatics6030028

[23] S. Sheth, J. Bell, and G. Kaiser, "Increasing Student Engagement in Software Engineering with Gamification," Columbia University Computer Science Technical Reports. Department of Computer Science, Columbia University, 2012, Retrieved from http://academiccommons.columbia.edu/catalog/ac:154509

[24] S. Grand-Clement , "Digital learning. Education and skills in the digital age," [online], available at: https://www.rand.org/content/dam/rand/pubs/conf_proceedings/CF300/CF369/RAND_CF369.pdf [Accessed 15. January 2020]

[25] Z. Fitz-Walter, D. Tjondronegoro, and P. Wyeth, "Orientation Passport," proceedings of the 23rd Australian Computer-Human Interaction Conference, ACM Press, New York, USA, pp. 122-125, 2011.

[26] A. Decker and E. Lane Lawley, "Life's a Game and the Game of Life: How Making a Game Out of it Can Change Student Behavior," proceeding of the 44th ACM technical symposium on Computer science education. ACM Press, New York, USA, pp. 233-238, 2013

Continuous Summative Assessment Sessions as a Motivational Tool for STEM students: a Case Study

G. Durovic
University of Rijeka, Rijeka, Croatia
gdurovic@uniri.hr

Abstract - **While using a classical approach to summative assessment that was based on mid-term exams, it was observed that overall results for each consecutive generation of STEM students at the end of the semester failed to improve. To promote a deep learning approach among STEM students, that can be achieved through a continuous approach to learning, activities for continuous summative assessment were introduced. Students have been given the opportunity to take the test for each of the parts of the course twice in two week period with the best of the two results used for grading. A comparison of the obtained results and results of the previous generations of students was made and it showed that students that participated in the experiment achieved significantly better overall results and understanding of the course materials at the end of the semester. In order to expand the experiment into the online environment, a model for an Education Recommender System with assessment capabilities for STEM education was proposed. Based on the proposed model of the system, the prototype of the system was built, and online summative assessment sessions based on the offline experiment were designed for the next cycle of the research.**

Keywords - summative assessment; STEM; Education Recommender System, ELARS

I. INTRODUCTION

Student motivation for learning is one of the basic elements needed for achieving successful learning results [1]. In order to help students to maintain or increase the level of their motivation, a number of different motivational strategies for STEM (science, technology, engineering, and mathematics) education have been explored in the past decade [2, 3].

Within the education process, assessment sessions can have a great impact on students' motivation. Successful formative assessment results can further motivate students for future learning activities [4], while summative assessment results can have a great impact on career and academic opportunities of students [5]. With the advancement of ICT, the ways in which student knowledge is evaluated have been expanded from the traditional classroom environment toward the online environment. Online evaluation systems that have been developed and tested in real educational environments showed that these systems can also be used for motivational purposes [6, 7].

This paper presents three-year research about the possibility of motivating STEM students to start to learn continuously during the semester. A model of summative assessment that can be used to increase students' motivation for continuous learning is proposed. The model was used with STEM students at the University of Rijeka, Croatia. Two consecutive generations of students were evaluated using classical mid-term exams, while the third generation of students was evaluated using the proposed weekly organized summative assessment sessions. The obtained results showed that students included in the proposed weekly organized summative assessment sessions positively accepted such an assessment model and achieved significantly better overall results. The paper also presents a prototype of an online evaluation system that was build within Educational Recommender System ELARS to enhance the implementation of the proposed assessment model for larger groups of students.

II. PREVIOUS RESEARCH

One of the indicators that can be used for measuring the success of the students' coursework during the semester is the course points score that students have achieved at the end of the semester (before the final exam). Table I. shows the percentage of students with insufficient course points score in two STEM courses at the University of Rijeka (Undergraduate University Study program of Electrical Engineering and Undergraduate University Study of Polytechnics, in both cases for obligatory course *Electrical*

TABLE I. PERCENTAGE OF STUDENTS WITH INSUFFICIENT COURSE POINTS SCORE AT THE END OF THE SEMESTER

Academic Year	**Number of enrolled students**	**Number of students with insufficient** course points **score**	**%**
The University of Rijeka, Faculty of Engineering, Undergraduate University Study of Electrical Engineering			
2015./16.	175	66	37,71 %
2016./17.	171	54	31,58 %
2017./18.	162	64	39,51 %
The University of Rijeka, Undergraduate University Study of Polytechnics			
2015./16.	21	6	28,57 %
2016./17.	30	10	33,33 %

engineering 2.). From the presented data it can be observed that a great number of students in each generation haven't achieved at least minimal course points score needed for taking the final exam (roughly a third of students or more in each generation).

On the other hand, while conducting research on the use of digital tools for learning by STEM students [8], it was observed that students want to be additionally motivated by their teachers. Although the introduction of digital tools for learning within STEM courses was positively accepted by students, it did not significantly affect their motivation for learning the course content [9]. Without sufficient motivation, it was not possible to get students to start learning continuously and to adopt a deep learning approach as their learning practice. The majority of students continued to approach their learning with the aim to just complete the tasks and memorize the minimal amount of presented course content in a so-called surface approach to learning [10].

III. METHODOLOGY

The aim of the research presented in this paper is to explore the possibility of motivating students to start to learn continuously during the semester. Since summative assessment can be used as a motivational strategy [6], the classical approach through mid-term exams can be replaced with weekly organized summative assessment sessions. To organize these sessions, the teacher must predetermine the time required for solving a task (that can be different for each topic), number of attempts for each topic within course, and the time period during which each topic should be in the focus of students learning activities. Therefore, a classical approach to summative assessment through several mid-term exams was replaced by weekly organized summative assessment sessions. The following research questions were explored:

RQ1: Will the introduction of weekly organized summative assessment sessions motivate students to start to learn continuously during the semester?

RQ2: Which form of the summative assessment (through classical mid-term exams or through weekly organized summative assessment sessions) will be better suited for students?

Participants of the research were three consecutive generations of students at the University of Rijeka enrolled in the Undergraduate University Study of Polytechnics that attended the obligatory course *Electrical engineering 2*. The course is taught through weekly lectures that encompass both theoretical part and math-based tasks developed for a deeper understanding of the presented theoretical concepts. In total 78 students were included in the research that was carried out offline in a traditional paper-based approach to conducting exams.

The research was conducted during the academic years 2015./16., 2016./17. and 2017./18. Knowledge assessment of students enrolled in obligatory course *Electrical engineering 2* in academic years 2015./16. and 2016./17. was carried out through the classical summative assessment approach using mid-term exams with math-based tasks. In the academic year 2017./18. knowledge assessment was conducted through introduced weekly organized summative assessment sessions. Starting from the second week of the semester, weekly organized summative assessment sessions started. These sessions were designed for grading smaller parts of the course content. They consisted of the math-based task related to the course content that was presented at the lecture a week earlier. Students had the opportunity to take the weekly exam twice for each topic of the course content during two consecutive weeks. The best result from both attempts was used for grading.

To answer the research questions, results obtained from the third generation were compared with the results obtained from the first and second generations of students. Also, at the end of the semester, the students from the third generation were surveyed using a paper-based questionnaire. Gained results were used for building the online prototype of the proposed system.

IV. RESEARCH RESULTS

When starting to attend universities, students tend to adopt a surface approach to learning resulting in a superficial accumulation of knowledge [11]. This approach can result in inadequate mastering of the course material since previously covered course content must be mastered at the expected level and used while learning the contents that follow. Previous research showed that one of the main contributing factors is students' habit of non-continuous learning [8].

To motivate students to approach their learning in a continuous way is one of the steps that can lead to achieving a deep understanding of the course material [9]. Since the summative approach can have a significant impact on students' career, exam format and number of exams must be carefully planned [5].

While comparing students' results achieved through three mid-term exams for the consecutive generation of students, it was observed that a significant number of students do not achieve enough course points during the semester. For each course, a minimal number of course points are set for the successful completion of coursework during the semester. If the student does not achieve at least a minimal score, that student is excluded from taking the final exam, thus having the obligation to enroll in that same course in the next academic year.

In Table II. the results for achieving at least a minimal course points score at the end of the semester for the three generations of students that participated in the research are shown.

TABLE II. COURSE POINTS SCORE OF STUDENTS AT THE END OF THE SEMESTER

Academic Year	**Number of enrolled students**	**Number of students with insufficient course points score**	**%**
The University of Rijeka, Undergraduate University Study of Polytechnics, obligatory course Electrical Engineering 2			
2015./16.	21	6	28,57 %
2016./17.	30	10	33,33 %
2017./18.	27	1	3,70 %

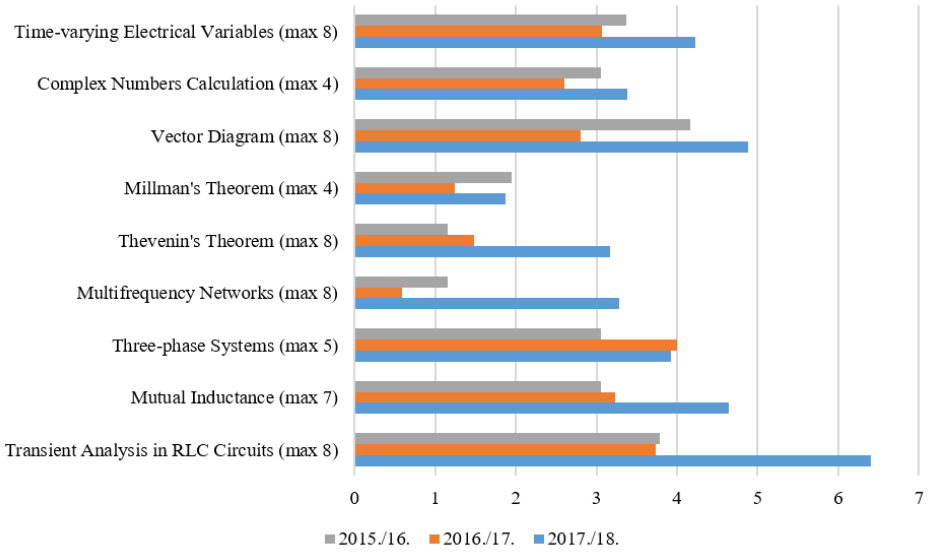


Figure 1. Comparison of overall mean score values of achieved points for nine topics of the course content.

TABLE III. SURVEY RESULTS

I usually prepare for the mid-term exam as follows:	
I start studying just before the exam	60,87 %
I start studying more than a week before the exam	0,00 %
I study continuously throughout the semester	4,35 %
I start studying some of the course content just before the exam while I study other content continuously	34,78 %
The above-mentioned method of preparation for the mid-term exam is most influenced by:	
A perennial habit of learning that way	30,43 %
My colleagues' learning habits	4,35 %
The number of duties I have during the semester	52,17 %
Something else	13,04 %
Did the weekly organized summative assessment sessions encourage you to study continuously during the semester?	
It encouraged me	52,17 %
It did not encourage me	4,35 %
It did not affect my usual way of learning	43,48 %
Does it suit you that you can take the test more than once and that only your best results are used for grading your work?	
It suits me	100,00 %
It does not suit me	0,00 %
I don't have a preference	0,00 %
Which form of knowledge assessment, which is carried out in writing during the semester, is more appropriate for you?	
2-3 mid-term exams covering larger sections of the course material	13,04 %
Continuous weekly organized summative assessment sessions of the smaller content of the course material	82,61 %
I don't have a preference	4,35 %
Would you rather prefer that the knowledge assessment is conducted online through weekly summative assessment sessions or through classical mid-term exams conducted offline?	
Online weekly summative assessment sessions	52,17 %
Classical offline mid-term exams	17,39 %
I don't have a preference	30,43 %
If you can choose between online weekly summative assessment sessions and online mid-term exams what form of knowledge assessment would you choose?	
Online weekly summative assessment sessions	78,26 %
Online mid-term exams	4,35 %
I don't have a preference	17,39 %

In order to compare students' achievement within the course content, weekly summative assessment sessions were organized using the same math-based task used in classical mid-term exams carried out with the previous two generations of students. This approach enabled a comparison of results for each of the nine topics of the course content that is evaluated through summative assessments and is used for grading.

During academic years 2015./16. and 2016./17., when classical mid-term exams were used, three topics were part of each of the mid-term exams. Students had two hours to finish the exam and solve three math-based tasks, each corresponding to one of the topics. Each mid-term exam could have been taken only once, and the achieved results were used for grading. The same maximum numbers of points were assigned to the topics in the academic year 2017./18. when weekly organized summative assessment sessions were introduced. Students had the opportunity to take the tasks two times in two consecutive weeks that followed the week in which lecture for that topic has been held. The result from the better of two attempts was used for grading students. For each math-based task, students had half an hour at their disposal. Results for each of the course content topics are shown in Figure I.

At the end of the semester students that were part of the experimental group were asked to express their views regarding their experience with weekly organized summative assessment sessions. With the aim of getting feedback information, a paper-based questionnaire consisting of a set of single-answer multiple-choice questions was devised and conducted. Out of 27 students that were part of the experiment, 23 of them completed the questionnaire.

Detailed results obtained for each prepared question are presented in Table III.

A. Discussion of the results

As can be seen in Table II., respectively 28,57 % and 33,33 % of students in the first two generations have failed to achieve a minimal number of course points required

during the semester. A summative assessment for all students in those academic years was conducted through three classical mid-term exams. When weekly organized summative assessment sessions were introduced, only 3,70 % of enrolled students failed to achieve at least a minimal course points score at the end of the semester. From these results, it can be concluded that a significant improvement in overall results achieved by students through coursework during the semester can be seen.

When mean score values for each of the nine topics that were part of the summative assessment procedures are compared (Figure 1.) it can be seen that for seven topics the results obtained from the experimental group were higher than in two previous academic years. For two topics one of the previous generations of students achieved a slightly better overall mean score (0,07 points difference for Millman's Theorem and 0,08 points difference for Three-phase Systems). However, in topics in which the experimental group achieved better results over previous generations of students, the difference in mean score value ranged from 0,33 points difference (for topic Complex Number Calculation) up to 2,62 points difference (for topic Transient Analysis in RLC Circuits).

Survey results (Table III.) show that the majority of surveyed students start learning just before the exams, mainly because of their other obligations. Weekly organized summative assessment sessions prompted more than half of the surveyed students to start to learn continuously. All of the surveyed students stated that the possibility to take the tasks for each topic more than once suited them. Also, a great majority of them stated that they do like weekly organized summative assessment exams better than traditional mid-term exams. When asked to choose between online weekly organized summative assessment sessions versus offline classical mid-term exams, more than half of surveyed students chose online weekly organized summative assessment sessions. On the other hand, when surveyed students have been asked to choose between online weekly organized summative assessment sessions versus online mid-term exams, a majority of them chose weekly organized summative assessment sessions.

Regarding RQ1, from results obtained through weekly organized summative assessment sessions, it can be concluded that the offline part of the proposed experiment has successfully motivated a number of students in a group to start learning continuously.

Through the conducted paper-based questionnaire used for getting an answer to RQ2, students' opinions showed that the majority of them do prefer the introduced weekly organized summative assessment sessions over classical mid-term exams. Obtained results are also the foundation for building a prototype of the online system based on the previously designed model [9].

V. AN ONLINE EVALUATION SYSTEM PROTOTYPE

The model for an Online Evaluation system for STEM education was designed using previously obtained research results [9]. Results of the offline case-study experiment that are presented in this paper were used in order to build a prototype of the proposed system.



Figure 2. Example of the math-based task in the system prototype.

The model of the proposed system and its' prototype were designed around Educational Recommender System ELARS that was developed at the Department of Informatics at the University of Rijeka. A standard Educational Recommender System (ERS) structure that ELARS was build upon was modified in order to add online math-based tasks testing [12].

Within the prototype of the online system, for each topic that is to be graded, two sets of math-based tasks were envisioned. The first set was planned to be used for formative assessment in preparation for taking the exam. The second set was designed for conducting weekly organized summative assessment sessions. In this way, the prototype of the system will be able to allow the students to use the system in order to practice for the exam by solving math-based tasks of approximately equal complexity as the tasks that will be used for grading.

In both sets of the math-based tasks, a task is randomly selected for each student in each session. The overall number of tasks that students can access is defined by the teacher, according to the overall number of tasks within the set. The same task will not be selected more than once for each student. In Figure 2., a screenshot of the math-based task in the system prototype is presented.

The prototype of the system enables the teacher to define the time period during which students will be able to access each topic, separately for each set of math-based tasks. Also, the teacher can define a time period for solving a task based on the tasks' complexity.

Since the prototype of the online system is built around the ERS structure, the system has the capability to present students not only with the information about the accuracy of the provided answers but also with feedback information about concepts that are graded through tasks. This feedback information that is available for students during the formative knowledge assessment within the practice set of math-based tasks can be used by students in order to identify the parts of the topic content that they have to master at the higher level.

The prototype of the online system is built with the basic statistical capabilities intended for teachers. These capabilities could be used in order to follow students' progress (individual students and the group as a whole).

Since one of the main problems with online testing is the possibility of fraudulent activities, the basic statistical capabilities of the system can also be used in order to detect these types of behavior. For example, since the teacher can see time periods used for solving the math-based tasks for each student and compare them, it can be seen if the student has entered the results in too short or too long a time interval relative to other students. In this case, the teacher can forewarn the student thus preventing the abuse of the system in future sessions, or exclude the suspicious results for selected math-based tasks from consideration when grading the students' knowledge.

VI. Conclusion and Future Research

Using classical mid-term exams for summative assessment for STEM students showed that overall results at the end of the semester were similar and unchanged for consecutive generations of students. Out of all enrolled students, between 30% and 40% of them didn't, through coursework during the semester, achieve minimal course points score needed for taking the final exam. In order to address the observed problem, weekly organized summative assessment sessions were introduced.

As the preparation for building an online prototype of the system for conducting summative assessments in the proposed fashion, the offline experiment was conceived and conducted. The results gained from the experiment showed that students were motivated to learn continuously throughout the semester and that almost all of them achieved at least minimal course points score needed for taking the final exam. Also, feedback information gathered through the paper-based survey at the end of the semester showed that the proposed way of conducting a summative assessment in a weekly fashion was positively accepted. Students also indicated that most of them prefer weekly organized summative assessment sessions over mid-term exam approach, especially if these weekly organized summative assessment sessions are to be conducted online. If conducted offline, weekly organized summative assessment sessions can be extremely demanding for the teacher, especially when a large number of students are enrolled in the course. In order to successfully implement the proposed assessment approach in a real educational environment, weekly organized assessment sessions must be conducted online using adequate online technologies.

Based on the initial results, a prototype of the online evaluation system was built. A prototype was designed around the Educational Recommender System structure based on the previously developed model of an Online evaluation system for STEM students. By combining the ERS structure with math-based testing capabilities, the prototype of the online system could be used for conducting weekly organized summative assessment session and to provide students with quick feedback information regarding their success. The system can be used both for formative and summative assessment, and the gathered data can be used for providing teachers with basic statistical analysis. A combination of recommendations and statistical analysis based on the students' achievements during formative and summative assessment sessions conducted through the system should provide both students and teachers with feedback information that can be used for fine-tuning their learning and teaching activities. In the next phase of the research, the prototype of the system will be used in the real educational environment. The prototype will be refined according to the feedback information gained from students and teachers that will participate in the research.

Acknowledgments

The research has been co-funded by the University of Rijeka under the project “Digital games in the context of learning, teaching, and promoting inclusive education” (uniri-drustv-18-130).

References

[1] M. Prosser, and K. Trigwell, Understanding Learning and Teaching; the Experience in Higher Education, SRHE and Open University Press, Philadelphia, USA, 1999.

[2] P. Lynch, S. Sangelkar, G. Demeo, and B. Morkos: “The I-C-D-M methodology improving undergraduate engineering student motivation, satisfaction, and performance“, in Proceedings of the 2017 IEEE Frontiers in Education Conference (FIE), 18-21 Oct. 2017., Indianapolis, IN, USA, pp. 365-371.

[3] M. Kapoor, S. Hua, and D. C. Anastasiu, “Improving Student Motivation through Competitive Active Learning“, in Proceedings of the 2018 IEEE Frontiers in Education Conference (FIE), 3-6 Oct. 2018., San Jose, CA, USA, pp. 275-279.

[4] T. O'Mahony, “Assessing engineering learning: formative, summative and dialogical“, in Proceedings of the 2013 IEEE Global Engineering Education Conference (EDUCON), 13-15 March 2013, Berlin, Germany, pp. 990-995.

[5] D. Hernandez-Leo, and V. Moreno Oliver, “Fine-tuning formative and summative assessment in Bachelors' Final Projects“, in Proceedings of the International Symposium on Computers in Education (SIIE), 12-14 Nov. 2014., Logrono, Spain, pp. 41-46.

[6] M. Naghizadeh, and H. Moradi, “A model for motivation assessment in intelligent tutoring systems", in Proceedings of the 7th Conference on Information and Knowledge Technology (IKT), 26-28 May 2015, Urmia, Iran, pp. 323-328.

[7] N. A. Azmi, K. Mohd-Yusof, and F. A. Phang, “Impact of Effective Assessment towards Students' Motivation in Computer Programming Course“, in Proceedings of the 7th World Engineering Education Forum (WEEF), 13-16 Nov. 2017, Kuala Lumpur, Malaysia, pp. 415-419.

[8] G. Durovic, M. Holenko Dlab, and N. Hoic-Bozic, “Research on the Use of Digital Tools by STEM Students at the University of Rijeka”, TEM J., vol. 8, no. 2, pp. 636–641, 2019.

[9] G. Durovic, M. Holenko Dlab, and N. Hoic-Bozic, “A Model of an Online Evaluation System for STEM Education” in 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2019, pp. 677–680

[10] M. Svedin, O. Balter, M. Scheja, and K. Pettersson, “A Surface Approach to Learning Rewards First-Year Engineering Students", in Proceedings of the 2013 Learning and Teaching in Computing and Engineering, 21-24 March 2013, Macau, China, pp. 143-147.

[11] A. Desierto, C. D. Maio, J. O'Rourke, and S. Sharp, "Deep or Surface? The learning approaches of enabling students in an Australian public university" in Students Transitions Achievement Retention & Success (STARS), 2018., pp 140 (1-10)

[12] G. Durovic, M. Holenko Dlab, and N. Hoic-Bozic, N. "Motivating STEM Students to use Web 2.0 Tools for Learning: a Case Study" in Proceedings of the IEEE International Conference on Information Management and Processing, 2018., pp 140.-144.

Increasing the Learning Efficiency in Decision-Making Field using the Workshop Activity in Moodle

Nikola Kadoić* and Barbara Šlibar *
* University of Zagreb, Faculty of Organization and Informatics Varaždin, Pavlinska 2, 42000 Varaždin, Croatia
nkadoic@foi.hr, bslibar@foi.hr

***Abstract* - The decision-making field is one of the most critical fields in education for entrepreneurs. Students must adopt different concepts, methods, and tools that improve both - their analytic and intuitive decision making. One of the most used decision-making methods is the analytic hierarchy process (AHP). Students must know how to recognize that a specific problem requires AHP analysis and to apply it correctly. This study presents the implementation of AHP education using the Workshop activity supported by the learning management system Moodle. In the first part of this activity, the students have to make a paper of AHP analysis of one real strategic or tactical problem related to any organization they choose: they have to structure the problem using the structuring methods, collect pairwise comparisons from decision-makers, calculate the group priorities and make the sensitivity analysis. In the second part of the activity, the students do peer assessment – they evaluate other students' projects using the rubrics and qualitative assessment. The analysis reveals better results in understanding the AHP when compared to traditional ways of learning (direct assessment of papers by teachers without peer assessment). The different ways of learning are evaluated by comparing the points achieved for the AHP papers, points achieved for the AHP test, and using a short survey – all of them are implemented in Moodle as well.**

Keywords* - *learning, efficiency, Moodle, Workshop, decision making, AHP

I. INTRODUCTION

Course *Business decision analysis* covers several topics in the area of multi-criteria decision making and decision making under uncertainty and risks at the Graduate Study Economic of Entrepreneurship at the Faculty of Organization and Informatics Varaždin, University of Zagreb. One of them is the Analytic Hierarchy Process (AHP). Teachers on that course pay massive attention to this topic because AHP is the most often used method in decision making. There are several steps in this method [1], [2]:

1. Structuring decision-making problem – in this step, a problem from real context has to be structured in the form of the hierarchy, which is a form that AHP can analyze. At the top of the hierarchy, there is a decision-making goal (ex. to identify the best candidate applied on the call; to identify the best software for digital transformation, to select the most promising market where the products will be offered, etc.). Below the decision-making goal, there are the main criteria, which can be then decomposed to sub-criteria. At the lowest level, there are alternatives. Figure 1 presents the example of the hierarchy: two cities want to select the best way of future cooperation. There are three forms of cooperation available (right side of Figure 1), and five main criteria (left side of Figure 1) [3].
2. Pairwise comparisons procedures. In this step, elements from the hierarchy are pairwise compared with respect to superior elements in the hierarchy.
3. Calculation of the priorities. Based on pairwise comparisons (step 2), criteria weights and local priorities of the alternatives are calculated. Finally, using criteria weights and local priorities – global priorities are calculated.
4. Sensitivity analysis.

Firstly, the method with all steps has been presented to them in much detail and demonstrated with different examples. That is done in lectures and seminar hours. Then, students have to do a small project:

- First, they have to visit a real enterprise, association, NGO, public body, or any other type of organization they want, and through discussion define a real strategic or tactic problem related to the selected organization.
- After the problem has been identified, students have to apply the AHP method and make a decision proposal. The assignment is defined in detail describing the elements they have to implement.
- Finally, students have to make a report with a problem description and all steps in AHP implementation.

Until the academic year 2019/2020, the activity has been implemented in a way that students became a detailed description of the work they have to do, and when they finished their work, they uploaded the work on the Moodle and teachers evaluated their work and gave the grades. Also, some papers from previous years were available to them as demo-examples of the final projects.

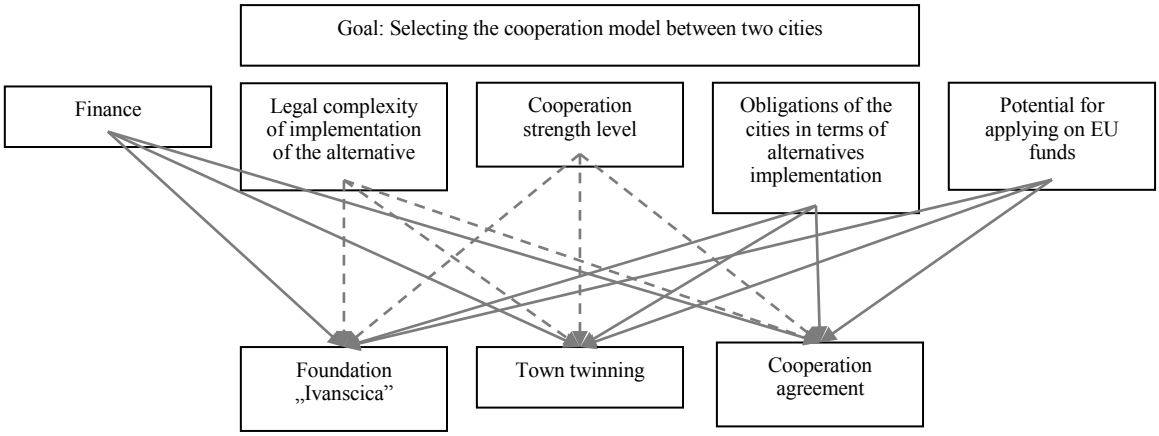


Figure 1. AHP model (goal, criteria, alternatives; sub-criteria not shown)

In addition to that, in case of any questions related to the problem selection as well as to the implementation of the AHP, they were able to consult teachers.

In the 2019/2020 academic year, the Workshop activity in Moodle was introduced to implement the AHP project activity. Workshop activity is a type of peer-assessment evaluation [4]. In this activity, the following is possible:

- Teachers can define a rubric instrument for evaluation of students' projects,
- Teachers can upload demo-projects, i.e., students' projects from the previous year,
- Students can upload their projects, but before that, they can consult rubric or demo-projects and use them in doing their projects,
- Students can evaluate (assess) other students' projects using the rubric (or another evaluation instrument),
- Self-assessment is also possible,
- Students can assess demo-example to practice assessment process,
- Teachers can assess students' projects,
- Teachers can assess students' evaluations of other students' projects.

In this activity, students receive two types of grades: grades for their submissions and grades for assessment (quality of their feedbacks on other students' projects).

In this paper, we will demonstrate how has the workshop activity been applied and what are the results in terms of understanding the AHP when compared to the previous way of implementing the AHP activity.

II. THEORETICAL BACKGROUND

Assessment is an inherent part of the learning process in higher education. As such assessment can be interpreted as a way to evaluate, assess, measure the nature, quality, ability of student learning in order to further enhance learning as well as a teaching process. Assessment of learning could be denoted as a traditional approach were teachers assess students, while assessment as learning as a contemporary one where students assess their own work or work of their peers. Seeing assessment and learning as one process is very promising for learning development [5], [6]. By that seeing, assessment becomes indispensable from student learning and students become an intrinsic element of the assessment process. There is increasing interest regarding self and peer assessment in the literature. The connection between self-assessment and peer assessment is indisputable. Both assessments include feedback(s) and/or grade(s) as the final result, but they differ in terms of whose work has been evaluated. Students assess their own work or work of other students or rather their peers with the help of predefined quality criteria. Since self-assessment and peer assessments have been widely explored many advantages and disadvantages have been highlighted in the literature [5]–[7]. Proved beneficial parameters are raise of effective overall learning, raise of effective overall teaching, improved learning atmosphere, greater satisfaction with given assessment and feedback, etc. [5], [7]. The most criticized issue is the consumption of time regarding the perceived [5], [8].

Rubrics are tools that help us to evaluate work for which there is no particular and strict answer but rather a more complex one which can vary based on various parameters. Various definitions of rubrics are proposed in the literature [9]–[12]. The following definition covers all the important aspects. Rubrics are schemes utilized for evaluation of student work and some detail about performance levels and those schemes are based on the list of criteria.

In most reviewed papers the rubrics are categorized differently. Mostly they are described as holistic, analytic or scoring. Based on levels of detail rubrics can be grouped as follows [10], [13]:

- Holistic rubric – provides detail for performance level, but it does not contain a list of criteria.
- Single-point rubric – provides detail for multiple criteria. It does not provide detail for performance level, but instead of that, it gives description when

performance would be accepted based on defined multiple criteria.

- Analytic rubric – provides detail for both performance level and multiple criteria.

Although, there is no unique way that is simultaneously the best way to create and apply certain rubric, there are guidelines on how to use this measurement tool for self and peer assessment [5], [10], [11]. Some of the tips may be briefly described as follows [10], [11]:

- The rubric should match your style, your students, your task – Students should be able to understand the grading system which you set by selecting the certain type of rubric. Students should be able to understand the chosen rubric because otherwise, they would not be able to operate with a rubric and make an evaluation. Choose a rubric that will work best for you because each type of rubric has pros and cons with regard to the task.
- Rubrics should be fashioned in such a way to catch your learning objectives – If you decided to use the rubric as an assessment tool in your class first make an exhaustive description of the assignment and then write down all components of outstanding work. Also, you should consider what makes a student works unacceptable with regard to the assignment. Try to think about common student mistakes to further polish the rubric.
- Rubrics could be shaped in collaboration with students - This could be accomplished by presenting objectives of your assignment to students, explaining the theoretical background of such an assessment tool to students, asking students to provide feedback, and then shaping a draft of the rubric based on gained feedbacks.
- Encourage a brief tract about the utilization of rubrics in the learning process – By pointing out all the possibility of such assessment tolls students can improve their knowledge as well as skills, and apply all gained more effectively in the future.

The rapid development of information technology has led to the introduction of new and original learning models as well as teaching ones. The utility of introducing information technology (IT) for the learning process is visible through learning management systems (LMS) which are widely applied in higher education [14]–[17]. It is clear that IT drastically changed ways of learning and teaching. Therefore, the importance of student acceptance of LMS, benefits and drawbacks of using such systems is widely explored in the literature [12], [14]–[19].

Given the broad application and popularity of Moodle in practice, it is a commonly researched learning management system [16], [19]. This is the reason, why student perception is widely explored. Some previous researches investigate the factors which influence effectiveness and use of Moodle in higher education from student perspective [19], the student satisfaction with the online learning platform Moodle and the learning environment itself [15], the student acceptance of using Moodle for learning in Higher education [16].

There are previous studies which had been developed rubrics and investigate the influence of developed assessment tool on various aspects within the learning management system Moodle [12], [14], [17], [18], [20].

III. IMPLEMENTATION OF THE WORKSHOP ACTIVITY IN CASE OF TEACHING AHP

The implementation of AHP activity using the Workshop in Moodle obtained the following steps:

1. The deadline for the submission of projects was set up. Also, demo-projects were uploaded. The analytic rubric for (self-)assessment was defined.
2. During their projects' creation, students were able to consult rubric to guide them through AHP implementation and documentation. Also, they were able to consult demo-examples and make test-assessment on them.
3. After the submission deadline was met, the assessment phase started. By using the random function in Workshop activity, to each student three projects were assigned. Also, their project was assigned to them for self-assessment.
4. After students finished their work in terms of assessment, the teacher assessed all projects. Teachers' grades on students' projects have the highest impact on students' grades for submission.
5. Teachers evaluated all students' assessments grading them using a direct assessment method. Also, for each assessment, teachers have to define the weight. These weights have an influence on students' grades for submission. If some students made proper assessments of other students' works, then the weight of his/her assessment will be high and vice-versa.
6. In the last step, Workshop activity calculated the finale grades for submissions and assessments for each student respecting the students' and teachers' assessment and student assessments' weights.

After the grades were given to all students:

- Students were tested by using the Test option in Moodle with questions related to understanding the AHP,
- Students responded to the survey also implemented in Moodle using the Feedback option – the survey was related to the implementation of the AHP activity. The questions for the survey were based on previous researches and are upgraded to meet the needs of our research [5], [9]. The feedback questions can be seen in the results section.

IV. RESULTS

The new way of implementing the AHP activity was evaluated in three ways: (1) quantitative analysis, (2) qualitative analysis, and (3) through the survey.

A. Quantitative analysis

Quantitative analysis is implemented in two ways: (1) comparing the grades achieved for submission before and after changing the way of implementing the AHP activity; and (2) comparing the grades achieved on the AHP test, which followed the AHP activity. The results are presented in Table 1.

TABLE 1. QUANTITATIVE RESULTS

	% of max possible points, on average	
Activity	**Before Workshop (2018/19; N=76)**	**After Workshop (2019/20; N=69)**
Submission grades	68.4	73.9
AHP test points	39.5	44.4

According to the results in Table 1., both submission grades and AHP test points, are approximately 5% higher in the scenario in which AHP activity was implemented by using the Workshop in Moodle. Even though it is not a huge difference, it is a good starting point for making future upgrades of how the AHP activity should be implemented. Teachers identified some possible upgrades based on their experience. Besides, the students gave their proposals and observation as part of the survey which is presented in the C subsection of the results section.

B. Qualitative analysis

Qualitative analysis is performed by a teacher who participated in the implementation of AHP activity in both ways, not using and using the Workshop activity in Moodle. The analysis is divided into two parts: one is related to the role and effects of the rubric in the assignment, and the other is related to peer-assessment implementation.

The effects of the rubric in the Workshop activity are multiple:

1. Since the rubric is published at the beginning of the assignment, students were able to consult it during the whole process of creation of their submission – and they did it. The rubric helped students to implement each phase of AHP activity, guiding them through the work they must do.
2. The rubric helped students to organize the documentation of the AHP implementation. Indeed, the students' documents were organized in sections that followed the criteria in the rubric.
3. The analyzed decision-making problems were now structured much better describing all the methods that were applied in the structuring phase: students visited the organizations whose problems they analyzed, talked with responsible persons, and worked harder on problem structures - when compared to the old way of AHP activity implementation.
4. By using a rubric, it was much easier to explain the weak points of the submitted papers. Also, those explanations were more understandable to the students. After the final grades for submission were published, there were almost no requests for additional explanations for the grades that are achieved for the submission.

In terms of peer-assessment, the effects were also multiple:

1. Students had to assess their submissions. This was sometimes a cold shower when they realized that their work is not so brilliant. Self-assessment contributed to the creation of more realistic expectations in terms of the grades they will achieve.
2. When doing assessments, students were able to understand the teachers' role in the teaching process better.
3. Even though students evaluated the submissions of their colleagues and friends, they had to be objective, because the quality of the assessment had a direct influence on their grades for assessment. In practice, there were examples of too high grades for submission given to other students. This was another lesson for them – they have to learn to separate private from a business – those situations will often appear in their real jobs.

The teacher's main conclusion was that the AHP activity was implemented at a higher-level comparing to the old way of implementation – even though the grades do note show such significant progress. Now, students became much more aware of the quality of their submissions – they have clear and argumentative feedback they understand – that is a good starting point for further learning in this field.

Finally, students promoted the AHP method in the organizations they visited, which might be useful for the organizations in dealing with other decision-making problems they are facing.

C. Students evaluation of workshop activity

The purpose of the survey was to examine students' opinions on peer and self-assessment through a designed rubric for the AHP activity. Questions in the designed survey are grouped into three sections: Making my own work, Works assessment, and Gained grades:

- The first section covers questions related to the making of the assignment.
- Questions in the second section are directed to the process of the assessment.
- The third section contains questions about gaining grades for work as well as for peer and self-assessment.

There are two types of questions that are used for the questionnaire. Most of the questions belong to the Likert Scale question type (24 questions in total) since the purpose is to evaluate the students' opinions and attitudes. There are 5 satisfaction levels where 1 represents *Strongly disagreement,* while 5 stands for *Strongly agreement*. Only two questions can be addressed as Text question type for which participants could enter only the number.

The survey was completed by 59 students and their feedback is analyzed by 4 statistical measures: arithmetic mean (M), median (Mdn), standard deviation (SD), and coefficient of variation (RSD). The overall results are provided in Table 2.

TABLE II. STUDENT EVALUATION OF WORKSHOP MANY

No	Questions	M	Mdn	SD	RSD
Section A: Making my own work					
A1	The evaluation criteria which were published before I could start doing on my own work (rubric) helped me in making my own work.	3.92	4	0.95	24%
A2	The rubric used for the assessment and its descriptions were clear enough.	3.86	4	1.06	27%
A3	Works examples of previous generations helped me in making my own work.	4.00	4	0.96	24%
A4	The possibility of a demo-examples assessment has been helpful to me in making my own work.	3.08	3	1.10	36%
A5	I had been modifying my work according to the criteria defined in the rubric.	3.56	4	1.18	33%
A6	The time I spent on making my own work (in minutes)	601.68	500	360.87	60%
Section B: Works assessment					
B1	Peer assessment helped me learn how to provide constructive feedback to peers.	3.29	3	1.07	32%
B2	Reviewing the works of others helped me to broaden my knowledge of the given topic.	3.42	4	1.16	34%
B3	I was confident in evaluating my own and others' works.	3.10	3	0.98	32%
B4	I am convinced that my peers are capable enough to evaluate my work.	3.34	3	0.84	25%
B5	I am convinced that I am capable enough to evaluate the work of my peers.	3.24	3	0.99	31%
B6	The rubric used for peer assessment has allowed me an entry of detailed feedback about the assignment of my peer.	3.95	4	0.90	23%
B7	The rubric used for peer assessment helped me to focus on certain details in the work of my peer.	3.93	4	0.94	24%
B8	The rubric description helped me to understand the teacher's expectations in the evaluation report.	3.88	4	0.98	25%
B9	The rubric used for peer assessment has made a review of the assigned task clearer.	3.85	4	1.03	27%
B10	The rubric used for peer assessment has made the examination process more transparent.	3.80	4	0.83	22%
B11	I think that a student does not need to know the author's names whose work he/she assesses.	4.07	5	1.14	28%
B12	I think that a student does not need to know the names of students who have evaluated her/his work.	4.14	5	1.12	27%
B13	The time I spent on the work assessment (my own work and three works of my peers, in minutes)	159.66	120	127.04	80%
Section C: Gained grades					
C1	I am satisfied with gained points for my work through peer assessment.	3.69	4	0.91	25%
C2	The points I gained for my work are realistic.	3.86	4	0.90	23%
C3	The points I gained correspond to the points I have given myself through self-assessment.	3.44	3	1.00	29%
C4	The feedback gained from peers helped me to spot mistakes in my own work.	3.56	4	0.95	27%
C5	The received feedback (descriptive part, the argumentation of points) is accurate and credible.	3.42	3	0.99	29%
C6	I am satisfied with the points I have gained for the assessment of my peers.	3.59	4	1.21	34%
C7	The points I gained for the assessment of my peers are realistic.	3.75	4	1.25	35%

Results disclosed that students' have an uppermost agreement with assertions of Section B and to those assertions which relate to anonymity in the assessment process (B12, B11). Text questions (A6, B13) are excluded from the comparison. Students' have an uppermost disagreement with assertions that relate to the possibility of trial evaluations (A4) and confidence in their ability to conduct such an assessment (B3, B5). Comparison of students' agreement of assertions by sections (i.e. A, B, C) varies between 3.61 and 3.68.

The highest relative variability has text questions where a participant could enter a number related to the time he/she spent on the making of assignment (60%) or work assessment (80%). The relative variability of other questions is lower or equal to 36%.

The survey results confirmed the qualitative analysis of qualitative analysis. Students also agree that rubric (as an assessment tool) was instrumental in the phase of creating the AHP project (A1-A3): it helped in organizing the final document, but also, it served as a control mechanism for completeness of each part of the project (A5).

Since peer and self-assessments were the new activities in their study process, students were slightly reserved in terms of confidentiality (B3-B5). However, using rubrics made the assessment easier (B6-B9).

Generally, students were satisfied with the grades they achieved (C1, C6) and they find them realistic (C2, C7).

V. CONCLUSION

The coalescence of information technology and teaching, as well as the learning process, is inevitable. There are many examples of the integration of IT into existing teaching and learning models in higher education. Also, this is conspicuous by LMS systems that are widely applied in practice. To explore the impact of IT on the learning and teaching process the, Workshop activity was created for student project activity. This project activity has been performed for years as part of the course at the graduate level, but in the traditional way in which the teacher evaluates the student's work. The implementation of the AHP project activity as Workshop activity is evaluated through qualitative analysis, quantitative analysis, and student feedback questionnaires. Student progress by the new way of implementing the AHP activity is visible from the results of all three conducted evaluations.

Generally, the analysis showed better results in terms of grades when the activity is implemented through the Workshop when compared to the traditional way. Students' satisfaction with the activity is also high. During the implementation – some minor issues were identified and will be changed in the next year. They are related to the explanations of the rubric degrees per some criteria. There are also thoughts of implementing the two-step workshop: one for the problem structure and the other for the whole project because, in some cases, the decision-making problem is not structured very well. We want to put an even higher focus on the quality of the problem structure because, if the problem is poorly structured, the solution cannot be right.

REFERENCES

[1] T. L. Saaty, "Decision making with the analytic hierarchy process," *International journal of services sciences*, vol. 1, no. 1, pp. 83–98, 2008.

[2] M. G. Marković, N. Kadoić, and B. Kovačić, "Selection and prioritization of adaptivity criteria in intelligent and adaptive hypermedia e-learning systems," *TEM Journal*, vol. 7, no. 1, 2018.

[3] N. Kadoić and I. Kedmenec, "Using ICT tools for decision making support in local government units," in *39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2016, pp. 1533–1538.

[4] "Using Workshop - MoodleDocs." [Online]. Available: https://docs.moodle.org/38/en/Using_Workshop. [Accessed: 03-Dec-2019].

[5] T. Wanner and E. Palmer, "Formative self-and peer assessment for improved student learning: the crucial factors of design, teacher participation and feedback," *Assessment & Evaluation in Higher Education*, vol. 43, no. 7, pp. 1032–1047, 2018.

[6] G. Thomas, D. Martin, and K. Pleasants, "Using self- and peer-assessment to enhance students' future-learning in higher education," *Journal of University Teaching & Learning Practice*, vol. 8, no. 1, 2011.

[7] P. Black, C. Harrison, C. Lee, B. Marshall, and D. Wiliam, "Working inside the Black Box: Assessment for Learning in the Classroom," vol. 86, no. 1, pp. 8–21, Sep. 2004.

[8] S. J. Hanrahan and G. Isaacs, "Assessing Self- and Peer-assessment: The students' views," *Higher Education Research & Development*, vol. 20, no. 1, pp. 53–70, May 2001.

[9] A. M. F. Yousef, U. Wahid, M. A. Chatti, U. Schroeder, and M. Wosnitza, "The Effect of Peer Assessment Rubrics on Learners' Satisfaction and Performance Within a Blended MOOC Environment," in *Proceedings of the 7th International Conference on Computer Supported Education - Volume 2: CSEDU,* 2015, pp. 148–159.

[10] C. J. Brame, *Science Teaching Essentials: Short Guides to Good Practice*. Academic Press, 2019.

[11] P. Dickinson and J. Adams, "Values in evaluation – The use of rubrics," *Evaluation and Program Planning*, vol. 65, pp. 113–116, Dec. 2017.

[12] S. Domínguez-García, M. I. Garcia-Planas, R. Palau Martín, and J. Taberna Torres, "EVALUATING AN E-PORTFOLIO FOR A LINEAR ALGEBRA COURSE USING RUBRICS," in *INTED2015 Proceedings*, Madrid, Spain, 2015, pp. 1366–1372.

[13] F. Chowdhury, "Application of Rubrics in the Classroom: A Vital Tool for Improvement in Assessment, Feedback and Learning," *International Education Studies*, vol. 12, no. 1, p. 8, 2019.

[14] F. González-Andrés and B. Urbano, "THE USE OF RUBRICS TO ASSESS COMPETENCES," in *INTED2018 Proceedings*, Valencia, Spain, 2018, pp. 4255–4261.

[15] A. Horvat, M. Dobrota, M. Krsmanovic, and M. Cudanov, "Student perception of Moodle learning management system: a satisfaction and significance analysis," *Interactive Learning Environments*, vol. 23, no. 4, pp. 515–527, Jul. 2015.

[16] C.-H. Luk, K.-K. Ng, and W.-M. Lam, "The Acceptance of Using Open-Source Learning Platform (Moodle) for Learning in Hong Kong's Higher Education," in *Technology in Education. Innovative Solutions and Practices*, Singapore, 2018, pp. 249–257.

[17] V. A. Nguyen, "A peer assessment approach to project based blended learning course in a Vietnamese higher education," *Education and Information Technologies*, vol. 22, no. 5, pp. 2141–2157, Sep. 2017.

[18] D. Dolezal, R. Motschnig, and R. Pucher, "Peer Review as a Tool for Person-Centered Learning: Computer Science Education at Secondary School Level," in *Teaching and Learning in a Digital World*, Cham, 2018, pp. 468–478.

[19] V. Damnjanovic, S. Jednak, and I. Mijatovic, "Factors affecting the effectiveness and use of Moodle: students' perception," *Interactive Learning Environments*, vol. 23, no. 4, pp. 496–514, Jul. 2015.

[20] M. Tobajas, C. B. Molina, A. Quintanilla, N. Alonso-Morales, and J. A. Casas, "Development and application of scoring rubrics for evaluating students' competencies and learning outcomes in Chemical Engineering experimental courses," *Education for Chemical Engineers*, vol. 26, pp. 80–88, Jan. 2019.

Studying dynamic mathematics software in the professional training of teachers of computer science, mathematics, and IT specialists

I. V. Shishenko*, V. H. Shamonia*, V. S. Loboda*, V. V. Punko*, Yu. V. Khvorostina* and A. A. Voitenko*
* Makarenko Sumy State Pedagogical University, Sumy, Ukraine
shiinna@ukr.net, shamona@gmail.com, vasaloboda27@gmail.com, vladpunkovlad@gmail.com,
y-y-y@fizmatsspu.sumy.ua, voitenko@fizmatsspu.sumy.ua

***Abstract* – The article deals with the need to study dynamic mathematics software by future mathematics teachers, informatics teachers, and IT-specialists. Emphasis is placed on the ability to organize an effective learning process using dynamic mathematics software. The authors raise the issue of practical training of mathematics teachers, students of specialty 014 Secondary education (Mathematics, Informatics), and 122 Information's Technology of Makarenko Sumy State Pedagogical University (Ukraine), an integral part of which is the formation of skills of mathematical objects cognitive modeling. Examples of tasks aimed at developing knowledge and ability to use the tools of dynamic mathematics software in the future professional activity-are given. It is noted that the formation of such skills occurs, in particular, when writing a text of a lesson fragment using dynamic mathematics software, with its subsequent presentation and discussion. The dynamic of qualitative changes in the content of the laboratory work is shown.**

Keywords - dynamic mathematics software; studying dynamic mathematics software; DMS; using DMS; mathematics preparation; professional preparation of teachers IT-specialty, professional preparation

I. INTRODUCTION

Informatization of mathematical education has led to the use of various information aids in mathematics education, among which a separate group should be allocated to dynamic mathematics software (DMS). This software provides dynamic manipulation of mathematical objects, which is based on the preservation of the basic design algorithm. In other words, in the environment of this software, it is possible to study objects in dynamics, which is achieved by interactively operating some elements of the structure (for example, changing the position of the points of the object, the length of the elements, etc.). Among such programs, there are GRAN (GRAN 1, GRAN 2d, GRAN 3d), DG [1], MathKit [2], The Geometer's SketchPad, Cabri [3], GeoGebra [4] and more. Their work is intuitive and identical - basic objects are built, then they can be dynamically modified and observed for their specific qualitative properties and quantitative characteristics.

The possibility of interactively changing the created structure while changing the position of the original objects has attracted the attention of scientists and methodologists, who study the problems of modern mathematical preparation and often recommend involving the mentioned software in the educational process. Scholars and practitioners point out not only the need to use such software to facilitate the work of the teacher in the classroom but also the advisability of using such software to create their learning world for each student, which should provide personal searches and their discoveries. These are the software where constructive approaches, informal solutions, boundary case studies are possible, that allows us to penetrate deeply enough into the essence of the phenomenon under study, focus on constructing models and interpreting the results and move away from the routine (computational) side of the problem.

II. DYNAMIC MATHEMATICS SOFTWARE

A. Analysis of current research on the problem of the use of DMS in mathematical preparation

The involvement of such means in the teaching of school mathematics confirms its effectiveness, as evidenced by the work of both national (Ukrainian) and foreign scientists: E. F. Vinnichenko, Y. V. Goroshko, M. I. Zhaldak [5], O. P. Zeleniak [6], S. A. Rakov [7], O. V. Semenikhina, M. G. Drushlyak (Ukraine) [8], V. N. Dubrovsky [9], M. V. Shabanova, T. S. Shyrykova (Russia) [10], I. S. Khrapovitsky (Belarus) [11], M. Hohenwarter (Austria) [12, 13], and others.

A more detailed analysis of the works of the mentioned scientists revealed the following. The study of the willingness and the readiness of future mathematics teachers to use DMS in their professional activity is described in the work [14].

The peculiarities of introducing a praxeological approach in the process of formation in future computer science teachers' skills of rationally choice the DMS when solving a particular class of tasks are highlighted in [15].

In [16], the authors investigate the problem of using GeoGebra to improve the quality of students' learning achievements by applying a constructive approach to solving geometric extremum problems.

The results of [17] reveal the essence of such a modern educational trend as the BYOD approach by using the GeoGebra cloud service.

The aforementioned and other findings confirm the relevance of the use of DMS in mathematical training of specialists, which is why we experimented on the use of such software in the mathematical training of mathematics teachers, computer science teachers, and IT professionals.

The purpose of the study: to substantiate the feasibility of using DMS in the training of mathematics teachers, computer science teachers, and IT specialists.

The purpose of the study determined the need to solve following tasks: 1) to analyze the practical usage of DMS in preparation; 2) to develop a special course of DMS, in particular, to clarify it;' content, and conduct an expert assessment of it; 3) to conduct a pedagogical experiment in purpose to examine the effectiveness of this special course.

For solving the research tasks authors used theoretical (analysis of training plans, analysis of the content of training courses, analysis of computer tools of DMS, generalization of obtained results) empirical (observation, expert evaluation) and statistical (Pearson's criterion) research methods.

B. The practice of using DMS computer tools

The use of DMS in Makarenko Sumy State Pedagogical University began in 2006 when the special course "Computer Science in Mathematics" was introduced for future math teachers at first. The mentioned special course has been modernized for ten years, and its study has shown the expediency of using DMS in mathematical training not only for mathematics and computer science teachers but also for the training of specialists in the field of information technology (IT specialists).

The introduction of DMS in mathematical preparation is preceded by the study of DMS itself (since 2016, there is software such as MathKit, The Geometer's SketchPad, Cabri, GeoGebra), which is more focused on mastering computer tools and comparative analysis of such tools in different software.

In addition to the usual tools for constructing elementary geometric figures (point, line, segment, ray, circle, triangle, polygon, middle segment, bisector, etc.), the focus is on tools for constructing geometric locus ("Trace" and "Locus"), setting parameters ("Parameter"), statistical tools (sample characteristics, statistical distributions, etc.).

The DMS study course is followed by the study of mathematical disciplines based on the use of DMS. These include elementary mathematics courses, computing practicum, some topics in projective geometry, probability theory, and statistics. The technology of using DMS is to bring computer tools to lectures to demonstrate examples and to engage software themselves as research environments in the classroom.

But it should be noted that the organization of learning requires the teacher's ability to reformulate mathematical problems, turning them into research: instead of "prove that ..." is put "to confirm that ..." or "to find exceptions to ..." or "to refute ...".

Let us dwell on this in detail.

1. The theorem "The medians of a triangle intersect at one point and is divided by the point of intersection in the ratio of 2: 1, starting from the vertex" can be reformulated as a problem: "Investigate in what relation the medians of the triangle are divided by their intersection point (from the vertex)."

Students are invited to perform an experiment, complete a table, draw a conclusion, and formulate a hypothesis (Fig. 1).

2. The theorem on the coincidence of the median and the height of an isosceles triangle can be reformulated as: "In the triangle, the median is perpendicular to the side. What will be the triangle?"

Students are encouraged to independently construct such a triangle with arbitrary input data and explore its characteristics (lengths of sides, angles).

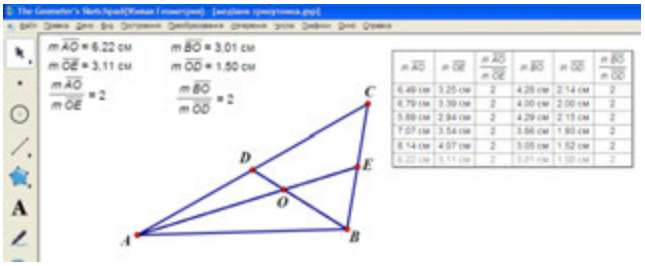


Figure 1. Organization of the study on the planimetry lesson

With DMS, geometric visualizations are refined at a higher level. A constructive approach to the definition of a cone as a body formed by the rotation of a right triangle around the cathetus is given in Fig. 2. Students are encouraged to construct other geometric shapes on a structural basis (sphere, hyperboloid, paraboloid cube), which requires them to refine the plane to be rotated and the axis around which the object rotates.

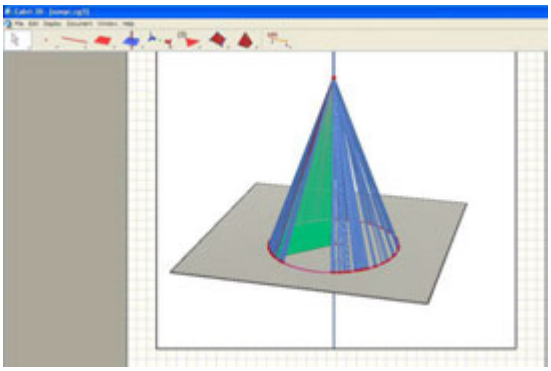


Figure 2. Visualization of the cone design

It should also be noted that it is not appropriate for all tasks and in all situations to use dynamic mathematics software. These include computational problems (find the length, volume, calculate the determinant, find the boundary, etc.), analytic conversion tasks (simplify an expression, highlight full square, etc.).

C. *Features of construction and study of the special course*

The special course is built on a practical basis. It is designed for 3 ECTS credits (of which 10 hours are lectures and 26 laboratory hours). This study involves mastering the topics listed in Table I.

The development and modification of the DMS themselves, as well as the improvement of curricula for the training of specialists, required changes in the content of the special course in which the list of laboratory works was changed. Thus, in 2016 there were two separate laboratory works "Solving Geometric Research Problems" and "Solving Locus Problems" (2 exercises in each of 12 variants), as well as tasks for independent work on filling and supplementing the table for comparative analysis of computer tools of dynamic mathematics software (Table II).

TABLE I. TOPICS OF LABORATORY WORKS (2016)

№ lab. work	The topic	Number of tasks
1	Solving equations, inequalities, and their systems	5
2	Solving tasks with parameters	3
3	Using the computer to learn the beginnings of analysis	4
4	Statistical calculations in DMS	4
5	Solving geometric research problems	2
6	Additional computer tools	3
7	Locus problems solving	2
8	Step-by-step demonstrations solving constraction problems	3
9	Study of the topic "Geometric transformations" with DMS	2
10	Using the coordinate method in the DMS	4
11	Solving stereometric problems in DMS	4
12	Construction of sections of polyhedra in The Geometer's ScetchPad	3
13	Topic "Geometric transformations in space" with DMS	3

TABLE II. TOPICS OF LABORATORY WORKS (2019)

№ lab. work	The topic	Number of tasks
1	Solving equations, inequalities, and their systems. Using your computer to learn the beginnings of analysis. Solving tasks with parameters.	5
2	Test laboratory work.	not scheduled for 2016
3	Statistical calculations in dynamic mathematics software	4
4	Solve geometric research and locus problems.	3
5	Step-by-step demonstrations when solving constraction problems	2
6	Test laboratory work.	not scheduled for 2016
7	Additional computer tools. Organization of automated knowledge control.	2
8	Solving stereometric problems in dynamic mathematics software	3
9	Construction of sections of polyhedra in The Geometer's ScetchPad	3
10	Test laboratory work.	not scheduled for 2016

But it turned out that during the preliminary joint discussion of theoretical material and detailed analysis of examples with a focus on possible typical mistakes, students complete the proposed tasks fairly quickly. Therefore, taking into account their own experience and expert judgment, some laboratory work was combined into one and the number of control works increased (Table III).

Since 2017, laboratory tests have been introduced where students are asked to solve mathematical problems on different topics. Here are examples of tasks for several variants of test lab work on planimetry (Table IV).

TABLE III. TASKS OF LABORATORY WORKS 2016 AND 2019

	The proposed tasks for the laboratory work in 2016
1	The angle based on the diameter of the circle is straight.
2	Investigate the properties of an angle whose vertex lies outside the circle and the sides of the angle intersect the circle.
	The proposed tasks for independent work in 2016
3	Complete a table for comparative analysis of dynamic mathematics software and complete it.

	The proposed tasks for the laboratory work in 2016
1	The two vertices of the triangle are fixed at points A and B, AB = c. Its third vertex C moves in a circle of radius b centered at point A. What line traces the point D of the intersection of the bisector of angle A with the side BC? Use different dynamic mathematics software to demonstrate.
2	Find the locus of the points whose distances to the fixed point and the fixed line are constant and equal to 3.
	The proposed tasks for independent work in 2016
3	Complete a table for comparative analysis of dynamic mathematics software and complete it.

	The proposed tasks for the laboratory work in 2019
1	Develop demonstrations using a dynamic expression in Gran2d, DG. The centerline of the trapezoid is half the sum of the bases.

2	Develop demonstrations using the dynamic expression in The Geometer's SketchPAd, GeoGebra. Investigate the properties of external one-sided, external versatile, and corresponding angles formed by the intersection of two parallel lines by the third one.
The proposed tasks for independent work in 2019	
3	Solve the problem with MathKit, GeoGebra. AB is a chord of a given circle. Find the locus of the midpoints of all the chords of this circle whose length is AB.
4	Fill in the comparative table of tools in different DMS.

TABLE IV. TASKS OF THE TEST LABORATORY WORK ON PLANIMETRY

B-т	**The Tasks**
The task for proof. Prepare the task for proof by pre-reformulating it as a research task. Create a model with dynamic expression and tips (if needed). Develop a table to record research findings and conclusions.	
1	A circle is inscribed in a square. Show that the sum of the squares of distances from the point of the circle to the vertices of the square does not depend on the choice of this point.
2	Chords are drawn through the point inside the circle. Show that the product of the segments of each of the chords to which it divides a given point is constant for a given circle.
3	Show that if the bisectors of a triangle break it into six triangles of equal areas, then this triangle is equilateral.
4	Show that in an arbitrary quadrilateral the product of diagonals is equal to the sum of the products of opposite sides (Ptolemy's theorem).
The Locus Problem. Record the step-by-step algorithm for constructing the desired GMT. Play the build progress using the buttons. Demonstrate (if possible) the differences of the GMT built using the *Track*, *Locus*, and *Parametric Color* tools.	
5	If you mark the point *P* in the triangle and connect it to the vertices, then the triangle will break into three smaller triangles. Find the locus of points for which the sum of the areas of two of these triangles will be equal to the area of the third one.
6	Find the locus of points *M* that lie inside the *ABCD* rhombus, for which the sum of the angles *AMD* and *BMC* is 180^0.

D. The results of the pedagogical experiment

The base of the research was Makarenko Sumy State Pedagogical University. The experiment itself pursued two goals: 1) to determine whether the level of educational achievement of students in terms of their mathematical preparation can be tested by the developed list of tasks; 2) whether the use of DMS affects the level of students' educational achievement in their mathematical preparation.

Achieving the first goal involved analyzing the opinions of working teachers who expertly evaluated the list of proposed tasks.

Achievement of the second goal was based on comparing the distribution of learning outcomes for the two samples: the experimental group (EG) studied the special course and the following mathematical courses using DMS, and the control group (CG) mastered the special course with the DMS, but further courses in the study of other mathematics disciplines using traditional approaches (without DMS). At the end of the study course (the second year of study) and the end of the fourth year, we made an assessment of the educational achievements in mathematical preparation. This assessment was based on the objectives of an external independent assessment of mathematics knowledge in 2018 [11].

Let's describe the statistical analysis of the results in detail.

1. Experts (19 persons from the working teachers of mathematics, methodologists, and university teachers) at the beginning of the pedagogical experiment were given a list of exercises, the implementation of which was to provide sufficient mathematical training for further study. After familiarizing themselves with the list of tasks for the independent work, the experts were asked the following questions.

1. Does the given system of exercises cover the main sections of elementary mathematics (at the basic school level)?

2. Are the proposed tasks sufficient to master the DMS tools?

3. Do we need to change the number of proposed tasks in general? within the study of a particular topic?

4. Is it necessary to solve the proposed number of tasks in general? as part of a particular topic?

The obtained expert assessments (Table 5) determined the quantitative and qualitative choice of tasks to support the study of the basic topics of the elementary mathematics course. The final distribution of topics and the number of tasks is shown in Table V.

The peer-review agreed on the quantitative and qualitative composition of the tasks, which were then offered to students to determine the level of their academic achievement.

2. Comparison by Pearson's criterion of the results of the academic achievements of students of KG and EG revealed their statistical similarity before studying the special course at the significance level of 0.05 ($T_{exp} = 1.86 < T_{crit} = 3.84$).

At the same time, the comparison of the results at the end of the experiment confirmed the alternative hypothesis about the positive impact of DMS on the level of students' educational achievement in mathematics ($T_{exp} = 4.45 > T_{crit} = 3.84$), which cannot be explained by random factors.

TABLE V. EXPERT EVALUATIONS

	Question 1		**Question 2**		**Question 3 (in general)**		**Question 4 (in general)**	
	Yes	No	Yes	No	Yes	No	Yes	No

I expert evaluation, (2016)	42%	58%	53%	47%	89%	11%	84%	16%
II expert evaluation, (2019)	84%	16%	79%	21%	21%	79%	95%	5%

III. CONCLUSION

To develop the skills of working with the DMS tools, it is necessary to use a specially selected system of tasks, which should be consistent with the course of elementary mathematics and ensure the involvement of the maximum number of tools of different DMSs.

The developed list of tasks is effective in the context of preparing the future teacher of mathematics and computer science to use the DMS tool, which is confirmed by a positive assessment of its quantitative and qualitative content by experts.

The developed system of tasks is effective in the context of mathematical training of mathematics and informatics teachers, as well as future IT specialists, which is confirmed by Pearson's statistical criterion at the significance level of 0.05.

It should be noted that DMS involvement must be well-considered, appropriate, and justified. Dynamic mathematics software should not be used as a tool for constructing a static drawing, that is, used where there are enough blackboard and chalk or paper and pencil. It is necessary to find such “points” of dynamic mathematics software in the mathematical preparation, which would reveal the zest of such software, bring the idea of dynamism for the development of mathematical imagination, stimulate creative thinking, stimulate mathematical discovery, provide a basis for reflection with further conclusions.

REFERENCES

[1] Dynamic Mathematics Software DG. Retrieved from: http://dg.osenkov.com/index_ru.html.

[2] 1C: Learning Software. Retrieved from: http://obr.1c.ru/mathkit.

[3] CABRILOG. Innovative Maths Tolls. Retrieved from: http://www.cabri.com.

[4] GeoGebra. Retrieved from: http://www.geogebra.org.

[5] Zhaldak M.I, Goroshko Y. V., Vinnichenko E.F. Mathematics with computer. A guide for teachers. K .: RNNU "DINIT", 2004. 252p.

[6] Zeleniak O.P. Technologies of application of dynamic geometry environments. Information Technologies and Learning Tools, 2013. Is. 36, №4. Retrieved from: https://journal.iitta.gov.ua/index.php/itlt/article/view/861/657.

[7] Rakov S.A. Mathematical education: competence approach using ICT: Monograph. H.: Fakt, 2005. 360 p.

[8] O. Semenikhina, M. Drushlyak. “On the Results of a Study of the Willingness and the Readiness to Use Dynamic Mathematics Software by Future Math Teachers”, 11th International Conference on ICT in Education, Research, and Industrial Applications: Integration, Harmonization, and Knowledge Transfer (ICTERI 2015), pp.21-34, May 14-16, 2015.

[9] Dubrovsky V. Learning to work with the "Mathematical Designer". Mathematics, 2009. No.13. P. 2-48.

[10] Shabanova M.V., Shyrykova T.S. Training in proof using an interactive geometric environment. Yaroslavl Pedagogical Bulletin, 2012. Vol. 2, No. 3. P. 86-92.

[11] Khrapovitsky I.S. Heuristic polygon. Computer tools in education, 2003. No1. P. 15-26.

[12] M. Hohenwarter, J. Keith “Ways of Linking Geometry and Algebra: The Case of GeoGebra”, Proceeding of the British Society for Research into Leaning Mathematics, 2007, 27, 3, pp. 126-131.

[13] M. Hohenwarter, J. Hohenwarter, Z. Lavicza “Introducing Dynamic Mathematics Software to Secondary School Teachers: The Case of GeoGebra”, Journal of Computers in Mathematics and Science Teaching, 2008, 28, 2, pp. 135-146.

[14] O. Semenikhina, M. Drushlyak, I. Shishenko, V. Zigunov. “Using a praxeology approach to the rational choice of specialized software in the preparation of the computer science teacher”, TEM JOURNAL – Technology, Education, Management, Informatics, 2018, Vol.7, No.1, pp 164-170.

[15] O. Semenikhina, M. Drushlyak, I. Zigunova, D. Budyanskiy. “Geogebra as means of improving the quality of education”, 14th International Conference on ICT in Education, Research, and Industrial Applications: Integration, Harmonization, and Knowledge Transfer (ICTERI 2018), pp. 331-345, May 14-17, 2018.

[16] E. Semenikhina, M. Drushlyak, Yu. Bondarenko, S. Kondratiuk, N. Dehtiarova. “Cloud-based service GeoGebra and its use in the educational process: the BYOD-approach”, TEM JOURNAL – Technology, Education, Management, Informatics, 2019, Vol.8, No.1, pp. 65-72. DOI: 10.18421/TEM81-08.

[17] Education.ua. Retrieved from: https://ru.osvita.ua/doc/files/news/608/60848/Matematyka-Osnovne-ZNO_2018-Zoshyt_1.pdf

The automated control of students achievements by using paper clicker Plickers

M. G. Drushlyak*, O. V. Semenikhina*, S. M. Kondratiuk*, T. M. Kryvosheya**, A. V. Vertel* and N. M. Pavlushchenko*

* Makarenko Sumy State Padagogical University, Sumy, Ukraine

** Vinnytsia Mykhailo Kotsiubynskyi State Pedagogical University, Vynnytsia, Ukraine

e.semenikhina@ fizmatsspu.sumy.ua, marydru@fizmatsspu.sumy.ua, kondratiuks.m@gmail.com, krivosheja.tatiana@gmail, antonvertel@ukr.net, pavnm1972@ukr.net

***Abstract* – The article deals with the problem of organizing student achievement control using an online testing based on the Plickers software. It is established that schools' teachers in Sumy region (Ukraine) and future mathematics teachers (Makarenko Sumy State Pedagogical University, Ukraine) are mostly unfamiliar with such technology. The usefulness of modernizing the content of educational curricular and master classes for working teachers to integrate the testing technology by using the mobile application Plickers into the educational process is confirmed. The article describes the process of preparation and experience of organizing an online testing of future mathematics teachers by using Plickers based on the study of the discipline "Mathematical logic and theory of algorithms". The advantages and disadvantages of using such a tool in the educational process are noted. The results of the effectiveness of the use of Plickers in the educational process are given.**

Keywords – paper clicker; Plickers; online testing; student achievement; educational process

I. INTRODUCTION

Digital transformation of the educational sphere is necessary and irreversible process. This is emphasized in "Vision of Ukraine – 2030" by Ukrainian Institute of the Future [1], which correlates with the ideas of a perspective portrait of education of the 21st century – Education 4.0. It states the urgent need to revise the curricula of higher education institutions in order to integrate digital technologies into the educational process that meet the requirements of Industry 4.0. Outdated teaching methods, inaccessibility of digital technologies for the educational process have led to extremely low levels of digital literacy in all available segments of the state education system (preschool, primary, secondary, and higher). Therefore, the idea of introducing digital technologies, in particular, the means of automated control of students' achievements is currently relevant.

The aim of the article is to describe the process of preparation and experience of organizing an online testing of future mathematics teachers by using Plickers, to prove the effectiveness of the use of Plickers in the educational process.

II. MOBILE LEARNING TECHNOLOGIES

One of the problems that teachers of Ukrainian educational institutions face is the lack of computers and limited access to computer classes. That is why the involvement of digital technologies in the educational process is "in bits and pieces", is fragmented – they are used either by the teacher himself or by the student at home during independent work. At the same time, young people often use their own mobile devices not only to communicate on networks but also to support their own educational activities. Therefore, it is important to involve your own mobile devices (smartphones, tablets, netbooks, etc.) in the educational process. This approach is called BYOD (Bring Your Own Device). Implementation of the BYOD approach should be considered in the context of mobile learning. The authors have already addressed the problem of implementing the BYOD approach when using the GeoGebra cloud service in teaching natural sciences and mathematics [2].

At the same time, it is advisable to consider the use of mobile applications within the BYOD approach not only in teaching, but also from the positions of students' achievement control. Let us note the mobile application *Plickers* in this context.

III. MOBILE APPLICATION PLICKERS

A. The analysis of current research on the problem of using the mobile application Plickers

When organizing digital exams, student response systems (online formative assessment, audience response systems) including *Clickers, Plickers, Kahoot, and Socrative* are becoming more popular recently. Plickers has a number of advantages over other applications of this type [3]. Let us note that *Clickers* costs around $ 1000-1500, each set containing 32 cartridges for 32 students. *Plickers* is an online software that works like a classroom response system, and it is free. This application is easy to use, does not need any hardware devices. *Plickers* require printed-paper cards and only one mobile phone or tablet to scan. Students' answers are automatically collected, stored, analyzed on the company website, and made available to the teacher.

The feasibility and effectiveness of using the Plickers mobile application is highlighted in the following findings.

J. R. De Thomas, V. López-Fernández, F. Llamas-Salguero, P. Martín-Lobo, and S. Pradas analyzed the relationship between educational achievement, engagement rate, and student creativity when using *Plickers*. They claim that active participation in the *Plickers* survey improves the quality of student knowledge [4].

T. A. Wood, K. Brown, and J. M. Grayson [5] examined the perception of *Plickers* by senior technology students as a method of achievement control. They found that this technology improves the audience, perceives the students as exciting, unusual and fast in terms of test results. S. Wuttiprom, K. Toeddhanya, A. Buachoom and K. Wuttisela [6] obtained similar results.

M. G. McCargo [7] examined the impact of the use of *Plickers* technology on the behavior of high school students and the perception of Plickers high school teachers as a socially sound method of combating student behavior. The author found out that *Plichers* provide every student with opportunities to respond, which an effective instruction strategy for promoting learning. Based on the results were mixed ratings of social validity from the teachers participating in this intervention.

A. Gürişik [8] analyzed the opinions of high school students regarding the use of *Plickers* and the problems that generated by students using this technology. It was found that *Plickers* makes a positive contribution to the lesson when it is used in education. The author approved that students find using *Plickers* in lessons enjoyable. They think that after the test, *Plickers*' showing the correct and incorrect answers instantly has corrected their mistakes. Plickers enables a significant number of high school students to study more. A considerable number of high school students think that their success increases in *Plickers* tests.

O. Demirkan, A. Gürişik and O. Akin [9] explored teachers' opinions regarding the use of *Plickers* in their professional activities. It was revealed that using *Plickers* saves time, shows the overall success of the class, increases class participation and competition, attracts students' attention. It was determined that this program will enhance the motivation of students, students will enjoy it, students will get instant feedback, success and attendance will increase, and students will answer more questions. It was obtained that almost all of the teachers expressed that they would use Plickers in their class. It has been concluded that almost all of the teachers think that such trainings are very useful.

E. A. Michael, I. E. A. Ejeng, M. A. Udit, and M. M. Yunus [10] showed teachers' opinions regarding the positive impact in assessing pupils' understanding of reading comprehension classroom in a fun and stress-free way by using *Plickers.*

At the same time, there has been no systematic study of the use of the *Plickers* app in the educational process in Ukraine; in particular, in mathematics teaching, so scientific research in this area is relevant.

B. The organization of automated achievements control by using the mobile application Plickers

To start using the application, you need to register at http://plickers.com and enter the alphabetical list of the group (+ New Class).

Next, you must attach the appropriate tests to each group, that is, to add the tests to the so-called queue (Add to Queue). The queue is a series of questions to ask the group students at the next lesson. Once a question is asked, it is removed from the queue. Each time you need to update the queue of questions you need.

The teacher prints out the cards. Card layouts are available for download on the official site *Help / Get Plickers Cards*. Five different sets of cards are offered. Each card is unique and has its own number that corresponds to the student's number in the list. Therefore, the survey is personalized.

Each side of the card has a different answer – A, B, C, D (Fig.1). Cards can be used for every test, so it is worth printing them on paperboard.



Figure 1. Students are answering the question

The teacher must download the Plickers app on his device. During the testing, the teacher should go to his own page and select a group and a test. At the same time, he should select *Now Playing* mode at the website on computer.

The teacher asks the question (the question is parallel displayed on the screen and on the mobile phone). The question can be selected from a mobile phone, the teacher does not have to be near the computer all the time, all control is from the phone (Fig.2).



Figure 2. The teacher is reading the question

Students choose the answer by lifting the cards upwards. Using the mobile application, the teacher scans students' answers in real time and the results are stored in the database (Fig.3). The student can change his answer, but only the one that was at the time of scan will be counted.



Figure 3. The teacher is scanning student answers

The results are available both on the mobile app and on the site (on screen) for instant announcement and processing (*Scoresheet*) (Fig.4).

Name	Total	Які з наступних тверджень не	Висловлення А: "Існує від'ємне ціле	_____ двох висловлень і це	Яка логічна операція позначається	Формула називається _____
Class Average	51%	60%	73%	53%	67%	7%
Бакланов		-	-	-	-	-
Гаррибаєв		-	-	-	-	-
Джуманазаров	30%	A	A	C	D	C
Диченко		-	-	-	-	-
Захарченко	65%	B	B	D	B	B
Заєць	65%	B	B	D	B	B
Какаджанов	15%	D	B	C	A	A
Кирюз	65%	B	D	D	B	B

Figure 4. The detailed test results

The teacher can also print the test results for both the whole group (*Reports*) (Fig.5) and for each student to work on mistakes (*Students Reports* or pre-clicking on the student's name). You can export the results table to an MS Excel file (*Export Data to CSV*).

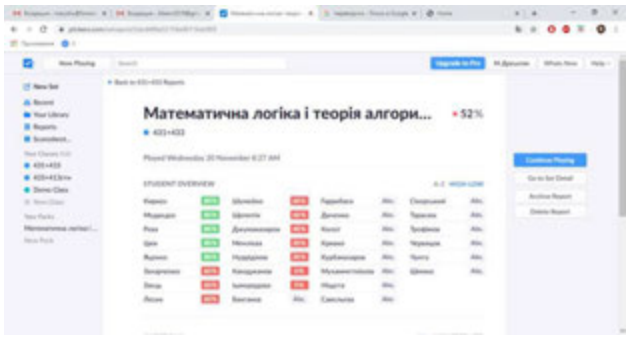


Figure 5. Test results

C. *Advantages and disadvantages of organizing achievements control by using the mobile application Plickers*

Here are the organizational and methodological advantages and disadvantages of using the Plickers mobile application.

Organizational advantages:

- it is ease of use (simplicity);
- it is free;
- it does not demand student training to be used;
- it does not require students' devices, only the teacher's mobile device is required;
- answer cards can be reused.

Among the organizational disadvantages of the Plickers app, the following can be distinguished taking into account the specific features of mathematics lessons:

- inability to insert formulas into question text or answers (only as drawings in the text of the question);
- there are only two types of closed-ended questions - single choice and multiply choice or true / false format;
- one test can have a maximum of five questions, so the teacher will need to create several tests and run them one by one. In this case, the results are provided for both a separate test and a series of tests;
- it requires a stable channel of access, both on the computer and on the mobile phone;
- it is not suitable for large student groups because the camera of the phone cannot fix all the answer cards.

Methodical advantages:

- the use of Plickers intensifies the learning process, brings elements of interactivity into the control process;
- it is a kind of fun for students;
- the teacher immediately sees who of the students answered correctly and who did not;
- students can change their answers, only the answer that was at the time of scanning is fixed;
- students do not see the answers of others;
- test results are instant, they are summarized in a table both for students of the whole group and for each student separately;
- for each student, you can print test results for each question to make corrections.

Methodological disadvantages:

- it cannot be employed in distance education, only face-to-face education.

The main ways to use the Plickers app in the classroom are a face-to-face testing at the end or at the beginning of the lesson, independent work as well.

IV. EXPERIMENTAL RESULTS

A. Research Questions

We needed to consider two problems: 1) whether the Plickers testing required more time than computer-based testing; 2) whether the computer testing is identical control form to the Plickers survey for evaluating student learning outcomes?

B. Methods

Until 2017, the educational process was provided for the organization of control of students' educational achievements in the form of computer testing. For each topic, 20 test questions were developed with four options for each. Computer testing was carried out at the end of each topic. The test was designed for 30 minutes of time.

With the advent of Plickers, the idea arose to introduce validation of academic achievement based on this tool because of its novelty as a means of control and interest in the technology of questioning by students.

Two years of experimentation, according to those teachers who used the tool, confirmed the possibility of its use, along with computer testing. However, the expediency and effectiveness of using Plickers required a pedagogical experiment.

After studying each topic (5 topics in total), students had two tests: computer-based testing and a Plickers survey. The questions were similar, but the order of the answer options, the order of the tasks themselves, and the numerical values of the output of the task conditions were changed.

We describe the results of the experiment on the example of the study of the discipline "Mathematical logic and theory of algorithms", where we studied:

Topic 1. Boolean algebra.

Topic 2. Propositional logic.

Topic 3. Predicate logic.

Topic 4. First-order logic.

Topic 5. Elements of Theory of computation.

C. Participants

The pedagogical experiment lasted for 2017-2019. It was attended by 101 persons – students of specialty 014 Secondary education (Mathematics and Computer Sciences).

D. Results

Statistical difference in time of control measures was checked according to Student's t-test (Table I).

TABLE I. AVERAGE TIME FOR TASKS (MIN.)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Computer testing	30	26	30	27	28
Plickers	21	24	24	25	23

The Excel spreadsheet and the Data Analysis package were used to process the results (Table II).

TABLE II. THE RESULTS OF AVERAGE ANALYSIS ACCORDING TO STUDENT'S T-TEST

	Variable 1	Variable 2
Average	28,2	23,4
Experimental	5	5
Hypothetical mean difference	0	
df	4	
t- statistics	3,638871	
P(T<=t) one-sample	0,010993	
t critical one- sample	2,131847	
P(T<=t) two- sample	0,021987	
t critical two- sample	2,776445	

The results confirm a statistical difference between the average time spent on computer testing (28.2 minutes) and the Plickers survey (23.4 minutes).

The impact of the control form on learning achievements was tested according to McNamara's test.

Controls were assumed to be based on the same material: first, students were computer-tested, then the same set of questions and answers was conducted using Plickers. Based on the results, students were divided into two categories: mastered – not mastered. The first group included students who answered correctly 13 or more questions. Other students were assigned to the second group.

The results of the double control of students' achievements represent the measurement on the scale "Yes – No". The results are presented in Table III.

TABLE III. THE RESULTS OF DOUBLE STUDENT'S ACHIEVEMENTS CONTROL

		Results using Plickers		
		mastered	**not mastered**	
Computer test results	**mastered**	a	b	a+b
	not mastered	c	d	c+d
		a+c	b+d	a+b+c+d

Under these conditions, McNamara test may be used to identify the significance of the difference in student distribution by the level of academic achievement.

Hypothesis H_0 is tested: the control form does not impact the distribution of students by the level of their academic achievement. Then the alternative hypothesis H_a is "the distribution of students by the level of their academic achievement depends on the chosen control form". Under these conditions, McNamara's two-sided test is used. For $n=b+c>20$ the value of the statistics T_{exp} is calculated by the formula $T_{exp} = \frac{(b-c)^2}{b+c}$. At the significance level $\alpha = 0{,}05$ we have a critical value of $T_{crit}=3{,}84$.

Table IV shows the results for 2017, 2018, 2019 for topic 1.

TABLE IV. THE RESULTS OF THE PEDAGOGICAL EXPERIMENT BY MCNAMAR TEST

Year	a	b	c	d	T_{exp}	Note
2017	51	14	25	11	3,10	H_0 is accepted
2018	57	10	16	18	1,38	
2019	62	11	16	12	0,92	

Taking into account that $T_{exp} < T_{crit}$ for each year and each topic, we had no reason to reject the H_0 hypothesis at the significance level $\alpha = 0{,}05$.

E. Discussion

Establishing statistical difference in time of control measures we found out a smaller gap in the mean time for topics 2 and 4 is evident, when the theoretical knowledge is largely tested. A larger gap in average time for topics 1 and 3 when testing skills to solve practical tasks. Results on topic 5 are mixed.

The shortening of the average time in the Plickers survey is explained by the peculiarity of using this technology – fixing of students' answers is instantaneous, as is the change of questions on the screen, while in computer testing each student at his own pace answers the questions.

When interviewing with Plickers, students answer in their own pace. Students with low educational achievements or a phlegmatic or melancholy psychotype are guided by the response rate of students with high and intermediate educational achievements and often thoughtlessly answered.

In computer testing, there have been cases where time is running out and the student has not yet answered all of the test questions. With the use of the Plickers technology, such situation is impossible.

The average time of computer-based testing is largely determined by the pace of completion of the tasks of the majority, i.e. students with average academic achievement. The average interview time with Plickers depends on the pace of task solving for students with high levels of academic achievement.

Studying the impact of the control form on learning achievements we found out that the control form does not impact the distribution of students by the level of their academic achievements.

V. CONLUSION

The results of the experiment support the authors' hypotheses that computer-based testing takes longer than testing using *Plickers*. It is statistically confirmed that the form of control (computer-based testing or *Plickers*) does not impact the distribution of students by their level of academic achievements.

The generalization of the other scientists' findings confirmed that similar studies were not conducted, hypotheses of this type were not put forward.

Taking into account the statistical analysis of the results, *Plickers* should be recommended as an alternative to computer-based testing, because it (with prior mastery of organization technology and testing) takes less time and results are adequate to students' achievements.

The prospects of further research are seen in the use of *Plickers* as a means of control future teachers' educational achievements, but also as an object of study in order to implement this technology in their future professional activities.

REFERENCES

[1] V. Andrusiv, A. Amelin. Vision of Ukraine – 2030. Ukrainian Institute of the Future. Kyiv : Summit-Book, 2019. 136 p.

[2] O. Semenikhina, M. Drushliak and Yu. Khvorostina, (2019). "Use of GeoGebra cloud service in future math teachers' teaching," Information technologies and learning tools, vol. 73, no 5. pp. 48-66, 2019, DOI: https://doi.org/10.33407/itlt.v73i5.2500.

[3] Plickers. [website]. Available at http://plickers.com, [accessed: 04.06.2020].

[4] J. R. De Thomas, V. López-Fernández, F. Llamas-Salguero, P. Martín-Lobo and S. Pradas, (2016). "Participation and knowledge through Plickers in high school students and its relationship to creativity". In UNESCO-UNIR ICT & Education Latam Congress, pp. 113–123.

[5] T. A. Wood, K. Brown and J. M. Grayson, (2017). "Faculty and student perceptions of Plickers". In ASEE Zone II Conference. Available at http://zone2.asee.org/sessions/program/3/84.pdf, [accessed: 04.06.2020].

[6] S. Wuttiprom, K. Toeddhanya, A. Buachoom and K. Wuttisela, (2017). "Using Plickers cooperate with Peer Instruction to promote students' discussion in Introductory Physics Course". Universal Journal of Educational Research, 5(11), pp. 1955–1961. https://doi.org/10.13189/ujer.2017.051111.

[7] M. G. McCargo, (2017). "The Effects of Plickers As Response Cards On Academic Engagement Behavior In High School Students". Master's Thesis. College of Education and Psychology, University of Southern Mississippi. Available at http://aquila.usm.edu/masters_theses/300, [accessed: 04.06.2020].

[8] A. Gürişik, (2019). "Opinions of high school students about plickers: one of the online formative assessment tools." International Journal of Scientific Research and Innovative Technology, vol. 6, no. 1, January 2019, pp. 11-25.

[9] O. Demirkan, A. Gürişik and O. Akin, (2017). "Teachers' Opinions About "Plickers" One Of The Online Assessment Tools." Educational Research And Practice, pp. 476-486. Available at https://www.researchgate.net/profile/Gokhan_Duman/publication/326902788_Educational_Research_and_Practice/links/5b6b1c2f45851546c9f6d0e7/Educational-Research-and-Practice.pdf#page=484, [accessed: 04.06.2020].

[10] E. A. Michael, I. E. A. Ejeng, M. A. Udit and M. M. Yunus, (2019). "The Use of Plickers for Language Assessment of Reading Comprehension." International Journal of Academic Research in Business and Social Sciences, 9(1), pp. 637–645.

The use of electronic textbooks in the learning process: a statistical analysis

O.M. Udovychenko*, M.M. Ostroha*, A.E. Chernysh*, O.Yu. Kudrina*, Yu.A. Bondarenko and A.V. Kurienkova*
* Makarenko Sumy State Pedagogical University, Sumy, Ukraine
udovich_olga@fizmatsspu.sumy.ua, mariia.ostroha@fizmatsspu.sumy.ua, anna_chernysh@ukr.net, olya.dina@gmail.com, bondarenkosspu@gmail.com, ann.kurenkova.91@gmail.com

***Abstract* - With the active introduction of information technology in the educational sphere, approaches to the textbook as the main means of presenting educational material have changed. Along with printed publications, electronic ones began to be actively used, which in their development have already gone from a simple text document to a complexly structured system that includes various ways of presenting educational material (text, audio, video, graphics). Their use today requires detailed study and reasonable implementation in professional training of specialists. The purpose of the article is to describe a pedagogical experiment on the effectiveness of using an electronic textbook in the training of computer science teachers. To achieve this goal, the methods of mathematical statistics for the quantitative and qualitative analysis of the results of the pedagogical experiment (χ^2 Pearson, Student) were used. The corresponding results for the author's ET "Information Systems" by criteria χ^2 Pearson, Student are given. Calculations for the positions are described: 1) availability (clarity) of the author's ET; 2) evaluation of the results of independent work based on the author's ET.**

Keywords* - *electronic textbook; electronic educational resources; professional training; computer science teacher; learning process.

I. INTRODUCTION

Modern education faces the challenges of the information society, which cause the active introduction of modern technologies in the process of specialists' training. Young people actively use portable, mobile devices, computer equipment, and therefore resources, which in their bases provide for such use, become popular. This led to the emergence and development of electronic educational resources (EER), which are now understood as educational, scientific, information, additional materials and tools developed in electronic form and presented on media of any type or placed in computer networks, and which are reproduced using electronic digital technical tools and are necessary for the effective organization of the educational process, in terms of its content with quality teaching materials [7].

Technologies of introduction and use of EER in the educational process are relevant today, and therefore the issues of training specialists on the basis of their use are extremely relevant [1; 4: 6]. Electronic textbook (ET) requires special attention as a popular electronic educational resource in this case [5; 13].

Various aspects of the problem of training computer science teachers to the professional activity were considered in the works of L. Breskina (professional training based on network information technologies) [15], V. Demianenko (teaching hardware and system software) [16], O. Zhmud (formation of specific competence in computer architecture and configuration of computer systems) [17], M. Zolochevska (preparation for the research methods use) [18], S. Ovcharov (an individual and differentiated approach in professional training) [19], V. Oleksiuk (methodological bases of application of educational network complexes) [20]. V. Bykov, V. Lapinskyi, A. Gurzhiia, V. Vember devoted their works to the issue of implementation of EER in the educational process, and the works of O. Balykina, N. Kononets, V. Hasov, O. Grytsenchuk, O. Gurkova, I. Pustovalov, T. Yakovenko, V. Yasynskyi and others are devoted to e-books in particular.

At the same time, we have found that the training of computer science teachers using EER technologies is not enough considered in the Ukrainian educational space. ET as a means of modern education is described in the system of training of social pedagogues, teachers of history [2], the Ukrainian language [8], agro-technicians [14] etc., and the question remains open for bachelors, computer science teachers training, on the basis of EER, including ET, as electronic educational resource, which provides: program compliance; consistency, scientific character and completeness of educational material presentation; accounting for psychological and pedagogical characteristics of the subjects of training; simultaneous presentation of educational material in various forms; the level of development of information technologies in the ways of presentation of educational material and organization of work with the resource; continuity and completeness of the educational cycle; individualization and differentiation of training [12].

Based on modern trends in the field of education and understanding the need and demand for ET, we have implemented a project to create such type of electronic product, which has its own pdf-version, as well as presented in the form of a complex structured educational resource and is positioned by us as a modern ET with multimedia content, which includes, in addition to theoretical material, visual support (in the form of diagrams, tables, animation, video), glossary and test module for self-testing.

The aim of the article is to describe the pedagogical experiment on the effectiveness of the electronic textbook in the training of computer science teachers.

Theoretical training of bachelors in the specialty 014.09 Secondary education (Computer Science) provides for the study of a number of Computer sciences disciplines. As the analysis of methodical systems of training of specialists, future computer science teachers has showed that various forms, methods and means of training are applied for their studying. Along with the traditional we use an electronic textbook "Information systems". Its structure is indicated in works [9; 10; 11].

To achieve this goal, the following methods were used: *theoretical*: analysis of special literature on the introduction of electronic educational resources; generalization and systematization of domestic and foreign experience in the use of electronic textbooks in the training of computer science teachers, personal pedagogical experience in the use of electronic educational resources in the educational process; *methods of mathematical statistics* for quantitative and qualitative analysis of the results of the pedagogical experiment.

II. PRESENTATION OF THE MAIN MATERIAL

We have conducted a study of the feasibility of using the author's ET in independent work on the basis of the nonparametric method of χ^2 Pearson.

The results of training in two groups were investigated:

Group I – the study of the course "Information systems" was carried out on the basis of EER, in particular, the students were offered lectures-presentations, author's ET "Information systems" with built-in dictionary, graphics, text content, fragments of videos, tests for self-testing of residual knowledge; provided a list of additional electronic publications and videos, links to open educational resources that are related to this course.

Group II – the study of the course "Information systems" was supposed to use lecture notes written by students at the teacher, the printed version of the textbook "Information systems", graphic designs in the form of tables and diagrams, as well as additional literature available in the library.

The composition of the groups contained 109 and 94 respondents among students and 12 and 15 respondents-teachers, respectively.

Evaluation of the results of the experimental training took place in several positions: the level of performance of students modular and semester control measures; answers of students and teachers to oral control questions; conversations with teachers and students regarding the accompanying educational materials.

Consider the method of comparing the answers of teachers to the question: "Is a textbook accessible (understandable) as a whole for self-mastery, that is, does the textbook help to learn the training material during the self-study of some topics?"

The inclusion of such a question in the questionnaire is explained by the fact that in order to identify the availability of educational materials, in particular ET, a specific list of topics for self-study by students of both groups was determined. Based on the comparison of the opinions of the experimental group teachers, we tested the hypothesis of no differences in the availability of the ET and its printed version.

The opinion of teachers on the quality of educational materials as accessibility was measured on a scale of names, which has two categories: yes and no. Both teacher samples were random and independent. Therefore, all conditions for the application of the χ^2 criterion are fulfilled in the experimental conditions. Since the scale of names has only two categories, we use a variant of the two-sided criterion, which is applied for cases where the results are summarized in a table of type 2x2.

The answers of the teachers (12 of them used ET) were divided in the following way (table 1).

TABLE 1. DISTRIBUTION OF TEACHERS' ANSWERS

	Yes	*No*	
Sample №1	$O_{11} = 7$	$O_{12} = 5$	12
Sample №2	$O_{21} = 5$	$O_{22} = 10$	15
	12	15	

Denote: p_1 – the probability that the teacher of the group №1 considers the ET available, p_2 – the probability that the teacher of the group №2 considers the printed textbook available.

Based on the data from the previous table, we check the null hypothesis:

H_0: $p_1 = p_2$, that is both versions of the textbooks are available.

Then the alternative hypothesis will be:

H_a: $p_1 \neq p_2$, that is some of the tutorials are less accessible.

All values of absolute frequencies in the table are not less than 5, but all values are less than 10. Therefore, according to the terms of calculation of the criterion we have [3]:

$$T_{\text{exp.}} = \frac{N\left(\left|O_{11}O_{22} - O_{12}O_{21}\right| - \frac{N}{2}\right)^2}{n_1 n_2 (O_{11} + O_{21})(O_{12} + O_{22})} =$$

$$= \frac{27\left(\left|7 \cdot 10 - 5 \cdot 5\right| - \frac{27}{2}\right)^2}{12 \cdot 15 \cdot 12 \cdot 15} = 0.83$$

With the conditions of application of the bilateral criterion χ^2 (degrees of freedom – 1, the significance level of 0.05) we find $T_{crit.}$: $T_{crit.} = 3.84$. Because $T_{exp.} < T_{crit.}$, then accept the null hypothesis about the availability of both textbooks to study the course "Information systems". In other words, the results of the questionnaire of teachers of the two samples do not provide sufficient grounds for rejecting the assumption of the same availability of

educational materials in the self-development of the course topics.

The second part of the experiment concerned the method of comparing the results of control works (unit testing and semester testing).

Samples of 50 papers ($n_1 = 50$, $n_2 = 50$) were selected with random method. According to the evaluation criteria, the work was divided into four categories: excellent, good, satisfactory, unsatisfactory.

Let's build a hypothesis H_0: an electronic textbook contributes to a better assimilation of knowledge (that is, students of the first sample on average will receive higher grades compared to students of the second sample, where only printed educational materials were used).

The work samples were random and independent, with the assimilation rate having a continuous distribution and measured on a scale of order of four categories. In this case, all the conditions for the application of the Wilcoxon-Mann-Whitney test, which allows checking the formulated assumption, are fulfilled. However, due to the small number of categories of the scale, a significant part of the experimental data is a chain of the same values, so again we use the two-sided criterion of χ^2 for table 2×4 (table 2).

TABLE 2. DISTRIBUTION OF STUDENTS' ANSWERS

	Unsatisfactory	*Satisfactory*	*Good*	*Excellent*	
Sample №1	$O_{11} = 4$	$O_{12} = 15$	$O_{13} = 20$	$O_{14} = 11$	50
Sample №2	$O_{21} = 8$	$O_{22} = 20$	$O_{23} = 20$	$O_{24} = 2$	50
	12	35	40	13	

In the table, O_{ij} means the number of works of the i-th sample, who got the mark j.

Denote p_{1j} – for the probability of performing the work by the student of the first sample on the mark j; p_{2j} – the probability of performing the work by the student of the second group on the mark j.

Then check the hypothesis

H_0: $p_{1j} = p_{2j}$ for all four categories

(that is $p_{11} = p_{21}, p_{12} = p_{22}, p_{13} = p_{23}, p_{14} = p_{24}$)

alternative:

H_a: $p_{1j} \neq p_{2j}$ at least for one of four categories.

To calculate the statistics, we use the formula [2]

$$T = \frac{1}{n_1 n_2} \sum_{j=1}^{4} \frac{\left(n_1 O_{2j} - n_2 O_{1j}\right)^2}{O_{1j} + O_{2j}},$$

$$T = \frac{1}{50 \cdot 50}\left(\frac{\left(50 \cdot 8 - 50 \cdot 4\right)^2}{4+8} + \frac{\left(50 \cdot 20 - 50 \cdot 15\right)^2}{15+20} + \right.$$

$$\left. + \frac{\left(50 \cdot 20 - 50 \cdot 20\right)^2}{20+20} + \frac{\left(50 \cdot 2 - 50 \cdot 11\right)^2}{11+2}\right) = 8.28$$

For the level of significance $\alpha = 0{,}05$ and the number of degrees of freedom $\upsilon = 4 - 1 = 3$ find the critical value of statistics: $T_{crit.} = 7.815$.

We got a situation, where $T_{crit.} < T_{exp.}$ (7.815 < 8.28), therefore, we must accept the alternative hypothesis that the sample averages differ significantly and this cannot be explained by random reasons.

In other words, the results of self-study of the special course "Information systems" on the basis of the author's ET give grounds not to reject the alternative hypothesis that students who were studying using an electronic textbook, learned the course better (with a high average score).

The results are divided into levels in the following way (table 3).

TABLE 3. DISTRIBUTION OF STUDENTS BY LEVELS

Levels / *Groups*	*Actual (0–60)*	*Operational (61–74)*	*Analytical-synthetic (75–82)*	*Creative (83–100)*
Sample №1	$O_{11} = 4$	$O_{12} = 15$	$O_{13} = 20$	$O_{14} = 11$
Sample №2	$O_{21} = 8$	$O_{22} = 20$	$O_{23} = 20$	$O_{24} = 2$
	12	35	40	13

A significant difference between the averages is confirmed by the Student's criterion at the same level of significance, which is confirmed by calculations in Excel (table 4).

TABLE 4. ESTIMATION OF AVERAGES FOR STATISTICAL ANALYSIS OF EFFICIENCY OF ET USE (AFTER EXPERIMENT)

	Control group	*Experimental group*
Average	76.14	79.76
Observation	50	50
Hypothetical average difference	0	
t-statics	–2,123923442	
t critical double-sided	1,984984312	

III. DISCUSSION

The introduction of computer-aided learning at universities is in line with the innovative processes that are taking place in Ukraine to modernize the education system. The current computer market places particular demands on the professionals' training and offers, at the same time, electronic educational resources that are currently actively, but often not systematically, used by students.

Orientation of universities to a high-level automation of the educational process and the use of electronic educational resources in teaching requires scientifically substantiated support and availability of appropriate educational and methodological support, which actualizes the improvement of future computer science teachers' training towards active use of electronic educational resources.

We consider the ET as a means that is more comfortable for young people to use. Our ET is not only filled with text, but is also accompanied by a large number of diagrams, tables and figures. Such visual support helps to better memorize and understand the learning material. Its use, on the one hand, is a means of teaching at the university, and on the other, it demonstrates to future teachers a means of their professional activity.

The described ET received praise from teachers who were involved in the pedagogical experiment. Thanks to their advice, in particular, tests of various topics were added to it, and the topic "Office software package" was excluded.

The analysis of the results of the pedagogical experiment confirms the opinion that it is advisable to involve electronic rather than paper textbooks in the educational process. At the same time, the development of an ET requires the work of a whole team (IT specialists, psychologists, teachers) and is time consuming.

Currently, a similar experiment is underway to implement an electronic computer-based textbook based on the use of QR codes. Preliminary results give a positive assessment of academic achievement compared to traditional learning.

IV. CONLUSION

As the pedagogical experiment shows, training of computer science teachers based on EER, in particular, ET, is effective. Students often turn not to print, but to electronic educational resources while preparing for classes, which provide consistency and thoroughness of knowledge, the development of visual memory and visual thinking, and therefore the author's electronic textbook is an effective electronic resource in the training of computer science teachers.

The study does not claim to be a definitive solution to the problems of using the ET in the educational process.

Further scientific research is directed to the solution of the problem of organizing students' independent work on the basis of the ET.

REFERENCES

[1] Alan Brinkley, Betty Dessants, Michael Flamm, Cynthia Fleming, Charles Forcey, and Eric Rothschild, "Using Electronic Resources for Teaching: an excerpt from The Chicago Handbook for Teachers: A Practical Guide to the College Classroom", 1999, 196 p.

[2] E.N. Balykina, "Electronic educational edition on historical disciplines: definition, structure, model". State and development of methodological research in historical science of the Republic of Belarus and the Russian Federation: Coll. scientific Art. Grodno, pp. 119–127, 2008.

[3] M.I. Grabar and K.A. Krasnyanskaya, "The use of mathematical statistics in educational research. Non-parametric methods". M .: Pedagogy, 1977.

[4] Kelefa Mwantimwa, Emmanuel Elia, and Esther Ndenje-Sichalwe, "Utilisation of E-Resources to Support Teaching and Research in Higher Learning Institutions, Tanzania". University of Dar es Salaam Library Journal, vol. 12, No 2, pp. 98-123, 2017.

[5] L.M. Gorghiu, G. Gorghiu, M. Bîzoi, and A.M. Suduc, "The electronic book – a modern instrument used in teachers' training process". Procedia Computer Science, vol. 3, pp. 563-567, 2011.

[6] R. Singh, and S. Jindal, "Promoting the Use of E-Resources in Teaching, Learning and Research: A Case Study of DULS". 7 th International CALIBER-2009, Pondicherry University, Puducherry, February 25-27. pp. 430-438, 2009.

[7] Regulations on electronic educational resources. Retrieved from http://zakon4.rada.gov.ua/laws/show/z1695-12#n13.

[8] L.V. Ruskulis, "An electronic textbook as an effective information technology in the process of preparing the future teacher of the Ukrainian language". Modern information technologies and innovative teaching methods in the training of specialists: methodology, theory, experience, problems, 38, pp. 431–434, 2014.

[9] O. Semenikhina, V. Shamonya, O. Udovychenko, A. Yurchenko, "Electronic Textbook in the Context of Educational Trends and Modern Internet Technologies". Zhurnal ministerstva narodnogo prosveshcheniya, 2(2), pp. 99–107, 2014.

[10] O.V. Semenykhina, O.M. Udovychenko, A.O. Yurchenko, "Electronic textbook "Information Systems" as a demanded educational resource in the practice of modern higher educational institutions". Informatics and information technologies in educational institutions, 3(51), pp. 15–22, 2014.

[11] O.M. Udovychenko, A.O. Yurchenko, "From the experience of creating an electronic textbook as a means of supporting the learning process". Physical and Mathematical Education, 1(6), pp. 210–214, 2014.

[12] O.M. Udovychenko, "Electronic textbook in the system of information teaching aids". Quality Strategy in Industry and Education: IX International Conference, Varna, pp. 533-535, 2013.

[13] V. Liesaputra, and I.H. Witten, "Realistic electronic books". International Journal of Human-Computer Studies, vol. 70, iss. 9, pp. 588-610, 2012.

[14] A.B. Zueva, "Electronic textbook of agrarian direction in the training of agricultural technicians". Scientific and methodological provision of vocational education and training: materials of the Report scientific-practical conference, 2, pp. 114-116, 2014.

[15] Breskina L.V. The professional training of the future computer science teacher on the basis of modern network information technologies. Candidate's thesis. Kyiv, 2003.

[16] Demyanenko V.M. Methods of training of future teachers of informatics of hardware and system software science. Candidate's thesis. Kyiv, 2003.

[17] Zhmud O.V. Formation of subject competence of computer architecture and configuration of computer systems in the future teachers of Computer Science. Candidate's thesis. Kyiv, 2015.

[18] Zolochevska M.V. Methodological training of the future teacher of Computer Science for using research methods in teaching. Candidate's thesis. Kyiv, 2011.

[19] Ovcharov S. M. Individually-differentiated approach in vocational training of future teachers of computer science. Candidate's thesis. Zhytomyr, 2005.

[20] Oleksyuk V.P. Methodological bases of educational network complexes application in the process of future computer science teachers training. Candidate's thesis. Kyiv, 2006.

Influence of Moodle and MS Teams on teaching-learning-studying (TLS) processes

Marjan Krašna*, Igor Pesek**
*University of Maribor, Faculty of Arts, Department of Pedagogy, Maribor, Slovenia
**University of Maribor, Faculty of Natural Science and Mathematics, Department of Physics, Maribor, Slovenia
marjan.krasna@um.si, igor.pesek@um.si

Abstract – **In two decades, the Moodle established itself as favorable LMS. It is used on all levels of education and many lecturers have gained skills to use it effectively. Moodle becomes mature and many lecturers have found it to be their first choice Learning Management System (LMS). Educational systems around the world have always problems with money and the Moodle is free; it has good multi-language support; and it implements many functionalities covering most lecturers' needs. But in the recent years some new alternatives emerge. They do not want to copy the Moodle and this makes them interesting enough to try them in practice. In recent time the Microsoft Teams incorporating the Microsoft OneNote becomes increasingly popular and gained users in many fields of work. Both systems were tried in different courses. It is clear that the Moodle is the winner in traditional educational tasks (assessments, peer to peer assessment, quizzes, categorization, extendibility). The MS Teams excels in different types of communications (chat, videoconferencing, lecturing, screen shareing), and is superior in general content creation. There is no clear winner in the assignment, mathematics (symbol), and file distribution.**

Keywords – E-learning, learning management system, Moodle, Microsoft Teams, comparison

I INTRODUCTION

E-learning paradigm has always depended on learning materials, online accessibility, student engagement and communication. Our interest in e-learning has been very long, beginning with the e-learning materials in the course Application of computer in education – the first multilanguage e-book. At that time tests and development of the e-lecture room [1] was done too. It proved very valuable addition to our e-learning materials, but the pilot project was later dismissed because of lack of funds.

Few years later; the project "Partnership of Faculties and Schools" (2004-2005) resulted in research of publicly available LMSs, where one of them (Moodle [2]) was newly available. The project was successful and continued in follow up project "Partnership of Faculties and Schools 2" (2006-2007) where the Moodle established itself as an LMS of choice for primary and secondary schools. During that time, guidelines and user manuals on how to use Moodle was prepared and different courses were prepared and used in practice.

The Moodle grew in functionality and popularity and years later our university also decided to centralize e-learning on the central Moodle's platform. The Moodle proved to be very useful during last two decades and it is now used on different usage levels depending on teachers' skills. Teachers can use them from document repository to advanced blended learning courses. Researchers report different use scenarios. Thomas [3] reports that Moodle use at their University is limited primarily to two activities – quizzes and assignments; on the other hand there is the report that at the secondary level in Catalonia teachers primarily use activities [4] and the reason for them is *because they provide new educational scenarios*.

Instant messaging has a long history starting with IRC (Internet Relay Chat). In recent years instant messaging gained momentum when teams needed some tool that will help them to be more organized. New platforms emerged with the Slack and the Discord as their frontrunners. If the Slack was used mostly within software companies, then the Discord was used by players in gaming industry. The other users' groups quickly found their potential; scientists use the Slack [5] and lecturers use them in undergraduate classes [6].

The Microsoft introduced the Teams, which works on the same principles as before mentioned platforms, with the one big difference. Microsoft integrated the Teams in the Office365 into seamless user experience. Educators quickly saw its potential to integrate it into their educational process. [7] [8] [9]

In the recent years, observations have shown that teachers sometimes struggled with the Moodle, especially if they wanted something more (more fluid communication with the students). If teachers wanted to correct students work and send them feedback, it is possible in the Moodle but it is not very user friendly. The Moodle version 3.5 and later can convert all regular office files into the Portable Document Form (PDF). A teacher can annotate converted PDF page by page and send the feedback back to the students. The conversion process is not simple. Currently there are two ways to convert office file into PDF: (1) use the Google's service – not an option in EU because of the General Data Protection Regulation (GDPR) law that forbids to send personal data outside EU, (2) use the Openoffice or Libreoffice conversion tool which requires significant server CPU usage, especially if used by many teachers at the same time. Additional drawback of such conversion to the PDF is that each user file (assignment) is now in two or more versions (original, pdf, and annotated pdf) that proves to be huge burden on the Moodle's servers' disk space.

The communications students are using today are very different what they were using 15 years ago. Students consider forums and e-mails archaic and they do not used them very frequently and they use them only if they have to. They use mobile devices for communication, collaboration and also for learning [10]. It is interesting that student blur the lines between social networking tools (also chat platforms) and mobile devices.

Therefore, the decision to try the MS Teams in our classrooms and compare it to the Moodle has matured. In the academic year 2019/2020 have started to use the MS Teams for six courses in winter semester (the article was prepared before COVID-19 pandemic). In these courses the Moodle and the MS Teams were equally used. At the end of the semester students were asked about their opinions and attitudes about both platforms.

II METHODOLOGY

Comparison of most common tasks that teachers or lecturers do in the pedagogical process was made. The tasks were selected based on the findings of the project PICT.UM where different didactic specialists from different educational area worked for two years to prepare guidelines to use ICT in education [11].

Students from the Faculty of Natural Science and Mathematics used both system in the course of Information and Communication Technology (ICT) and student of general pedagogy from the Faculty of Arts used both system in courses of Digital security and Information support of didactics strategies. The MS Teams was used exclusively in the courses of Multimedia for Physics and E-learning in physics (Faculty of Natural Sciences and Mathematics); and Applied computer science for translation studies (Faculty of Arts).

For the comparison of the educational tasks between two systems, the text analysis of students' reports (reflection of their thoughts and learning experience at the end of the course) were used and the teachers' experiences using both systems.

The online survey was used to gather data from students at the end of the courses. Students were given 8 questions where they could decide which system they use are preferred to them. The 5 scale response was used. The grade level was 1 – Moodle (strong); 2 – Moodle (weak); 3 – equal; 4 – MS Teams (weak); and 5 – MS Teams (strong).

Because the sample is small the descriptive statistics and table data presentations are used in this article.

III MOODLE VS. MS TEAMS COMPARISON

The comparison of both platforms are done from teachers' point of view combined with students' reflection on using both products; whereas students' views on the usability will be presented in separate section. Comparisons is made in tables based on tasks and comments for the Moodle and the MS Teams are provided based on our experience and students feedback. If the winner could be declared it is labeled. No label of the winner means a tie. The process of declaring the winner or a tie is done by the teachers that are proficient users of both systems. The discussion group of teachers read the students' reflections and consider their own experiences with both systems and declare a verdict.

Both systems have mobile client for Android and iOS but students rarely use the Mobile Moodle client in our classrooms. Most of our students have installed the MS Teams mobile client though.

A Moodle

The Moodle is free online Learning Management System (LMS), still actively being developed, with current version 3.8. Its main use is in the field of education and can be utilized and used in different types of e-learning. To meet different educational demands, it can be extended with various plugins. Teachers can share content with students, give them assignments and different tasks, prepare questionnaires, grade their work, give feedback and communicate. Whole environment is HTML based and operates only online.

B MS Teams

The MS Teams [11] is the Microsoft product and is a unified communication and collaboration platform, that provides persistent chat, a file storage and an integration with MS Office 365 suite. It integrates also with the other MS Office applications, most notably with the MS OneNote, which is already used in classrooms. Environment is available both online and as the application. It works also offline and synchronizes across all platforms when online. For a data storage it uses the SharePoint server or the OneDrive. The OneDrive functionality is available in one-to-one communication and allows single document sharing where both parties can edit it simultaneously (and have voice communication too).

The MS Teams has mobile client that works also as communication client (like Skype, Viber …) but with the advantage of not sharing a phone number. Calls can be made and answered in the device that runs a program (either pc, tablet, or any phone user may have). Multiple clients and be used at the same time. Therefore, using a computer(s) and a mobile phone at the same time is seamless.

C Content manipulation

One of the most common tasks that teacher does is to prepare and publish various educational content.

Task 1: Write and publish a page with some lists and images.

Moodle	Teams (winner)
Moodle has a rich text editor, can add pictures in text. Using HTML 5 even multimedia elements can be inserted but it requires access to the web server and writing HTML code which is generally not known among our teachers.	Since Teams integrates with OneNote there is also rich text editor, any file can be added by just "drag and drop". High learning curve for anyone who uses MS Office applications.

Task 2: Publish files of different types (ppt, pdf) in the online classroom

Moodle	Teams
Files can be uploaded separately or in bulk. On the page, links can be added to a file or to a folder (archaic mode).	Files can be added to the common space. It can be organized as folders. Some files can be read-only. There is also possible to link it to the cloud storage (e.g. the OneDrive). Files can also be uploaded to the notebook in the OneNote.

Task 3: Writing special symbols (e.g. for the mathematics)

Moodle	Teams
Symbols can be written using the Latex notation. It needs to be enabled across the site.	No direct input with Latex notation, but handwritten math is possible using tablets. Additionally, math problems can be solved (with all the necessary steps).

D Communication

Main function of online learning systems is not to post content but engagement of students, which can be achieved with communication between a teacher and a student and different activities.

Task 1: Teacher communicating with all students on a topic (synchronous and asynchronous)

Moodle	Teams
Group communication in the Moodle uses forums. There may be multiple forums in each classroom. Settings of the forum allow users to subscribe, attach files with limited sizes. Moodle uses primarily asynchronous communication. Chats are available but rarely used. Moodle sends e-mail notification for all forum posts if user is subscribed to the forum.	Teams uses Posts for group communication as primary communication channel. Multiple channels can be added with restrictions. In the channels arbitrary files can be uploaded; the file size is generally no restriction for the documents. Communication can be synchronous but can be used also as asynchronous.

The MS Teams applied the popular responses know from other social networks that allow students to give instant feedback on the note without need to reply, which is more convenient for them.

Task 2: Communication in a team (teacher is also present)

Moodle	Teams (winner)
Usually students use other means of communication for group talk. In the Moodle there is no dedicated tool for such tasks, it can be simulated with forums.	In the MS Teams it is very easy to make a new private group and invite team members to talk about the assignment.

Task 3: Individual chat between teacher – student

Moodle	Teams (winner)
Moodle uses messages for individual communication between student and teacher. It is well known type of communication which also sends e-mail notifications. The problem may arise when teacher replies with the e-mail instead from within the Moodle. Consequently, it breaks the topic because two communication channels are used instead of one.	Individual communication in Teams is done using chat. Teams can send notifications to the e-mail, but it is required that whole communication is done in chat. Because communication is stored in one place it gives great overview of chat history. Chat allows adding a file that both can view simultaneously if file format is in Office 365 compatible list. Chat support video and audio calls and screen sharing. This is very valuable for solving problems on a student's computer.

Task 4: Video or audio calls for the online tutorial

Moodle	Teams (winner)
It is not a part of the Moodle to have audio or video chat, it can be done via paid plugins. Third party applications are used instead (e.g. Big Blue Button).	Audio and video calls works as good as in the Skype and allows also desktop sharing. Conference calls with multiple users is supported. Screen sharing functionality not available in the online version of the MS Teams. Only 9 video streams are visible at the same time.

E Activities

This section is focused on the different activities that engage students.

Task 1: Ask a question, all students need to answer

Moodle (winner)	Teams
It can be done using Choice activity. There are different options how to execute this task (student do not see other responses, it can be anonymous, etc.). Alternatively, this could also be done using Forum.	A quick poll can be made in the chat, but cannot enforce, that all students answer. It has fewer options how to execute a poll.

Task 2: Prepare a quiz with different types of questions (true/false, multichoice, etc.)

Moodle (winner)	Teams
Moodle has built in support with different types of questions and can be extended with plethora of plugins, some of them are very powerful (e.g. STACK for mathematics). Questions can be stored in Question Bank and reused in different quizzes.	Teams integrated MS Forms that enables teachers to prepare quizzes. Questions are stored in forms and cannot be reused.

Task 3: Assessment with grades of student work

Moodle (winner)	Teams
Moodle has built in very powerful grading system, which can be used in different scenarios. Sometimes it is even too complex for a not tech savvy teacher to use it. Almost every activity in Moodle can be graded, e.g. in current version of Moodle (v3.8) it is possible to grade also student responses in forums.	Assignments can be graded and are collected in gradebook. It is not as complex and powerful as Moodle grading system. For grading assignments that are Office files the MS Teams is superior to Moodle.

Task 4: Peer to peer assessment

Moodle (winner)	Teams
Workshop activity is built in Moodle and enables teachers to implement peer assessment. Workflow is not very intuitive.	In Teams it is not possible to implement peer assessment. It can be simulated but it is not easy and needs some workarounds.

Students upload their work as files, text, etc. It can be done with due date. Work can be graded. Feedback is also possible.

Task 5: Assignments

Moodle	Teams
Assignments are a core functionality in Moodle. It has different options to enable teachers to fine tune each assignment. Assignments can have due date, can be graded and teacher can give feedback to the individual student. Not easy to upload images from phone (e.g. photos of hand written documents).	Teacher can prepare assignments and also set due date. Assignments can be graded, and teacher can give individual feedback to the students. With mobile Teams students can take photos of their work and upload it to the Teams.

F System administration

Sometimes teacher needs some modification of original LMS. This section explores two tasks that are usual for teachers but they still need to be done in the institution.

Task 1: Add a plugin that extends functionality of a tool.

Moodle (winner)	Teams
Moodle community has a large repository of plugins that can extend functionality of Moodle. Plugin install is possible only if teacher is also administrator.	Teams has also plugins although not so large as Moodle. Plugins can also work as bots in chat and can extend usability of chats.

Task 2: Server administration and upgrades

Moodle	Teams (winner)
Moodle is not easy to administer. Usually schools use central installation but are consequently to use Moodle as is. Such approach is viable from financial point of view but disables some educational opportunities in specific subjects (e.g. Math teacher needs different specific tools compared to language teacher).	Teams are part of Office 365 and cannot be installed on own servers. It is always managed centrally.

IV STUDENT SURVEYS

Students taking the survey had earlier experiences with Moodle, because in Slovenia the Moodle is the de facto LMS in high schools and universities. The MS Teams and MS OneNote were used in first semester of academic year 2019/2020. The MS Teams was used exclusively on three courses and three courses both system were used at the same time. At the end of semester, survey was send to all participating students asking them about the various aspects of their attitudes regarding use of both LMS's.

Between January 16 and January 23, 22 out of possible 46 students from academic year 2019-2020 provided valid answers for the survey questionnaire. Questionnaire had eight (8) questions, which could be graded with one of five (5) choices. The grade level was 1 – Moodle (strong); 2- Moodle (weak); 3 – equal; 4 – MS Teams (weak); and 5 – MS Teams (strong).

Questions were:

- Search for the data inside the classroom
- Communication with the teachers
- Notifications
- Assignment uploading
- Preparing notes
- Use on mobile device
- Ease of user
- Which of the system you like the most?

Table 1: The responses of students by topics.

Topic	Moodle (strong)	Moodle (weak)	Equal	MS Teams (weak)	MS Teams (strong)	Total
Search for the data inside the classroom	6	3	9	2	2	22
Communication with the teachers	0	0	3	1	18	22
Notifications	0	0	3	2	17	22
Assignment uploading	9	2	6	2	3	22
Preparing notes	1	2	5	6	8	22
Use on mobile device	3	2	3	1	13	22
Ease of user	5	0	9	2	6	22
Which of the system you like the most	5	2	4	3	8	22

Table 2: Sorted answers of the student responses.

Topic	mean grade
Communication with the teachers	4.68
Notifications	4.64
Use on mobile device	3.86
Preparing notes	3.82
Which of the system you like the most	3.32
Ease of user	3.18
Search for the data inside the classroom	2.59
Assignment uploading	2.45

The data have confirmed our suspicions. The Moodle is still valued for its better structure of the data inside a classroom and assignment uploading. But the Moodle clearly loses to the MS Teams in the area of communications and notification. This is probably because the mobile phone client for the MS Teams is much better than the mobile client for the Moodle and looks like the social network that is the favorite type of communication for younger generations of students.

V DISCUSSION AND CONCLUSION

The results are conclusive – the MS Teams cannot replace the Moodle yet. The Moodle is superior in the assessment and in the clear structure of the course. But Moodle's archaic design has lost to the MS Teams in the communication's topics.

The Moodle is still standing strong. The infrastructure of many educational institutions is based on using Moodle and switch to another LMS is far from simple. Sometimes it was difficult to assess the winner in some topic. Especially in the assignments – Moodle's traditional approach is favored by the students who express that assignment in the MS Teams could be easily overlooked. From the teacher's point of view, where assignment needs the feedback to the student, the MS Teams is clear winner because it requires far less operations and the assessment process is much faster. Therefore, no winner was declared under those circumstances.

One interesting aspect is the question about preparing personal notes in both products. At least three students have given us an answer that the Moodle is better in this topic which cannot be true. On the other hand, students in one course need to prepare report about the both product and in their free text reports it was evident that they are still reluctant to make digital notes. This could change in a few years. In the MS Teams the teacher is able to see students' notes about the course and get much better overview of the transfer of the knowledge than he could ever get from asking questions during the lectures. It is true that this would require much more lecturer's work to check all students' notes but if an exemplary student is checked then all other students would get corrected notes through students' note dissemination. Error detection is valuable feedback for teacher to know which topics needs special attention because it is prone to students' wrong understanding.

At the end we have learned the most important thing that both products can be used simultaneously and use the one that is better suited for the specific task.

VI REFERENCES

[1] I. Gerlič, "Didactic model of internet application in education," in *WebNet World Conference 99*, Honolulu, Hawaii, 1999.

[2] [Online]. Available: https://moodle.org/.

[3] D. Thomas, "Moodle Usage Case Study at a University," 2018.

[4] A. Badia, D. Martin and M. Gomez, "Teachers' Perceptions of the Use of Moodle Activities and Their Learning Impact in Secondary Education," *Technology, Knowledge and Learning volume,* pp. 483-499, 2019.

[5] P. Jeffrey, "How scientists use Slack," 29 December 2016. [Online]. Available: https://www.nature.com/news/how-scientists-use-slack-1.21228.

[6] S. Hussain, A. Mills and A. K. Sanders, "Use of Slack in the Undergraduate Classroom as a Means of Fostering Student/Faculty Engagement," *Qatar Foundation Annual Research Conference Proceedings,* 2018.

[7] P. Davidson, E. Long, A. Molnar, T. M. Chui and C. Y. Ting, "MS Teams and Google classroom: Preliminary qualitative comparisons & user feedback," in *5th Pre-University Sunway Academic Conference 2018*, Selangor Darul Ehsan, Malaysia, 2018.

[8] L. Martin and D. Tapp, "Teaching with Teams: An introduction to teaching an undergraduate law module using Microsoft Teams," *Innovative Practice in Higher Education ,* vol. 3, no. 3, pp. 58-66, 2019.

[9] L. Riding, "From consumers to creators: developing learner autonomy using Microsoft Teams," Technology Enabled Learning & Teaching at UCLan, 7 1 2020. [Online]. Available: https://teltblog.uclan.ac.uk/2020/01/07/from-consumers-to-creators-developing-learner-autonomy-using-microsoft-teams/. [Accessed 13 2 2020].

[10] J. Gikas and M. Grant, "Mobile computing devices in higher education: Student perspectives on learning with cellphones, smartphones & social media," *The Internet and Higher Education,* vol. 19, pp. 18-26, 2013.

[11] University of Maribor, Strokovne podlage za didaktično uporabo IKT, Maribor: University of Maribor, 2018.

[12] [Online]. Available: https://products.office.com/en-US/microsoft-teams/group-chat-software.

The impact of distance learning on student success for electrical engineering professional courses

Mladen Sokele, Trpimir Alajbeg, Frane Brkić
Department of Electrical Engineering
Zagreb University of Applied Sciences Zagreb, Croatia
mladen.sokele@tvz.hr, trpimir.alajbeg@tvz.hr, frane.brkic@tvz.hr

***Abstract* - Electrical Engineering undergraduate study at Zagreb University of Applied Sciences for professional courses is performed in blended learning environment combining the advantages of both face-to-face teaching and computer-assisted learning. Usually, teaching process of a professional course consists of lectures, exercises and laboratory exercises. The data sets related to attendance and overall exam success were analyzed to validate the impact of blending learning components on the learning success with the focus on distance learning one. Distance learning proves to be a useful method as a complement to face-to-face teaching. In addition, it can serve as the main learning method for the late exams preparation, but with lower achievement success.**

Keywords - Exam data analysis; Knowledge acquisition; Blended learning; Asynchronous online learning; Face-to-face teaching; LMS; Share of self-learning in average success

I. INTRODUCTION

Professional courses for undergraduate study of electrical engineering at Zagreb University of Applied Sciences are performed in blended learning environment, combining the advantages of face-to-face teaching and computer-assisted learning (i.e. asynchronous online learning) [1]. Face-to-face (F2F) teaching is dominant for lectures and exercises. Laboratory exercises are held in a standard laboratory environment where students work mostly independently with assistance of lecturer. All above mentioned teaching types are supported by different levels of learning management systems (LMS) for the purpose of group and individual student preparation, knowledge assessment and automatic grading during lecture period [2], [3]. LMS teaching contents can be reached remotely, so asynchronous online learning is enabled for each student's preparation during all exam terms [4].

In previous research, it is shown that despite of all the trends in the development and implementation of technology mediated learning, the role of lecturer is still very important. Recent paper presents the positive correlation of students' success on early exam terms and their attendance of F2F lectures and exercises for fundamental telecommunication professional courses [5]. Data modeling of students' success on the exam and their attendance was performed by piecewise linear regression ("hockey stick" shaped model). High values of determination coefficient R^2 (0.82 – 0.92) are obtained for all analyzed courses, confirming strong relationship between student attendance and their success at exam terms within 2 and/or 4 weeks after lecture ends.

For all analyzed courses, "hockey stick" shaped model has the first linear segment with higher slope and the second segment with lower slope. According to hypothesis presented in [5] the first segment represents knowledge adoption of basic terminology, systematization and concepts related to course curriculum where F2F teaching ensures systematization into meaningful knowledge entities and their further clarification. The second segment lower slope indicates that students, after acquiring basic terminology and essential knowledge may use all the benefits of technology mediated individual (self) learning.

Abovementioned previous research excludes late exam terms results since the focus was on F2F teaching impacts. As part of the internal project: *Research on the importance of blended learning components for student exam success* continuation, data related to the late exam terms results are collected, processed and analyzed. Results and conclusions are presented in this paper. During lecture period, students acquire knowledge via F2F teaching and all ways of individual learning. In general, acquired knowledge fades over the time, therefore students that attend late exam terms have to put more effort on individual learning [6], [7], [8]. For students who were unable to attend lectures, the same applies for late exam terms as well as for early exam terms.

The aim of this paper is to examine individual learning and F2F impacts on achievement during all exam terms (early and late) during a single academic year. Due to the fact that for electrical engineering professional courses self-learning is enabled also by distance learning, the conclusions of the paper could provide directions how to improve asynchronous online learning to achieve benefits of F2F teaching (lecture attendance in a class or synchronous online learning).

The rest of the paper is organized as follows. In Section II, input data collecting, preparation and its graphical representation are described. In Section III, data analysis procedure is presented as well as results of

analysis in tabular and graphical form. Finally, in Section IV obtained results and possible guidelines for future development are discussed.

II. Data Collection, Preparation and Representation

The scope of this research are professional courses for undergraduate study of electrical engineering. For this purpose fundamental telecommunication courses *Digital circuits*, *Signals theory and processing* and *Information theory and coding* were selected. Teaching for these courses consists of lectures, exercises and laboratory exercises. Lectures and exercises are held in a F2F manner. Laboratory exercises are performed in a standard hardware or software oriented laboratory environment with F2F assistance of lecturer.

Courses and academic years

Data about students' attendance on lectures and exercises as well as corresponding exam success data are collected for the following academic years:

- *Digital Circuits* – 2017/2018 and 2018/2019;
- *Signals Theory and Processing* – 2018/2019;
- *Information Theory and Coding* – 2017/2018 and 2018/2019.

In cases where data is available for more than one academic year, all course data are aggregated to obtain a more representative sample.

Attendance

Data for *Attendance A* are normalized to scale from 0% (no attendance) to 100% (attendance on all lectures and exercises). Attendance on laboratory exercises is mandatory for all abovementioned courses, so their data are excluded as irrelevant for finding correlation of students' success on exams in relation to their attendance. Moreover, for students who did not complete the laboratory exercises successfully, data records are excluded in further processing. Students that re-enrolls courses are not taken in further processing, too.

Exam terms

During single academic year, knowledge assessments have the following time frames:

Early exam terms

- 1st term: Within 2 weeks after the end of lecture period. Includes grades from continuous course assessment from two mid-term exams held during lecture period.
- 2nd term: Within 4 weeks after the end of lecture period.

Late exam terms

- 3rd term: Approx. 4 months after the end of lecture period.
- 4th term: Approx. 8 months after the end of lecture period.

It is worth to mention that if students do not pass exams in the offered terms, they have to re-enroll the course in the next academic year.

Grades

Obtained *Grades G* from exam terms are normalized to scale from 0% (no achievement) to 100% (maximum possible achievement). It is important to note that grades are in fact achieved points that have fine resolution (as opposed to ordinary grading in education). For students who have accessed more than one knowledge assessment, each attempt is recorded and included in analysis.

Graphical representation of grades over exam terms

Initially, a cumulative distribution function (CDF) was chosen to provide a complete view of student achievement across for all exam terms. However, the complementary cumulative distribution function (CCDF) has been shown to be more convenient for data interpretation [9].

$$S(g) = P(\{G \geq g\}) \tag{1}$$

Graphs presenting (g , $S(g)$) pairs for all analyzed courses are given in Fig. 1.-3. so that each point represents one student at certain exam. For example, on the first exam term for *Digital Circuits* course, there were 88% students that had 50% or more achievement ($g = 50\%$; $S(g) = 88\%$) – see Fig. 1.

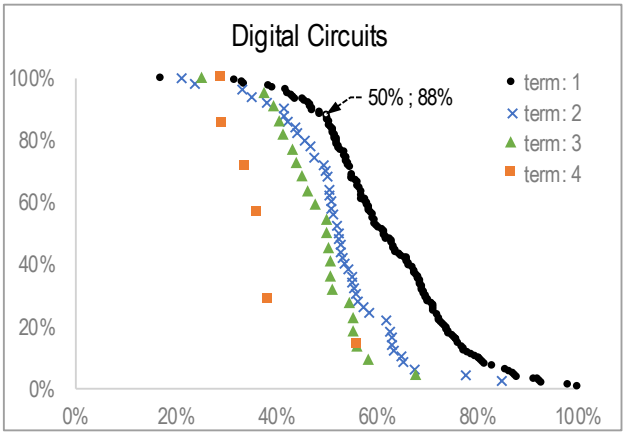


Fig. 1. Complementary cumulative distribution function for all exam terms for course ***Digital Circuits***; *x-axis*: grades G; *y-axis*: percentage $S(g)$ of students with $G = g$ or more achievement

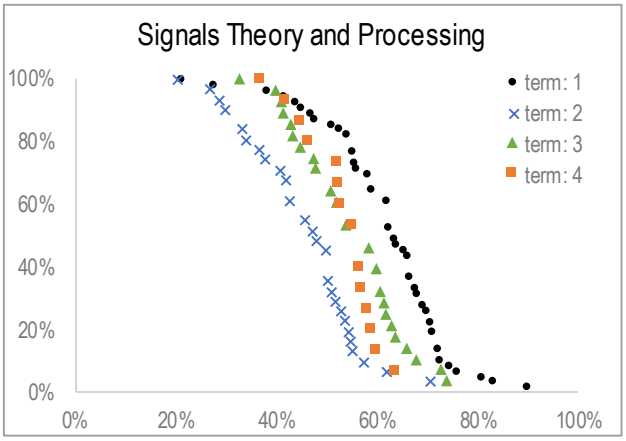


Fig. 2. Complementary cumulative distribution function for all exam terms for course ***Signals Theory and Processing***; *x-axis*: grades G; *y-axis*: percentage $S(g)$ of students with $G = g$ or more achievement

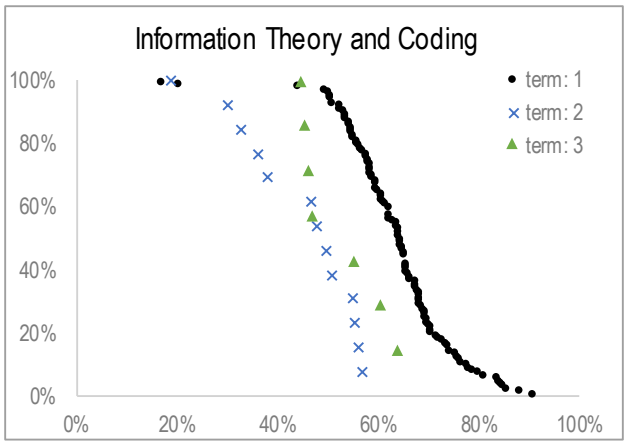


Fig. 3. Complementary cumulative distribution function for all exam terms for course ***Information Theory and Coding***; *x-axis*: grades G; *y-axis*: percentage $S(g)$ of students with $G = g$ or more achievement

According to graphical representations shown in Fig. 1.-3., the highest grades are achieved at the 1st exam term. In addition, this representation provides information about number of attendances showing e.g. that early exam terms have denser dots. Relationship between attendance A and grades G is presented in the next chapter. By using CCDF (1) it is possible to make a quick comparison of students' success on exams for all exam terms.

III. Data Analysis and Results

From the graphic representation of average student success through obtained grades G vs. their attendance of lectures and exercises A for all exam terms for *Digital Circuits* course is shown in Fig. 4.-7. the monotone growth is noticed. Increase in attendance correlates with an increase of exam success, which is more noticeable for earlier exam terms. For showing a growth trend, without entering deeper into area of sophisticated modeling, linear trend line was applied. It is worth to say that the linear trend was calculated from all exam term result, not just from the average values of G as presented in Fig. 4.-7.

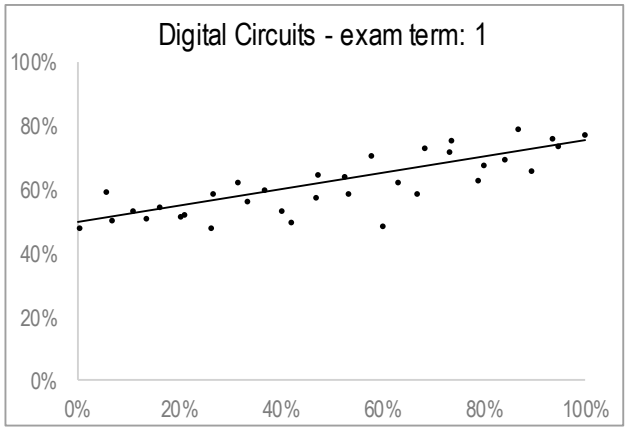


Fig. 4. Average exam success G (*y-axis*) vs. Attendance A (*x-axis*) for the first exam term of course ***Digital Circuits***

It is also noticeable that if the exam term is later, the slope decreased and the intercept increases. In addition, it is noticeable that the majority of students who attend late exam terms belong to a group that rarely attended lectures. Namely, on the graphical representation, the points move to the left side of graphs, i.e. their higher density is near $A = 0$ (rare or no attendance).

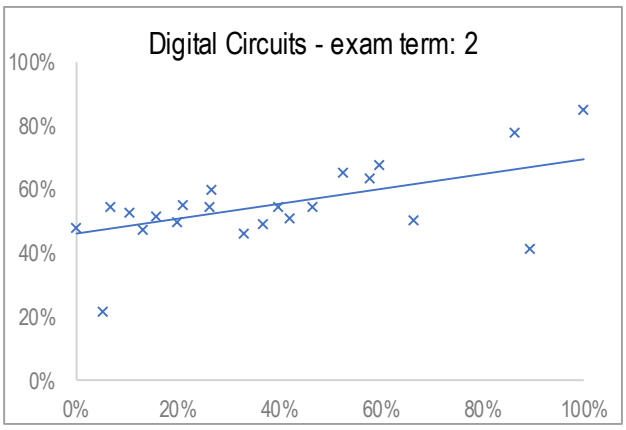


Fig. 5. Average exam success G (*y-axis*) vs. Attendance A (*x-axis*) for the second exam term of course ***Digital Circuits***

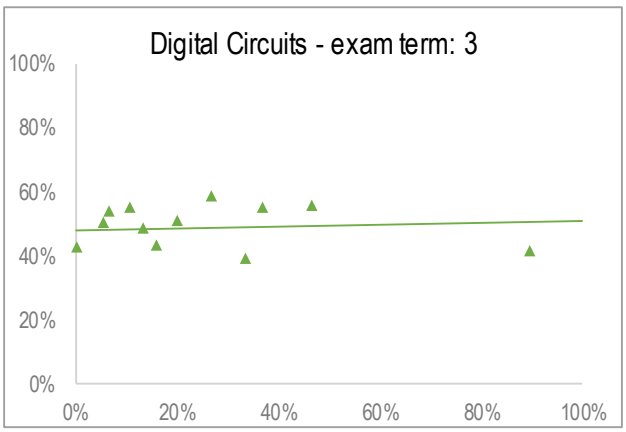


Fig. 6. Average exam success G (*y-axis*) vs. Attendance A (*x-axis*) for the third exam term of course ***Digital Circuits***

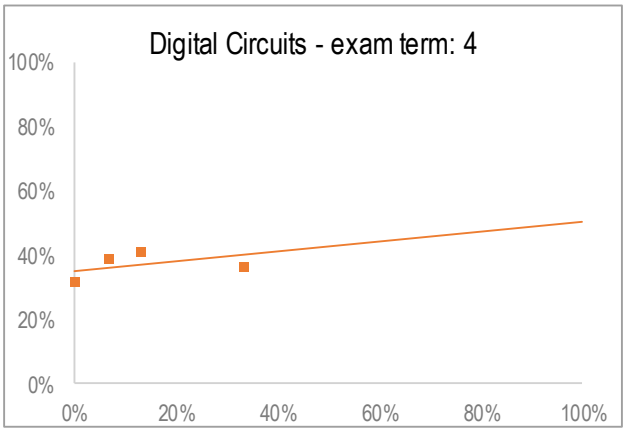


Fig. 7. Average exam success G (*y-axis*) vs. Attendance A (*x-axis*) for the fourth exam term of course ***Digital Circuits***

The graphical representations and trend lines of the data for the remaining two analyzed courses give almost identical results, so their figures are omitted here. Obtained values for intercept and slope for all analyzed

courses, as well as other exam statistical results are given in Tab. I.-III.

TABLE I. DIGITAL CIRCUITS – EXAM STATISTICS

	Term: 1	*Term: 2*	*Term: 3*	*Term: 4*
Intercept	0.497	0.463	0.479	0.353
Slope	0.257	0.232	0.027	0.151
Students at exam	179	50	22	7
Average attendance	0.502	0.262	0.160	0.114
Average success	0.626	0.523	0.483	0.370

TABLE II. INFORMATION THEORY AND CODING – EXAM STATISTICS

	Term: 1	*Term: 2*	*Term: 3*	*Term: 4*
Intercept	0.549	0.377	0.539	0.185
Slope	0.202	0.253	-0.134	-0.030
Students at exam	118	13	7	9
Average attendance	0.445	0.266	0.144	0.105
Average success	0.639	0.444	0.520	0.181

TABLE III. SIGNALS THEORY AND PROCESSING – EXAM STATISTICS

	Term: 1	*Term: 2*	*Term: 3*	*Term: 4*
Intercept	0.479	0.391	0.495	0.590
Slope	0.259	0.188	0.174	-0.345
Students at exam	57	31	28	15
Average attendance	0.519	0.317	0.280	0.178
Average success	0.613	0.451	0.544	0.528

TABLE IV. AGGREGATED EXAM STATISTICS

	Term: 1	*Term: 2*	*Term: 3*	*Term: 4*
Intercept	0.515	0.430	0.490	0.382
Slope	0.234	0.207	0.126	0.071
Students at exam	354	94	57	31
Average attendance	0.486	0.281	0.217	0.142
Average success	0.628	0.488	0.517	0.392
Share of intercept in av. success	81.9%	88.1%	94.7%	97.4%

It is important to note that in the case of a small number of students at exam, unexpected negative slope can occurred. This is unexplainable and the reason should be sought in too small a sample (usually in case of late exams). Therefore, in Tab. IV. aggregated exam statistics for all courses in scope are presented. Such aggregation is justified because all courses belong to the group of fundamental telecommunication courses.

As can be seen from the Tab. IV., values for *Average success* and *Intercept* decrease on late exam terms. In order to have an insight into the share of self-learning in exam success, the introduction of new indicator *Share of intercept in av. success* is proposed, calculated as follows:

$$Share_of_intercept_in\ av._success = Intercept\ /\ Av._success \quad (2)$$

Values of new indicator is added in the last row of the Tab. IV. According to aggregated exam statistic data from Tab. IV., comparing the results of earlier exam terms with later one it can be concluded the following:

a) *Slope* – decreases;

b) Number of *Students at exam* – rapidly decreases;

c) *Average attendance* – decreases;

d) *Average success* – decreases;

e) *Share of intercept in av. success* – increases.

In the abovementioned comparison, comments about intercept trend is intentionally omitted. On first site, the expected increase of *Intercept*, which can be a measure of self-learning amount, is not evident due to the decrease of *Average success*. However, the introduced indicator *Share of intercept in av. success* (2) shows that *Intercept* has an increasing share in *Average success.* Accordingly, in the acquired knowledge that has been assessed at the exam, the part corresponding to the self-acquired knowledge increases in later exam terms.

Therefore, in the acquired knowledge that has been assessed at the exam, the part corresponding to knowledge obtained through F2F teaching (dependent on the attendance) decreases in later exam terms. At the same time, the part corresponding to the self-acquired knowledge (independent on the attendance) increases.

IV. DISCUSSION

Based on results from the previous Chapter, the following findings arise for the role of distance learning (i.e. asynchronous online learning) in self-learning part of knowledge acquisition:

The need for self-learning for students who rarely attend lectures and access late exam terms

The results confirm that students who access the exams at a later term are, as a rule, rarely attending lectures (Monotone decline of *Average attendance* from Tab. IV.). The rise of *Share of intercept in av. success* (which is in fact share of self-learning in acquired knowledge) for late exams confirms that there is a considerable need for self-learning for students who rarely attend lectures and access late exam terms.

The need for continuous and repetitive learning

Also, students who attended lectures regularly, need to repeat/review acquired knowledge if they access to late exam terms. Therefore, the role of self-learning is for acquiring of the new knowledge (for students who rarely attend lectures) as well as for repetition (for students who attended lectures regularly).

Repetition/review is necessary because acquired knowledge fades soon after initial learning and the rate of memory retention (%) then gradually declines [7]. Fig. 8. illustrates the effect of the review.

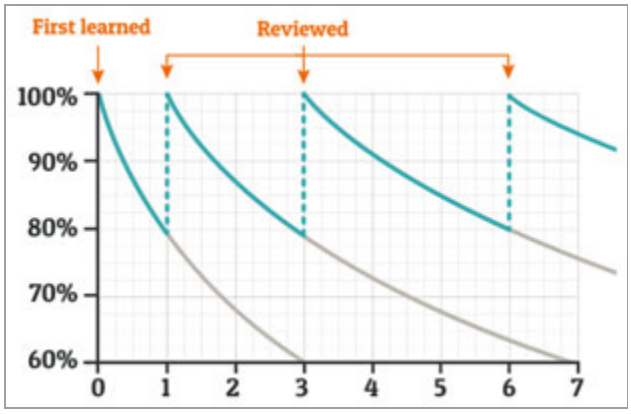


Fig. 8. Alteration of the forgetting curve through repetition/review *x-axis* – elapsed time, *y-axis* – memory retention [7]

Students who access late exam terms have lower average success

Monotone decline of *Average success* is noticed from Tab. IV. for late exam terms. In fact, students who attend lectures rarely, cannot compensate their absence with self-learning neither with asynchronous online learning. Therefore, asynchronous online learning have potentials for improvement.

On the other hand, the demands of the typical student in higher education include the use of ICT in teaching methods, flexible timing and location to complete studies, and real world applicability of courses [10]. In the era the ubiquitous Internet (anywhere, anytime and on any device) instant access to distant learning materials and supporting content is a starting point that need to be provided for students.

Consequently, development of online learning should go in three directions:

- Improving education content and content delivery enabling additional attention;
- Improving motivation for students to acquire knowledge;
- Finding optimal balance of asynchronous and synchronous online learning.

V. CONCLUSION

The data sets related to attendance and overall exam success for three professional courses of electrical engineering undergraduate study were analyzed. The results showed the following:

- Average success is higher if face-to-face lecture attendance is higher;
- Students who access late exam terms have lower average lectures attendance;
- Students who access late exam terms have lower average success;
- Share of self-learning is higher for late exam terms.

In case of frequent absence from lectures, the benefits from face-to-face lectures attendance or synchronous online learning are difficult to compensate. As the support for self-learning, Zagreb University of Applied Sciences offers distance learning via LMS. Based on all stated findings, possible directions for development of online learning are given.

Currently, the main task is to find the optimal balance of asynchronous and synchronous online learning, especially as response to challenges in Covid-19 pandemic crisis. Certainly, the crisis has resulted in a paradigm shift in education. Further research will validate and supplement stated findings by analyzing data collected during pandemic crisis, when F2F learning in the classroom was substituted by synchronous online learning.

REFERENCES

[1] T. Horvat, T. Alajbeg, S. Predanić: Experiences and Practices in Blended Learning Environment, Proceedings of 38th International Convention MIPRO 2015 / Computers in education (CE), pp. 1039 – 1043.

[2] T. Alajbeg, D. Ćika, T. Ražov: Using Moodle to Automatically Grade Laboratory Exercises, Proceedings of 34th International Convention MIPRO 2011 / Computers in education (CE), pp. 1343-1347.

[3] T. Alajbeg, D. Martinić, M. Sokele: Evaluation of the Impact of Changing the Number of Test Tasks on Successful Completion for Automated LMS Assessments, Proceedings of 41st International Convention MIPRO 2018 / Computers in education (CE), pp. 595-599.

[4] I. Hakala, M. Myllymäki: From Face-to-Face to Blended Learning using ICT, IEEE Global Engineering Education Conference (EDUCON), UAE, 2016, pp. 409-418.

[5] M. Sokele, T. Alajbeg, I. Lujo: The importance of face-to-face teaching at fundamental telecommunication courses, Proceedings of the 15th International Conference on Telecommunications ConTEL 2019.

[6] J. M. J. Murre, J. Dros: Replication and Analysis of Ebbinghaus' Forgetting Curve, PLoS ONE journal 10(7), 2015.

[7] B. A. Chun, H. J. Heo: The Effect of Flipped Learning on Academic Performance as an Innovative Method for Overcoming Ebbinghaus' Forgetting Curve, ICIET '18: Proceedings of the 6th International Conference on Information and Education Technology, 2018.

[8] P. Zarevski: Psihologija pamćenja i učenja, Naklada Slap, Jastrebarsko, 1997.

[9] T. Funke, T. Becker: A Tool for an Analysis of the Dynamic Behavior of Logistic Systems with the Instruments of Complex Networks, Proceedings of the 6th International Conference LDIC 2018, pp 418-425, Springer Nature, Bremen, 2019.

[10] S. Hansen (ed.): Capacity Building in a Changing ICT Environment 2018, International Telecommunication Union (ITU), Geneva, Switzerland, 2018.

Use of Open Training Portals to Host Developed STEM Courses

T. Hryhorova*, V.P. Lyashenko*, I. Hvozdeva ** and I. Getman*
* Kremenchuk Mykhailo Ostohradskyi National University, Kremenchuk, Ukraine
** National University "Odessa Maritime Academy"/Department of Electrical Engineering and Radio Electronics, Odessa, Ukraine
grital0403@gmil.com

One of the topical areas of innovative development of education is STEM education. It involves the formation of a future specialist spatial understanding of the logical connections between events or processes with approaches to their solution. Such formation can be realized through interdisciplinary communication, considering how some fundamental knowledge is put into practice. As a rule, basic and professional courses are taught separately, so most students do not form a holistic picture of knowledge. In this way, students slowly learn their professional skills. Therefore, a training course was created that combines two disciplines Probability Theory and Information Theory between which there is a strong link. Open training portals allow you to use similar courses. The course is being tested in groups of students of the specialty Secondary Education Informatics.

Keywords - STEM education; training course; combines disciplines; education portal

I. INTRODUCTION

In today's society, education is becoming one of the key areas of society in which its future intellectual wealth is shaped. The rapid growth of information volumes requires rapid updating of content, forms, methods and learning tools. The priority in the preparation of future specialists was the need to develop the student's creative thinking, to develop his communication skills and practical preparation for active life. STEM education is one approach to learning that meets these needs. STEM education is a low or sequential course or course of study that prepares students for successful employment, requiring different and more technically complex skills, including applying an interdisciplinary and applied approach, mathematical knowledge and scientific concepts [1]. It is constantly evolving and taking on new forms. A new element is added to each of these forms, which fills the education with new content to develop a creative approach to acquiring new knowledge and the ability to use them. The idea behind STEM education is to combine basic knowledge with their immediate use to solve technological or engineering applications. The works of many authors working in the field of STEM-education are devoted to this [2-6]. It shortens study time, increases interest in knowledge acquisition and develops a creative approach that is the key to successful employment.

A. Open educational web portals

Open educational portals contribute to the development of qualitatively new Internet projects in the field of education. The goals of the educational portal are to create a new communication space and information field of the educational professional community through the Internet [7, 8]. STEM education is best suited to these goals. The development of educational portals contributes to: improving the links between existing resources; exchange of information; observance of copyrights, protection of intellectual property; announcing competitions for the best online course, educational technology, educational service, etc. announcement of vacancies; rating of open educational establishments or establishments using distance education technologies; placement of specialty ratings depending on labor market demand. Educational portals have the ability to change and update their content automatically and include all the latest communications and training services over the Internet in a virtual information space: audio and video conferencing, automatic data sharing, electronic message boards, and e-mail. All these opportunities make it possible to introduce training courses to a wider audience. Ukraine uses the Ukrainian online education portal http://prometheus.org.ua/.

B. Reasons

All disciplines taught in the university are divided into three cycles: humanitarian and socio-economic training, mathematical and natural science training, professional training.

Skills and knowledge acquired in the study of basic disciplines are used in the study of professional disciplines. However, the teaching of the disciplines is separate, most students have the idea that the knowledge acquired in the study of these disciplines is in no way interconnected, and in the study of professional disciplines, they have difficulty using the knowledge already acquired in the basic disciplines. STEM education involves the formation of a spatial understanding of the logical connections between processes using approaches to solving them. Such formation can be implemented through the use of interdisciplinary communication in the learning process. Integrate core teaching and training courses that are closely related. Particular attention should

be given to the examples, thus moving away from the general examples used in most of the basic textbooks.

II. PREVIOUS RESEARCH

The aim of this work was to prepare a training course that would integrate disciplines from the block of basic and professional disciplines and place it in the e-learning system of the university and in the open web portal.

After analyzing the basic disciplines and disciplines of professional training, the disciplines that are closely related were identified all the theoretical knowledge obtained in the basic discipline is the basis of an integral part of the professional discipline.

A close link was found between the disciplines of Probability Theory and Information Theory, so it was suggested to create a training course that would combine the two.

Probability Theory is a further development of the school mathematics course and aims at an in-depth study of this discipline.

After examining the list of topics and the examples given in the textbooks on Probability Theory and Mathematical Statistics, it was found that they are all built on almost identical general purpose examples. In the following, only the topics of Probability Theory were considered. Example:

- In the textbook Probability Theory and Mathematical Statistics, by Zhyltsov O. B. – more examples are given, such as: tossing a dice; tossing a coin and fixing the face it will fall into; random taking out of white or black balls from the box, etc.
- In the textbook Probability Theory and Mathematical Statistics, by Zaitsev E. P. – more theory than examples. Examples include: probability of snow or rain, determining which shooter is better, tossing a coin, etc.
- In the textbook A Guide to Solving Problems in Probability Theory and Mathematical Statistics, by Gmurman V. E. – there are more examples than the theory, among the examples: are throwing two dice, determining the intended number, determining the best students in the group, etc.

In any textbooks is hasn't been found any examples that could be used when solving problems that arise in professional activity.

The course Information Theory and Coding is part of the professional training cycle of specialists of educational and qualification level "bachelor". Only the first part of Information Theory was considered.

Analyzing textbooks in the discipline of "Information Theory and Coding" it was found that they are all built in almost the same way. However, in some textbooks such as: "Information Theory and Coding" Zhurakovsky Yu. P., Poltorak V. P.; "Information theory and coding in problems" Zhurakovsky Yu. P., Gnilickiy V. V.; "Information Theory and Coding" Kozhevnikov V. L., Kozhevnikov A. B., there is more theoretical material, but few examples for practical tasks.

To prepare the course, the textbooks that formed the basis for the course Probability and Information Theory were selected:

- Probability Theory and Mathematical Statistics by Zhiltsov A. B. [9] namely part I "Random events and their probabilities" and part II "Random quantities", because it is in this textbook more structured material, after each topic there is a task for self-study, self-test questions.
- A Guide to Solving Problems in Probability Theory and Mathematical Statistics, by Gmurman V. E. [10]
- The Book of Problem on Information Theory and Coding, by Tsymbal V. P. [11]
- A Collection of Examples And Tasks on The Basics of the Theory of Information and Message Coding, by Zverev E. N., Lebedko E. G. [12]
- Fundamentals of the Theory of Information and Coding, Krushniy V. V. [13]

III. RESULTS AND DISCUSSION

The course curriculum consists of the following topics.

The first topic is random events. The basic concepts of Probability Theory which including:

- events and their varieties;
- what the experiment is and what it is related to;
- sets, as a series of elementary events and operations on them;
- statistical probability of events and its attributes;
- probabilistic model;
- measure of the set.

Basic concepts that are part of the mathematical model of information transmission, in the discipline of Information Theory which including:

- information measure;
- entropy of information,
- amount of information.

The second topic is the concept of Probability Theory – elements of combinatory and their applying in the calculation of probabilities. The formulas of addition, subtraction, and multiplication of probabilities:

- elements of "combinatory";
- rules of sum and product;
- probability addition theorem;
- conditional probability;

- what events are independent;
- probability multiplication theorem;
- probability addition theorem for events where events are compatible.

From the discipline of Information Theory:

- send signals through communication lines.

The third topic is Probability Theory – Full Probability Formula and Bayes Formula:

- full likelihood formulas;
- hypothesis;
- Bayes' formula.

From the discipline of Information Theory:

- the communication channel;
- capacity;
- noise or obstacle;
- general scheme of information transmission.

The fourth topic is the concept of Probability Theory is the scheme and formula of Bernoulli. Bordering theorems in the Bernoulli scheme:

- the Bernoulli scheme;
- the Bernoulli formula;
- the rules of large numbers;
- central bordering theorems;
- local Laplace function;
- Bernoulli's theorem;
- Laplace integral function;
- the local Moivre - Laplace theorem;
- P. L. Chebishev's theorem.

From the discipline of Information Theory:

- entropy of typical discrete distributions;
- basics of the theory and practice of message encoding in technological information transmission systems.

The fifth theme is the concept of Probability Theory is a random variable whose numerical characteristics and functions, continuous random variables.

- random variable;
- a discrete random variable;
- continuous random variable;
- the law of distribution;
- number of distribution;
- types of discrete distributions;
- distribution function;
- absolutely continuous and singular continuous random variables;
- types of absolutely continuous distributions;
- numerical characteristics.

From the discipline of Information Theory:

- notification items;
- entropy of continuous messages;
- entropy formulas for continuous messages.

The sixth theme is the concept of Probability Theory is a system of two random variables:

- two-dimensional random variable;
- distribution function;
- the density of distribution;
- conditional law;
- numerical characteristics of conditional laws;
- conditional density.

From the discipline of Information Theory:

- methods for effective coding of the correlative sequence of characters.

Examples of all topics include definition tasks in the Information Theory discipline, which include the basic concepts of the Probability Theory discipline.



Figure 1. The university's distance learning system

The developed course was placed include to the university's distance learning system Fig. 1.

In half a year, more than one hundred students have studied it. This includes not only the students who study at the Informatics and Higher Mathematics Department, but also the students who do course at other departments. Students' interest indicates that the combination of disciplines enables them to understand how to apply the knowledge gained from basic disciplines in their professional activities, and also facilitates the resolution of examples. The study process quality check for disciplines Probability Theory and Theory of Information showed that the quality of training increased by 32% compared to when the courses were taught separately. For the test, an experimental group of students was created which studied the course "Probability Theory and the Theory of Information", instead of the general program, like everyone else. This has become possible due to the

fact that, at first, the basics of Probability Theory are studied at secondary school, Math Statistics is a separate block, and, third, Coding Theory is also separated. Later students of the experimental group joined the rest of the other students. The average score by the five-point system in the experimental group was 3.8 points, and among the rest students, who were not included from this group – 4 points. At the end of the semester, students passed the common test. The test was divided into four blocks: Probability Theory, Math Statistics, Theory of Information and Coding Theory. The results of this test showed that the average score on the Theory of Probability in the experimental group was 4.5 points and in the regular group 3.8 points, on the Theory of Information 4.6 points and 3.6 points, respectively. Score on Math Statistics and Coding Theory were almost the same. But to get more accurate results on effectiveness of this course, we need to use an open learning portal. Then everyone will be able to access it and the statistics will be more accurate. We are currently working on this issue.

IV. CONLUSION

Basic and professional training disciplines were analyzed as well as isolated disciplines that are closely related. A close link was found between the disciplines of Probability Theory and Information Theory. It was suggested to create a training course that would combine the two disciplines.

The textbooks that formed the basis for the training course Probability and Information Theory were selected. Theoretical and practical materials were developed according to the course plan. The integrated course is based on the principles of STEM education.

The developed course was tested among university students who got acquainted with it in the e-learning system of the university. Testing results show that students learn the material better. Further testing of the course will be held in an open education portal, which will give more reliable results.

REFERENCES

[1] Dennis Vilorio. “STEM 101: Intri to tomorrow″s jobs.” 2014. US: Bureau of Labor Statistics (BLS).

[2] H. A. Diefes-Dux, In. S. Purzer, J. Strobel, & M. Cardella (Eds.) “In-service teacher professional development in engeneering education: Early years.” 2015. Engeneering precollege setting: Synthesizing research, policy, and practices. Lafayette: Purdue Universite Press.

[3] So Yoon Yoon, Yi Kong, Heidi A. Diefes-Dux, and Johannes Strobel “Broadening K-8 Teachers″ Perspectives on Professional Development in Engineering Integration in the United States.” 2018. International Journal of Research in Education and Science, Volume 4, Issue 2, pp. 331-348.

[4] D. Duncan, H. Diefes-Dux, & M. Gentry “Professional development through engineering academies: An examination of elementary teachers,, recognition and understanding of engineering.” 2011. Journal of Engineering Education, pp. 520-539.

[5] D. Duncan, , E. Oware, M. Cox, & H. Diefes-Dux “Program and curriculum assessment for the Institute for P-12 Engineering Research and Learning (INSPIRE) summer academies for P-6 teachers.” Honolulu, HI, USA, 2007. [Proceedings of the 114th American Society for Engineering Education Annual Conference and Exposition].

[6] S. Y. Yoon, M. Dyehouse, A. Lucietto, H. A. Diefes-Dux, & B. M. Capobianco “The effects of integrated science, technology, and engineering education on elementary students,, knowledge and identity development.” 2014. School Science and Mathematics, 114(8), pp. 380-391.

[7] Matt Andrews and Nick Manning. “Peer learning in public sector reforms.” 2015. Paris: Effective institutions platform (EIP).

[8] T. Hryhorova, V. Lyashenko and D. Kalinichenko “Approaches for implementation of distance learning in higher education.” Opatija (Croatia), pp. 1070-1074, 25-29 may, 2015. [Digests 38 International convention on information and communication technology, electronics and microelectronics “MIPRO 2015”].

[9] A. B. Zhiltsov, Probability Theory and Mathematical Statistics. Kyiv: Kyiv University. B. Grinchenko, 2015.

[10] V. E. Gmurman, A Guide to Solving Problems in Probability Theory and Mathematical Statistics. Moscow: Higher School, 2004.

[11] V. P. Tsymbal, The Book of Problem on Information Theory and Coding. Kyiv: Higher School, 1976.

[12] E. N. Zverev, E. G. Lebedko, A Collection of Examples And Tasks on The Basics of the Theory of Information and Message Coding. St.Petersburg: NIU ITMO, 2014.

[13] V.V. Krushniy, Fundamentals of the Theory of Information and Coding. Snezhinsk: SGFTA, 2005.

Internationalization and Student Mobility: STEAM Students Case Study

Petra Cuculić*, Tihana Babić* and Marcela Pavlica*
*Algebra University College, Zagreb, Croatia
petra.cuculic@algebra.hr, tihana.babic@algebra.hr, marcela.pavlica@algebra.hr

Abstract - **Higher education institutions have been increasingly shifting their focus towards various internationalization strategies in order to boost the numbers of international students studying abroad both as full time students and exchange students. As a result of their efforts, there has been a growing number of students who see mobility as an excellent opportunity to experience living and studying in another country. Students who undertake short and long-term mobility programmes profit from student mobility not only regarding knowledge gained but also from experiencing living abroad, being independent, exploring new cultures. However, mobility trends among STEAM students show different results. In other words, there is a negative tendency in outgoing mobility among STEAM students who are reluctant towards joining mobility programmes. At the same time, professionals who work in international office think that STEAM students would immensely profit from short and long-term mobility programmes. This paper thus addresses key challenges experienced by international office staff when trying to motivate students to undertake mobility programmes. It also explores reasons behind negative trends in outgoing mobility of STEAM students by evaluating the results of a survey among these students.**

Key words: internationalization, student mobility, STEAM students, international office

I. INTRODUCTION

In academic circles of today's globalized world, internationalization has become one of the buzzwords. But what does it really stand for? Jane Knight defines internationalization as "the process of integrating an international, intercultural or global dimension into the purpose, functions and delivery of post-secondary education" [1]. In other words, Knight sees internationalization as an all-encompassing and comprehensive concept, which pervades various aspects of higher education. Despite its multifaceted nature, this paper particularly focuses on one of the key concepts of internationalization – student mobility, which according to Knight "continues to be a high priority of internationalization" [2].

Student mobility is a term that refers to free movement of students between higher education institutions in different countries. It is usually limited to a defined period of time after which students return from the host university to their home university. Data show that the number of mobile students has been growing steadily over the years. According to Tse-Mei Chen, George A. Barnettand and OECD data in Jane Knight, "The growth in the number of mobile students from about 238,000 in the 1960s to 3.3 million in 2008 is staggering. If forecasts are correct, this number will double in another ten to fifteen years" [3].

Depending on the direction of mobile student flow, student mobility can either be described incoming or outgoing. Student mobility programmes can be divided into several categories. Based on the duration of the programme students take part in, student mobility programmes can be defined either as short or long-term. Semester and yearlong mobility programmes are those during which students stay at a foreign host institution for a semester or whole academic year, while short-term programmes last about three or four weeks, usually during summer vacation period. Popular short-term programmes include summer schools, which offer various theme-specific courses or intensive language courses. Knight thinks that short term mobility programmes, such as semester or year abroad programmes are popular mostly among students who are unable to afford studying abroad as full-time students or do not have the time to do it[4]. However, when examining reasons why students decide to study abroad as full-time students, the rationale is far more complex. In their scientific paper "The 'Push-Pull' Factors Influencing International Student Selection of Education Destination", Tim Mazzarol and Geoffrey Soutar discuss a selection of various push and pull reasons that motivate students to study abroad. As a starting point, they identify three stages in the decision-making process of students who decide to study in a foreign country. Firstly, the student has to make a conscious decision to study abroad and not in his/her home country. After deciding that studying abroad is the right choice, the student selects the host country where he/she will pursue the studies. The final decision the student has to make is the selection of the host university, which again depends on the institution's reputation regarding quality, courses selection, expertise of teaching staff [5].

When comparing these two sets of reasons behind the decision to study abroad, it can be argued that, from the point of view of students, it is much easier to make a decision to participate in mobility programs than to decide to study abroad as full-time students. Mobility programs offer students a possibility to experience living abroad in a foreign country for a limited period of time, after which they can return to their home university and home country. During their mobility period, students can profit from learning a foreign language, exploring the culture and country they are living in, meeting young people from

different parts of the world, becoming independent and more open-minded.

II. EXAMPLE OF BEST PRACTICE: ERASMUS PROGRAMME

In the European Union, it is clear that mobility as a general concept plays an important role in its core values and policies. In fact, in 2006 the European Parliament and the Council of the European Union passed the European Quality Charter for Mobility in which it is stated that, "Mobility is […] one of the main objectives of the European Union's action in the field of education and training, based both on common values and on respect for diversity. It is an essential tool for creating a genuine European area of lifelong learning, for promoting employment and reducing poverty, and for helping to promote active European citizenship"[6]. Having in mind this premise, it is understandable why the EU devotes considerable effort in maintaining its widely known and recognized programme Erasmus, which was established more than 30 years ago, in 1987. The programme has since served as an example of best practice when it comes to the mobility of students and higher education academic and non-academic staff. According to the publication of the General Secretariat of the Council as many as 33 programme countries and around 90% of European universities are involved in Erasmus+ scheme. In 2014 Erasmus evolved into Erasmus+, the EU's programme for education, training, youth and sport and it was allocated a budget of 14.7 billion euro for the period 2014 – 2020. By the year 2020, its aim is, among other things, to support 2 million higher education students in mobility exchange [7].

As mentioned earlier, Erasmus+ is widely known and recognized, even outside of the European Union. What is more, the programme overwhelmingly dominates among other less popular or famous mobility and exchange programmes in Europe. The Special report 22/2018 of the European Court of Auditors even established that, "Erasmus+ is a well-known, successful EU brand. Erasmus+ plays a key role in building up learning mobility abroad and has a positive effect on participants' attitudes towards the EU. Countries would not be able to achieve such effects acting alone" [8]. In other words, Erasmus+ is popular among students to the extent that it has indeed become a recognized EU brand. It has won its popularity because it provides students with the opportunity of studying at one of the partner universities for a defined period of time, providing them a financial grant that helps them cover their living costs and offering the opportunity of recognizing credits they collect during their mobility period. In other words, students can experience living and studying abroad, while at the same time retaining all the rights and student status they had at the home university. When they return to their home university, they can simply pick up where they left of. During the period of time they spend abroad, students have the opportunity to hone their language skills, experience living on their own in a foreign country; moreover, they become more independent, open-minded towards different cultures and lifestyles, which is important for citizens of a multicultural Europe. Besides, nowadays employers prefer employees with an international study or work experience, which positions Erasmus students in a more competitive place on the labour market. In fact, Erasmus+ Higher Education Impact Study found that, "students who complete an Erasmus+ mobility for studying or training boost their employability skills, with a large majority (72%) saying it had been beneficial or highly beneficial in finding their first job. Their mobilities increased their technical, inter-personal and inter-cultural skills and competences, as well as their self-confidence, ability to achieve goals, and social and cultural openness" [9].

Having in mind the popularity of Erasmus+ and its far-reaching influence, it could even be argued that in the EU, its name is identified with mobility as a general concept. In other words, when someone says "Erasmus+" what they actually mean is "mobility". However, leaving aside the general popularity of Erasmus+, studies show that this EU brand is not equally popular among students of all study programs. The EU survey Erasmus Facts, Figures & Trends: The European Union support for student and staff exchanges and university cooperation in 2013-2014 found out that, "Students of social sciences, business and law made up the biggest share (41 %) of those on exchanges. The second biggest share was made up of students of humanities and arts (22 %). Students of engineering, manufacturing and construction (15 %); science, mathematics and computing (7.5 %); and health and welfare (6 %) continue to participate actively, though in proportionately lower numbers compared to the overall number of students taking these subjects" [10]. Considering these results, it seems that students of humanities and arts are more prone to undertaking mobility programmes and consequently constitute a major share of students willing to participate in student mobility programmes.

III. STUDENT MOBILITY TRENDS AT ALGEBRA UNIVERSITY COLLEGE

Algebra University College in Croatia has six undergraduate and seven graduate study programmes, specialized mainly in the area of IT. On the undergraduate level, there are three subspecializations in computer science, namely, Software Engineering, System Engineering and Multimedia Computing, and two more subspecializations on the graduate level: Data science and Game Development. Apart from these study programmes, Algebra University College carries out Digital Marketing study programme on both undergraduate and graduate levels. Moreover, in the area of design, Algebra University College carries out Visual Communications Design and 3D Design on the undergraduate level and Design and Communications Management on the graduate level of studies.

A. Incoming Students

According to the data provided by the International Office of Algebra University College, the institution has so far signed bilateral mobility agreements and memoranda of understanding with 124 institutions all

around the world, out of which 88 are Erasmus+ Inter-institutional Agreements and 36 are bilateral agreements signed outside of Erasmus+ framework. The first incoming Erasmus+ students studied at Algebra University College in the academic year 2013/2014 and up to the academic year 2019/2020 incoming students have shown great interest for spending their semester abroad at Algebra University College.

B. Outgoing Students

Outgoing mobility, that is outgoing students, at Algebra University College have been active since the academic year 2012/2013 and up to the winter semester of the academic year 2019/2020 the International Office has marked a steady exponential growth in the number of students who participated in long and short-term mobility programmes. In the academic year 2019/2020, the International Office has also recorded a marked growth of interest among students towards short-term mobility programmes (summer and winter schools) and student internships. When it comes to the distribution of Algebra University College students regarding their participation in long and short-term mobility programmes based on their study programs, there was almost an equal proportion of STEAM and Digital Marketing students when considering the period from the academic year 2012/2013 up to the current academic year 2019/2020.

However, in the academic year 2019/2020 and compared to the previous years, the International Office marked a significant growth of interest of Algebra University College students for participation in long-term mobility programmes. Already in the winter semester, a substantial number of Digital Marketing students participated in Erasmus+ student mobility for studies and more Digital Marketing and Design students will participate in Erasmus+ student mobility for studies in the summer semester of the academic year 2019/2020. Compared to previous years, this is a substantial increase in participation of students in student mobility for studies programme, more precisely, an increase of 300%. However, most of the students who have undertaken student mobility for studies belong to the area of digital marketing, while a smaller number of students belong to STEAM area. The question thus is what are the reasons behind the lower interest of STEAM students towards undertaking short and long-term student mobility programmes?

IV. RESEARCH AIM AND QUESTIONS

The general research aim of this article is to determine why STEAM students of Algebra University College are less inclined towards undertaking short and long-term mobility programmes.

To fulfill research aim more easily, the following research questions were defined:

1. Research question: What motivates STEAM students for participating in student mobility programmes?
2. Research question: What do STEAM students consider as advantages of participating in student mobility programmes?
3. Research question: What do STEAM students consider as disadvantages of participating in student mobility programmes?

A. Research Sample

The research was conducted on the voluntary sample of 116 STEAM students of Algebra University College who have not yet taken part in student mobility programmes. The term STEAM stands for the fields of science and technology combined with the field of arts, in the case of Algebra University College, these are the following study programs: Software Engineering, System Engineering, Multimedia Computing, Game Development, Data Science, Design and Communications Management and Visual Communications Design. Out of the total number of survey participants, 46 (39.7%) of them were Software Engineering students, 28 (24.1%) Visual Communications Design students, 20 (17.2%) System Engineering students, 15 (12.9%) Multimedia Computing students. Other participants were 3 (2.6%) Game Development students, 3 (2.6%) Design and Communications Management students and 1 (0.9%) Data Science student.

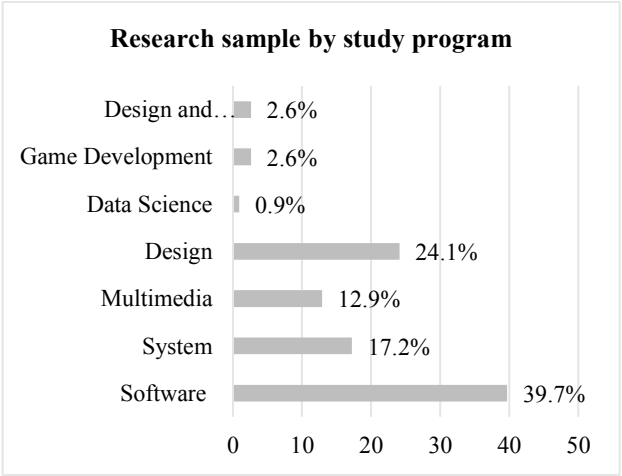


Chart 1. Research sample by study programme expressed in percents, N=116.

The majority of participants were on the undergraduate level of their studies; 112 (96.5%) of them, while 4 (3.5 %) of the participants were on the graduate level of their studies. The majority of participants were male, 86 (74.1%) of them, while there were 30 (25.9%) female participants.

B. Research Methods

The research was conducted using an anonymous voluntary survey during the winter semester of the academic year 2019/2020. The survey was designed as a Google Forms questionnaire with 12 questions: seven closed-ended questions, one open-ended question and two statements. Four questions at the beginning of the questionnaire were related to demographics. For some

questions and statements, the participants had to indicate the extent to which they agree or disagree with them. The level of agreement or disagreement was defined using Likert scale.

In two questions, the participants had to indicate whether they would like to participate in mobility programmes and in which type of mobility programmes (short or long-term) they would prefer to participate. Two questions were asked to check how well students were informed about mobility options and who they thought they should contact to get more information. Finally, to analyse the results in Google Forms quantitative method was used. For the purpose of this article, the emphasis is put on the analysis of six selected questions, which are listed in the following paragraph.

C. Research Results – Interpretation

To examine the first research question more easily, students were asked the following set of questions in the questionnaire:

1. *I am familiar with the possibilities of participating in student mobility programmes*

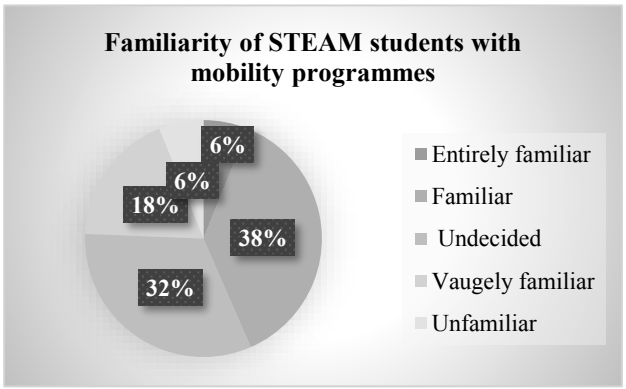


Chart 2. The rate of agreement/disagreement of STEAM students with the statement: *I am familiar with the possibilities of participating in student mobility programmes* expressed in percents N = 115.

When asked to state to what extent they were familiar with the possibilities of participating in student mobility programmes, only 6% of total 115 respondents say they are entirely familiar. However, a fairly large percentage – 38% of STEAM students say they are familiar with the possibilities of participating in student mobility programmes. This means that in total almost 50% of the respondents think they are well informed about the possibilities of participating in student mobility programmes.

On the other hand, 18% of the respondents say they are vaguely familiar with the possibilities, while 6% are completely unfamiliar with the possibilities of participating in student mobility programmes. Finally, 32% of students said they are undecided regarding this statement. The distribution of answers on familiarity of STEAM students with participation in mobility programmes is presented in Chart 2.

2. *I would like to participate in student mobility programmes.*

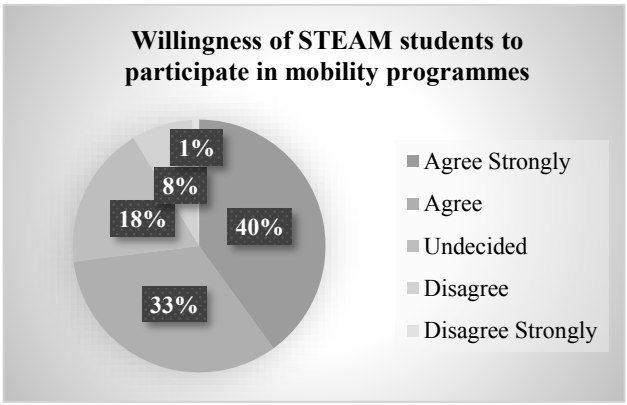


Chart 3. The rate of agreement/disagreement of STEAM students with the statement: *I would like to participate in student mobility programmes* expressed in percents N = 115.

As many as 40% of the respondents strongly agree with the statement "I would like to participate in student mobility programmes.", in other words, 40% of STEAM students show their willingness and motivation to participate in student mobility programmes during their studies. Moreover, another 33% of them agree with the statement. In other words, 73% of all the participants in the survey stated that they are inclined towards undertaking mobility programmes in the future. On the other hand, 18% of STEAM students are undecided about whether they would like to participate in student mobility programmes, while 7.8% disagree and 0.9% strongly disagree with the statement, and thus show their unwillingness to participate in student mobility programmes. The rate of agreement/disagreement of STEAM students with the statement: "I would like to participate in student mobility programmes" is shown in Chart 3.

3. *In which type of student mobility programmes would you like to participate most?*

When asked in which type of student mobility programmes they would like to participate most, the total number of 111 respondents replied as follows: 38% percent of them would like participate in Erasmus+ student mobility for studies (SMS), while 30% of them would like to participate in Erasmus+ student mobility for traineeships (SMT). Regarding other responses, 21% of the participants said that they would like to participate in summer school programs, while 11% of them would like to take part in intensive language courses. The results of this Chart confirm the popularity of Erasmus+ programme as mentioned in one of the previous sections, since most students identify Erasmus+ programme with sole mobility. The Distribution of answers to the question "In which type of student mobility programmes would you like to participate most?" is presented in Chart 4.

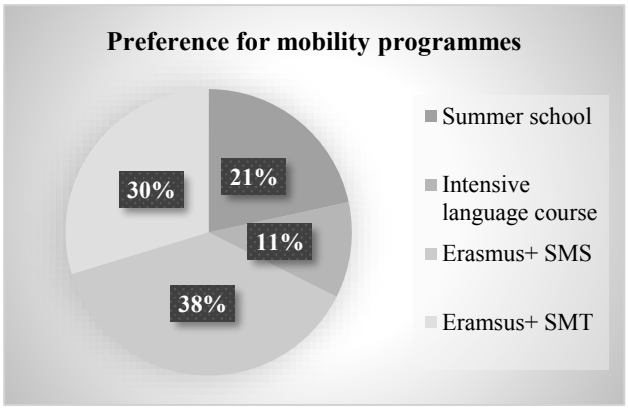


Chart 4. Distribution of the answers to the question: *In which type of student mobility programmes would you like to participate most?* Results are expressed in percents, N=111.

To examine the second research question more easily, students were asked the following question in the questionnaire:

4. *What do you think are the biggest advantages of participating in student mobility programmes?*

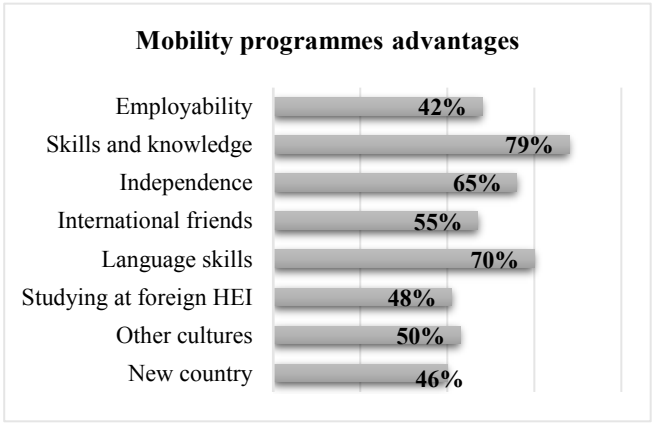


Chart 5. Rating of the biggest advantages of student mobility programmes, expressed in percents, N = 116.

Regarding advantages of participation in student mobility programmes, *acquiring new skills and knowledge* was rated as a major advantage based on the opinion of 79% of STEAM students. On the second place among advantages, 70% of STEAM students put d*eveloping foreign language skills*. Moreover, *becoming more independent* was rated as an advantage by 65% of STEAM students. In addition, 42% of the participants think that student mobility programmes can help them in *finding their job*, while 55% of them think that it can help them *make international friendships*. Furthermore, 50% of the participants think that student mobility programmes can help them *explore new cultures*, while 48% of the participants see the sole opportunity of *studying at a foreign higher education institution (HEI)* as an advantage. Finally, the total number of 46% of the students see the opportunity of *living in a foreign country* as an advantage.

The ranking of advantages of student mobility programmes participation done by STEAM students can be seen in Chart 6.

To examine the third research question more easily, students were asked the following set of questions in the questionnaire:

5. *What do you think are the biggest disadvantages of participating in student mobility programmes?*

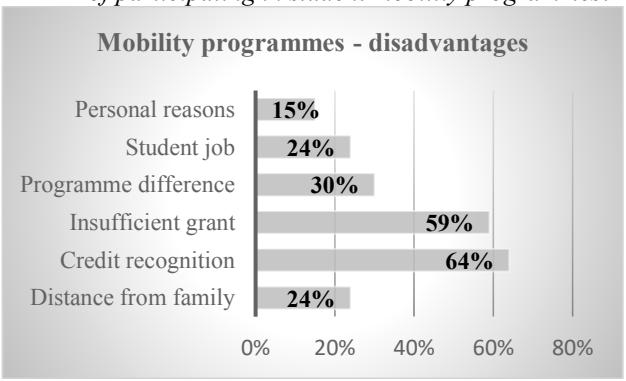


Chart 6. Rating of the biggest disadvantages of student mobility programmes participation expressed in percents, N = 116.

When considering major disadvantages of student mobility programmes participation as ranked by the STEAM students who participated in the survey, *credit recognition* was placed first with 64% of participants evaluating it as a major disadvantage. It was followed by *the insufficient amount of grant*, which is considered to be a disadvantage by 59% of STEAM student. The *difference between study programmes at home and host universities* is seen as a disadvantage by 30% students, while 24% of them think that student mobility programmes could have a *negative influence on their student job* or *distance them from their family*. Finally, 15% of students listed *other personal reasons* as disadvantages of participating in student mobility programmes.

The ranking of disadvantages of student mobility programmes participation done by STEAM students can be seen in Chart 6.

6. *What are other obstacles that prevent you from participation in student mobility programmes?*

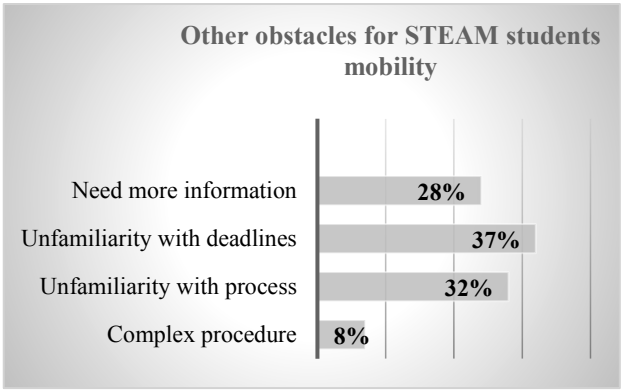


Chart 7. Rating of administrative reasons that discourage STEAM students from applying for student mobility programmes, expressed in percents, N = 116.

The final question STEAM students were asked was what are other obstacles that could prevent or demotivate

them from participation in student mobility programmes. They were given a list of possible answers from which they could assess more and less probable ones. The results have shown the following: the total number of 37% of the respondents say that they have not applied so far because they are *not familiar with deadlines for application*. Similarly, 32% of them say that they are *not familiar with the application process*, while 28% of them think they *need more information about how the application process looks*. Only 8% of the participants stated that their reason for not applying is *complex application process*.

Chart 7 shows how STEAM students ranked their reasons for not applying for student mobility programmes so far.

V. CONCLUSION

The results of the survey conducted among 116 STEAM students of Algebra University College have shown that the majority of students are familiar with the possibilities of participating in student mobility programmes and clearly see benefits of undertaking such programmes during their studies. Moreover, the survey also showed that there is a great interest among STEAM students to participate in student mobility programmes.

At the same time, STEAM students see credit recognition and insufficient amount of grant as main disadvantages of participation in student mobility programmes, which discourage them from participating in mobility programmes. Also, some of them listed some other obstacles such as unfamiliarity with the deadlines for application process or not being familiar with the application process.

Even though a vast majority of the EU universities participate in Erasmus+, as the most widely recognized mobility programme in the EU, and a considerable amount of funds is reserved for student mobility, higher education institutions still face challenges when implementing this and other mobility programmes. As shown in the survey, students see poor recognition of credits and insufficient amount of grant as the main reasons that discourage them from taking part in student mobility programmes. Taking into account the results of the EU survey Erasmus Facts, Figures & Trends, according to which students of social sciences, business and law are a group of students more inclined to taking part in student mobility programmes, a special attention should be devoted to students of other, less inclined student groups.

At Algebra University College, the results are in line with the abovementioned study: Digital Marketing students make up the biggest share of students who take part in student mobility programmes. Therefore, when it comes to STEAM students, who are by definition a less inclined group, special care should be put on informing them about the application process details, important dates and motivating them to undertake mobility during their studies.

Moreover, higher education institutions should also strive to overcome the disadvantages, such as being more flexible in recognizing students' credits after the end of their mobility periods. If there is a difference in course content, at least some learning outcomes should be recognized and students should get an opportunity to pass other learning outcomes or exams on time so that they can keep the pace with their home university colleagues. That way we could motivate more students to participate in student mobility programmes. This practice has recently been implemented at Algebra University College and shows better results among students who are more motivated into undertaking mobility programmes.

REFERENCES

[1] J. Knight, "Higher Education in Turmoil: The Changing World of Internationalization", Sense Publishers, 2008.

[2] J. Knight, "Student Mobility and Internationalization: Trends and Tribulations", in Research in Comparative and International Education, Volume 7, Number 1, 2012.

[3] J. Knight, "Student Mobility and Internationalization: Trends and Tribulations", in Research in Comparative and International Education, Volume 7, Number 1, 2012.

[4] J. Knight, "Student Mobility and Internationalization: Trends and Tribulations", in Research in Comparative and International Education, Volume 7, Number 1, 2012.

[5] T. Mazzarol and G. Soutar, "The 'Push-Pull' Factors Influencing International Student Selection of Education Destination", International Journal of Educational Management, January 2002.

[6] Recommendation of the European Parliament and of the Council of 18 December 2006 on transnational mobility within the Community for education and training purposes: European Quality Charter for Mobility, Official Journal of the European Union, L394/5, 2006. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32006H0961 [accessed: 1/11/2019].

[7] Celebrating 30 years of the Erasmus Programme, The General Secretariat of the Council, 2017. Available at: https://www.consilium.europa.eu/en/documents-publications/library/library-blog/posts/celebrating-30-years-of-the-erasmus-programme/ [accessed: 23/10/2019].

[8] Special report 22/2018: Mobility under Erasmus+: Millions of participants and multi-faceted European Added Value, however performance measurement needs to be further improved, European Court of Auditors, 2018. Available at: https://www.eca.europa.eu/Lists/ECADocuments/SR18_22/SR_ERASMUS_EN.pdf [accessed: 7/10/2019]

[9] Erasmus+ Higher Education Impact Study - Final Report, Publications Office of the European Union, 2019. Available at: http://www.sepie.es/doc/comunicacion/publicaciones/2019/erasmus_higher_education_impact_study.pdf [accessed: 3/11/2019]

[10] Erasmus Facts, Figures & Trends: The European Union support for student and staff exchanges and university cooperation in 2013-2014, Unit B1 'Higher education', Directorate-General for Education and Culture, European Commission, Brussels, 2015. Available at: https://ec.europa.eu/assets/eac/education/library/statistics/erasmus-plus-facts-figures_en.pdf [accessed: 12/10/2019]

Educating the Educators for Introducing Internet of Things to Primary and Secondary Schools' Curriculums

Aleš Jaklič
University of Ljubljana/Faculty of Computer and Information Science, Ljubljana, Slovenia
ales.jaklic@fri.uni-lj.si

***Abstract* - We describe an approach to continuous education course at the University of Ljubljana for training of computer science and technology teachers in primary and secondary schools. These teachers face numerous challenges how to present computer science and engineering as a dynamic and interesting field to their students. Our approach is based on idea to introduce computer science and engineering by building Internet of things devices. A low-cost open source NodeMCU hardware and firmware platform based on ESP8266 is used, that provides WiFi connectivity and Lua interpreter as programming language. Raspberry Pi is used as a central hub for IoT devices. Teachers build an IoT system for monitoring indoor temperature and humidity, outdoor temperature and controlling an air-conditioner using IR control. Finally, we present the course evaluation by the teachers.**

Keywords - IoT; teachers; primary school; secondary school; ESP8266; NodeMCU; smart home

I. INTRODUCTION

Ministry of Education, Science and Sport in Republic of Slovenia organizes continuous education courses for teachers in primary and secondary schools. The courses are offered at the universities, which apply each year to a call for proposal for courses - workshops. The courses typical length is 24 hours spread over several days, for example 3 x 8 hours or 4 x 6 hours. The courses are usually delivered on Saturdays so that they not interfere with teachers' pedagogical processes in their schools.

Faculty of computer and information science at University of Ljubljana each year organizes Summer school for children from primary and secondary schools as an outreach program. Due to limited class sizes in the summer school programs we decided to scale out the provided curriculum to primary and secondary educational level.

Ideally, all these efforts should combine to establish an engineering pipeline [1] as a pathway of activities geared towards motivating students all through primary and secondary school to choose an engineering discipline at university level.

Numerous recent publications [2]–[4] in the field of education stress the importance of educating the Internet-of Things generation about IoT technologies since they present important opportunities for novel applications in various fields [5].

The rest of the paper is organized as follows: in section II we present a practical use case used for the course in detail, in section III we present the evaluation of the course, followed by section IV with conclusions.

II. USE CASE FOR THE COURSE – SMART APARTMENT

To increase motivation of participants we selected a practical use case of IoT application in a smart home. We wanted to remotely monitor temperature inside an apartment, outside temperature and be able to turn on/off an air-conditioner to heat or cool the apartment. In this way we can return home to cooled or heated apartment after summer or winter holidays.

The use case offers ample opportunities to teach basics of IoT, that is TCP/IP protocol, IP addressing, HTTP protocol, WWW servers, basics of HTML, Lua programming language, event driven programming and basics of PHP programming.

Figure 1 depicts the hardware system architecture of a smart apartment.

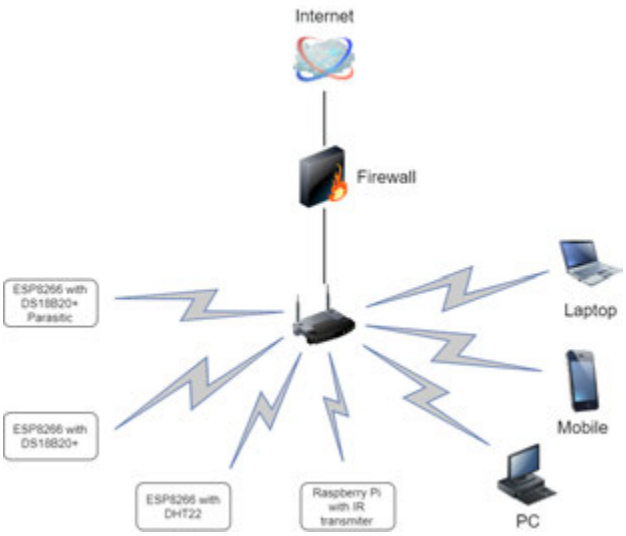


Figure 1. System architecture of a smart appartment

The whole system consists of two subsystem parts:

1. temperature and humidity sensors and
2. an actuator part to turn on/off heating or cooling (IR transmitter and a smart plug).

A. *Temperature and humidity sensors*

Three ESP8266 microcontrollers (two Wemos D1 minis (Figure 1), one ESP-01) with WiFi interface were used as sensor IoT nodes to monitor temperature in two rooms and outdoors. For details about implementation see [6].

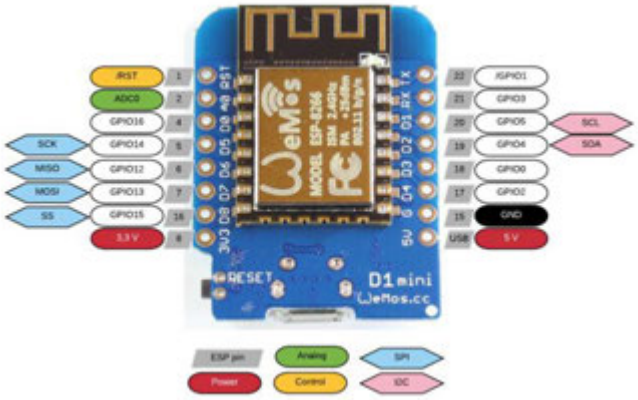


Figure 1. Pinout of a compact version of NodeMCU hardware development board – D1 mini

The IoT node that measures outdoor temperature uses a sensor DS18B20+ with parasitic power harvesting, so that we can use only two existing wires for connecting the sensor on the outer wall of the apartment. The other two sensor IoT nodes use the DTH22 and a regular DS18B20+ sensor locally on a protoboard. Figure 2 shows one of the sensors nodes implemented with ESP-01 using existing wiring for the temperature sensor.



Figure 2. IoT node with ESP-01 and parasitically powered temperature sensor DS18B20

A Raspberry Pi single board computer running Linux is used as a central IoT node that collects data from sensor nodes over HTTP and also serves collected data in a form of a Web page (Figure 3). Each of the sensor nodes has a small WWW server implemented in Lua [6].

A simple PHP script that runs on Raspberry Pi Apache web server provides aggregated data to the user, together with uptimes of the system's components (Figure 3).



Figure 3. Temperature and humidity measurements presented through a Web interface

B. *Actuators – IR transmitter on Raspberyy Pi and smart plug*

The Raspberry Pi also serves as an actuator by sending IR commands to the air-conditioner. Instead of reverse engineering the IR protocol to control the air conditioner, we recorded the IR signal from the IR remote, that was set to cooling or heating. Then we played back these signals using IR transmitter LED diode. Figure 4 shows a setup with an IR receiver and air-conditioner IR remote for signal capture.



Figure 4. IR command protocol recording on a Raspberry Pi 4

We used *sudo ir-ctl -d /dev/lirc1 –receive* command to record the signal from the IR remote controller in a form of pulses and spaces with corresponding timing parameters. The sequence was then further processed in MS Excel to make the measured timing parameters consistent. Finally, the sequence was included into open source IR slinger library [7] for sending the IR command

through IR diode connected to Raspberry Pi through a transistor (Figure 5).



Figure 5. IR LED diode connected to Raspberry Pi as a part of IR transmitter

To provide additional layer of security and power savings we connected the air-conditioner through a smart plug (Figure 6.). In this way we can monitor with an Android mobile application Smart Life the power consumption of the air-conditioner and check to see if it is really turned on or off (Figure 7.)

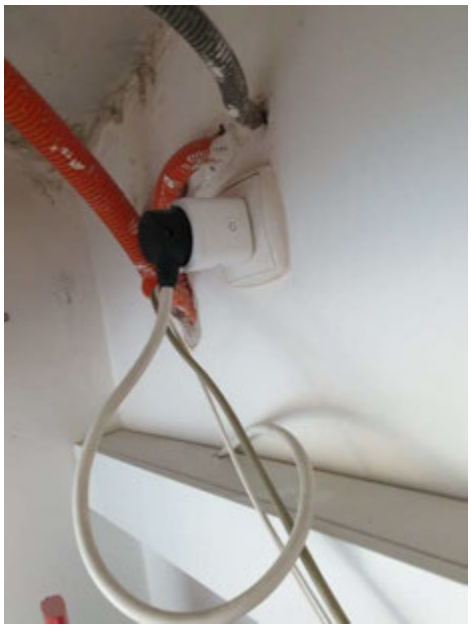


Figure 6. Wi-Fi smart plug XP1030 used to control and monitor AC power to the air-conditioner

Control of the air conditioner is achieved by a simple WWW user interface (Figure 8) implemented in PHP on Raspberry Pi, that runs corresponding IR slinger commands.



Figure 7. Screencapture of Android application to switch AC power on/off and monitor AC power to the air-conditioner



Figure 8. Screen capture of simple Web GUI to control the air-conditioner

Finally, we can monitor how temperature changes over time using ThingSpeak [8] visualizations of stored temperature and humidity data (Figure 9 and Figure 10). Each sensor node sends its measurements to ThingSpeak cloud over HTTP.

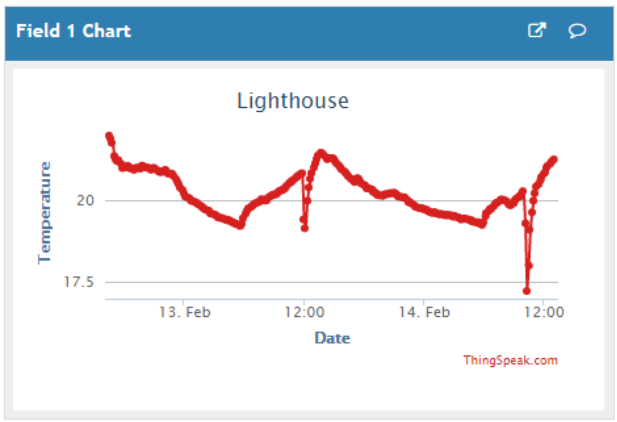


Figure 9. Change of temperature inside the appartment over last 48 hours as presented by ThingSpeak

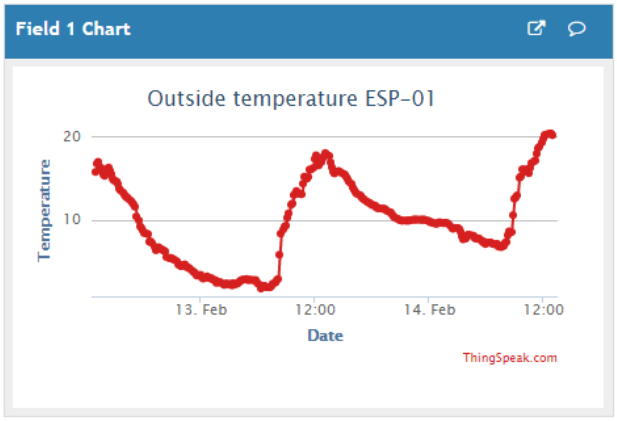


Figure 10. Change of temperature outside the appartment over last 48 hours as presented by ThingSpeak

III. EVALUATION OF THE COURSE BY TEACHERS

In total 13 teachers, 7 from primary schools, 8 from secondary schools, participated in the course over the two years in 2018 and 2019. Teachers were given source code templates to accelerate the development of required Lua code.

Tables below presents the evaluation of the course by the teachers regarding: a) content and presented methods (Table I), b) applicability of the presented approach (Table II).

TABLE I. SURVEY OF CONTENT AND METHODS OF WORK

No	Sample size (N = 13, Grade scale 1 – 10)	
	Statement	***Average Grade***
1.	Program introduces new and fresh ideas	8,84
2.	Previous experience of participants were considered	9,00
3.	I will be able to use new knowledege in practice	8,84
4.	Course materials are of apropriate quality	8,69
	Average grade for content and methods	8,84

TABLE II. SURVEY OF APPLICABILITY OF THE EDUCATION PROGRAM

No.	Sample size (N = 13, Grade scale 1 – 10)	
	Statement	***Average Grade***
1.	With the program I gained new theorethical knowledege	8,69
2.	The program motivated me to introduce changes in my work	8,31
3.	The program ofered practical solutions for work	8,77
4.	The program opens new questions and encourages continuing education in the field	8,08
	Average grade for applicability	8,46

The results show a strong viability of our approach to use creation of useful IoT devices as an introduction to computer science and engineering.

IV. CONLUSIONS

The approach presented in the paper was well received by teachers. The use case of smart apartment introduces all basic IoT technologies and components. Also, the cost of all components is very low and thus suitable for implementation in classes and workshops. The ESP8266 WiFi modules running NodeMCU proved to be reliable with uptimes of several months. The Raspberry Pi running Linux as a central hub also displayed remarkable robustness with good uptimes.

However, there are also a few drawbacks when using ESP8266. The first one is relatively large energy consumption due to WiFi transmission, and this need special consideration when using battery power. The second drawback is that is its rather hard to use HTTPS in NodeMCU Lua since it requires a significant amount of RAM to establish a HTTPS session. We might alleviate this problem by using compiled Arduino language instead of Lua interpreter of NodeMCU or using ESP32

REFERENCES

[1] T. Odun-Ayo and T. Obafemi-Ajayi, "Building the K-12 engineering pipeline: An assessment of where we stand," in *Proceedings - Frontiers in Education Conference, FIE*, 2017, vol. 2017-October, pp. 1–9.

[2] A. Mavroudi, M. Divitini, F. Gianni, S. Mora, and D. R. Kvittem, "Designing IoT applications in lower secondary schools," in *IEEE Global Engineering Education Conference, EDUCON*, 2018, vol. 2018-April, pp. 1120–1126.

[3] J. R. Byrne, K. O'Sullivan, and K. Sullivan, "An IoT and Wearable Technology Hackathon for Promoting Careers in Computer Science," *IEEE Trans. Educ.*, vol. 60, no. 1, pp. 50–58, Feb. 2017.

[4] D. Glaroudis, A. Iossifides, N. Spyropoulou, I. D. Zaharakis, and A. D. Kameas, "STEM Learning and Career Orientation via IoT Hands-on Activities in Secondary Education," in *2019 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2019*, 2019, pp. 480–485.

[5] J. Voas and P. Laplante, "Curriculum considerations for the internet of things," *Computer (Long. Beach. Calif).*, vol. 50, no. 1, pp. 72–75, Jan. 2017.

[6] A. Jaklič, "IoT as an Introduction to Computer Science and Engineering: A Case for NodeMCU in STEM-C Education," in *2020 IEEE Global Engineering Education Conference (EDUCON)*, 2020, pp. 91–95.

[7] "GitHub - bschwind/ir-slinger: A small C library for sending infrared packets on the Raspberry Pi." [Online]. Available: https://github.com/bschwind/ir-slinger. [Accessed: 14-Feb-2020].

[8] "IoT Analytics - ThingSpeak Internet of Things." [Online]. Available: https://thingspeak.com/. [Accessed: 09-Dec-2019].

An Algorithm for Assessment of Students Using Gamification

D. Keremedchiev[*] [**], D. Borissova[**] [***] and G. Tuparov[*]
[*] New Bulgarian University/Department of Informatics, Sofia, Bulgaria
[**] Bulgarian Academy of Sciences/Institute of Information and Communication Technologies, Sofia, Bulgaria
[**] University of Library Studies and Information Technologies, Sofia, Bulgaria
delyan.keremedchiev@gmail.com

***Abstract* - The use of game design and elements in non-gaming context known as Gamification presents one of the newest ways to improve traditional assessment. There are various tools for self and final assessment of students in the modern e-learning systems. In this paper, we propose a grading algorithm composed of two parts: self-assessment and final exam. In the self-assessment phase, students must solve tests and fill a bar (cover scores) according to their results. The tests are generated using the "Intelligent system for generation and evaluation of e-learning tests using integer programming". This approach allows students to choose between different levels of question scores and different number of questions in the generated test. The required score for filling the bar and successful self-assessment can be obtained either with a small number of questions with higher score or a large number with a lower score. Only after passing successfully all self-assessment tests students are allowed to do final exam.**

Keywords - component; formatting; style; styling; insert (key words)

I. INTRODUCTION

With the development of information and communication technologies, ease of internet access and the rise of internet-based applications many traditional activities go online. This refers to the learning processes in the modern e-learning systems as well. The assessment is a common way to give a feedback to students about the acquired knowledge. Assessments can be done in many ways. Various tools for self and final assessment exist. A popular method of assessment is the test. Tests can consist of different test elements depending on the complexity and purpose of the test. Many of those elements can be automatically checked and evaluated. Creating a test using those elements will give the students the opportunity to get an instant feedback about the acquired result. There are researches which conclude that the students who use tests for self- assessment have better performance compared with those who do not use them [1]. The result is 85.7% vs. 77.8% for the students who have used the tests. The use of game design and elements in non-gaming context, formally named gamification, as well as serious games has become very popular in the last few years. Combining these approaches can lead to qualitative leap in the efficiency of these processes. In recent studies like [2] and [3] the results from peer and self-assessment in the Moodle e-learning environment prove that gamification in e-learning provides better learning experience and better learning environment. Gamification can also be applied to assessment as a tool for instant feedback. The game-based quizzes have recently become more interesting and popular. Modern web browsers supporting HTML5, CSS3 and JavaScript can be used as an environment for e-learning tasks, assessments, quizzes and various interactive elements including gamification. The browsers with such advantages are available on all modern computer and mobile operating systems.

In this paper we propose a grading algorithm for student assessment, which combines several gaming elements. The algorithm includes "Intelligent system for generation and evaluation of e-learning tests using integer programming" [4] as a main part in the process of test assembly. This approach allows the students to control the level of complexity of the tests, the number of questions and question weights. The use of game elements will make the assessment more attractive. Ranking elements will provoke the students to join the race and work harder to obtain higher results. The algorithm itself is implemented in a web-based system for student SQL knowledge and skill assessment.

II. RELATED WORKS

An increased interest in gamification research in scientific databases has been noticed in the last few years. A literature searches in SCOPUS database which indexes relevant databases such as ACM, IEEE, Springer, etc. with the query: TITLE-ABS-KEY for the word [gamif*] is made in [5]. Some results of this paper analysis are shown in Table I.

Apart from empirical researches different algorithms for gamification exist in education. The authors of [6], [7], [8] propose different combinations of technologies and gamification elements in e-learning and report successful impact on students. According to [9] the gamification techniques considered the most potentially effective in an educational setting are: "badges (appealing to masterminds), progress bars and leaderboards (appealing to conquerors), a storyline and a visual (appealing to the seekers)" [9].

The study is funded by Bulgarian National Scientific Research Fund, Contract DN-05/10, 2016

TABLE I. PAPERS STUDIED IN THE EMPERICAL RESEARCH (EXTRACTED FROM [5])

Achievement/progression	Count
Points, score, XP	67
Challenges, quests, missions, tasks, clear goals	53
Badges, achievements, medals, trophies	47
Leaderboards, ranking	47
Levels	35
Quizzes, questions	25
Progress, status bars, skill trees	19
Performance status, performance feedback	18
Timer, speed	13

With the usage of a combination of progress bars and leaderboards the impact of the proposed algorithm will grow. Progress bars allow the students to keep track of their points and to track and display the overall goal progression. They are used as a display mechanism to motivate students who are close to achieving the goals. In other implementations of gamificated educational systems students are usually graded with badges for reaching a certain level. Badges indicate that the covered score is in a certain known scope. Using scores will add more precise tracking of the students' knowledge. Progress bars indicate students' overall progress. Leaderboards allow students to view their achievement compared to others. Due to privacy needs there must be an option to make the leaderboard anonymous. The names of other participants must be hidden in these cases and the only information shown must be the summary for the ranking.

There are automated grading tools for SQL knowledge and skills like SQL-Lab [10], aSQLg [11] and MySQL Sandbox for Processing SQL Statements [12]. These tools give an opportunity to test students for their SQL skills. Queries are executed and the results of the execution are compared to the correct answer. Select queries are the most common queries used. Queries for modifying the database, tables, indexes and data are more complicated to manage. Allowing students to write and test queries for modifying the structure and data can lead to security vulnerability, cheating options and decrease of the system stability, even system crash. Using preconditioned questions will reduce the number of topics that can be covered by the test. Tools like "Web-based system for automatic evaluation of SQL queries" [13] can generate a question bank in an exchangeable format for use in other learning and assessment systems. These tools give an opportunity to combine questions and database statements. With the use of these combinations queries can be run against different database statements and the correct answer depends on the current state.

III. GRADING ALGORITHM

An original algorithm for the assessment of students using gamification is presented in Figure 1. The algorithm starts with a test on the very first topic. The next and all the other topics are hidden/inaccessible until the minimum test score requirement for the previous topic is covered. The test is generated using the "Intelligent system for generation and evaluation of e-learning tests using integer programming" [4]. Students can choose between different levels of complexity for the test and different number of questions with different weights. The test can be solved as

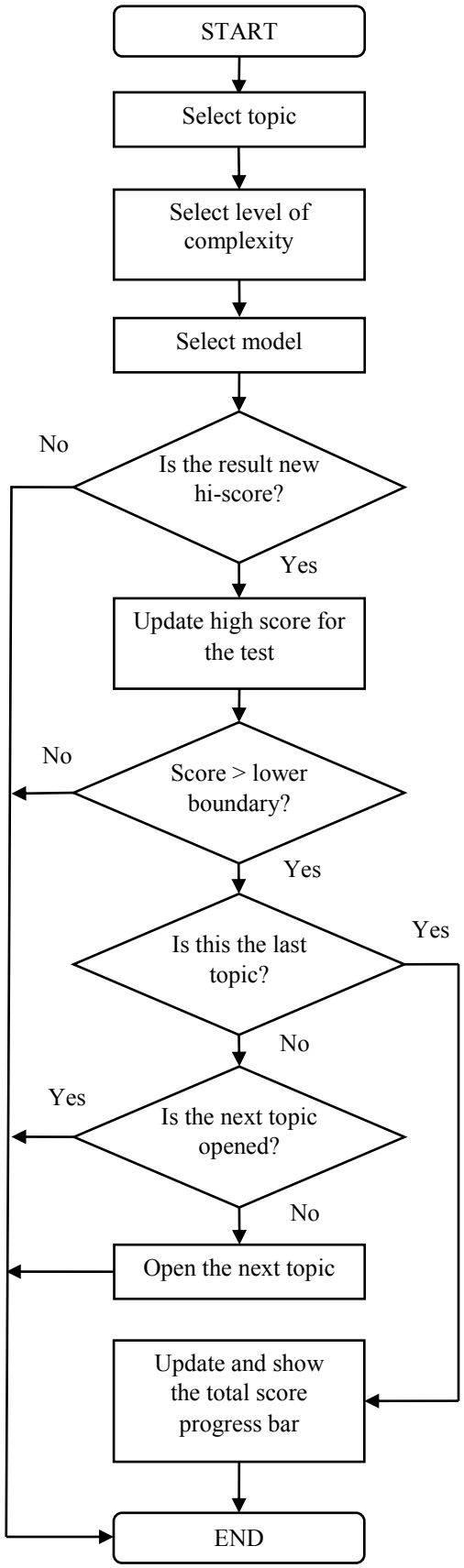


Figure 1. Algorithm for the assessment of students using gamification

many times as necessary. At every attempt, students can change their preferences for the test generation options. When the minimal score set for the test is covered, the next test opens and is available for solving. The final result for every test is the best result obtained. After successful solving all the tests, students' total score is calculated as a sum of the results obtained from all the tests. Students can do the final exam when the total score, calculated as the sum of all scores is over the minimum set.

For the implementation of the algorithm, the following gamification and graphical elements are chosen:

- Fill the progress bar – the score of every test is presented as an empty progress bar. With the solving of the test students obtain scores. The obtained score is visualized as filling the progress bar. Once the student covers the minimal amount of points needed for the test completion, the progress bar is thoroughly full and the next stage is accessible. After all stages are completed an additional progress bar shows up. This additional progress bar contains the sum of all the scores obtained by the student for the different topics during the tests minus the points already used for progress bar filling.
- Leaderboard and ranking – students can see their position in the ranking compared to the results of the other students. This leaderboard can help the students see their rank in the community. The ranking includes position, list of participants and their score. Ranking can be made anonymous due to privacy requirements and in this case only the student's own position, score and points insufficient for the student to reach the leading position will be shown. Leaderboard and ranking is created and updated for every single score and for the total score obtained.

IV. IMPLEMENTATION

A classical web-based system implementing algorithm is created. "Assessing SQL knowledge and skills" is chosen as a main topic. Questions are created using "Web-based system for automatic evaluation of SQL queries" [13]. The topics for the tests are:

- T1: Single table query;
- T2: Multiple table query (JOINS);
- T3: Filtering data;
- T4: Grouping and aggregating data;
- T5: Functions and stored procedures;
- T6: Sub-queries;

Every question has score points set by the teacher. The test scores are set in the range [0-200]. Test generation is made using "Intelligent system for generation and evaluation of e-learning tests using integer programming" [4]. The minimal score for successful test achievement is 50% or 100 points. In the test score only the correct answers are calculated. The test score must be adjusted into marks. The marks used for the course grade are set in the range Poor (2) - Excellent (6). Scores less than 50% are considered as failure and the mark is Poor (2). For successful are counted all results with 100 or more points. There are no limitations for the complexity ranges in the test generation algorithm and the complexity levels can be as many as needed. A range of points and their adjustment into marks is shown in Table II:

The use of "Intelligent system for generation and evaluation of e-learning tests using integer programming" algorithm for test generation gives the students the opportunity to choose among different levels of complexity of the test. The interface of a test generation system can be seen on Figure 2.

TABLE II. POINT TO MARK AJUSTMENT

Complexity level	Mark	Obtained points
L1	Excellent (6)	175-200
L2	Very Good (5)	150-174
L3	Good (4)	125-149
L4	Satisfactory (3)	100-124
-	Poor (2)	0-99

The chosen intelligent system for the test generation emphasizes on the optimal selection of questions for the generation of tests with different levels of complexity. As described in [4] a predefined set of questions for the test generation is needed. Those questions must have different weights and points for a proper test generation. This optimization problem is solved with the use of integer programming. When the student chooses a L1 test the upper boundary of the generated test will be set to 200 and

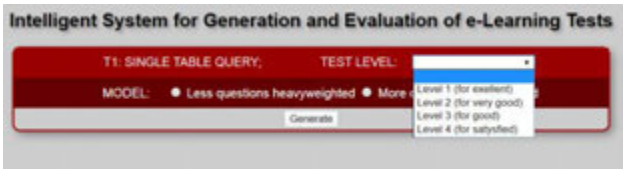


Figure 2. Interface for test generator

the lower boundary will be set to 175. Changing the number of questions and their weight preference also influences the sum of maximum points which can be obtained from the test. The authors of the algorithm have noticed that choosing the "more questions but light-weighted" will result in a generated test with a sum of points closer to the lower boundary. While choosing the "less but heavily-weighted questions" option results in a generated test with a sum of points closer to the upper boundary. In Table III the approximate scores of the generated tests are shown:

TABLE III. GENERATED TEST SCORES

	T1	T2	T3	T4	T5	T6
L1 less questions	192	198	195	195	196	197
L1 more questions	178	177	180	178	177	179
L2 less questions	168	170	170	171	173	172
L2 more questions	155	152	154	155	151	152
L3 less questions	145	145	148	146	149	144
L3 more questions	130	128	125	128	127	127
L4 less questions	120	119	122	119	121	124
L4 more questions	105	103	106	106	101	104

As can be seen in Table III the approximated results have points within the boundaries but never equal to them. Moreover, a different test generation can result in different points for the test. This randomness will lead to more attempts to get the best result possible. After a test is completed, an animation which fills the progress bar with the obtained result is started. The progress bar width is set to the minimum of 100 points for opening the next topic required. The result can be seen as shown on Figure 3:

After the successful completion of all stages the additional progress bar shows up. This progress bar is larger than the previous and requires more points. Initially this progress bar is filled up with the obtained points minus the points already used. The filled sections are the same color as the colors of the completed tests.



Figure 3. Filled bars and results

Furthermore, there are labels showing which test points are used to fill the section. This solution helps the student to easily find out which topics have more and which have less points. When the whole additional progress bar is filled the student can have access to the final exam. Using a progress bar composed by parts previously completed with the sum of the results obtained is more sensitive and robust compared to badge-based algorithms, where the requirements are such as "Obtain at least 3 bronze and 3 silver badges". Badge-based algorithms set the lower boundary and do not estimate results higher than this boundary. In score-based algorithms all acquired points are calculated in the final result. These algorithms will motivate students who have not completed the assessment bars, but their total result is close to the minimum requirements. Moreover, using a requirement such as "Total score of XXX points" is very clear. Translated to badges it may be presented as: "3 silver and 3 bronze badges or 1 gold, 1 silver and 4 bronze badges or 1 platinum and 5 bronze badges, etc." Figure 4 shows the partially filled final progress bar:

The detailed leaderboard per test is just on a click and it shows the student's result compared to all obtained results for the test. Figure 5 shows the leaderboard interface:

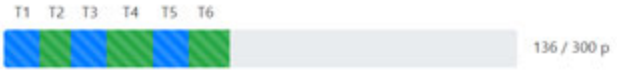


Figure 4. Final progress bar

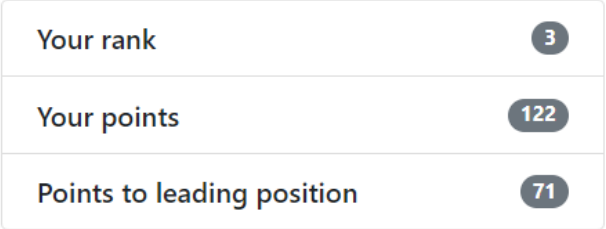


Figure 5. Leaderboard and ranking interface

The anonymous ranking interface is seen on Figure 6.

The front-end of the proposed system and all visual elements such as progress bars, rankings and leaderboards are created using Bootstrap. Bootstrap is a front-end framework initially developed by Twitter and one the most popular open source frameworks in the world [14]. In Bootstrap 4 there are important features such as

#	First	Last	Points
1	Student	One	192
2	Student	Two	187
3	Current	Student	122

Figure 6. Leaderboard

responsive functionality and mobile first approach. Using flexbox as a main design technique is modern browser oriented. Bootstrap has a predefined set of icons, badges, crates, accordions and many other graphical elements ready to use. Using Bootstrap will lead to rapid web development and faster creation of ready-to-use front-end interfaces.

Every completed test can be started again either with a different set of options or with the same one. The mixing of different levels of complexity and the number of questions according to their weight can produce many different tests. It can be seen on Table III that the highest scores for a test can be obtained with the "More questions but heavily-weighted" option. Another feature of the test generation algorithm is the randomness provided. This randomness can cause variations in the maximum test result. In such cases full-score solved tests by different students can obtain different results. This may produce additional activity from the best students and they will do the tests again and again until they obtain the maximum

possible result. This technique and user behavior can be seen in social gaming platforms, where users do the same level many times to beat their friends high-score and to lead the ranking [15].

V. Conclusion and Future Work

VI. Implementation

The current article describes an algorithm for the assessment of students using gamification. The selection of a different level of complexity and a different number of questions with different question weights by the students gives them the feeling of full control of the process. Using "Intelligent system for generation and evaluation of e-learning tests using integer programming" algorithm for question selection in the test assembly process adds an additional randomizing element to the possible maximum test score obtainable. This will result in additional activities by the best students to obtain the highest possible result. Using gamification elements and animations for the interface makes the implementation of the algorithm more sophisticated and with better appearance.

The proposed algorithm can be easily implemented as a classical web-based application as seen in IV. The algorithm is flexible and can be set with a wide range of options. A scheme for points-to-mark adjustment according to test complexity levels can be included during the setup. The modern attention-grabbing techniques and interactive interfaces add additional impact to the assessment process. The algorithm is based on scores and with this approach every single point matters. Progress bars will help the students to track closer their knowledge and skills.

It will be interesting to compare the results and students' behavior obtained using this algorithm with the results obtained by badge-based algorithms. Data such as number of tries, combination of selected options, acquired points for the initial progress bars, tests that are re-taken after the total progress bar shows up and number of tries after leaderboard checks can be collected and examined. These student activities can be recorded to xAPI [16] server for further analysis.

Creating a mathematical model for completing the tasks with a minimal number of tries and solving the model can also be an interesting problem. The maximum possible test result variations and the acquired points according to the selected test generation options and student knowledge may vary in a larger range.

Acknowledgment

The study is funded by Bulgarian National Scientific Research Fund, Contract DN-05/10, 2016.

References

[1] Stefanie Hennig, Christine E. Staatz, Jacqueline A. Bond, Debbie Leung, Judith Singleton, Quizzing for success: Evaluation of the impact of feedback quizzes on the experiences and academic performance of undergraduate students in two clinical pharmacokinetics courses, Currents in Pharmacy Teaching and Learning, Volume 11, Issue 7, 2019,pp 742-749

[2] Tuparov G., Keremedchiev D., Tuparova D., Stoyanova M.: Gamification and educational computer games in open source learning management systems as a part of assessment. In: 17th International Conference Information Technology Based Higher Education and Training (ITHET), pp. 1-5. IEEE, Olhao, Portugal, (2018).

[3] Paunova-Hubenova E., Terzieva V., Dimitrov S., Boneva Y.: Integration of game-based teaching in Bulgarian schools State of art. In Proc: 12th European Conference on Game-based Learning ECGBL 2018, Sophia Antipolis, France, 4-5 October, pp.516-525 (2018)

[4] Borissova D., Keremedchiev D. (2020) Intelligent System for Generation and Evaluation of e-Learning Tests Using Integer Programming. In: Simian D., Stoica L. (eds) Modelling and Development of Intelligent Systems. MDIS 2019. Communications in Computer and Information Science, vol 1126. Springer, Cham, pp. 97-110

[5] Majuri, J., Koivisto, J., & Hamari, J. (2018). Gamification of education and learning: A review of empirical literature. In 2nd International GamiFIN Conference, Pori, Finland (pp. 21-23).

[6] Manal M. Alhammad, Ana M. Moreno, Gamification in software engineering education: A systematic mapping, Journal of Systems and Software, Volume 141, 2018, pp 131-150

[7] Talingdan J., Llanda C., Assessment of the effectiveness of learning theories using gamifie dandroid app in teaching C programming, IOP Conference Series: aterials Science and Engineering 482 (2019) 012030

[8] Aparicio M., Oliveira T., Bacao F., Painho M., Gamification: A key determinant of massive open online course (MOOC) success, Information & Management, Volume 56, Issue 1, 2019, pp 39-54

[9] S. O'Donovan, J. E. Gain, and P. Marais, "A case study in the gamification of a university-level games development course.," in SAICSIT, 2013, pp. 242–251.

[10] P. Brusilovsky, S. Sosnovsky, M. V. Yudelson, D. H. Lee, V. Zadorozhny, and X. Zhou, "Learning SQL Programming with Interactive Tools: From Integration to Personalization," Trans. Comput. Educ., vol. 9, pp. 1-15, 2010.

[11] C. Kleiner, C. Tebbe, and F. Heine, "Automated grading and tutoring of SQL statements to improve student learning," in Proceedings of the 13th Koli Calling International Conference on Computing Education Research, Koli, Finland, 2013, pp. 161-168

[12] K. Atchariyachanvanich, S. Nalintippayawong, and T. Permpool, "Development of a MySQL Sandbox for processing SQL statements: Case of DML and DDL statements," in 2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2017, pp. 1-6.

[13] Tuparov G., Stanchev M. (2019) Assessing students' SQL knowledge and skills in an e-learning environment. In: Computer Science and Education on Computer Science (CSECS 2019), pp.77-79

[14] About Bootstrap https://getbootstrap.com/docs/4.5/about/overview/ checked on 01.08.2020

[15] Hedegaard M., Kerschbamer R.,Müller D., Tyran J., Distributional Preferences Explain Individual Behavior Across Games and Time (May 9, 2019).

[16] Rustici Software, xAPI Overview, https://xapi.com/overview/, last visit on 01.02.2020

Effective Decision Making: the Added Value of Including Humanities in STEM Studies

Marcela Pavlica*, Tihana Babić* and Petra Cuculić*
*Algebra University College, Zagreb, Croatia
marcela.pavlica@algebra.hr, tihana.babic@algebra.hr, petra.cuclic@algebra.hr

Abstract - **The rapid ascension of STEM programs in education came accordingly with the needs of the labour market. Although technology made its way into our everyday lives, it is difficult to find a balance between the empowerment it brought by making our everyday tasks easier and our dependency on it that came with it. In search of that balance, there is a vast need in the contemporary digital revolution for computer engineers to be able to make responsible political, ethical and social decisions while contributing to the global economy and the future of the world. Besides technical skills they learn as a part of their curriculum, STEM students and experts in the field need decision making skills for their personal and professional development. The importance of implementing courses from the field of humanities in education has been recognized by educators and employers. The aim of this research is to determine whether the students themselves are aware of the added value such courses could bring to them which will be presented in this paper on the example of Algebra University College students. The research was conducted through a voluntary Google Forms survey comprised of 10 questions that have shown that students are not fully aware of the added value of including humanities in STEM studies, however, they do believe that learning about effective decision making during their studies would be of benefit to their professional development.**

Keywords – *humanities, effective decision making, added value, STEM students, soft skills*

I. INTRODUCTION

The high numbers in Internet usage are just one out of many arguments that prove not only how dependent we are on technology, but also explain the reason of a large increase in STEM field study programs and job vacancies. However, the switch to STEM focused education and employment has brought to question the place of humanities in the higher education curriculum. On the topic of humanities vs. STEM it is always our concern how to best educate ourselves, our children and our students to be able to acclimate to the current ever-changing world of technology. Moreover, it is on us to be able to make correct and effective decisions in our future lives and careers and share that knowledge to those who will join the demanding and fast changing labour market. Technical skills that STEM students are taught in their education provide them with the knowledge to become trained professionals in the field of their interest and apply that knowledge to their job positions. Nevertheless, that is not a guarantee that a person will succeed. The job market is highly competitive, and students require a well-rounded education in order to become attractive to employers and be successful in their field. Making decisions in their professional path will become an everyday challenge for the future IT experts and it is up to educators to prepare them to be able to bring responsible decisions with positive outcomes. The lack of non-technical skills in STEM students has been widely recognized by employers, however, the challenge is to determine whether the students themselves see the added value of learning non-technical skills as a part of their curriculum.

II. EFFECTIVE DECISION MAKING: TOP SKILLS FOR SUCCESS

A. Labour market needs

Many researchers have already explored the importance of soft skills in the 21st century by analyzing current labour market trends that show a vast need for university graduates to bring more than the technical skills acquired during their studies. For instance, an employer survey by the OECD concluded that job specific skills are no longer enough for graduates to meet the needs of the current labour market [1]. Furthermore, a research made by the American Management Association showed that "soft skills" such as critical thinking, decision making and problem solving, creativity and innovation, collaboration and communication are becoming increasingly important in the current global economy [2]. Moreover, the Wall Street Journal survey in 2015 reported that out of nearly 900 executives last year, 92% said soft skills were equally important or more important than technical skills. However, 89% said they have a very or somewhat difficult time finding people with the necessary competences [3].

STEM degrees have increased in numbers in the past decade, arguably due to the needs of the current labour market that lacks professionals skilled in STEM fields. A report by the European Parliament on overview of the labour market situation in STEM occupations showed that STEM skilled labour is increasing and is expected to grow even more. The number of STEM university graduates has increased in 15 EU Member States and in entire Europe since the mid-2000s. Employment in STEM-related sectors is predicted to grow by 6.5 % between 2013 and 2025, from which a growth of 8 % is expected in computing sector [4]. Pursuing a degree or a position in the STEM field requires technical skills and knowledge and universities have been focusing on

checking off boxes that equip their students with the right set of skills for that field. However, schools, businesses, and community leaders have recently started to realize that technical skills alone are not sufficient for the competitive and fast-paced STEM field. People with outstanding technical skills may be able to deliver in a familiar environment, but it is not enough for them to become top of their field. Whether they are trying to move up or are required to be on a cross-functional team, soft skills are needed by every STEM professional to compliment the hard skills.

B. Soft and people skills

The terms "hard" and "soft" skills often come up in the employment industry as vital requirements of the current labour market. Hard skills, or technical skills, are defined as "those skills acquired through training and education or learned on the job and are specific to each work setting, such as programming languages, operating system skills, networks and communications, etc. [5]. On the other hand, "soft skills" are defined as "the cluster of personality traits, social graces, language skills, friendliness, and optimism that mark each one of us to varying degrees" [6]. Non-technical, or "soft skills" are often recognized in the industry by other terms, such as "people skills", "emotional skills", "employment skills" etc. The Higher Education Council of Australia defines soft skills as "the skills, personal attributes and values which should be acquired by all graduates, regardless of their discipline or field of study" [7]. More abstractly, some researchers point out that soft skills are interpersonal, human, people or behavioural skills needed to apply technical skills and knowledge in the workplace [8]. The Assessment and Teaching of 21st Century Skills (ATC 21) organization has presented a framework for organizing different types of 21st century skills. It includes four classes of soft and hard skills:

1. Ways of Thinking, which encompasses creativity and innovation; critical thinking, problem solving, and decision-making; and metacognition or learning to learn
2. Ways of Working, which includes communication and collaboration or teamwork
3. Tools for Working, which addresses information literacy and information and communication technology (ICT) literacy
4. Living in the World, which includes citizenship, life and career skills, and personal and social responsibility [9].

Conversely from "hard skills" that are job specific, soft skills are transferable between learning and work settings. For example, a Software Engineer can transfer his problem solving and decision making skills from his or her role in one industry to the other. However, the person would need to adapt those skills to the current work setting. The problem solving and decision making skills stay the same, but the context in which they are applied is different, so when developing these skills for students it is important for the teaching process to be diverse, so that the transferring of the "soft skills" across different contexts would be easier.

C. Effective decision making

As one of those skills, effective decision making (often in research synonymous with problem solving) became a crucial skill for students to develop in order to become a better skilled worker for the current job market. We can define effective decision making as a process that facilitates our intellectual, moral and spiritual decision-making and helps us determine the simplest and most efficient way to make the right decision.

The decision making process is an important skill for employers as it saves them time and money by hiring individuals with this skill as they can effectively plan and organize their team by following a structured process instead of relying solely on their intuition without taking the time to analyze the possible outcomes of their decision. University of Massachusetts divided this process into 7 steps:

1. IDENTIFYING THE DECISION – Defining the nature of the decision
2. GATHERING RELEVANT INFORMATION – Collecting information from the best sources available depending on the decision
3. IDENTIFIYING ALTERNATIVES – Identifying several possible paths of action and listing all possible alternatives
4. WEIGHING THE EVIDENCE – Drawing on information collected and imagining of each alternative listed. Through this step you can place the alternatives in the order of priority, based upon your own value system
5. CHOOSING AMONG ALTERNATIVES – selecting the alternative that is the best one for you, or a combination of alternatives.
6. TAKING ACTION – Starting to take some positive action and implementing the alternative chosen in step 5.
7. REVIEWING YOUR DECISION & ITS CONSEQUENCES – considering the results of your decision and evaluating whether it resolved the problem you identified in step one [10].

Following this process helps us make a decision that is more thought out than just relying on our instincts and assuming that a certain decision will bring positive results. Instead, a structured process can help an individual and in the end the business to experience positive outcomes from making a right decision. However, as many other "soft skills", effective decision making has to be acquired by students during their university education so they can finish their studies fully prepared for what the current market demands. The only way to accomplish such goal is to include subjects from various academic disciplines into the STEM curriculum.

III. THE ADDED VALUE OF HUMANITIES

The Stanford University Humanities Center defines humanities as "the study of how people process and document the human experience," going on to explain that "since humans have been able, we have used philosophy, literature, religion, art, music, history and language to understand and record our world" [11]. The word humanities derives from the Latin word humanus (humanity), which comes from the word homo (human). The humanities deal with man, morality and ethics, man's thirst for knowledge, love of wisdom, culture, languages, history and art. Therefore, courses in the humanities concern, for example, anthropology, philosophy, ethics, theology, history, art history, and the science of art.

There are many layers to humanities, and they are more than merely a study of human culture. Humanities help us think critically and creatively, think about "why" of the problem and connect the dots to get an answer to the "why" question. Deborah K. Fitzgerald, a professor of History of Technology at MIT stated that by including humanities in their studies, students learn that "most human situations defy a single correct answer, that life itself is rarely, if ever, as precise as a math problem, as clear as an elegant equation" [12]. Moreover, she continues, students "gain historical and cultural perspectives, and critical thinking skills that help them collaborate with people across the globe, as well as communication skills that enable them to listen, explain, and inspire".

In education, humanities enhance our moral beliefs and often make us question the arguments and situations beyond our own understanding. These skills are equally important for STEM students. In order to be truly innovative, scientists must consider how their work will affect the world and its inhabitants [13].

In the past, humanities were deemed as pillars of our society and were principal in higher education. Philosophy, ethics, literature and languages dominated the higher education industry and education in humanities was what created a Renaissance man. However, the shift to a technology focused world and STEM focused education has challenged the position of humanities in higher education. Nevertheless, there is a reason why humanities reigned the world of higher education before the appearance of technology and why they should still be included in the curriculum of STEM students. As humanities are a broad term and subjects in the field are numerous, the higher education industry has to create a suitable interdisciplinary study of humanities in cooperation with the job industry to make sure students receive education from the field of humanities that will add value to their education and create a complete, job market ready graduate. Arguably, there is no need for a Software or System Engineer to necessarily study in depth all subjects that make up humanities. There is no added value to his studies if the universities fill up the already crowded STEM curriculum with courses such as poetry, archeology or history of arts. The added value that is referred to in this context is the addition, or the benefit a student will have from a certain course in the field of humanities for his further professional or personal development. In addition, if we examine all of the disciplines that comprise the field of humanities, on the surface, we will find very few that add value to the future career of a Software Engineer. So how can we determine which humanities disciplines add value to a STEM curriculum?

By staying on top of current trends in the employment industry. Universities have a great responsibility to respond to the criticisms of the employment industry that students are not equipped with the skills of the 21st century, that is, the above-mentioned "soft skills". Even though seemingly not a discipline in humanities, soft skills originate from all of the disciplines in the field of humanities. The soft skills that emerged from humanities can help give us answers on some of the most complex questions and problems in science and technology. Studies within humanities teach students to evaluate complex and sometimes flawed or fragmentary information by skeptically and carefully considering evidence and considering more than one side of every question. By comparing information with different points of view and using critical and creative thinking, students are taught to make a subjective and effective decision, which is one of the key skills of the 21st century job market. To be able to make the right decision for your problem, you need to use other "soft skills" that are a part of the decision making process. Such skills do not come to an individual naturally, and by creating a curriculum that will allow students to gain these humanities based skills we are preparing them for success.

IV. STEM STUDIES CURRICULUM: FUTURE CHALLENGES

STEM studies curriculum will surely have to adapt to the dissatisfaction of the job industry with the level of soft skills STEM students bring to the industry. STEAM was in its way a response to the demands of the industry and has successfully tackled one of many challenges the higher education sector faced. Adding Arts to the already stone set Science, Technology, Engineering, and Mathematics world opened up an array of possibilities for students interested in those four disciplines to acquire skills from the field of arts. Nevertheless, there are more challenges that universities will face, some of which are:

1. LACK OF RESEARCH – the current state of STEM curriculum in regards to the representation of humanities is important to determine, as it is the foundation to understanding and responding to other challenges that the higher education industry will have to surpass. Not much research has been made on the statistical representation of non-technical courses in the STEM curriculum. In order to establish a well-rounded curriculum, the universities will require past research to help them with the process.

2. MOTIVATING STUDENTS – for some students it is difficult to see the added value of something that they do not fully understand and that is not a requirement of the profession. It will be challenging to motivate students to see the importance of developing their effective decision making skills for their personal and professional development.
3. INCLUDING HUMANITIES IN THE CURRICULUM – with the curriculum already being extensive, it will be demanding to find a form and space for the humanities skills. Universities can choose many options, such as the development of soft skills through stand-alone subjects or by embedding it in existing courses.
4. PLANNING ACTIVITIES – skills such as effective decision making require a dynamic and practical approach to be used to its full potential. That requires educators to be creative and think carefully about the activities they can present to students that will keep them motivated and help them learn the effective decision making process.
5. KEEPING TRACK OF THE INDUSTRY – the process of changing a curriculum of a study program does not happen over night. There are many factors to consider that differ from university to university and from country to country. Making those changes is an administrative and sometimes slow process so the challenge will be for universities to keep up with the demands of the labour market.

V. THE RESEARCH METHODOLOGY

A. Research Objectives and Methods

The main research objective was to examine and determine do Algebra University College STEM students realize the added vakue of effective decision making as the part of their curriculum.

The research was conducted through a voluntary Google Forms survey during the winter semester of the 2019/2020 academic year. The survey was anonymous and consisted of 10 questions, 4 of which refer to demographics. For 6 questions answers were predefined with the 5 degrees of importance according to the Likert scale. A quantitative method was used for the analysis of the research results and the results were processed by Microsoft Excel tool.

B. The Research Sample

The research was conducted among Algebra University College students in STEM studies.

The total number of population sample was 60, N = 60. The majority of participants are Software Engineering students (57 %), while other students who participated come from System Engineering (23%), Multimedia Computing (13%) and Game Development (7%) study programs.

The majority of participants were undergraduate students; 90 % of them, while 10 % of participants were graduate students. More that 80 % students were male students. The structure of all surveyed students according to the study program, level of study and gender is presented in Table.1

Study Program	**Software Engineering**	**System Engineering**	**Multimedia Computing**	**Game Development**
Number of students	34	14	8	4
Gender				
Male	32	9	6	4
Female	2	5	2	0
Study level				
Graduate	2	1	3	4
Undergraduate	32	13	5	0

Table 1. The structure of the participants by Study Program. N=60

C. The Research Results

The research objective was examined through the research questions. In order to answer the research objective, 3 research questions were asked:

1. *Do and to what extent Algebra University College STEM students realize the added value of including courses from the field of humanities in their STEM curriculum?*
2. *Do the STEM students think that soft skills will benefit them in their professional development?*
3. *Do the STEM students think that learning about effective decision making is important for their professional development?*

In order to answer the first research questions, the questions asked in the survey were as follows:

1. *Do students find that the courses from the humanities field are represented enough in their curriculum?*

The research results have shown that only 2% of the total number of surveyed students finds courses from the humanities field very important for their curriculum. Conversely, 37 % of students believe that such courses are not important. Distribution of the answers for the importance of humanities courses in the curriculum can be found in Chart 1.

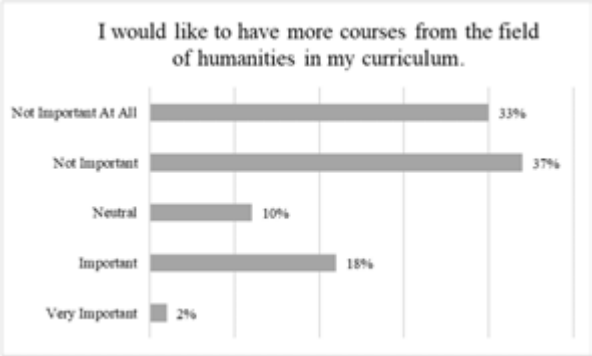


Chart 1. Distribution of the answers to the question: How important do you find the inclusion of courses from the field of humanities in your curriculum? N=60

2. *Do students wish they had more courses from the field of humanities in their curriculum?*

20% of the students expressed the wish to study more humanities based courses, while almost 40% of the students see no need in learning more humanities courses than they do. Furthermore, 10 % remained neutral. Chart 2. presents answers of participants to the above mentioned question.

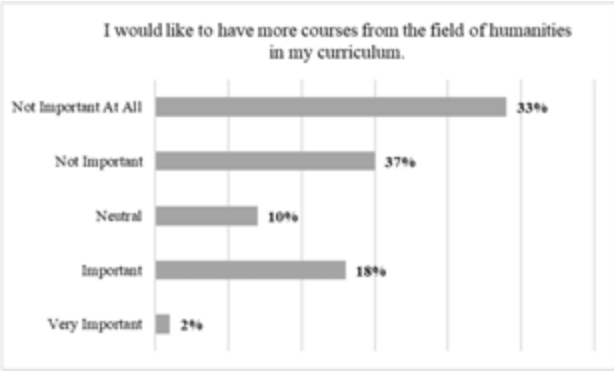


Chart 2. Distribution of the answers to the question: *Would you like to have more courses from the field of humanities in your curriculum? N=60*

In order to answer the second research question, the questions asked in the survey were as follows:

1. *How important do students find courses from the field of humanities for their professional development?*

For 18 % of participants the courses in humanities are important for their professional development, while 25 % of students believe they are not. Moreover, 22% of students remained neutral to this question and while 25 % of students find courses from humanities not important at all for their professional development, and 10 % very important. Distribution of the answers on whether the students consider courses in humanities relevant for their professional development is presented in Chart 4.



Chart 4. Distribution of the answers to the question: *How important do you find courses from the field of humanities for your personal development?*

2. *What factors do students find most valuable when looking for a job?*

Students were presented with four different skills and knowledges that could help them find a job: their degrees, technical skills and knowledges, soft skills and connections with the professionals in the field. Presented with the Likert scale of importance they had to choose how valuable they find each of those skills on the scale of 1 to 5 (1 = Not important At All, 2 =Not Important, 3 = Neutral, 4 = Important, and 5 = Very Important). Nearly 80% of students found technical skills and knowledges (in case of STEM studies that are skills such as alghoritms, data structures, working with databases, etc.) very important in finding a job and 35% of the students found soft skills very important for their job search. 5% of the students found that their degree and their soft skills are not relevant to their job search.

Important skills in job search:	***5***	***4***	***3***	***2***	***1***
Degree	30%	33%	17%	15%	5%
Technical skills and knowledges	77%	17%	5%	2%	0%
Soft Skills	35%	45%	10%	5%	5%
Connections from the field of interest	23%	22%	12%	3%	0%

Table 2. How important do you find the following when applying for a job? N=60

3. *What skills do students find the most important for their professional development?*

Students were presented with 9 different skills and asked to rate the skills from most to least important in regards to their professional development. Presented with the Likert scale of importance they had to choose how valuable they find each of those skills on the scale of 1 to 5 (1 = Not important At All, 2 =Not Important, 3 = Neutral, 4 = Important, and 5 = Very Important). Almost

half of the students rated all skills as very important. Communication skills were rated as very important with 27 students ranking them the highest. Teamwork, business communication and leadership skills followed with 45% of students rating them as very important. Conversely, conflict resolution skills, business ethics skills and business communication skills were rated not at all important by 5% of students.

Skills important for professional development:	***5***	***4***	***3***	***2***	***1***
Teamwork skills	45%	37%	12%	3%	3%
Communication skills	55%	35%	5%	3%	2%
Presentation skills	40%	33%	13%	12%	2%
Conflict Resolution skills	43%	30%	13%	8%	5%
Business ethics skills	37%	30%	15%	13%	5%
Negotiation skills	40%	42%	10%	5%	3%
Business communication skills	45%	30%	10%	10%	5%
Leadership skills	45%	30%	13%	8%	3%
People skills	40%	35%	15%	7%	3%

Table 3. How important do you find the following skills for your professional development?

In order to answer the send research questions, the questions asked in the survey were as follows:

1. *How important do students think learning about effective decision making would be for their professional development?*

The majority of the students found that learning about the process of effective decision making would be important for their professional development. 37 students found learning about effective decision making either very important, or important. 17% of students found that learning about effective decision making would not help their professional development.

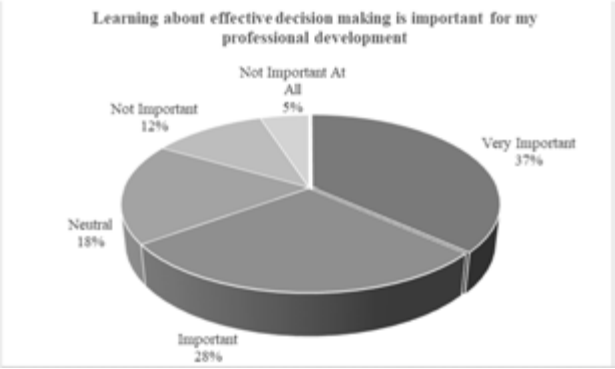


Chart 4. Distribution of the answers to the question: *How important do you think learning about effective decision making would be for your professional development?*

VI. CONCLUSION

The research has shown that students are not fully aware of the added value of including humanities in STEM studies, however, they believe that learning about effective decision making during their studies would be of benefit to their professional development. Effective decision making is an important process included in all aspects of humanities and soft skills, therefore students realize that it is a skills worthy of learning. Furthermore, there is not enough research on the number of courses from the field of humanities in the STEM curriculum, which makes it more difficult for researchers and educators to determine how much change is needed in the STEM curriculum in regards to humanities. Future generations of students will have to make important moral and intellectual decisions for their entire career and it is important to equip them with the necessary skills to do so. Even though the students may not realize the added value of humanities in their curriculum, it is up to employers to highlight the importance of soft skills for future graduates and, and it is up to educators to make sure that the curriculum is always up to date with the current labour market. Further research can be made on this topic, starting with analyzing the number of courses from the field of humanities in Higher Education institutions which will be a starting point for solving the challenges of STEM studies curriculum.

REFERENCES

[1] OECD, "OECD Skills Outlook 2013: First Results from the Survey of Adult Skills", 2013, Online: https://www.oecd.org/skills/piaac/Skills%20volume%201%20(eng)--full%20v12--eBook%20(04%2011%202013).pdf [Accesed: 02.2.2020]

[2] American Management Association (AMA), "Executives say the 21st century requires more skilled workers", 2010, Online:. https://www.amanet.org/assets/1/6/2012-critical-skills-survey.pdf [Pristupljeno: 03.2.2020]

[3] K. Davidson, "Employers Find Soft Skills Like Critical Thinking in Short Supply", 2016, Online: https://www.wsj.com/articles/employers-find-soft-skills-like-critical-thinking-in-shortsupply-1472549400 [Accesed: 03.2.2020]

[4] European Parliament, “Encouraging STEM Studies for the Labour Market”, 2015, Online: https://ec.europa.eu/eurostat/statistics-explained/index.php/Being_young_in_Europe_today_-_digital_world [Accesed: 28.1.2020]

[5] Snyder, L. A., Rupp, D. E. & Thornton, G. C., “Personnel Selection Of Information”, in J. J. Martocchio (Ed.), Research in personnel and human resources management: Vol. 25, Elsevier Science/JAI Press, 2006, pp 305–376

[6] Litecky, C.R. & Arnett, Kirk & Prabhakar, Bipin, “The paradox of soft skills versus technical skills in IS hiring”, Journal of Computer Information Systems. 45, 2004, pp 69-76.

[7] Treleaven, Lesley, and Ranjit Voola. “Integrating the Development of Graduate Attributes Through Constructive Alignment.” Journal of Marketing Education 30, no. 2, 2008, pp. 160–73.

[8] Weber, Melvin R., Dori A. Finley, Alleah Crawford, and David Rivera, “An Exploratory Study Identifying Soft Skill Competencies in Entry-Level Managers.” Tourism and Hospitality Research 9, no. 4, 2009, pp. 353–61

[9] Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., & Rumble, M, “Defining 21st century skills. Assessment and teaching of 21st century skills draft white paper.” The University of Melbourne, 2010

[10] University of Massachusetts, “Decision-making process”, Available on: https://www.umassd.edu/fycm/decision-making/process/, [Accesed: 28.1.2020]

[11] University of Stanford, “What are the humanities”, Available on: http://shc.stanford.edu/what-are-the-humanities, [Accesed: 29.1.2020]

[12] Association of American Colleges & Universities, “At MIT, the Humanities Are Just as Important as STEM”, available on: https://www.aacu.org/aacu-news/newsletter/mit-humanities-are-just-important-stem, [Accesed: 28.1.2020]

[13] E. Koprowski. “Five Reasons STEM Students Should Study Humanities.” Available on: https://www.masterstudies.com/article/Five-Reasons-STEM-Students-Should-Study-Humanities/, [Accesed: 3.2.2020]

Using BBC micro:bit in primary and secondary schools for creating simple smart home

P. Voštinár and J. Knežník
Matej Bel University, Department of Computer Science, Banská Bystrica, Slovakia
patrik.vostinar@umb.sk

***Abstract*— The concept of a smart home is getting more and more popular for people of all ages. You can meet the term smart home not only on the web, but also in schools. Students can solve their seminar works on smart home topics such as "Is my home smart" or "how will the future home looks like", etc. With popular educational board BBC micro:bit students can create their own simple smart home concept and practice programming thinking. This contribution deals with using BBC micro:bit in primary and secondary schools to create a simple smart home concept, which will use BBC micro:bit, crowtail STEAM edu kit for micro:bit, 1.8 inch colorful display and smart home kit.**

Keywords—BBC micro:bit, programming, teaching informatics, smart home

I. INTRODUCTION

Almost all countries in the world signalize that in IT sector are missing people for the jobs as a software engineer. Many researches show, that tech jobs will likely to keep growing in the next several years and they will be paid very well. According Kiersz research, job as an applications software developer will be the second-best job in the future [1].

Nowadays, computers are using in almost each type of job position - in the shops as cash desk, to controlling machines and production processes in factory, in the hospital for storing patient data, in the school for presenting interactive teaching materials, etc. Computers are also used to controlling devices in many households. These households are calling Smart Homes. Residents are using mobile devices e.g. to control heating, turn on/off lights or other devices such as coffee machine, etc. Smart Homes could also do automatic heating - turning on/off heating, draw the blinds down or up or controlling height of the water in pool.

Many jobs or devices have to be programmed by IT specialist. The question is, where can we find these people? Many researches focus on motivation of pupils to learn computer science.

The solution could be using motivational hardware, such as Lego EV3, Makecode robots, BBC micro:bit etc. Physical computing in cooperation with a block-based environment (e.g. Microsoft MakeCode) have become more and more attractive, especially for presenting programming to younger students – beginners [2].

BBC micro:bit could be one of the option, to motivate pupils learn computer science.

II. BBC MICRO:BIT

The educational board BBC micro:bit consists of 25 individually-programmable LEDs, 2 programmable buttons (A,B), physical connection pins, light, temperature and motion sensors, compass, battery socket, radio and Bluetooth antenna [3]. The main advantage of micro:bit is his very low price. Another huge advantage of micro:bit is the possibility to use it in primary and secondary schools, due to supporting programming languages like block-based language, JavaScript, Python and language C.

III. BBC MICRO:BIT EXPANSIONS

For teaching with BBC micro:bit it is possible to use more expansions like smart home kit, temperature sensors, crocodile cables, 1.8 display, special set of sensors, etc. Most of these extensions have low prices similar to micro:bit.

It is low-cost device, that was assigned for 11-12 years old pupil in the UK in 2016 [4]. Independent research tracked the progress of the BBC micro:bit in the UK during its first year of its deployment and found, that 85% of teachers who used this device agreed, that it makes Computer Science more enjoyable for their students, and 90% of the children who used it said, it helped to show them ,that anyone can code [5].

In comparison with more sophisticated piece of hardware (such as Arduino and Raspberry), BBC micro:bit is powerful even without any extensions [6].

A. ELECFREAKS Smart Home kit

Smart Home kit in fig. 1 shows simple room with various sensors in our small model of smart home.



Figure 1. Smart Home model

ELECFREAKS micro:bit Smart Home Kit contains executed components commonly used at home as TMP36 temperature sensor, sound sensor, light sensor, servo, DC motor, soil moisture sensor and OLED screen.

Fig. 2 shows the code, which is responsible for that our sensors are working. We have created two variables, to store sensors values in "forever" loop. Fan and light are triggered after crossing a certain value through calling a function further in loop.

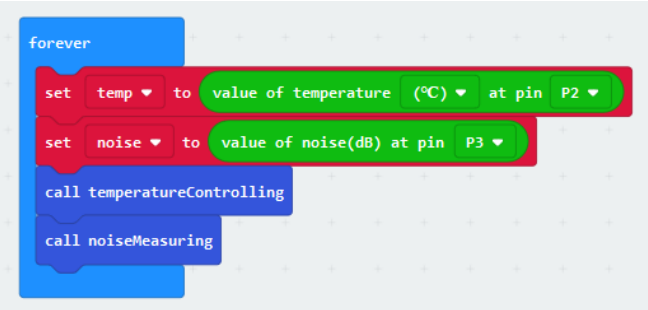


Figure 2. Reading values from sensors

Figure 3 shows the function which detects temperature values. If temperature is higher than 28C the fan located at pin1 is triggered for 5s. Noise measuring function detects the value of noise in room. If value is higher than 70 the light is triggered.

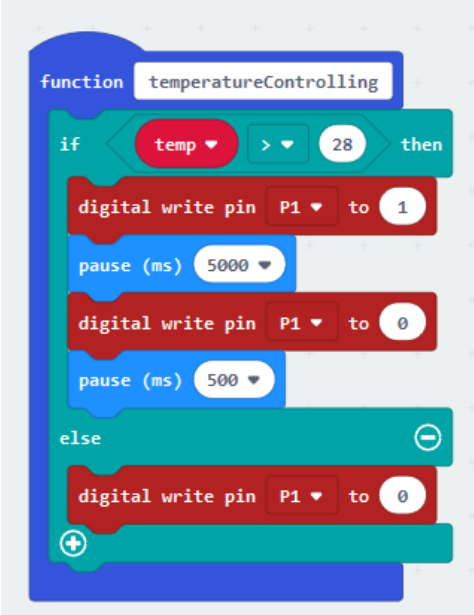


Figure 3. Temperature sensor code

Another interesting sensors we used from this micro:bit kit are submersible pump and soil moisture sensor (Fig. 4).

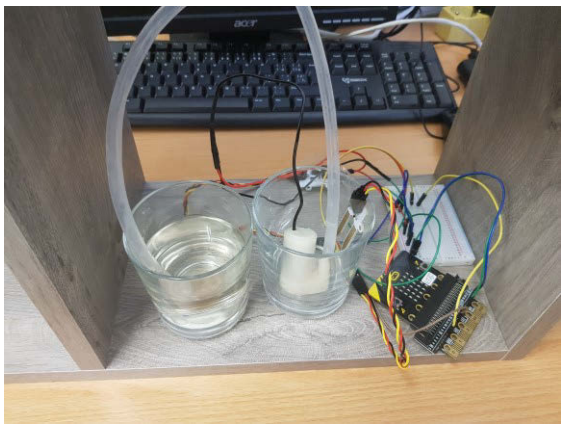


Figure 4. BBC micro:bit submersible pump and soil moisture sensor

If the height of water in glass (pool imitation) is higher than should be, the pump draws off the water. Moisture sensor detects the height of a water. Fig. 5 shows the code of this functionality.

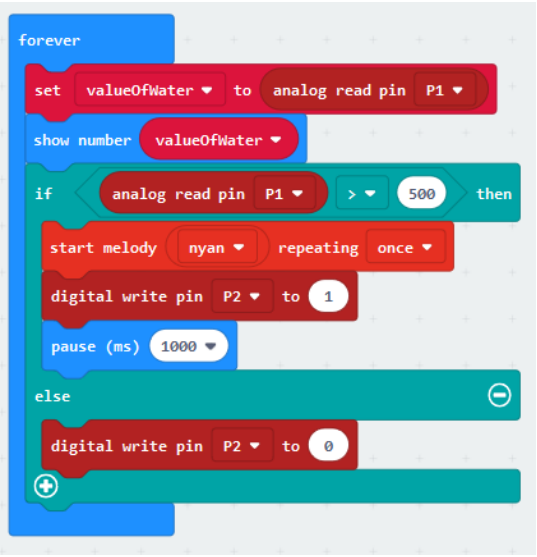


Figure 5. The code of submersible pumb and soil moisture sensor

B. 1.8inch display, crocodile cables, temperature sensor

1.8inch colorful display module consist of 160x128 pixels, capable of displaying 65K colors (Fig. 6).

We mounted this LCD display into a living room wall, looking like a TV set. The display shows a screen with the logo of the Department of Computer Science FNS UMB and a second screen with the text "Current temperature is" and the value that is received via radio communication with the Sensor Board every 5 seconds.



Figure 6. 1.8inch display in micro:bit room

Fig. 7 shows sensor Board for micro:bit, which allows to sense sound level, temperature and light level. These were connected easily to micro:bit with crocodile leads. Crocodile cables are suitable for quick connections to test circuit. Each lead is 50cm long with a rubber shrouded crocodile clip on each end.

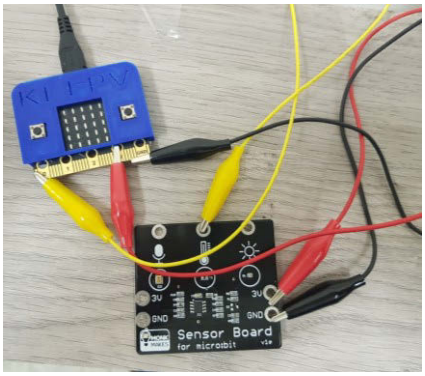


Figure 7. Sensor board with micro:bit

Fig. 8 shows code for measuring actual temperature in Fahrenheit degrees. We change this value to Celsius degrees.

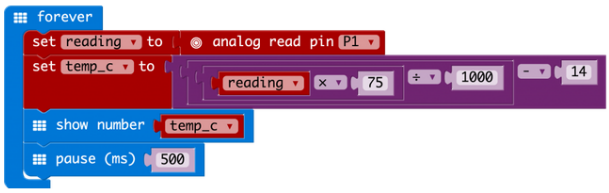


Figure 8. Sensor board – the code of reading temperature

C. *The Crowtail STEAM Edu Kit for Micro:bot*

The Crowtail STEAM Edu Kit for micro:bit contains useful equipment such as buzzer or twelve sensors e.g. LED, sound, noise sensor etc. which allows making interesting projects.

We used this kit for opening and closing window (in our case the window is cd case). The window is open by servo motor (Fig. 9).



Figure 9. Micro:bit opening window with servo

Fig. 10 shows source code of opening and closing window.



Figure 10. The simple code for opening window

Except servo motor we used also Led diods, noise detector, buzzer and touch sensor.

In one room (hall) is micro:bit kit, noise sensor is connected to this kit through PIN0, touch sensor through PIN01 and PINS 12 and 16 are reserved for LED diodes (green and red). As default there is in first room a red led turned on. This shows that the alarm is on. In this mode is alarm turned on in both rooms with detecting move in one room and noise in another. If we touch the touch sensor we turn of the alarm, by another touch we activated it again (Fig. 11).

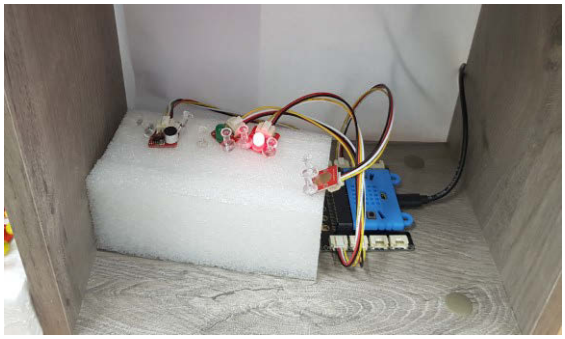


Figure 11. Micro:bit set for detecting sound

Fig. 12 shows the code, which set radio group (the same value should be in micro:bit in both rooms), turn on red LED light (PIN12), create variables for storing values from touch and sound sensors.

If alarm is on and noise in the room is detected, micro:bit sends:

- Value (number) 2 to second micro:bit (another room) while alarm is turned on and noise detected.
- If alarm is turned on, sends value 1 (noise sensor does not detect noise)
- If alarm is turned off, sends value 0.

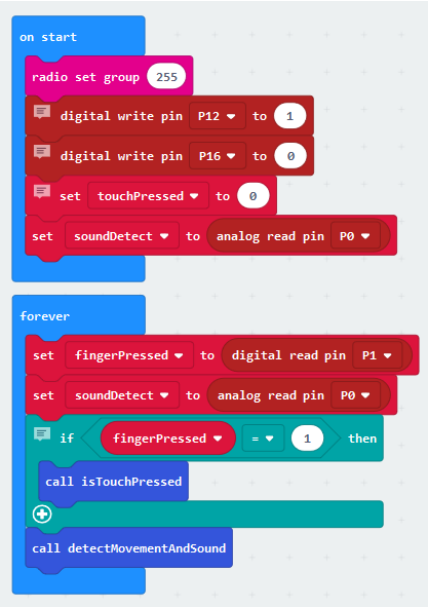


Figure 12. Micro:bit code for detecting touch sensor and sound noise

Fig. 13 shows function *isTouchPressed()*, which sends value if we touched the touch sensor.

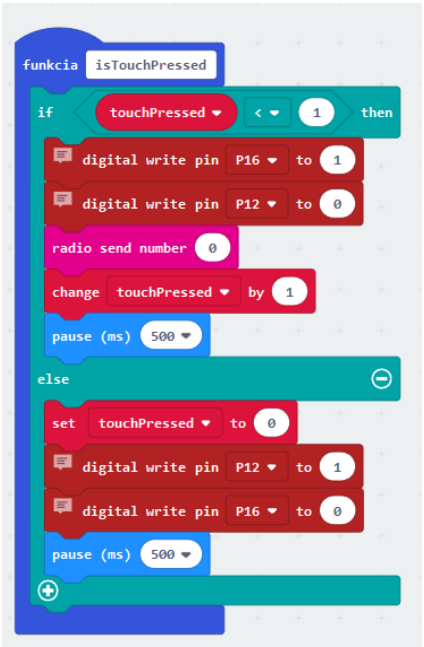


Figure 13. Source code of touch sensor

There is another microbit in second room, which recieves values from the first one. This one performs an action according to recieved value:

- 0 alarm is turned off , no action
- 1 alarm is turned on , move sensor has been detected so it triggers a buzzer and turns off/on diodes for 5s.
- 2 alarm is turned off , sound sensor has been detected, and perform same action as in previous case.

Fig. 14 shows second room with another micro:bit with sensors.

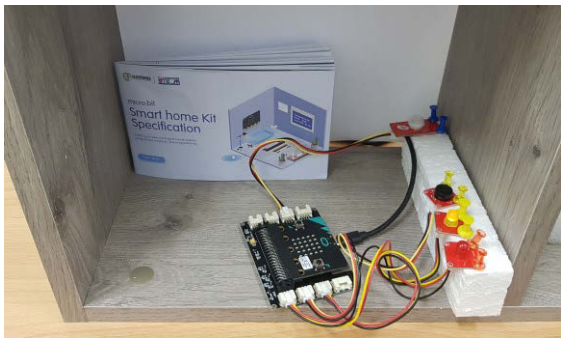


Figure 14. Second room with micro:bit sensors for detecting movement and turning on/off LEDs, buzzer.



Figure 15. Micro:bit code for second room

IV. Methodology and Results

We used BBC micro:bit expansions for teaching programming in our extracurricular activity for Computer Science. This activity is free of charge and it is suitable for primary school children (Fig. 16). We organize this activity twice a week for 1,5 hours. Pupils can choose the day of their attendance. Each time at least 7 pupils.



Figure 16. Extracurricular activity for Computer Science programming BBC micro:bit LEDs

Since we have only one piece of smart home kit, 3 sets of crowtail kit, we decided to make only one house, with two pupils working in each room.

Firstly, we used this expansion separately, later we shown to students, that they can create simple smart home concept with their knowledge from previous lessons. We measured the interest of each expansion from chapter 3 by questionnaire.

V. Conclusion

By this article, we wanted to show our simple smart home model, which is based on the educational board BBC micro:bit with various sensors such as crowtail STEAM edu kit for micro:bit, 1.8 inch colorful display and smart home kit. We built this house in our Extracurricular activity for Computer Science. Using educational board BBC micro:bit with sensors described in this article could increase motivation for teaching Computer Science, especially teaching programming.

Acknowledgment

This contribution has been processed as part of the grant project Interactive Applications for Teaching Mathematics at Primary Schools, project no. 003TTU-4/2018

References

[1] Business Insider. (2020). The 30 best high-paying jobs of the future. [online] Available at: https://www.businessinsider.com/best-jobs-future-growth-2019-3 [Accessed 14 Feb. 2020].

[2] K. H. Jin, H. Eglowstein, and M. Sabin. “Using Physical Computing Projects in Teaching Introductory Programming”, In Proceedings of the 19th Annual SIG Conference on Information Technology Education (SIGITE '18). ACM, 2018, New York, NY, USA, 155-155.

[3] M. Cápay and N. Klimová, “Engage Your Students via Physical Computing!,” *IEEE Global Engineering Educatin Conference (EDUCON)*, 2019, pp. 1250 – 1257.

[4] S. Sentance, J. Waite, L. Yeomans, and E. MacLeod, “Teaching with physical computing devices: the BBC micro:bit initiative”, In Proceedings of the 12th Workshop on Primary and Secondary Computing Education (WiPSCE '17), Erik Barendsen and Peter

[5] Hubwieser (Eds.). ACM, 2017, New York, NY, USA, 87-96.

M. Cápay, “Enhancing the teaching of informatics through engaging experience,“ *Proceedings of the 11th international conference on computer supported education (CSEDU 2019)*, 2019, pp. 453-460.

Using mBot robots for the motivation of studying computer science

P. Voštinár*
* Department of Computer Science, Matej Bel University, Banská Bystrica, Slovakia
patrik.vostinar@umb.sk

***Abstract* - Students in primary and secondary schools in Slovak republic are not very interested in study computer science, especially programming. At our department, we are trying to motivate primary and secondary school students to learn programming through multiple activities, such as computer science extracurricular activity and doing workshops in primary and secondary schools. This contribution deals with using Makeblock mBot robots for teaching programming in primary and secondary schools.**

Keywords – mBot; teaching informatics; programming;

I. INTRODUCTION

Due to the massive expansion of computers and mobile devices into almost all areas of everyday life, it is required to have more and more people skilled in computer thinking. Students use mobile devices and computers almost constantly, but they do not have a particular interest in studying programming. Therefore, it is useful to motivate them teaching programming since the primary school. Nowadays we can find enough great tools and teaching aids for programming, such as Scratch, Lego Mindstorms, Kodu, BBC micro:bit, Ozobot robots, etc. For teaching computer thinking it does not matter which tool or didactic aids we use. Almost all of them are appropriate for teaching programming. The problem could be low interest of students to learn programming.

Recently, an increasing number of educators have considered educational robotics as a promising field for applying the embodied cognition view [1]. Researchers and teachers are popularizing Computer Science and its practical usage in Slovakia as well [2-5].

Robotics for STEM (Science, Technology, Engineering and Mathematics) is an interdisciplinary emphasizing collaborative learning, and robots have been perceived as a potentially powerful vehicle, providing opportunities for construction, experimentation and collaboration between learners [6].

Ts. St. Georgiev from Ruse University did research about using MIT App Inventor programming environment at different levels of education. [7]. S.-W. Kim and Y. Lee have conducted a study [8] on the possibility of including App Inventor in the national curriculum for primary school students in South Korea. The results show, that the use of this program does not improve the ability of small group students to creatively solve problems and also that more time is needed for learning.

Identify applicable sponsor/s here. If no sponsors, delete this text box.

The solution could be using STEM robots and online tool App Inventor for increasing students' interest about teaching programming. The worldwide most popular STEM robots are Lego Mindstorms robots, but the main disadvantage of these robots is their price. The solution could be to use Makeblock robots which are significantly cheaper compare to Lego Mindstorms.

II. COMPUTER SCIENCE EXTRACURRICULAR ACTIVITY

We are using various types of STEM robots in our computer science extracurricular activity at Department of Computer Science FNS Matej Bel University in Banská Bystrica, Slovakia. This activity is focus on gaining additional teaching experience for our students from department and to teach programming to primary school students. Its free and its organized twice a week (different students on each group) since the academic year 2017/2018. This academic year, we have 27 students from primary schools and 20 students from our department (as teachers). Fig. 1 shows programming mBot in computer science extracurricular activity at our department.

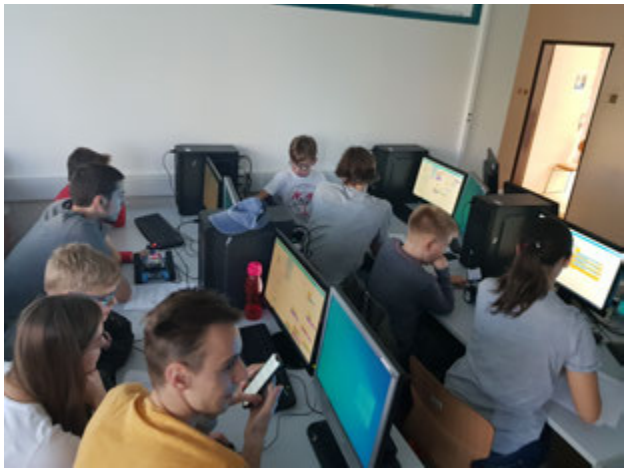


Figure 1. Computer science extracurricular activity - programming mBot

III. WORKSHOPS

To increase interest of learning programming we offer free workshops for primary and secondary schools in Banská Bystrica county. Each school could choose workshop topic (programming microcontrollers Arduino, programming mobile applications, programming robots, etc.).

Due to pandemic of covid-19 this year, we did only two workshops of programming robots topic. One of them was in the Gymnázium Jana Chalupku in Brezno and the other was in the Gymnázium Milana Rúfusa in Žiar nad Hronom. Both of these schools were eight-year secondary grammar school. In the Gymnázium Milana Rúfusa in Žiar nad Hronom we had two-day workshops.

In the teaching process we started with simple robots Phiro Pro and the main time we focus on using mBot robots. Fig. 2 shows workshop with mBot in Brezno.

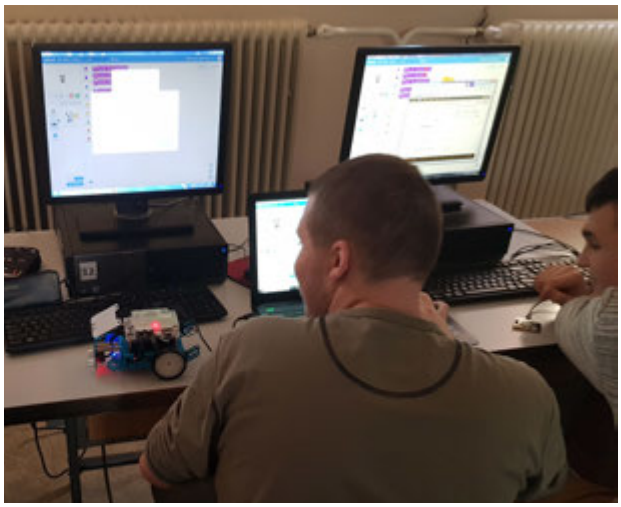


Figure 2. Gymnázium Jana Chalupku – workshop

Fig. 3 shows workshop with mBot in Žiar nad Hronom.



Figure 3. Gymnázium Milana Rúfusa - workshop

IV. MAKEBLOCK ROBOTS

Makeblock Co., Ltd, founded in 2013 is targeting the STEAM education and entertainment markets for schools and everyone who want to learn programming. Makeblock provides comprehensive hardware, software, content solutions, and top-notch robotics competitions, with the aim of achieving deep integration of technology and education.

The company offer various types of robots which could be programmed in the mBlock app (for computers[1]) or for mobile devices Makeblock app (iOS [2] and Android[3]).

We used for teaching mBot Ranger and mBot robots. Basic version of this robots is mBot, which is entry-level educational robot kit. mBot Ranger is educational robot, which is improved version of mBot, that comes with 3 preset forms, the tank-like Off-road Land Raider, the three-wheel racing car Dashing Raptor, and the Self-balancing Nervous Bird. Both of them contain Bluetooth module, line follower sensor, ultrasonic sensor, gyroscope, lights and mainboard contain microcontroller Arduino. mBot is cheaper version (99€) as mBot Ranger (150€). Fig. 4 contains robots: mBot Ranger and mBot.

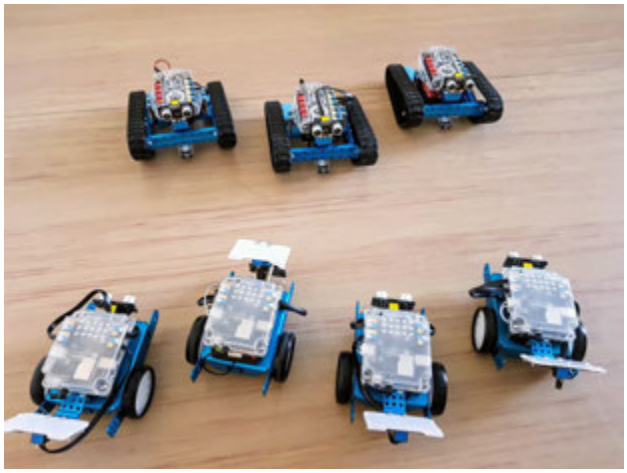


Figure 4. mBot robots

V. MBOT TASKS

We prepared the various types of tasks for mBot and mBot Ranger. The robots are almost the same with very small difference in functionality. Because of that, we can use them for the same tasks. All students already have some previous experience with Scratch block environment, these experiences were useful for them.

All tasks and teaching materials are available at our webpage[4].

In this paper we describe three types of tasks, students from extracurricular activities and workshops worked on as an example.

[1] https://www.mblock.cc/en-us/

[2] https://apps.apple.com/us/app/makeblock-play-stem-robots/id918804737

[3]https://play.google.com/store/apps/details?id=cc.makeblock.makeblock

[4] http://www.fpv.umb.sk/katedry/katedra-informatiky/informaticky-kruzok-umb-pre-ziakov-zs-a-ss/materialy-k-vyucbe-pre-zs.html

Activity name:	Programming mBot robots
Target group:	5.-9. year primary school
Expected time:	2 hours
Goal:	**Cognitive goal** - The student learns to program using visual programming. **Affective goal** - The student is able to respond correctly to regulations and rules.
Methods and forms:	• investigative and research methods, • methods of individual work of students, • narration, description, explanation, • question and answer method, • methods of practicing and improving skills.
Preparation, teaching aids:	**Materials:** Computer with internet access, robots mBot software: Off-line: Mblock (Windows), Makeblock (Android, iOS)
The course of the activity	

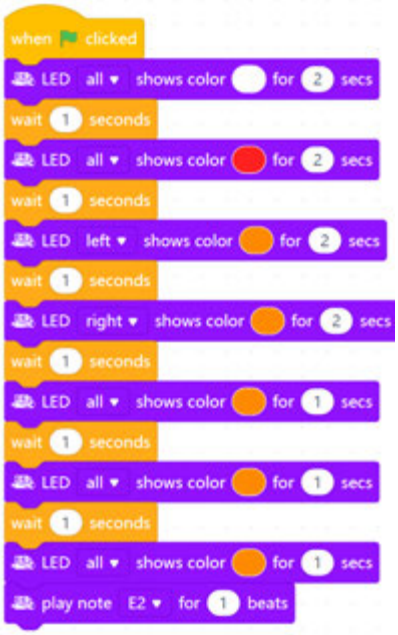
Task 1 - checking lights, turn signals, horn before driving

The first thing we should do, when we sit down in the car is to check the lights, turn signals, brakes, horn. It's similar with our mBot. This task is focus on the basic functionality of the car:

1. Turns on white lights (LED)
2. The brake lights come on for 1 sec - red
3. Turns on the left turn signal for 1 sec - orange
4. Turns on the right turn signal for 1 sec- orange
5. Starts the horn for 2 seconds

Students read the assignment and the teacher verifies, that students understand the assignment. Students work on tasks individually or in pairs. The teacher watches the students and if necessary, he explains the ambiguities.

Into the block "When the green flag clicked", we will gradually insert a commands for displaying lights.

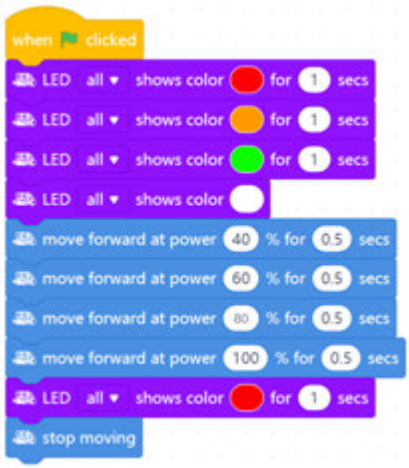


Extension options:

- set longer lights (e.g. 3 seconds)
- Add a command, that delayed the lights for 2 seconds (white light, 2 seconds wait, red light, 2 seconds wait, left turn signal, 2 seconds wait, right turn signal, 2 seconds wait, trumpet)
- Finally, add flashing turn signals. Program them, so that they will flashing (eg both at the same time).

Task 2

After pressing the button (green flag), the red light on the mBot lights up for 1 second, then the orange light for 1 second and finally the green light for 1 second. After the green light, the lights turn on - the white light (which will always be on) and the mBot will start first at a speed of 40% (0.5 seconds), then 60%, 80% and finally 100%. When the speed reaches 100%, mBot lights up the bridles (red light) and mBot stops.



Task 3 – cycles

Edit previous example # 3 to create a variable, that you initialize to 40. This variable will store the vehicle speed mBot. Modify the "move forward" calls, so that you only use this command once in the code (use a loop, don't forget to increase the value of the global variable by 20)



VI. METHODOLOGY AND RESEARCH

Our research was held in the academic year 2019/20 and it still lasts. In our research, we used didactic aids - robots for learning programming in two eight-year secondary grammar school in Brezno and Žiar nad Hronom (Slovak republic). The main aim of our research was to find out, whether we can increase students' interest to study programming by using didactic aids - robots. We used the questionnaire for measuring data of our research. We asked 41 students several questions in the questionnaire. The questionnaire filled 28 boys and 13 girls. The respondents were 7 and 8 class students (13-14 years old). Students were selected according to their interest (except extracurricular activity at our university).

First question after the gender selection was focused to whether they use didactic aids - robots before. All of them said, that they used Lego Mindstorms.

Question *How much did you enjoy programming mBot robots?* was aimed to measure interest of using mBots. On a scale from 1-5, they chose options - 1 meant *did not enjoy at all* and 5 *had a lot of fun.* Fig. 5 shows, that almost all of students chose the option "had a lot of fun".

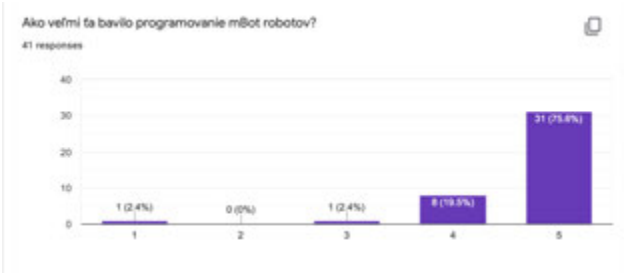


Figure 5. How much did you enjoy programming mBot robots?

Question *Were the activities difficult for you?* was aimed to complexity of tasks. Fig. 6 shows that, 41,5% students said no, 51,2% said "rather no" and only few of responders said "yes" or "rather yes"(6%).

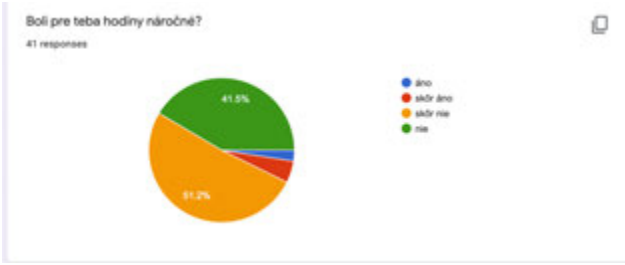


Figure 6. Were the activities difficult for you?

We asked same questions also our students in Computer science extracurricular activity at our university. The questionnaire filled 20 students, all of them were boys.

On the same question as in the secondary grammar school *How much did you enjoy programming mBot robots?* Fig.7 shows, that all of the students answered positively.

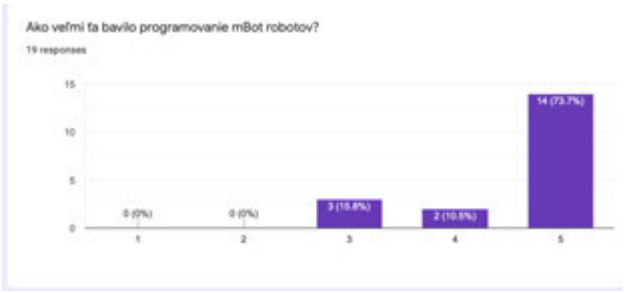


Figure 7. How much did you enjoy programming mBot robots?

35% of students of Computer science extracurricular activity answered question *Were the hours difficult for you?* "no" and 65% "rather no".

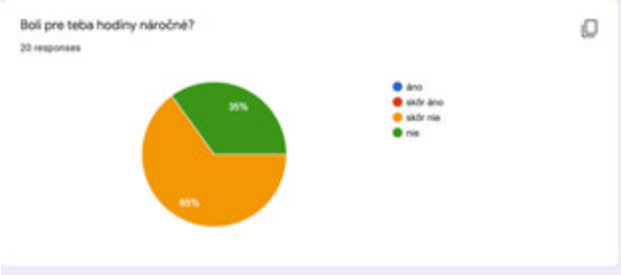


Figure 8. Were the hours difficult for you?

VII. Conlusion

In this paper we focused on using mBot robot for teaching programming in two secondary grammar schools and in the Computer science extracurricular activity. The main aim was to motivate pupils and teach them some programming basics. We received from almost all participants positive feedback. We described 3 tasks as an example of methodological sheets, students worked on.

Acknowledgment

This contribution has been processed as part of the grant project Interactive Applications for Teaching Mathematics at Primary Schools, project no. 003TTU-4/2018 and Implementation of Blended Learning into Preparation of Future Mathematics Teachers and Future Computer Science Teachers, project no. 001UMB-4/2020.

References

[1] A. Merkouris and K. Chorianopoulos, “Programming touch and full- body interaction with a remotely controlled robot in a secondary education STEM course,” Proceedings of the 22nd Pan-Hellenic Conference on Informatics - PCI 18, 2018.

[2] E. Klimeková, et al., “Case Study on Physical Computing with NodeMCU on Summer School,” International Conference on Robotics and Education RiE 2017, 2018.

[3] G. Lovászová, V. Palmárová, “Location-based games in informatics education,” International Conference on Informatics in Schools: Situation, Evolution, and Perspectives, 2013.

[4] P. Voštinár and M. Klimová, ”Experience with Using Robots for Teaching Programming,” IDT Proceedings, 2019.

[5] J. Fialová, “Tvorba a využitie aplikácií pre mobilné telefóny v školskej matematike,” EME Proceedings, 2018.

[6] D. Alimissis, “Educational robotics: Open questions and new challenges,” Themes in Science and Technology Education, 2013.

[7] T. St. Georgiev, "Students’ Viewpoint about Using MIT App Inventor in Education," 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 611-616.

[8] S.-W. Kim and Y. Lee, “A study of educational method using App Inventor for elementary computing education”, Journal of Theoretical and Applied Information Technology, Vol.95, No.18, 2017, pp.4376-4384.

Investment of European Enterprises in ICT Education of Their Employees: Relationship with Country Competitiveness and Digital Economy Development

M. Pejić Bach*, J. Zoroja * and I. Strugar*
* University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia
mpejic@efzg.hr, jzoroja@efzg.hr, istrugar@efzg.hr

Abstract— **Rapid growth and usage of different information and communication (ICT) tools influence of private and business activities. In order to sustain their position on the market, companies invest in their employees, especially in their lifelong education in the area of developing information skills. In this paper, the authors investigate the importance of ICT education in companies. The goal of the paper is to examine to what extent enterprises in European countries provide training in order to develop or upgrade the ICT skills of their employees. Data about enterprises that provide training regarding the development of ICT skills were collected from Eurostat for the year 2018. Data about Global Competitiveness Index (GCI) and its subindexes and data about the Digital Economy and Society Index (DESI) index have also been collected for the year 2018 in order to conduct comparative analysis. Results indicate that there is a positive trend regarding training for developing ICT skills, however, mostly in developed European countries. Although enterprises in developing countries do not substantially lag behind, their investment in this area is evidently lower.**

Keywords: information literacy and information skills, Global Competitiveness Index (GCI), Digital Economy and Society Index (DESI), enterprises, K-means, the digital divide

I. INTRODUCTION

Education is crucially important for companies achieving convergence with a dynamic business environment and ICT changes [1; 2]. Digital transformation becomes a new business mantra and quality and expected skills and knowledge of employees become one of the important issues in achieving this goal [3; 4]. Internet and e-commerce increase internationalization and widens markets, increase fierce competition of the existing markets, and opens for competition even the closest markets and almost every previously isolated local business niche [5; 6]. This development pressures the government of different countries, not only to nourish existing but to reestablish and conclude new international trade agreements and alliances with different countries.

On the other side, strong demographic changes appear in the workforce market as well [7]. The refugee crisis, baby boom generation retirement process and new generation Z entering on the workforce market is just a few important factors that require attention and fast reaction not only from companies but from government and especially from EU Countries as well [8; 9].

The baby boom generation, evaluated as a highly educated generation, with great business skills and experience are retrieving from the business arena and leaves an important gap in the workforce market [10]. All these processes require fast reactions from the companies and government as well in the form of continuous education.

The main goal of the paper is to investigate to what extent enterprises in European countries provide training for their employees regarding improving ICT skills and knowledge. We also tried to examine the relationship between country competitiveness and ICT education in enterprises as well as the relationship between digital economy development and ICT education in enterprises. We focused on small, medium and large enterprises European countries aiming to detect ICT skills of employees. Two main contributions of this paper could be highlighted: (i) homogenous groups of European countries are identified that have similar characteristics regarding offering ICT training for their employees; (ii) positive relationship between the digital economy development and ICT education in medium enterprises was confirmed as well as a positive relationship between country competitiveness and ICT education in medium enterprises.

The paper consists of eight sections. After the introduction, in the second section of the paper, we presented the importance of ICT usage in education for individuals and for enterprises regarding e-learning. In the third section of the paper, the methodology is described, while a descriptive analysis of the data and cluster analysis are presented in the fourth and in the fifth section of the paper. In the sixth section relationship between country competitiveness and ICT education in enterprises is presented, while in the seventh section relationship between digital economy development and ICT education in enterprises is described. The conclusion outlines the limitations and future research directions.

II. ICT EDUCATION IN ENTERPRISES

Data showed that in 2015 the Member States has spent over €716 billion on education which is around 5% (4.9%) of the EU's GDP [11].

In European Union, Denmark, Sweden and Belgium have the highest share of expenditure on education. In 2015, the ratio to GDP of government expenditure for education for 18 of the 28 Member States was 5% or more [12]. Average expenditure on education in 2015 for European Union countries was €1405 per inhabitant with leading countries: Luxembourg (€4685), Denmark (€3368) and Sweden (€2977) [6].

However, it has to be noted that any enterprise-wide change like digital transformation cannot succeed unless [13]: (i) changing mindset of the organization culture; (ii) acquiring new skills and knowledge by employees; (iii) including digital into organization processes.

III. METHODOLOGY

A. Data sources

Data were collected from Eurostat where were presented results from the research "Community Survey on ICT Usage and E-Commerce in Enterprises" [14]. Our focus in this paper is on the following question: "Did your enterprise provide any type of training to develop ICT related skills of the persons employed, during 2017?".

The respondents were ICT managers or senior professionals in the ICT department. We focus to different size of enterprises: all, small, medium, and large, defined by Eurostat: (i) All enterprises (10 persons employed or more), %; (ii) small enterprises (10-49 persons employed), without financial sector, %; (iii) medium enterprises (50-249 persons employed), without financial sector, %; and (iv) large enterprises (250 and more persons employed), without financial sector, % (Table 1). Research variables used in the analysis are for the year 2018 and refer to the % of small, medium and large enterprises that provide training to their personnel to develop their ICT skills.

TABLE 1. RESEARCH VARIABLES, EUROPEAN COUNTRIES, 2018

Variable name	Statistical units
Enterprise provided training to their personnel to develop their ICT skills	
ALL_2018	All enterprises (10 persons employed or more), %
SMALL_2018	Small enterprises (10-49 persons employed), without financial sector, %
MEDIUM_2018	Medium enterprises (50-249 persons employed), without financial sector, %
LARGE_2018	Large enterprises (250 and more persons employed), without financial sector, %

Source: Authors, Eurostat (2018)

B. Statistical analysis

Statistical analysis is conducted in three steps: (i) descriptive statistics for all the variables (including all investigated European countries; (ii) k-means cluster analysis; (iii) coefficient correlation of GCI index and DESI index across the clusters.

IV. DESCRIPTIVE ANALYSIS

Table 2 presents the results of the descriptive analysis regarding indicators of ICT education in small, medium and large enterprises in European countries for the year 2018.

The highest mean value (60,00) has medium enterprises which implicit that ICT education is used the most in medium enterprises. Small (38,00) and large (41,00) enterprises have lower medium value regarding ICT education in enterprises. Box plot diagram of variables measuring the ICT education in enterprises (small, medium and large) confirms presented mean values (the highest mean value is for medium enterprises) (Figure 1).

TABLE 2. DESCRIPTIVE STATISTICS OF ICT EDUCATION IN ENTERPRISES, EUROPAN COUNTRIES, 2018

	N	Min	Max	Mean	Std. Dev.
Enterprise provided training to their personnel to develop their ICT skills					
ALL_2018	29	22.38	5.00	42.00	8.89
SMALL_2018	29	17.83	4.00	38.00	8.09
MEDIUM_2018	29	39.03	7.00	60.00	13.44
LARGE_2018	29	21.03	4.00	41.00	8.73

Source: Authors, Eurostat (2018)

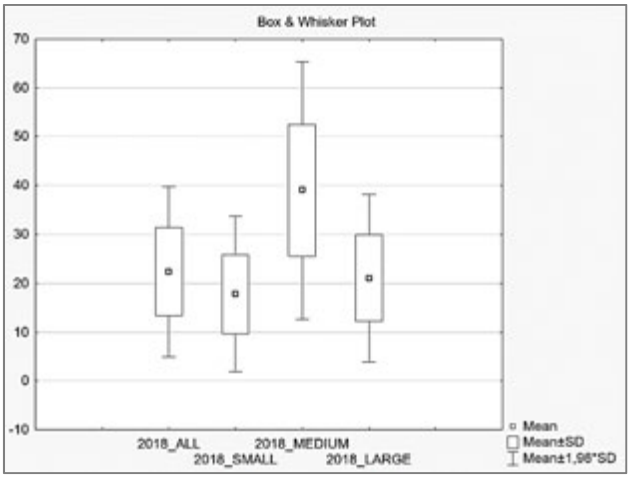


Figure 1. Box plot diagram of variables measuring the ICT education in enterprises; Source: Authors, Eurostat (2018)

Figure 2 presents a graph for the variable ICT education in small, medium and large enterprises and in all enterprises in selected European countries.

The graph showed that medium enterprises in almost every selected European country foster the most ICT education of their employees and countries like Belgium, Norway and Finland are leaders. The situation is similar for small and large enterprises while they have quite similar values and lower than medium enterprises regarding developing ICT skills of their employees.

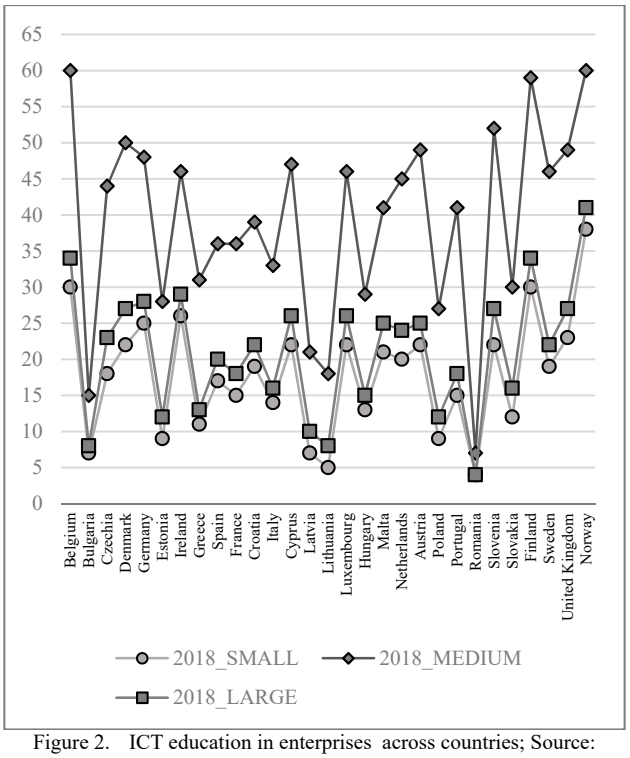


Figure 2. ICT education in enterprises across countries; Source: Authors, Eurostat (2018)

V. CLUSTER ANALYSIS

In this analysis, we used the K-means clustering algorithm to the variables regarding the development of ICT skills of employees in 31 European countries. Statistica Data Miner was used to conduct cluster analysis and applying the linear transformation, all variables were normalized. Furthermore, the graph of cost sequence presenting the error function for the various number of clusters was used (Figure 3) aiming to detect the optimal number of clusters which is five.

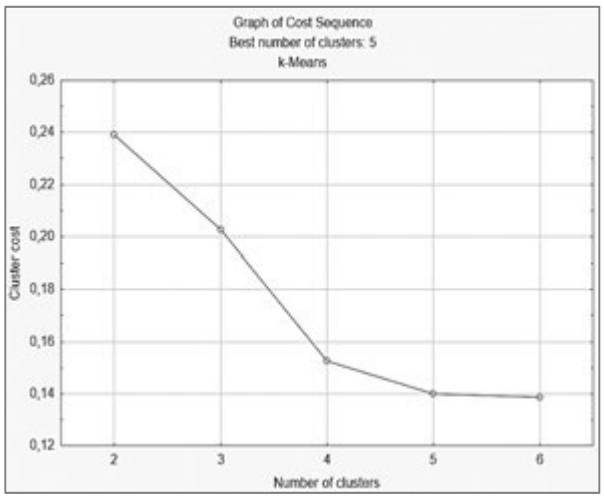


Figure 3. Cost sequence graph; Source: Authors, Eurostat (2018)

Table 3 presents the results of ANOVA analysis for the variables regarding training for developing ICT skills used for developing cluster solution, which indicate that all the variables have statistically significant different values across clusters. Therefore, the decision to use the five-cluster solution is justified.

TABLE 3. ANOVA ANALYSIS OF VARIABLES USED IN K-MEANS CLUSTER ANALYSIS

	df	F	p-value
Enterprise provided training to their personnel to develop their ICT skills			
ALL_2018	24	93.379	0.000***
SMALL_2018	24	77.762	0.000***
MEDIUM_2018	24	67.394	0.000***
LARGE_2018	24	83.923	0.000***

Source: Authors
Note: ***Statistically significant at 5%

TABLE 4. COUNTRY MEMBERSHIP ACROSS CLUSTERS

Cluster	Countries
Cluster 1	Belgium, Finland
Cluster 2	Bulgaria, Latvia, Lithuania, Romania
Cluster 3	Czechia, Denmark, Germany, Ireland, Croatia, Cyprus, Luxembourg, Malta, Netherlands, Austria, Slovenia, Sweden, United Kingdom
Cluster 4	Estonia, Greece, Spain, France, Italy, Hungary, Poland, Portugal, Slovakia
Cluster 5	Norway

Source: Authors

Table 4, as well as Figure 4, presents the countries, which are members of a specific cluster.

It can be noted that the first cluster comprises two of the most developed north European countries especially in the area of ICT (Belgium and Finland). The second cluster includes four countries which are geographically and economically similar (Bulgaria, Latvia, Lithuania, Romania). The third (Czechia, Denmark, Germany, Ireland, Croatia, Cyprus, Luxembourg, Malta, Netherlands, Austria, Slovenia, Sweden, United Kingdom) and the fourth (Estonia, Greece, Spain, France, Italy, Hungary, Poland, Portugal, Slovakia) cluster are more heterogeneous according to the economic development and geographical position, while cluster 5 contains only one country (Norway), which was often the outlier according to the highest level of developing ICT employees' skills.

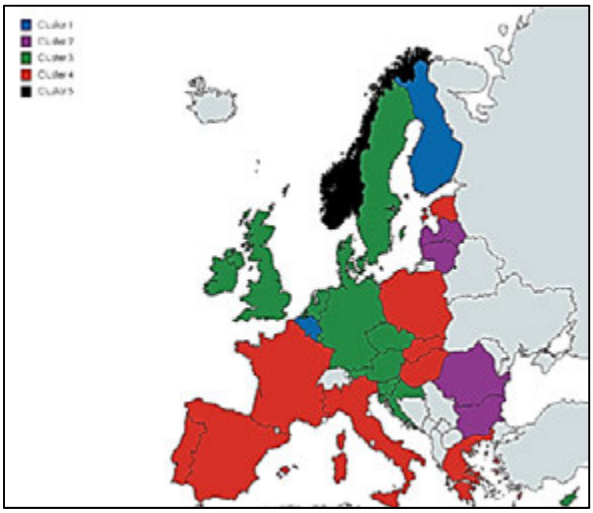


Figure 4. Countries membership across clusters; Source: Authors, Eurostat (2018)

Figure 5 and Table 5 presents the mean values of research variables across 5 clusters. It can be concluded that mean values for medium enterprises are the highest in all five clusters.

Furthermore, the highest mean values have Cluster 1 (Belgium, Finland) and Cluster 5 (Norway) which are leaders regarding ICT development and usage. Members of cluster 2 (Bulgaria, Latvia, Lithuania, Romania) are lagging behind the most indicators of ICT education in small, medium and large enterprises in European countries.

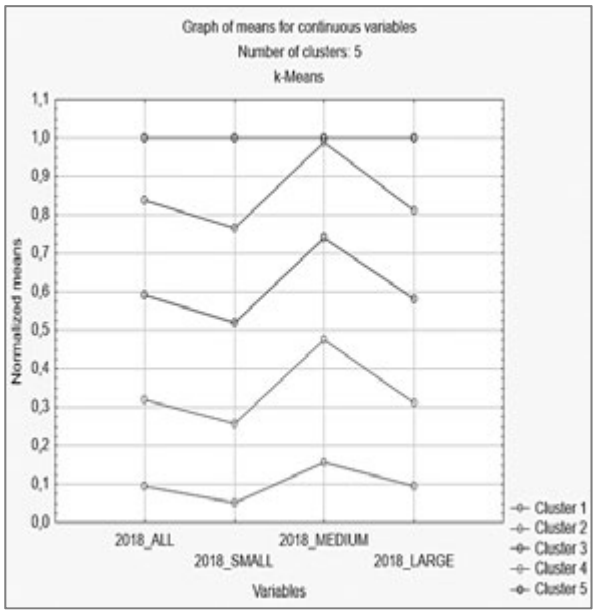


Figure 5. Mean values of research variables across clusters; Source: Authors, Eurostat (2018)

TABLE 5. VARIABLE MEANS ACROSS CLUSTERS

Online learning indicator	Cluster value				
	1	2	3	4	5
Enterprise provided training to their personnel					
ALL_2018	36.00	8.50	26.92	16.78	42.00
SMALL_2018	30.00	5.75	21.62	12.78	38.00
MEDIUM_2018	**59.50**	**15.25**	**46.31**	**32.22**	**60.00**
LARGE_2018	34.00	7.50	25.46	15.56	41.00

Source: Authors

VI. RELATIONSHIP BETWEEN COUNTRY COMPETITIVENESS AND ICT EDUCATION IN ENTERPRISES

In order to examine if there is a positive relationship between country competitiveness and ICT education in medium enterprises, we calculated the mean values of the GCI index across clusters (Table 6) [15].

Members of clusters 1 and 5 had the highest level of ICT education in medium enterprises, and they have the highest values of GCI index, which indicates that there is a possible positive relationship between the level of country competitiveness and ICT education in medium enterprises.

Other clusters, which were lagging behind according to ICT education in medium enterprises also lagged behind according to the GCI index. However, the Kruskal-Wallis test (H=10.355, p-value=0.035) confirmed that the found differences are statistically significant. Therefore, the hypothesis that there is a positive relationship between country competitiveness and ICT education in medium enterprises was confirmed.

TABLE 6. GCI ACROSS CLUSTERS

	Descriptives			Kruskal Walis test	
	Means	N	Std.Dev.	H	p-value
Cluster 1	78.45	2	2.62	10.355	0.035**
Cluster 2	65.10	4	1.83		
Cluster 3	74.88	13	7.29		
Cluster 4	69.49	9	4.86		
Cluster 5	78.20	1			
Total	72.22	29	6.86		

Source: Authors
Note: ** statistically significant at 5%

Box plot diagram of GCI index in medium enterprises confirms presented mean values (the highest mean value is for Cluster 1 and Cluster 5, which is an outlier) (Figure 6).

GCI index consists of four dimensions: (1) Enabling environment; (2) Human Capital; (3) Markets; (4) Innovation ecosystem and twelve pillars. In this analysis, we have investigated the correlation between ICT education in medium enterprises and two pillars of GCI index: ICT adoption (Enabling environment) and Skills (Human capital).

The correlation coefficient for ICT education in medium enterprises and ICT adoption is 0.5602 while ICT education in medium enterprises and Skills is 0.7091. Figure 7 presents scatterplot of ICT adoption and Skills and ICT education in medium enterprises, which indicate that selected variables are positively correlated and they move in the same direction.

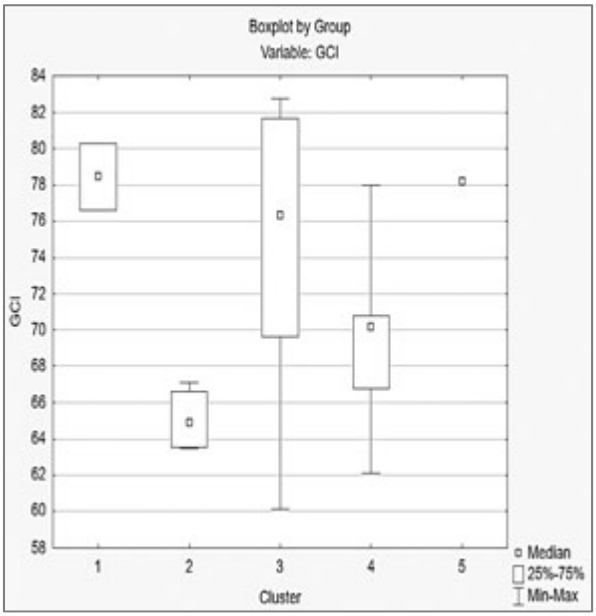


Figure 6. Box plot diagram of GCI in medium enterprises; Source: Authors, Eurostat (2018)

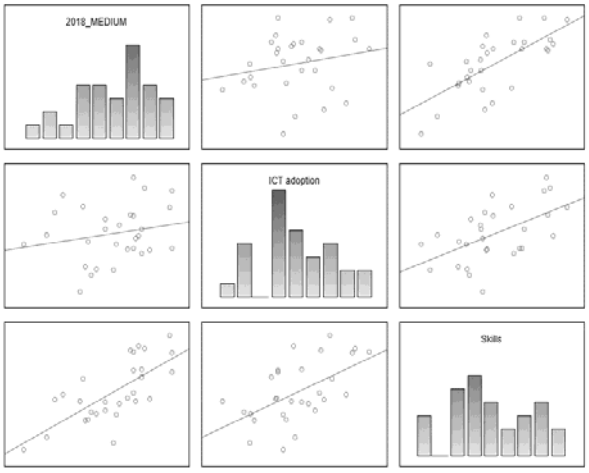


Figure 7. Scatterplot of ICT adoption, skills and ICT education in medium enterprises; Source: Authors, Eurostat (2018)
Figure 8.

VII. Relationship between digital economy development and ICT education in enterprises

In order to examine if there is a positive relationship between digital economy development and ICT education in medium enterprises, we calculated the mean values of the Desi index across clusters (Table 7) [16].

Members of clusters 1 and 5 had the highest level of ICT education in medium enterprises, and they have the highest values of the Desi index, which indicates that there is a possible positive relationship between the level of digital economy development and ICT education in medium enterprises. Other clusters, which were lagging behind according to ICT education in medium enterprises also lagged behind according to the values of the Desi index.

However, the Kruskal-Wallis test (H=12.530, p-value=0.014) confirmed that the found differences are statistically significant at 5%. Therefore, the hypothesis that there is a positive relationship between digital economy development and ICT education in medium enterprises was confirmed.

TABLE 7. DESI ACROSS CLUSTERS

	Descriptives			Kruskal Walis test	
	Means	N	Std.Dev.	H	p-value
Cluster 1	65.40	2	6.65	12.530	0.014**
Cluster 2	46.48	4	8.79		
Cluster 3	59.38	13	8.31		
Cluster 4	49.50	9	6.80		
Cluster 5	69.60	1			
Total	55.30	29	9.84		

Source: Authors
Note: ** statistically significant at 5%

Box plot diagram of the Desi index in medium enterprises confirms presented mean values (the highest mean value is for Cluster 1 and Cluster 5, which is an outlier) (Figure 8).

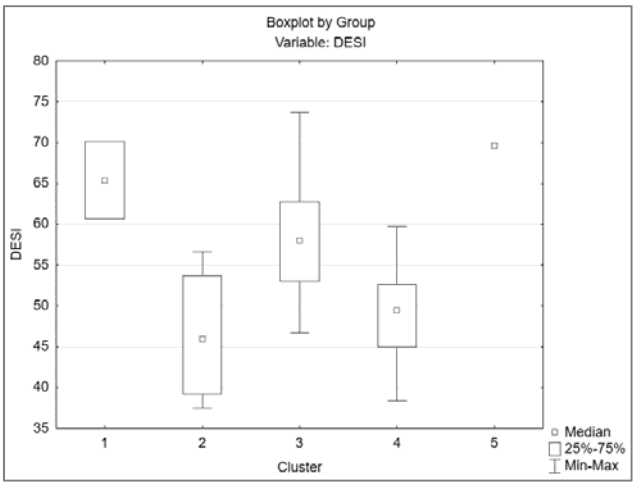


Figure 9. Box plot diagram of DESI in medium enterprises; Source: Authors, Eurostat (2018)

DESI index consists of five dimensions: (1) Connectivity; (2) Human Capital; (3) Use of Internet; (4) Integration of Digital Technology and (5) Digital Public Services.

In this analysis, we have investigated the correlation between ICT education in medium enterprises and two dimensions of the Desi index: Human Capital and Use of the Internet. The correlation coefficient for ICT education in medium enterprises and Use of the Internet is 0.5255 while for ICT education in medium enterprises and Human Capital is 0.7206. Figure 9 presents a scatterplot of Use of the Internet and Human capital and ICT education in medium enterprises, which indicate that selected variables are positively correlated and they move in the same direction.

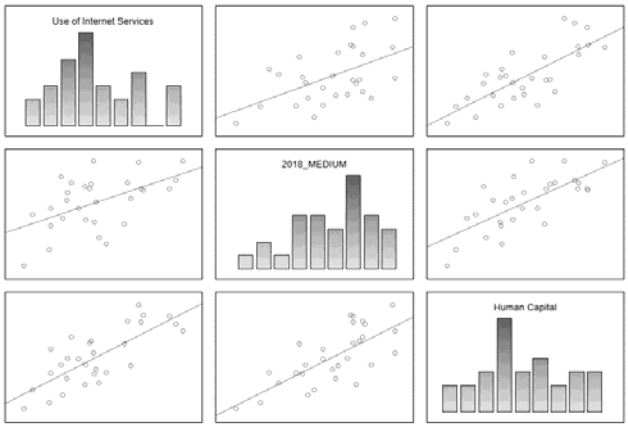


Figure 10. Scatterplot of Use of the Internet and Human capital and ICT education in medium enterprises; Source: Authors, Eurostat (2018)

VIII. Conclusion

The relationship between education and competitiveness is detected and explained in the number of researches on the company as well as on the national level [17; 18]. Fast technological and ICT development make this relationship much more important and as a result requirement for fast knowledge, adoption becomes one of the key issues in fast-developing markets. Since the government task is to improve country competitiveness

different measures in achieving this goal have to be developed and coordinated; development centers for small and medium companies, research areas centers, national and regional science parks, are just some of them [19].

Various strategies contribute to achieving a competitive advantage, but human capital investment is the major contributor to achieving innovation and increase company competitiveness in the marketplace [20]. Thus, the company is motivating entrepreneurship, fostered innovation in company culture, enabling employees with resources and abilities to solve the company problems that lead to increased productivity and profitability [21].

Our research has several limitations regarding collected data and methodology used which further implicit the future research directions. First, we collected data for the only one the year 2018 and we focused on selected European countries. Second, we used K-means cluster analysis. Therefore, future research should focus on larger samples regarding the time and different types of ICT education of employees. Furthermore, different clustering methods should be conducted and it would be important to include all Desi index dimensions and all GCI pillars.

IX. References

[1] N. Kusumaningtyas, and D. H. Suwarto, "ICT Adoption, Skill and Use Differences among Small and Medium Enterprises Managers Based on Demographic Factors", Procedia - Social and Behavioral Sciences, vol. 169, pp 296-302, 2015.

[2] J. Kukkamaki, S. Vesa, and H. Ruohoma, "Development of ICT education in digitalizing business environment", Acta Technica Napocensis-Series: Applied Mathematics, Mechanics, And Engineering, vol. 61, no. 4, 2018.

[3] C. Matt, T. Hess, and A. Benlian, "Digital transformation strategies", Business & Information Systems Engineering, vol. 57 no. 5, pp. 339-343, 2015.

[4] S. J. Berman, "Digital transformation: opportunities to create new business models", Strategy & Leadership, 2012.

[5] K. Zhu, S. Dong, S. X. Xu, and K. L. Kraemer, "Innovation diffusion in global contexts: determinants of post-adoption digital transformation of European companies", European journal of information systems, vol. 15, no. 6, pp. 601-616, 2006.

[6] H. Matlay, and M. Addis, "Adoption of ICT and e-commerce in small businesses: an HEI-based consultancy perspective", Journal of small business and enterprise development, 2003.

[7] Population structure and Ageing, Statistic explained (2019). Available at: https://ec.europa.eu/eurostat/statistics-explained/pdfscache/1271.pdf [February 2020]

[8] The Deloitte Global Millennial Survey (2019). Available at: https://www2.deloitte.com/global/en/pages/about-deloitte/articles/millennialsurvey.html?zd_source=hrt&zd_campaign=2159&zd_term=chiradeepbasumallick [February 2020]

[9] K. Breu, C. J. Hemingway, M. Strathern, and D. Bridger, "Workforce agility: the new employee strategy for the knowledge economy", Journal of Information technology, vol. 17, no. 1, pp. 21-31, 2002.

[10] S. El Kadiri, B. Grabot, K. D. Thoben, K. Hribernik, C. Emmanouilidis, G. Von Cieminski, and D. Kiritsis, "Current trends on ICT technologies for enterprise information systems", Computers in Industry, vol. 79, pp. 14-33, 2016.

[11] European Commission, Education and Training Monitor, Croatia (2019). Available at: https://ec.europa.eu/education/sites/education/files/document-library-docs/et-monitor-report-2019-croatia_en.pdf [January 2020]

[12] Eurostat, How much do Member States spend on education? (2020). Available at: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/DDN-20170828-1 [January 2020]

[13] Deshpand, A. "Why HR is at the Center of Digital Transformation" (2019). Available at: https://www.hrtechnologist.com/articles/digital-transformation/why-hr-is-at-the-center-of-digital-transformation/, [February 2020]

[14] Eurostat, Community Survey on ICT Usage and E-Commerce in Enterprises (2018). Available at: https://circabc.europa.eu/sd/a/1fbef4a1-4c31-4b6a-afe8-19ee6d7e3b0f/ICT-Entr%202018%20-%20Model%20Questionnaire%20V%201.2.pdf [November 2019]

[15] K. Schwab, "The Global Competitiveness Report", Geneva: World Economic Forum (2018). Available at: http://www.cdi.org.pe/pdf/IGC/2018/The_Global_Competitivenes s_Report_2018.pdf [December 2019]

[16] DESI Methodology (2018). Available at: https://ec.europa.eu/digital-single-market/en/desi [November 2019]

[17] M. Delgado, C. Ketels, M. Porter, and S. Stern, "The Determinants of National Competitiveness", Working paper series (National Bureau of Economic Research), 2012. Available at: https://www.researchgate.net/publication/236903847_The_Determinants_of_National_Competitiveness [February 2020]

[18] S. Krammer, "Science, technology, and innovation for economic competitiveness: The role of smart specialization in less-developed countries", Technological Forecasting and Social Change, vol. 123, pp. 95-107, 2017.

[19] J. Kowal, and G. Paliwoda-Pękosz, "ICT for global competitiveness and economic growth in emerging economies: Economic, cultural, and social innovations for human capital", Transition economies, pp. 304-307, 2017.

[20] S. Bubenik, "Can HR Strategies Create A Competitive Advantage For Your Company?", 2019. Available at: https://www.forbes.com/sites/scholleybubenik/2019/02/18/can-hr-strategies-create-a-competitive-advantage-for-your-company/#2605c06f5856 [February 2020]

[21] J. O. Ogbor, and E. G. Eromafuru, "Regional trade blocs, location advantage and enterprise competitiveness in the global economy", Archives of Business Research, vol. 6, no. 6, 2018.

Project Based Teaching with Digital Tools in Primary Education

J. Mezak*, P. Pejic Papak*
* Faculty of Teacher Education, University of Rijeka, Rijeka, Croatia
jasminka.mezak@uniri.hr
petra.pejic.papak@uniri.hr

***Abstract* - The aim of the project based teaching as a form of integrated classroom teaching is to direct students to research a particular topic so that they can gain new insights through interactive learning. It implies a constructivist approach by which the process of learning is based on the student's activity, interaction, sharing previous knowledge and taking responsibility for research work. The contemporary approach to project based teaching involves the use of various digital tools in the research process. Within the GLAT project, a focus group of primary school teachers used some digital tools to carry out a project activity with their students. This paper analyses project activities and teachers' views on the use of digital tools, methods and forms of work in the preparation and implementation of project activities, as well as their students' assessments of the level of fun and learning using these tools in classroom.**

Keywords – project activities, digital tools, teachers, primary school students

I. INTRODUCTION

Project teaching is based on interesting content or a problem that requires the student to set up a research, investigate, and solve previously appointed problem. It is characterized by interaction between students and teachers, but also between students themselves who actively use their previous knowledge and skills in order to construct new knowledge with meaningful training operation. In search of the most effective ways of achieving student-centered project-based learning goals, Creitaru [3] points out that the project method is flexible and easily correlated with other didactic methods. The connection with problem-oriented teaching should be emphasized. Problem solving as a teaching method has the intention to enable students to explore, understand and evaluate the content they are learning [10]. Problem solving competency implies that the student will perceive the problem from different perspectives, practice different problem solving strategies and use different digital tools [11]. The use of technology, or digital tools, in the educational process has become indispensable in order to be able to carry out certain activities more successfully and more quickly and thus to support active learning in students. Project teaching can also be perceived as a form of Inquiry Based Learning, an active approach to learning that requests students' investigations and finding answers for assigned questions. This paper presents the results of the Erasmus+ GLAT project with the aim of involving algorithmic thinking in teaching different subjects in primary school in a fun and attractive way. One of the important activities of the project was the professional development of teachers in primary schools through different innovative teaching methods using digital tools and technology.

Co-funded by the Erasmus+ Programme of the European Union, Key Activities 2, Strategic Partnerships for school education, Project „Games for Learning Algorithmic Thinking“ (2017-1-HR01-KA201-035362)

II. STUDENTS' ORIENTED PROJECT TEACHING

Čudina-Obradović [4] explains the project as one of the basic forms of integrated teaching, which she defines as planning and organizing teaching in which different areas are interconnected in order to achieve a deep understanding of certain content and at the same time mastering critical and creative thinking skills. The implementation of the project particularly emphasizes the student's activity. As emphasized by Vujičić et al. [13], project teaching means the use of different methods, procedures, forms of learning and techniques and its overall aim is not only to improve teaching, but for students to work independently and learn. Project-based teaching can be motivating for students, if the teacher organizes it well and allows the student the freedom to self-organize learning. Project-based learning is important for realizing complex learning goals, so Cindrić et al. [1] emphasize that project teaching is "a well-planned, well-designed teaching designed to gain insight by exploring a particular situation through interactive learning". Interactive learning is the process of constructing knowledge.

The constructivist approach to teaching "rests on the premise that the learning process takes place on the basis of the personal construction and reconstruction of knowledge that results from students' interactions with the natural world in a particular sociocultural context, with the dynamic mediation of their prior knowledge" [8]. The project encourages teamwork and enhances the development of students' research, communication, organizational and critical skills. This form of learning encourages curiosity and learning with understanding and correction of misconceptions [5]. Barron and Darling-Hammond [1] define project learning as: central to the curriculum, organized around questions that guide students toward the central concepts or principles of a course, and focused on study that involves research and knowledge building. The student must be directed to find

information from multiple sources independently, as well as to use digital technology both as a source of information and as a means of learning through prepared digital materials. Structured and guided approaches are recommended for young learners in collaborative project learning. Song [12] also emphasizes the "productive-failure" as "a learning design that affords students opportunities to generate solutions to a novel problem that targets a concept they have not learned yet, followed by consolidation and knowledge assembly where they learn the targeted concept". This kind of instructional design first involves students in an independent activity of finding solutions based on prior knowledge, which often leads to failure to solve problems. However, students will also use this information later to identify and gain new knowledge after the teacher helps them deal with misconceptions. Such approach can be useful for young learners to foster their active learning process. Furthermore, collaboration as a form of teaching is another important factor in project teaching in younger students, as different critical thinking and individual differences in prior knowledge can encourage better achievement [10].

III. THE GLAT PROJECT

The Erasmus+ project GLAT - Games for Learning Algorithmic Thinking, which ended in October 2019, included, as one of the goals, the professional development of primary school teachers with various innovative teaching methods using digital tools and technology. Primary school teachers were participating in the education through three workshops and the online part with the Moodle based e-learning system. After each workshop, teachers-education participants created learning scenarios according to the current national curriculum, using teaching methods and digital tools that were pointed out in the workshop. Subsequently, teachers applied such learning scenario in classroom with their students. A scenario consists of a subject and lesson, level of complexity, key concepts, learning outcomes and all activities description supplemented by teacher and student materials and resources [6], [9].

The second and the third workshop included Project-based teaching activities with digital tools. The topic of the second workshop, held on August 2018, was Problem Based Learning (PBL), online quizzes and logical tasks. As the name of the workshop points out, the focus was on examples of learning scenarios and supporting materials for the implementation of problem based activities and logical tasks. The topic of the third workshop, held on January 2019, was Games and Tools for Programming. The emphasis was on the learning scenarios that will include educational games and algorithmic thinking for the implementation of inquiry based learning activities for different school subjects.

When explaining a problem as algorithm, it is broken down in smaller parts that can be more familiar and thus solved using a set of rules (algorithms). Abstractions can be also used to generalize those solutions to similar problems [14]. For example, in early grades, teachers can explain the algorithm by highlighting the steps involved in solving any problem. Algorithmic thinking can be developed through any school subject with the appropriate teaching methods, with or without technology. For example students can play educational digital games to learn some new concepts as well as board games without technology. Younger students can understand and use the concept of the algorithm by writing a sequence of commands that solve a particular everyday situation they encounter (such as preparing breakfast, preparing for school, the way from home to school, etc.). In project teaching that includes problem solving and group work in lower grades, different tasks without technology can be assigned to individual groups: one group can design a gymnastic exercise and present it with pictures, while the other group than perform the exercise according to the presented algorithm. Any learning situation can be improved by incorporating such methods into the appropriate learning scenario from the first grade of primary school, and this can only be carried out by teachers who are qualified. [9].

A. Learning scenarios for Project Based Learning

A learning scenario is a document by which teachers plan or structure a learning situation. It can be made for one or more teaching units. It should have a strong connection with other subjects, that is, thoughtful activities that will emphasize and stimulate connection with other subjects. For example, one of the scenarios, created at the second workshop, aimed to understand the importance and value of cultural landmarks in the coastal area. The activities were designed in such a way that teacher created a quiz and a memory game on a computer /tablet for the introductory part of the lesson, to help students memorize key terms and define a problem question for group work. Through project-based learning and group work on a computer, students will explore cultural landmarks, present the gathered information, and replicate the learning unit by solving an interactive computer worksheet. The applications used for carrying out online quizzes and logical tasks were Kahoot! quiz, Match the Memory game, Wizer.me and LearningApps interactive worksheets [10].

The other example is one of the scenarios developed on the third workshop for the first grade students. The goal was to recognize and distinguish the characteristics of the seasons, specific to our climate. Project assignment included group work with the specific part of a project assignment (season). The specificities of the seasons and their seasonal fruits were explored partly at school, with the supervision of the teacher, and mostly at home with the supervision of the family members. Each group makes their own poster with the theme of the given season and present it. For remembering specifics of each season students design a storyboard for a digital story that includes a game: design a scene, main character, define the moves of the main character, what he/she is doing, saying, determine the ultimate goal of the game, describe the main character's tasks, define the way the points are collected in the game and determine when the game is over. Besides the learning scenario, the teachers, in collaboration with the students, created a "Game/story scenario" (Figure 1.) according to which the game in Scratch was programmed. Since the lower grades elementary school students who had not encountered

programming so far were involved in the design of the game/story, the program was created according to their instructions by the students of Informatics from University of Rijeka.

Games for Learning Algorithmic Thinking

Game/story scenario

Title of the game	Seasons
Type *(Scratch or micro:bit)*	Scratch story with game elements
Course/Grade	Science: Seasons / Health - 1st grade
Learning outcomes	Learning outcomes focused on general subjects • Recognize the seasonal fruits of particular seasons • Associate appropriate clothes with specific seasons Learning outcomes focused on algorithmic thinking • Search, find, and extract relevant information
Goal of the game	Dress the main character in season-appropriate clothing to eventually harvest the fruits characteristic of that season.
Characters and their role	Mia – the main character Family (mom, dad, 2 sisters, 1 brother) - supporting characters
Description of the game flow	The main character introduces the player into the story. Little girl Mia comes from Africa to Croatia in search of a better life. The goal is to teach Mia all about the seasons specific to the Croatian climate. For each season, it is necessary to first dress appropriately Mia, which is also a condition for continuing the game. Successful dressing is followed by the second part of the game, in which fruits specific to a particular season are collected. In that way the main character goes through all four seasons. By collecting the correct fruit, the player wins 1 point, while by making the wrong choice, he loses 1 point. Accordingly, by collecting seasonal fruits in all seasons, the player can collect a total of 36 points, out of which a minimum of 20 points is required for successful completion of the game.
List of scenes/backgrounds	1. Park - an introduction to the game 2. Africa 3. Summer - Dubrovnik, beach, sea 4. Laundry rope - choose summer clothes 5. Sailboat 6. Summer - beach - selection of summer fruits 7. Tree (displaying 4 seasons) - entry of the season following the summer 8. Autumn – forest 9. Laundry rope - choose autumn clothes 10. Autumn - forest - selection of autumn fruits 11. Tree (displaying 4 seasons) - entry of the season following the autumn 12. Winter - mountains, snow, cottage 13. Laundry rope - choose winter clothes 14. Winter – selection of winter fruits 15. Tree (displaying 4 seasons) - entry of the season following the winter 16. Spring – Meadow 17. Laundry rope - choose spring clothes 18. Spring – selection of spring fruits 19. Promenade - end, display of points won, arrival of Mia's family
Logical mini-games within the story *(Note: select games that are aligned with your learning outcomes)*	Recognizing the characteristics of the seasons. The first part of the game for each season: Choosing the right clothes for the main character according to the current season. Condition to continue the game - collecting points. Minimum required for each season: • Summer: 4 points • Autumn: 3 points • Winter: 4 points • Spring: 3 points The second part of the game for each season: Collecting fruits specific for specified season. By collecting fruits, a maximum of 9 points can be collected in each season.
End of the game	If all four levels are successfully played and the minimum number of points (20 points) is collected, we get to know the whole family of the main character. If the player does not collect enough points, Mia instructs him/her to repeat the game from the beginning.

Figure 1. Example of Game/story scenario developed by the teacher participant at GLAT education

IV. TEACHERS' AND STUDENTS' ATTITUDES ON THE USE OF DIGITAL TOOLS IN PROJECT TEACHING

The activities designed by the teachers who participated in the GLAT project were evaluated by examining the attitudes of their students and through the insight of their impressions of the performed activities. The evaluation in whole is presented in document "GLAT Project Dissemination" [6] of project web site. This paper describes only the part considering teachers' and students' attitudes on the project based teaching.

Assessment of the effectiveness of teacher education in the GLAT project was conducted using the initial and final questionnaires. A total of 27 teachers (only one male teacher) participated in the evaluation study. The main objective of the initial questionnaire was to identify the extent to which teachers are familiar with the terminology, methods and tools relevant to the use of ICT in teaching. In the final questionnaire, teachers indicated changes in relation to the same questions in the initial questionnaire.

The questionnaires consisted of four parts. For the purposes of this paper, the results of second part: "Familiarity with the possibilities of adapting, creating and using methods, contents and tools" consisting of six questions, were analyzed. Teachers expressed their progress in answering a series of the same six questions in the initial and final questionnaires for their ability to create teaching materials and apply various digital tools and teaching principles related to the GLAT project. Progress is evident by comparing the Likert-type responses in both questionnaires. Teachers reported progress in response format from 0 = "not at all" to 4 = "to a great extent".

Figure 2 shows the results regarding initial knowledge of the possibilities for adapting, creating and using digital teaching content and self-assessment of progress in the final questionnaire for the same terms. Three questions were identified relevant to the topic of this paper. It can be seen that the highest indicator of progress is related precisely to the area in which the participants showed a relatively high score in the initial questionnaire: creating online quizzes and logical tasks. However, it is more important to highlight an item with a relatively lower initial score that has a relatively high self-assessment of progress, such as "creating digital content using Web 2.0." Generalizing the progress from the initial situation, it can be said that the capability of the "digital teaching content creation" has doubled.

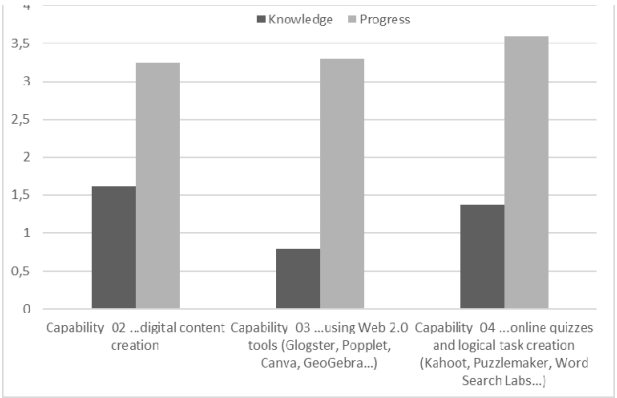


Figure 2. Comparison for initial status and progress in capabilities

To analyze the success of the implementation of GLAT activities with students, a questionnaire was designed for students in 3rd and 4th grades. Students evaluated the level of interest and the level of fun using digital learning tools. The questionnaire consisted of 5 statements about all the tools used in the class that

students evaluated with a 5-point Likert scale in which 1 indicated "I didn't have fun at all" and 5 indicated "I had a lot of fun". The respondents' sample consisted of 114 3rd and 4th grade students from six primary schools, of which 55.26% (N = 63) were girls and 44.74% (N = 51) were boys.

The Table I shows that Kahoot! is the best rated tool by level of entertainment, followed by Scratch. The tool with the lowest rating is LearningApps.

TABLE I. EVALUATION OF THE LEVEL OF ENTERTAINMENT WITH THE DIGITAL TOOLS

Tool	Match the Memory	Kahoot!	Wizer.me	Learning Apps	Scratch
N	111	113	114	98	114
Min	0	0	0	0	0
Max	5	5	5	5	5
Mean	3,26	4,67	2,86	1,97	4,3
Std. Dev.	1,818	0,881	1,886	1,976	1,212

A T-test was conducted to assess whether gender or age influences the level of entertainment in using digital tools. The results showed that there were no statistically significant differences between male and female respondents in assessing the level of entertainment with individual digital tools. A statistically significant difference in the estimate of the level of entertainment by gender (F = 8,885, p 0.023) was observed only with the Scratch tool, where the female examinees rated the Scratch tools higher (M = 4.54) than the male respondents (M = 4.00).

TABLE II. EVALUATION OF THE LEVEL OF ENTERTAINMENT REGARDING TO GRADE

	Grade	**N**	**Mean**	**Std. Deviation**
Match the Memory	3.	51	3,90	1,418
	4.	60	2,72	1,949
Kahoot!	3.	52	4,44	1,195
	4.	61	4,87	0,386
Wizer.me	3.	53	3,36	1,582
	4.	61	2,43	2,029
LearningApps	3.	39	2,21	1,720
	4.	59	1,81	2,129
Scratch	3.	53	4,47	1,137
	4.	61	4,15	1,263

The analysis of the level of entertainment of computer tools in relation to the class that students attend shows that 3rd grade students statistically significantly rated the Match the Memory and Wizer.me tools from 4th grade students, while the results for Kahoot! quiz were opposite (Table II). When choosing the tool they learned the most from, students mostly specified Kahoot! quiz (73.42%), while tools LearningApps and Match Memory have the lowest rating of 1.26%.

V. CONLUSION

Today, more attention is being paid to modernize the teaching process with digital technology. As technology has become one of the most important didactic tools in all school subjects, their use is no longer restricted to Computer science teachers. Timely with the help of digital tools is possible to carry out new, more interesting activities that enhance the educational process and enable active learning for students who achieve better results. However, teachers will be able to lead effective teaching with technology once they have been sufficiently educated to apply it [10]. Therefore, the training of teachers in this regard is very important. On the other hand, it is very important to modify the educational process to the needs and interests of students. In order to do this, it is important to examine the interests and opinions of students with a purpose of improving the quality of educational practice. Certainly, further research on this topic will bring in results that will make contributions to changes and innovations in teaching practice.

With the GLAT project we tried to encourage teachers to apply contemporary methods and teaching strategies together with digital tools and to implement problem-solving skills, logic and creativity into the daily learning. Project based teaching method can contribute to the development of creativity and logical thinking in young students which may consequently enable them to carry out what they have learned in new situations.

ACKNOWLEDGMENT

This paper is co-funded by the Erasmus+ Programme of the European Union, Key Activities 2, Strategic Partnerships for school education, Project "Games for Learning Algorithmic Thinking" (2017-1-HR01-KA201-035362)

REFERENCES

[1] B. Barron, L. Darling-Hammond ,Obeti in izzivi za pristope k učenju, temelječe na raziskovanju. In. Hanna, Dumont., David, Istace., i Francisco, Benavides. (ur.). O naravi učenja. Uporaba raziskav za navdih prakse. OECD Publishing (pp. 183–207), 2013.

[2] M.Cindrić, Đ.Miljković, V.Strugar, Didactics and the Curriculum. Zagreb: IEP-D2, 2010.

[3] A.Creitaru, The Use of the Project Method in Teaching Mechanics as a Technical Discipline. Seria Stiintele Educatiei, 61(2), (pp.119–126), 2009.

[4] M.Čudina-Obradović, S.Brajković, Integrated Teaching. Zagreb: Pučko otvoreno učilište Korak po korak, 2009.

[5] V.Fabijanić, Projektna nastava: Project-based learning: Application in the Writing of Student's Research Papers. Educatio biologiae, 1 (pp.89–96), 2014.

[6] "GLAT project dissemination", N. Hoić Božić and M. Holenko Dlab, Eds. 2019. Available at: https://glat.uniri.hr/wp-content/uploads/2019/12/GLAT_Project_dissemination.pdf [15.12.2020.]

[7] C. Hursen & F. G. Fasli, "Investigating the Efficiency of Scenario Based Learning and Reflective Learning Approaches in Teacher Education", European Journal Of Contemporary Education, vol. 6 (2), 2017, pp.264-279, doi:10.13187/ejced.2017.2.264

[8] R. Jukić, Konstruktivizam kao poveznica poučavanja sadržaja prirodnoznanstvenih i društvenih predmeta. In *Pedagogijska istraživanja,* 10, 2 (pp. 241-265), 2013.

[9] J.Mezak, P. Pejić Papak, "Learning scenarios and encouraging algorithmic thinking". In 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 0760-0765). IEEE. (2018, May)

[10] J.Mezak, P. Pejić Papak, "Problem Based Learning for Primary School Junior Grade Students Using Digital Tools". In 2019 42st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 805-810). IEEE. 2019.

[11] PISA 2015 COLLABORATIVE PROBLEM-SOLVING FRAMEWORK 2017. Available at: https://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf [14.01.2019.]

[12] Y. Song, "Improving primary students' collaborative problem solving competency in project-based science learning with productive failure instructional design in a seamless learning environment", Educational Technology Research & Development, 66(4), pp. 979-1008, 2018. Available at: https://doi.org/10.1007/s11423-018-9600-3 [14.01.2019.]

[13] L.Vujičić, P. Pejić Papak and M. Valenčić Zuljan, Okruženje za učenje i kultura ustanove, Učiteljski fakultet Sveučilišta u Rijeci, Rijeka, 2018.

[14] J. M. Wing, "Computational thinking". Communications of the ACM, vol. 49 (3), 2006, pp. 33-35. available at: http://denninginstitute.com/pjd/GP/Wing06.pdf [20.01.2018.]

University teachers' needs for the efficient use of ICT in the pedagogical process

N. Špur**, E. Kranjec*, M. Puhek*, K. Breznik*, L. Klasinc* and S. Frumen*
* University of Maribor, Maribor, Slovenia
**Corresponding author: natalija.spur@um.si

Abstract - **The appropriate use of Information and Communication Technology (ICT) along with Modern Teaching Practices (MTPs), such as flipped learning and project-based learning, can increase the efficiency of teaching and learning in the higher education (HE) process. To ensure effective implementation of ICT and MTPs, the University of Maribor (UM) has established institutional support in the form of the Centre for Teaching Support of the University of Maribor (CTSUM), which is available to university teachers of all 17 UM faculties. The CTSUM provides support for the use of ICT and implementation of MTPs through individual consultations and workshops. This study aims to assess teachers' interests and needs in the field of ICT and MTPs along with the self-evaluation of their existing ICT skills. It was conducted as an online survey with responses from 154 teachers (approx. 26% response rate). The results show that a significant majority of teachers are interested in the provided support and need training on the use of ICT for preparing interactive assignments, voting and providing feedback, performing project-based learning, designing graphically rich study materials, and organising course materials online. The obtained information helps the CTSUM prepare customised workshops and consultations aimed at strengthening teachers' digital literacy and didactic skills.**

Keywords - consultations; ICT; institutional support; university teachers; workshops

I. INTRODUCTION

The use of ICT can increase the efficiency of teaching and learning [1]. However, to achieve a positive effect, teachers also need to understand when and how to use ICT in support of their pedagogical practices. With the rapid development of technology, acquiring new knowledge and constantly upgrading existing skills and competences is crucial.

The teaching staff of the University of Maribor (hereinafter UM) has the possibility to collaborate with the Centre for Teaching Support of the University of Maribor (hereinafter CTSUM) to exploit the potential of ICT and, consequently, to improve the pedagogical process. The CTSUM is a support service established under the Didakt.UM project. Apart from providing ICT support (e.g. individual support, face-to-face workshops, and online workshops) to university teachers, the CTSUM's tasks are linked to the development of didactic recommendations for the use of ICT, technical support for the use of ICT, and support in implementing modern teaching approaches with ICT. The CTSUM also disseminates good practices through educational events and encourages transmission of teaching experiences among teachers. Last but not least, the CTSUM runs a simultaneous and longitudinal verification of effectiveness in the field of modern education and participates in the scientific research field.

Since the CTSUM provides institutional support to approximately 900 teachers of all 17 UM faculties, it is imperative to learn about their interests and needs in the field of ICT tools to provide effective support. Therefore, every workshop is evaluated by the participants, allowing the CTSUM to assess the level of satisfaction with the performed work and content while at the same time giving the teaching staff the possibility to express their needs by suggesting ICT tools that the CTSUM has not yet offered.

A. Aim of the study and research questions

The study aims to assess the interest of the UM pedagogical staff in the use of ICT and evaluate their self-perception on the use of ICT. The CTSUM used the online survey as an enquiry to gather information on the current needs and interests of UM teachers. Based on the acquired data, it can organise further activities and workshops, tailored according to teachers' interests, needs, and existing skills, for a specific tool within a learning practice (e.g. flipped learning, project-based learning). The initial goal of the study was to determine the trainings the teachers need. The survey included the following research questions: (i) Which ICT tools do teachers want to use in support of the pedagogical process? (ii) How do teachers evaluate their ability to use different ICT tools (as "basic users" or "proficient users")? (iii) For which ICT tool would teachers want to obtain additional training/education? (iv) Is there a statistically significant correlation between interest in specific ICT tools and self-evaluation of ability to use ICT tools? (v) Is there a statistically significant difference in self-evaluation of ability to use ICT tools and interest in using ICT based on the years of working experience?

II. METHODS

To achieve our goals, we gathered information on the extent to which teachers are acquainted with and interested in specific ICT tools, and how they assess their skills for the use of ICT elements (see Ch. Structure of the questionnaire). The analysis was conducted among UM teachers via an online survey.

A. Data collection

Data were collected via the online survey system 1KA[1] in December 2018. The link to the survey was sent via email to all (approximately 900) UM teachers. The anonymity of the respondents was guaranteed.

B. Structure of the questionnaire

The questionnaire consisted of three parts.

I. Interest in specific ICT tools:

Respondents were asked about their knowledge of and interest in the following 23 ICT tools for which the CTSUM prepared didactic and technical guidelines[2] in 2017 and 2018:

1. Asana: Support for project-based learning. Organising individual and group work, creating tasks and subtasks, etc.
2. Avogadro: Editing and visualising chemical particles.
3. Desmos: Graphing calculator.
4. Edpuzzle: Editing videos enhanced by adding various types of interactive questions.
5. LearningApps: Creating different activities: sorting answers in the correct order, quizzes, etc.
6. Educreations: Recording the whiteboard and oral explanation.
7. Mentimeter: Gathering participants' answers and providing immediate feedback.
8. Kahoot: Gathering participants' answers and providing immediate feedback.
9. Socrative: Gathering participants' answers and providing immediate feedback.
10. Moodle UM: Centralised learning management system of the UM, used by all faculties.
11. Padlet: Distance collaboration via informative dashboard by adding different elements (e.g. pictures, videos, links).
12. Amara: Video editing with subtitles.
13. Audacity: Audio recorder and editor.
14. BioDigitalHuman: Visualisation of human body parts.
15. Canva: Editing infographics, invitations, posters, etc.
16. Duolingo: Language learning.
17. Mahara: E-portfolio.
18. GeoGebra: Programme for teaching mathematical content.
19. Mendeley: Organising, searching, and machine citing sources.
20. Microsoft Publisher: Creating stickers, leaflets, certificates, etc.
21. Microsoft Visio: Creating flowcharts, organisation charts, building plans, etc.
22. Shotcut: Video editing.
23. Arnes VOX (based on Adobe Connect): Organising and conducting videoconferences with interactive tools for distance collaboration (e.g. e-slides, screen and file sharing, chat).

The 5-point scale was used to measure the respondents' knowledge and interest in the above-mentioned ICT tools:
(F1) "I am not familiar, and I am not interested."
(F2) "I am not familiar, but I would like to use it."
(F3) "I am familiar, and this is enough for me."
(F4) "I am familiar, but I would like to learn more about it."
(F5) "I am familiar and do not need extra education."

II. Self-evaluation of the ability to use ICT elements:

The scale consisted of 10 categories of ICT elements on a 4-point scale (see details of self-evaluation explained for each category in [1]):

1. written communication,
2. learning management systems (e.g. Moodle),
3. videoconference systems,
4. interactive whiteboards with didactically prepared materials,
5. collections of educational resources (e.g. OpenLearn, Web of Science, and local DKUM[3]),
6. voting applications and feedback,
7. production of graphically rich learning materials (e.g. animations, images, schemes, infographics, storytelling, conceptual maps, and mind maps),
8. production of e-slides and written materials (e.g. Prezi, Microsoft PowerPoint, Microsoft Word, and LaTeX),
9. screen recording, oral explanation, and video editing (e.g. Microsoft PowerPoint), and
10. recording pedagogical activities.

III. Demographics:

Respondents were asked to answer questions about their gender (male or female) and years of working

[1] The survey was administrated using the OneClickSurvey tool (1KA) (http://english.1ka.si/).
[2] Didactic and technical guidelines for particular ICT tools are hosted on the CTSUM's web subpage Learning Stack (https://didakt.um.si/stolpic/).
[3] DKUM: The Digital Library of the University of Maribor (DKUM) is an institutionalised repository. The DKUM supports open access to scientific, research, and professional works, as well as research data, resulting from the research and education at the UM.

experience (up to 5, 6-11, 12-16, 17-21, 22-26, 27-31, 32-36, and more than 36 years, Table I).

C. Statistical procedures

Interest in ICT tools and self-evaluation of the skills for the use of ICT elements is presented with descriptive statistics with frequencies (F) in Tables II and III. Differences in interest in ICT tools and self-evaluation of ability to use ICT tools between separate groups based on years of working experience were examined with nonparametric statistics through Kruskal-Wallis H test and pairwise comparison analysis [2]. Spearman's correlations (r_s) among items were examined to determine the association between interest in ICT tools and ability to use ICT [3]. Collected data were analysed with IBM SPSS Statistics 24 software.

III. RESULTS

We collected 154 filled in or partially filled in questionnaires (response rate of 17.11%). More than half of the teachers were male (58%). The majority had been working at the UM for 17 to 21 years.

TABLE I. DEMOGRAPHIC CHARACTERISTICS OF PARTICIPANTS

Gender	F (%)
Male	88 (58%)
Female	65 (42%)
Years of working experience (at home faculty)	F (%)
less than 5 years	23 (15%)
6–11 years	26 (17%)
12–16 years	20 (13%)
17–21 years	33 (21%)
22–26 years	15 (10%)
27–31 years	13 (8%)
32–36 years	10 (6%)
more than 36 years	14 (9%)

Table II shows which ICT tools teachers are not familiar with but would be keen on using. For example: LearningApps (n = 47%), Edpuzzle (n = 46%), Educreations (n = 40%), Mentimeter (n = 37 %), Socrative (n = 37%), Asana (n = 34%), Kahoot (n = 34%), Microsoft Visio (n = 34%), Canva (n = 32%), Duolingo (n = 32%), and Amara (n = 32%). Although most respondents are familiar with Moodle UM and feel no need for extra education/training, there are about 30% of teachers who would like to learn more.

Moreover, for some tools that fall into the same category (e.g. Mentimeter, Kahoot in the voting and feedback tools category), correlations emerge. For example, teachers who show interest in Mentimeter also express interest in Kahoot ($r_s = 0.55$, $p < 0.001$) and Socrative ($r_s = 0.68$, $p < 0.001$).

TABLE II. INTEREST IN SPECIFIC ICT TOOLS FOR WHICH THE CTSUM HAS PREPARED DIDACTIC AND TECHNICAL RECOMMENDATIONS

ICT tool	**F1 (%)**	**F2 (%)**	**F3 (%)**	**F4 (%)**	**F5 (%)**	***n***
LearningApps	44	47	7	2	0	137
Edpuzzle	49	46	3	1	1	137
Educreations	56	40	4	1	0	136
Mentimeter	40	37	9	7	6	139
Socrative	52	37	1	7	3	135
Asana	51	34	10	4	1	140
Kahoot	42	34	9	9	6	139
Microsoft Visio	51	34	10	4	1	140
Amara	64	32	2	1	1	136
Canva	58	32	4	4	1	139
Duolingo	45	32	12	6	6	139
Mendeley	26	29	9	16	20	137
Microsoft Publisher	41	29	13	10	8	140
Padlet	58	28	3	7	4	139
BioDigital Human	72	26	1	0	1	138
Desmos	70	25	4	1	1	139
Audacity	46	25	14	4	12	138
Mahara	69	25	4	1	0	136
Arnes VOX	70	25	4	1	1	139
Avogadro	77	22	1	1	0	139
Shotcut	77	22	1	1	0	139
GeoGebra	71	16	6	3	4	139
Moodle UM	1	5	19	30	45	150

Note. F1 = I am not familiar and I am not interested, F2 = I am not familiar, but I would like to use it, F3 = I am familiar and this is enough for me, F4 = I am familiar, but I would like to learn more about it, F5 = I am familiar and do not need extra education, n = total number of answers.

Table III shows that most teachers rank themselves as proficient users of written communication tools (n = 61%), production of electronic slides and written material (n = 50%), the learning management system Moodle UM (n = 48%), and collections of educational resources (n = 32%).

The results also show that most teachers rank themselves as independent users of videoconference systems (n = 34%) and production of graphically rich learning material (n = 33%), and as basic users of interactive whiteboards with didactically prepared materials (n = 50%), screen recording, oral explanation and video editing (n = 38%), and ICT tools for recording pedagogical activities (n = 51%).

Interest in additional training in the Moodle tool (Table II) is statistically significantly associated with the self-evaluation of the ability to use learning management systems (Table III, $r_s = 0.43$, $p < 0.01$) and ICT for written communication ($r_s = 0.28$, $p < 0.01$). A similar correlation is found for interest in Audacity and self-evaluation of ability to use tools for screen recording, oral explanation, and video editing ($r_s = 0.29$, $p < 0.01$), tools for recording pedagogical activities ($r_s = 0.33$, $p < 0.01$), and self-evaluation of ability to use an interactive whiteboard with prepared didactic materials ($r_s = 0.28$, $p < 0.01$). A correlation is also found between interest in the video editing tool Shotcut and self-evaluation of ability to use recording equipment for pedagogical activities ($r_s = 0.7$, $p < 0.01$).

TABLE III. SELF-EVALUATION OF ABILITY TO USE ICT TOOLS IN 2018

ICT elements	F1 (%)	F2 (%)	F3 (%)	F4 (%)	F5 (%)	*n*
Videoconference systems	19	34	21	20	7	151
Production of graphically rich learning materials	22	33	23	16	6	147
Voting applications and feedback	39	26	16	9	10	148
Collections of educational resources	13	24	27	32	3	151
Interactive whiteboards with didactically prepared materials	50	21	12	3	14	151
Screen recording, oral explanation, and video editing	38	21	15	18	7	150
Recording pedagogical activities	51	19	10	11	9	149
Learning management systems	1	18	31	48	1	150
Production of electronic slides and written materials	1	17	31	50	1	150
Written communication	1	9	28	61	1	149

Note. F1 = Not at all, F2 = Basic user, F3 = Independent user, F4 = Proficient user, F5 = I am unable to assess. Explanations of the proficiency categories can be found in [4], n = total number of answers.

Kruskal-Wallis H test shows no statistically significant differences ($p > 0.05$) in interest in ICT tools and perceived ability to use ICT based on years of working experience. Additionally, participants were classified into four new groups based on the years of working experience: 1) less than 12 years, 2) 12-21 years, 3) 22-31 years, and 4) more than 31 years. Finally, statistically significant difference between ability to use tools for screen capture, oral interpretation, and video processing according to years of working experience was obtained (χ^2 (3) = 9.0; $p = 0.03$). Differences in perceived ability to use tools for screen recording, oral explanation, and video editing are found between group 1 and group 3. Teachers with less than 12 years of working experience perceive themselves less capable (n = 29, R = 34.78) to use these tools than those with more years of working experience (n = 14, R = 56.32).

IV. DISCUSSION

The present study has achieved the initial goals. It reveals (i) which ICT tools teachers want to use in the pedagogical process, (ii) their self-evaluation of ability to use ICT tools, (iii) and which are the ICT tools the teachers would like to learn more about. It also determines (iv) association between interest in specific ICT tools and self-evaluation of ability to use ICT tools (v) as well as differences between self-evaluation of ability to use ICT tools and interest in ICT tools based on years of working experience.

A. Familiarity with ICT tools and interest in using them in pedagogical process

The examination of university teachers' familiarity with specific ICT tools and interest to use them shows that the majority of teachers are not familiar only with two ICT tools that they would like to use: LearningApps and Mendeley. LearningApps can be useful for interactive educational materials, therefore the CTSUM has found a similar installation within the learning management system Moodle UM (i.e. H5P) which experts consider particularly appropriate in combination with modern teaching approaches (e.g. blended learning, flipped classroom) [5]. With this in mind, the CTSUM is also preparing an online learning workshop on the use of H5P plugin for interactive content. Similarly, the workshop "Organizing Citations and Bibliography with Mendeley" was conducted, targeting the development of bibliography citation skills and the preparation of reference lists in a text editor.

Regarding ICT tools that teachers are unfamiliar with and do not show interest in using (e.g. Avogadro, Shotcut, BioDigital Human, and GeoGebra), a possible reason could lie in teachers' judgement about appropriateness of using individual tools based on their study area, or in their interest for using alternative tools that are not included on our list. Therefore, as mentioned in the introduction section, teachers can suggest training for ICT tools that have not yet been offered by the CTSUM.

The CTSUM strives to provide support also to smaller groups of teachers that are interested in using ICT and need individual support. Through this customised support, the CTSUM can increase the ability of using ICT tools of all UM teachers, consequently improving the pedagogical process in different study areas at the UM. According to our data, teachers are interested in using ICT tools that significantly contribute to more effective implementation of modern education. These are the tools for producing graphically rich learning material (e.g., Microsoft Publisher, Canva), assessment of knowledge and feedback (e.g. Mentimeter, Kahoot, Socrative), and videoconferencing systems (e.g. Arnes Vox). Following the expressed interest in use, the CTSUM conducted the workshop "Effective Feedback with Mentimeter" where participants learned about the components of effective and immediate feedback and were tested in practical situations. In addition, the workshop »Combine Teaching with Microsoft Teams" (i.e. a comparable tool for video conferencing systems) in support of blended and distance learning and the workshop "Visualize Study Materials with Canva" for creating graphically rich learning materials are available. The results of our study also indicate teachers' interest in using ICT tools that support modern teaching practices. Currently, teachers can participate in the workshop "With an Asana-assisted Project-based Learning", which aims at acquiring the basic features of project-based learning and its implementation into the teaching practice.

B. *Self-evaluation of ability to use ICT tools*

Current results show that the majority of teachers do not use the equipment for recording pedagogical activities, interactive whiteboards, and voting applications, therefore they could not evaluate their ability to use them. Past internal analyses found that ICT tools for screen recordings and oral explanations as well as equipment for preparing video materials were often not available to them. There is still a small proportion of teachers who are unable to assess some tools (e.g. equipment for recording pedagogical activities), which led the CTSUM to purchase a video camera with associated equipment (e.g. camera stand, wireless microphone system) teachers can borrow for educational purposes. Additionally, the CTSUM is preparing a workshop on the effective use of such equipment in the pedagogical process. During the workshop, teachers will be trained in tools that enable preparing screen recordings (e.g. PowerPoint) and editing video materials (e.g. Shotcut, Amara). Teachers need professional support in editing didactically and technically appropriate video materials, and this can only be provided by services such as the CTSUM. With the subsequent transfer of good practices, we can assume that more and more teachers will be involved in these activities, especially considering the findings that the increasing use of ICT in the teaching process is a consequence of the immense and growing pressure resulting from educational reforms [6].

C. *Interest in additional training/education for ICT tools within modern teaching practices*

The investigation of teachers' self-evaluation of ability to use ICT tools and their interest in using them has shown that most participants are familiar with the open-source learning management system Moodle UM and do not need extra education. Despite such insights, some of them still pointed out the need for further training. The CTSUM has therefore organised the first online learning workshop named "Didactically Organized Learning Unit in Moodle UM," available to teachers since 2018. Through the workshop, teachers can gain knowledge in the editing of the unit-based course and its components (e.g. activity naming, head design, use of tag modules, HTML and calendar blocks, embedding elements, and more). Based on the identified lack of knowledge of gamification in higher education [4], the workshop "Gamification of Learning Unit in Moodle UM" was also prepared. Participants explore the role of gamification in education and are presented with the possibility of gamification elements in Moodle UM. Participation in this workshop is especially useful to teachers who would like to learn more about Moodle UM and its options. The CTSUM also prepared the workshop "With ICT-supported Flipped Learning" where participants can learn about the phases of flipped learning and are trained for using different activities in Moodle and creating video lectures with PowerPoint for which most teachers consider themselves to be basic users of (Table III: Screen recording, oral explanation, and video editing).

Such supported distance and open learning show many benefits, especially in adjusting the time and place of activities, which help teachers to overcome one of the critical challenges of introducing ICT into the teaching process [7].

D. *Association between interest in specific ICT tools and self-evaluation of ability to use ICT tools*

Based on found correlations between interest in specific ICT tools and self-evaluation of ability to use ICT tools, the CTSUM can prepare differentiated activities for using ICT tools according to perceived degree of ability to use them (e.g. basic and proficient users). For proficient users, workshops involving more ICT tools can be prepared, because significant correlations between interest for different ICT tools in the same category were found (e.g. voting apps such as Mentimeter, Kahoot, Socrative). This kind of workshops can encourage discussion about using the appropriate tool for specific didactic scenarios (e.g. using Socrative for group competition). Teachers should be aware of proper technological tools for a particular task and the strategies for using the tools [8]. They should be able to evaluate ICT tools and materials, so they can adapt or develop suitable activities for their study area. However, workshops including more ICT tools should be well considered and planned to ensure meaningful connectivity between tools and prevent cognitive overload.

E. *Years of working experience as a significant factor in the perceived ability to use ICT*

An important factor in the perceived ability to use ICT is also teachers' seniority. The results of our analysis show statistically significant differences, particularly among teachers with the lowest and highest seniority in relation to the ability to use ICT tools. Based on this, the CTSUM started to carry out differentiated activities for technically experienced users who are more didactically oriented (e.g. through the workshop "With an Asana-assisted Project-based Learning," the CTSUM demonstrated practical and didactic examples without technical details). Moreover, collaboration with less experienced teachers (i.e. lower seniority) should be established from the beginning of their teaching career to encourage them to develop learning environment with appropriate ICT in accordance with the didactic guidelines.

Successful implementation and more effective trend-following also require high-quality promotion and networking with similar institutions [9]. In addition to the above-mentioned workshops and prepared recommendations for different ICT tools, the CTSUM has established different ways of promoting and informing about selected ICT tools and prepared other workshops

(e.g. Canva, Microsoft Teams, Mendeley). One of the examples of indirect promotion is the so-called "CTSUM News," delivered in an online format. It emphasises the importance of ICT and modern teaching practices in the field of education and includes recommendations for the use of a selected ICT tool. In addition, it provides an overview of ICT and educational events in Slovenia and around the world. For upgrading the knowledge and ICT skills, teachers can use tips from the "IKTarna" – a section on the CTSUM's webpage *didakt.um.si* with short posts about the practical use of specific ICT tools. The promotion of the CTSUM is almost entirely tied to website activity, which is slowly outgrowing the node with information on the use of ICT in higher education.

V. CONCLUSION

In the present study, we achieved initials goals and answered all five research questions. Based on the obtained results, the CTSUM can effectively address teachers' needs through customised workshops and didactic materials with guidelines. The research revealed the tools teachers are interested in, the tools they seek additional training for, and the level of their perceived ability to use ICT. The CTSUM support service takes on an important role in supporting teachers to acquire skills for the pedagogical use of ICT. Since the field of educational technologies is evolving rapidly, institutions must mindfully take steps towards assisting teachers in how to use and implement ICT in their teaching. In the future, the CTSUM expects requests for support from teachers with diverse needs and interests.

The performed analysis includes self-evaluation of ability to use ICT, which by itself might be less objective. However, it is an important factor for teachers in deciding to participate in the workshops and upgrade their skills. We assume that most teachers who participated in the research are interested in the field of ICT more than an average user and were therefore more attracted by the title and keener to participate. It is possible that the outcomes would vary if the number of participants were greater.

Regardless, our results point out the teachers' needs that the CTSUM will consider as a priority in the future. In planning educational events, the CTSUM considers implementing similar periodic analyses, identifying the needs for new knowledge of ICT tools and MTPs that can change rapidly. Since the CTSUM was established only three years ago, it should be stressed that it is still in the process of gaining teachers' trust. It is expected that insights into the interests and needs of teachers will increase and become more accurate over time.

ACKNOWLEDGMENT

This paper has been prepared as a part of the Didakt.UM project with funding from the Republic of Slovenia and the European Union under the European Social Fund.

We would like to thank all the UM faculties for their cooperation; dr. Tina Balič and Petra Usar for proofreading.

REFERENCES

[1] S. L. Comi, G. Argentin, M. Gui, F. Origo, L., and Pagani, "Is it the way they use it? Teachers, ICT and student achievement", *Economics of Education Review*, vol. 56, pp. 24-39, 2017.

[2] A. Field, "Discovering statistics using SPSS, 3nd edition". London: Sage Publications, 2009.

[3] J Hauke, and T. Kossowski, "Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data", *Quaestiones Geographicae*, vol. 30, issue 2, pp. 87–93, 2011.

[4] K. Breznik, N. Špur, A. Creelman, and S. Frumen, "Didactical use of information communication technology (ICT) and modern teaching practices (MTP) in higher education". In N. Gajšt (ed.), Book of abstracts of the 2ND International scientific conference "Teaching methods for economics and business sciences". Maribor: Univerzitetna založba Univerze v Mariboru, 2018. Available: http://press.um.si/index.php/ump/catalog/view/334/295/531-1 [Jan 20, 2020].

[5] S. Wilkie, G. Zakaria, T. McDonald, and R. Borland, "Considerations for designing H5P online interactive activities. ASCILITE 2018". In conference proceedings: "35th International Conference of Innovation, Practice and Research in the use of Educational Technologies in Tertiary Education: Open Oceans: Learning Without Borders", Geelong, Australia pp. 543–549, 2017.

[6] Cox, M, "The effects of attitudes, pedagogical practices and teachers' roles on the incorporation of ICT into the school curriculum". In Dowling, C., and Lai, K. (eds.), Information and communication technology and the teacher of the future, pp. 145–158. Boston: Springer, 2003.

[7] K. Bingimlas, "Barriers to the successful integration of ICT in teaching and learning environments: A review of the literature", Eurasia Journal of Mathematics, Science and Technology Education, vol. 3, pp. 235–245, 2009.

[8] Chen, Y-L, "A Mixed-method Study of EFL Teachers' Internet Use in Language Instruction", *Teaching and Teacher Education*, vol. 24, pp. 1015-1028, 2008.

[9] N. Kalinenko, V. Krasnopolsky, V. Kukharenko, and A. Serebryakov, "Using of social networks in educational process". In "The 7th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications", Berlin, Germany, pp. 12–14, 2013.

Analysis of Social Networks Usage among Students

M. Lončar and D. Kermek
Faculty of Organisation and Informatics, Varaždin, University of Zagreb, Croatia
e-mail: mloncar@foi.unizg.hr

***Abstract* - This paper gives an insight into the implementation and results of the research on the habits and reasons for using social networks among students of the Faculty of Organization and Informatics in Varaždin. For the purpose of conducting the research, a survey questionnaire was created to examine the habits and reasons for using Facebook, Instagram and Twitter. Among other things, the use of social networks for academic purposes is examined. Data obtained through the survey over 14 days was meticulously analysed and a comparison of habits and reasons for the use was made between Facebook, Twitter and Instagram. An analysis of 217 survey responses showed which the most popular social network among college students was, which social network students used the most for academic use and for entertainment purposes. The analysis covered dependences on the number of friends and followers for three social networks.**

Keywords - social networks; questionnaire; research; Facebook; Instagram; Twitter;

I. INTRODUCTION

New technologies have always been of interest to younger generations. Young people are more inclined to explore new things because they are more open minded and are more likely to accept and adapt to new technologies. The Internet as a technology has enabled the emergence of recently unimaginable products and services. Different ways of communicating that were not possible before have become a part of everyday life. People of all generations from all over the world interact with each other by exchanging text, image and video content. They often use some form of social software to communicate. Social networks are certainly the most popular social interaction form today, relying on the Internet and, more recently, mobile technologies. Social networks have become an integral part of the lives of not only the young but also the older generations over the last ten years. They are used for information, entertainment, communication with other users, but also as tools for education, business events, marketing and more.

In order to understand how the students of the Faculty of Organization and Informatics experience social networks today, a survey was conducted on their use of social networks. The research was conducted with a questionnaire on a sample of students of the Faculty of Organization and Informatics. The analysis of the research results will show to what extent, how and for what purposes students use social networks.

The rest of the paper is structured as follows: Section 2 describes related work and Section 3 presents research on using social networks among students. Section 4 presents data analysis and research results. Section 6 offers a conclusion.

II. RELATED WORK

Social networks and students as their users has been rather interesting area for research in last fifteen years. Many of them had focus on one of the most popular social networks platform/site Facebook and examine its relationship to formation and maintenance of social capital [1], its pedagogical potential [2], why students use it [3], privacy concerns [4], etc.

Spreading the potential of social networking to different media types results in popularity of other social networks platforms like Twitter, Instagram, and Snapchat, just to mention some of the most popular. There are many papers that were dealing with one of those social network platforms e.g. [5] motivation to use Instagram but here we prefer those researches that included at least three of them. Some authors [6] examined motivations and use of four social network platforms, other [7] bridging and bonding social capital, etc.

Salem and Mengyan [8] surveyed 363 students at Mitchigan State University in an online survey. They questioned the use of Facebook, Twitter, Instagram and Snapchat. The questions used in the survey are based on a 7-point Likert scale. Analysis of the results showed that students on average per day in minutes maximum use Instagram (108.73 min), followed by consecutive Snapchat (107.15 min), Facebook (106.35 min) and Twitter (88.92 min). Of all respondents, 97.2% have an account on Facebook, 79.1% on Twitter, and 87.1% on Instagram. The students responded to the "I use Facebook to share information" statement on a 7-point scale with an average of 4.39 (with a standard deviation of 1.67). They responded to the same statement, but related to Instagram, with an average of 4.00 (with a standard deviation of 1.81), related to Twitter 4.08 (with a standard deviation of 1.83). The result for the "I use ___ to meet new people" statement for Facebook was 2.94 (with a standard deviation of 1.83), on Instagram 3.20 (with a standard deviation of 1.89), and on Twitter 2.94 (with a standard deviation of 1.82). The result for the "I use ___ to relax" statement for Facebook was 3.69 (with a standard deviation of 1.76), on Instagram 4.58 (with a standard

deviation of 1.70), and on Twitter 3.78 (with a standard deviation of 1.80).

In their research, Karal and Kukoc [9] investigated the purposes of use and frequency of student population activity when using Facebook. Using the questionnaire method, 1818 students were examined at undergraduate (1190) and graduate (628) levels of education. Analysing the results, 79.2% of undergraduate students have 0 - 200 friends on Facebook, and similarly with graduate students, 73% of them have 0 - 200 friends on Facebook. In terms of daily time spent on Facebook, 42.1% of Facebook students use less than 30 minutes a day. 34% of Facebook students use 1 hour a day, 17.8% of Facebook students use 2 - 3 hours a day, and 6.1% of students use it 4 or more hours a day. Like the previous research, the Likert scale was used for some questions but this time 5-point scale. The students responded to the statement that they use Facebook to search for friends on average from a 2.00 to a 5-point scale (with a standard deviation of 0.82). For using Facebook for video and photo sharing, the average score was 2.73 (with a standard deviation of 0.83), for participating in group discussions 1.76 (with a standard deviation of 0.83), and for sharing study materials 2.04 (with a standard deviation of 0.88).

Tomljanović [10] analysed the impact of 2.0 tools in learning programming based on a survey conducted on 160 respondents. The survey covered various issues related to the use of social software and the Internet in general. 71.25% of survey respondents said they use Facebook. One of the aims of the research was to prove that the use of social software in learning has a positive effect on the achievement of learning outcomes and the mastery of programming materials. The students were divided into 2 groups. The first group was preparing for the second midterm exam in the classic way, and the second group was preparing using Web 2.0 tools, including Facebook. Everyone was preparing for the first midterm exam without using Web 2.0 tools. Analysis of the results showed that the second group significantly improved their score at the second midterm exam compared to the results from the first midterm exam. The first group of students didn't see progress in the second midterm exam.

Regarding the objective benefit of using social media to educate students, Faizi, El Afia and Chiheb [11] state that social media improves communication and interaction among students, as well as interaction between students and professors. Thanks to such platforms, students and professors communicate with each other during and outside classes. They also state how social networks help shy students share their ideas and express opinions without face-to-face communication. An additional advantage is the collection of works and ideas in one place where they can be worked on in groups.

III. RESEARCH ON USING SOCIAL NETWORKS AMONG STUDENTS

Today, students in the Republic of Croatia are on average between 19 and 24 years old, and most of them have already used some of social networks in elementary school. This makes them excellent candidates for research on the use of social networks at the university level. For the purpose of this paper at the Faculty of Organization and Informatics (hereinafter: FOI), students of FOI were selected as respondents.

A. Questionnaire

The questionnaire contains questions regarding the use of Facebook, Instagram and Twitter. Gupta and Bashir [12], in their work on a questionnaire related to the use of social networks among students, describe the development of a questionnaire of a total of 56 questions, which during the validation were reduced to 19 questions in the form of a 5-point Likert scale. According to the authors, four dimensions of the use of social networks are examined through questionnaire questions [12]: academic, socialization, entertainment, and informative. In addition to questions based on the 5-point Likert scale, the survey includes demographic questions about respondents, year of study, amount of time spent on individual activities on the social network, number of followers and friends on the social network, etc.

B. Method of data collecting

The survey was conducted among FOI students over a 14-day period (May 17, 2019 to May 31, 2019). The survey was created in the Google Forms tool and distributed to students via students' Facebook groups and FOI e-learning system (to inform students who are not using social networks or are not in the groups where the survey was published). Google Forms offers the ability to create dynamic surveys that greatly facilitate surveys. The questionnaire is designed dynamically so that questions are generated for each user separately depending on their answers to some key questions.

IV. DATA ANALYSIS AND RESEARCH RESULTS

During the 2 weeks of the survey, a total of 217 responses were collected. Of the total number of respondents, 54% were female students and 46% were male students. 58% of the students were students of the university undergraduate study programs in informatics, 20% of students were studying in the vocational undergraduate study program in informatics, 13% of students were students of university undergraduate study program Economics of Entrepreneurship, and the other 9% of students were in some of 5 graduate study programs. Most respondents, 46% of them, were at the time of the survey attending the 3rd year of undergraduate study. 29% of respondents were from the 2nd year of undergraduate study, and 16% of respondents were from the 1st year of undergraduate study. 4% of students were in the 1st year and 5% of students were in the 2nd year of graduate studies. 44% of the respondents were 20 or 21 years old, which corresponds to a student of 2nd or 3rd year of study.

At the beginning of the survey on the use of each social network, the respondent was asked if he or she had an account on that social network, in the first place, in order to know whether to ask any further questions about the use of the same social network. Out of 217 respondents, only 4% of them had no Facebook account. As for Instagram, 23 of respondents did not have an

Instagram account. Twitter was the least popular social network among FOI students, with only 19% of students among the respondents having an account.

Following the question of having an account on a particular social network, the students were asked how much time they spend on weekdays on Facebook, Instagram and Twitter, and how much on weekends. Of the 209 respondents who had a Facebook account, 2 did not use Facebook at all and 1 used it only during weekends. Of the respondents using Facebook, 25% answered that they only used Facebook for 11 to 30 minutes a day on weekdays, and 3% of respondents answered that they used Facebook for more than 5 hours a day on weekdays.

Of the 168 respondents who had an Instagram account, 5% of them did not use Instagram at all, a significantly higher number than on Facebook. As with Facebook, only 1 user did not use Instagram on weekdays. Although Instagram has fewer users than Facebook, Instagram was used more than Facebook over the week. Even 23% of respondents used Instagram 1-2 hours a day on weekdays, which is a significantly higher percentage than in comparison to Facebook for the same answer. The number of respondents who used Instagram for more than 5 hours a day (3% of respondents) is relatively similar to the number of users who spent more than 5 hours on Facebook. To make good comparisons of Facebook usage on weekdays with Instagram usage on weekdays, respondents' results on their usage should be displayed in the same graph. In order to create such a graph, the data to take into account must be from the respondents who used both social networks. The collected data on the use of social networks are in interval form, and numerical values are required to create a graph, so that the intervals are converted to arithmetic mean. Figure 1 shows the link between daily Facebook usage and Instagram on weekdays in minutes. The y axis shows the minutes of daily use on weekdays and the x axis represents respondents. It can be seen in the picture that the line representing Facebook usage is growing as it approaches the last user. This is because the data is sorted first by the Facebook usage and then by the Instagram usage. The data becomes useful when we read the graph in such a way that for each respondent on the x axis we look at how much Facebook uses and how much Instagram. The users who spent less time using Facebook spent more time with Instagram, while users who spent more time using Facebook spent less time with Instagram.

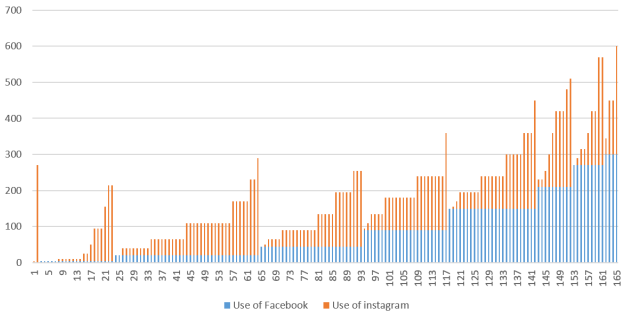


Figure 1: The relationship between daily use of Facebook and Instagram on weekdays in minutes

Table I shows the matrix of correlations between daily usage of Facebook, Instagram and Twitter on weekdays. On the main diagonal of the matrix, all the coefficients are 1.0. This is because the correlation is computed over two exactly the same variables among which the highest possible correlation is. The correlation coefficient between Facebook use on weekdays and Instagram on weekdays is 0.13, so we know that their connection is almost non-existent. Comparing the correlation between Facebook and Twitter usage on weekdays, the correlation coefficient is -0.32. Such a correlation coefficient shows a weak correlation between the observed variables and its negative sign indicates a negative relationship between the variables. It can be interpreted that by reducing the daily use of Facebook among respondents, the use of Twitter was increased on weekdays. As for the link between using Twitter and Instagram, a correlation coefficient of 0.07 shows that there is almost no correlation between the two.

Table I: Matrix of correlations between daily Facebook usage, Instagram and Twitter on weekdays

	Facebook daily use on weekdays	Instagram daily use on weekdays	Twitter daily use on weekdays
Facebook daily use on weekdays	1.00	0.13	-0.32
Instagram daily use on weekdays	0.13	1.00	0.07
Twitter daily use on weekdays	-0.32	0.07	1.00

With only 19% of users among the students surveyed, Twitter was not nearly as popular as Facebook and Instagram. 25 respondents (61%), more than half of those who own a Twitter account did not use Twitter at all. It can be assumed that the users opened a Twitter account with the intention of using it, but it probably did not meet their expectations. Of those who used Twitter (39% of the total number of respondents who had a Twitter account), 5% of the total number of respondents with a Twitter account used it only on weekends. The top Twitter users spent between 1 and 2 hours a day on weekdays.

As for the use of social networks over the weekend, it is assumed that the number of hours of daily use will increase in relation to the use of weekdays. The results show that the time of daily use on weekends, with respondents using Facebook and on weekends, is higher than the use on weekdays. More respondents said that Facebook uses 2 to 3 hours on weekends, than 2 to 3 hours on weekdays. The number of respondents who used Facebook from 4 to 5 hours a day had also increased, and the number of those who spent more than 5 hours a day on Facebook was equal on weekends and weekdays. A similar situation was with Instagram and Twitter, where there was an increase in the weekend usage.

The survey asked respondents how much of the total hours they spent on Facebook (weekends or weekdays) playing games. 90% of respondents to this survey did not play games on Facebook. Of the respondents who played

games, none spent more than 1 to 2 hours a day playing games (both on weekends and weekdays). Most gaming respondents spent up to 30 minutes a day on this activity. Men are more often associated with computer games today, so it would be interesting to examine the gender correlation of respondents with playing games on Facebook on weekdays and weekends. Figure 2 shows the number of gender respondents who played games on Facebook on weekdays. It is very interesting that female respondents who played games on Facebook spent more time each week on this activity than the male ones.

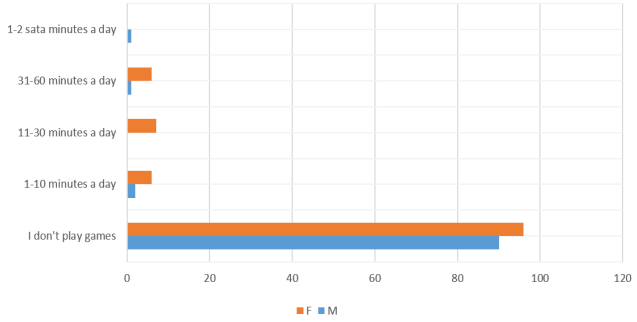


Figure 2: Number of respondents by sex who play games on Facebook weekdays

On weekends, the number of male respondents who play games was higher than on weekdays, but the female gender still prevailed in the time spent playing games, although there were more female respondents who did not play games on Facebook on weekends at all.

20% percent of Facebook respondents had between 501 and 1000 friends, and fewer (3%) had more than 1500 friends. Also, fewer people had up to 50 friends on Facebook (3%). Facebook offers the ability to follow people, in addition to friendship. The difference is that friendship requires confirmation from both users, while tracking does not. Instagram offers connectivity to users solely through monitoring. As noted above, user tracking does not have to be two-sided, so the number of people's followers and the number of people they follow are usually different. The results show that, on average, respondents followed fewer people on Instagram than they had friends on Facebook. For a better understanding of the relationships between the variables of the number of friends on Facebook and the number of followers on Instagram, the scatter diagram shown in Figure 3 was made. A certain grouping of the respondents' answers is visible and it can be said that there is a significant correlation between these two variables, that is, there is a positive dependence between them.

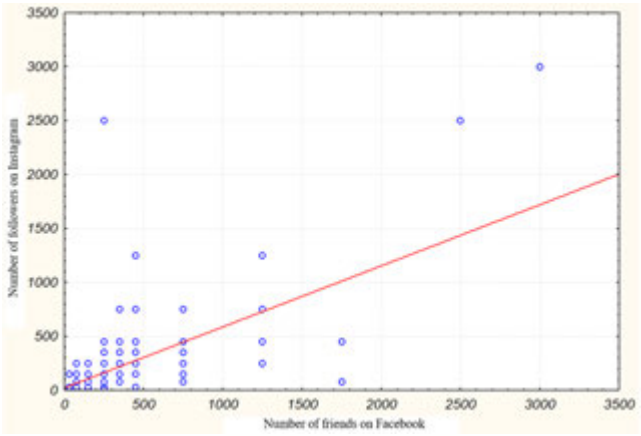


Figure 3: Scatter diagram of the number of friends on Facebook and the number of Instagram followers

Twitter was the least popular among respondents, which was probably one of the reasons why Twitter was the only social network (out of 3 social networks in this survey) for which several users replied that they did not follow anyone on Twitter and that no one was following them. From this, it can be concluded that Twitter may have "lost" these users at the very beginning of the use, even before searching for the people they wanted to follow. On the other hand, it is likely that these respondents did not follow any of the individuals, i.e. people, but followed some organizations, portals, websites, etc. Regarding the people who the respondents know personally outside Twitter, 44% of respondents knew no one outside of Twitter, 54% of respondents knew up to 10 people, and only 2% of respondents knew 11 to 50 people personally outside of Twitter. Only 8 respondents answered that they followed up to 10 FOI students on Twitter, and all others did not follow anyone from FOI.

Depending on the number of friends on Facebook, Facebook users are likely to have a certain number of minutes they spend on Facebook on weekdays. Someone with more friends may spend more time on Facebook because they have more information generated to process, and on the other hand, someone with a small number of friends may spend a lot of time on Facebook for other reasons, such as playing games or reading news. While it could be expected that there was a significant correlation between the values of the Facebook friends counts and the amount of Facebook daily usage time, the scatter diagram showing Figure 4 says the opposite. It shows that the data are not clustered at all, which concludes that the presumed connection does not exist.

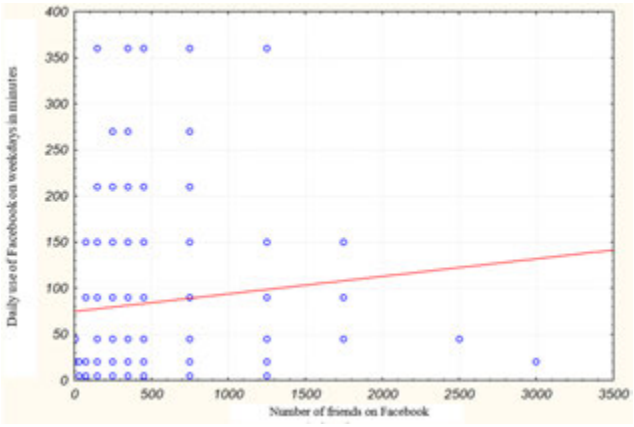


Figure 4: Scatter diagram of the number of Facebook friends and the amount of Facebook daily use

Regarding the devices that the students used to access social networks, tablets were the least used to access Facebook because 91% of respondents answered that they did not use the tablet at all for Facebook access. Only 2% of respondents answered that the tablet was used more than 50% of the time. Smartphones were definitely leading the way in the use of Facebook access, and desktops and laptops were slightly less represented. Instagram was initially just a mobile app for smart mobile devices and there was no desktop version. Because of this, the first Instagram users only used mobile devices to access Instagram. Most people use only a smartphone to access Instagram. 57% of respondents answered that in 100% of cases Instagram was accessed via smartphone. Only 1 respondent (0.6 percent) answered that he/she used a desktop or laptop computer in 100% of cases. The frequency of tablet use was similar to that of Facebook.

Using the 5-point Likert scale for each social network, the respondents were asked 19 statements related to the use of social networks. The analysis of the answers once again confirms the poor use of Twitter among the respondents, because Twitter has the highest percentages of "Never" in all statements. The statements are divided into academic, socialization, fun and informative.

Most respondents used Facebook to solve various study problems. 81% percent of the respondents said that they never used Instagram for this purpose, and 98% of the respondents related to Twitter. The respondents mostly used Facebook, then Instagram, and Twitter was extremely under-represented. A large number of the respondents interacted with friends through social networks to prepare for the exam, and they most often used Facebook, and less often Instagram. The respondents almost never communicated with friends regarding exam preparation via Twitter. For group learning, the students most commonly used Facebook, for which they averaged 2.42 on a 5-point scale, slightly less Instagram (1.23) and Twitter (1.00).

Students are often stressed out by a large amount of study obligations, so they use social networks to relieve stress. According to the survey, Instagram was the first choice to help relieve the stress of FOI students (with an average response of 3.19), followed by Facebook (2.44) and Twitter (1.70). Instagram was also a leader among students in sharing pictures and watching video content. The funny content of the people they are connected with is the most viewed by the respondents on Instagram.

When it comes to the informational dimension of using social networks, such as reading news and finding job related information, the respondents most commonly used Facebook. Interestingly, to share new ideas, the respondents liked to use Instagram a little more than Facebook.

Regarding the socialization dimension, the respondents most commonly used Instagram to become more social (average response is 2.18), followed by Facebook (2.06) and Twitter (1.49). Facebook also led the way with students who preferred to use the social network for social gathering. The results show that Instagram was ahead in the entertainment dimension, and Facebook in all other dimensions. Twitter was by all dimensions very underrepresented.

The number of the respondents' reactions to the statement "I use Facebook to ask for help from a professor" might be more interesting if we put it in the context of the year of study that the respondent was attending at the time. That relationship shows Figure 5. From the graph we can see that the help of professors was most often sought by students of the 1st and the 2nd year of the undergraduate study. It is possible that the cause of such a result was a lack of study-related experience among lower-level college students.

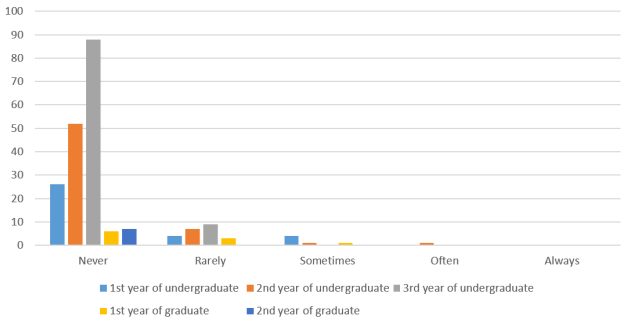


Figure 5: Number of reactions to the statement "I use Facebook to ask for help from professors" by year of study

Some students consulted with colleagues through social media during exam preparation. The question is whether students from all study programs at FOI used Facebook equally to communicate with friends to prepare for the exam. Figure 6 shows the dependency of the study programs and reactions to the statement "I communicate with friends via Facebook for exam preparation". The graph shows that the students were happy to communicate with other colleagues regarding the preparation of exams, and this is mostly the study program of Information and Business Systems. According to the graph, the respondents studying for graduate studies (at all but Information and Software Engineering) all answered that they always or often communicated with friends via Facebook to prepare for the exam. The reason for this result is the very small number of respondents from these study programs (from 1 to 3 students, depending on the study program).



Figure 6: Dependence study programs and responses to the statement "I communicate with friends via Facebook for exam preparation"

V. CONCLUSION

In order to investigate the use of Facebook, Instagram and Twitter by FOI students, a survey questionnaire was created to survey 217 FOI students over a 14-day period using Google Forms. Almost the same number of respondents of both sexes were surveyed, more than half of whom were undergraduate students in the Information and Business Systems program. Response analysis showed that Facebook was the most popular social network among survey students with more than 96% of respondents being Facebook users, while Instagram uses about 77% of them and Twitter is just under 19%. Surveyed students spent more time per day on weekends than weekdays, and interestingly, Instagram users spent more time on that social network per day than Facebook users using Facebook. Looking at the number of friends that respondents had on Facebook, the largest number, over 20%, had 501-1000 friends, while on Instagram, the number of people whose profiles followed the respondents was significantly smaller, so the largest group, more than 26%, was those who follow 101-200 profiles.

Facebook was also used by some students for academic purposes, and as a break from the study and stress it causes, FOI students most commonly used Instagram. The survey examined, among other things, the academic, socialization, entertainment and informational dimensions of social networking. Among students, Instagram is the leader in the entertainment dimension, and Facebook in all other dimensions.

Research conducted by Salem and Mengyan [8] showed that Instagram was the most used social network, not Facebook as was the case here. Twitter was also the least used social network in this research. The results partially match the statement about using the social network for stress relief. Both studies showed that Instagram was ahead in its use for stress relief purposes. However, the results of a survey conducted by Salem and Mengyan show that Twitter was right behind Instagram by use for stress relief, and in this research, the results show that it is the least used for these purposes. In terms of responding to the statement about using social networks to become more social and meet new people, both studies confirmed that Instagram was the most used for these purposes. In this research, Twitter was in the last place, and in the research conducted by Salem and Mengyan, Twitter and Facebook are used equally for the purpose of meeting new people.

Looking at the number of friends that the respondents have on Facebook, the largest number, over 20%, have 501-1000 friends and only 8% of respondents have less than 200 friends. Karal and Kukoc [9] found in their research that 79% of the undergraduate respondents had 0-200 friends on Facebook, which is significantly less than in this survey. One of the main reasons must be the earlier stage of social networks usage in the year 2012 which was 7 years before this research.

In his paper on the impact of 2.0 tools in programming learning, Tomljanović [10] shows how social software has a positive impact on learning and on learning outcomes. In this paper, the average response to the statement about using the social networking site Facebook for group learning is on a 5-point scale of 2.42. We can assume that students use social networks even more for independent learning. This confirms that students are aware of the benefits of using social networks for learning purposes and that is why they use it.

REFERENCES

[1] NB. Ellison, C. Steinfield, C. Lampe, *The benefits of Facebook "friends:" Social capital and college students' use of online social network sites.* Journal of computer-mediated communication, 12(4), pp. 1143-1168, Jul 2007.

[2] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, N. Christakis, *Tastes, ties, and time: A new social network dataset using Facebook.com*. Social networks, 30(4), pp. 330-342, Oct 2008.

[3] CM. Cheung, PY. Chiu, MK. Lee, *Online social networks: Why do students use facebook?*. Computers in Human behavior, 27(4), pp. 1337-1243, Jul 2011.

[4] AL. Young, A. Quan-Haase, *Information revelation and internet privacy concerns on social network sites: a case study of facebook*, InProceedings of the fourth international conference on Communities and technologies, pp. 265-274, 2009.

[5] E. Lee, JA. Lee, JH. Moon, Y.Sung, *Pictures speak louder than words: Motivations for using Instagram* Cyberpsychology, behavior, and social networking, 18(9), pp. 552-556, Sep 2015.

[6] S. Alhabash, M. Ma, A tale of four platforms: Motivations and uses of Facebook, Twitter, Instagram, and Snapchat among college students?, Social Media+Society.;3(1), Feb 2017.

[7] J. Phua, SV. Jin, JJ. Kim, Uses and gratifications of social networking sites for bridging and bonding social capital: A comparison of Facebook, Twitter, Instagram, and Snapchat. Computers in Human Behavior, 72, pp. 115-22, Jul 2017.

[8] A. Saleem and M. Mengyan, *A Tale of Four Platforms: Motivations and Uses of Facebook, Twitter, Instagram, and Snapchat Among College Students?* Michigan State University, USA, pp. 1-13, January-March 2017.

[9] H. Karal, and M. Kokoc, *Social Networking Site Usage among University Students: Differences of Educational Level,* Center for Research and Application in Distance Education, Karadeniz Technical University, Croatian Journal of Education, pp. 629-654, May 2012.

[10] J. Tomljanović, *The impact of web 2.0 tools for learning programming,* Varaždin: Faculty of Organization and Informatics. 2011.

[11] R. Faizi, A. El Afia and R. Chiheb, *Exploring the potential benefits of using social media in education*, International Journal of Engineering Pedagogy (iJEP), 3(4), 50-53, 2013.

[12] S. Gupta, and L. Bashir, L. *Social Networking Usage Questionnaire: Development and Validation in an Indian Higher Education Context*. Turkish Online Journal of Distance Education, 214-227, 2018.

Digital technologies in education of preschool children – preparing for future

J. Gunčaga*, E. Severini**, B. Kožík Lehotayová** and P. Ostradický**
* Comenius University in Bratislava, Faculty of Education,
Department of Didactics of Natural Sciences in Primary Level,
Račianska 59, 813 34 Bratislava,
Slovak Republic
** Comenius University in Bratislava, Faculty of Education,
Department of Pre-primary and Primary Education,
Račianska 59, 813 34 Bratislava,
Slovak Republic
guncaga@fedu.uniba.sk severini@fedu.uniba.sk lehotayova@fedu.uniba.sk ostradicky1@uniba.sk

***Abstract* – This contribution presents research findings about possibilities of using digital technologies in education in pre-primary level in Slovakia. Many discussions are devoted to the impact of digital technologies on the quality of pre-primary education, but also to the use of these technologies in kindergartens. The implementation of digital technologies and electronic digital aids in pre-primary education is one of the alternatives to developing complex literacy for preschool children. Observation aided research brings research findings that primarily support the promotion of the learning process of pre-school children manifested by increased internal motivation and self-reliance in the learning process. Research findings create the basis of the subsequent research direction oriented to the learning outcomes of children during the period of several years.**

Keywords – child; digital competences; digital technologies; pre-primary education; teacher

I. INTRODUCTION

The digital technology is an umbrella notion for all kinds of communication devices. Development in this area does not stop over time but expands and it brings new quality for many parts o society. In practice, it means preparing new digital content that requires continuous learning. Society is becoming a huge learning community. Digital world connectivity is created, and the new competences acquired allow for quick access to ideas and experiences from a wide range of people, communities and cultures. Our planet becomes increasingly "smaller" as nations in our post-modern era, are under the influence of digital technologies. We can see many convergent tendencies in our culture and states around the world. A cosmopolitan meaning is formed in the society, education and culture.

We can understand digital technologies in the context of education as a set of means, tools, environments and practices that are used for teaching and learning, communication, collaboration, presentation, creation and other factors, which make influence for the comprehensive development of a child in the learning process. We have seen great progress in the development of digital technologies in the last years of the 20th century and the beginning of the 21st century. The tools for recording, processing, transmitting and reproducing data have been improved to a level, which we had never known years ago. We can speak today about the time of digitization, if we observe the development of society.

Digital technology is an important and essential part of the learning process around the world. This process is characterized by the fact that the child learns globally and acquires new experiences, knowledge, values and attitudes every day through digital technologies that help in his / her personal development. For this reason, we include among important competences also digital and information competencies that need to be developed in pre-primary education. Their continuation is developed in further education.

The child meets itself very early in the nowadays globalised society with development and using of digital technologies (in the surrounding of his / her family). Partially, digital technology has place also in the teaching process in schools. For this reason, we connect digital technologies with development of digital literacy of children.

A. Developing competences by subjects in the teaching and learning process through digital technologies

The competences of the child are conglomerates, which are content-related to the whole teaching process. They express the current state of competence in the every stage of the solving of the various learning situations of children. They are based on the child's individual potential and level of its current development. The competencies of the child are compact units that express how the child can handle the process, content, goals and other learning situations in the reality. These components are connected with the teaching in relation to the specific school context and knowledge of current sciences. We can observe these competences in various complexity, because there exist competences in the form of metacompetencies as well as microcompetencies.

According to [9] competences form clusters and according to actual use, some competencies can be integrated under other competences. They regroup according to current problem solving and also, they dynamically change and develop (are constructed and reconstructed according to the dissemination of the child's thinking in the context of his/her activity) and often they are intertwined in cognitive part of the competence (one part of the cognitive structure of one competence overlaps with the cognitive structure of another / other competence).

Acquisition and improvement of competences is considered as a lifelong learning process of the human. This process is realized not only in the family, cultural and social life of some human, but also in school. The main goals nowadays include identifying and defining new transferable competencies, how and whether these competences can be better integrated into curricula, how they can be maintained and learned throughout the whole life of some human.

According to [4] we can generally understand digital literacy as a set of competences of knowledge by subjects. These subjects transfer their knowledge to appropriate, safe and productive use of digital technologies for teaching and learning. It is a set of capabilities related:

- to the use of digital resources for own needs, knowledge, presentation and comprehensive development;
- to effective solution of problem situations in the digital environment; decision making and use of appropriate digital technologies for information acquisition, processing, usage and presentation;
- to critical assessment and analysis of knowledge obtained through digital resources;
- to understanding the socio-cultural implications (security, privacy and ethics), which are caused by the use of digital technologies in the digital world.

Digital competence involves the confident and critical use of information society technologies (tools, etc.) for various purposes. This competence is based on basic capabilities for the use of digital technologies. The usage of a computer for the acquisition, evaluation and assessment, storage, creation, presentation and transfer of information is an important part of this competence. We can add here the capability of mutual communicative exchange and participation in collaborative (and social) networks via the Internet. The basic knowledge, competences, values and attitudes related to this competence are as follows: digital competence requires an appropriate understanding of the nature, roles and opportunities of using technologies in the information society; in everyday contexts, in personal and social life, as well as in various professional activities. The mentioned aspects include major computer applications such as word and spreadsheets, databases, information storage and management, and an understanding of the opportunities and potential risks associated with the Internet and electronic media communications (e-mail, network tools, social networks) in the professional life, leisure, information sharing or networking, education and research.

Individuals should also understand how information society (IS) technologies can promote creativity and innovation and be aware of issues related to the validity and reliability of the available information, as well as the legal and ethical principles of interactive usage of these technologies. The necessary competencies include the ability to search, to group and to process information; to exploit them in a critical and systematic way; to evaluate and to assess relevance; whether to distinguish between reality and the virtual world; as well as recognize links.

Individuals should be able to use IS tools to create, present and understand complex information and to access, to search and to use internet-based services. Individuals should also be able to use digital technologies (in today's IS) to support their own critical thinking, creativity and innovation. The usage of digital technologies requires a critical and reflective attitude towards available information and responsible and safe use of interactive media. This competence is also developed by interest in participating in (socially shared) communities and internet / intranet networks used for cultural, social and / or professional purposes.

We can understand according to [11], [14] digital competence as a two-stage competence. We speak at the first (basic) level about capabilities, which (together with their attitudes, values and knowledge) enable an individual to use digital technologies to acquire, to concentrate, to sort, to select, to evaluate, to store, to create, to protect and to exchange information, to communicate and to participate at the usage of the Internet. The second stage is characterized by the independent, confident and critical use of digital technologies (including Internet services), digital content and digital media. This level is characterized by critical (rational) thinking in the use of digital technologies in a higher level and in the identically developed level of the communication competence.

A competent teacher is a necessary factor for the effective teaching. A 21st century teacher, in the context of the usage of digital technologies, must be competent:

- to create and to edit digital sound;
- to exploit, socially back up and to share resources with / among children; to use blogs and wikis for creating of some online learning platforms focused on their learning subjects;
- to use digital images for use in a learning group; to use audio-visual and video content together with children; to use computer graphics to visually stimulate children;
- to connect with colleagues via social networks to ensure your own professional growth; to create and to provide presentations and training; to create an e-portfolio for his / her own and learning subject development;
- to have knowledge of online safety and security; to be able to detect plagiarism in their children's

products; to create videos, images, and video tutorials; to select web content suitable for teaching and learning process in a learning / playing group;

- to provide children with the skills needed for the usage of necessary tools such a "Task Manager".

Such teacher should have the possibility to organize and to plan optimal teaching; to know to select software options. Software can be used for example, for creating a real-time questionnaire for the learning group.

Such teacher understands issues related to copyright and fair use of materials. He / she is able to use computer and console games for educational purposes;

- to use digital technologies to create evaluation tools; to use collaborative tools for designing and editing / changing texts;
- to search and to rate web content; to use mobile devices (e.g. tablets, etc.);
- to identify safe online educational resources for children; to use digital technologies for making adequate and rational use of time;
- to know and to discuss the usage of YouTube and its potential in the learning group;
- to know the notation tools and to share content with his children in learning / playing group.

It is important to give / to share digital technology websites and resources after discussion in the learning group;

- to make usage of graphic, online and press organizers; to use online notebooks to get interesting ideas;
- to use the tools for creating and sharing tutorials and image recordings for movie recording;
- to use online tools in the learning group for reporting; to search effectively using the Internet in the shortest possible time; to carry out research via using digital technologies;
- to make usage of tools for sharing files and documents with children.

It is possible according to [8] to expand these competences for following activities:

- to know, to teach and to learn through the usage of plans and tables;
- to know and to teach research methods through digital technologies together with information management competencies;
- to know and to use digital image editing technologies;
- to reflect and to teach children the practical, critical and ethical use of network applications;
- to understand and to manage virtual classrooms; to use digital forums with children;
- to create, to compile and to publish digital books.

Digital technologies in education are a tool for obtaining information in connection with digital communication-based technologies, and their integration into education provides information resources, opinions, possible concerns. They allow to share learning experience of children through digital technologies, to learn from others, but especially to learn together with others (concept of collaborative learning); to interact with others for enriching each other.

II. QUALITATIVE RESEARCH

A. Action research

Action research in education according to [13] is one of the arts of scientific research in which teachers systematically examine their own didactic practice in order to improve it. It is usually carried out by teachers (often in collaboration with researchers) in the frame of their teaching practice, which is a part of their teaching process. Action research responds to the actual needs of didactic practice (sometimes only partial, which the researcher does not consider because he / she considers them less relevant). This research proposes solutions, which are validated directly in the learning process and in the case of effective result, the solutions are transferred to the learning process. Action research should be a part of the process of teaching of each teacher, who wants to fulfil the quality system of his / her teaching – developing of his / her teaching competencies.

The realization of the action research is in connection with the following statement according to [6]: „If it must to come to a conceptual change in the work of teacher, then it is necessary to intervene deep layers of the teacher's professional structure. The teacher can understand in this case: the existence of his / her own concept (content, structure, boundaries, etc.); the need and nature of the expected changes; ways of change - real activities, which enrich or partially reconstruct elements of the original individual concept".

We speak here according to [12] about the teachers, who will at the end change themselves, so that they can understand the current school system. If teachers participate in action research, they will understand the teaching process itself. What they (will) learn, will have a big impact to the learning process in the children's learning groups and in schools. At the same time, the development of teachers' professional development is the preparation of curriculum / themes / contents (focusing on the competence profile) by teachers and improving the initiative of schools. The goal of pedagogical research is to develop attentive reflection in order to support teachers' professional success.

It is possible to say according to [3] that: "School institutions are not only places where culture and ideology "are reproducing", but also where they are produced. The researcher should explicate the theoretical framework of his / her professional activity. He / she should clearly define which effects are acting on him / her, which phenomena he / she will study, which limits are possible

to apply to paradigms, and which theories frame his / her professional activity."

It should not be forgotten that pedagogical research, as an applied research, is conditional on the priority objective to help via reflection and criticism to improve the quality of the teaching process. Action research is according to [5] "targeted research conducted by a group of researchers (teachers) under the conditions of the teaching process."

This research according to [2] is characterized by spiral cycles of problem identification, systematic data collection, reflection, analysis, procedures which are managed on the basis of acquired data and redefinition. The goal of action research, as a part of the teacher's work, is to obtain practical results that can be used immediately in a didactic situation, while not excluding their contribution to the development of broader generalizations, or – as a higher degree – to the development of concepts and scientific theories. Research findings are disseminated as a rule by exchange between teachers within professional associations at school, lifelong learning education institutions and through higher education institutions. The methodology of the action research process is represented in a cyclic form, in a spiral or in a circle.

The decision for this type of research came from the knowledge that the phenomena under investigation are according to the angle of view of some human, which has influence on the choice of methodology and it affects the overall nature of the investigation. Subjects of our study were teachers of pre-primary education and we had the interest to know their opinions on the subject of study, which is also closely connected to other subjects of the teaching process (children). We preferred to identify opinions, beliefs, ideas, etc. that are of a human nature.

The phenomena examined in our research speak about a big variety of different phenomena. In order to understand the spectrum of the phenomenon of the child's independent action, it is necessary to give possibility, that the phenomenon will be conceptualized by the teacher. It is also important to note that the subjects of the qualitative research are researched in a natural environment, which gives both research participants the opportunity for cooperation. The essence of qualitative research is the collection of qualitative data in the form of words, images, which provide a detailed description of the phenomenon (situation) and they lead to the creation of a new theory. Qualitative research according to [15] gives possibility to find "the quality of certain educational activities, relationships and situations and their holistic (overall, complete) picture."

B. Research problem

A dynamic and changing world around us gives demands on education that have little to do with the educational goals of the second half of the 20th century, more precisely, on the educational practice that still exists in schools today. The current educational process is oriented to the innovation of the content of education and methods in the teaching process, but especially on preparing teachers of the pre-primary education with new competences for the 21st century.

It is possible according to [1] to use digital technology in the teaching practice in a systematically and appropriate way, in order to disseminate information and expand learning and teaching opportunities, as well as to help teachers to reduce a negative impact of the media and multimedia on the child.

C. The goal of the research

We would like to identify the impact of digital technologies and digital devices on developing children's competence and autonomy.

D. Research questions

- How is the competence of children in the teaching and learning developed, supported by digital technologies and digital aids?
- Which kind of educational offer, with the use of digital technologies and digital aids, is created and delivered by the teacher of pre-primary education to children?

E. Realization of the research

The research was based on a systematic reflection on:

- professional situations; research into pedagogical reality;
- initial diagnosis of children prior to investigation (indirect observation);
- study of literature on the subject; curriculum project planning;
- the implementation of a curricular project in the process of teaching, in which the continuous diagnosis of children was carried out;
- pedagogical reflection on realized teaching activities, as output diagnostics of children.

Unstructured observations in the framework of action research were organized in Slovakian kindergartens in the years 2016 and 2017.

We observed teachers, as well as children, in their learning / playing groups in the number (2 + 3 + 2 learning / playing groups) with a total of 7 observations.

The number of children (in groups) in the research according to state-recognized legislation (registered / children present) ranged up to 102.

Direct observations were conducted in the morning hours with a time span of two weeks. Microteaching in the framework of action research was organized in the same kindergartens in Slovakia in the years 2017 and 2018. We identified the presence of quality assertion of children's autonomy, as well as the nature of the emerging model to support the development of children's autonomy, competence and collaboration, through the use of digital technologies and through the implementation of microteaching in learning / playing groups, which were observed directly and indirectly in the mentioned kindergartens. Direct observations were made in the morning hours within the one-week time interval. The behaviour of the subjects and the search for phenomena

that were indicative of the presence of support and development of digital literacy of children were observed directly in the process of learning and teaching.

F. Processing of the research material

Analysing data was a very difficult activity, via which we obtained a large amount of research material. Its processing represents countless hours of thorough analysis and repeated entry into the terrain. Qualitative methodology requires the researcher to interpret the subjective concepts that research subjects own and conceptualize them in the form of mental abstraction so that they cannot in any way deviate from the original concepts of the observed subjects.

This requires according to [7] time and maximum concentration together with the researcher's responsibility. Thanks to that, there is a need to deal with such issues as for example, freedom, consciousness, wisdom, morality and, above all, the significance of human realities on the background of socio-cultural and historical constructing, the understanding of which is key to explaining lasting, valid and socially acceptable knowledge. Permanent (as a derivative of mentally elaborated experience), valid (as a socially valuable) and socially acceptable knowledge is continuously developed by members of society and at the same time legitimized through qualitative research of human realities by constructing consensus in dialogue (discourse) and intersubjectivity.

G. Analysis and interpretation of data from observation and microteaching

We focused in the analysis on a detailed description of the research material obtained by direct and indirect observation. We considered events and phenomena as indicators of phenomena to which we gave conceptual designation. Systematic reflection of professional situations, examination of pedagogical reality and initial diagnosis based on the evidence of children in the learning / playing group preceded the examination (indirect observation). Realization of continuous diagnosis based on the evidence of children in the learning / playing group was performed in the course of the teaching process and the outcome diagnosis based on the evidence of children in the learning / playing group was carried out in the pedagogical reflection on the realized teaching and learning activities within the microteaching.

During the research, a lot of material was obtained which took the form of video recordings and pictures / photographs. We repeatedly played / analysed them in the frame of indirect observations, focusing on teaching activities. We searched for significant elements which can document the development of children's autonomy, competence and collaboration ability through the usage of digital technologies.

PROTOCOL I. EDUCATION ACTIVITY: SEARCHING FOR SYMBOLS IN THE CITY / VILLAGE (ILLUSTRATION)

Description of the didactic situation in the teaching activity	*Usage of digital technologies in education*
Pre-active phase: The teacher evokes children by comparing the city / town map (printed and electronic) together with the children. Subsequently, the teacher presents the goal of the first phase in the teaching activity to the children, which was to search for and identify symbols on the map and to record the found symbols. *Interactive phase:* The children are divided into groups. Each group has its own map. At this stage, learners search, record, and identify symbols on the map with the help of a digital toy Bee-bot. The teacher is in the role of a partner, who advises the children in the identifying of a symbol that is unknown to the children. In the activity, children get to know the legend that serves to explain the symbols on the map. Learning subjects record the number of found symbols. They compare their results with other groups, because each group has the same map. *Post-active phase:* Children look for symbols identified on the map in the real world during a walk around the kindergarten. They record them later through digital technology-photographing machine, camera. They present the acquired videos and photographs through an interactive whiteboard at the end of the teaching activity in the class.	 Fig. 1 Pre-active phase- recording of symbols. The teacher applies in the teaching activity a model of teaching that focuses on asking open questions to the learning group of children. The usage of digital technologies (computer, software searching tool Google, digital toy Bee-bot, digital photographing machine, digital camera, interactive whiteboard) by children is at maximum level. The presentation of the topic of education is in framework points. The usage of digital technologies is effective, creative, non-directive, self-contained, competent, thought-out and planned by the learners in the learning process. Multidirectional communication is applied in the learning group of children, together with encouraging talk, discussion, justification, argumentation, approval and evaluation. It is supported by self-decision-making of learners in the use of digital technologies. Other supporting activities are: promoting learning group coordination and collaboration; enabling active, group participation in the curriculum, taking into account the learning aspects of children; assessment and (co) evaluation were part of the post-engagement (active) phase.  Fig. 2 Post-active phase – reflection of findings.

The reference framework was abstracted through the findings of direct and indirect observations (see: Table I.). Registration of the research results was evaluated on the base of the established reference framework with evaluation comments, which were used in the interpretation of the research results.

TABLE I. REFERENCE FRAMEWORK

Integration	Children are completely independent. They needn't any help. They fully develop their digital competences. They are fully competent and fully independent.
Contextualizing	Children are starting to be independent. They act independently, but they are independent not completely. They are partially independent. They haven't full digital competences / they are not fully competent.
Decontextualisation	Children learn to act independently and gradually they become independent. They are outside the scope of digital competences.
Coordination of points of view	Children are partially independent. They have minimal independence and they minimally develop their digital competences.
Decentralisation	Children haven't got any independence. They depend on the teacher and on instructions, models and decisions defined by teacher.

The microteaching analyses looked at the details, in particular the micro-facts, from which the nature of the emerging model of the support the development of digital literacy of children were deduced and inferred. We studied also the interaction of children and teachers, and their actions. Microteaching analyses provided facts, not only impressions or opinions and attitudes. We registered the actual behavior of subjects in the teaching process.

The basis of microanalyses of the educational processes of teaching according to [15] is an interaction fact that has the following attributes:

- the interaction is recorded on the medium (video recording);
- the captured interaction is thus fixed, allowing it to return to it and subject this fact to analysis, judgment and interpretation;
- captured and fixed interaction fact can be analyzed by more exact methods compared to classical observation, using mainly qualitative methods;
- chronological recording of interactions enables tracking of sequences, procedures, process ability of the educational process of teaching.

We realized constant comparison as a phase of qualitative analysis of realized qualitative research. Kolb in [6] argues, that if we apply the constant comparative method, "it is not the intention to verify universality, nor to confirm the reasons or other input data. This procedure ensures the saturation of the necessary information. The goal of the constant comparative method is not to verify but to generate theory. We observe the size and deepness of the information in the text until the saturation level is reached".

The constant comparative method according to [7] has the following levels: comparing events applicable to some category; integrating categories and their characteristics; definition of theory; conceiving theory.

PROTOCOL II. SIGNIFICATES FOR THE SUPPORT OF THE DEVELOPMENT OF CHILDREN´S DIGITAL LITERACY (SAMPLE)

Description of the teacher's activity during the teaching process	*Interaction fact*	*Description of children´s activity during the teaching process*
Teacher discusses with the children different proposals, how to realize the topic. Teacher writes children´s suggestions on the blackboard. Teacher accepts proposals for activities proposed by the children.	T ⇆ CH	Children propose the teacher to vote on the proposals, so they choose only one solution for realization of the chosen topic. Children shall respect the voting rules agreed by them jointly.
Teacher supports children in their decision-making, evaluation and solving. Teacher brings up a learning problem into the teaching process. Teacher encourages children to express their knowledge verbally for solution of the problem.	T → ⦸ CH	Children cannot decide and think independently. Children (as a learning group) are unable to suggest solutions for the solution of the problem. They cannot discuss how to solve the problem.
The teacher does not give children the opportunity to experience something. Children cannot fix their new knowledge based on their own experience. Teacher does not provide sufficient space for children to create their own ideas related to the subject of interest.	T ⦸ ← CH	Children gain new knowledge through their own experience in their area of interest. Children graphically represent the link between the concepts of interest. Children create a conceptual map via linking meanings, introducing concepts and relationships between them.
The teacher intervenes in the activity of children without asking. He / she helps children and He / she provides them material for production. Teacher does not let children create independently; he/she presents them possible methods of creation.	T → ← CH	Children prepare their own working material. They draw on a large wrapping paper a possible procedure for constructing the creation prepared with help of digital tools. Children evaluate and present independently their own product.
Teacher does not create enough space for independent activity and influence of children. He / she does not ask open questions for children for developing a discussion, but he / she presents his / her own knowledge obtained from experience.	T ← → CH	Children search in the real-world new symbols, which they identified on the map and record them with the camera. Children create a conceptual map on the base of their own initial concepts.
Explanations: T-teacher, CH-children, T ⇆ CH – identity in acting; T → ⦸ CH – without the answer to some acting; T ⦸ ← CH – without the answer to some acting; T → ← CH – conflict in the acting; T ← → CH – discrepancy in the acting.		

During the initial phase of direct and indirect examination we noticed partial support of the child's digital competence and independence by the teacher. The teacher applied non-directive approach in his / her teaching. We evaluate this phase via decontextualization according to our reference framework. Children learn to act independently, and they become independent step by

step. They are outside the scope of digital competence. The usage of digital technologies in the activities of learners is partially represented. We evaluate the middle phase of our research by contextualization. Children are starting to be independent, they act alone, but they are not completely independent. They are partially independent, lacking full digital competence. They are not fully digital competent in the teaching process that has been carried out through the usage of digital technologies. Teacher supported the autonomy, competence and collaboration of the learning subjects in the teaching process during the final phase of our research. We evaluate this phase by integration. Children are completely independent, they do not need help, they fully develop their digital competences. They are fully competent, fully independent. The usage of digital technologies by researchers has been maximized.

III. CONCLUSIONS

The most important research finding is that the development of autonomy and competence by children depends on many other characteristics of children, as well as teachers and adults (for example a close relationship between independence, competence, self-awareness, decision-making, participation, etc. that results in collaboration) and vice versa.

These facts answer the first research question: How is the digital competence of children in teaching via using digital technologies and digital aids developed? Research findings also point to the importance of child's personality development trajectory in supporting the development of children's digital competence as a fundamental idea of bringing "dependent" individuals (and hence children) into a state of relative "independence" that creates space for cooperation in some learning / playing group of children.

Our findings bring the result, that it is important to ensure education with high quality and digital accessibility in schools by developing of the digital literacy of teachers in the teaching process so that digital technologies can be used effectively in the educational process. It is equally important as the didactic incorporation of the use of digital technologies in the teaching process, because the use of digital technologies is not limited to one particular area of child development.

Another noted research finding is the need for the teacher / adult to respect children's interests in knowledge, role play, realized different activities, as well as respect the efforts to be self-contained, based on the children's inner need to learn, explore and discover new / unknown to them. We expect at the same time, that the children can use everything, what they have acquired. They will be able to use these things for their own benefit so they are able to solve different challenges of life.

It is important according to [10] that in the process of deliberately promoting the development of digital literacy, children are spontaneous and participate in the creation of a process in which they develop thinking, communication, creativity, self-awareness, respect for themselves and others; the ability to adapt, to agree and to cooperate in joint activities within the activities / plays and optimally assert in the learning / playing group of children.

Teachers involved in research have begun to consciously support the development of children's digital literacy in the framework of action research. What educational offer with the use of digital technologies and digital aids for children is created by a pre-primary education teacher? In the process of the educational offer from the side of teachers, many kinds of efforts have been made for integration of digital technologies and digital aids with an emphasis on independence, collaboration, imagination and complexity of view.

It was accepted and implemented fully as according to [16] in free decision-making and acting, responsibility, self-evaluation etc. in individual and group learning processes. Preference was more given to role-playing, debating than to some kinds of cooperative and collaborative learning. The more support and less external control in self-regulation the children receive from the adult (the teacher at school), the more structured, more comprehensive and purposeful their development is. Children become less dependent and more independent, competent, authentic and responsible.

Children have according to [17] the potential to learn and the teacher has different possibilities and didactically corresponding opportunities for full-scale learning and teaching. The teacher is a person with a greater and qualitatively different socio-cultural experience, the children are full of developmental goals and potentials, which was didactically reflected in the teaching process with a close relationship with the curriculum, as well with transforming and supporting components. The experiences of children from the pre-primary level of education are important for their learning at primary and secondary level.

ACKNOWLEDGMENT

Supported by grants KEGA 020KU-4/2018 and VEGA 1/0079/19.

REFERENCES

[1] N. Andjelković, Dete i računar u porodici i dečijem vrtiću. Beograd: Beoknjiga & CNTI & Savez informatičara Vojvodine, 2008.

[2] T. Gil Hernández, "El Aprendizaje Comprensivo y Creativo a Partir de la Investigación-Acción Como Estrategia Didáctica Epistémica en la Educación Básica," in Laurus-Revista de Educación. Universidad Pedagógica Experimental Libertador. Caracas: Venezuela, 2007, Año 13, número 023, pp 11–35.

[3] P. J. Goetz and D. M. LeCompte, Etnografía y diseño cualitativo en investigación educativa. Madrid: Morata, 1988.

[4] I. Kalaš, Spoznávame potenciál digitálnych technológií v predprimárnom vzdelávaní. Analytická štúdia. Bratislava: Ústav informácií a prognóz školstva, 2011.

[5] S. Kemmis and R. McTaggart, The action research planner. Victoria Australia Deakin: University Press, 1982.

[6] M. S. Kolb, "Grounded Theory and the Constant Comparative Method: Valid Research Strategies for Educators," in Journal of Emerging Trends in Educational Research and Policy Studies, 2012, 3 (1), pp. 83–86.

[7] D. Kostrub, Základy kvalitatívnej metodológie. Keď interpretované významy znamenajú viac ako vysoké čísla. Bratislava: Univerzita Komenského, 2016.

[8] D. Kostrub, E. Severini, and M. Rehúš, Proces výučby a digitálne technológie. Bratislava: Alfa print, s. r. o., 2012.

[9] D. Kostrub and S. Kikušová, "Perspektívy pregraduálnej prípravy učiteľov predprimárnej a primárnej edukácie z aspektu nadobúdania profesijných kompetencií," in A. Klim-

Klimaszewska et al. Kompetencje współczesnego nauczyciela. Siedlce: Uniwersytet Przyrodniczo-humanistyczny wsiedlcach, 2012, pp. 326–336.

[10] I. Pavlov, Profesijný rozvoj pedagogického zboru školy: námety na školocentrický a personocentrický model kontinuálneho vzdelávania učiteľov. Prešov: MPC, 2002.

[11] J. Pavlovkin and M. Ďuriš, Príprava učiteľov pre informačnú spoločnosť. Banská Bystrica: FPV, Katedra techniky a technológií UMB, 2012.

[12] J. Rudduck, "Changing the world of the classroom by understanding it: A review of some aspects of the work of Lawrence Stenhouse," in Journal of Curriculum and Supervision, 1988, 4 (1), pp. 30–42.

[13] E. Severini and D. Kostrub, Kvalitatívne skúmanie v predprimárnom vzdelávaní. Prešov: Rokus, 2018.

[14] Štátny vzdelávací program pre predprimárne vzdelávanie v materských školách. Schválilo Ministerstvo školstva, vedy, výskumu a športu Slovenskej republiky dňa 6. júla 2016 pod číslom 2016-17780/27322:1-10A0 s platnosťou od 1. septembra 2016. Bratislava: Raabe, 2016.

[15] Š. Švec, Metodológia vied o výchove. Bratislava: Iris, 1998.

[16] J. Záhorec, A. Hašková, and M. Munk, Particular results of a research aimed at curricula design of teacher training in the area of didactic technological competences. International Journal of Engineering Pedagogy, Vol. 8, No. 4, 2018, pp. 16–31. https://doi.org/10.3991/ijep.v8i4.8184

[17] J. Záhorec, A. Hašková and M. Munk, Teachers´ Professional Digital Literacy Skills and Their Upgrade. European Journal of Contemporary Education, Vol. 8, No. 2, 2019, pp. 378–393.DOI:10.13187/ejced.2019.2.378

Impact of education on communication etiquette

Ph.D. Giovanna Kirinić*, Professor, Ljubica Bakić-Tomić, Ph.D.**

*Elementary school „Ivo Andrić“, Zagreb
** University of Applied Sciences Baltazar Zaprešić, Zaprešić
gkirinic91@gmail.com
lj.bakic-tomic@ffzg.hr

***Abstract* - In the context of learning there is a strong consensus that students' perception about their interaction with teachers affects their motivation, outcomes and achievements. Although the communication process is embedded in all other common processes, few studies focus on the quality of communication, especially in education. Communication etiquette requires a greater number of communication skills due to the constant adaptation of situations in the environment. The survey was conducted on a sample of 257 teachers of primary education of the City of Zagreb and Zagreb County. The conducted research showed that primary education teachers show a noticeable higher level of communication etiquette (71.8%) than e.g. managers of social areas (43.21%). The reason can be found in two communication levels: the content and relation, whereby during initial education the relation level is often omitted without which there is no quality communication. It can be assumed that the student-professor relationship is more or less the same at all higher education institutions in Croatia, but that the social ones may differ from the technical ones regarding the content of courses which may contribute more to the development of communication culture.**

Key words* - *Communication culture; Communication etiquette; Teachers; Managers

I. INTRODUCTION

Despite the growing flow of non-standard (non-formal) communication in professional and private life, more relaxed communication and relationships, openness to different influences, and the intention that global communication becomes "easily understandable" to all actors involved in the process, more than ever there is a need to maintain ethical boundaries in communication in order to retain the correct relations as well as its further development. In the same way that we spontaneously learn our native language, we learn about the culture of the environment and cultural behavior through the communication process. In fact, communication is the only process in which people are acquiring and transferring their culture, through developed communication competences.

In the context of the school environment, the teacher has the role of a regulator of communication quality, where, adapting his way and the style of communicating to the student creates a quality starting point for the upgrading of the communication activity of the students [1]. The role of the teacher is multiple and complex, because "he educates by what he knows, and upraises by what he is" [2] which can be interpreted as teachers teach students more by the way they were taught, than by the way they have undergone professional education. Therefore, the importance of the teacher as a role model is undeniable. People are born with the capability of vocalization, but not with the knowledge, attitudes and skills that define communication etiquette or communication skills. The ability to learn how to communicate effectively and correctly exists, therefore it can and should be learned [3]. Notably for reasons of underdeveloped communication etiquette, because research results suggest that two-thirds of people believe that communication abilities are innate, and that they cannot and should not be taught and developed [4]. We can easily compare or connect educators from different fields of work, e.g. managers and teachers who have similar management, leadership, and organizational tasks and who basically use communication as their primary means of work. The success of a manager or teacher will largely depends on the development of communication knowledge and skills. However, when it comes to improving teacher communication skills, the problem is not only in the absence of adequate programs, but also in the need to personalize such education.

II. COMMUNICATION ETIQUETTE

Teaching is a form of communication [5], but mere transmission of information is not sufficient for the contemporary teacher. Communication skills are not innate to teacher, and some authors [6] believe that such competences consist of "intellectual skills and attitudes controlled by cognitive strategies." Once learned, they have a very wide application, so teachers need education to use communication skills tailored to the learning environment and learners. The research of the communication climate in the school shows that teachers are important communicators who create a positive climate that strongly influences the improvement of success [7]. Within the framework of the communication etiquette, its most important components being investigated are assertiveness [8],[9] active listening [10], congruency [11],[12] and empathy [13],[14],[15] in the communication process of teachers and pupils.

However, no conversation can happen without certain rules. The communication etiquette is the one that makes the recommendations for quality and efficient communication, or the rules of successful interpersonal communication, or an inventory of communication skills. Thus, Wahlroose [16] expanded Watzlawick's axioms into 20 rules of successful interpersonal communication, and Brajša [17] in 10 characteristics of interpersonal relations in the field of communication. This paper starts from a theoretical model based on Watzlawick's [18] definition of six axioms of communication, the basic characteristics in the observation of the communication relationship:

1. The individual should be observed in relation to others.
2. Interpersonal relationships should be studied through the communication process, since we interact with each other and experience ourselves in relation to others and to ourselves.
3. Feedback is essential to the behaviour and conditions of the behaviour of each individual in the communication.
4. The mutual influence of persons in a relationship is conditional on the situation and acts on communication itself.
5. There are varying degrees of awareness that condition the behaviour of an individual, which is important for a proper and objective approach to human behaviour.
6. Communication constitutes an individual's behaviour affecting the behaviour of others (Watzlawick, 1967).

Interpersonal communication takes place at the level of the relation(ship) and at the level of the content simultaneously. At the level of the relation, successful communication implies the existence of trust and cooperation, and also discusses the unconscious communication that accounts for 80% of overall communication. On the other hand, the level of the content requires the consistency of signs and symbols and the knowledge of their denotative (direct) and connotative (indirect) meanings. Therefore, in line with the content level of communication, the relation level is indispensable, because it speaks of the attitude towards the content of communication and towards the interlocutor. Educationally desirable communication in class is related to the realization of collaborative, symmetrical, two-way communication with the aim of the student's complete development.

III. METHOD

Communication culture, i.e. communication etiquette, leads to the need for a wider number of communication skills that enable constant adaptation of situations in the environment. In studies of Watzlawic axioms more often, individual axioms or certain aspects of etiquette are chosen, contrary to mentioned this work is done by researching the entire inventory, all aspects of etiquette, and very little research has been done within the aforementioned parameters.

The aim of this research is to determine the level of communication competences of Croatian teachers through the adoption of communication etiquette, which represents an inventory of communication skills to carry out quality communication among teachers and students.

A. Sample research

The survey was conducted in 2017 on a sample of 257 teachers in elementary schools of the City of Zagreb and Zagreb County. The basic set of respondents, primary education teachers in the Republic of Croatia, consisted of approximately 31500 teachers. At the time of the sample determination in the Republic of Croatia, 851 elementary schools were active. Due to the large number of schools and teachers, the sample is stratified and comprises 1% of the teachers' population in Croatia (approx. 300 teachers), working in the City of Zagreb and Zagreb County area (11.63% of schools in Croatia). Therefore, 17 primary schools of the area were selected, and to achieve a greater representativeness of the sample, the school selection criteria were: the location of the school center-sub-urbia, the size of the school (small, medium and large), geographical orientation east, west, north and the south. The sample is represented by teachers of the teaching of all educational areas: social and linguistic area, science area and pedagogy area according to National Framework Curriculum (Nacionalni okvirni kurikulum, NOK).

B. Research methods, instruments and hypotheses

The paper has used inventory of claims covering communication etiquette including the core components of the communication process: receiving messages, sending messages, and prejudices and blockages that make communication effectiveness more difficult. The inventory for testing the communication etiquette, edited by Walroose [16], and adapted to the Croatian language and culture by Bakić-Tomić [4], contains 30 claims of which 13 are correct and 17 incorrect. Inventory examines the self-assessment of knowledge and the application of quality communication of teachers. All claims can be divided into three categories: receiving messages, sending messages, and blockages and prejudices. The following are the null-hypotheses verified by this research:

H1 – Teachers do not achieve high-quality communication with students, i.e. they not respect the communication etiquette staying below of a minimum of 80% of the expected correct clearance of inventory.

H2 - Teachers belonging to various educational areas (socio-linguistic area, science and pedagogy area) do not differ significantly from each other in applying communication etiquette.

IV. RESULTS AND DISCUSSION

In this inventory communication etiquette, communication habits and teachers' skills that are most commonly used are examined. Regarding to docimology, the correct resolution should be at least 80% out of the expected 100%, in order to establish a satisfactory development of skills and habits. Taking into account the above criterion, the following characteristics of the communication etiquette of teachers are highlighted. The highest number of teachers responded to the following claims according to the expectation: Humor contributes to greater success in communicating 95.00%; In communication we are successful if we are flexible and spontaneous 91.90%; By gestures and body movements, one expresses far more than words 90.30%; Sometimes we are unaware of the message we sent, and the interlocutor still understands 88.40%; In an official conversation it is useful to adhere to objective benchmarks in order to communicate on an equal footing 88.00%; In conversation I often use analogies, comparisons or examples 81.40%.

Then the teachers correctly recognized the incorrect assertions: Humor is a sign of inseriousness and an escape from responsibility 93.40%; The boss needs his employees more carp than brag if he wants success in the business of 86.40%; When we listen to the speaker, attention should be paid only to the spoken message, and not to his behaviour 82.90%; Not trusting others is a good starting point in communicating 82.60%; I cannot accept the opinion of a co-speaker with whom I am not in good relations 81.40%; I speak more in conversation than I listen to other 80.20%.

For the following claims, only a small number of teachers responded according to expectations: I give the messages to the speaker so that the interloca must interpret them correctly 19.80%; I am trying to immediately clear up the negativity that I notice in the conversation 18.60%; Criticism undermines interpersonal relations 31.40%. In these few claims, communication prejudices have been observed. In the process of communication due to the subjective noises of the sender or the recipient of the message (prejudices, stereotypes, fears, expectations, plans, desires, interests), there is often misunderstanding, miscommunication and wrong comprehending of the message or information.

Some negative habits have been domesticated by teachers and are pronounced proper behavior, and could be substantiated by the motto "everyone does so". For example, the assertion: In conversation with the interlocutors, I insist on small things, it is an unfair communication technique that teachers have marked as correct (53.50% of teachers). On the other hand, in the inventory, even 68.60% of teachers find criticism in communication as good technique, which is incorrect because it is the use of negative messages. Also, for the communication technique that belongs to manipulation and maddening communication (is good to confuse the interlocutor), the teachers graded equally as accurate and inaccurate.

In the statement: I try to immediately clear up the negativity that I notice in the conversation for which 81.40% of teachers have opted out, it is noted that teachers are not listening carefully and do not use communication tools for mutual understanding (e.g. they are not sufficiently empathic).

Results show that teachers are insufficiently aware of the characteristics of quality communication as well as techniques to be used for this purpose. Teachers have a certain level of knowledge, but not competency, which is why certain claims remain only at the declarative level. For example, they say they can listen to others, but they do not apply it. Also, they only partially apply: metacommunication, feedback, assertive and congruent communication. The results showed that the Croatian teachers solved 71.82% of the inventory of communication etiquette, but still below the desirable 80% of the knowledge and skills of communication etiquette. Still, that is noticeably greater level of communication culture then e.g. managers in social areas (43.21%) measured by the same inventory [4]. Although the percentage of expectations is quite high, it has not reached a limit that would indicate quality teacher communication with students.

Hypothesis 1: *Teachers do not achieve high-quality communication with students, i.e. they not respect the communication etiquette staying below of a minimum of 80% of the expected correct clearance of inventory*, is confirmed.

A simple analysis of the One-way ANOVA with the educational areas as independent, and etiquette as a dependent variable was performed. The Scheffe test was used as a post-hoc test, and the cells for the Scheffe test showed p-values.

The results (Table 1.) have shown there is no statistically significant difference in applying communication etiquette between teachers of the pedagogy area and teachers of socio-linguistic area. There is also no difference between the teachers of science and socio-linguistic areas. The only statistically significant difference was found between the teachers of the pedagogy and the science areas and that in favor of the teachers of the pedagogy area ($F = 4.29$; $p = 0.01$).

TABLE 1. ANALYSIS OF DIFFERENCES IN EDUCATIONAL AREAS IN THE APPLICATION OF COMMUNICATION ETIQUETTE

Variable	Group	N	M	SD	F (p)	Scheffe	
						2	3
Educational area	Soc-ling.	108	21,53	2,84	4,29 0,01	0,40	0,19
	Sci.	88	20,92	3,46		-	0,01
	Edu.	61	22,44	3,07		0,01	-

The teachers of all educational areas concerning to communication etiquette, do not differ from each other excepting the teachers of pedagogy area and their higher quality of communication over the other two groups of

TABLE II. ANALYSIS OF DIFFERENCES BETWEEN TEACHERS OF DIFFERENT GROUPS IN THE AREA OF COMMUNICATION ETIQUETTE (RECEIVING, SENDING MESSAGES, PREJUDICES AND BLOCKAGES IN COMMUNICATION)

Variable	Group	N	M	SD	F (p)	Scheffe	
						2	3
Etiquette – receive messages	Soc-linq.	108	8,06	1,51	1,33 (0,26)		
	Sci.	88	7,75	1,89			
	Edu.	61	8,16	1,62			
Etiquette - sending messages	Soc-linq.	108	7,16	1,34	2,27 (0,10)		
	Sci.	88	7,11	1,63			
	Edu.	61	7,59	1,36			
Etiquette – prejudices and blockages	Soc-linq.	108	6,31	1,50	3,48 (0,03)	0,48	0,25
	Sci.	88	6,06	1,45		-	0,03
	Edu.	61	6,69	1,28		0,03	-

teachers (socio-linguistic and science), but only in relation to prejudice and blockage in communication where statistically significant differences were found ($F = 3.48$, for $p = 0.03$) (Table 2).

Hypothesis 2: *Teachers of various areas (socio-linguistic area, science and pedagogy area) do not differ from each other significantly regarding application of communication etiquette*, is not confirmed.

Yet, the teachers of pedagogy area are statistically significantly better in communication etiquette than teachers of other two areas. It is important to point out that teachers of all three educational groups remain under expected level of the knowledge and skills of communication etiquette.

V. CONCLUSION

Croatian teachers have solved 71.82% out of the inventory of communication etiquette from the expected 80%, therefore, no limit was reached that would indicate the quality of teacher's communication with students. Teachers use communication techniques that block communication which arise from bad communication habits. Results below the minimum expected knowledge and skills of communication etiquette showed that teachers sometimes know how to communicate at a declarative level, but do not have sufficiently developed listening skills, empathy as well as insufficient assertive communication. The biggest obstacles are the blockages and prejudices, which can be identified in applying of elements of manipulative communication, negative messages and criticism in the communication process. It can be assumed that the student-professor relationship is more or less the same at all higher education institutions in Croatia, but that the social ones may differ from the technical ones regarding the content of courses which may contribute more to the development of communication culture. This research was based on a questionnaire of teachers' self-assessment of communication culture, while for future more objective and complete research, it should examine how teachers and their communication culture are seen by their students. For future improvement, perhaps introducing or strengthening communication knowledge and skills courses not only in the field of future teacher education, but also in other areas of higher education institutions, should be strongly suggested.

REFERENCES

[1] I. Markić. Socijalna komunikacija među učenicima, Pedagogijska istraživanja, 7 (2), pp. 307–317, 2010.

[2] M. Bratanić. Paradoks odgoja, studije i eseji, Hrvatska sveučilišna naklada, Zagreb, pp. 43, 2000.

[3] S.P. Morreale, M.M. Osborn, J.C. Pearson. „Why communication is important: A rationale for the centrality of the study of communication," Journal of the Association for Communication Administration, 29, pp. 1-25, 2000.

[4] LJ. Bakić-Tomić. "Komunikološko-menadžerski profil rukovoditelja u policiji," doktorska disertacija, Sveučilište u Zagrebu, 2003.

[5] T. Wubbels, M. Brekelmans. „Two decades of research on teacher-student relationships in class". International Journal of Educational Research, 43, pp. 6-24, 2005. doi:10.1016/j.ijer.2006.03.003

[6] K.L. Medsker, J.P. Fry. „Acquisition of interpersonal communication skills: a research-based approach," Telematics and Informatics," 14, 3, pp. 214, 1997. doi:10.1016/s0736-5853(97)00005-1

[7] M.T. Wang, J. L. DeGol. „School Climate: a Review of the Construct, Measurement, and Impact on Student Outcomes," Educ Psychol Rev, Vol. 28, Issue 2, pp. 315–352, June 2016. DOI 10.1007/s10648-015-9319-1

[8] Goleman, D. Emocionalna inteligencija, Geopoetika, 1997.

[9] R. Urea. „The Impact of Teachers' Communication Styles on Pupils' Self- safety Throughout the Learning Process", Procedia - Social and Behavioral Sciences, Volume 93, 21, pp.164-168, October 2013. https://doi.org/10.1016/j.sbspro.2013.09.171

[10] H. Weger. „Instructor Active Empathic Listening and Classroom Incivility. International Journal of Listening," 32(1), pp. 49–64, 2007. doi: http://dx.doi.org/10.1080/10904018.2017.1289091

[11] H.G. Ginott. Teacher and child: A book for parents and teachers, New York, NY: Macmillan, 1975.

[12] K. Sakamoto, T. Laine, F. Ilya. „Deciding whether to deceive: Determinants of the choice between deceptive and honest communication", Journal of Economic Behavior & Organization, September, pp. 392-399, 2013. dostupno na: https://doi.org/10.1016/j.jebo.2013.05.003

[13] S.R. Covey. Sedam navika uspješnih ljudi, Mozaik knjiga, Zagreb, 1995.

[14] M. Bratanić. Mikropedagogija-interakcijsko-komunikacijski aspekt odgoja. Treće izdanje. Zagreb: Školska knjiga, 1993.

[15] P. Swan, R. Riley. „'Mentalization': A Tool to Measure Teacher Empathy in Primary School Teachers," AARE APERA International Conference, Sydney 2017. dostupno na: https://files.eric.ed.gov/fulltext/ED544520.pdf

[16] S. Wahlroos. Familiengluck kamm jeder lernen, Fischer Frankfurt/M, 1974.

[17] P. Brajša. Pedagoška komunikologija. Zagreb: Školske novine, 1994.

[18] P. Watzlawick. Menschliche Kommunikation. Bern: Hans Huber, 1967.

Harmonization of curriculum with the needs and requests of the Fourth Industrial Revolution: Case of Faculty of Economics and Business Rijeka

Heri Bezić, PhD*, Davorin Balaž, MEcon* and Barbara Buljat, MEcon*
* University of Rijeka, Faculty of Economics and Business Rijeka
** Rijeka, Croatia
heri.bezic@efri.hr, davorin.balaz@efri.hr, barbara.buljat@efri.hr

***Abstract* – The Fourth Industrial Revolution and its disruptive technologies shape everyday life and future of society. They have and will have a significant impact not only on the economy, but on each area of human life. Higher education that prepares students for life and professional careers represents a vital part of each economy. It is impacted by the Fourth Industrial Revolution and its technologies. The objective of this research is to determine the level of curriculum harmonization at the Faculty of Economics and Business Rijeka with the needs and requests of the Fourth Industrial Revolution and predictions about the future needs of the labor market. The research data has been collected through questionnaires filled by students and the teaching staff at the Faculty of Economics and Business Rijeka. The results will reveal obstacles and possibilities to increase the level of harmonization.**

Keywords - Fourth Industrial Revolution, Faculty of Economics and Business Rijeka, education, curriculum, harmonization

I. INTRODUCTION

Nelson Mandela said that education is the most powerful weapon people can use to change the world. The Fourth Industrial Revolution changes the world and affects education. Academia must respond to the challenges of the Fourth Industrial Revolution. Reference [14] states that education can increase the labor force's human capital, which will improve labor productivity growth. Increasing investments in education by 1%, causes an economic growth increase of 0.14%. Reference [7] states that growth differences are the results of differences in knowledge gained during studies. Education should adapt to changes caused by the Fourth Industrial Revolution. It can be done by harmonizing it with predictions about labor market. Literature provides different definitions of the Fourth Industrial Revolution. Reference [21] states that humanity stands at the brink of a technological revolution that will fundamentally alter the way people live, work and relate to one another. Reference [13] agrees with [21] and concludes that pace of changes is increasing. How can education adapt to changes caused by the Fourth Industrial Revolution? Reference [27] refers to Jack Ma's speech at the annual meeting of the World Economic Forum in Davos. Jack Ma states that teaching methods should be changed because they are knowledge based. Reference [4] estimates that the Fourth Industrial Revolution will create 123 million new jobs, but will cause the disappearance of 55 million existing jobs. Estimation is that 65% of children which attend primary school will have a job that doesn't exist today. In determining investment locations, 74% of the companies prioritize the availability of skilled local talent. Talent will be created through education that can provide a necessary set of skills for labor market of the 21st century. The question is which skills will be necessary? Reference [20] emphasizes these skills: analytical thinking and innovation, active learning and learning strategies, creativity, originality and innovation, system analysis and evaluation technology designing and programming, critical thinking and analysis, complex problem solving, leadership and influence, emotional intelligence, reasoning, problem-solving and ideation. This paper wants to examine the level of curriculum harmonization at the Faculty of Economics and Business Rijeka with the needs and requests of the Fourth Industrial Revolution. One objective is to reveal obstacles and possibilities to increase the level of harmonization. Through analysis, it will try to provide guidelines not only for the Faculty of Economics and Business Rijeka, but for all faculties.

II. LITERATURE REVIEW

The Fourth Industrial Revolution and education are subjects of huge scientific interest. Reference [25] states that the educational achievement of a country's population is a key determinant of its economic growth. Improving educational attainment is an urgent priority for all countries. Reference [23] concludes that between 75 million to 375 million people will need to switch occupational categories. The technological transformation is clear and now the question is how universities can respond. Reference [15] concludes that the technological age has opened a bright horizon for education offering tools to improve its quality. The computer revolution enabled greater access to learning and larger teaching opportunities. It allows individual progress towards learning goals. However, a problem appears in the period of rapid technological progress, when the gap between technological invention and teacher utilization becomes more noticeable. Reference [26] states that partnerships between governments, educators, training providers, workers and employers have to be enhanced. It leads to better managing of the transformative impact of the

Fourth Industrial Revolution on employment, skills and education. The evolving nature of work together with technologies of the Fourth Industrial Revolution will be the biggest change driver. Current technological trends are bringing about an unprecedented rate of change in the core curriculum content of many academic fields, with nearly 50% of subject knowledge acquired during the first year of a four-year technical degree outdated by the time students graduate. Reference [12] agrees with [4] about the necessary skills, which were already mentioned in chapter introduction. Reference [12] states that the skills students will need for the society in which they will work and live shouldn't be thought of as "one more thing to teach", but rather training integrated across all curriculum. References [1] and [16] agree with [12] and conclude that in preparation for their future life, students need to deal with the real-life problems during classes. Through the integration of science, technology, engineering and math, curriculum exemplifies the cross-curricular learning that is the foundation of the 21st century education. Reference [22] states that the Fourth Industrial Revolution changes everything. Education as one of the main goals of a sustainable development strategy should adapt to the new situations. Everything becomes digital: consumers, enterprises and operation waves. Reference [22] agrees with previously referred authors about the necessary skills that should be taught. Reference [19] concludes that education must adapt to the Economy 4.0 era, as it was changing due to industrial revolutions. Education 4.0 caters to the needs of society in an innovative era. Learning management must help to develop the learner's ability to apply new technologies which will contribute to increase their knowledge and skills according to the changes in society. Reference [8] agrees with this perspective and mentions nine trends in Education 4.0: learning can take place anytime and anywhere; learning is personalized; students have a choice in determining how they want to learn; students will be exposed to more project-based learning; to more hands-on learning through field experience such as internships, mentoring projects and collaborative projects; to data interpretation in which they are required to apply their theoretical knowledge to numbers and use their reasoning skills to make inferences based on logic and trends from given sets of data; students will be assessed differently, and the conventional platforms to assess students may become irrelevant or insufficient; student's opinion will be considered in designing and updating the curriculum; students will become more independent in their learning, thus forcing teachers to assume a new role as facilitators who will guide the students through their learning process. Reference [2] states that industries and private sector will have to collaborate with the universities to adapt to the requirements of the Fourth Industrial Revolution, because Education 4.0 will combine the real and virtual world. Reference [18] states that each industrial revolution caused changes in education due to demanded skills and technologies and by founding new institutions. The digital environment changes everything. Reference [18] emphasizes the program Stanford 2025 as an example of adapting higher education to the needs and requests of the Fourth Industrial Revolution. Reference [18] agrees with [9] which states that information technology (IT) will be highly incorporated into education. Technologies like virtual reality (VR) or augmented reality (AR) will be not only part of curriculums, but will provide the way students gain new knowledge. Massive open online courses (MOOCs) will become the future of the classroom experience.

Various authors stress the importance of curriculum harmonization with the Fourth Industrial Revolution. The curriculum has to be changed and more adapted to the real world. One of the questions for Education 4.0 is how to prepare students for the world that nobody knows how it will look like. Universities have to respond and one of the answers is Stanford 2025 [24].

III. METHODOLOGY

To examine the level of curriculum harmonization with the needs and requests of the Fourth Industrial Revolution, a questionnaire was used. The sample-set consisted of 255 people: 46 teaching staff, 209 undergraduate and graduate students at the Faculty of Economics and Business Rijeka. While constructing questions, relevant sources were referenced, such as The World Economic Forum, Stanford University, etc. In interpreting the results, questions were grouped into five elements: knowledge about the Fourth Industrial Revolution, knowledge about professions related to the Fourth Industrial Revolution, usage of technology, skills and exams. Curriculum harmonization with the needs and requests of the Fourth Industrial Revolution will be examined through each element. Knowledge about the Fourth Industrial Revolution (know) summarizes questions related to the estimated knowledge about terms and technologies of the Fourth Industrial Revolution. Knowledge about professions related to the Fourth Industrial Revolution (knowpro) summarizes questions related to the estimated level of familiarity with professions created by the Fourth Industrial Revolution. Usage of technologies (tech) summarizes questions related to the usage of the Fourth Industrial Revolution technologies in the teaching process. Skills (skill) summarizes questions related to developing skills necessary for the labor market of the Fourth Industrial Revolution. Exams (exam) summarizes questions related to the harmonization of assessments, questions and tasks with the needs and requests of the Fourth Industrial Revolution. The Likert scale from 1 to 5 was used to measure estimations from respondents. For elements knowledge about the Fourth Industrial Revolution and knowledge about professions related to the Fourth Industrial Revolution, value 1 means that respondents have no knowledge about term or profession. Value 2 represents a basic level where respondents are familiarized with the definition and existence of certain term or profession. Value 3, or intermediate level, means respondents possess theoretical knowledge of the term and job description of the profession. Value 4, or advanced level, means respondents are familiarized with theoretical, technical and basic practical knowledge related to term or profession. Value 5 means respondents possess high theoretical, technical and practical knowledge about the term or profession. For elements technology, skills and exams, values are composed in following way: 1 – strongly disagree, 2 – mostly disagree, 3 – neither agree,

nor disagree, 4 – mostly agree, 5 – strongly agree. In the last group of questions, respondents had to choose between two options about the type of studies they prefer. Option A represents the actual type of studies. Option B represents studies adapted to the Fourth Industrial Revolution. Option B is based on the Stanford 2025 program, which is a new curriculum based on the needs and requests of the Fourth Industrial Revolution created by Stanford University [24]. Descriptive statistics was used to determine the level of curriculum harmonization.

IV. RESULTS AND DISCUSSION

Using descriptive statistics, the authors plan to provide insights into the level of curriculum harmonization at the Faculty of Economics and Business Rijeka with the needs and requests of the Fourth Industrial Revolution. The last group of questions will be analyzed just to determine the preferences of respondents regarding the type of studies.

TABLE I. DESCRIPTIVE STATISTICS

Measure	know	knowpro	tech	skill	exam
MODE	4	3	1	4	3
MEAN	2.51	2.94	2.61	3.54	3.16
ST.DEV	0.80788	0.12746	0.15678	0.07395	0.07971

It is interesting to examine the level of knowledge about the Fourth Industrial Revolution and its technologies. Since this element contains 18 questions, it will be divided in two sections: knowledge about the Fourth Industrial Revolution and its technologies, and knowledge about the terms and indicators related to the Fourth Industrial Revolution. Questions in each section have a common denominator. For both sections, the common denominator is the similarity of results in each question and subject of the question. As it is written, the first section is more related to technologies, while the second is more related to indicators and theoretical terms. For this element, mean equals 2.5, while mode equals 4. The reason for this is the very low level of knowledge about terms and indicators related to the Fourth Industrial Revolution. For this section, mean equals 2.1, while mode equals 1. Terms related to the Fourth Industrial Revolution are: Global Competitiveness Index 4.0, Industry 4.0 Index, Global Connectivity Index, Digitalization Index, Index of Economic Complexity, Global Innovation Index, Global Talent Competitiveness Index, Global Information and Communications Technology Development Index, cognitive capitalism and massive open online courses. For the section related to knowledge about technologies of the Fourth Industrial Revolution, mean equals 2.96, while mode equals 4. Technologies of the Fourth Industrial Revolution in this research are: Internet of Things (IoT), Artificial Intelligence (AI), robotics, smart systems, big data, blockchain, 5G, virtual and augmented reality. Respondents do not possess enough knowledge about the indicators that provide insight into digital development of a certain economy. It is very interesting to notice that more than 50% of respondents estimate their knowledge about smart systems, VR, AR, and AI to be at an advanced or expert level. This means they can apply these technologies on a practical level. Respondents are most familiarized with the Global Competitiveness Index 4.0, Global Innovation Index and Digitalization Index. Even here, results show a very low level of familiarity. Among respondents, 60 - 70% of them are not familiarized with these indexes or just know about their existence. According to responses, the Global Innovation Index is the indicator respondents are most familiarized with. 16.4% of respondents estimate their knowledge about this indicator to be on advanced or expert level. Professions included in this questionnaire are: community manager, content manager, app developer, big data analyst, social media manager, influencer, YouTuber, cloud broker, digital marketing specialist, design team leader, communication assistant, iOS engineer, digital architect, 3D specialist, blockchain manager, bitcoin trader, AR manager, VR developer. Regarding professions, mean equals 2.94, and it is almost at the intermediate level of knowledge with mode value of 3. More than 50% of respondents possess basic, advanced or expert knowledge. For professions like YouTubers or influencers, more than 50% of respondents estimate their familiarity on expert level. Blockchain manager and AR manager are professions with the lowest level of familiarity. 63.9% of respondents are not familiarized at all or are only aware of the existence of these professions, while 10.1% of them estimate their knowledge on advanced or expert level. 66.6% of respondents are not familiarized at all or are only aware of the existence of an AR manager, but only 8.3% of them consider to have advanced or expert level knowledge. However, it is questionable why respondents estimate their level of knowledge and familiarity at such levels. For the second element, usage of technologies of the Fourth Industrial Revolution in the teaching process, mean equals 2.61, while mode equals 1. Respondents strongly agree that information is available online and online studies is enabled. They are generally satisfied with the quality of IT equipment. One of the reasons for this are investments. Faculty invested 337.000 € in hardware and software over the last five years [6]. Respondents strongly disagree with the fact that technologies are used in the teaching process. According to results, Internet of Things has the highest usage rate in the teaching process. Mode equals 3, while mean equals 2.36. For all other technologies, mode equals 1, while mean is smaller than 2. Smart systems are exception, with mode value 3 and mean value 2.36. 91% of respondents strongly or mostly disagree robotics is used in the teaching process, while only 2.8% of them strongly or mostly agree with the statement. 72% of respondents strongly or mostly disagree that AI is used, while nobody strongly agrees that AI and blockchain are used in the teaching process. 78% of respondents strongly or mostly disagree that blockchain is used in the teaching process. 80% of respondents strongly or mostly disagree with the fact that VR and AR are used in the teaching process. Skills included in this research are: creativity, teamwork, critical thinking, discussion and interaction, argument-based conclusion, personalization, problem-solving, working on projects, initiative, analytical thinking, innovativeness and complex problem-solving. Regarding skills necessary for the Fourth Industrial Revolution, mean equals 3.54, while mode equals 4. Respondents consider that discussion and interaction,

teamwork, critical thinking and argument based conclusion are skills that are mostly developed in the teaching process. According to respondents, the skill that is the least developed is innovativeness. 30% of respondents strongly or mostly disagree with the statement that this skill is developed in the teaching process. However, 40% of respondents totally or mostly agree that it is developed in the teaching process. Element exams includes: assessments, types of questions and tasks, such as midterms, seminars, case studies, calculus, essay questions, short questions and multiple choice. Midterms also include estimations about final exams because they are very similar in question types and examining student knowledge. For element exams, mean equals 3.16, while mode equals 3. 53.7% of respondents consider seminars are useful in the preparation for the labor market, while 49.8% of them consider the same for case studies. According to 45.9% of respondents, multiple-choice questions are considered useful for preparation for the labor market of the Fourth Industrial Revolution. 33% of respondents strongly or mostly agree that essay questions are useful in the preparation for the labor market of the Fourth Industrial Revolution, while 34.1% of them consider the same for short questions. These results are not in accordance with relevant literature about developing skills and preparing students for the labor market of the 21st century. Reference [3] confirms that the skills questioned in this research are necessary for the Fourth Industrial Revolution. Reference [3] states that standardized facts and procedures belong to the past. Multiple choice and short questions are used to test knowledge of standardized facts which do not belong to the age of the Fourth Industrial Revolution.

While analyzing the total questionnaire, the authors divided it into two parts: knowledge about the Fourth Industrial Revolution and the teaching process. Knowledge about the Fourth Industrial Revolution includes two elements: knowledge about technologies and terms related to the Fourth Industrial Revolution, and knowledge about professions related to the Fourth Industrial Revolution. The teaching process includes elements: usage of the Fourth Industrial Revolution technologies, developing skills and harmonization of assessments. Common denominators to divide elements into sections are: Likert scale, what it measures and the influence of respondents. In the first two elements, respondents estimated their level of knowledge and familiarity with terms, technologies and professions of the Fourth Industrial Revolution. The Likert scale is organized by the level of knowledge: no knowledge, basic, intermediate, advanced and expert level. The level of knowledge is a result of the teaching process and informal education. Respondents, especially students, have a higher possibility of influence on their knowledge. The teaching process produces knowledge. The Likert scale measures the level of respondent agreeance with the statements related to elements included in this section. Student influence is very low or doesn't exist at all, as they don't define literature, teaching plans, assessment types or questions. The teaching staff has a higher influence because they define literature and teaching plan, but influence is not absolute. The assessment types and lessons that will be taught are defined by the rules of the Faculty, University and Ministry of Education.

For both sections, mode equals 3. Mean for knowledge equals 2.72, and for the teaching process equals 3.11. The majority of respondents possess basic level of knowledge about terms, technologies and professions related to the Fourth Industrial Revolution. It implies that there is a lack of practice and simulations of real situations in education. Reference [11] states that, in order to increase the employability of students, it is necessary to concentrate on what happens in the real or virtual classroom and changing the formal curriculum. Employers generally prefer to hire people with work experience. One way of increasing student's competitiveness in the labor market is to incorporate practical content into degree programs. Practical usage and real-world simulations can increase student involvement and interest. For the second section, mode and mean value show that the majority of respondents neither agree, nor disagree with the statement. It is a result of the lack of knowledge about the Fourth Industrial Revolution and its requirements. It is especially expressed in responses to the questions regarding element exams, where respondents are not sure, which types of questions are harmonized with the needs and requests of the Fourth Industrial Revolution. These answers imply necessity to combine practical and theoretical lessons in curriculum and train teaching staff for teaching and using technologies of the Fourth Industrial Revolution. Education produces human capital and for this reason it is one of the key determinants of national economy competitiveness. It has to follow trends from the real world. The increasing level of cooperation with private sector can be one of the guidelines to incorporate new technologies and theoretical aspects of the Fourth Industrial Revolution.

The last group of questions was the type of studies respondents prefer. The first type, Option A, is the actual type. The second type, Option B, is harmonized with the needs and requests of the Fourth Industrial Revolution. This type is based on the Stanford 2025 program that is emphasized as a guideline for Higher Education 4.0. The majority of respondents have chosen Option B. Instead of immediately choosing a study and program strictly divided into topics, respondents prefer choosing interdisciplinary projects which would allow them to work during their studies. A university organized in competency hubs where project-based lessons are performed in specialized and equipped classrooms and laboratories is considered a better option than a university strictly divided on faculties and departments where lessons are performed in one room. Respondents prefer stronger cooperation with Alumni, whose members stay in contact with the Faculty as lecturers, project leaders and donors instead just as lecturers only on particular events. Through such projects, students are enabled to study social and global impacts of their projects, communicate with Alumni and future employers instead of being unaware of impacts of their projects and maintain actual communication with employers only through Curriculum Vitae (CV). In this way, employers could track student advancement in adopting necessary skills and students can become more employable. Instead of a job that is unrelated or partially

related to their field of study, respondents prefer an option where their projects are basics for their professional career. In two questions, respondents prefer Option A. Studying three to five years immediately after high school is a better option for 55.7% of respondents than studying six years during whole professional career. Attending defined number of classes through semester that lasts 15 weeks is a better option for 51.8% of respondents than class based on solving project tasks in micro cycles which last maximum one month.

V. CONCLUDING REMARKS AND CONTRIBUTIONS

In this paper, the authors wanted to examine if curriculums at the Faculty of Economics and Business Rijeka are harmonized with the needs and requests of the Fourth Industrial Revolution. According to the data, authors got through statistical analysis, the curriculum is mostly harmonized in the element skills. This means the curriculum prepares students to be competitive in the labor market of the Fourth Industrial Revolution. According to relevant literature, Faculty follows global trends. The element with the lowest level of harmonization is the usage of technologies of the Fourth Industrial Revolution in the teaching process. The authors noticed an interesting comparison of the results for knowledge about technologies and usage of technologies. For element knowledge, mode equals 4, while mean equals 2.97. For element usage of technologies mode equals 1, while mean equals 1.5. This raises the question: How did respondents learn about these technologies? The question remains unanswered and the authors consider it as one of the limitations of this research. In these elements, the curriculum needs certain improvements in order to be harmonized with the needs and requests of the Fourth Industrial Revolution. Lack of knowledge about the indicators like Global Competitiveness Index 4.0, Global Talent Competitiveness Index, Index of Economic Complexity indicates that the Faculty should increase efforts in educating students about them. These indexes and all other indicators are very important for the estimation of a certain market and students should be familiarized with them. Regarding knowledge about professions created by the Fourth Industrial Revolution, mode and mean value prove that students are familiarized with theoretical basics and job descriptions, but there is a lack of practical knowledge. This is another reason to incorporate these technologies into the teaching process. According to respondents, assessments need certain improvements. Before answering the question about the possibility of improvements, the authors would like to mention other limitations of this research. One limitation can be respondent biases. The authors can't be sure if respondents gave honest answers, especially in the element exams. Answers about the type of questions can be exposed to biases, especially because respondents can prefer a certain type. Another limitation is the knowledge of the first year, undergraduate students, about the teaching process after just one semester. The last limitation is the previous knowledge of respondents about elements taken into consideration. This is why their answers might be a result of a lack of knowledge. This is especially emphasized in the element exams, where they can't be sure if a certain type of assessment, task or question is adapted to the Fourth Industrial Revolution. This is one of the reasons why the majority of respondents answered that they neither agree, nor disagree with the statement.

The main contribution of this paper is revealing respondent preferences in the organization of the studies and harmonization of curriculum with the needs and requests of the Fourth Industrial Revolution. By providing answers on different types of questions, Faculty could see in which areas it is necessary to increase harmonization efforts. With the last group of questions about the type of studies respondents prefer, paper provides insights about the possibility of changes the Faculty could take in the future. Recommendation for the Faculty would be the creation of a strategic plan for transformation to Education 4.0. Its main goals would be an adaptation to the new technological age and increasing the competitiveness of the Faculty. One of the key pillars of the Faculty competitiveness should be the higher competitiveness level of its students within the labor market, especially in the area of technology usage. These technologies are already incorporated in the curriculums of other universities and colleges. It is very important to develop and maintain stronger relations with Alumni and private sector not only in organizing classes, but also through projects. Efforts should be increased in all elements of the curriculum in order to increase knowledge about the Fourth Industrial Revolution and its technologies. It will enable transformation towards project-based learning. This type of learning enables better simulation of real situations and better development of skills necessary for the labor market [17]. Project-based learning can solve problems like the lack of theoretical knowledge regarding indicators of the Fourth Industrial Revolution through case studies, seminars and workshops. Knowledge of market research will always be required in companies. Reference [10] states that it is a key mechanism through which service companies understand their current as well as potential customers. It is the functional link between marketing management and an organization's ultimate customer base. As globalization level increases, service companies will need to know how to utilize market research approaches that enable them to stay close to these worldwide and diverse customer segments. It is impossible to expect a transformation process will last for a short period. It demands training of the teaching staff and preparation for new types of teaching, adaptation of learning outcomes, preparation of students and cooperation between faculties, universities and private sector.

REFERENCES

[1] S. Beers, 21st Century Skills: Preparing Students for their Future, Non – Journal, 2011, online: https://cosee.umaine.edu/files/coseeos/21st_century_skills.pdf ,

[2] A. Benešova and J. Tupa, Requirements for Education and Qualification of People in Industry 4.0, Procedia Manufacturing, vol.11, pp. 2195–2202, June 2017

[3] G. Brown-Martin, Education and the Fourth Industrial Revolution ,Groupe Media TFO, August 2017, unpublished, online: https://www.groupemediatfo.org/wp-content/uploads/2017/12/FINAL-Education-and-the-Fourth-Industrial-Revolution-1-1-1.pdf ,

[4] E. De Boer and H. Leurant, The Next Economic Growth Engine Scaling Fourth Industrial Revolution Technologies in Production, White paper, World Economic Forum, January 2018, online: http://www3.weforum.org/docs/WEF_Technology_and_Innovation_The_Next_Economic_Growth_Engine.pdf

[5] E. Dunwill, 4 Changes That Will Shape The Classroom Of The Future: Making Education Fully Technological, Non – journal, March 2016, online: https://elearningindustry.com/4-changes-will-shape-classroom-of-the-future-making-education-fully-technological

[6] Faculty of Economics and Business Rijeka, Internal documentation of the faculty, December 2019

[7] E. Hanushek Will more higher education improve economic growth? Oxford Review of Economic Policy, vol. 32, pp. 538–552, Winter 2016

[8] A. Hussin, Education 4.0 Made Simple: Ideas for Teaching, International Journal of Education & Literacy Studies, vol. 6, pp. 92-98, July 2018

[9] S. Jaschke and U. Heintze, Higher Education 4.0 – Trends and Future Perspectives for Teaching and Learning, March 2014, Non – Journal, online: https://www.researchgate.net/publication/284717997_Higher_Education_40_-_Trends_and_Future_Perspectives_for_Teaching_and_Learning ,

[10] R.G. Javalgi, C.L. Martin and R.B. Young, Marketing research, market orientation and customer relationship management: a framework and implications for service providers, Journal of Services Marketing, vol. 20, pp. 12–23, January 2006

[11] P. Knight and M. Yorke, Employability and Good Learning in Higher Education, Teaching in Higher Education, vol.8, pp. 3-16, 2003

[12] L. Larson and T. Miller, 21st Century Skills: Prepare Students for the Future, Kappa Delta Pi Record, vol. 47, pp. 121–123, July 2012

[13] H. Lassi and F. Petke, Industry 4.0, Business and information system engineering, vol. 6, pp. 239–242, August 2014

[14] L. Liao, M. Du, B. Wang, Y. Yanni, The Impact of Educational Investment on Sustainable Economic Growth in Guangdong, China: A Cointegration and Causality Analysis, MDPI Journal Sustainability, vol. 11, pp. 1-16, February 2019

[15] G. McMeen, The Impact of Technological Change on Education, Educational Technology, vol. 26, pp. 42-45, February 1986

[16] National Science Teachers Association, Quality Science Education and 21st-Century Skills, National Science Teachers Association, 2011, Non - Journal online: http://static.nsta.org/pdfs/PositionStatement_21stCentury.pdf ,

[17] pblworks,org, 2020, online: https://www.pblworks.org/what-is-pbl

[18] B.E. Penprase, The Fourth Industrial Revolution and Higher Education, Book chapter, Higher Education in the Era of the Fourth Industrial Revolution, Chapter 9, Palgrave Macmillan, 1st edition, vol.1, June 2018, pp. 207-229

[19] V. Puncreobuthr, Education 4.0: New Challenge of Learning, St. Theresa Journal of Humanities and Social Sciences, vol.2, pp.92–97, July-December 2016

[20] W. Ratcheva, T. Leopold, 5 things to know about the future of jobs, Non-Journal, September 2018, online: https://www.weforum.org/agenda/2018/09/future-of-jobs-2018-things-to-know/

[21] K. Schwab, The Fourth Industrial Revolution : what it means, how to respond, January 2016, online: https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

[22] A. Selamat, D. Paduka, M. Puteh, R. Alias,, Higher Education 4.0: Current Status and Readiness in Meeting the Fourth Industrial Revolution Challenges, University Teknology Malysia, IEEE Computer Society Malaysia, Non-journal, December 2017, online: https://www.moe.gov.my/muat-turun/teks-ucapan-dan-slide/2017/1527-redesign-he-4-0-higher-education-4-0-current-status-and-readiness-in-meeting-the-fourth-industrial-revolution-challenges/file

[23] P. Sherghold, How can we best prepare students for the future of work?, Non – journal, Western Sidney University, 2019 online: https://usap.sydney.edu.au/wp-content/uploads/2019/05/peter-shergold.pdf

[24] stanford.2025.com, 2020, online: http://www.stanford2025.com/ ,

[25] D. William, How do we prepare students for the world we cannot imagine, ETS, Listening, Reading, Institute of Education, Univeristy of London, Non-Journal Article, December 2011, online: http://www.nnstoy.org/download/international-comparisons/Optimizing+Talent+-+Wiliam.pdf ,

[26] World Economic Forum, Global challenge insight report, The Future of Jobs Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution, January 2016, online: http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf

[27] K. Whiting, Jack Ma wants to go back to teaching, Non-Journal article, September 2018, online: https://www.weforum.org/agenda/2018/09/jack-ma-wants-to-go-back-to-teaching/

A Comparative Study of Gamification in Progamming Education in a Croatian High School

Marinela Schatten
Dugo Selo High School
Ul. S. Ferenčaka 25, 10370, Dugo Selo, Croatia
Email: marinela.schatten@skole.hr

Markus Schatten
Artificial Intelligence Laboratory
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42000 Varaždin, Croatia
Email: markus.schatten@foi.unizg.hr

***Abstract*— During the European Code Week event in 2019. a comparative study on using a gamification platform (namely CodeCombat) for programming education was conducted in 4 distinct 1st and 2nd grade classes (8 groups of students). The groups were from three distinct study programmes: computer technician, economist and car mechanic. With each group two surveys have been conducted one on self-assessment (before using a gamification platform) and one on self-evaluation (after using the gamification platform) in order to answer three simple questions: (1) what is the perception of students on the importance of programming knowledge and skills in everyday life (before and after using the platform), (2) did the students perceive that their knowledge and skills have been improved during the lesson (if yes, how much), and (3) did their motivation to learn programming increase due to using gamification? A total of 79 responses has been collected and the analysis of the results show that there is an increase in motivation for learning programming for the economist study programme, a decrease of motivation for the car mechanic study programme and an indifference in motivation for the computer technician study programme.**

Keywords*— *gamification, codeweek, code combat, survey, motivation, high school, programming, education

I. Introduction

The term gamfication, which has raised major interest from both academia and industry is usually defined as "*the use of game design elements in nongame contexts*" [6], [5]. Gamification in the most broadest sense is often a successful method for fostering motivation among people, supporting user engagement and enhancing positive behavioral patterns on various services like increase of activity, quality of work, socialization and hence productivity [4].

One of the most interesting applicative fields of gamification, often with promising results is of course education as numerous studies have shown [1], [2], [7], [3]. Due to the nature of gamification, especially from a videogaming perspective, computer science educators on all levels of formal education were often the first to adopt gamification processes into their lectures and the field of programming education has been very fruitful as shown in various studies (i.e. [8], [9], [10] and others).



Fig. 1: Screenshot of the CodeCombat gamification platform (student's view)

The European CodeWeek which is "*a grassroots initiative which aims to bring coding and digital literacy to everybody in a fun and engaging way*" [11] which has been held from October 6th to 20th 2019 and had 4.2 million participants and more than 72,000 activities from 80 countries from all over the world. One of these activities has been held in a Croatian high school with the objective to use gamification as a facilitator to educate various levels of high school students in programming. The event was organizied due to a previous success of using gamification (see [12] for details). In particular the gamification platform CodeCombat (available at https://codecombat.com/) shown on figure 1 was used.

From a game genre perspective CodeCombat is a dungeon crawler game in which a player has to navigate through a dungeon by avoiding enemies and traps, collecting items like keys and similar in order to advance through the game. The twist to education is that the player character has to be navigated using programming code as can be seen on figure 2. In the current version the Python and JavaScript programming languages are supported. We have chosen to use Python since it is a common choice for teaching programming in Croatian schools. Each level usually introduces some core programming concepts like basic syntax, function arguments,

strings, understanding of documentation, loops etc.

Due to the fact that CodeWeek has been targeted towards all kinds of students regardless of age, educational background or even learning programme, we have chosen to conduct the activity in four distinct 1st and 2nd grade high school classes in which 8 groups of students have participated. We have conducted two surveys and have analyzed the results achieved by the students during this activity.

In this paper we would like to report on this activity by firstly introducing the methodology in section II. Afterwards we shall present the obtained results and provide a brief discussion on findings in section III. In the end we draw our conclusions in section IV.

II. Methodology

The coding activity was devided into three disting parts: (1) a survey on self-assessment (before using a gamification platform), (2) using the gamification platform to solve as much levels as possible, (3) a survey on self-evaluation (after using the gamification platform). The two surveys were quite similar in order to allow us to compare the results before and after using the gamification platform. Students had to answer three simple questions: (1) what is their perception on the importance of programming knowledge and skills in everyday life, (2) did they perceive that their knowledge and skills have been improved during the lesson (if yes, how much), and (3) did their motivation to learn programming increase due to using gamification? The questions had to be answered on a scale of 1 (not at all) to 5 (a lot). Additionally, in the second survey an open-ended question was added: Name one useful thing you have learned during the lecture. The surveys were anonymous and were conducted using an on-line surveying tool.

The activity has been conducted with 8 groups of students: two groups of first grade car mechanics, two groups of first grade economists, two groups of secon grade economists and two groups of second grade computer technicians. For each group the students were instructed to firstly fill out the first survey. Afterwards they were instructed to open the CodeCombat platform (a special class inside CodeCombat has been created for each group) and try to solve as many levels as possible with and additional motivation that students who score best shall be awarded. The students had a total of 30 minutes to solve levels after which they were instructed to fill out the second survey.

III. Results & Discussion

A total of 79 responses for each survey as well as 79 playing results have been acquired during the activity. In the following we will present a short analysis of results. In table I a comparison of answers to the question regarding motivation to learn programming before using the gamification platform and after using it. The abbreviations are defined as follows: 1CM - first grade car mechanic, 1EC - first grade economist, 2EC - second grade economist, 2CT - second grade computer technician, 1 - first grade student, 2 - second grade student and EC - economist student.

TABLE I: Statistical analysis of motivation to learn programming

		Before		After		
	n	$\overline{x}_B$	σ_B	$\overline{x}_A$	σ_A	$\overline{x}_B - \overline{x}_A$
1CM	21	3.29	1.19	2.90	1.67	-0.38
1EC	19	3.05	1.31	3.11	1.45	0.05
2EC	16	2.88	1.36	3.13	1.26	0.25
2CT	23	4.04	1.07	4.09	1.12	0.04
1	40	3.18	1.24	3.00	1.55	-0.18
2	39	3.56	1.31	3.69	1.26	0.13
EC	35	2.97	1.32	3.11	1.35	0.14

If we compare the results one can see that there is an increase in motivation for learning programming for the economist study programme (the first grade students show an indifference, whilst the second grade students show an increase in motivation), a decrease of motivation for the car mechanic study programme and an indifference in motivation for the computer technician study programme.

These results are intriguing since as it comes out, just by using a (very interesting) gamification platform, perceived motivation to learn programming does not increase automatically, but seems to depend on age and previous knowledge. First grade students seem to have a lower motivation in learning then their older peers, but this assumption has to be tested with comparable groups. Car mechanic student, which have the least previous knowledge in programming and computer science in general lost their motivation after using the gamification platform. This might be the case due to frustration since additional knowledge would be required to actually start understanding the concepts and techniques covered in the platforms tasks.

On the other hand, while first grade economist students have shown indifference in regard to motivation, their second grade peers have shown an increase in motivation after using the platform. This might be since they have a bit more knowledge about programming, having covered some programming fundamentals in the first grade, but still not enough to do something useful. The gamification platform seemed to provide an application domain for their knowledge and motivated them to learn further.

The hypothesis of usefulness might also explain the indifference of computer technician students - the knowledge provided by the gamification platform is something they have already, more or less, internalized in their previous studies, so even if the platform is interesting and funny, it isn't useful enough to motivate them to study further.

In table II the perceived knowledge of programming before and after using the gamification platform is analyzed. As one can see from the table, all groups of students perceive that their knowledge has increased after using the gamification platform.

It is interesting to note, however, that first grade students have a very similar level of perceived knowledge increase regardless of study programme, whilst their second grade peers have quite a different perspective. The second grade economist



Fig. 2: Example task in CodeCombat

TABLE II: Statistical analysis of perceived knowledge in programming

	Before		After		
	$\overline{x}_B$	σ_B	$\overline{x}_A$	σ_A	$\overline{x}_B - \overline{x}_A$
1CM	2.19	1.08	2.81	1.44	0.62
1EC	2.53	1.12	3.16	1.07	0.63
2EC	2.00	1.21	2.88	1.20	0.88
2CT	2.83	0.83	3.13	0.97	0.30

students perceived that they have learned more then the computer technician students. This might be since second grade computer technician students have already learned some of the basic concepts used in CodeCombat's initial lessons, whilst economist students have not. Also, it is worth mentioning that second grade computer technician students perceived they have learned the least of all other groups of students.

TABLE III: Keyword analysis in open-ended question

Keyword	Comments
programming	23
algorithms	11
nothing	9
playing games	9
logic thinking	2

It might be useful to review what students have perceived to have learned during the lesson. In table III a summary of keywords used by students in the open-ended question, dealing with useful things learned is provided. As one can see from the table, most students have answered that they

TABLE IV: Statistical analysis of performance

	Levels		Concepts	
	$\overline{x}_L$	σ_L	$\overline{x}_C$	σ_C
1CM	5.55	2.06	3.09	1.11
1EC	5.58	1.87	3.05	1.13
2EC	7.00	2.14	3.80	0.77
2CT	8.61	2.98	4.26	0.92

have learned some aspects of programming (programming, algorithms, logic thinking), but, on the other hand, there were quite a number of students which answered that they have not learned anything (a total of 9 students). This indicates that gamification does not always provide the necessary motivation for any student. Additionally, also 9 students have answered that they had learned to play games, not perceiving that the game was only a method of delivering programming knowledge. This can be interpreted twofold: (1) from a negativist perspective, students do not perceive a gamified programming environment as serious enough to learn anything useful, or (2) from a positivist perspective, students are not aware that they have learned very useful knowledge since it was as simple as a game for them.

In the following the performance of students during their use of the platform is analyzed. Table IV shows a comparison of student groups in regard to numbers of levels solved and numbers of programming concepts internalized.

As one can see from the table, the computer technician student had performed best, as one would expect. It is interesting to observe that both groups of first grade students,

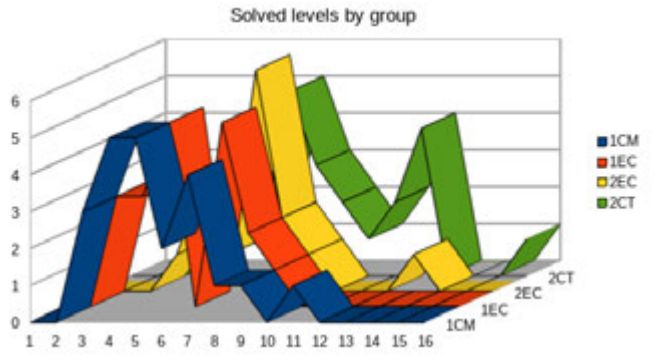


Fig. 3: Solved levels by student group and number of students which have solved the given level

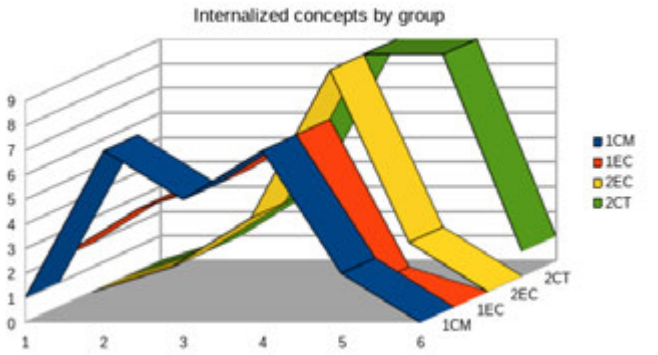


Fig. 4: Internalized concepts by student group and number of students which have internalized the given number of concepts

regardless of study programme performed equally well – there is almost no difference in their results in both number of solved levels and number of covered concepts. Since the study was conducted during the start of the school year it seems that both groups of students had on average a similarly well preexisting knowledge in computer science and programming coming from their primary education. It would be beneficial to repeat the study with first grade computer technician students and compare the results to see if there would be any significant difference.

It is also interesting to note that second grade economist student outperformed their first grade peers significantly, almost as mush as computer technician students outperformed them.

Figure 3 gives a side-by-side comparison of student groups in regard to solved levels. The best of all students (a computer technician) was able to solve 16 levels. The best of all second grade economist students was able to solve 13 levels. A first grade car mechanic student was able to solve 11 levels whilst the best first grade economist student was able to solve 9 levels. One can observe visually that most students in all study programmes have solved between 3 and 7 levels. What differentiates computer technician students from the other groups is that they have a number of above average students which has shifted their distribution to the right.

Figure 4 shows a similar side-by-side comparison, but this time in regard to internalized concepts. As one can see from the figure, only one student (again a computer technician) was able to internalize 6 concepts. The distribution of car mechanic students is shifted to the left since quite a number of students internalized only 2 or 3 concepts. On the contrary the distribution of computer technician students is shifted to the right due to a number of more advanced students. The distributions of both first and second grade economic students are very similar, but the distribution of first grade students is flatter indicating on average worse results.

IV. Conclusion

In this paper we have reported on a comparative study in programming education which was conducted as an activity of the European Code Week 2019. in a Croatian high-school. We have conduced two surveys and a gamification platform activity for four distinct groups of first and second grade students in which 79 students have participated. The surveys were aimed upon studying the impact of gamification on the motivation of students to learn programming.

The results have shown that gamification is not a "*silver bullet*" and just by introducing it into the classroom does not necessarily increase students motivation to learn a complicated topic like programming. As we have discussed, the benefits of gamification come only to light if the lessons are being perceived useful by the students and if the students have adequate preexisting knowledge to understand the topics being covered. Namely, car mechanic students, which had the least background in computer science perceived a loss of motivation to learn programming. Economist students had a slight increase in motivation, especially second grade students. Computer technician students have shown an indifference in motivation before and after using the gamification platform, which might be due to the fact that they are already motivated to learn programming since they chose the study programme but also due to the fact that the lessons did not provide them with much additional knowledge in regard to their existing skills.

An interesting observation was that first grade students had very similar results when using the gamification platform regardless of their particular study programme. Since only car mechanic and economist students in the first grade participated in the activity, a future study should be conducted that shall also include first grade computer technicians to see if there will be any significant difference in results.

Nevertheless, all groups of students perceived that their knowledge in programming has increased, and we believe that by tailoring the lessons and activities to be aligned with the particular study programmes that the results would have been even better.

Our future research will be aimed towards more profound studies of gamification effects over longer periods of time.

ACKNOWLEDGEMENT

The research was conducted within the project "e-Schools: Development of the System of Digitally Mature Schools (Second Phase)" being financed from the European Regional Development Fund, European Social Fund and the state budget.

V. REFERENCES

[1] D. Dicheva, C. Dichev, G. Agre, G. Angelova *et al.*, "Gamification in education: A systematic mapping study." *Educational Technology & Society*, vol. 18, no. 3, pp. 75–88, 2015.

[2] S. de Sousa Borges, V. H. Durelli, H. M. Reis, and S. Isotani, "A systematic mapping on gamification applied to education," in *Proceedings of the 29th annual ACM symposium on applied computing*. ACM, 2014, pp. 216–222.

[3] K. M. Kapp, *The gamification of learning and instruction*. Wiley San Francisco, 2012.

[4] J. Hamari, J. Koivisto, H. Sarsa *et al.*, "Does gamification work?-a literature review of empirical studies on gamification." in *HICSS*, vol. 14, no. 2014, 2014, pp. 3025–3034.

[5] K. Huotari and J. Hamari, "Defining gamification: a service marketing perspective," in *Proceeding of the 16th international academic MindTrek conference*. ACM, 2012, pp. 17–22.

[6] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "Du game design au gamefulness: définir la gamification," *Sciences du jeu*, no. 2, 2014.

[7] W. H.-Y. Huang and D. Soman, "Gamification of education," *Research Report Series: Behavioural Economics in Action, Rotman School of Management, University of Toronto*, 2013.

[8] B. S. Akpolat and W. Slany, "Enhancing software engineering student team engagement in a high-intensity extreme programming course using gamification," in *2014 IEEE 27th Conference on Software Engineering Education and Training (CSEE&T)*. IEEE, 2014, pp. 149–153.

[9] F. L. Khaleel, N. S. Ashaari, T. S. Meriam, T. Wook, and A. Ismail, "The study of gamification application architecture for programming language course," in *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*. ACM, 2015, p. 17.

[10] F. Panagiotis, M. Theodoros, R. Leinfellner, and R. Yasmine, "Climbing up the leaderboard: An empirical study of applying gamification techniques to a computer programming class," *Electronic Journal of e-learning*, vol. 14, no. 2, pp. 94–110, 2016.

[11] "Eu code week." [Online]. Available: https://codeweek.eu/

[12] M. Schatten and M. Schatten, "Gamification of game programming education: A case study in a croatian high school," in *Proceedings of the Central European Conference on Information and Intelligent Systems*, 2019, pp. 13–18.

Computer-Based Education in the Course "Digital Electronics" Teaching the Topic "Adders-Subtractors"

A.N. Borodzhieva*, I.I. Stoev**, I.D. Tsvetkova*, S.L. Zaharieva** and V.A. Mutkov**
*University of Ruse "Angel Kanchev"/Department of Telecommunications, Ruse, Bulgaria
**University of Ruse "Angel Kanchev"/Department of Electronics, Ruse, Bulgaria
{aborodzhieva, istoev, itsvetkova, szaharieva, vmutkov}@uni-ruse.bg

***Abstract* – The paper presents different computer-based tools for teaching the topic "Adders-Subtractors" in four courses for students-bachelors of five specialties in the University of Ruse. These courses introduce the students to the main issues of digital electronics. The aim of these courses is connecting the functional capabilities of the digital components with their microelectronic base and their applications in constructing pulse and digital devices. For the purposes of the courses, students use different software tools for studying the topic. The paper presents a way of presenting the topic of arithmetic circuits to the students making the classes more attractive using active learning.**

Keywords – Computer-based education, Adder-subtractor, Active learning, MS Excel, Logisim, FPGA, Xilinx.

I. INTRODUCTION

During the last years, computers have changed the way people live, study and work. There is no field in which computers do not play a very important role. Education has grown beyond learning from textbooks [1].

Computer-Based Education (CBE) refers to teaching methodologies using computers as a key component of information transmission. For example, computers can be used to present the content of the lesson to the students in more interesting way or used in test-taking to simplify grading and evaluation [2, 3].

The paper presents various computer-based tools for teaching the topic "Adders-Subtractors" in the courses "Synthesis and Analysis of Logic Devices", "Digital Circuits", "Pulse and Digital Circuits" and "Pulse and Digital Devices" introducing the students-bachelors of the specialties "Computer Systems and Technologies", "Electronics", "Internet and Mobile Communications", "Computer Control and Automation", and "Information and Communication Technologies" in the University of Ruse "Angel Kanchev" to the main issues of digital electronics making the connection between the functional capabilities of the digital elements, their microelectronic base and their applications in constructing various digital devices.

For the purpose of these courses, students use software tools for studying the topic "Adders-Subtractors", such as MS Excel [4], Logisim [5], ISE Project Navigator [6], etc.

During the classes the students create an application in MS Excel illustrating the principle of operation of 2-bit, 3-bit and 4-bit adders-subtractors, built on the basis of full adders and additional XOR logic gates. Then the students implement these devices in Logisim, "an educational tool for designing and simulating digital logic circuits" [5] for exploring the operation of the device when changing the input signals. In the end, the students use FPGA design of these devices, built on the basis of full adders and additional XOR logic gates on a FPGA laboratory board.

The paper describes a way of presenting the topic of arithmetic circuits realizing the operations of addition and subtraction of binary numbers to the students making the classes more alluring for students using active learning.

II. ACTIVE AND INTERACTIVE LEARNING METHODS USED IN THE TRAINING PROCESS

The passive teaching method is a form of interaction between the teacher and the students where the teacher is in the center of the lesson and the students are passive listeners. Feedback in such lessons is organized through tests, independent problems, surveys, etc. The passive teaching method is considered as the most ineffective method in terms of material usage, but the advantages of this method include the ability to prepare less labor-intensive and less time-consuming lessons in advance and provide plenty of information in a short time [7].

The active teaching method is a form of teacher-student interaction in which both the teacher and the students participate equally. In this form of learning, the students are no longer passive learners, but active participants in the training process. Many teachers choose this method of teaching due to its advantages. In some cases, this method is efficient when used by experienced teachers with students who have clearly defined learning objectives [7].

This paper is supported by the Project 2020-FNI-FEEA-02 „Research and development of specialized electronic industrial process control systems based on statistical analysis" with financial support of the National Science Fund of University of Ruse and the National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)", financed by the Ministry of Education and Science of Bulgaria.

The interactive method can be considered as a modern version of the active teaching methods, focused on closer student-teacher relationships, where students are more active in the learning process. The difference between active and interactive methods is that, interactive learning involves the interaction of students not only with the teacher but also with each other, i.e. it implies interaction in dialogue mode. In other words, the interactive teaching method is a form of learning and communication activities in which students participate in the learning process and reflect on what they know and what they think of. Unlike the traditional teacher-oriented teaching method, whose primary function is to support and facilitate the students, interactive learning focuses on the needs, abilities, and interests of students. Based on the teacher's knowledge and experience, the students categorize, analyze, synthesize, accept opinions, acquire new skills and develop their attitudes to facts and events [7]. Some interactive teaching activities include self-learning, while others – pairs or groups of students, like shown in Fig. 1 [8, 9].

Individual student activities are:

- Short written (one-minute) exercises, known as exit slips – students need to write for one minute on a specific question, a generalization of "what is the most important thing learned today".
- Correct answer when given a false fact, known as misconception check – students need to identify the correct answer, when given a false fact.
- Green, yellow, red card – each student has three cards (green, yellow and red), representing how students feel about the material studied. For example, the green card means "I get it! I can do this by myself and even explain to others", the yellow one means "I need a little more help" and the red one is "I don't get it, I need a lot of help".
- Ask the winner – students first solve a problem individually. After finding the answer, those who got it right will raise their hands (and keep them raised). Then, all other students will talk to someone with a raised hand to better understand the question and how to solve it next time.

Examples for students' pair activities are:

- Pair-share-repeat – the students are paired up and each pair of students has enough time to reach a conclusion and share it in voice; also, students can find a new partner and share the wisdom of the old partnership to this new partner.

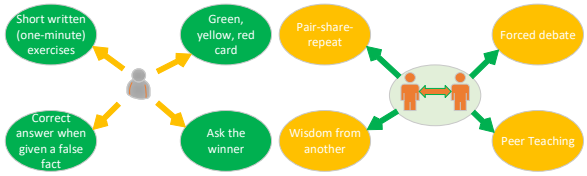


Figure 1. Interactive teaching activities

- Wisdom from another – after a creative activity or an individual brainstorm, students can be paired to share their results. Students are often more likely to share the work of their fellow students publicly than their work.
- Forced debate – in this situation, students debate in pairs and they must defend the opposite opinion of their personal ones; it stimulates them to move away from their own beliefs and teaches them to look through different colored glass from time to time.
- Peer teaching – students need to work as a teacher to teach their peers considering a given question.

The active and interactive teaching methods were introduced in the considered courses for the first time 3 years ago. Previously, the students did not like the courses and they showed poor results during the exam. The disciplines were classified as so-called "barrier". In addition, there was no possibility of exemption from the exam, thus demotivating students to study during the semester, which further complicates the learning process. All this motivates the teaching team to change the approach for teaching, from teacher-centered to student-centered, and for this purpose various computer-based tools were introduced.

Especially for students majoring in "Computer Systems and Technologies" and "Computer Control and Automation", studying the disciplines "Digital Circuits" and "Pulse and Digital Devices", the widespread use of computer-based tools during their studies made the disciplines closer to the topics they studied and their professional orientation, as they showed greater interest in the material and, accordingly, showed better results during the tests and the classes.

III. Synthesis and Analysis of Adders-Subtractors Using Computers in Education

The theory and implementation of arithmetic circuits realizing the operations of addition and subtraction of binary numbers are presented in more details in [10 – 12]. The half-adders, full one-bit adders and two-bit adders are considered in [10, 11] and the half-subtractors, full one-bit subtractors and two-bit subtractors are given [10, 12]. Their implementation in Logisim and FPGA design based on the schematic view as well as active learning methods used in the training process for attracting the students' attention to the topic "Arithmetic circuits for addition and subtraction" are presented in [11, 12]. Using this research [11, 12], MS Excel, Logisim and FPGA implementations of 2-bit, 3-bit and 4-bit adders-subtractors, built on the basis of full adders and additional XOR logic gates are presented in this paper.

During the practical exercises in the courses "Digital Circuits" and "Pulse and Digital Devices", the students are divided into groups of four students and their training is based on project and team work. The process of preliminary preparation for the lesson is illustrated in Fig. 2. Each team should familiarize themselves with the material that will be considered during the exercise in advance (at home). To do this, students use Internet or

reference literature provided by the teacher. Students can communicate with each other to clarify the material, as well as consult with the teacher by e-mail in case of confusion in the material. At the end of their preparation, they must answer the question "What did I learn about this exercise?" within about 2-3 minutes. About the topic "Adders-Subtractors", students should recall from the past exercise the issues about synthesis and analysis of half-adders / half-subtractors, full one-bit adders / subtractors and two-bit adders / subtractors, as well as students should become familiar with the theoretical formulation from the textbook used for their preliminary preparation [10], including the following issues: 1) *controlled inverter* (Fig. 2, block 1); 2) *adder-subtractors* (Fig. 2, block 2); 3) *four-bit adder-subtractor* (Fig. 2, block 3) built with integrated circuits from TTL (Transistor-Transistor Logic) family [13], the library 74xx [14] (integrated circuits of Texas Instruments).

For checking the preliminary preparation of students for the exercise the lecturer uses short written one-minute exercises. During the classes, pair-share-repeat, wisdom from another and debate are used, which is dictated by the fact that students are divided into groups. Also, the teacher gives the students false facts, which they need to correct and this helps the lecturer to check if students understand the topic.

During the class, each student determines the output functions of one of the stages of the 4-bit adder-subtractor, built on the basis of full one-bit adders and XOR logic gates, with an additional SUB input to control the operation to be performed: for addition – $SUB = 0$ and for subtraction – $SUB = 1$, as follows student S1 – for stage 0 (D_0, C_{o0}), student S2 – for stage 1 (D_1, C_{o1}), student S3 – for stage 2 (D_2, C_{o2}), and student S4 – for stage 3 (D_3, C_{o3}), then exchange information with each other and check their authenticity. The determination of the output functions is based on the output functions of the full one-bit adder considered in the previous exercise. This process is illustrated in Fig. 3.

Students learn best by doing, and if one student is ready with his/her task, he/she can teach others, so peer teaching is also used during the exercise. The usage of interactive learning methods helps students to understand better and quicker the essence of the studied material.

The activities of each student during the classes are shown in Fig. 4 ... Fig. 7, representing the implementation schemes, respectively the process of education on the topic of adder-subtractors in the subjects. Within the schemas (Fig. 4 ... Fig. 7), the actions of the individual students as well as the software used are indicated.

During the class, the first student S1 must develop a MS Excel application, as shown in Fig. 4, using the tools for: 1) validating the data (Fig. 4, block 1); 2) inserting formulas (Fig. 4, block 2); 3) conditional formatting (Fig. 4, block 3).

The MS Excel-based application allows the user to select from a drop-down menu the operation to be performed, addition or subtraction, and the bits of the two operands A and B, also using drop-down menus to select one of the two possible values 0 or 1 (Fig. 4, block 1). After selecting an operation, the value of the SUM parameter (column 1 of the table in MS Excel) is automatically changed: SUM = 0 for addition and SUM = 1 for subtraction. The output functions of each stage are implemented by using built-in functions in MS Excel that allow the use of Boolean functions, such as OR, AND, and XOR, and IF function in order to correctly visualize the result 0s or 1s (instead of False and True) (Fig. 4, block 2). Finally, conditional formatting is used in order to display 0s and 1s with different font and background (Fig. 4, block 3) for better presentation of the information. The formulas defined in the cells of MS Excel are based on the formulas determined by the students (Fig. 3). Using software for this type of exercise gives students confidence that if they make a mistake, nothing will be actually broken, so they can use these mistakes to learn better the principles of adder-subtractors.

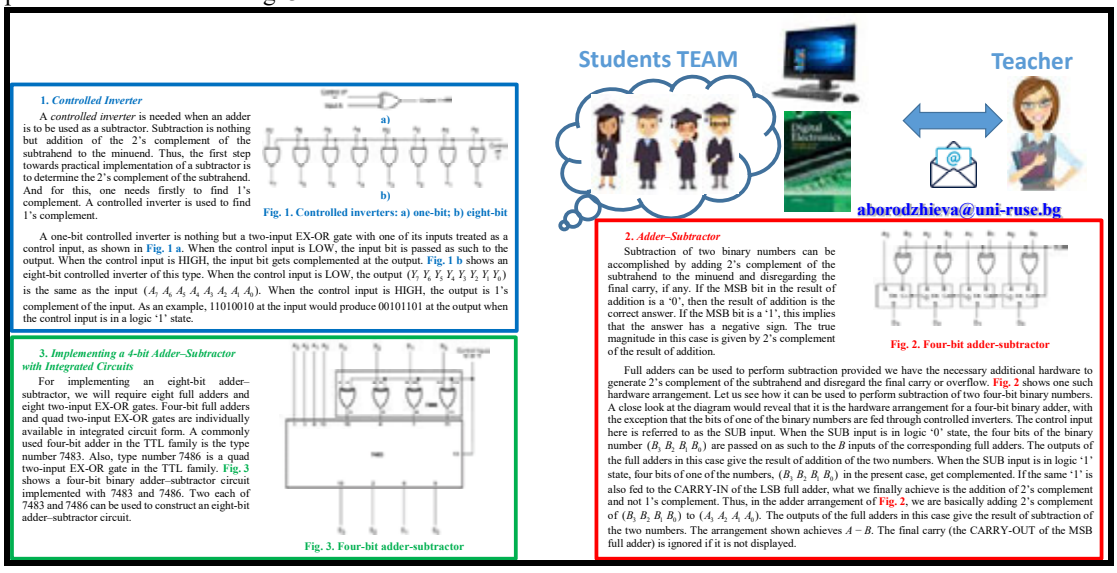


Figure 2. Preliminary preparations of the students for the lesson

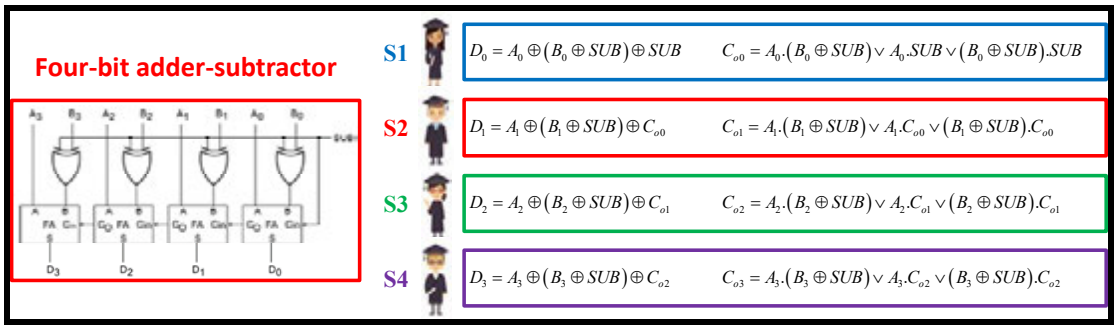


Figure 3. Determining the output functions of a 4-bit adder-subtractor

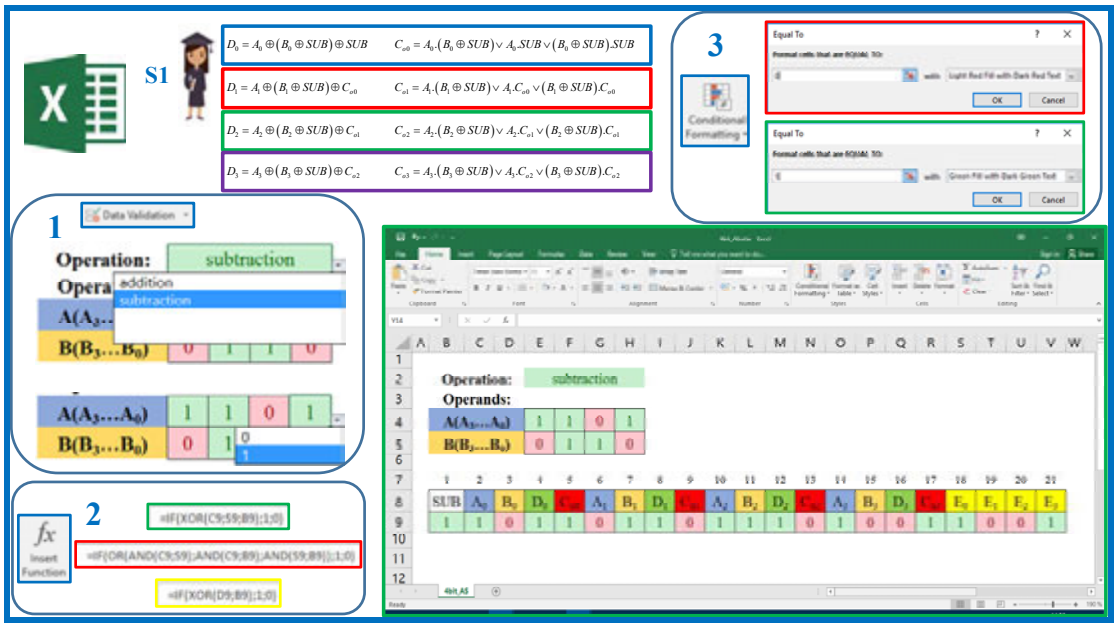


Figure 4. Activities of the student S1 during the classes

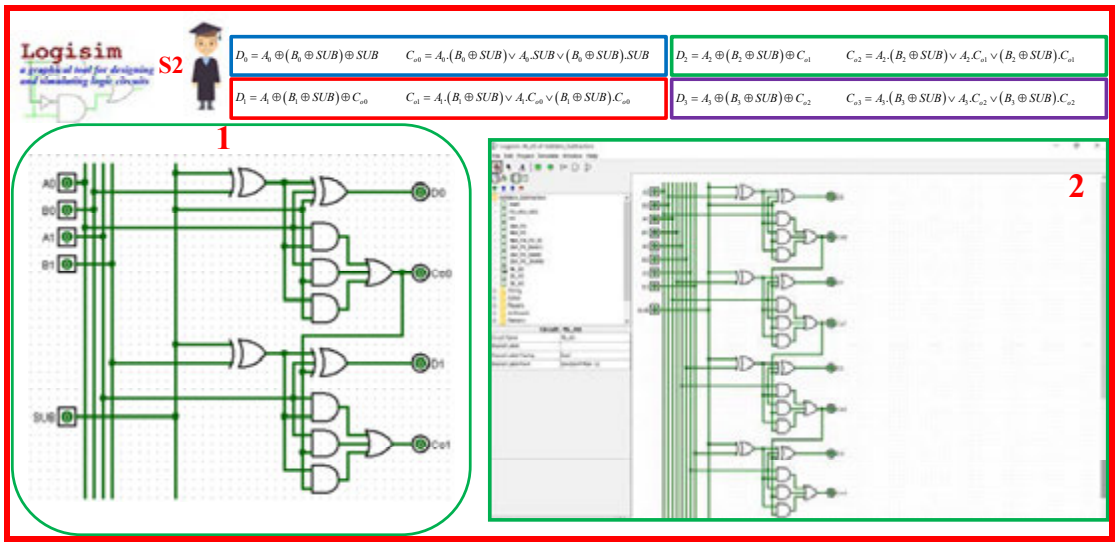


Figure 5. Activities of the student S2 during the classes

During the classes, the second student S2 draws the circuit of the 2-bit adder-subtractor (Fig. 5, block 1) in Logisim based on the output functions of the individual stages previously determined by the students (Fig. 3). For this purpose, the student S2 uses logic gates of type AND, OR and XOR. Based on this 2-bit adder-subtractor,

they can easily build the circuits of 3-bit adder-subtractor and 4-bit adder-subtractor (Fig. 5, block 2).

During the classes, the third student S3 is familiarized in advance with the two used integrated circuits (Fig. 6, block 1), the integrated circuit SN7483 (4-bit binary full adder with fast carry) [15] and SN7486 (Quad 2-Input Exclusive-OR Gate) [16], and draws the circuit of the 4-bit adder-subtractor using TTL integrated circuits of the library 74xx mentioned above (Fig. 6, block 2), using the circuit given in the textbook [10].

During the class, the fourth student S4 examines the circuit implemented in ISE Project Navigator (Fig. 7, block 1) in advance and determines the output functions of the device (Fig. 7, block 2). Only the first two output functions are determined Y_0 and Y_1 (Fig. 7, block 2). Then, the student looks for a match between the inputs (x0…x9) and outputs (Y0…Y9) of the laboratory layout and the inputs $(A_i, B_i, SUB, \text{for } i = 0...3)$ and outputs $(D_i, C_{oi}, \text{for } i = 0...3)$ of the device (Fig. 7, block 3). The following activity of the student S4 is testing the device with the laboratory layout for combinations given by the teacher using the "bit" file, generated in the environment of ISE Project Navigator using the schematic view of the device (Fig. 7, block 1).

After using various software products, which gives students more confidence, they have the opportunity to use real equipment. This will provoke them to acquire practical skills to work with application software and specialized equipment.

A photo of the laboratory board is presented in Fig. 8. It is based on Spartan-6 FPGA Family and it is developed at the University of Ruse "Angel Kanchev" by the authors of the paper. This board allows FGPA implementation of different devices of combinational and sequential type.

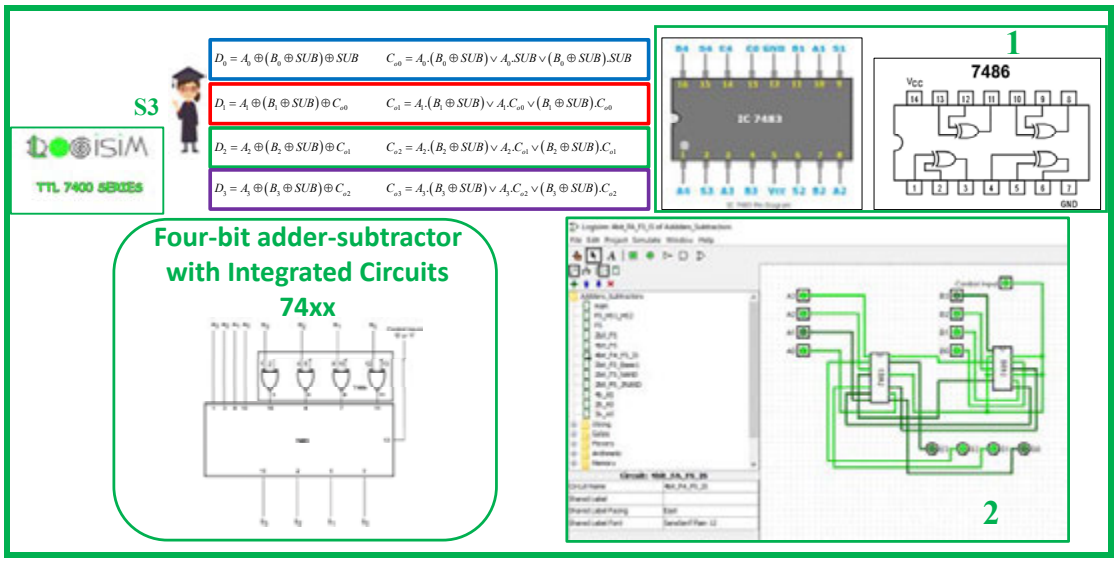


Figure 6. Activities of the student S3 during the classes

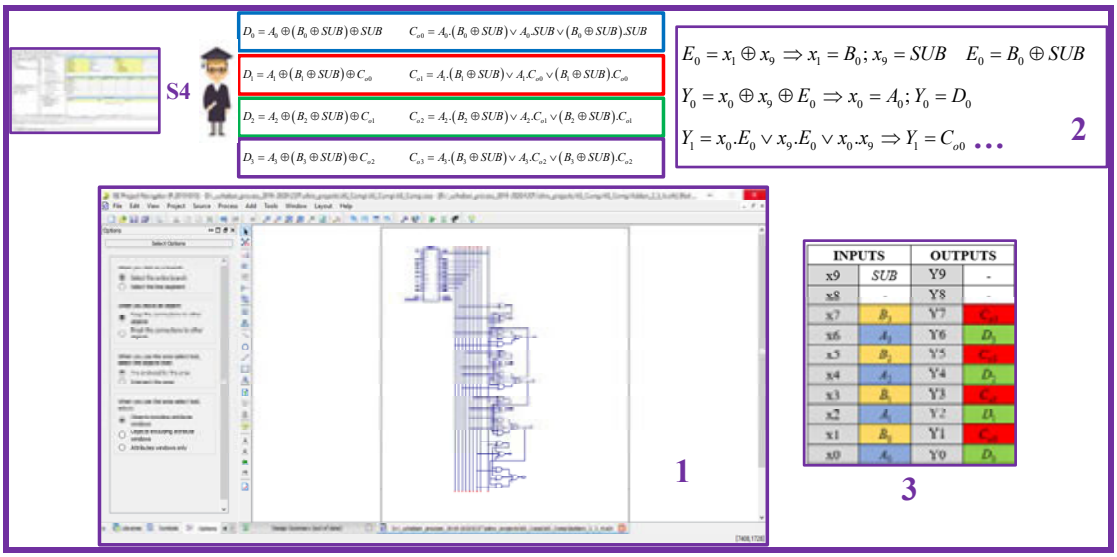


Figure 7. Activities of the student S4 during the classes

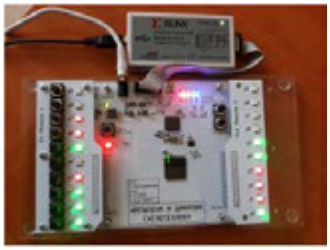


Figure 8. A photo of the laboratory board used in the learning process

The teacher assigns different combinations for the bits of the two four-bit numbers to the students, and the students test and record the results of the two operations (addition and subtraction), comparing the results they received – each with their own computer-based tool.

In this part of the exercise students are encouraged to discussions and debates, also to ask themselves questions and to look for answers among them.

In the end of the exercise the teacher uses the activity ask the winner for summarise the material and to check if all the students understand it.

Next week, each student uses another computer-based tool for investigating digital devices (for example, the student S1 uses Logisim, instead of MS Excel, etc.).

The experiment conducted with our students in the 2019/2020 academic year also proves that students perceive the material more easily by using active teaching methods. About 89 % of the students showed enviable results (Excellent and Very Good) in studying the discussed topic, in addition to their increased interest. The results, unified for all students studying the considered disciplines and the respective topic for adders-subtractors, applying the approach described in the paper, are presented in Fig. 9.

Marks	Excellent	Very Good	Good	Satisfactory	Poor
Number, %	76	13	6	3	2

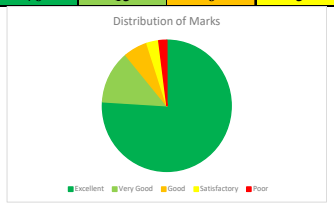


Figure 9. Results unified for all students studying the disciplines

CONLUSION

During the practical classes in the courses of Digital electronics at the University of Ruse, students develop an application in MS Excel illustrating the principle of operation of 2-bit, 3-bit and 4-bit adders-subtractors, built on the basis of full adders and additional XOR logic gates. Then the students implement these devices using Logisim where they can see the operation of the real devices changing the input signals. In the end of the classes, the students use FPGA implementation of these devices, built on the basis of full adders and additional XOR logic gates on a laboratory board, based on Spartan-6 FPGA Family and developed at the University of Ruse.

The practical exercises using different computer-based tools for investigating real digital devices help the students to develop skills important for the future engineers, such as problem-solving, working in groups (teams), applying project-based approaches for learning to use various software products and engineering equipment. They gain skills for synthesizing, analyzing, and evaluating scientific information and problems. Therefore, it is really important for the university lecturers to develop effective pedagogical methods to involve students in the teaching and learning processes based on active learning strategies.

The use of interactive teaching methods provides the participation of the students in the training process, which is the main source of learning. The main difference between the traditional and the interactive activities is that the students do not only revise and enhance their knowledge, but also construct and complement it with new material.

REFERENCES

[1] O. Iskrenovic-Momcilovic, “Using computers in teaching in higher education”, Mediterranean Journal of Social Sciences, vol. 9, no. 4, 2018, pp. 71-78.

[2] A. M. Albarkati, “The application of computer in education system and its significance to teaching and learning”, International Journal of Computer Applications, vol. 134 (9), 2016, pp. 30-33.

[3] R. Raja, P. C. Nagasubramani, “Impact of modern technology in education”, Journal of Applied and Advanced Research, vol. 3, no. S1, 2018, pp. 33-35.

[4] Microsoft Excel, https://products.office.com/en/excel

[5] Logisim, http://www.cburch.com/logisim/

[6] Xilinx documentation, https://www.xilinx.com/content/xilinx/en/downloadNav/design-tools/v2012_4---14_7.htm

[7] M. Giorgdze, M. Dgebuadze, “Interactive teaching methods: challenges and perspectives”, IJAEDU – International E-Journal of Advances in Education, Vol. III, Issue 9, December 2017, pp. 544-548, DOI:10.18768/ijaedu.370419.

[8] S. Senthamarai, “Interactive teaching strategies”, Journal of Applied and Advanced Research, 3 (Suppl. 1), 2018, pp. S36-S38.

[9] W. Long, J. Wang, Y. Li, “Discussing the application of interactive teaching in classroom”, Proceedings of the 2018 International Seminar on Education Research and Social Science (ISERSS 2018), 2018, pp. 323-326, ISBN: 978-94-6252-540-5.

[10] A. K. Maini, Digital Electronics: Principles, Devices and Applications. John Wiley & Sons, Ltd., 2007.

[11] A. Borodzhieva, I. Stoev, V. Mutkov, “Active learning methods applied in the course "Digital electronics" on the topic "Arithmetic circuits using FPGA design"”. 29th Annual Conference of the European Association for Education in Electrical and Information Engineering (EAEEIE) 2019, 4-6 September 2019, University of Ruse, Ruse, Bulgaria, Conference Proceedings, pp. 310-313, ISBN: 978-1-7281-3221-1.

[12] A. Borodzhieva, I. Stoev, V. Mutkov, “FPGA implementation of arithmetic circuits for subtraction used for educational purposes”, 12th International Conference of Education, Research and Innovation, Seville (Spain), 11-13 November 2019, Proceedings of ICERI2019 Conference, pp. 10241-10250, ISBN: 978-84-09-14755-7, ISSN 2340-1095.

[13] Transistor-Transistor Logic (TTL), https://www.elprocus.com/transistor-transistor-logic-ttl/

[14] Related links, Libraries, 7400 series Logisim library from Ben Oztalay, http://www.cburch.com/logisim/links.html

[15] SN7483, 4-bit binary full adder with fast carry, https://www.data-sheetarchive.com/sn%207483-datasheet.html

[16] SN7486, Quad 2-Input Exclusive-OR Gate, http://www4.ujaen.es/~gnofuen/Hoja%20caracteristicas%207486.pdf

Low Cost Rapid Control Prototyping – a useful method in Control Engineering Education

W. Werth, L. Faller, H. Liechtenecker and C. Ungermanns
Systems Engineering
Carinthia University of Applied Sciences, Villach, Austria
w.werth@cuas.at
l.faller@cuas.at
h.Liechtenecker@cuas.at
c.ungermanns@cuas.at

***Abstract*—In this paper, we will report about the adaption of our control education by implementation of Rapid-Control-Prototyping methods in our study program. The first step in this adaption process was a detailed analysis of the main difficulties for students with respect to the implementation of a control task. It turned out that the continuous usage of the proposed software tool chain (MATLAB/Simulink) in class presents a goal-oriented approach when finally implementing a control task also in the lab. We point the students to use MATLAB/Simulink for controller design, simulation and plant identification (data acquisition task) and likewise to also deploy the controller on the specified hardware. Thus, Rapid-Control-Prototyping is a process that allows students to develop, modify and test control strategies on a real-world system.**

As a main result, we re-designed a specific lab exercise related to digital control in the Bachelor program: we introduced a low-cost rapid prototyping system based on Arduino Uno for position control of a DC-motor. The paper also includes the adaption of a course "Embedded Control Systems" in the consecutive Master program following the idea of integrating lectures with exercises.

***Keywords*—Didactic concept, control systems, control labs, engineering education, Rapid Control Prototyping, HIL-simulation, MATLAB/Simulink, automated code generation**

I. Introduction

Undoubtedly, the importance of the field of control are evidenced by the number of international conferences and publications and its increasing impact on industrial applications that touch the lives of everyone [1]. Thus, Control Engineering has become an important and integral part in engineering education in various domains (study programs) and at different levels. On the theoretical side, control systems engineering requires advanced mathematical knowledge and a profound understanding of system theory. On the practical side, a certain 'feeling' for dynamic real-world devices is necessary. This combination makes control systems a challenging topic for many students. To support the learning process we integrated the MATLAB/Simulink® Software tool-chain in all the given control courses. Originally, the intention was to simplify the calculation effort for complex control problems.

In our Bachelor program "Systems Engineering" three control courses (Systems Engineering (SE), Control Systems 1, Control Systems 2) are obligatory to all students (see Fig.1). The course Systems Engineering (SE) Basics covers a practical oriented introduction to MATLAB including simple calculations, making plots and functions. The course Systems Engineering covers the main elements in system modelling (ode) and simulation of physical systems, system descriptions (transfer functions) and system responses. Control Systems 1 and 2 deal with classical control design methods (PID, bode plot, root locus) and controller implementation. In both courses control laboratories try to keep the balance between theory and practice. Furthermore, actual implementation of laboratory experiments offer the opportunity of validating the results.

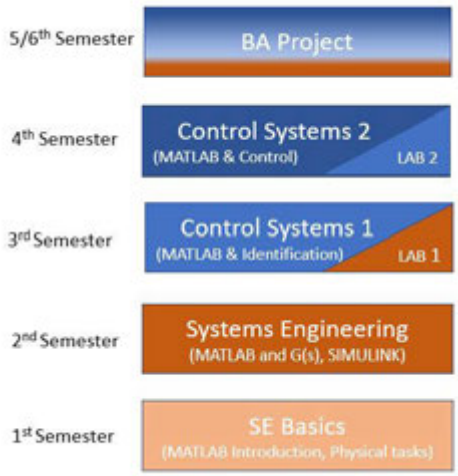


Fig. 1: Current Control Courses in curriculum SE

In a Bachelor Project students will show their ability to solve an overall control task including many aspects like modeling, identification, simulation, analysis, design and implementation.

Traditional teaching methods in control engineering provide therefor the following main steps:

- Explaining the theoretical content (see e.g. [2])
- Elaborating on examples in class or as homework (by hand or MATLAB guided)

- Controller design using MATLAB/Simulink
- Simulation of the feedback system using MATLAB/Simulink as a support tool for complex problems
- Applying the controller to specific hardware (in our Control Lab) and evaluate the performance

We identify three major problem areas for students that keep coming back. First, even though the main idea of a specific topic (e.g. frequency response, transfer function) is understood quite well in theory it seams to be difficult to verify the content when applied to a real world problem. ”Give a suitable interpretation of the poles of a transfer function related to a DC motor ?” is seen as a challenging question. Secondly, MATLAB-exercises are performed in class, but when it comes to redo that in a lab environment students spend a disproportionate amount of time to use already known MATLAB commands again. Finally, the implementation of a real-time controller usually requires good knowledge in hardware design and software programming (e.g. coding).

Software tools like MATLAB/Simulink allow automatic generation of real-time code for implementing a controller or block developed in the simulation diagram since years. This is the key element of Rapid Control Prototyping (RCP) which eliminates exhausting and error-prone hand coding procedures.

II. Rapid Control Prototyping

The term RCP defines a set-up or phase of real-time simulation in which a re-programmable controller is used for fast development of a control algorithm without worrying about writing codes. This process lets engineers quickly test the performance of the controller and validate the system requirements to iterate their control strategies [3]. Rapid Control Prototyping is discussed since a couple of years (see for example [4]). Mathematical models (e.g. the controller) are automatically deployed with e.g. MATLAB/Simulink on a real-time machine with real I/O interfaces to connect to real-world systems (plant). Thus, it decreases development time by allowing corrections to be made early in the product process while they are still inexpensive. RCP methods can be used with different real-time solutions in hardware and software such as Speedgoat, dSPACE, Quanser, National Instruments and others.

Rapid Control Prototyping as a tool for control systems engineering is one of the most powerful technologies for speeding up product development time in industry but also for increasing the learning progress in study programs. Therefore, students' ability to master this method is an indispensable outcome of a practice-oriented education. (see also [5], [6], [7])

The proposed workflow in the overall design process allows an application to be developed according to the so-called ”V-Cycle” (see Fig.2). In a model based design process students start with modelling and simulation of the system (plant) and of course controller design. The feedback system is then tested in a virtual environment usually by desktop-based simulation using Simulink. Then by RCP the plant model is replaced with hardware drivers, which interface with the proposed physical system. The controller-code is automatically generated form the modified Simulink-model and deployed to the controller. Now students can evaluate by a real-time simulation if the controller behaves as expected in the physical environment. Furthermore, this procedure is used to improve the trust in the mathematical model derived. We observed that the students recognise the benefits of block oriented simulation, which is used in a similar way for different purposes.

Following the V-Cyle additional simulation, i.e hardware-in-the-loop (HIL) testing, might be performed when the plant model is running on a real-time target machine (e.g. Speedgoat). This step is currently not performed with the students.

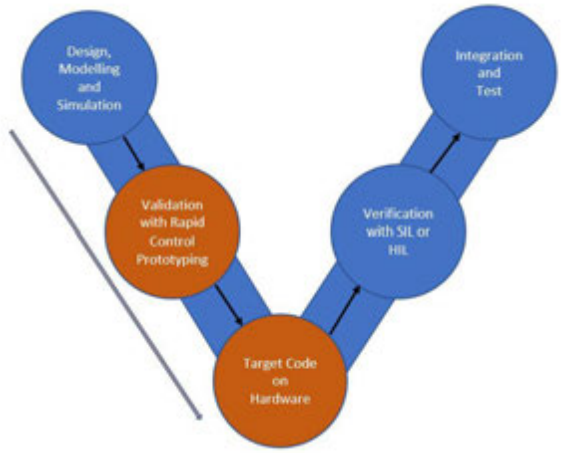


Fig. 2: Model-Based Design with Real-Time Testing Workflow according to V-Cycle

In our control laboratory we currently use Quanser hard- and software as real-time test bench. The main drawbacks of the set-up at the time are:

- The hardware configuration is quite expensive
- One needs additional software beside MATLAB (e.g. QUARC) subject to charge
- Also for real-time siumlation a Desktop-PC is needed
- The set-up is rarely used in industry

Nowadays, various low-cost devices, i.e Arduino based micro-controllers or a Raspberry Pi computer, together with open-source software are suitable solutions for a lot of at least ”small” control applications. We use them for educational purposes related to control problems in our new designed lab exercises. In addition, MATLAB/Simulink provides for many popular hardware devices hardware-support packages. This software empowers the student to communicate and work with hardware without having knowledge in hardware device programming. Also data logging becomes an easy and less error-prone task. Most important feature for our purpose is the support of automatic code generation.

III. Adapted Control Lab Exercises

The lectures in control systems are accompanied by a lot of laboratory exercises. This section represents a classical control

task in details, which has been implemented in our course. This example is one of many, that students perform at CUAS during their control education. One requirement of the setup was to partly use available Hardware Quanser components.

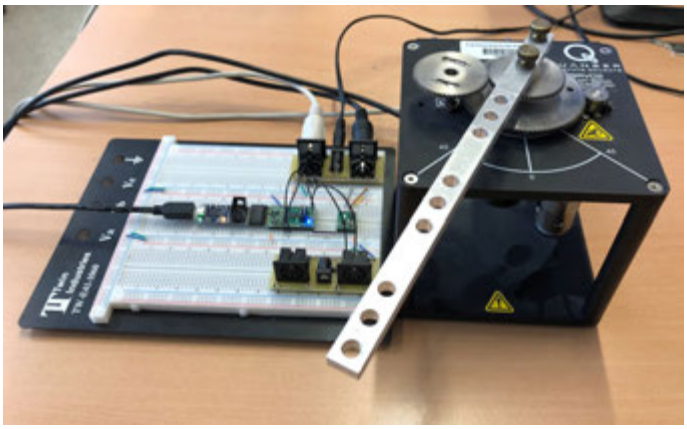


Fig. 3: DC-Motor experiment

A. DC Motor Control LAB Setup

The main task of the experiment represents the speed and position control of a DC motor, which is a central component in applications like electric vehicles. This servo system allows the demonstration of fundamentals of motor control in a variety of ways. The overall experimental setup is illustrated in Fig.3. The system is composed by a typical DC motor equipped with an incremental encoder. A removable bar is mounted on the shaft to simulate different loads and make the angle control visible to the students. An Arduino-Nano

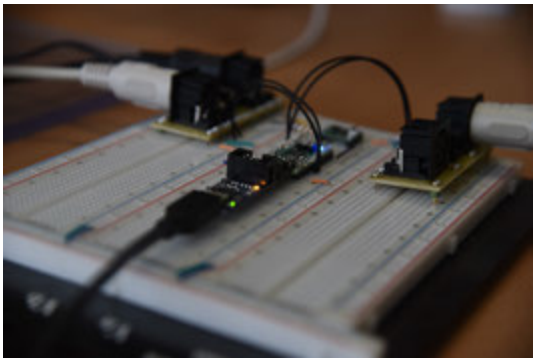


Fig. 4: Breadboard Setup

compatible A-Star 328PB (Pololu) based on an ATmega328PB Micro controller is used as central control unit. Thereby the board needs 5V supply voltage and runs on 16 MHz. Furthermore a MAX14870 Single Brushed DC Motor Driver Board (Pololu) controls the proposed DC motor (see Fig.5). These low-cost hardware components are fixed on a breadboard (see Fig.4). The controller design is done on a standard desktop-PC in MATLAB/Simulink environment. Furthermore an USB-connection guarantees the data communication between PC and controller. This is needed for both measuring specific signals (e.g. step response), logging data and uploading the control law to the controller.

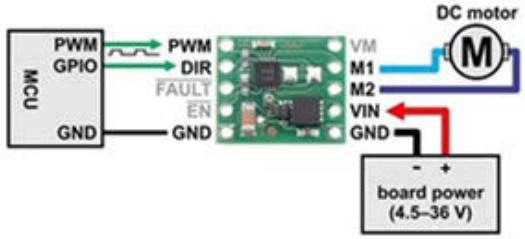


Fig. 5: MAX14870 Single Brushed DC Motor Driver Board [8]

B. Modelling and Control

A suitable simplified description of the plant is well known and given by its transfer function $G(s)$:

$$G(s) = \frac{V}{s(T_1 s + 1)} \tag{1}$$

Basically the plant parameters (V and T_1) can be calculated using the data sheets and analysing of the corresponding differential equations. In the lab exercise these values have to be found by evaluation of suitable measured data. In details students capture the step response of the real plant (see Fig.6) and extract the corresponding characteristics, i.e. slope k_i and delay T_V. The calculation of the step response $y(t)$ of $G(s)$

$$y(t) = -T_1 V + Vt + T_1 V e^{-\frac{t}{T_1}} \tag{2}$$

leads directly to

$$V = k_i \tag{3}$$

$$T_1 = T_V \tag{4}$$

The task of the students is then to calculate different controllers, i.e. P and PD controller (see also [9], [10]). As a first (natural) approach we use simple tuning rules, i.e.

$$K_p = 0.6 \frac{1}{k_i T_V} \tag{5}$$

$$T_D = 0.5 T_V, \;\; T_N = 0.1 T_D \tag{6}$$

The PD controller is designed as

$$C(s) = K_p \frac{(T_D s + 1)}{(T_N s + 1)} \tag{7}$$

The results in Fig.7 show the real world system performance improved by the closed-loop PD controller. During the exercises different control design methods have to be applied to the position control problem. Thereby students take the found transfer function $G(s)$ and guarantee some specifications (e.g. little overshoot (5%) and a suitable settling time $t_s = 200ms$) to calculate a suitable controller $C(s)$.

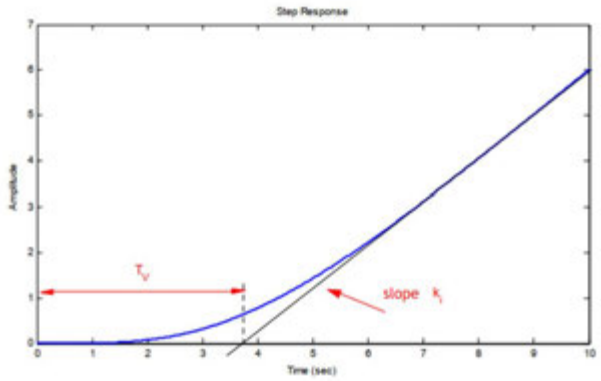


Fig. 6: Step response of a DC motor

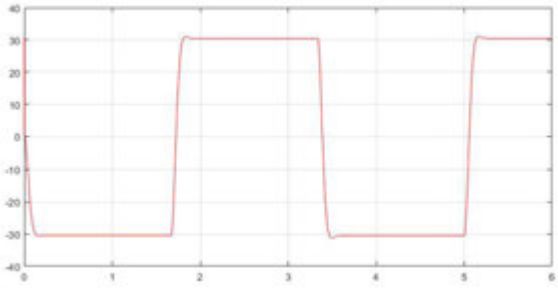


Fig. 7: Position controlled DC motor

C. Automated code generation

Students have to develop sophisticated software to at least test their control system design on embedded hardware. When evaluating the performance of a closed loop control system, an incorrect or incomplete software may add an additional level of complexity and act as an error source. Students are used to evaluate and develop control designs using tools like Simulink, hence implementing a Simulink-Model in C/C++ Code on a resource constrained embedded controller is a major task and a must have skill in a modern control systems development environment. The main advantage of the proposed setup is the availability of comfortable tools (MATLAB/Simulink) for automated code generation. We make use of the hardware support package for Arduino controller.



Fig. 8: Simulink Code for automated code generation

A typical Simulink Code for automated code generation is illustrated in Fig.8. More complex control strategies including practical issues like Anti-Windup are implemented easily. Furthermore, the influence of the used sampling time T_S can be evaluated by the students in details. It turned out that $T_S = 20ms$ is a suitable sampling time for the proposed device (DC motor), whereas $T_S = 100ms$ led to an unsatisfactory performance.

D. Expansions to the concept

The proposed setup is also used for digital control exercises. Thereby the control transfer function $C(s)$ is transformed to a discrete time representation and then implemented as C-code.

Students consider the same configuration also for more sophisticated control problems. They implemented a state space controller for controlling the upright position of an inverted rotary pendulum (see Fig.9). This represents an important lab exercise task in the Master course "Embedded control Systems".



Fig. 9: Inverted pendulum experiment

IV. Validation

The courses discussed were subjected to an evaluation task by the students. Compared to the past years we noticed an improvement in the overall assessment (Mean value decreased form 1.8 to 1.36) illustrated in Fig.10. The results confirm the success of the conversion to the modified laboratory units and MATLAB exercises. *"The laboratory unit was very instructive, vivid and interesting - you could finally see the results of all the calculations!"* was a typical individual feedback of a student. On the other hand, professors found that the use of MATLAB as scientific tool in the subsequent bachelor projects and thesis required much less support than in previous years. The Covid-19 committed effects on laboratory exercises

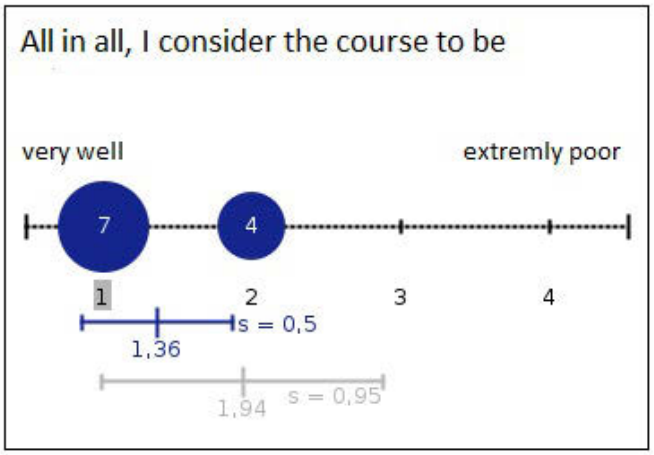


Fig. 10: Students' evaluation result of control systems 1 course, 3rd-semester

(smaller group size and individual work) will be evaluated in autumn. As much as automated code generation simplifies controller implementation for complex systems, direct coding also has its advantages. The manual coding method contributes greatly to the understanding of the internal control processes. Therefore, in the introductory simple exercises in control engineering implementation of controller C-code, for example, is still a fixed part of the course.

V. Conclusion

In this paper, we described possible elements to improve control engineering education in the study program "Systems Engineering". First of all this should help to overcome difficulties for today's students. Secondly, we want to prepare our students so that they are able to meet current and future demands in working life. This was mainly done by the adaption of computer-aided exercises and the implementation of new lab exercises. Feedback from students showed the increasing understanding for control design and controller implementation. The concept of Rapid Control Prototyping by using low-cost components is one of the trendsetting features in modern control education. Additionally, the concept can also be used in students' Bachelor and Master Thesis. The consideration of industrial Rapid Control Prototyping will be investigated in details in future work.

References

[1] N. Kheir, K. Åström, D. Auslander, K. Cheok, G. Franklin, and M. Masten, M.and Rabins, "Control systems engineering education," *Automatica*, vol. 32, p. 147–166, February 1996.

[2] G. Fränklin, J. Powell, and A. Emami-Naeini, *Feedback Control of Dynamic Systems*, 8th ed. Pearson Education Limited, 2019.

[3] The MathWorks, Inc., "Real-time testing with simulink real-time and speedgoat hardware," 2017.

[4] K. Kowalczyk, H. Karkosch, P. Marienfeld, and F. Svaricek, "Rapid control prototyping of active vibration control systems in automotive applications," *Conference on Computer Aided Control Systems Design*, pp. 2678–2682, 2006.

[5] Q.-C. Zhong, C. Matthews, P.-L. Nguyen, and S. Clarke, "Low-cost rapid control prototyping paradigm," *UKACC International Conference on CONTROL 2010*, p. 1269–1273, 2010.

[6] W. P. Milam, "A distributed computing approach to rapid prototyping for embedded controllers," *Proceedings of 36th Midwest Symposium on Circuits and Systems*, p. 1155–1158, 1993.

[7] D. Hercog, M. Curkovic, and K. Jezernik, "Dsp based rapid control prototyping systems for engineering education and research," *Proceedings of the IEEE Conference on Computer Aided Control Systems Design*, p. 2292–2297, 2006.

[8] Pololu Corporation, "Product," https://www.pololu.com/category/239/a-star-328pb-micro, December 2019.

[9] K. Aström, *PID Controllers : Theory, Design and Tuning*. ISA: The Instrumentation, Systems, and Automation Society, 1996.

[10] K. Aström and T. Hägglund, *Advanced PID Control*. ISA: The Instrumentation, Systems, and Automation Society, 2006.

Digital literacy as a part of continuing education library efforts

R. Vrana
Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
rvrana@ffzg.unizg.hr

Abstract - **As the information technology (IT) continues being a part of everyone life, or, at least, of everyone being in position to use any device connected to some type of a communication network including mobile networks, it has become extremely critical for a person to become digitally literate. Younger generations receive education in digital literacy education during their schooling, while adults are frequently lacking such competences once they leave formal education and find employment. Public libraries aim to bridge the digital divide and provide access to critical information and activities to anyone interested in acquisition of knowledge and skills for use of IT and use of digital information sources. The purpose of the research conducted in form of a survey was to confirm an important role of public libraries as places of informal and continuing education in digital literacy in an environment different from schools and universities (places of formal education). The research results in this paper show that public libraries are active facilitators and educational partners in introducing digital literacy to their users. In spite of some obstacles, public libraries remain determined to provide access to critical information sources and activities, which would lead to greater proficiency in digital literacy.**

Keywords – digital literacy; public library; continuig education; Croatia

I. INTRODUCTION

Being digitally literate is a necessary state in the modern society as it changes constantly in its every segment, and these changes require proficiency in different literacies including digital literacy. Those who are not digitally literate will be at a distinct disadvantage as the world is being significantly impacted on by digital technologies [1]. Each new generation of users of digital devices grew up with a different device: games consoles, the Internet and subsequently the smart phone [2]. Each of these technologies has left its mark on life of individuals and development of their abilities to be active workforce and to contribute to the development of society.

The need for digitally literate workforce has become more evident as the development of technology speeds up and as the technology influences development of the society. Rapid changes in the world are putting pressure on already challenged labour markets in many countries [3]. According to the same author, Technology is fundamentally altering the employment relationship and each individual seeking employment should reskill himself or herself. The reskilling revolution aims to provide skills to 1 billion workers within the next five years [3]. Schools and universities around the world are aware of this situation and they change their study programmes according to the needs of the labour market. While this activity might seem logical (and it is) and obvious (still is), it happens over time, and sometimes too slowly (in the eyes of employers). The labour market demands students to have digital skills in order engage better not only with their learning but also contribute to a rapidly changing work environment after their graduation [4]. The employers also expect that their employees will continue their education in the coming years and contribute to the development of the company or organization they work for.

II. DIGITAL LITERACY

Digital literacy can be defined differently according to different sources. According to The American Library Association [5], digital literacy is the ability to use information and communication technologies to find, evaluate, create, and communicate information, requiring both technical and cognitive skills. Another definition offers another view of the matter: Digital literacy can be defined as a social practice that involves reading, writing and multimodal meaning-making through the use of a range of digital technologies. It describes literacy events and practices that involve digital technologies, but which may also involve non-digital practices [2]. A digitally literate person: possesses the variety of skills—cognitive and technical—required to find, understand, evaluate, create, and communicate digital information in a wide variety of formats; is able to use diverse technologies appropriately and effectively to search for and retrieve information, interpret search results, and judge the quality of the information retrieved; understands the relationships among technology, lifelong learning, personal privacy, and appropriate stewardship of information; uses these skills and the appropriate technologies to communicate and collaborate with peers, colleagues, family, and on occasion the general public; uses these skills to participate actively in civic society and contribute to a vibrant, informed, and engaged community [6]. Digital literacy is usually talked or written about in terms of skills sets. Digital marketing institute [7] listed 10 digital skills educators should offer in their portfolio to meet student and industry needs: social media, search engine marketing, analytics, content marketing, email, mobile connectivity, strategic planning, social selling, pay-per-click marketing (PPC), video. Another document What are digital skills? A comprehensive definition for modern

organisations [8] lists another set of digital skills: fluency in inputting information digitally on a range of devices and software packages; manipulating data into less cognitively-taxing and more understandable form; confidence moving between devices and using time equally; reproducing and manipulating different types of digital information; research, analysis and search skills; understanding how information can be effectively and ethically used and manipulated; confident and competent transacting via digital channels while being aware of the limitations; effectively using and managing digital troubleshooting mediums; setting up, modifying and personalising digital devices and software; positive, sceptical and safe use of digital channels, devices and information; an awareness of how human beings consume content in digital mediums; fluency in participating in social channels; being able to manage virtual interpersonal relationships; an understanding of what digital collaboration looks like; an open mind to future digital developments; an awareness of digital trends in your (and related) domains; comfortable with combining experiential and formal learning for new technologies; how to be data-driven (and the limitations of being data-driven); how automation can help and hinder operational excellence; know-how and experience in stopping technology from controlling you. Acquisition and possession of required digital skills qualifies a person for existing jobs and prepares him / her for new and yet non existing jobs in emerging sectors so people with more advanced digital skills can take advantage of an even wider range of opportunities brought about by ongoing advances in digital technologies, platforms, and devices [9]. Digital skills are important because of the changing nature of the work environment, including the sharp growth in the use of freelancers and people participating in the gig economy, as well as broader structural changes profoundly impacting future jobs [10].

III. DIGITAL LITERACY AND LIBRARIES

Libraries in general and public libraries in particular around the world offer different types of activities including those related to acquisition of knowledge and skills in digital literacy. Information sources on the Internet list different skills classified as digital skills in different areas of market. If individuals cannot acquire these skills during their formal education, they start seeking additional education elsewhere including the Internet as a virtual place of cheap education and public libraries as physical spaces equipped and ready to facilitate informal education. Libraries have always had a crucial role in the dissemination and application of knowledge and in providing a place for (informal) life-long learning [11]. Public libraries are place of continuing education i.e. education for adults in a variety of subjects, most of which are practical, not academic [12]. Today, public libraries offer a variety of information sources, printed and digital as well as a plethora of activities aimed at different user groups and meeting their needs for knowledge. Digital literacy is a topic of common interest to all public libraries wishing to facilitate better use of their existing sources in order to contribute actively to the development of society seeking skilful individuals who will fit into the industry 4.0 framework. Knowledge and skills about digital technology are transferred in public libraries in different forms: lectures, workshops, demonstrations etc.

IV. RESEARCH

The purpose of the research was to confirm an important role of public libraries in Croatia as places of informal and continuing education in digital literacy in an environment different from schools and universities (places of formal education). The aim of the research is to collect data about sources and activities offered in public libraries related to education in digital literacy. The research sample was formed from the data available at the portal of public libraries in Croatia at https://www.knjiznica.hr/. Total of 220 public libraries were selected as the sample in this research. The online survey consisting of 15 questions was chosen as a research method because of the geographically dispersed respondents and low cost of carrying out the research. The first part of the survey aimed at collecting data about carrying out digital literacy activities in public libraries in Croatia, including types of activities, their frequency, and financing of such activities. The next part of the survey included topics in digital literacy carried out in public libraries in Croatia. The final part of the survey included questions about ways of carrying out digital literacy activities and abilities of public libraries to offer such activities to their users.

All identified public libraries in Croatia were invited to participate in the research on January 2nd 2020 by sending e-mail invitations to their e-mail addresses. Total of 14 invitation were not delivered due to technical issues on the side of public libraries which gives total of 206 libraries to which invitations were successfully delivered by e-mail. The research was closed on January 17th 2020. The total of 95 public libraries in Croatia participated in the research which is 46,11% of the total number of public libraries to which invitations were successfully delivered.

V. RESULTS AND COMMENTS

In the introductory question (Q1), public libraries were asked to confirm whether they have activities related to digital literacy as part of their annual plans of activities. Of the total 95 libraries, 71 (74,7%) public libraries confirmed that they included digital literacy related activities in their annual plans and 24 of them (25,3%) denied it. The overall result is positive which suggests that public libraries has accepted this type of activity and have added them to the already long lists of activities available to library users in public libraries in Croatia.

Q2. What activities carried out in your public libraries were related to digital literacy? The total of 23 pre-defined answers were offered.

Table 1. Activities carried out in public libraries related to digital literacy (N=95) (multiple answers possible)

	N	%
Searching the Internet	88	92,63
Communication by use of digital technology	75	78,95
E-mail	67	70,53
Data analysis and search	65	68,42

Problem solving by use of digital technology	65	68,42
Use of social media	65	68,42
Understanding of new digital trends	61	64,21
Presentation design	59	62,11
Safe use of electronic devices and information sources	59	62,11
Word processing	58	61,05
Data processing	54	56,84
Digital data inputting	52	54,74
Graphical material processing	48	50,53
Understanding of technology management	47	49,47
Manipulation of different types of data	45	47,37
Ethical use of information	44	46,32
Collaboration in digital environment	44	46,32
Understanding how information are transferred in communication channels	42	44,21
Video editing	40	42,11
Connecting PC to the Internet	37	38,95
Management of interpersonal relations in digital environment	35	36,84
Ability of using multiple electronic devices	32	33,68
Setting of electronic devices and software	26	27,37

Answers to this question were formed mostly from the available literature on digital literacy and activities that are directly related to increasing proficiency in digital literacy. Activities in Table 1. indicate the need of the respondents for digital communication, use of digital technology for problem solving and data analysis and search. More traditional activities like word processing or video editing are not so highly ranked anymore suggesting that the respondents' priorities in knowledge have shifted.

Q3. Types of education activities related to digital literacy offered by public libraries.

Table 2. Types of education activities related to digital literacy offered by public libraries. (N=94) (multiple answers possible)

	N	%
Workshops	74	78,72
Lectures	42	44,68
Demonstrations	17	18,09
We don't offer any educational activity	16	17,02
Tribunes	7	7,45
Panels	3	3,19
Other	3	3,19

The Table 2. shows two distinctive activities – workshops and lectures – carried out in public libraries according to their abilities to carry out such activities, At the same time the results suggest popularity of these two activities as public libraries wouldn't be offering them without users' interest for them. Sixteen libraries do not offer any of suggested activities and there were 3 answers "Other": individual work with library users, training for use of E-citizen system, and help to users in use of PCs.

Q4 Information sources related to digital literacy offered in public libraries.

Table 3. Information sources related to digital literacy offered in public libraries. (N=94) (multiple answers possible)

	N	%
Written printed guidelines / tutorials	47	50,00
Written digital guidelines / tutorials	25	26,60
Multimedia Web tutorial of the Web page of the library	25	26,60
Multimedia Web tutorial of the Web page of another institution or library	18	19,15
Quiz	12	12,77
Video about digital literacy related activities	9	9,57
We don't offer any information sources related to digital literacy	8	8,51
Other	1	1,06

This question aimed at discovering types of information sources related to digital literacy present in public libraries as a means of help when taking part in digital literacy related activities. The results in Table 3. indicate that half of the public libraries in this research use traditional printed guidelines and tutorials while only a quarter of libraries use digital guidelines and tutorials and multimedia Web tutorial of the Web page of the library. Generally, digital information sources and Web are underused and should be better support in acquisition of digital literacy knowledge and skills. There was only one alternative answer in the open part of the question material for workshops, but, clearly, this last answer does not provide any new information about use of information sources as a support to library activities enumerated in previous questions.

Q5 Frequency of receiving queries from users about information sources and activities related to digital literacy offered in library.

Table 4. Frequency of receiving queries from users about information sources and activities related to digital literacy offered in library. (N=94)

Never				Every day	
20	30	32	9	3	N
21,28	31,91	34,04	9,57	3,19	%

Answers to this question present the real demand for information sources and activities related to digital literacy. User queries about information sources and activities related to digital literacy tend to be rather infrequent. Additional promotion of these specific information sources and activities could attract more users to public libraries as places offering education in digital literacy.

Q6 Duration of offering activities related to digital literacy.

Table 5. Duration of offering activities related to digital literacy. (N=94)

Less than 1 year	1-2 years	3-5 years	More than 5 years	We don't offer such activities	
6	19	18	35	16	N
6,38	20,21	19,15	37,23	17,02	%

In spite of infrequent user queries about digital literacy information sources and activities in the previous question, over one third of public libraries in this research offer such sources and activities for longer than 5 years and over half of public libraries (in total) in this research offer such sources and activities for three years and

longer which is encouraging and suggest successful history of education of uses in digital literacy. This is a strong argument when advocating libraries as places for education.

Q7. Topics offered in public library as part of digital literacy related activities.

Table 6. Topics offered in public library as part of digital literacy related activities (N=95) (multiple answers possible)

	N	%
Searching the Internet	81	85,26
Searching the library online catalogue	76	80,00
Use of PCs in the library	76	80,00
E-mail	62	65,26
Word processing	60	63,16
Connecting devices to the Internet	58	61,05
Use of social media	50	52,63
Use of government information sources	41	43,16
Graphical content editing	41	43,16
Database (professional and scientific information) search	40	42,11
Presentation editing	39	41,05
Web page design	8	8,42
Web don't offer such activities	6	6,32
Other	3	3,16
E-book design	1	1,05

High percentages of majority of answers in this question suggest versatility of public library user interests when seeking additional education about use of ICT. Learning how to use an Internet service in addition to more traditional skills like use of PC and word processing is dominating the current topic landscape in public libraries. The topics are easily adapted to the interests of library users and public libraries have always been very cooperative in providing support for self-learning, informal learning, or continuing learning.

Q8. Segments of public library users' lives would be improved if they had higher level of digital literacy.

Table 7. Segments of public library users' lives would be improved if they had higher level of digital literacy. (N=95) (multiple answers possible)

	N	%
Education	82	86,32
Interaction with the government	80	84,21
Preparation for search for a job	78	82,11
Free time / hobbies	76	80,00
Preparation for a workplace	75	78,95
Informing oneself about health	49	51,58
Growth of Internet economy (buying, selling)	33	34,74
None	1	1,05

This question aimed at discovering segments of public library users', which would benefit from knowledge and skills in digital literacy. The results suggest that public libraries perceive a few segments as important in users' lives, which could lead to conclusion about the growing importance of digital literacy in one's life regardless of the segment of one's life.

Q9. Age of library users asking for content related to digital literacy.

Table 8. Age of library users asking for content related to digital literacy.(N=95)

	N	%
Pre-school users	0	0,00
Elementary school users	28	29,47
High school users	14	14,74
University students	10	10,53
Young employees	8	8,42
Middle aged users	15	15,79
Elderly users	19	20,00

According to the results in Table 8., activities in acquisition of knowledge and skills in digital literacy have been sought by many generations of library users. Participation in such activities by younger generation of public library users could suggest lack of such activities in schools and universities, while participation of middle aged and elderly users could suggest insufficient quantity and intensity of such activities during their earlies education. Therefore, public libraries fill in these gaps perfectly by servicing many generations of users regardless of their educational background. What counts is that users are determined to become digitally literate citizens.

Q10. Cooperation of public libraries in carrying out digital literacy-oriented activities.

Table 9. Cooperation of public libraries in carrying out digital literacy-oriented activities. (N=94) (multiple answers possible)

	N	%
With individuals from the Croatian elementary and high schools	42	44,68
With other public libraries in Croatia	20	21,28
With schools in Croatia	20	21,28
With other school libraries in Croatia	19	20,21
No, all activities are carried out by our library alone	19	20,21
We don't have this type of activity in our library	16	17,02
Other	10	10,64
With scientists (in general)	9	9,57
With individuals from the Croatian universities	8	8,51
With the Croatian Ministry of culture	7	7,45
With scientific institutions in Croatia	6	6,38
With other libraries outside Croatia	2	2,13
With other high-education libraries in Croatia	2	2,13
With the Croatian Ministry of science, education and sport	1	1,06

In cases in which public libraries do not have the necessary sources, knowledge, or adequate number of employees, they cooperate with individuals and institutions. The results in Table 9. show that public libraries cooperate most frequently with individuals from the Croatian elementary and high schools. It is rather surprising that public libraries have very infrequent cooperation with scientists and scientific institutions in Croatia. Ten public libraries chose the answer "Other": Microbit:content (irrelevant answer); different associations; associations like Croatianmakers, MakeX, WER; With the institute for development and innovations for youth (2x); volunteers – students; IRIM; only one librarian in the library doing all jobs; volunteers; private

sector. All these answers except “We don't have this type of activity in our library” are positive for public libraries and library users counting on the former to offer them additional education activities they are missing otherwise. More experienced public libraries will be offering education in digital literacy to even more individuals in future and institutions will be in position to cooperate with them in this type of activities.

Q11. Is public library well equipped with information sources necessary for education of users about digital literacy?

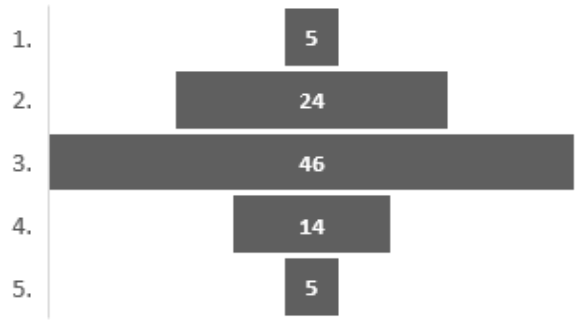


Figure 1. The degree to which public library is well equipped with information sources necessary for education of users in digital literacy. (1=The library doesn't have the necessary information sources at all; 5=The library has all the necessary information sources) (N=94)

Public libraries in Croatia in this research were given an opportunity to estimate the degree to which they are well equipped with information sources necessary for education of users in digital literacy. On the scale from 1 to 5 they indicated that there is a space for improvement and that public libraries could acquire additional information sources (printed or digital) to support their digital literacy activities. One should have in mind that public libraries offer substantial number of versatile activities and that they should be equally prepared and equipped for all of them, which is extremely difficult task for the management of any public library.

Q12. Is public library well equipped with the ICT equipment necessary for education of users about digital literacy?

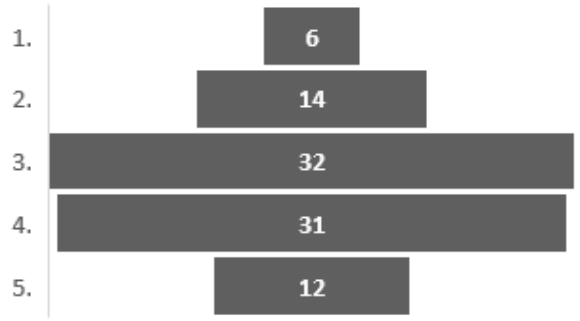


Figure 2. The degree to which public library is well equipped with the ICT equipment necessary for education of users about digital literacy. (1=The library doesn't have the necessary ICT equipment at all; 5=The library has all the necessary ICT equipment) (N=94)

In comparison to answers in the previous questions, the situation with the ICT equipment in public library necessary for carrying out digital literacy activities is less than optimal but it is not bad to the extent that public libraries wouldn’t be able to carry out the planned activities. Answers to the questions 13 and 14 provide more insight in segments of public libraries’ operation, which could be improved.

Q13. How well is public library prepared to carry out digital literacy related activities by itself?

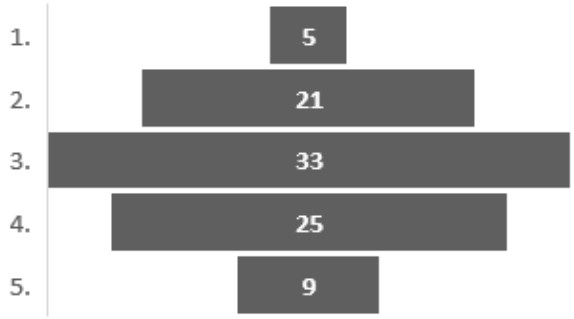


Figure 3. Preparedness of public library to carry out digital literacy related activities by itself. (1=Totally unprepared; Totally prepared) (N=93)

Although it as an estimation, the answers in this scale indicate that situation is not bad it can be improved. This estimation includes all the elements necessary for carrying out digital literacy activities including confidence in abilities of librarians for participation in such activities. The overall situation is good, but it could be even better. Questions 11, 12 and 13 provide us with three very similar estimations that help us understand the degree to which public libraries could be included in digital literacy activities alone or in cooperation with others (Question 10.).

Q14. Improvements in public library in order for the library to be able to offer better information sources and activities related to digital literacy.

Table 10. Improvements in public library in order for the library to be able to offer better information sources and activities related to digital literacy. (N=95) (multiple answers possible)

	N	%
More librarians included in digital literacy related activities	63	66,32
Additional financing of digital literacy related activities	56	58,95
Additional education for librarians	51	53,68
Newer personal computers	42	44,21
Larger library space	39	41,05
More personal computers	37	38,95
New software for personal computers	24	25,26
Faster Internet connections in libraries	16	16,84
Other	2	2,11
Nothing	0	0,00

This question aimed at collecting answers about concrete improvements public library would like to see in future in order to be able to offer better information sources and activities related to digital literacy. Most public libraries

(two thirds) in this research need more personnel or librarians, as they feel understaffed. Public libraries also need better financing, additional education for librarians, newer personal computers (as well as more of them), larger library space, newer software, and faster Internet connections. All these improvements and requests are frequent in public libraries. More librarians, additional financing and larger library space are present in every research of quality of library operation. The positive side of these results is enthusiasm present in public libraries participating in this research, helping librarians to overcome usual inadequacies present in every public library. There were also two answers in category "Other": all the answers and more user needs (irrelevant answer). It is worth noting that not a single public library chose the answer "Nothing".

Q15. Opinion about library as a place of education and carrying out of digital literacy activities.

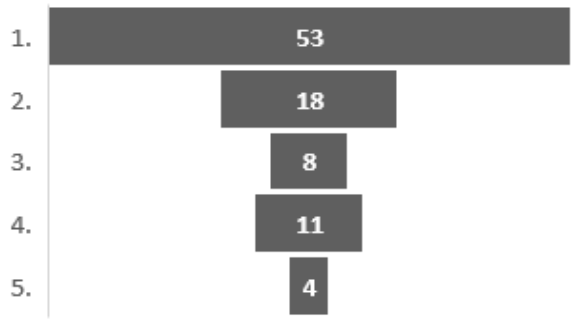


Figure 4. Opinion about library as a place of education and carrying out of digital literacy activities. (1=Totally positive; 5=Totally negative) (N=94)

The final question in the research aimed at finding out opinion of the participating public libraries as places of education and carrying out of digital literacy activities. The answers to this question were grouped on the positive side of the scale indicating that public libraries have positive attitude towards this question, which is a guarantee for the continuation of such activities in future.

VI. CONCLUSION

Digital literacy has become one of the critical literacies for development of individuals and society in general. Customary places for acquisition of knowledge and skills related to digital literacy are schools and universities. However, not all students end their formal education with sufficient knowledge and skills necessary for rapid employment and advancement in career. In the world of emerging industry 4.0, individuals need continuing education to keep abreast of constant changes in society and to remain relevant on the labour market. As formal education system sometimes cannot provide all the necessary knowledge and skills, public libraries step in and offer additional information sources and activities, which will bring future workforce to employment and those seeking new job or just wanting to improve their knowledge and skills closer to their goal. The research in this paper showed that public libraries are active players in the educational arena. They understand what digital literacy means; they have solid resources to carry out digital literacy related activities. However, they face certain obstacles in carrying out their digital literacy related activities like lack of library staff, more financial means, larger space and equipment. In spite of these problems, public libraries remain relevant places for education for many generations of library users and they plan to remain so in the coming years.

REFERENCES

[1] D. Van der Westhuizen, G. Barlow-Jones. "The Role of Digital Literacy in the Academic Performance of First-Year Students in an Information Technology Course", Retrieved December 20, 2019. from https://www.researchgate.net/profile/Duan_Van_Der_Westhuizen/publication/261725496_The_Role_of_Digital_Literacy_in_the_Academic_Performance_of_First-Year_Students_in_an_Information_Technology_Course/links/55951a4808ae21086d1f0d2f/The-Role-of-Digital-Literacy-in-the-Academic-Performance-of-First-Year-Students-in-an-Information-Technology-Course.pdf

[2] J. Sefton-Green, J. Marsh, O. Erstad and R. Flewitt. "Establishing a Research Agenda for the Digital Literacy Practices of Young Children: A White Paper for COST Action IS1410", 2015. Retrieved December 20, 2019. from http://digilitey.eu/wp-content/uploads/2015/09/DigiLitEYWP.pdf

[3] P. Hummelgaard. "The jobs forecast is unsettled. It's time for a reskilling revolution", 2019, Retrieved December 5, 2019. from. https://www.weforum.org/agenda/2020/01/denmark-reskilling-revolution-future-of-work/

[4] F.J.L. Handley. "Developing Digital Skills and Literacies in UK Higher Education: Recent Developments and a Case Study of the Digital Literacies Framework at the University of Brighton, UK". Publicaciones, Vol. 48, pp. 97–109., 2018.

[5] "Framework for Information Literacy for Higher Education", Retrieved December 20, 2019. from http://www.ala.org/acrl/standards/ilframework

[6] "Digital literacy, libraries, and public policy: Report of the office for information technology policy's digital literacy task force", Retrieved December 20, 2019. from https://districtdispatch.org/wp-content/uploads/2013/01/2012_OITP_digilitreport_1_22_13.pdf

[7] "10 Digital Skills That Can Make Students Instantly Employable", Retrieved December 5, 2019. from https://digitalmarketinginstitute.com/en-eu/blog/10-digital-skills-that-can-make-students-instantly-employable

[8] "What are digital skills? A comprehensive definition for modern organisations", Retrieved December 5, 2019. from http://www.skillsoft.com/assets/white-papers/Skillsoft_whitepaper_What-are-digital-skills-a-comprehensive-definition.pdf

[9] "Digital Skills Toolkit", 2018. Retrieved December 20, 2019. from https://www.itu.int/en/ITU-D/Digital-Inclusion/Documents/ITU%20Digital%20Skills%20Toolkit.pdf

[10] "World Economic Forum, The Future of Jobs: Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution", 2016. Retrieved December 5, 2019. from http://reports.weforum.org/future-of-jobs-2016

[11] "IFLA Statement on Digital Literacy", 2017. Retrieved December 5, 2019. from https://www.ifla.org/publications/node/11586

[12] "Continuing education", Retrieved December 5, 2019. from https://www.collinsdictionary.com/dictionary/english/continuing-education

The Model of Collaborative Terminology Contribution to the *Art & Architecture Thesaurus*: Application and Improvement of Crowdsourcing Methods in an Educational Context

Goran Zlodi*, Tomislav Ivanjko*, Petra Štefičar* and Marta Marochini*
* University of Zagreb, Faculty of Humanities and Social Sciences / Department of Information and Communication, Zagreb, Croatia
gzlodi@ffzg.hr

***Abstract* - The paper elaborates the model of collaborative terminology contribution to the Art & Architecture Thesaurus (AAT) through the cooperation of university teachers, students and museum professionals. The elaborated model can accelerate the otherwise time-consuming process of developing multilingual thesauri through greater student engagement while achieving multiple educational goals on real-world tasks. This model includes a statistical selection of the most frequent Croatian art and heritage domain terms and their links to relevant AAT concepts (accompanied with references to relevant literature and lexicographical sources). This process is conducted by students. Subsequent quality control methods include peer-checking and teacher supervision. Mapping of Croatian and English terms for specific concepts was further implemented through semantic technologies (SPARQL queries). Concepts that are particularly demanding are examined by scholars and museum professionals in the corresponding field. After the quality check, the terms would be included in the AAT and made available to open access as Linked Open Data (LOD), provided by the Getty Research Institute. The methods of quality control and student engagement are applied from crowdsourcing projects methodologies, but they were revised and improved concerning the specific needs and requirements of the educational context.**

Keywords* - *thesaurus, linked open data, crowdsourcing, citizen science, education, learning outcomes, Bloom's taxonomy

I. INTRODUCTION

The Getty Vocabulary Program is a program of The J. Paul Getty Trust, a California charitable trust, a cultural and philanthropic institution dedicated to the preservation, research, presentation and interpretation of the world's artistic heritage. That program has developed a structured vocabulary called the Art & Architecture Thesaurus (AAT), containing concepts and terms, hierarchical, equivalent and associative relations, sources, and scope notes on objects, materials, techniques, styles, periods, and other concepts related to art, architecture and the broader field of cultural heritage. The AAT is particularly important in the contemporary information and communication environments (e.g. Semantic Web), as an important resource that enhances access to information about art, architecture, and other material cultural heritage. AAT is available to the public under the Open Data Commons Attribution License (ODC-By) 1.0. as LOD (Linked Open Data) on the SPARQL endpoint at vocab.getty.edu.

In November 2018, the University of Zagreb, Faculty of Humanities and Social Sciences (FHSS) and The J. Paul Getty Trust signed the *Agreement For Vocabulary Contributions*, according to which the FHSS, as the contributing institution would contribute approximately two thousand (2000) terms translated in the Croatian language to the AAT.

II. THE PROJECT OF TRANSLATING OF AAT TERMS TO CROATIAN

Thus, according to the methodology described by the AAT editorial guidelines for contributing large translations [1], the project (hereinafter referred to as Project) of translating terms from source (AAT) to the target language (Croatian) was established. According to the same methodology, this type of mapping represents selective mapping or selective translation as a type of partial mapping, as AAT source terms are chosen as most frequent terms from the collection management databases of partner heritage institutions.

The method of translation and mapping applied in this Project is a direct mapping. This type of mapping refers to the matching of concepts one-to-one in the source and target languages and assumes that there will be the same meaning and level of specificity between the two terms representing the concept in each language [1]. In the process of mapping or translating involved are only descriptors in the source language and descriptors in the target language. The Project uses a hierarchical structure from the AAT.

A. Participants

The importance of the controlled terminology and the need for its alignment with the AAT Thesaurus has long been discussed in the Croatian museum community. The last systematic approach to creating terminology in a museum context was in 1987 through the publication of the

Classification of museum and gallery objects by type of object [2]. As this is a very time-consuming process and which requires the involvement of a significant number of experts, no systematic project has been established so far.

Thus, the initiative to launch the project came from the FHSS, which involved the collaboration of academics, researchers, museum professionals and students.

As stated in the AAT editorial guidelines, "It is critical that translation work is carried out by authorized experts in content, language, and thesaurus construction. Identifying multilingual equivalents is not a simple matter of providing literal translations or looking up words in a dictionary." [1] Table I shows expertise required and how they are distributed to participants.

The expert group includes university professors and museum professionals from partner institutions (for example, the Ethnographic Museum Zagreb, contributed with previously translated terms from past projects).

The project is managed and coordinated by the FHSS Information and Communication Sciences Department. Students from two courses participated in the project: *Museum Documentation I* (Undergraduate program (bachelor's level)) and *Museum Documentation II* (Graduate program (master's level)). Two student groups are involved in the project: Student Group 1 (SG1), which includes all students in courses, and Student Group 2 (SG2), which includes specially motivated students who have agreed to invest more of their free time than they were required by regular study program.

TABLE I. DISTRIBUTION OF REQUIRED EXPERTISE TO THE PARTICIPANTS

Expertise required	**Participants**	
	Experts (Museum professionals, academics, researchers)	***Student Group 2 (SG2) – advanced students)***
Content experts	Museum professionals, academics, researchers.	Students in humanities study groups (e.g., art history, archaeology, ethnology, cultural anthropology).
Language experts	Translators, museum professionals, academics, researchers.	Students of language study groups (e.g. English, Italian).
Thesaurus experts	Academics (information sciences and museology professors).	Students of information sciences (e.g. museology, librarianship, archivists).
Technical experts	Academics (information sciences and museology professors).	Students of information sciences.
Project coordinators	Academics (information sciences and museology professors).	Students of information sciences.

In the approach presented, most of the work can be done by students (mapping, finding references), while final quality control and solving complex conceptual and terminological problems are performed by experts. As the applied approach entails greater student engagement (especially in the first stages of the project), and the term translation projects require expertise, additional methods of quality checking were provided.

III. WORKFLOW OF TRANSLATION AND MAPPING

A. *Stage 1. – Selection of terms*

Firstly, a criterion was established to select terms from the following domains: materials, techniques, styles and periods, object names and roles of agents. This criterion was applied because the terms in these categories are present in the structure of catalogue entries used in printed or online catalogues. Subsequently, the most frequent terms from the above categories were selected from the partner institutions' collection management databases. The selected categories of terms, examples of terms and AAT top facets and hierarchy names are specified in Table II.

TABLE II. SELECTED CATEGORIES OF TERMS AND CORRESPONDING AAT TOP FACETS

Selected categories of terms	***Example***	***AAT top facet and hierarchy name***
Materials	marble (rock)	Materials Facet
Techniques	direct carving	Activities Facet - Processes and Techniques
Object Names (types of objects)	sculpture (visual works)	Objects Facet
Roles of Agents	sculptor	Agents Facet
Styles and Periods	Early Renaissance	Styles and Periods Facet

B. *Stage 2. – the selection of reference sources*

In order for a Croatian term to enter the AAT and join other terms in other languages that refer to the same AAT concept, it is necessary to find references to that term in at least three relevant reference sources in Croatian. This phase of the project consisted of collecting these sources, based on an analysis of the required and optional literature of a series of courses in the departments of art history, archaeology and ethnology and cultural anthropology and further consultations with experts.

C. *Stage 3. – terms translation and mapping*

This stage of translation and mapping was carried out by Student group 1 (SG1).

- Each student is assigned several terms in Croatian for the translation and mapping of which he or she is responsible.
- Each individual Croatian term (e.g. *litografija*) is searched in the following online encyclopaedias and lexicographic sources:
    - The Croatian encyclopaedia [3]
    - The Proleksis encyclopaedia online [4]
    - STRUNA - Database of Croatian Special Field Terminology [5]
    - Croatian Wikipedia

- A reference (record URL) is recorded for each term found.
- STRUNA often includes information about equivalents in other languages (e.g. *lithography*), which can serve as a starting point for finding terms in AAT. Also, using the term in Croatian Wikipedia, one simply follows the link to the record in other languages (students also note a reference to the term in English Wikipedia).
- Then, students search for that term on the online version of the Arts & Architecture Thesaurus. If they find it, they add the following data from the AAT:
    - The Subject_ID for the master AAT concept.
    - The preferred term from the AAT (usually in the plural form, often with a qualifier in brackets).
    - The alternative descriptor (AD) – the term most often found in the singular form in the AAT.
    - The term which would be displayed in catalogue entries – in the AAT it is most often the alternative descriptor (without the qualifier) – as the equivalent to the Croatian term which would be found in reference sources.
- Mapping of Croatian terms for specific concepts was further implemented through semantic technologies by SPARQL queries.
- Notes are recorded in case of problems with the translation of a particular term.

D. Stage 4. – proof-checking and quality control

This stage of translation and mapping was carried out by Student group 2 (SG2).

- Students first proof-checking the terms that SG1 students had worked on. During the proof-checking, the English translation of the term provided by an SG1 student was compared to the translations found in bilingual dictionaries and other reference literature. In case of incorrect translation, the error is recorded and categorized in order to further improve the system of assigning terms to students of the particular study groups. The correct term is recorded and mapped according to the appropriate AAT concept.
- Students conduct selected literature research (while SG1 students consult only online sources, SG2 students consult also extensive printed sources).
- If students find references, they note the following:
    - Reference to do the specific source.
    - The number of the page on which reference was found.
- Definitions of concepts from each source are also recorded.
- Notes are recorded in case of problems with the translation of a particular term.

E. Stage 5. – Review and final verification

This stage of quality check and translation is carried by experts.

- Concepts and terms that are categorized by students as demanding are given for review and translation to experts of different specialities, depending on the domain to which the concepts belong.
- After review and verification terms are designated as suitable for export to the AAT by the required XML format.

IV. EDUSOURCING AND BLOOM'S TAXONOMY

Since the proposed model relies heavily on the inclusion of students in the process one of the main concerns (apart from ensuring the quality of translation and mapping), is ensuring student engagement while at the same time achieving multiple educational goals on real-world tasks. Similar efforts involving the assistance of a large number of volunteers to address various research problems in the field of science have often been brought together under the term *citizen science* and have become a reliable and scalable solution that supports scientific endeavours [6]. A large number of domains are used by them, making the most of the work of volunteers to assist professionals and scientists in processing a large amount of research data that they enhance their work. The combination of human computing and social computing has proven to be very effective, not only in achieving original scientific goals but also in reaching unforeseen discoveries made by community members [7]. This approach has become extremely popular in recent years, and a large number of papers have explored the various applications of systems and initiatives gathered under the term crowdsourcing [8]. Accordingly, an increasing number of implemented projects base their business models on the work of the many, including science and education.

For applying such approaches in the educational context, we propose a term *edusourcing*. Edusourcing is an approach that applies methods and practices derived from crowdsourcing projects to student engagement activities and is tailored for addressing the specific needs and requirements of the educational context.

Since one of the main goals of the educational process is reaching the desired learning outcomes at the highest possible cognitive levels, there is a need to examine the model in that aspect. One of the most widely applied taxonomies of learning outcomes is Bloom's taxonomy that identifies "six levels within the cognitive domain, from the simple recall or recognition of facts as the lowest level, through increasingly more complex and abstract mental levels, to the highest order which is classified as evaluation" [9]. Table III lists the desired outcomes

according to the aforementioned taxonomy form the lowest to the highest level.

TABLE III. COGNITIVE DOMAINS OF EDUSOURCING ACCORDING TO BLOOM'S TAXONOMY

Level	Actions	Outcome
Knowledge	define, list, name, order, recognize, relate, recall, repeat	Gathering term definitions enables students to define, list and recognize basic field terminology
Comprehension	classify, discuss, explain, identify, indicate, report, review, select	Translation of the terms enable students to classify, discuss and select appropriate terminology within the field
Application	apply, choose, demonstrate, sketch, solve, use, write	Linking the terms to the AAT enables students to apply and demonstrate their knowledge in a structured environment
Analysis	analyse, calculate, compare, contrast, discriminate, examine, experiment	Comparing the work process and results between different students enables them to analyse and compare their own knowledge
Synthesis	assemble, construct, create, design, develop, formulate, prepare, propose, write	Incorporating the results within a coherent AAT structure enables students to gain valuable insight into the construction and maintenance of controlled vocabularies
Evaluation	assess, attach, choose, compare, predict, rate, select, evaluate	Performing quality control enables a student to compare and evaluate different approaches through peer-checking

As we can see from the Table III, the proposed model of edusourcing can successfully reconcile the contribution through the quality translation of multiple multilingual thesaurus terms, while at the same time reach desired learning outcomes at the highest possible cognitive levels through greater student engagement.

V. PERCEIVED BENEFITS

In this chapter, we would like to discuss how the participants (students, heritage professionals, scholars) perceive different aspects of vocabulary development and benefits from the project.

A. Benefits for vocabulary

Among other contributions vocabulary we can single out the following:

- Enriching the Art & Architecture Thesaurus with the first contributions of Croatian terms.
- Contribution to world knowledge in the context of Linked Open Data (LOD) and Semantic Web.

B. Benefits for students

Student benefits can be viewed from several complementary perspectives:

- The Project provided information science students with an insight into some of the tasks of heritage professionals via practical work, and an experience of participating and contributing the vocabulary from different perspectives (as a user/as a professional). For example, students of museology have gained various experiences and knowledge such as getting to know AAT in more detail, getting acquainted with problems and limitations in translation, a deeper understanding of the need to standardize museum terminology and documentation, resolving doubts and problems in a group with an interdisciplinary approach etc.
- It enabled interdisciplinary work for double major students who study both information sciences combined with languages or humanities (art history, archaeology, ethnology and cultural anthropology).
- To the students of humanities studies, this project has brought various experiences and knowledge such as an expansion of their own vocabulary of English terms, developing the skill of critically reviewing the literature used for references, experience in interdisciplinary problem-solving, etc.
- Language students have gained experience in translation of professional terminology thus connecting their two different majors.
- For both groups the Project enabled the practical application of theoretical concepts, principles and methods to solve a problem or task in a real-life situation, creating new and integrating existing knowledge into a functional whole.

C. Benefits for university professors and study program

Benefits for the educational process can be categorized into the following:

- Developing a scalable solution for different courses and educational levels.
- Supervising the educational process by controlling the quality of work.
- Possibility of better insight into the applied knowledge of students.
- More active collaboration between students and teachers and increasing student motivation.
- Connecting different study groups and perspectives through the prism of digital humanities.

D. Benefits for Croatian cultural heritage community

Among the other perceived benefits are the following:

- Identification of most frequent terms used for indexing of cultural heritage in Croatian museums.

- Translation and mapping Croatian terminology to AAT.
- Improvement of multilingualism which enables greater visibility and accessibility of Croatian cultural heritage.

VI. CONCLUSION

The proposed model of collaboration between academics, museum professionals, and students can successfully facilitate the contribution of terms to a multilingual thesaurus. By applying edusourcing approach, greater student engagement can accelerate translation and mapping of terms and ensure the quality of the process, while at the same time reach desired learning outcomes at the highest possible cognitive levels.

REFERENCES

[1] P. Harpring, "Contributing large translations : guidelines for multilingual equivalency work". J. Paul Getty Trust, 2015. [Online]. Available: http://www.getty.edu/research/tools/vocabularies/guidelines/aat_4_4_appendix_d_contributors.html#4_4_2. [Accessed: 27-Jan-2020].

[2] "Dokumentacija i klasifikacija muzejskih i galerijskih predmeta", Muzeologija, vol., br. 25, str. 3-117, 1987. [Online]. Available: https://hrcak.srce.hr/95072. [Accessed: 27-Jan-2020].

[3] „Hrvatska enciklopedija". [Online]. Available: http://enciklopedija.hr/. [Accessed: 27-Jan-2020].

[4] „Proleksis enciklopedija". [Online]. Available: https://proleksis.lzmk.hr/. [Accessed: 27-Jan-2020].

[5] „Struna : Database of Croatian Special Field Terminology". [Online]. Available: http://struna.ihjj.hr/. [Accessed: 27-Jan-2020].

[6] Rick Bonney i dr., „Next steps for citizen science", *Science* 343, 6178 (2014): 1436-1437.

[7] Ramine Tinati i dr. „An investigation of player motivations in Eyewire, a gamified citizen science project", *Computers in Human Behavior*, 73, (2017): 527-540.

[8] Daren C. Brabham, *Crowdsourcing*. (Cambridge: The MIT press, 2013)

[9] J. Rutkowski, K. Moscinska, and P. Jantos, "Application of Bloom"s taxonomy for increasing teaching efficiency–case study," unpublished, International Conference on Engineering Education ICEE-2010, July 2010. Available: http://www.ineer.org/Events/ICEE2010/papers/W13A/Paper_1292_1417.pdf . [Accessed: 27-Jan-2020].

ICT for Innovative Education and Science: Smart Environment for Networked Strategies

V. Omelyanenko[*, **], O. Kudrina[*], H. Shevtsova[**], O. Prokopenko[***, ****], V. Petrenko[*****]
[*] Sumy State Pedagogical A.S. Makarenko University, Sumy, Ukraine
[**] Institute of Industrial Economics of NAS of Ukraine, Kyiv, Ukraine
[***] Collegium Mazovia Innovative University, Siedlce, Poland
[****] Teadmus OÜ, Tallinn, Estonia
[*****] Melitopol Bohdan Khmelnitsky State Pedagogical University, Melitopol, Ukraine
omvitaliy@gmail.com

***Abstract* – This research addresses the issues of information and communications technologies (ICT) support of Higher Education Institutions (HEIs) education and scientific activities management in innovation networks by indicating some strategies and approaches for ICT infrastructure management. The main issues and cases of ICT application in HEIs for technology transfer management thought the creation of smart environment of HEIs researches programs for business were considered. Empirically examining of these cases produced a more complete understanding of ICT-based HEI cooperative marketing networked strategies.**

Keywords – innovation networks, smart environment, cooperative marketing, ICT, education.

I. INTRODUCTION

All sectors of the economy are undergoing a digital transformation, which will become a driver for changing existing educational institutions and creating new ones.

All new trends inexorably capture the educational market and change it. Trends in the implementation of digital technologies in educational and research activities are set by commercial organizations (private universities, business schools, corporate universities etc.). But public universities and institutions are starting to think more and more about digital transformation. It is important for them to prepare, compete with the wave of educational startups that emerged with the digital revolution. The formation of an online environment is an area that still requires serious investment from universities. But in the modern conditions the areas of information and communications technologies (ICT) application are expanded within the networked strategies. Modern trends do not even give clear idea of what the university will look like in 20 years, how the campuses or laboratories of universities will look like, how much education and research activities will move into the virtual reality.

In our study we will discuss a wide range of issues of an economic and organizational nature in order to understand what new university models look like now, what are the prospects and future steps needed to reform a modern university and create fundamentally new models of higher education based on ICT. This objective requires considering the limitations and possibilities of new university models as part of network strategies.

II. LITERATURE REVIEW

Modern innovation transformations require focus on universities, which are open to contestation concerning their role in public sphere and the politics of knowledge production.

Where universities are discussed, higher education is focused on, because this is undergoing an intense audit-driven neoliberal re-engineering [1]. It's common knowledge that higher education, science and technology policy discourse is a key point of innovation system development [2]. Research activities have expanded research-intensive academic departments [3]. It relates to existing research on organization in academic leadership in many ways [4].

The paper [5] develops an analytical framework for analyzing the factors of global flows in higher education. Here we can note that attention would be primarily given to issues if Industry 4.0 impacts on higher education, including the problems of rapid learning and innovations development. Therefore one of the most striking features of this problem is re-assembling knowledge production in university [6]. Industry 4.0 revolution causes the transformation of social and economic systems [7] and higher education institutions (HEIs) can be considered as a source of innovations. Research [8] deals with the problem of education space that supports the learning needs of students. In the same in study [9] particular attention is given to distance education. It is clear from these observations that higher education should be based on innovation management and cooperative marketing. A number of key issues arise from the statement.

For instance, in research [10] a diverse aspects of universities influence on the development of the society is considered. E.g. it includes countless smart city and community initiatives. These cases bring together city planning, industry, universities and citizens to improve the urban life of individuals and organizations with the integrated use of versatile information, digital and communication technologies. Within these issues the quality management system in universities, which comprehensively covers education, research and social collaboration activities, and which can ensure the universities' successful partnership in smart city projects, were studied. In [11] an analysis of three challenges to

established understandings of higher education: digitalization, commodification and precarity were proposed. Universities remain the most important organizations involved in developing knowledge and providing means of social mobility.

The study [12] aims to identify and analyze factors that determine the implementation of Information and Digital Technologies of smart manufacturing. Digitalization enables an administrative solutions, data security systems, research data storage and learning resources, as well as opportunities for better collaboration [13].

From these facts, we may conclude that digitalization provides opportunities for the exchange of accumulated experience and knowledge, which allows different stakeholders to learn more and make more informed decisions. ICT support of higher education institutions participation in innovation networks [14; 15], including international networks [16] is important issue. Examples of such ICT support can be found in nanoinformatics [17] application for R&D and industrial analysis.

The given points confirm the idea that strategies and approaches for ICT infrastructure management substantiation is a complex task, proposed by authors in previous studies [18]. Therefore the objective of this study is to consider the issues of ICT support of cooperative marketing in innovation networks by indicating some strategies for ICT infrastructure management.

III. Universities Networked Strategies Smart Environment Levels

Based on experience of leading universities and business schools [19], it is possible to form a conceptual model of smart environment for universities networked strategies, which consists of five levels.

The first level is the most important and is represented by scientific and pedagogical workers, students, industry and academic partners of the university, graduates and applicants. The first level is, in fact, the internal and external stakeholders of the university.

The second level is represented by basic information services. Their task is to create a single information space for digital interaction within the university using flexible tools. Examples of such services are video screens for lectures and seminars, wireless communications throughout the university, cloud storage for storing and exchanging data, professional printing, etc.

The third level includes services that greatly facilitate the lives of students and teachers in a modern university. For foreign teachers and students, they are already an obligatory element of the university; a number of services in most universities are still at the initial stage of implementation.

The fourth level is the most resource-intensive in terms of implementation, but at the same time it allows the university to get the highest added value. It consists of services such as digital marketing, research project management, procurement management, interaction with applicants and students.

The fifth level consists of digital technologies, which are highly likely to be widely used in the university environment (virtual reality, network labs etc.).

Advanced levels of smart environment for its development require appropriate infrastructure, and smart environment.

IV. ICT-based Innovation Education Issues

According to modern innovation policy approach, to ensure the educational services competitiveness, HEIs should implement appropriate quality assurance procedures, develop innovation management policies, standards of practice-oriented educational programs, form and implement a strategy for continually improving the quality of education through R&D interaction with business and society. At the same time, the innovation management in modern HEI should include the digitization the main directions of innovation activity.

Management of formation and practical application of intellectual resources becomes a key task for HEIs in the framework of new role in development. Therefore, ICT tools that are involved in the creation of intellectual resources are largely decisive. Experience of number of leading world universities and expert's reviews confirm that critical thinking and communicative skills are the two core competencies of Industry 4.0.

As a consequence, the role of the communication component of the HEIs interaction with its environment is growing, which forms relevant requests for educational strategies and can facilitate the application of various academic resources for the purposes of socio-economic development. In the context of the foregoing it is expedient to consider the peculiarities of strategic role of academic institutions, in particular in the context of educational and research missions. The analysis shows the practical absence of system processes in research and social missions, which leads to the loss of development potential and impairs the implementation of an educational component that is not related to practice.

For the purposes of ICT tools role analysis we consider the innovation networks. Today, innovative networks are actively involved in modern business processes, offering effective ways to implement complex projects. The growing number of multifunctional networks offers high availability, security and service quality for business processes. But at the same time business and public sector face with a number of challenges related to development of network infrastructure and introduction of ICTs, because digital transformation leads to a number of problems (especially security problems) and opportunities (formation of interaction synergy).

From an innovative point of view, we propose to consider the role of ICT within a systematic approach. In Stanford, the innovation system is defined as inter-organizational, political, economic, technological and non-military systems, through which a favorable environment for business development is formed, supported and developed. From this definition we can define four main components of innovation system:

1. Purposeful process: the innovation system is built

around the process of commercialization of innovations, which has a clear result in the form of innovative products in the market;

2. Subjects interacting with the process of commercialization of innovations;

3. Relationships between them;

4. An environment that provides resources for subjects of commercialization of innovations.

Therefore the strategic development of the HEIs networked-based innovation system is a factor that equally contributes to both increasing the scientific reputation and its financial sustainability, attracting students, new employees and innovative resources. To do this, we need to orient the system to specific commercial outcomes.

Specialized ICT offer the new platforms (applications) for interaction and integration the educational & research processes. Domestic and international R&D cooperation, which becomes effective through modern ICT tools, is crucial for solving modern research problems. In general, collaboration tools and services create new opportunities for future research challenges overcoming.

It can be accurately noted that individual ICT solutions, network infrastructure and scalable solutions will play major role in HEIs R&D and technology transfer. Such ICT can be considered as a key for educational processes reengineering. Thus, the higher education system digital transformation is important to maintain competitiveness in the era of Industry 4.0.

The role of innovation networks is growing significantly as future education reduces the distance between employer and employee, and the educational services market will increase its degree of flexibility. As a result, ICTs in innovation and education networks will participate in the following education and R&D areas:

- large corporations will be able to order certain competencies (projects) without mediation of specialized government or private firms, including project approach (learning in R&D projects);
- organizations that are in particular in need of collective intelligence and team building can enter the educational market themselves and offer specialized courses or services;
- carriers of new interactive educational technologies will be able to combine into the virtual professional communities and to form the community colleges with their own standards and teaching methods. Subsequently, these structures can compete with traditional universities;
- to implement a model of future education, a high-tech infrastructure is needed that will be developed based on up-to-date databases based on Big Data or cloud technologies;
- there are prototypes of stock exchanges for educational opportunities for joint and individual learning, platforms for educational content stores;
- mass attestation tests to determine the overall educational level will gradually lose relevance. They will be replaced with online evaluation and certification systems that allow for an external, verified assessment of the availability of individual competencies and skills.

In these conditions, the educational sector needs to achieve significant progress in digitization and the use of leading innovations in ICTs in new tasks solving (Fig. 1).

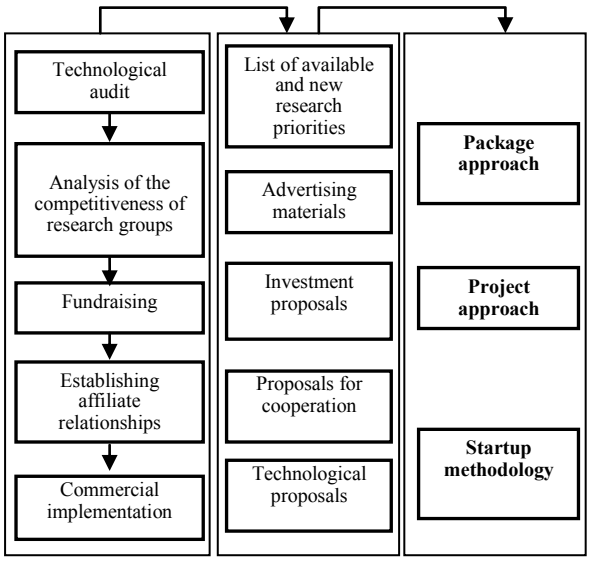


Figure 1. ICT application framework in HEIs innovation system (authors contribution)

Thus, the role of ICT for the development of innovation and educational processes of modern universities can be defined in the following areas:

- creation of educational programs based on unified scientific-educational and innovative process with application of interdisciplinary problem-oriented and project-oriented educational technologies;
- organizational structure and project management methods, which are appropriate to the new task, have been developed;
- diversified sources of HEI funding and effective system of crowdsourcing;
- increase in the number of cooperation projects and practical oriented R&D and educational services;
- developed infrastructure of interaction of HEI with the external environment, which includes agents from the regional, national and international dimensions.

For development of smart environment appropriate for networked strategies purposes we propose to consider the elements of system design methodology [20; 21]. Among other things it includes a comprehensive analysis and evaluation of the customer's task, goals and objectives of complex technical systems, the formation of a source data system for system design, the development of an interconnected system of criteria and indicators for innovation and its components. Rational Unified Process (RUP) model supports the basic principles of system

design and Model-Driven Systems Development (MDSD), which is used both for education and R&D.

Virtual reality technology, e.g. IC.IDO, also has significant potential, which, in combination with engineering systems for product design, is used both by large international corporations and medium and small companies to increase the competitiveness of products on the market. The IC.IDO system is a model of a real product with all its specific characteristics and behavior in real conditions, thereby providing the most accurate teaching aid for conducting trainings and advanced training courses for academic staff in applied areas of innovation management (e.g. [22; 23; 24]. Such methodological foundations can solve the problem of ICT support of universities R&D strategies.

V. CONLUSION

In research it was concluded that universities should perceive the use of new educational technologies as a tool to increase their own competitiveness within the innovation networks. A general conclusion is made concerning that processes of testing and implementing such technologies develops the university's digital educational system but at the same it requires individual strategy that should corresponds with R&D strategy.

As a result of models of ICT support of universities R&D strategies considering the main issues and cases of ICT application in universities for technology transfer management thought the creation of smart environment of HEIs researches programs for business were outlined.

Empirically examining of some cases has produced a more complete understanding of ICT-based university networked strategies. Discussion of the present findings has confirmed the smart environment for networked strategies of universities not only helps students to form unique competencies, but at the same time expands the creative opportunities for teachers and researchers.

ACKNOWLEDGMENT

The publication was publicly funded by Ministry of Education and Science of Ukraine for developing of research project № 0117U003855 «Institutional and technological design of innovation networks for Ukraine national security systemic providing», research project № 0119U100179 «Development of scientific and methodological foundations and practical tools for evaluating of the product innovations commercial (market) prospects».

REFERENCES

[1] J. Cruickshank, and R. Abbinnett, "Neoliberalism, Technocracy and Higher Education: Editors' Introduction," Social Epistemology, vol. 33, iss. 4, pp. 273–279, 2019.

[2] T. Nokkala, "Making the case for policy – persuasiveness in higher education, science and technology policy discourse," European Journal of Higher Education, vol. 6, iss. 2, pp. 144–158, 2016. DOI: 10.1080/21568235.2015.1086664

[3] G. Gibbs, C. Knapper, and S. Piccinin, "Disciplinary and contextually appropriate approaches to leadership of teaching in research-intensive academic departments in higher education," Higher Education Quarterly, vol. 62 (4), pp. 416–436, 2008.

[4] C. E. Tømte, T. Fossland, P. O. Aamodt, and L. Degn, "Digitalisation in higher education: mapping institutional approaches for teaching and learning," Quality in Higher Education, vol. 25, iss. 1, pp. 98–114, 2019.

[5] S. Marginson, and E. Sawir, "Interrogating global flows in higher education," Globalisation, Societies and Education, vol. 3:3, pp. 281–309, 2005. DOI: 10.1080/14767720500166878

[6] C. Muellerleile, and N. Lewis, "Re-assembling knowledge production with(out) the university," Globalisation, Societies and Education, vol. 17, iss. 1, pp. 1–10, 2019.

[7] L. Melnyk, O. Kubatko, I. Dehtyarova, O. Matsenko, and O. Rozhko, "The effect of industrial revolutions on the transformation of social and economic systems," Problems and Perspectives in Management, vol. 17 (4), pp. 381–391, 2019.

[8] B. Taggart, "Modern higher education students within a non-traditional higher education space: not fitting in, often falling out," Research in Post-Compulsory Education, vol. 21:1–2, pp. 86–97, 2016. DOI: 10.1080/13596748.2015.1126420

[9] O. Zawachi-Richter, and S. Naidu, "Mapping research trends from 35 years of publications in distance education," Distance Education, vol. 37 (3), pp. 245–269, 2016.

[10] J. Anttila, and K. Jussila, "Universities and smart cities: the challenges to high quality," Total Quality Management & Business Excellence, vol. 29:9-10, pp. 1058–1073, 2018.

[11] J. Holmwood, and Ch. Marcuello Servós, "Challenges to Public Universities: Digitalisation, Commodification and Precarity," Social Epistemology, vol. 33:4, pp. 309–320, 2019.

[12] M. Ghobakhloo, "Determinants of information and digital technology implementation for smart manufacturing," International Journal of Production Research, 2019.

[13] J. Khalid, B. R. Ram, M. Soliman, M. Khaleel, and M. S. Islam, "Promising digital university: a Ali pivotal need for higher education transformation," International Journal of Management in Education, vol. 12, iss. 3, pp. 264–275, 2018.

[14] T. Tirto, and Y. Ossik, "ICT Support for Industry 4.0 Innovation Networks: Education and Technology Transfer Issues," Lecture Notes in Mechanical Engineering. Springer, Cham, 2020. DOI: 10.1007/978-3-030-22365-6_36

[15] O. Prokopenko, V. Omelyanenko, T. Ponomarenko, and O. Olshanska, "Innovation networks effects simulation models," Periodicals of Engineering and Natural Sciences, vol. 7, no. 2., pp. 752–762, 2019. DOI: 10.21533/pen.v7i2.574

[16] O. Prokopenko, Yu. Eremenko, and V. Omelyanenko, "Role of international factor in innovation ecosystem formation", Economic Annals-XXI, vol. 3-4 (2), pp. 4–7, 2014.

[17] V. Omelyanenko, and D. Volodin, "Nanoinformatics application framework for R&D and industrial analisys," 2017. DOI: 10.1109/NAP.2017.8190183

[18] O. Prokopenko, O. Kudrina, and V. Omelyanenko, "ICT Support of Higher Education Institutions Participation in Innovation Networks," CEUR Workshop Proceedings, vol. 2387, pp. 466-471, 2019.

[19] Sidorov G., "Digital University: the use of digital technology in modern educational institutions," ITWEEK, 2017. https://www.itweek.ru/idea/article/detail.php?ID=192831

[20] New high technology in technology: Encyclopedia. Part 21. Cosmonautics and solving the problems of the development of civilization in the 21st century. Moscow: ENTSITEH, 2002.

[21] M. Cantor, Rational Unified Process for Systems Engineering, IBM Rational Software, 2003.

[22] I.O. Bashynska, "Using SMM by industrial enterprises," Actual Problems of Economics, 12 (186), pp. 360-369, 2016.

[23] O. Kovalova, I. Bashynska, O. Malovichko, and O. Shirobokova, "Risk Management of Innovative Socially Significant Projects (On the Example of Urban Passenger Transport)," International Journal of Advanced Research in Engineering and Technology, 11(4), pp. 294-305, 2020.

[24] I. Bashynska, I. Lytovchenko, and D. Kharenko, "Sales tunnels in messengers as new technologies for effective Internet-marketing in tourism and hospitality," International Journal of Innovative Technology and Exploring Engineering, Vol. 8, Iss. 12, pp. 594-598, 2019.

Student Social Media Usage and Its Relation to Free-recall Memory Tasks

Violeta Vidaček Hainš*, Maja Kućar ** and Ratko Kovačić***
* Faculty of Organization and Informatics, University of Zagreb, Varaždin, Croatia
**Electromechanical School, Varaždin, Croatia
***Electromechanical School, Varaždin, Croatia
E-mails: vvidacek@foi.hr, maja.kucar@skole.hr, ratko.kovacic@skole.hr

Abstract - Some studies reported that higher usage of social media is associated with lower performance in memory or other cognitive ability tests or lower academic achievement. The purpose of this research was to provide descriptive data regarding student social media usage and examine the relationship between the amount of time and frequency a certain platform was checked with the performance of memory tasks. The participants were high-school and university students aged between 16 and 20. They took part in a free-recall working memory task (WMT), after which they completed a Social Media Questionnaire (SMQ). Using regression analysis, it was found that Facebook/Instagram usage could not predict the task result and neither could the frequency of checking. Models tested separately for high-school and university students and found that frequent Instagram checking was a positive predictor of the free-recall task for university students. Additionally, Facebook usage related to Instagram usage and the frequency of Facebook checking related to Instagram checking. The results did not provide evidence to support the idea that social media is damaging to cognitive abilities, quite the opposite in the instance of Instagram checking by university students. A consensus on this topic is yet to be made and there is more research to be done.

Keywords - Social media; memory; cognitive abilities; Instagram; Facebook

I. INTRODUCTION

A Theoretical background

The Internet and social media has become an important part of modern-day communication, especially amongst young people. Research explores relation between **frequency and duration of internet and social media usage and other variables such as cognitive ability, memory capacity and academic performance**. Social media usage and completing school work at the same time requires some multitasking skills. Media **multitasking** includes engaging in more than one media or non-media activity simultaneously [1]. Engaging in Facebook, texting and using instant messaging while trying to complete schoolwork means that students are **multitasking** [2].

Different scientists have provided a variety of result in this research field. Research conducted by [3] found that university students who multitask and use social media for academic purposes have the same academic performance as students who do not multitask with social media. However, students who multitask and use social media for non-academic purposes such as gaming have a lower academic performance. Researchers [2] have found that college students who multitask with ICT by using Facebook (FB) and texting during school work have a lower academic performance when measured by GPA. According to [4], there are some differences between academic performance and interpersonal relationships for students who are heavy and non-heavy internet users. Heavy internet users are more depressed, physically ill, lonely and introvert than non-heavy users. The same authors found that non-heavy internet users have better relationships with academic staff, better academic grades and have more learning satisfaction than heavy internet users. Additionally, some research has found that students who multitask are very often less efficient. But, in instances where students use appropriate learning platforms, their successful learning strategies can enhance their acquisition of knowledge [5]. In general, high-school students consider social media a useful learning tool [6].

Research that focused on the link between **memory capacity and the frequency and duration of social**

media usage showed that social media has differing impacts. Those impacts depend on variety of variables, such as personality traits, motivation, purpose of viewing and other factors. One of the positive effects of social media usage is that young people who used FB for more than a year had better results when tested for verbal ability, working memory and spelling, while the impact of social networks when measured by frequency of checking was not related to cognitive scores [7]. Research conducted by [8] showed differing Big Five personality traits between FB users and non-users. Facebook users aged 19-44 are more extravert and narcissistic in comparison to FB nonusers. Additionally, FB users have lower results on the conscientiousness dimension and are socially lonely.

Additionally, FB viewing strategy is linked to the motivation of FB users. Participants reported that they are more focused on images in instances when they are searching the profiles of their friends. If they are looking for potential employees however, they are more focused on text content [9].

An important research question is whether **social media has any effects on memory capacity ability to provide different answers** and this depends on variety of factors. Some research proved that there is **no relation** between **multitasking and free recall testing** or memory capacity. The free recall test means the participants need to write down all the stimuli from the task, without any cues given. This type of task will be used in this research as a type of working memory task. For example, [10] proved that is no 'google effects' on the memory of social media users, meaning that viewing text messages and Snapchat does not reduce memory capacity. On the other hand, respondents' cognitive self-esteem is positively related to the memory of social media users, which means that respondents with a higher level of self-esteem have better results in memory tasks. Multitasking is not related to working memory capacity, but young adults who multitask have divided attention [11]. Memory capacity measured by the dynamic of free recall is value-directed and depends on the importance of information, as people better encode and remember information that is important to them [12]. On the other hand, there are **negative relations between multitasking and memory capacity**. For example, [13] found that students who multitask, meaning they sent instant messages during school work, needed more time to read texts in comparison to students who did not multitask during reading activities. Furthermore, students who multitasked on Facebook for one or more hours per day and used social media during classes have lower results on the free recall test [14]. Divided attention and encoding reduced free recall memory [15]. Student working memory capacity, storage information, processing information and academic achievement can be increased while students are relaxed [16]. To conclude, and overview of available literature performed by [1] suggested that there is a gap in research results that explained the relation between multitasking and cognitive control.

While the majority of researchers proved that media multitasking is associated with failures of everyday executive functioning and changes in cognitive control, some researchers proved that people who are multitasking have better cognitive control. This area of research needs to be further investigated and explored.

B. Research Goals and Hypothesis

Four hypotheses were made for the purpose of this research:

a) Students who spend more time per day on Facebook and Instagram will remember less words in WMT.

b) Students who check Facebook and Instagram more frequently per day will remember less words in WMT.

.

c) Students who spend more time on Facebook per day also spend more time on Instagram per day.

d) Students who check Facebook more often per day will also check Instagram more often per day.

The first two hypotheses relate to social media usage and Working Memory Task (WMT), while the other two correlate the time spent and the frequency of checks made by students on two popular social media platforms - Facebook and Instagram.

I. METHODOLOGY

A. Participants

There were 117 participants with a mean age 17.33 years (SD=1.6). Most participants were male (M=100, F=12, undefined=5). Participants were students of:

a) The Varaždin Electromechanical School, ESS (86 students, 72.3%), or

b) The Faculty of Organization and Informatics, University of Zagreb FOI (31 students, 26.1%)

All ESS students were in the same program (IT specialist) and all FOI students were in the same program (Applied informatics in business). The ESS and FOI programs were somewhat similar.

B. Data collection

The participants were told that the research was examining technology usage and its connection to memory. First, they were presented words - nouns in the Croatian language (WMT) - after which there was a short math task which was used to "empty" the working memory so it can be examined how many words are retained in long-term memory. Participants had to write down all the words from the list that they remembered. The whole task was repeated. After the memory task, participants completed a technology and social media usage questionnaire.

C. Instruments

Social Media Questionnaire
A questionnaire featuring 24 questions was created for the purpose of this research. Initially students answered questions relating to their gender, age, their previous year's GPA and what number their school/college program was on their priority list of possible programs. Also, students were asked if they owned a personal computer, a smartphone or had internet access. All students answered "yes" on those three questions.
Students answered questions separately for each of the three popular social networks (Facebook, Instagram and Snapchat). They were asked if they have a "XY" profile, how much time do they spend on it (in minutes) and how often they checked "XY" (in minutes). Also, they were asked if there was another social media platform where they spent a significant amount of time. Other questions related to checking social media accounts while studying or in class, reading habits and taking notes, but they do not form the topic of this paper.

Working Memory Task (WMT)
The words presented were all nouns of the Croatian language. There were 35 words in both tasks and one word was presented for three seconds. None of the words from first list were repeated in the second list. Before writing down words they remembered, students completed short math problems (addition). Following this, participants free-recalled the nouns presented to them.

II. RESULTS AND INTERPRETATION

A. Social Media Usage - Descriptive Analysis

Facebook
As expected, almost all participants had a Facebook profile (98.3%, N=115). On average, they had had a Facebook profile for 6.55 years (SD=2.87). Most students claimed they spent less than 1 hour on Facebook per day (41.2%) and 36% claimed they spent 1-2 hours. On average, students checked their Facebook/Messenger every 64.6 minutes (SD=87.61).

Instagram
The majority of participants had an Instagram profile (81.2%, N=95). Most of those students (49.5%) spent 1-2 hours on Instagram per day and 26.3% spent less than 1 hour. Interestingly, 16.8% of students claimed they spent more than 3 hours on Instagram per day. On average, students checked Instagram every 92.06 minutes (SD=206.24).

Snapchat
Snapchat was a somewhat less popular than Facebook and Instagram. 38.5% of students (N=45) had a Snapchat profile, 61.5% of students (N=72) did not have a Snapchat profile, which made Snapchat the least-used media platform in comparison to Facebook and Instagram. Most students who had a Snapchat profile (62.2%) spent less than 1 hour per day using the app and 24.4% spent 1-2 hours. On average, students check their Snapchat account every 309.74 minutes (SD=625.23).

TABLE 1. PERCENTAGE OF STUDENTS HAVING A CERTAIN SOCIAL MEDIA ACCOUNT, TIME SPENT, AVERAGE CHECKING TIME (M) AND STANDARD DEVIATION (SD) OF M CHECKING TIME

	Facebook	Instagram	Snapchat
% of Students Using It	98.3%	81.2%	38.5%
< 1 Hour	41.2%	26.3%	62.2%
1-2 Hours	36%	49.5%	24.4%
3-4 Hours	12.3%	16.8%	11.1%

5-6 Hours	5.3%	7.4%	-
>7 Hours	5.3%	-	-
M Checking Time (Min.)*	64.4	92.06	309.74
SD Checking Time (Min.)**	87.61	206.24	625.23

The details of descriptive analysis are shown in Table 1; the percentage of students having a certain social media account, time spent, average checking time (M)* and standard deviation (SD)** of M checking time.

Working memory task (WMT)

Two WMT were conducted during the experiment, each WMT consisted of 35 words. In WMT1 the average number (M) of remembered words is 13.35 (SD=3.86) and in WMT2 the M is 14.60 (SD=4.37). The two WMT results were added, so the M used in regression analysis is 27.44 (SD=7.41).

Overall it is clear from the descriptive analysis that Facebook and Instagram were by far the most popular social media platforms. More participants had a Facebook account than an Instagram account, however those who had an Instagram profile spend more time on Instagram per day than on Facebook. Half of all Instagram users (49.5%) claimed they spend 1-2 hours per day on Instagram. The most frequently checked social media was Facebook, and the participants claimed to check Facebook every hour (M=64.4) which was expected because Facebook also contained the Messenger App which was used primarily for texting, while Instagram was not used primarily for texting. The average frequency of Snapchat checking was 309.74 minutes (around every 5 hours), but the variability was too big to form conclusive results with this information and compare it to Facebook or Instagram.

B. Regression and Correlation Analysis

For hypotheses a) and b) linear regression analyses were made separately for Facebook and Instagram.

Using linear regression, it was tested whether the amount of time spent on Facebook or Instagram per day could predict the success in the free-recall memory task (WMT). The regression models were tested separately for Facebook and Instagram. The regression model regarding amount of Facebook usage could not predict the result ($F_{(1,112)}$=.41, p>.05), but neither could the model regarding Instagram usage ($F_{(1,93)}$=3.69, p>.05). F value in regression analysis is the mean square regression divided by the mean square residual**,** determines the ratio between explained and unexplained variance and p is the statistical significance of the result or probability of error.

Two regression models regarding the frequency of Facebook and Instagram checking per day could also not predict the task result. The regression models were tested separately for Facebook and Instagram. The regression model regarding frequency of Facebook checking could not predict the result ($F_{(1,109)}$=.57, p>.05), but neither could the model regarding frequency of Instagram checking ($F_{(1,92)}$=.04).

When the regression models were tested separately for high-school and university students, it was found that frequency of Instagram checking was a significant predictor of WMT for university students ($F_{(1,25)}$=4.98, p<.05) - those who checked Instagram more frequently had a better result on the WMT (β=-.41). The given result could be considered somewhat surprising and certainly opposite to the stated hypothesis.

A correlational analysis was made to test hypotheses c) and d). Students' average usage of Facebook per day was positively correlated with students' average Instagram usage per day (r=.46, p<.01). Students' average frequency of Facebook checking per day was positively correlated with students' average Instagram checking per day (r=.26, p<.01)

C. Discussion

With the rapid rise of social media over the past 10 years, research on these topics also had to rise. It needs to be taken into consideration that, with time, social media and technology did not become more popular or widespread, but rather they became more complex and sophisticated, with new features appearing almost daily. That aspect of rapid change made the job more difficult for researchers - keeping up with social media and technology – and has become a serious challenge for experts in various fields. Various hypotheses have been posed regarding the effects of social media and technology on cognition, memory and various other aspects of behavior. The research conclusions that have been

made are not necessarily congruent and there is no clear consensus regarding effects of social media and technology on our cognition.

However, the results of previous research are more in favor of the negative effects of technology than not. The research team [14] examined the effects of Facebook usage on memory and found that high-level Facebook users had worse performance in the memory task than low-level Facebook users. The authors divided their participants into two groups - high and low Facebook users. That was not the case in this research so that the somewhat arbitrary divide of participants into two groups could be avoided. Consequently, the chosen statistical test was regression analysis and not a t-test because the participants were not divided into two arbitrary groups which would be suited for t-test. The regression analysis was used to possibly predict student performance in memory task based on their reported social media usage and their frequency of checking. They [14] published their paper in 2013 when Facebook was by far the most popular network and Instagram was not as popular. Today, the situation is significantly different and it was necessary to include Instagram and other social media platforms. Another novelty which has appeared over the past 6-7 years is the fact that all apps are now available on smartphones and are used primarily on smartphones. That means that the question is not how much time you spend on *XYZ* per day, but how often do you check *XYZ*? In instances of certain people and social media, it is a matter of minutes. For that reason, both aspects were covered in this research, namely the amount of time spent and frequency a certain social media was checked. The results did not confirm the hypothesis which was, in general, that social media is somehow damaging to our cognitive abilities. It cannot be said that results proved the opposite because, on the basis of the results in this research, social media has no effects on our memory abilities, at least not in the aspect of working memory capacity measured in WMT.

However, the limitations of this study should be mentioned. The first limitation was the mere fact that it is questionable how accurate the participants were when estimating their hours of usage and frequency of checking certain social media platforms. Some people were certainly more precise and accurate than others and the effects of possible social desirability should be mentioned as well. Some participants probably reduced the actual amount of time they claimed to have spent on social media in order to appear more conscientious, responsible, self-controlled or less "addicted" to the Internet and social media. Furthermore, it should be noted that although it may be true that social media usage does not affect cognitive processes included in this particular memory task, it is possible that the effects are more complex and are affecting parts of our cognition in ways we have yet to hypothesize. Also, it is worth mentioning that it is possible that any disturbing activity (e.g. frequent TV watching) would negatively effect the working memory capacity, but this paper focuses on social media.

The analysis in this research was primarily based on Facebook and Instagram and it should be noted that social media or smartphone usage in general has become a continuing process throughout our daily routine, with various apps overlapping and "fighting" for parts of our attention. It is suggested that researchers should find a more precise way of measuring user engagement with smartphones - in general and separately for various social media. Equally, engagement time could be measured through the app itself and not by asking the participant to make an estimation. Instagram is already offering that type of data for its users. This research showed that time spent on Facebook was positively correlated with time spent on Instagram and the same was found for checking frequency, which means that people who spent a lot of time on one network had a tendency to spend a lot of time on the other and vice versa. That is why it would be very useful to have a unique measurement of total social media usage.
Interestingly, when separate analyses were made for high-school and university students, the higher frequency on Instagram checking predicted a better result on the memory task. This notion brings us to the matter of causality in this instance. If it was true that individuals who checked social media more often had a better working memory capacity, then a question presents itself - which came first: the chicken or the egg? Do students who engage more in social media multitasking have a bigger working memory capacity exactly because their cognitive abilities allow them to do so, or did the frequent multitasking somehow positively affect the cognitive abilities related to memory? This question still remains unanswered and there is a lot of research yet to be done in this field.

D. Conclusion

In conclusion, the results did not provide evidence to support the idea that social media is damaging to cognitive abilities. Based on the results of this research, a conclusion can be made that there is no connection between the time spent or the frequency

of checking social media on working memory capacity. In the case of university students, the result was just the opposite of the research hypothesis – the frequency of Instagram checking was found to be a positive predictor of memory task results. Positive correlations were found between the time spent on Facebook with the time spent on Instagram and the frequency of Facebook checking with the frequency of Instagram checking. A consensus on this topic is yet to be made and consequently there is much more research to be done.

REFERENCES

1. D.A. Parry, D.B. le Roux, „Media multitasking and cognitive control: A systematic review of interventions“, Computers in Human Behavior, vol. 92, pp. 316-327, 2019
2. R. Junco, S.R. Cotten „No A 4 U: The relationship between multitasking and academic performance“, Computers & Education, vol. 59, no. 2, pp. 505-514, 2012
3. W.W.F. Lau „Effects of social media usage and social media multitasking on the academic performance of university students“, Computers in Human Behavior, vol. 68, pp. 286-291, 2017
4. Y.F. Chen, S.S. Peng „University students' Internet use and its relationships with academic performance, interpersonal relationships, psychosocial adjustment, and self-evaluation“, Cyberpsychology and Behavior, vol. 11, no. 4, pp. 467-469, 2008
5. M.L. Courage, A. Bakhtiar, C. Fitzpatrick, S. Kenny, K. Brandeau „Growing up multitasking: The costs and benefits for cognitive development“, Developmental Review, vol. 35, pp. 5-41, 2015
6. J. Mao „Social media for learning: A mixed methods study on high school students' technology affordances and perspectives“, Computers in Human Behavior, vol. 33, pp. 213-223, 2014
7. T. Packiam Alloway, J. Horton, A.G. Alloway, C. Dawson „Social networking sites and cognitive abilities: Do they make you smarter?“, Computers & Education, vol. 63, pp. 10-16, 2013
8. T. Ryan, S. Xenos „Who uses Facebook? An investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage“, Computers in Human Behavior. vol. 27, no. 5, pp. 1658-1664, 2011
9. G.G. Scott, C.J. Hand „Motivation determines Facebook viewing strategy: An eye movement analysis. Computers in Human Behavior“, vol. 56, pp. 267-280, 2016
10. A.S. Kahn, T.M. Martine „Text and you might miss it? Snap and you might remember? Exploring *Google effects on memory* and cognitive self-esteem in the context of Snapchat and text messaging“, Computers in Human Behavior, 104, 2020 in press
11. S.A. Himi, M. Bühner, M. Schwaighofer, A. Klapetek, A., S. Hilbert „Multitasking behavior and its related constructs: Executive functions, working memory capacity, relational integration, and divided attention “, Cognition, vol. 189, pp. 275-298, 2019
12. A. Stefanidi, D.M. Ellis, G.A. Brewe „Free recall dynamics in value-directed remembering“, Journal of Memory and Language. vol. 100, pp. 18-31, 2018
13. L.L. Bowman, L.E. Levine, B.M. Waite, M. Gendron „Can students really multitask? An experimental study of instant messaging while reading“, Computers & Education, vol. 54, no. 4, pp. 927-931, 2010
14. S.T. Frein, S.L. Jones, J.E. Gerow „When it comes to Facebook there may be more to bad memory than just multitasking“, Computers in Human Behavior, vol. 29, no. 6, pp. 2179-2182, 2013
15. M. Naveh-Benjamin, M.S. Brubaker „Are the effects of divided attention on memory encoding processes due to the disruption of deep-level elaborative processes? Evidence from cued- and free-recall tasks“, Journal of Memory and Language, vol. 106, pp. 108-117, 2019
16. R.K. Flor, K.C. Monir, A. Bita, N. Shahnaz „Effect of Relaxation Training on Working Memory Capacity and Academic Achievement in Adolescents“, Procedia - Social and Behavioral Sciences, vol. 82, pp. 608-613, 2013

Modernized Courses in Automotive Software Engineering

Ivan Kaštelan*, Bogdan Pavković*, Mario Vranješ** and Miroslav Popović*
* University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
** Josip Juraj Strossmayer University of Osijek, Faculty of Electrical Engineering, Computer Science and Information Technology, Osijek, Croatia
ivan.kastelan@rt-rk.uns.ac.rs

***Abstract* - In order to respond to the rapid expansion of the information-communication technology sector and the ever-increasing demands of the automotive industry, universities need to adapt and take active part in creating healthy business environment and provide education to the young generation which will be ready for challenging tasks in this rapidly developed field. The Faculty of Technical Sciences in Novi Sad and Faculty of Electrical Engineering, Computer Science and Information Technology in Osijek have participated in the common project DRIVE as part of the IPA Cross-Border Programme, with the goal of equipping the laboratories and developing the course materials in the field of automotive software engineering. Four laboratories have been set up and four new courses have been developed to support the project goals. This paper describes the newly developed courses covering the most important knowledge and skills for the automotive engineer of the future: functionally safe software design, software verification, machine learning, image and video processing in automotive systems. The developed courses are targeting the upper years of undergraduate studies and graduate studies.**

Keywords* - *automotive, automotive software, cross-border cooperation, laboratories

I. INTRODUCTION

Information and communication technology (ICT) sector is rapidly expanding. Although promising, it does not have a satisfactory business environment support that can adapt fast to trends and demands. In addition, one of the leading European employers, the automotive industry, has included in its focus the software and embedded systems. As a result, there is a growing need for engineers with adequate skills and ICT companies that can respond to the demand [1]-[2].

In order to respond to the fast-growing needs of automotive industry and enhance the competitiveness of the region in education of future engineers specialized for automotive software and hardware, University of Novi Sad, Faculty of Technical Sciences (FTN), Serbia and Faculty of Electrical Engineering, Computer Science and Information Technology (FERIT) in Osijek, Croatia have joined in a cross-border cooperation project "Modernizing Laboratories for Innovative Technologies" (DRIVE) as part of the European Union's INTERREG IPA program. The aim of this project is to equip the two partner institutions with latest laboratory equipment for development of automotive software and develop the curriculum with study materials for graduate-level education of computer engineers to specialize them for engineering in automotive industry. The project started on 15 July 2017 and lasted for 29 months, until 14 December 2019.

The project DRIVE aimed to increase the competitiveness of the region and make it closer to the already developed parts of Europe. By making the targeted region competitive in the automotive market, it aimed to decrease the difference between European regions, strengthening the entire Europe in the competitive world of automotive industry. Engineering of embedded computer systems applied to automotive systems is a new field within the automotive industry and the engineers working in this field have less than 20 years of experience. Nevertheless, it is one of the fastest-growing fields with a high impact on economy in terms of the Gross Domestic Product (GDP) [3]. The industry faces numerous challenged in infrastructure since the development suffers if the infrastructure does not keep pace with the market demands. Still, the infrastructure cannot be developed without skilled people. Skills are best acquired through formal studies, which stresses the importance of quality education programs.

In addition to increasing the competitiveness of the region, the project DRIVE aimed to allow younger generations to study on state-of-the-art technologies and gain practical skills on modern laboratory equipment which will make them ready for work in the automotive industry. [4]-[6] One of the main project results is the curriculum proposal for the education in the field of automotive software engineering, targeting final years of undergraduate academic studies and the graduate academic studies. Four courses which represent part of the proposed curriculum were fully developed in this project, covering the most important knowledge and skills for the automotive software engineer of the future: functionally safe software design, software testing and verification, machine learning, with the focus on deep learning and image and video processing in automotive systems and systems of autonomous networked vehicles.

The research leading to these results has received funding from the European Union's Interreg IPA CBC Croatia-Serbia Programme (2014-2020) under subsidy contract no. 2017HR-RS60.

The rest of the paper is organized as follows: section II gives a brief overview of the procured equipment and established laboratories; section III describes the four courses that were developed in the project and section IV provides the plans for post-project activities, as well as conclusions.

II. THE LABORATORIES

One of the main goals of DRIVE project was to equip four laboratories on two locations, two in Novi Sad and two in Osijek, which will serve as laboratories for teaching laboratory activities in the courses within the automotive software engineering field. Laboratories also served to host workshops for engineers which the faculties FTN and FERIT offered to institutions in the region.

At FTN and FERIT the following four laboratories have been established [6]:

- Laboratory for functionally safe automotive architecture and software design,
- Laboratory for hardware-in-the-loop testing, calibration and diagnostics.
- Laboratory for design and testing of automotive-related software,
- Laboratory for image and video processing in automotive applications.

The laboratories are equipped with state-of-the-art equipment for development and verification of software for automotive systems, which consists of:

- equipment that supports automotive application development in Automotive Open System Architecture (AUTOSAR) for automotive electronic control units (ECU) and supporting software, part of which is shown in Fig. 1,
- advanced driver assistance system (ADAS) development boards and supporting software (Fig. 2),
- equipment for automotive application testing based on hardware-in-the-loop (HiL) simulation.



Figure 1. Part of the equipment that supports the fundamental automotive application development based on AUTOSAR architecture

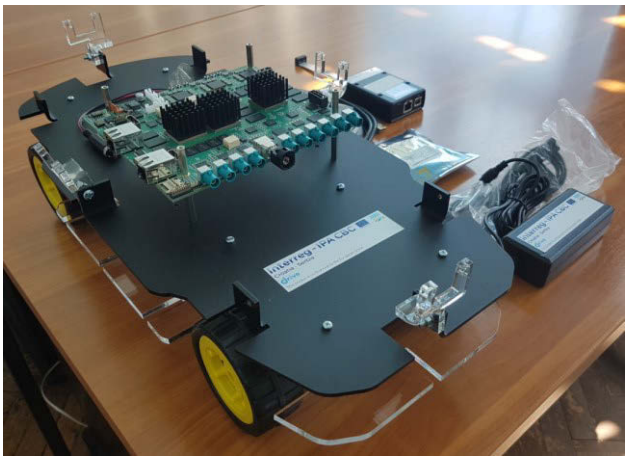


Figure 2. Development board for advanced driver-assistance system automotive application development with the automobile model ready for student use

III. THE COURSES

During the project, the curriculum for automotive software engineering education was proposed [6]. Part of this curriculum, which the project team considered to be the most important for providing the foundational knowledge and skills, was selected for complete development.

The four courses for which course materials and materials for laboratory exercises were developed are:

1. Digital image and video processing for autonomous vehicles,
2. Machine learning in systems of autonomous and networked vehicles,
3. Software architecture in safety-critical control systems,
4. Methods and techniques for automotive software testing.

This section provides description of the course materials that were developed in DRIVE project, and the examples of laboratory exercises the students are learning in the course.

A. Digital image and video processing for autonomous vehicles

The goals of this course are to introduce students to Advanced Driver-Assistance Systems (ADAS) features and to the features of digital images and video signals, as well as to the camera system in vehicles. The goal is to enable students to apply and develop advanced algorithms for processing of digital images and video signals, with an emphasis on algorithms for usage in autonomous vehicles.

Specifically, in the beginning of the course students are introduced to ADAS and in-vehicle camera systems. After that the topics regarding digital image acquisition, representing, filtering and noise removal are discussed. The next step includes the usage of image processing algorithms for edge detection, image segmentation and

object detection. Once students are familiar with how to detect an object, object tracking algorithms are explained. In that way they can practically detect all necessary objects in traffic, e.g. pedestrians, vehicles, traffic signs, etc.

The subsequent area is related to 3D scene reconstruction, where the stereovision and structure from motion (SFM) topics are presented and discussed. The last part regarding image processing explained within lectures is the analysis of complex ADAS algorithms, e.g. lane detection and lane departure warning (Fig. 3), pedestrian detection, surround view, driver monitoring, camera mirror replacement, etc.

At the laboratory exercises, when different image processing algorithms for different purposes are discussed, the real embedded platform (hardware and software) for their implementation is analyzed and explained. After that, the students are working in the lab on the platform in a way that they must implement different image processing algorithms used in ADAS, considering that the implemented solution must work properly in real-time.

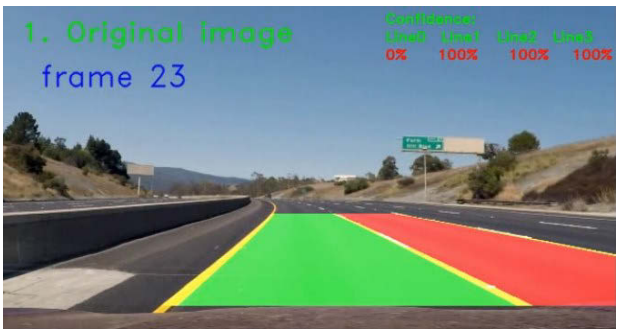


Figure 3. Visualization of the lane detection and lane departure warning algorithms the students are developing in the course laboratory exercises

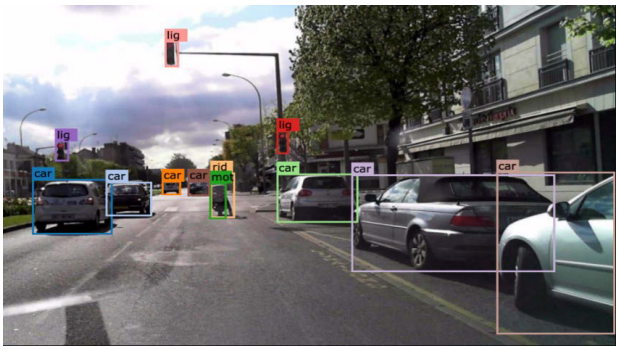


Figure 4. Visualization of the object detection and classification algorithms the students are developing in the course laboratory exercises

B. *Machine learning in systems of autonomous and networked vehicles*

This course is performed after the previous one. Namely, the goals of this course are as follows: (I) to introduce students to principles of data analysis and machine learning methods; (II) to enable students to apply machine learning methods in intelligent transport systems of autonomous and networked vehicles, focusing on image processing and deep learning; (III) to acquire appropriate skills for work with development tools for data analysis and machine learning, as well as with development tools that enable the implementation of the developed algorithms to the target platform.

At the beginning of the course students are introduced to the different types of machine learning (unsupervised, supervised learning, reinforcement learning) and different regression and classification methods are analyzed. The basics of deep learning followed by architectures and deep learning algorithms are discussed. Different types of deep neural networks, and specifically convolution neural networks, are presented to the students.

At the end of the lectures, different applications of machine and deep learning in intelligent transport systems, like segmentation, detection and classification of objects (signs, lines, vehicles pedestrians, etc.) in the image (Fig. 4), motion planning, centralized and distributed control of networked vehicles, are analyzed. Upon successful completion of lectures, students in lab are implementing different ADAS solutions based on deep neural networks on embedded ADAS platform.

C. *Sofware architecture in safety-critical control systems*

Designing and implementing software solutions for autonomous vehicles required special skills which are not required for most of the general software development tasks. This software must be safe and reliable with zero tolerance for errors as lives can be threatened if the software malfunctions while the vehicle is in motion. Therefore, students learning the automotive software must have training in the software architectures for safety-critical systems.

The aim of this course is to teach students to be able to understand and to design the architecture and safety critical software for automotive industry, and also to have a solid basic knowledge about several key concepts and standards necessary for understanding the safety aspect in automotive industry.

The course provides the participants with insights into the concepts and architecture of safety-critical control systems' software, explains to the participants the basics of AUTOSAR: concepts, architecture, methodology, building elements and ways of migration. The course also provides students with practical considerations: operating systems, software components, communication, input/output handling, handling of states, system services and memory, diagnostics. Students are introduced with the basics of the development of safe automotive software with a focus on ISO 26262 and basic requirements.

D. *Methods and techniques for automotive software testing*

The aim of this course is to teach students to be able to implement a framework for testing of automotive software and to have a solid basic knowledge about several key concepts necessary for understanding the design of testing process.

The students are first introduced to the testing of hardware-software real-world control loop systems and to the basic concepts of hardware-in-the-loop (HiL) testing, software-in-the-loop (SiL) testing, model-in-the-loop (MiL) testing, emulation, simulation, in-vivo, in-vitro and in-silico testing.

Overview of pros and cons with analysis of models follows in which students are taught about the transition between models: from HiL to MiL. All the models are studied using the examples from the automotive industry. Important components of the automotive system: electronic control unit (ECU), network, sensors and actuators are studied.

Testing process of an automotive system is specific, and the students are introduced to the entire testing process: modeling, configuration, calibration, testing, measuring and evaluation. An overview of existing testing tools is given.

At FERIT, the new study program "Automotive computing and communications" has already been established and more information about it can be found on an institutional webpage [7].

IV. Lessons Learned

The courses were implemented in the winter semester of the academic year 2019/20. FTN introduced the new courses as electives in the current master study program of Computing and Control Systems. FERIT introduced the new study program at the master level.

Since the courses are mainly practically oriented, evaluation was performed at the lab exercises. In order to examine students' success and motivation in carrying out the laboratory exercises, the short Lab Feedback Questionnaire (LFQ) was developed, which the students filled at the end of the course.

The survey showed some initial vulnerabilities of the platforms used in the courses, but it is an expected outcome in the pilot generation. Generally, students were satisfied with the easiness of usage of the laboratory platforms (67% responded with "agree" and "strongly agree" to the statement), and the value of learned materials (78% positive responses). The negative side was the time investment in the laboratory exercises and easiness to debug the problems in the solutions. This opens the door to opportunities for improvement in the following years.

V. Conclusions

The DRIVE project, part of the Interreg IPA CBC Croatia-Serbia program, allowed equipping the four laboratories in two institutions which will lead to the increase of competitiveness of the region and increase of the quality of teaching in the field of automotive software development.

DRIVE project has completed, but the results of the project will be applied in the years to come as the new study program has been established at FERIT and the complete set of new courses introduced at FTN which could pave the way for future accreditation of the corresponding study program there as well. Future research will be focused on expanding the set of courses and the set of laboratory exercises to cover more topics important for software development for automotive system. This field is rapidly developing nowadays, therefore it is necessary to regularly update the course materials to keep in pace with the state-of-the-art.

In addition, DRIVE project allowed establishing the collaboration between enterprises and academic institutions in the region which will create a positive feedback loop, benefiting both sides, as the better trained students will enter the workforce making the companies more competitive in this rapidly expanding industry, which will allow companies to offer even better opportunities for student internships and joint research with academia.

References

[1] F. Fabrini, M. Fusani, G. Lami, E. Silvera: "Software Engineering in the European Automotive Industry: Achievements and Challenges", IEEE International Conference on Computer Software and Application (COMPSAC), 2008, doi: 10.1109/COMPSAC.2008.140

[2] A. Haghighatkhah, M. Oivo, A. Banijamali, P. Kuvaja: "Improving the State of Automotive Software Engineering", IEEE Software, Vol. 34, Issue 5, 2017, pp. 82-86, doi: 10.1109/MS.2017.3571571

[3] P. Kaur, R. Sobti: "Current challenges in modelling advanced driver assistance systems: Future trends and advancements", IEEE International Conference on Intelligent Transportation Engineering (ICITE), 2017, doi: 10.1109/ICITE.2017.8056916

[4] M. Vranješ, G. Velikić, I. Kaštelan, D. Vranješ: „Projekt DRIVE – Put sveučilišta k automotiv industriji", 37th Conference on Transportation Systems with International Participation – Automation in Transportation, 2017

[5] I. Kaštelan, M. Popović, M. Vranješ, G. Velikić: "Work in Progress: Modernizing Laboratories for Innovative Technologies in Automotive", IEEE Global Engineering Education Conference (EDUCON), 2018, pp. 1700-1702, doi: 10.1109/EDUCON.2018.8363439

[6] I. Kaštelan, B. Pavković, M. Vranješ, M. Popović, G. Velikić: "Modernizing Laboratories for New Courses in Automotive Software Engineering", 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2019, pp. 667-670, doi: 10.23919/MIPRO.2019.8756648

[7] Sveučilišni diplomski studij „Automobilsko računarstvo i komunikacije", https://www.ferit.unios.hr/studiji/sveucilisni-diplomski-studij/#s52 (accessed on 04 February 2019)

Intuitive and rational information management

Irena Mikšić[*], and Tihana Babić[**], Ljubica Bakić-Tomić[***]
[*] Elementary school Popovača, Croatia
[**] Algebra University College, Zagreb, Croatia
[***] Baltazar Zaprešić University of Applied Sciences, Croatia
irenamiksic@gmail.com, tihana.babic@algebra.hr, ljbakictomic@bak.hr

Abstract - **Intuition is a term that can have different meanings. It can refer to intuitive knowledge as well as creative acts. Intuition belongs to the creative act and depends on unconscious resources, unburdened by rules of logic, contradictions, and dogmas of common sense. In the formal education, cognitive processes are used more, and the ratio is tied to logic, rules, and frames of thought, all of which must be justified, backed by reasons and arguments that serve to persuade about the truth or the delusion of the matter under consideration. Creativity and an exploratory spirit often involve intuitive processes. This research aims to determine whether students of computing, digital marketing, and market communications (Algebra University College) as well as students of economics (Baltazar Zaprešić University of Applied Sciences) rely more on ratio or intuition during different learning situations and are they threat different for each other? Although most of the students give priority to ratio, it has been shown that in reality they most often use a combination of intuition and ratio regardless of the field of study they are studying.**

Keywords – intuition, intuitive, ratio, rational, information management

I. INTRODUCTION

People's lives are the eternal movement from situation to situation, from interaction to interaction forming the everyday life of a man. Therefore, every person must constantly make decisions in different situations in life. Decisions similar to those made many times before seem easy to make, and big ones are life decisions that fundamentally change a person's life. Each person has their way of deciding. Some people base their decisions on information they receive from the outside world, through the senses, from their own previous experience, other people's experiences and rationally judge them, and some people base their decisions more on information coming from the "inside". This information is intuitive and depends on unconscious resources, unburdened by logic, contradictions, and reason.

Information from the outside can be conflicting and, if a person decides solely based on information he has previously collected and arranged, and now, after recognizing a similar "pattern" that our mind provides, then, the person is not sure what to decide and what is best for him. In the case of indecision and hesitation, many people reach for information from their insides, based on "foresight", often outside the logical pattern and contrary to the known rules of what would be considered "reasonable", that is, information from intuition.

II. INTUITION AND RATIO

A. INTUITION

Word comes from the Latin word "intuition" which means seeing, observing, knowing the truth and reality, regardless of experience or logical reasoning. Terms such as the sixth sense, extra-sensory perception, subconscious mind, unconscious mind, premonition, and hunch can be called intuition or often associated with it. It is the inner voice that serves as a kind of personal guide or advisor.

Agyakwa created a taxonomy of four different intuitive knowledge, and those are Extrasensory Perception and Preknowledge, Cognition including Abstract Logical and Arithmetic / Geometric Principles, Intuitive Insights into Aesthetic and Moral Areas, and Insights from Experts who Enter the Essence of Problems without Explicitly Rational Thinking [1].

According to Diane Brandon, there are seven types of intuition: presentiment, abdominal nausea, Aha! a moment of sudden surprise, vision, instinct, unpleasant feeling that does not pass, touch, smell, hearing, physical signs, telepathy. These are all ways in which intuition communicates with a person and transmits important information to him/her or for someone else [2].

B. RATIONALITY

The ratio is based on reason, logic, data, and information. Decisions are made after reviewing and evaluating the information or situation in which we are.

The main source of information that the mind processes are from sensory-perceived reality, from our earlier experiences, from other people's experiences. The main criterion is reasonable thinking, and living and behaving come from conscious decisions that follow from the precepts of reason.

Aristotle distinguishes between passive and active reason whereby the passive part accepts impressions and unites them, while in the active part, consciousness as a reasoning force creates thought connections and conceptual meanings. Kant, however, says that through various categories, reason interconnects the sensory data and information that then forms the objects of experience. Hegel, within his system, attributes to reason the formally logical actions while giving the mind greater and broader meaning that goes beyond reason itself [3].

III. INTUITIVE AND RATIONAL INFORMATION MANAGEMENT

Information management evolved in the 1970s and 1980s in response to the need to interpret, use, create, and store information in the burgeoning economy of the Western world and the accelerated computerization of society. Information management is about manipulating information, evaluating information, acquiring information, creating information, storing, interpreting, and using information.

The principle of managing information in organizations can also be used to personally manage information to respond to new situations as quickly as possible, thus saving time. In this sense, information management becomes a set of activities for a person to transform that information into knowledge. For information management to become knowledge management, it is important to keep in mind that not all types of information are equally valuable and can be managed in different ways.

In the present time, it has become necessary for every person to access and manage the information obtained in the most productive way, as it has been shown that skillful information management does not necessarily lead to increased productivity or innovation. With the informatization of society, the information became publicly available. Excess information in the mind creates noise that results in confusion and difficulties in managing information, knowledge, and decision making. Heuristic information management benefits the development of intuition and intuitive information management.

Scientific research has shown that while making the right decisions it is not a trick to accumulate information, but to discard it: one must know intuitively what one does not need to know by reason. John Naisbitt confirmed this by saying, "Intuition is becoming more and more important in the new information society, precisely because there is so much information around us" [4].

If intuition is a superconscious, irrational source of information, can we can consciously and rationally manage the information we receive from intuition at all?

We can consciously choose to "hear" intuitive information and process it rationally and either accept it or reject it. However, we have to be careful here because many people live in the space of their consciousness between instinct and intellect, so they are often unaware of the information they receive from intuition or those information collide with various emotions, and then they cannot recognize the real source of the information.

Also, to accept information from intuition one needs to trust it. If a person does not trust his or her intuition because of his/her upbringing or has decided that only rational information is relevant, the intuition information he / she receives will not be taken as true and will be rejected.

Gerard Hodgkinson said, "People tend to experience intuition when pressured for lack of time or in a situation where they have too much information to resolve, and conscious analysis of the situation is difficult or impossible" [5].

In which way the rational information and intuitive information in the form of creative and original momentum can produce the best results in decision-making and problem-solving was proposed in 1926 by American scientist Graham Wallas, a well-known scheme of the creative thinking process today. Wallas divided this process into four stages:

- the first phase – PREPARATION - involves gathering relevant information about a problem, finding a rational solution, and thinking about it.
- the second phase – INCUBATION - putting the problem aside. Period of apparent halt. In reality, deep, unconscious work on the task unfolds whereby at a conscious level one does not have to think about it at all.
- the third phase - ILLUMINATION - inspiration, discovery, intuitive information (Aha! Effect). It always comes unexpectedly, instantaneously, and similar to a fast gallop. At that moment, the solution is born in the form of symbols, images that are difficult to describe in words.
- phase four – CHECKING - the form is shaped in words, thoughts are put in a logical order, discovery is scientifically argued, the physical manifestation of a solution or decision arises [6].

In their scholarly work, People Like Logical Truth: Testing the Intuitive Detection of Logical Value in Basic Propositions, Hiroko Nakamura and Jun Kawaguci, and according to De Neys, W. Bais and Conflict, timely confirm that apart from managing information and making decisions based on by understanding the mind, we can equally validly manage information from intuition: Recent reasoning studies proposed that people possess intuitive logic as well as intuitive heuristics and that these two intuitive processes enable people to intuitively detect heuristic-logic conflicts and engage in analytic processes [7].

T. Babic, S. Papic, and M. Babic also suggest in their scholarly work Intuitive Communication versus Rational Communication among Computing Students at Algebra University College that decision making through information management using intuition is faster, more efficient, and more accurate. "When a decision-making process is based on intuition, decision-makers consciously recognize a problem through the perception of relevant cues and patterns, non-consciously activate all the cognitive schemas associated with the problem, non-consciously make holistic associations across cognitive schemas, and consciously generate a solution. The intuitive process includes problem definition, analysis, and synthesis, just like rational information processing, but these stages occur faster and are mostly non-conscious and deeply intertwined. Additionally, intuitive judgment is effectively charged and accompanied by a feeling of certitude and the perception that one's intuitions are correct, despite the lack of rational analysis" [8].

As far as business is concerned, businessmen are divided when considering the use of intuition in business decision-making, with some saying that there is not a single top manager who has not listened to their gut feeling and that they have had exceptional instinct (Forbes) [9]. But maybe those who say, like in the Harvard Business Review, that our gut is untrustworthy may not be wrong. If we

consider that, as Bruce Henderson said, founder of the Boston Consulting Group, in 1977, that the intuition is "the subconscious integration of all the experiences, conditioning, and knowledge of a lifetime, including the cultural and emotional biases of that lifetime, and researchers have shown that our unconscious desire to identify patterns is so strong that we routinely perceive them where they don't in fact exist" [10].

It can be concluded that an overwhelming desire for "superhuman power," as Eric Bonabeau calls it in Harvard Business Review, can lead to terrible decisions, because our brains, when confronted with a new phenomenon, try to categorize it based on our previous experiences, to fit it into one of the patterns stored in our memories. But, in making that fit, some people inevitably filter out the very things that make the new phenomenon new - and rush to recycle the reactions and solutions from the past [10].

How to distinguish between ratio-imposed solutions and intuitive solutions that come from "higher self" may be the biggest question of them all. After humans have mastered the tools of logical reasoning, perhaps scientists should embark on the development of tools that can help us in the deeper internal development of human beings, and thus the intuition that is placed "within" man.

V. THE RESEARCH METHODOLOGY

A. The Research Goals

The general research objective was to examine and determine whether students predominantly base the way they manage information relying on their intuition or their ratio, in their private, as well as in the professional environment.

1. *Do students rely on intuition in different life situations?*
2. *Do students base their information management on intuition or ratio?*
3. *To what extent do students consider intuition as a reliable information management tool?*

B. The Research Sample

The research was conducted among the Baltazar Zaprešić University of Applied Sciences and the Algebra University College students. A sample of participants was statistically significant.

The total number of population sample was 205 participants, which included 125 Baltazar Zaprešić University of Applied Sciences students of economics (61 %) and Algebra University College 80 of computing, digital marketing and market communications students (39 %). The majority of participants were undergraduate students; 170 of them (83.4 %), while 34 (16.6 %) of participants were graduate students. 139 of students (67.8 %) were female and 66 (32.2 %) were male.

The structure of all surveyed students according to the study program is presented in Chart 1.

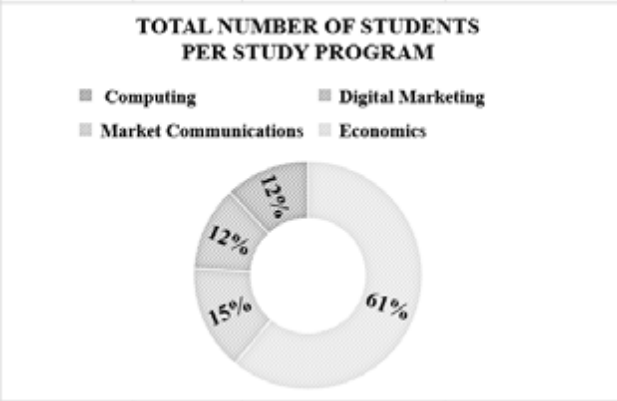


Chart 1. The structure of the participants per Study Program

C. The Research Methods

The research was conducted through an anonymous voluntary survey during the winter semester of the 2019/2020 academic year.

A specially designed questionnaire in the Google Forms tool had 12 closed-ended questions, of which 4 were related to demographics. One of these questions included 10 statements for which respondents could indicate the degree of agreement with the statements. For 4 questions answers were defined with the degrees of frequency according to the Likert scale. 3 questions had predefined 5 answers related to intuition and ratio. To ensure a clear understanding of the terms, a descriptive definition of terms of intuition, ratio, and information management was specified at the beginning of the survey.

A quantitative method was used for the analysis of the research results. Through the Google Forms tool, was made the data processing of the survey results.

D. The Research Results

1. *Do students rely on intuition in different life situations?*

The research results have shown that 57% of the total number of surveyed students often or always rely on intuition in everyday life, and one-quarter of them rely on intuition periodically (26 %). Only 1 % of students never rely on intuition in everyday life while 8 % rarely rely on it. Distribution of the answers if students rely on intuition in different life situations is presented in Chart 2.

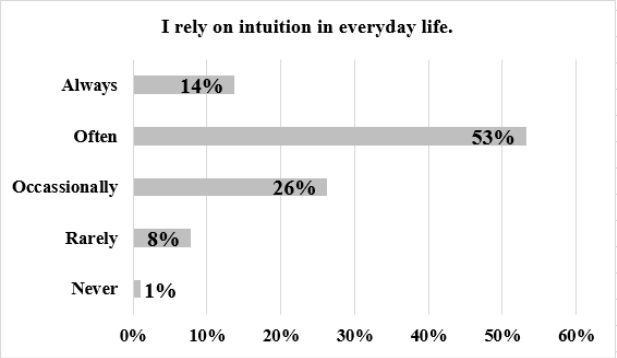


Chart 2. Distribution of the answers to the question: *How often do you rely on intuition in everyday life?*

When it comes to relying on intuition in different life situations 48.3 % of the students claim that they do it periodically and 9.8 % do it rarely. On the contrary, 38 % rely on intuition in most cases, while 4.4 % claim that they do it always. Only 1% of students never rely on intuition in different life situations.

More than half of the surveyed students (54.6 %) consider that other people periodically rely on intuition in different life situations and 22.4 % of the students claim that other people do it rarely. On the contrary, 21 % consider that other people in most cases, 2.9 % that they do it always, while 1% consider that other people never rely on intuition in different life situations. It is interesting to notice that other people believe their intuition more often than they believe their intuition.

2. *Do students base their information management on intuition or ratio?*

In the professional environment, 45.9 % of the total number of students base their information management on a combination of intuition and rationality, but rationality prevails, and 36.1 % of participants rely on the combination of intuition and rationality to the same extent. Only 8.3 % of participants base their information management in the professional environment exclusively on rationality, and only 1 % exclusively on intuition, while 8.7 % use a combination of intuition and rationality, but intuition prevails. Distribution of the answers if students base their information management on intuition or ratio in the professional environment is presented in Chart 3.

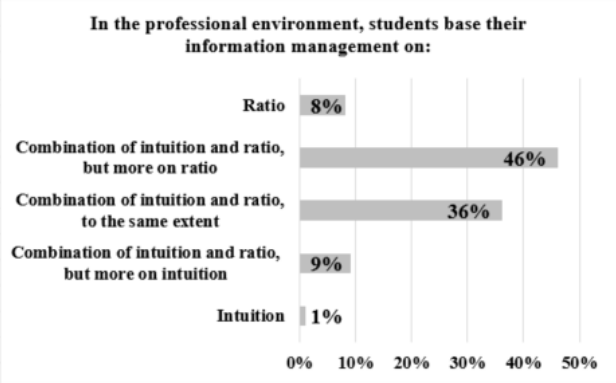


Chart 3. Distribution of the answers to the question: *In a professional environment, what is your information management based on?*

In the private environment, 33.2 % of total number of students base their information management on a combination of intuition and rationality, but rationality prevails, 30.2 % of students base their information management on a combination of intuition and rationality, but intuition prevails, while 27.8 % of participants rely on the combination of intuition rationality to the same extent. Only 5.4 % of participants base their information management in the private environment exclusively on rationality, and only 3.4 % exclusively on intuition. Distribution of the answers if students base their information management on intuition or the ratio in the private environment is presented in Chart 4.

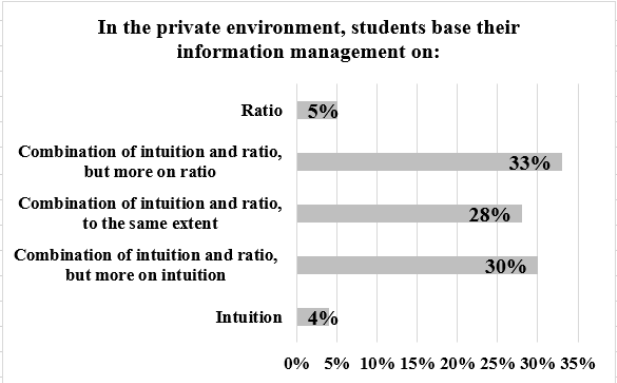


Chart 4. Distribution of the answers to the question: *In a private environment, what is your information management based on?*

Regarding the question does intuition help them to estimate if the received information is true, results have shown that the highest percentage of participants (47.8 %) consider that intuition helps them often, to 33.7 % helps periodically, and to 9.3 % of respondents intuition always helps. 8.8 % of students rarely rely on intuition to help them to estimate if the received information is true, and 1.5 % never do it.

In communication, 57.4 % of respondents agree that intuition and rationality have equal importance (of which 12.9 strongly agree), while 18.3 disagree (of which 4.5 % strongly disagree). 24.3 % of them neither agree neither disagree with the statement. In the decision-making process, 31 % of students agree that intuition is more important to them than rationality (of which 5.4 % strongly agree), 30 % disagree (of which 5.9 % strongly disagree), while 39 % neither agree neither disagree with the statement. In making any decision, 23.5 % of respondents agree that one should and can rely entirely on intuition, of which 5.5% strongly agree, 42 % disagree, of which 12.5 strongly disagree, and 34.5 % neither agree neither disagree with the statement. 26.9 % of respondents agree that for them making decisions for the future is impossible to judge rationally but purely intuitively, of which 4% strongly agree, 33.7% disagree, of which 9.5 % strongly disagree, but most of them; 38.7 % neither agree neither disagree with that statement.

3. *To what extent do students consider intuition as a reliable information management tool?*

For 12 % of students in total intuition is an unreliable tool, of which 4 % consider intuition very unreliable, but for most of them (50 %) intuition is a reliable tool, of which 9.3 % consider intuition very reliable. 38 % of students consider that is intuition neither a reliable nor an unreliable tool. Distribution of the answers if students consider that intuition is a reliable information management tool is presented in Chart 5.

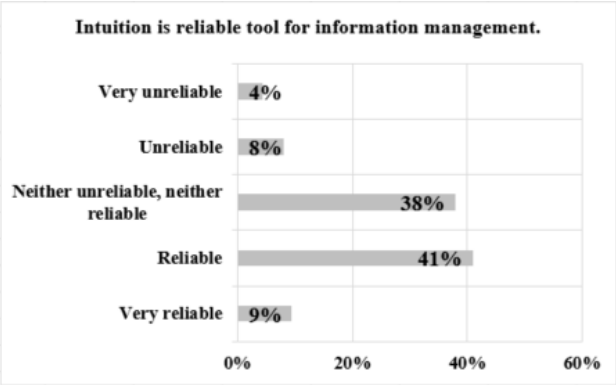


Chart 5. Distribution of the answers to the question: *According to your opinion, intuition is a reliable tool for information management?*

53.7 % of students agree that they always start with rationality, but if intuition persists to do the contrary - they listen to it, of which 15.4 strongly agree. 13.4% of students disagree with that statement, of which 2.4% strongly disagree, but 32.8 % neither agree neither disagree with the statement.

When it comes to work, 62.7 % of students agree that they give preference to rationality, of which 23.4 % strongly agree). 25.9 % neither agree neither disagree with the statement. Only 11.4 % of respondents disagree, of which 3.5 % strongly disagree. 72.4 % of students estimate that with the combination of intuition and rationality the best results are achieved, of which 31.2 % strongly agree with the statement. 90.5% disagree, of which 2.5 % strongly disagree, and 18.6 % neither agree neither disagree with the statement. Regarding the statement that intuition should be taken with a dose of the reserve because of the possible influence of emotions on it 66.4 % of students agree, of which 24.4 % strongly agree. 13.2 % disagree, of which 5% strongly disagree, and 20.3 % neither agree neither disagree with this statement.

It is interesting to notice that 56.3 % of student agree that whenever they ignore a strong inner feeling, always regret it later, of which 18.1 % strongly agree. 15.8 disagree, of which 5.5 % strongly disagree, 28.1 % neither agree neither disagree.

E. Limiting elements of the research and recommendations for further research

The study was conducted on a small sample, it would be advisable to conduct it on a larger sample.

Also, it would be advisable to consider longitudinal research of the development of students information management skills, which they have embraced under the influence of formal education – do they change attitudes toward the use of intuition and ratio with the increase of life experience, and what circumstances specifically influenced attitudes to change, if so.

Besides mentioned, in future research, the distinction could be made between male and female students, and it would be recommendable to check for statistically significant differences in relying on intuition between students from different scientific fields or study programs.

VI. CONCLUSION

The results of the conducted survey among 205 students in a total of the Baltazar Zaprešić University of Applied Sciences and the Algebra University College students showed that more than half of the surveyed students use intuition in everyday life.

It is interesting to notice that in the professional environment, almost half of the total number of students base their information management on rationality and more than half of surveyed students to give preference to the ratio in business. But in the private environment, they rely a lot more on intuition. It is worth mentioning that only every tenth student rarely or never relies on intuition.

Although they mostly agree that it is necessary to take intuition with a dose of the reserve because of the possible influence of emotions on it, it is interesting to notice that two-third of the total number of surveyed students confessed that whenever they ignore a strong inner feeling, always regret it later.

Unlike hard skills, which can be proven and measured, soft skills are intangible and difficult to quantify [11]. Precisely because of the difficulty of detection and measurability, initial surveys such as this must rely on the attitudes and assessments of the respondents, but this is not a reason to discontinue the research. Although the area of the so-called soft skills is still rather unexplored and blurred, this research aimed to contribute to clarifying the same, as experts agree that the future is no longer imaginable without them. When mastery-based expertise is mastered, what distinguishes the good from the great? Many great people throughout history have highlighted their inner feeling as a factor x that has raised them above average. Science still has a lot of work to do; how to determine if it is a skill that can be mastered, how to define it, how to measure it. But it is certainly important to continue research in this area because although computerization is based on smart devices, which could even master rationally based algorithms, it is precisely the "human", under whatever term we consider, will not be able to replace.

REFERENCES

[1] K. Agyakwa, “Intuition, Knowledge and Education”, Journal of Educational Thought, 22 (3), 1988.

[2] D. Brandon, „Intuicija za početnike“, Leo commerce, Beograd, 2015.

[3] Leksikografski zavod Miroslav Krleža “Razum”, 2020. Available at http://www.enciklopedija.hr/natuknica.aspx?id=52059 [accessed: 07.04.2020].

[4] D. Stevens, “How do you experience intuition in your life?” Posted on June 8, 2015. Available at https://yogaofthemind.com/how-do-you-experience-intuition-in-your-life/, [accessed: 07.04.2020].

[5] G. P. Hodgkinson, J. Langan-Fox, E. Sadler-Smith, “Intuition: A fundamental bridging construct in the behavioral sciences”, British Journal of Psychology, 99, 1-27, 2008.

[6] M. Popova, “The Art of Thought: A Pioneering 1926 Model of the Four Stages of Creativity”, Brainpickings, 2020. Available at https://www.brainpickings.org/2013/08/28/the-art-of-thought-graham-wallas-stages/, [accessed: 07.04.2020].

[7] H. Nakamura, J. Kawaguchi, “People Like Logical Truth: Testing the Intuitive Detection of Logical Value in Basic Propositions”, Published on December 30, 2016. Available at https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0169166, [accessed: 02.04.2020].

[8] T. Babić, S. Papić, M. Babić, "Intuitive communication versus rational communication among students of computing at the

Algebra University College," 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 731-736, doi: 10.23919/MIPRO.2019.8756971. Available at https://ieeexplore.ieee.org/document/8756971, [accessed: 01.04.2020].

[9] C. Wilson, "Trusting Your Gut Is The Best Business Tool You've Got -- If You Can Listen", Forbes. Posted on June 14, 2018. Available at https://www.forbes.com/sites/chipwilson/2018/06/14/trusting-your-gut-is-the-best-business-tool-youve-got-if-you-can-listen/#b304ff54b167, [accessed: 05.04.2020].

[10] E. Bonabeau, "Don't Trust Your Gut", Harvard Business Review. Posted in May 2003. Available at https://hbr.org/2003/05/dont-trust-your-gut, [accessed: 07.04.2020].

[11] D. Bortz, "Soft skills to help your career hit the big time", Monster, 2020. Available at https://www.monster.com/career-advice/article/soft-skills-you-need, [accessed: 23.08.2020].

Artificial Intelligence – a new topic in Computer Science curriculum at primary and secondary schools: challenges, opportunities, tools and approaches

Zuzana Tkáčová*,**, Ľubomír Šnajder * and Ján Guniš*
* Faculty of Science, Pavol Jozef Šafárik University in Košice, Košice, Slovakia
** Faculty of Education, Constantine the Philosopher University in Nitra, Nitra, Slovakia
zuzana.tkacova1@upjs.sk

***Abstract* - Within the context of the Industrial Revolution 4.0, the issue of Artificial Intelligence is becoming an important challenge throughout society. School plays an important role in preparing the young generation for emerging labor market changes. This paper presents opportunities for Computer Science curriculum innovations at primary and secondary schools that are currently being addressed in Slovakia. We present main ideas how to implement new available tools and approaches to teaching the topic of Artificial Intelligence. Based on the new educational materials we have developed, we have gained valuable experience from teacher training, pupil workshops as well as the pilot phase of testing directly in schools confirming that this topic can also be taught in an attractive and age-appropriate way.**

Keywords - artificial intelligence; curriculum; school informatics; teaching methodology

I. INTRODUCTION

Over the past few years, we have witnessed the significant rise of artificial intelligence (AI) in various areas of our daily life. The digital tools we encounter bring AI already built into their functionalities to simplify and enhance user experience - digital voice assistants, photo editing and recognition, optical character recognition, natural language processing, and more. Especially important is the impact of AI in industry and manufacturing, transport, medicine or even education. Good understanding of the current situation in AI and its possible prospects in the future raises new questions for educational institutions on new ways of learning, teaching and education [1] - the shift from strictly knowledge-based content to non-routine cognitive tasks (often referred to as 21st century skills) is increasingly resonating in recent pedagogical documents in many countries around the world. It also appears to be necessary to change the existing school curriculum with an emphasis on core concepts, interdisciplinary themes and competencies in order to make learning more enjoyable, meaningful and useful to pupils [2]. There is a clear need to define "AI competencies" beyond basic ICT competencies that would provide pupils with skills necessary for identification and solving problems using computing techniques, methods and technologies [3].

While educational robotics has already found its place in school informatics thanks to Arduino, LEGO and other programmable robotic kits, the AI-related curriculum is still under discussion. Touretzky et al. suggested 5 "Big Ideas" in AI that should be addressed in the K-12 classroom [4]:

Big Idea #1: Computers perceive the world using sensors

Big Idea #2: Agents maintain models or representations of the world and use them for reasoning

Big Idea #3: Computers can learn from data

Big Idea #4: Making agents interact comfortably with humans is a substantial challenge for AI developers

Big Idea #5: AI applications can impact society in both positive and negative ways

Currently, it is possible to find ideas for a variety of unplugged activities on AI [5] and machine learning [6], or examples of using AI tools for STEM projects [7]. Over time, new educational platforms are emerging, bringing appropriate didactically processed content and the necessary software (or hardware) tools to make it accessible, e.g. MIT Cognimates [8] suitable for kids aged 7-10, Machine Learning for Kids [9] for kids aged 10+, Microsoft AI School[10], or IBM TJ Bot [11] for upper secondary school pupils.

Popular educational programming tools also bring new extensions designed to work with AI/machine learning components. Scratch 3.0 [12] comes with two extensions - Text to Speech (collaboration with Amazon Web Services) and Translate (collaboration with Google). MIT App Inventor [13] allows image classification for mobile apps using a personal image classifier, which can be trained using our custom images.

In Slovakia, Computer Science is a compulsory subject from primary to upper secondary school. However, AI is not explicitly found in educational standards, which is also one of the reasons why teachers

do not address this topic in classrooms. Other reasons are the lack of teaching materials as well as the unpreparedness of teachers who have not met the AI topic during their university studies. As the results of our recently published secondary school survey [14] show, Slovak pupils perceive AI as a very interesting and important topic, but are aware of the fact that they lack knowledge, drawing all their information almost exclusively from movies and the Internet. Reflecting these results, we decided to create an AI extension to the existing national Computer Science curriculum followed by a basic set of educational materials, and to prepare teachers for their use in schools as part of the National Project "IT Academy - Education for the 21st Century" (accompanied by a set of subsequent minor projects).

II. AI CURRICULUM

We found the opportunity to broaden the existing curriculum in the educational standards section called "Information Society - Digital Technologies in Society". Our curriculum design is based on five pillars:

- **Tools:** *Which digital tools use AI? What abilities / features does AI add to these tools?*
- **Processes:** *How does AI work? How to embed AI functionality into a program?*
- **Data:** *Where does the data come from? What can we extract from the data? How do they affect AI during and after learning?*
- **Applications:** *Where is AI already used? What are the opportunities/limits?*
- **Impacts:** *What are the potential impacts on society? How might AI affect our life, work, study, career?*

Our aim was to give priority to active, experiential learning methods that would enable each pupil to build the necessary knowledge, skills and attitudes. Discussion, cooperation, active inquiry and creative work of all pupils must be an important part of the teaching process. The proposed pillars do not have to represent separate learning units - they should be continuously linked to create a comprehensive picture. From a practical application to a particular tool, to discovering how the tool works and what data it uses, to considering new opportunities, limits and impacts.

A. Educational materials

We designed the educational materials for the upper secondary Computer Science course as a general education subject in all types of schools (after some modifications some parts can be used also at the lower secondary school level). They are thematically divided into three areas each of which is elaborated into one lesson:

- *Discover artificial intelligence*
- *How a computer learns*
- *What we need data for*

We don't focus on specific software products - they serve only as demonstration tools for exploration. Also, we don't expect program implementation of neural networks by pupils - considering the planned time allocation, the main point is to show the possibility of using elements of artificial intelligence to extend the functionality of other programs (e.g. games, controls for Internet of Things, mobile applications, etc.). As specific software tools are gradually changing and evolving, we focus on the generalization of concepts and knowledge that arise during pupil research and experimentation itself.

A key part of the first lesson is the phase of exploring various web-based tools integrating AI:

- **Thing Translator** [15] for object recognition and translation into another language
- **Watson Speech To Text** [16] for speech processing and analysis
- **Watson Visual Recognit**ion [17] for recognizing objects in photographs
- **Watson Natural Language Understanding** [18] for text analysis

Pupils work in groups - each group has a specific tool. For each of these instruments, pupils have a brief instructional video and then explore the tool, its inputs, outputs, and suggest options for further use. Then, after a short presentation of the results of the pupil exploration, the activities continue with the **AutoDraw** tools [19] and **Quick, Draw!** [20], where they see the process of AI learning based on user interaction, which will be the starting point for the next lesson.

The second lesson is devoted to the presentation of the functioning of machine learning tools and neural networks. The aim is not to master these procedures in terms of programming, but to understand their principles, advantages, but also pitfalls or limitations. Initial inquiry activity demonstrates a way of recognizing faces in machine learning through feature extraction and then searching the database. The pupils first analyze the prepared photos of people and fill in the table with the monitored symptoms to identify who is in the new photo. The inquiry continues with group work again - pupils have to prepare suitable data for neural network learning, and then train and test the neural network. The **Teachable Machine** [21] web tool is used. After completing the activity, the pupils are explained the basic idea of neural networks. Their own inquiry should also lead them to the need to start to think intensively about the suitability of the data to be used in neural network learning, as well as the possible consequences of their inappropriate choice.

The last lesson demonstrates data and their use in machine learning and data mining. Pupils explore various examples of data usage on the Internet e.g. traffic monitoring in Google Maps, or using **Google Trends** [22]. The main activity is to interactively explore the network graph of pizza community using the **ConnectTheDots** [23] web tool from the Databasic.io portal, through which they discover the functioning of customer recommendation systems that pupils can encounter when shopping online.

B. Teacher training

As mentioned above, when introducing our educational materials into teaching, we had to think about teacher training as well, as the teacher plays a key role in successful implementation. In Slovak schools, an active, inquiry-based learning approach represents still an innovative way of teaching for Computer Science teachers, as instruction-based teaching still dominates schools. Therefore, in the pilot phase, we trained 90 teachers in an intensive course focusing on innovative approaches in teaching as well as working with our educational materials. The task of trained teachers is now to verify the effectiveness and appropriateness of the prepared educational materials and to provide valuable feedback for their further improvements.

C. Further steps

Since educational materials and teacher training were just the beginning of building our AI curriculum, there is now a demand from teachers for additional AI-based activities suitable for teaching. That is why we invited also IT companies and research institutes for further cooperation. We are working on an advanced collection of ideas for simple, but interesting programming tasks using AI components and the use of other AI-based teaching tools. Creative AI workshops (e.g. robotics, generative art, data analytics and machine learning) for teachers and pupils and webinars with AI experts are starting to take place all over Slovakia in the near future.

III. CONCLUSION

The aim of education is to prepare pupils for successful participation in society - and scientific predictions agree that the society of the future will be strongly influenced by AI. The school should be a place where pupils acquire the necessary knowledge, skills and attitudes to be beneficial and well-prepared members in this society. All countries will face this challenge - Slovakia, as a strong industrial country, will be significantly affected by the impact of Industry 4.0. That is why the National Project "IT Academy - Education for the 21st Century" is trying to reflect these trends and bring to schools a change in Computer Science curriculum, as well as innovative and attractive approaches that have the potential to reach and motivate pupils to further study of STEM subjects towards STEM-related professional careers.

We started to test the presented educational materials with pupils during workshops and with Computer Science teachers during teacher trainings, where we confirmed that when processed in an active and attractive way, the topic AI can attract both pupils and teachers and support their interest in discussing the topic of AI and its impact, benefits and pitfalls on society. Pilot verification in education is currently taking place at secondary schools throughout Slovakia. Teacher feedback is very promising and confirms the fact that AI as a topic should - along with progressive task-based digital skills - have a firm place in the curriculum.

ACKNOWLEDGMENT

This article was created in the framework of the National project "IT Academy - Education for the 21st Century," ITMS: 312011F057, which is supported by the European Social Fund and the European Regional Development Fund in the framework of the Operational Program Human Resources and in the frame of project KEGA 029UKF-4/2018 "Innovative Methods in Programming Education in the University Education of Teachers and IT Professionals".

REFERENCES

[1] I. Tuomi, " JRC Science for Policy report: The Impact of Artificial Intelligence on Learning, Teaching, and Education," Luxembourg: Publications Office of the European Union, 2018.

[2] W. Holmes, M. Bialik and Ch. Fadel, "Artificial Intelligence In Education: Promises and Implications for Teaching and Learning," Boston, MA: Center for Curriculum Redesign, 2019.

[3] F. Pedró, M. Subosa, A. Rivas, and P. Valverde, "Artificial intelligence in education: challenges and opportunities for sustainable development," in Working papers on education policy, vol. 7, UNESCO, 2019.

[4] D. Touretzky, Ch. Gardner-McCune, F. Martin, and D. Seehorn, "Envisioning AI for K-12: What Should Every Child Know about AI?," in The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19) Proceedings, 2019, pp. 9795-9799.

[5] A. Lindner, S. Seegerer, and R. Romeike, "Unplugged Activities in the Context of AI," in ISSEP 2019, LNCS 11913, S.N.Pozdniakov and V. Dagienė, Eds. Springer, 2019, pp. 123-135.

[6] E. Ossovski, and M. Brinkmeier, "Machine Learning Unplugged - Development of a Workshop About Machine Learning," in ISSEP 2019, LNCS 11913, S.N.Pozdniakov and V. Dagienė, Eds. Springer, 2019, pp. 136-146.

[7] M. Zimmerman, "Teaching AI: Exploring New Frontiers for Learning," Portland, OR: International Society for Technology in Education, 2018.

[8] "MIT Cognimates," http://cognimates.me/

[9] "Machine Learning for Kids," https://machinelearningforkids.co.uk/

[10] "AI School," https://aischool.microsoft.com/

[11] "IBM TJBot," https://www.research.ibm.com/tjbot/

[12] "Scratch," https://scratch.mit.edu/

[13] "Artificial Intelligence with MIT App Inventor," https://appinventor.mit.edu/explore/ai-with-mit-app-inventor

[14] Z. Tkáčová, "Umelá inteligencia ako súčasť kurikula informatiky na základných a stredných školách," in DIDINFO 2019 Proceedings, Banská Bystrica: Univerzita Mateja Bela, 2019, pp. 164-168.

[15] "Thing Translator," https://thing-translator.appspot.com/

[16] "Watson Speech to Text," https://www.ibm.com/cloud/watson-speech-to-text

[17] "Watson Visual Recognition," https://www.ibm.com/cloud/watson-visual-recognition

[18] "Watson Natural Language Understanding," https://www.ibm.com/cloud/watson-natural-language-understanding

[19] "AutoDraw," https://www.autodraw.com/

[20] "Quick, Draw!," https://quickdraw.withgoogle.com/

[21] "Teachable Machine," https://teachablemachine.withgoogle.com/

[22] "Google Trends," https://trends.google.com/

[23] "ConnectTheDots," https://databasic.io/en/connectthedots/

Inquiry-Based Python Programming at Secondary Schools

J. Guniš*, Ľ. Šnajder*, Z. Tkáčová* and V. Gunišová**
* Pavol Jozef Šafárik University in Košice, Košice, Slovakia
** Secondary grammar school of Ján Adam Rayman, Prešov, Slovakia
jan.gunis@upjs.sk

***Abstract* – Programming is a traditional part of informatics education from Primary to Upper Secondary Education in Slovakia. We have seen a growing interest of using Python in programming at our secondary schools during the past six years. However, the transition from using the still prevalent Pascal to Python is often only formal, regardless of the language specifics of Python. In addition, teaching programming is oriented primarily to mastering language constructions. In the frame of the National project IT Academy we have focused on inquiry-based programming teaching. In our teaching and learning materials we exploit the 5E learning cycle. Firstly, pupils are engaged in a topic, then they explore a topic and give explanation for their findings, then elaborate on their learning and finally are evaluated. Their learning is not focused just on acquiring skills and experience using suitable Python commands and data structures, but also on using various problem solving strategies. The paper presents examples of teaching materials and interesting findings and comments from pilot teachers.**

Keywords – Inquiry-based education; programming; Python; computational thinking; secondary school

I. INTRODUCTION

There are various programming paradigms, programming languages, various orders of teaching topics, as well as different teaching approaches used in teaching secondary school programming.

In Slovakia, teaching of programming has a long tradition. Algorithmic problem solving is one of five main areas of teaching informatics that is described in the National Education Programme (NEP) [1].

Under the topic Algorithmic problem solving, pupils from the 3rd grade of primary school to the last grade of secondary school solve algorithmic problems in various programming environments, starting with interactive iconic environments [2], through block environments (e.g. Scratch) to text programming languages (e.g. Python, Pascal, Logo, C, Java). Programming environments and languages are merely a means, not a learning objective; it is to develop the ability to solve problems using suitable tools of a programming language. The emphasis is gradually shifting from acquiring concrete experience and skills at primary school to mastering the basics of informatics at secondary schools, where this goal is already dominant.

Innovation of informatics is considered not only as content innovation (e.g. selection of suitable programming language or environment [3]), but also as innovation of learning methods. These methods according to [4] should make it possible to develop pupils' competences – creativity and innovation; communication and collaboration; research and information fluency; critical thinking, problem solving, and decision making; digital citizenship; technology operations and concepts.

These competences can be achieved by Inquiry-based learning (IBL) [5], which has no tradition in informatics teaching. IBL is naturally implemented in science education (e.g. Physics, Biology, Chemistry). In the implementation of IBL in programming teaching we have based on experience in implementing IBL in physics, chemistry and biology [6], research results in teaching programming [7, 8, 9] and our own experience with IBL in computer science [10, 11].

The results of researches aimed at teaching novice-programming were the most inspiring for our research. Kolouri et al. report in [12] as the main results of their research - the positive use of the syntactically simple Python language, teaching problem solving before teaching programming, the use of formative assessment tools. In their meta-analysis [13], Piteria and Costa mention programming topics from the highest to the lowest level of comprehension.

In the framework of the National IT Academy project, the authors of this article (two of them are didactics experts in informatics and two others are secondary school teachers) proposed the conception of teaching the subject and developed teaching and learning materials for the basic programming course with the IBL implementation. These results, achieved in cooperation with pilot teachers from participating secondary schools, are presented and discussed in the following parts of the article.

II. IMPLEMENTATION OF IBL IN TEACHING OF PROGRAMMING

A. Learning objectives

In specifying the content of education for the proposed teaching materials, we followed the NEP for secondary grammar schools with a four-year and five-year educational program for informatics subject [14]. Provided teaching materials for teachers cover the entire

content of the algorithmic problem solving area defined in the NEP. In each teaching material we specify more detailed objectives defined in the NEP, e.g. in the teaching material where pupils learn about the Python console, we stated:

Learning objectives defined in the NEP:

- To use mathematical expressions to express relationships.
- To solve problems where values need to be stored in memory, and then to use those stored values in expressions.

In our more detailed specification of learning objectives in the teaching material we add extra:

- To create and evaluate arithmetic expressions.
- To create and use a variable.
- To use appropriate, descriptive variable names.

In the informatics subject, especially in the area of algorithmic problem solving, we see great potential for the pupil development beyond the cognitive objectives defined in the NEP.

Computational thinking is a key competence for the 21st century [15]. It is the ability to solve problems, think and propose solutions of problems so that we can use computers efficiently to implement them. We tried to specify this thinking more precisely and to develop it through teaching in accordance with our teaching materials. For the purposes of our research, we used the framework defined in Barefoot [16]. Within this framework, 6 concepts of computational are distinguished:

- Logic – predicting and analyzing,
- Algorithms – making steps and rules,
- Decomposition – breaking down into parts,
- Patterns – recognizing and using similarities,
- Abstraction – removing unnecessary details,
- Evaluation – making judgements.

The concept of computational thinking is new to informatics teachers in Slovakia, they need to learn more about it and have also a set of practical examples how to develop it. Therefore, in our teaching materials we specifically describe the tasks (parts of their solutions) in which we intend to develop particular concepts of computational thinking, e.g.:

- Algorithms – creating of own problem solving algorithms (purchase price calculation).
- Decomposition – linear decomposition - linearly divide problems into smaller parts so that they can be used to achieve the goal (splitting the calculation into smaller steps).
- Abstraction – using essential elements of problems (solving verbal problems).

Pupils develop these computational thinking concepts in a particular lesson by solving a problem where they should suggest to the fast food retailer a way to calculate as quickly as possible the price of the goods sold, e.g.:

```
>>>hotdog = 0.55
>>>hamburger = 1.2
>>>fries = 0.7
>>>2 * hotdog + 5 * hamburger + 3 * fries
```

B. Content of innovative teaching materials

Python itself is not our goal. The dominant goal that we want to develop is pupils' ability to solve problems algorithmically. In our teaching materials we purposefully develop computational thinking of pupils. We achieve this by appropriate formulating problems (see Fig. 1.) and also by order of topics. E.g. we included topic of functions (defining and calling) among beginning lessons, because we want to develop ability of problem decomposition as an important concept of computational thinking. This ability allows pupils to solve more complex problems by programming smaller functions solving particular sub-problems.

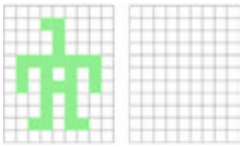


Figure 1. Task: Create the stamp by which can by filled the entire shape of the pawn in the image

We assume that approximately one third of the total hourly subsidy (3 x 33 hours) is devoted to algorithmic problem solving at secondary grammar school. We have designed a total of 27 teaching materials covering 32 lessons (Table I)

TABLE I. DESIGN OF LESSONS EXPLOITING INNOVATIVE TEACHING MATERIALS

Visualization of the calculation using turtle graphics (turtle module)
01. Introduction to programming, variables, calculations in console, errors
02. Sequence programming, turtle graphics, drawing simple geometric shapes
03. Definition of functions without parameters and without return value
04. FOR statement
05. Recapitulation of the thematic area and pupils testing
06. Definition of functions with parameters – drawing tasks
Console applications, problem solving
07. Definition of functions with parameters and with return value – numerical tasks
08. Errors in programs, debugging
09. IF statement
10. Recapitulation of the thematic area and pupils testing
11. Strings and manipulation with part of strings (indices, slices)
12. Strings and their methods, nested IF statements
13. Algorithms on strings
14. Exception handling
15. Determine problem parts of program, raising exceptions
16. Recapitulation of the thematic area and pupils testing
17. Lists and their methods
18. Lists – creating and modifications, random values
19. Algorithms on lists
20. WHILE loop statement
21. Nested statements (loops and ifs)
22. Recapitulation of the thematic area and pupils testing

Graphical User Interface (tkinter module)
23. Graphical User Interface – buttons, entries, labels, canvas
24. Graphical User Interface for own functions
25. Final project – determine problem, analyze of problem, solution of problem
26. Final project – implementation of Graphical User Interface
27. Final project – finalization, presentation and discussion

C. *Approaches and practices of teaching*

Informatics education is based not only on facts and knowledge, but also on methods and ways of acquiring this knowledge. We consider to be important for pupils to develop their inquiry skills (e.g. experimenting, predicting, verifying their predictions, critical reasoning, making conclusions, etc.) as well as a positive relationship to the scientific way of getting to know the world. We apply the 5E instructional model [17] in most of our teaching materials. The learning cycle consists of five phases (examples from the 4th lesson in which pupils encounter the loop concept):

Engage – motivation of pupils, finding out the pupils' initial knowledge about the studied topic. This phase can take the form of a discussion, a video presentation or a short story, solving an interesting problem, etc. The aim of this phase is to attract pupils to the topic of study and to increase their interest in the topic. For efficient use of cycles, it is important to recognize repeating parts/patterns of the problem, e.g. when drawing pictures (see Fig. 2.).

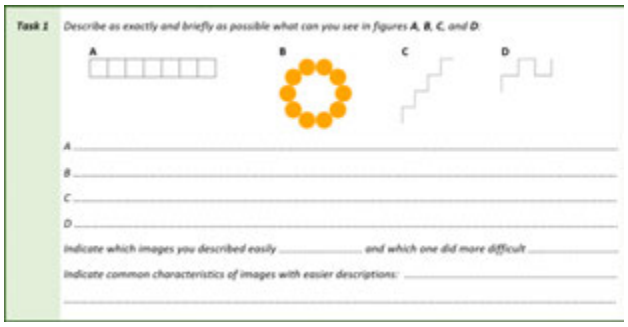


Figure 2. Task for recognizing repeating paterns in the images. (engage phase)

Explore – activities in which pupils come up with various ideas and hypotheses about how the studied system works. It is an independent work of pupils in which the teacher does not interfere or intervenes only minimally, e.g. in the case of guidance or clarification of the task. In this phase we often use Python console, which enables quick exploring and getting feedback from running parts of code without the need of writing complete code of program. Reading, comparing and comprehension of programming codes is the essential part of learning both programming language and programming as a process (see Fig. 3.)

Explain – the aim of this phase is to ensure that pupils linked their observations and findings from the previous exploration phase to their prior knowledge. Pupils should be able to explain what they have discovered, describe how system under study works (e.g. FOR statement). Emphasis is placed on substantial findings. At the end of the phase, the teacher summarizes the pupils' new knowledge or names it in the informatics vocabulary.



Figure 3. First touch with a new command (FOR statement) by comparing two program codes and further exploring the new command (changing the parameter in function range()). (explore phase)

Elaborate – at this phase, pupils practice and extend new knowledge (e.g. FOR statement). They solve various types of tasks (e.g. analytical, evaluation, synthetic), suggest possible improvements of solutions or other solutions of tasks. Typical task is to create a complete program for solving problem, e.g. to draw image with repeating pattern. (see Fig. 4.).

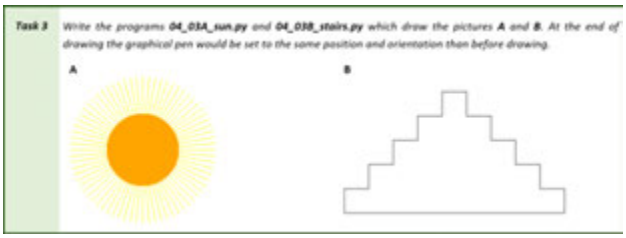


Figure 4. Solving tasks of varying complexity to practice FOR statement. (elaborate phase)

Evaluate – the aim of this phase is to obtain information on the level of pupils' acquired knowledge and misconceptions. This phase is often carried out by various forms of formative assessment, e.g. self-assessment test, self-assessment card, checklist. Typical misconception of control flow statement is that statement ends by a new line instead of indentation decreasing (see Fig. 5.)



Figure 5. Self-assessment test to obtain potential pupils' misconceptions about FOR statement. (evaluate phase)

After every 4 to 5 common lessons, we included a two lessons aimed at repeating and systematizing the acquired knowledge. Pupils do not learn new knowledge, they solve tasks (often graded) and fix and deepen the acquired knowledge. These two lessons also include a didactic test, its authorial solution and phenomena analysis. We have included project-oriented teaching materials in the final part of our Python course. The final lessons are focused on creating a product creation that is the result of a unique pupils' solution. Pupils take more responsibility for the

solution and in the case of group projects they choose which particular part of the project they will solve. An important part of the project solution is also the presentation of results and their evaluation.

D. Support for teachers and pupils

Teachers play an important role in implementing innovative teaching of programming, so we provide them with a number of supporting teaching materials. In the teaching materials we draw the teacher' attention to the key parts and also to the supposed problematic parts of the lessons. The final versions of the teaching materials will include teachers' experience in testing of the teaching materials and, where appropriate, alternative teaching approaches.

In addition to the teaching materials, we prepared two conceptual documents for teachers. One is more methodologically oriented and describes the whole concept of innovative teaching materials and teaching methods used (e.g. 5E learning cycle, formative assessment and project method). It also includes the concepts of computational thinking, which are described in more details. The second conceptual material is focused on the concept of teaching programming in Python where the specific content of the education and the reasons for the proposed lesson order are described.

We prepared worksheets, task collections and work files for pupils. The worksheets completed by pupils themselves can serve as notes from Python lessons. Our intention is that the worksheets do not restrict pupils; not to force them to strictly follow a defined sequence of steps, but to lead them to think about problems, use different problem-solving strategies, thoroughly analyze problems, and not just rely on trial-and-error approaches.

III. FINDINGS FROM TESTING OF THE TEACHING AND LEARNING MATERIALS

The teaching and learning materials have been verified at partner secondary schools in two runs. During the school year 2018/2019, there were 24 teachers involved in the verification and 20 teachers during the following school year. At least 8 teachers verified all 27 teaching and learning materials in each run. Based on their comments and recommendations, each year the original version of the materials was adapted.

The testing of the proposed teaching materials is currently still running at more than 25 secondary schools across Slovakia. We communicate with teachers via LMS Moodle. We receive feedback by:

- discussion forum posts,
- questionnaire for evaluation of the teaching materials after the lesson,
- quantitative and qualitative analysis of problems' solutions.

Based on feedback from teachers, here are the initial findings:

- Teachers appreciate comprehensive support for teaching (teaching materials, worksheets, teacher training, etc.),
- the level of pupils in the groups varies, that complicate teachers' pedagogical interventions,
- lessons are time consuming in cases where pupils do not have the necessary basic programming knowledge and skills,
- in the beginning pupils are not accustomed to IBL, they gradually adopt this approach and learn to use it effectively,
- pupils have difficulty using knowledge from other school subjects to solve problems with programming tools,
- pupils gradually learn to work together (by discussions, work in pairs),
- some teachers required additional tasks for their pupils,
- some teachers have proactively extended the content of the lesson, because of providing the complete learning content according to their opinions,
- only some teachers have identified themselves with inquiry-based teaching materials, especially those who have already had some knowledge and experience with inquiry-based teaching or have attended our training,
- teachers who have experience in IBL, as a student in our lessons, implement IBL better,
- pupils are not accustomed to using worksheets on informatics lessons,
- we initially assumed that teachers print worksheets for pupils, but because of the high printing costs, teachers asked for worksheets in editable electronic form, which some teachers find problematic because it is difficult for pupils to program and read and complete the worksheet on one screen at the same time,
- teachers tend to inappropriately transfer approaches from other programming languages to Python.

These and other findings from the testing we exploit in final versions of teaching materials.

IV. CONCLUSION

We consider solving problems with programming tools as a key area of school informatics. In this paper, we presented the proposal for an innovative way of teaching programming, using modern teaching approaches and practices developing inquiry skills and problem solving skills of pupils. We also respect specifics and approaches of Python programming.

Based on our experience [18] we recommend for better implementation of IBL in informatics, that the

teacher progress through the following roles: a pupil who explore, a teacher with theoretical bases of IBL using IBL lessons, a teacher adopting existing IBL lessons, a teacher creating own IBL lessons.

At the end of the project, efficacy of our methodological approach will be evaluated based on pre-testing and post-testing pupils' computational thinking level and feedback from teachers after their lessons. Final version of teaching materials will be released by the end of September 2020 under the CC BY SA license for use in all secondary schools.

Developed teaching and learning materials are unique in Slovakia and they provide comprehensive support for teachers and pupils. Ultimately, it is up to secondary school teachers to decide whether and to what extent they will use the materials in their programming lessons.

ACKNOWLEDGEMENT

This article was created in the framework of the National project IT Academy – Education for the 21st Century [19], ITMS: 312011F057, which is supported by the European Social Fund and the European Regional Development Fund in the framework of the Operation Programme Human Resources and in the frame of project KEGA 029UKF-4/2018 Innovative Methods in Programming Education in the University Education of Teachers and IT Professionals.

REFERENCES

[1] The National Institute for Education, National Education Programme, Bratislava, 2015, http://www.statpedu.sk/sk/svp/inovovany-statny-vzdelavaci-program/.

[2] I. Kalaš, A. Blaho, and M. Moravčík, Exploring Control in Early Computing Education, In: S. N. Pozdniakov and V. Dagiené (Eds.) ISSEP 2018, LNCS 11169, pp. 3-16, 2018, https://doi.org/10.1007/978-3-030-02750-6_1.

[3] M. Tomcsányiová and E. Klimeková, Case Study on the Process of Teachers Transitioning to Teaching Programming in Python, In: S. N. Pozdniakov and V. Dagiené (Eds.): ISSEP 2018, LNCS 11169, pp. 216-227, 2018, https://doi.org/10.1007/978-3-030-02750-6_1.

[4] International Society for Technology in Education, ISTE Standards Students, 2014, https://id.iste.org/docs/pdfs/20-14_ISTE_Standards-S_PDF.pdf.

[5] M. Rocard et al, Science Education NOW: A Renewed Pedagogy for the Future of Europe, 2007, http://ec.europa.eu/research/science-society/document_library/pdf_06/report-rocard-on-science-education_en.pdf.

[6] M. Kireš, Z. Ješková, M. Ganajová and K. Kimáková, Bádateľské aktivity v prírodovednom vzdelávaní, časť A, ŠPÚ, Bratislava, Slovakia, 2016, ISBN 978-80-8118-155-9.

[7] H. H. Hu and T. D. Shepherd, Using POGIL to help students learn to program, ACM Educ. 13, 3, Article 13 (August 2013), DOI: http://dx.doi.org/10.1145/2499947.2499950.

[8] J. Vaníček, Programming in Scratch using inquiry-based approach, In: A. Brodnik, J. Vahrenhold (eds.), ISSEP 2015, LNCS, vol. 9378, pp. 82–93, Springer, Cham, 2015, https://doi.org/10.1007/978-3-319-25396-1_8.

[9] A. Venkataraman, R. Ranawat, S. Vakil, J. Chen, S. Jagabathula and L. Subramanian, CollectiveTeach: Crowdsourcing Lesson Plans, Workshop on Advancing Education with Data, KDD'17, August 14, 2017, Halifax, NS,Canada, https://cs.nyu.edu/~ashwin/papers/kdd_collective_teach.pdf.

[10] Ľ. Šnajder and J. Guniš, Inquiry based learning of selected computer sciences concepts and principles, ICTE J. 1(1), 28–39, 2012, https://doi.org/10.1515/ijicte-2012-0003.

[11] S. Lukáč, Ľ. Šnajder, J. Guniš and Z. Ješková, Bádateľsky orientované vyučovanie matematiky a informatiky na stredných školách, Košice, Slovakia, Univerzita Pavla Jozefa Šafárika v Košiciach, 2017. ISBN 978-80-8152-487-5.

[12] T. Koulouri, S. Lauria and R. D. Macredie, Teaching introductory programming: A quantitative evaluation of different approaches. ACM Transactions on Computing Education (TOCE), Volume 14, Issue 4, February 2015, Article No.: 26, pp 1–28 https://doi.org/10.1145/2662412

[13] M. Piteira and C. J. Costa. Computer programming and novice programmers. ISDOC '12: Proceedings of the Workshop on Information Systems and Design of Communication, June 2012, pp 51–53, https://doi.org/10.1145/2311917.2311927.

[14] The National Institute for Education, Informatika, Bratislava, Slovakia, 2017, http://www.statpedu.sk/files/articles/dokumenty/inovovany-statny-vzdelavaci-program/informatika_g_4_5_r.pdf.

[15] J. M. Wing, Computational Thinking: What and Why? In: Center for Computational Thinking, Carnegie Mellon, Pittsburgh, United States, Carnegie Mellon University, 2010, https://www.cs.cmu.edu/~CompThink/papers/TheLinkWing.pdf.

[16] Barefoot project, Computational Thinking Concepts and Approaches, Barefoot Computing, 2019, https://www.barefootcomputing.org/concept-approaches/computational-thinking-concepts-and-approaches.

[17] R. W. Bybee, J. Taylor, A. Gardner, P. V. Scotter, J. Carlson, A. L. Westbrook and N. Landes, The BSCS 5E Instructional Model: Origins, Efectiveness, and Applications, BSCS Science Learning, 2006, https://www.researchgate.net/publication/242363914_The_BSCS_5E_Instructional_Model_Origins_Effectiveness_and_Applications.

[18] Z. Tkáčová, Ľ. Šnajder and J. Guniš, Inquiry-Based Learning in Computer Science Classroom, In: S. N. Pozdniakov and V. Dagiené (Eds.): ISSEP 2019, LNCS 11913, pp. 68–79, 2019, https://doi.org/10.1007/978-3-030-33759-9_6.

[19] Slovak centre of scientic and technical information, National project IT Academy – Education for the 21st Century. Bratislava, Slovakia, 2016, http://itakademia.sk/sk/domov/.

Sentiment analysis of open-ended student feedback

Timo Hynninen*, Antti Knutas** and Maija Hujala***
*South-Eastern Finland University of Applied Sciences / Department of Information Technology, Mikkeli, Finland
**LUT University / LUT School of Engineering Science, Lappeenranta, Finland
***LUT University / LUT School of Business and Management, Lappeenranta, Finland
timo.hynninen@xamk.fi, antti.knutas@lut.fi, maija.hujala@lut.fi

***Abstract*—In this paper, we perform a sentiment analysis on a large set of open-ended course feedback from university courses collected between 2016 and 2019. We used the R programming language and environment for statistical computing to categorize feedback texts by their sentiment values (positive, negative). Additionally, we calculate the NRC Emotion values, which categorise the feedback according to eight basic emotions. We present analysis on the trends of how the feedback evolved through the years. Finally we compare the findings from our data to existing literature.**

I. Introduction

Student evaluations of teaching (SET) are commonly used for quality control in higher education. Student evaluations from university courses are used to assess both teaching material and teachers themselves [1], [2]. While the bulk of evaluations usually use quantitative, Likert-type scales, open-ended feedback texts are also common. The use of the text feedback for quality control, however, is less common.

On the other hand, student evaluations are a difficult tool to utilize. For example, the validity of evaluations as a measurement of quality can in some settings be questionable [3], and evaluations do not necessarily reflect about students' learning [4]. In addition to the validity questions of the student evaluations, text feedback is challenging to analyze in large quantities.

This paper presents a study on the sentiment analysis of open-ended student evaluations from university courses. This work is a continuation to our previous study [5] in which we distinguished different themes from open-ended feedback texts using the topic modelling method. In our previous study we established that the feedback students give concerns comments about course arrangements, study motivation, course content, and dissatisfaction in the teaching methods. We found that these topics were in line with types of student feedback established in prior work (see for example [6] or [7]). However, one of the conclusions in our previous study was that in our topic modeling results we could not see a clear disposition towards either positive or negative feedback. This result is somewhat surprising, as previous work maintains that students' comments tend to be more positive than negative. Therefore, further investigation of our feedback data is necessary.

The main research question this paper addresses is, **What emotions can be distinguished from student feedback**? The main research question is further divided into sub-questions, which are listed as follows.

- How does the machine learning algorithm classify the basic emotions in student course feedback texts?
- Is the ratio between positive and negative feedback comments the same as established in SET literature?

Rest of the paper is structured as follows. Section II presents the related work on using open text student feedback and sentiment analysis. Section III presents the research methods, and the procedures for data collection and analysis. The main results are presented in Section IV, and further discussed in Section V. Finally, Section VI concludes.

II. Related Work

In recent years, there has been a growing interest in the analysis and characterization of responses to the open-ended questions of student feedback surveys. A few studies have proposed sentiment analysis approaches for classifying students' written comments as positive, (neutral) or negative [8], [9], [10], [11], [12]. Onan [10] aimed at finding an efficient sentiment classification scheme on feedback comments provided by students. He tested and compared several machine learning-based approaches and deep learning-based approaches to sentiment analysis and found out that the deep learning-based methods outperformed the traditional machine learning classifiers. In turn, Pong-inwong and Songpan [11] proposed a new sentiment analysis method (sentiment phrase pattern matching, SPPM) and tested it with open-ended student feedback. In addition to students' sentiment polarity, Jena [13] and Nimala and Jebukumar [14] used sentiment mining techniques to model and predict students' emotions based on open-ended student feedback. Jena [13] tested various emotion classifiers to identify eight emotions (amused, anxiety, bored, confused, engaged, enthused, excited, and frustrated) from students' written comments whereas Nimala and Jebukumar [14] proposed a sentiment topic emotion mining model which captures eight basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust). In addition to simply classifying students' written comments, Andersson et al. [8] compared

average hours of study outside of class with the average sentiment of student feedback data and found a moderate negative correlation between them. They concluded that the use of sentiment analysis for student feedback data could allow for the simplification of the feedback questionnaires. Furthermore, Ibrahim et al. [15] focused on distinguishing feedback related to specific issues, such as assessment, from student's comments. Cunningham-Nelson et al. [16], [17] and Pyasi et al. [18], for their part, proposed tools for visually exploring students' sentiment for analysing written student feedback.

III. Methods

In this section, we first detail the data collection and then data analysis process.

A. Data collection

The data used in this study comes from the student feedback surveys carried out at a Finnish university between three academic years, beginning in the autumn 2016 and ending in the spring 2019. The questionnaire in 2016-17 had one open-ended question: *"Other feedback about the course (for example, ways to enhance learning during the course)"*. The questionnaire in 2017-18 had five open-ended questions: *"What factors affected my level of motivation?"*, *"What factors affected how much I invested in my learning?"*, *"What factors affected the workload?"*, *"My feedback regarding the teaching methods:"*, and *"What factors promoted my learning and how could learning be supported better?"*. In turn, the questionnaire in 2018-19 had four open-ended questions: *"The course as a whole promoted my learning (1=very little; 5=very much) - If you chose 1-3, please give concrete examples"*, *"The course as a whole promoted my learning (1=very little; 5=very much) - If you chose 4-5, please give concrete examples"*, *"What aspects of the course most need improvement?"*, and *"What was best about this course?"*. In addition, all questionnaires included several 5-point Likert-scale questions about, for example, motivation, workload, and teaching methods.

The survey questionnaires were sent to students via email after they completed the courses. Responses were collected anonymously and voluntarily. This study is restricted to feedback written in English, so we included only those responses that contained answers to open-ended questions written in English. We used the langdetect library [19] with the Python programming language to detect the language, and verified the results by manual inspection.

TABLE I
Percentage of positive and negative comments each year

Year	positive	negative
2016-2017	81%	19%
2017-2018	83%	17%
2018-2019	84%	16%
All 3 years	83%	17%

The sample size for sentiment analysis was in total 4990 feedback responses. 1534 were collected in 2016-17, 1953 in 2017-18, and 1503 in 2018-19.

B. Sentiment analysis process

> "Sentiment analysis is defined as the task of finding the opinions of authors about specific entities." [20]

In the analysis process, we used the sentiment approach to detect emotions in text. We used the NRC Word-Emotion Association Lexicon [21], which is a list of words and their associations with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The lexicon was manually created using a crowdsourcing platform [22].

We performed the analysis using the R statistical computing environment [23] and its libraries, especially in sentiment analysis [24], text mining [25] and cleaning [26], and visualization [27].

The main part of the analysis was performed using the Syuzhet sentiment analysis library [24] and its included NRC lexicon [21]. The Syuzhet library uses a bag-of-words algorithm, where it compares words from the analyzed text string against the selected lexicon and provides sentiment or emotion data. The exact analysis steps are summarized as follows.

1) Import the data:
   i) Download student feedback data.
   ii) Sort the feedback by metadata such as course type and language.
   iii) Select the subset to be analyzed; in this case English language responses and specifc range of years.
2) Preprocessing:
   i) Divide each feedback response into an individual string in an array.
   ii) Remove punctuation, extra whitespaces, and other symbols not relevant to analysis.
   iii) Remove stopwords using the textmining [25] library's English stopwords list.
3) Sentiment analysis: Use the Syuzhet [24] library and the NRC lexicon [21] to calculate emotion values for each response.
4) Analyzing the findings: Export the results as a table and perform descriptive or statistical analysis.

We used the wordcloud [27] library to create a comparative wordcloud that displays the most characteristic words from each emotion and demonstrates the emotions' comparative prevalence in the dataset. The wordcloud was generated by first assigning each feedback response to their strongest detected emotion and then creating a term-document matrix based on emotions and the associated responses. This means that words co-occurring with the lexicon's emotion words are also included in the emotion's wordcloud segment.

TABLE II
PERCENTAGE OF EACH NRC EMOTION IN THE FEEDBACK TEXTS

Year	anger	anticipation	disgust	fear	joy	sadness	surprise	trust
2016-2017	3.4 %	21.7 %	3.6 %	7.4 %	15.1 %	6.8 %	9.3 %	32.6 %
2017-2018	2.9 %	21.0 %	2.8 %	5.9 %	17.8 %	5.5 %	11.3 %	32.7 %
2018-2019	2.5 %	21.7 %	2.6 %	7.0 %	15.7 %	6.1 %	8.8 %	35.7 %
All 3 years	2.9 %	21.6 %	3.0 %	6.9 %	15.9 %	6.2 %	9.5 %	33.9 %

TABLE III
MOST COMMON WORDS ASSOCIATED WITH EACH EMOTION (IN DESCENDING ORDER)

anger	anticipation	disgust	fear	joy	sadness	surprise	trust
students	lectures	lectures	lectures	lectures	lectures	lectures	lectures
lectures	time	students	work	work	work	work	work
time	work	time	students	students	case	students	students
work	students	work	case	time	students	time	time
exam	learning	group	time	learning	time	exam	learning
group	exam	lecture	exam	interesting	group	learning	interesting

IV. RESULTS

After collecting and preprocessing the feedback surveys given to students we were left with 4990 individual open feedback texts. After a manual inspection to ensure the preprocessing had properly cleaned up and formatted the data we proceeded with the sentiment analysis. First, we calculated the overall emotion values for each feedback text. The sentiment analysis yields a sentiment score which determines if the overall emotion in the text is negative (a negative score) or positive (a positive score).

In the analysis, we found an overwhelming number of positive feedback texts. Overall 83% of total feedback was positive and 17% was negative. The number of overall positive and negative comments is presented by year in Table I. The difference between the years is minimal, although the number of positive comments is slowly on the rise (81% in 2016, 83% in 2017, and 84% in 2018), and conversely the number of negative comments is slowly diminishing.

In order to distinguish more detailed emotions from the texts, we also performed an NRC emotion analysis. The analysis uses the NRC lexicon and classifies the text in one or more of the NRC categories (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) by calculating a sentiment score for the text in each category. Table II presents the percentages of each NRC emotion category distinguished by the analysis. The most prevalent emotions emerging from the feedbacks are *trust* (overall 33.9%), *anticipation* (21.6%), and *joy* (15.9%). Less prevalent emotions were *surprise* (9.5%), *fear* (6.9%), *sadness* (6.2%), *disgust* (3.0%), and *anger* (2.9%). Each emotion is present in less than 10% of the texts.

Most common words in the feedbacks related to the emotion categories are listed in Table III. As shown in the table, certain words, for example *lectures* or *time* appear in multiple emotion categories. In order to see how closely associated to the emotions the individual words are, we employed the word cloud approach, which is visualized in Figure 1. Of the most occurring concepts, other students are associated with anger, exercises and time with anticipation, homework with fear, learning with joy, cases with sadness, exams with surprise, and teaching and lectures with trust.

V. DISCUSSION

By using the Syuzhet [24] library in the R statistical computing environment [23] we were able to extract two dimensions of emotion in our open-text student evaluations of teaching. The overall dominant sentiment value is largely positive, and the NRC emotion analysis provided a breakdown of what those positive emotions likely are: The most common emotions were classified as trust, anticipation and joy.

Our results indicate that, overall, the emotions distinguished from the feedback data were generally mainly positive. During our three-year observation period more than 80% of the feedback was positive and less than was 20% negative. This is an important thing to consider, since usually when faculty evaluates student feedback of teaching, critical feedback is highlighted and negative points are evaluated. This finding is in line with the results of the study by Alhija and Fresko [6], which points out that student feedback is more often positive than negative. Also, according to the results of Sengkey et al. [12] and Cunningham-Nelson et al., [17] student feedback seems to be biased towards positive evaluations of teaching.

What comes to the NRC emotion categories, our results show that many common learning and university related words are the target of many different emotions. Lectures, work, time, and other students are the target of both positive (e.g. joy, anticipation and trust) and negative (e.g. fear, sadness and disgust) emotions. In our previous study on the use of topic modelling for analysing student feedback texts [5], we found that while topic modeling can be used to extract themes from student feedback, its results do not tell much about the overall level of student satis-



Fig. 1. A comparison wordcloud with most characteristic words per emotion and their relative prevalence

faction (or dissatisfaction). In contrast, sentiment analysis of the student feedback can be used as an indicator for student satisfaction, but it does not provide actionable feedback themes like other computational approaches, such as topic modelling. Both methods however seem to provide useful indicators of the quality of teaching in higher education, but to get the best overall understanding these approaches should be used in parallel.

VI. Conclusion

The objective of this study was to assess the use of sentiment analysis with student evaluations of teaching. We collected 4990 student feedback texts over the period of three academic years. The sample was limited to responses in English only, even though at the same time we did collect feedback in Finnish as well. We then processed the feedback texts using the R programming language and environment for statistical computing, and used the Syuzhet Package to determine the sentiment values (positive or negative) and NRC emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) for each feedback text.

The text mining methods, such as sentiment analysis and topic modeling, can be very useful in the context of processing thousands of feedback text. The automation of this process, of course, comes with certain limitations. For example, without manual reading and systematic analysis of all the collected data it is not possible to make strong conclusions of the data. Therefore, the results should be seen as exploratory and not confirmatory. Manual inspection in these high numbers is, in practice, infeasible. In future work, we should develop a mixed methods approach to combine the automatic analysis (text mining, sentiment analysis, topic modelling) with a lightweight manual inspection of the data using a systematic analysis method (such as thematic analysis).

References

[1] M. Shevlin, P. Banyard, M. Davies, and M. Griffiths, "The validity of student evaluation of teaching in higher education: love me, love my lectures?" *Assessment & Evaluation in Higher Education*, vol. 25, no. 4, pp. 397–405, 2000.

[2] F. Zabaleta, "The use and misuse of student evaluations of teaching," *Teaching in Higher Education*, vol. 12, no. 1, pp. 55–76, 2007.

[3] P. Spooren, B. Brockx, and D. Mortelmans, "On the validity of student evaluation of teaching: The state of the art," *Review of Educational Research*, vol. 83, no. 4, pp. 598–642, 2013.

[4] B. Uttl, C. A. White, and D. W. Gonzalez, "Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related," *Studies in Educational Evaluation*, vol. 54, pp. 22–42, 2017.

[5] T. Hynninen, A. Knutas, M. Hujala, and H. Arminen, "Distinguishing the themes emerging from masses of open student feedback," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2019, pp. 557–561.

[6] F. N.-A. Alhija and B. Fresko, "Student evaluation of instruction: what can be learned from students' written comments?" *Studies in Educational Evaluation*, vol. 35, no. 1, pp. 37–44, 2009.

[7] B. Brockx, K. Van Roy, and D. Mortelmans, "The student as a commentator: students' comments in student evaluations of teaching," in *Procedia: social and behavioral sciences*, 2012, vol. 69, pp. 1122–1133.

[8] E. Andersson, C. Dryden, and C. Variawa, "Methods of applying machine learning to student feedback through clustering and sentiment analysis," *Proceedings of the Canadian Engineering Education Association (CEEA)*, 2018.
[9] S. Gottipati, V. Shankararaman, and J. R. Lin, "Latent dirichlet allocation for textual student feedback analysis," in *Proceedings of the 26th International Conference on Computers in Education ICCE 2018*. APSCE, 2018.
[10] A. Onan, "Mining opinions from instructor evaluation reviews: A deep learning approach," *Computer Applications in Engineering Education*, 2019.
[11] C. Pong-Inwong and W. Songpan, "Sentiment analysis in teaching evaluations using sentiment phrase pattern matching (sppm) based on association mining," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 8, pp. 2177–2186, 2019.
[12] D. F. Sengkey, A. Jacobus, and F. J. Manoppo, "Implementing support vector machine sentiment analysis to students' opinion toward lecturer in an indonesian public university," *Journal of Sustainable Engineering: Proceedings Series*, vol. 1, no. 2, pp. 194–198, 2019.
[13] R. Jena, "Sentiment mining in a collaborative learning environment: capitalising on big data," *Behaviour & Information Technology*, pp. 1–16, 2019.
[14] K. Nimala and R. Jebakumar, "Sentiment topic emotion model on students feedback for educational benefits and practices," *Behaviour & Information Technology*, pp. 1–9, 2019.
[15] Z. M. Ibrahim, M. Bader-El-Den, and M. Cocea, "Mining unit feedback to explore students' learning experiences," in *UK Workshop on Computational Intelligence*. Springer, 2018, pp. 339–350.
[16] S. Cunningham-Nelson, M. Baktashmotlagh, W. Boles *et al.*, "Visually exploring sentiment and keywords for analysing student satisfaction data," in *29th Australasian Association for Engineering Education Conference 2018 (AAEE 2018)*. Engineers Australia, 2018, p. 132.
[17] S. Cunningham-Nelson, M. Baktashmotlagh, and W. Boles, "Visualizing student opinion through text analysis," *IEEE Transactions on Education*, vol. 62, no. 4, pp. 305–311, 2019.
[18] S. Pyasi, S. Gottipati, and V. Shankararaman, "Sufat-an analytics tool for gaining insights from student feedback comments," in *2018 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2018, pp. 1–9.
[19] *langdetect 1.0.7: Language detection library ported from Google's language-detection*, 2016. [Online]. Available: https://pypi.org/project/langdetect/
[20] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.
[21] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.
[22] ——, "Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon," in *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*. Association for Computational Linguistics, 2010, pp. 26–34.
[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/
[24] M. L. Jockers, *Syuzhet: Extract Sentiment and Plot Arcs from Text*, 2015. [Online]. Available: https://github.com/mjockers/syuzhet
[25] I. Feinerer, K. Hornik, and D. Meyer, "Text mining infrastructure in r," *Journal of Statistical Software*, vol. 25, no. 5, pp. 1–54, March 2008. [Online]. Available: http://www.jstatsoft.org/v25/i05/
[26] T. W. Rinker, *textclean: Text Cleaning Tools*, Buffalo, New York, 2018, version 0.9.3. [Online]. Available: https://github.com/trinker/textclean
[27] I. Fellows, *wordcloud: Word Clouds*, 2018, r package version 2.6. [Online]. Available: https://CRAN.R-project.org/package=wordcloud

Comparative Analysis of Students' Attitudes on Teaching Quality and Its Assessment in Higher Education

K. Pavlina*, A. Pongrac Pavlina* and V. Juričić*
* University of Zagreb, Zagreb, Croatia
kpavlina@ffzg.hr, apongrac@ffzg.hr, vjuricic@ffzg.hr

Abstract - **This paper presents the results of a study of students' attitudes about characteristics of teaching quality in higher education. The research was conducted using a survey method on graduate students who were asked to rank according to the importance certain aspects of quality that are checked during the teaching quality assurance procedures conducted by the University of Zagreb. Teachers' qualifications, teaching competences and personal qualities were listed as teaching qualities, while course content and course organizational characteristics were used for course assessment. Also students were asked to evaluate their own capability to participate in teaching quality assessment and that was ranked in section teaching quality assessment management. The paper will present a comparative analysis of collected results with the results of a survey conducted in year 2010. The paper will also present students' attitudes about procedure of conducting the teaching quality assessment, with special emphasis on the application of information and communication technologies for implementation of teaching quality assessment and the dissemination of the results.**

Keywords - teaching quality, higher education, ICT

I. INTRODUCTION

Universities around the world use teaching quality survey results as a benchmark for measuring their progress in strengthening teaching excellence and providing high-quality learning experiences for their students. Surveys are most used method for quality control, because they are time efficient means of collecting results, detect trends and monitor and improve current methods for effective teaching. Universities have also started exploring the use of teaching performance indicators for performance□based funding and for benchmarking purposes to evaluate their current performance, so it is important to explore attitudes about the quality of teaching of different participants in the educational process.

The issue of teaching quality is the subject of numerous researches. Research conducted by Malm [14] shows that a teacher needs to have professional knowledge and skills, but also values and norms, in order to perform their work well. Roberts and Dyer [17] are on the same track dividing these three framework categories in more detail into 33 teacher characteristics.

Blaškova, Blaško and Kucharčikova [15] present their model based on eight compulsory competencies *(moral and ethical; technological; scientific; teaching; recognized author; role of the model; mature personality; critical thinking; communication competencies and motivational competencies*) that a modern university professor needs to possess in order to be able to maintain quality of university teaching.

Young [16] considers that a teacher is effective in his work if he possesses pedagogical knowledge; he is able to prepare and plan his teaching, uses different teaching methods during teaching, and regularly monitors and evaluates the student's work.

II. METHODOLOGY

Main objective of this study was to investigate students' attitudes toward certain teacher and course characteristics which are measured in teaching assessment surveys conducted by the University of Zagreb. Secondary objective of this research was to explore students' attitudes about practical implementation of quality control procedures.

Third objective of this research was to explore students' attitudes about teaching quality assessment management process.

Students were asked to rank by importance teacher and course characteristics which influence quality of teaching. Survey was conducted by on-line questionnaire on 142 graduate students from Faculty of humanities and Social sciences in January 2020.

Since this is a longitudinal research, the results obtained in 2010 on a sample of 104 graduate students will be presented comparatively.

III. FINDINGS

A. General questions

First question asked students to express their competence for evaluation of quality of teaching on scale from one, which represents complete incompetence, to five, which represents complete competence.

TABLE I. STUDENT COMPETENCE FOR EVALUATION OF QUALITY OF TEACHING

	Description	Answers	%
1.	Completely incompetent	0	0%
2.	Relatively incompetent	4	3%
3.	Moderately competent	30	21%
4.	Highly competent	76	53%
5.	Completely competent	32	23%

Average result was 3,96 which present that students think that they are very competent to evaluate quality of teaching.

Following question asked students to estimate in what extent will their evaluation change teaching practice. They could answer on scale from 1, which represented no change at all, to 5, which represented complete change.

TABLE II. STUDENT ESTIMATE IN WHAT EXTENT THEIR EVALUATION CHANGES TEACHING PRACTICE

	Description	Answers	%
1.	No change	20	14%
2.	Minor changes	48	34%
3.	Moderate changes	46	32%
4.	Major changes	22	16%
5.	Complete change	6	4%

With average answer 2.62 it is obvious that students don't perceive that teachers change their teaching practice according to the results from quality control surveys. It is interesting to note that the average answer to this question in the research [5] conducted in 2010 was 2.15, so it can be concluded that there was an improvement in the students' perception of implementation of the improvement based on the quality of teaching survey.

Studies from Harvey [3] and Colford [2] showed that this attitude can be result of lack of feedback to students after implementation of improvements in teaching practice.

Questions that examined should good teachers be awarded, and should bad teachers be sanctioned, showed that more students think that good teachers should be awarded (96%) than that bad teachers should be sanctioned (80%). Comparing to the results from research conducted in 2010 these results represent significant difference because in 2010 89% of students would penalize bad teachers and 87% would award good teachers. This demonstrate that student perception about this matter have changed in ten years and change has resulted in lowered tendency to penalize bad teachers and increase in tendency to award good teachers.

B. Teacher assessment

Teacher quality characteristics assessment was divided into three categories: teacher expertise, teaching competence and personal qualities.

Teacher expertise characteristics represent teacher's expertise in his scientific field.

TABLE III. TEACHER EXPERTISE CHARACTERISTICS

Rank	Characteristic	Average rank 2010	Average rank 2020
1.	Gives good examples and exercises	1,7 (1)	1,63
2.	In teaching shows good knowledge of subject matter	1,9 (2)	1,83
3.	Answers question expertly	2,4 (3)	2,53

Students in this research ranked almost identically teacher expertise characteristics as students in research conducted in the year 2010.

Next set of teacher competence characteristics represents teacher's ability to create good working atmosphere and teach.

TABLE IV. TEACHING COMPETENCE CHARACTERISTICS

Rank	Characteristic	Average rank 2010	Average rank 2020
1.	Teaches course content clearly and self-explanatory	2,5 (1)	2,56
2.	Have good communication skills and creates pleasant working atmosphere	3,1 (3) ↑	3,25
3.	Clearly defines course goals and student duties	3,0 (2) ↓	3,32
4.	Lectures are interesting and dynamic	3,7 (5) ↑	3,66
5.	Lectures have good structure and time is rationally used	3,7 (4) ↓	3,84
6.	Raise the quality of teaching by use of modern technology	4,7 (6)	4,35

The most important teaching competence that teacher should have is to teach course content clearly and self-explanatory. There were slight changes in the ranking of the second and third characteristic. Such results may suggest changing student attitudes about importance of characteristic "Lectures have a good structure and time is rationally used" which rank have lowered.

Least important characteristic in both researches was that teacher raise quality of teaching by use of modern technology.

Last category of teacher's characteristics represents his personal qualities which influence quality of teaching.

Several studies show that teachers are still not inclined to overuse modern technology for the purpose of maintaining teaching at the higher education level. In their work, Manca and Ranieri [10] have shown that teachers limited use of social media in their teaching for pedagogical, institutional and cultural reasons. The use of modern technology for teaching purposes is also related to the teacher's digital competencies. Amhag, Stigmar and Hellstrom [11] showed that 54.8% of higher education teachers rated their digital competencies at a satisfactory average level. Also, 50.0% of teachers stated that they need further training in their digital skills. In addition, 55.4% of teachers attended professional development related to the digitalization of the teaching process. It is evident from the above that higher education teachers are aware of the growing trend of penetration of modern technologies into the teaching process and that they are

professionally improving themselves in order to modernize their teaching.

TABLE V. PERSONAL QUALITIES CHARACTERISTICS

Rank	Characteristic	Average rank 2010	Average rank 2020
1.	Regards students properly and respectfully	1,7 (1)	1,82
2.	Motivated for work and consciously fulfils his obligations	2,3 (2)	1,99
3.	Available for consultations	2,7 (3)	2,77
4.	Held lectures regularly and on time	3,4 (4)	3,47

Rank of teacher's personal quality characteristics remains the same as in research conducted in 2010 [5]. Students still think that most important personal quality is that teacher regards them properly and respectfully and least important personal quality is that teacher held lectures regularly and on time.

C. Course assessment

Assessment of course quality characteristics are divided into two categories: course content and course organization.

TABLE VI. COURSE CONTENT CHARACTERISTICS

Rank	Characteristic	Average rank 2010	Average rank 2020
1.	Course showed the importance of subject and application possibilities	2,0 (2) ↑	2,00
2.	Course gives insight in key problems and enables its understanding	1,9 (1) ↓	2,23
3.	Quantity of course contents match number of course lecture hours	2,7 (3)	2,65
4.	Course literature is useful for understanding subject matter	3,2 (4)	3,13

Results from 2020 in comparison with results from 2010 may suggest that there is change in perception of most important course content characteristic: this research revealed that students emphasize importance of application possibilities of newly acquired knowledge, while students in research [5] considered that most important characteristic is that course gives insight in key problems and enables its understanding. This could be explained by continuous tendency on strategic level to emphasize importance of application possibilities which have led students to express that this is most important course content characteristic.

TABLE VII. COURSE ORGANIZATION CHARACTERISTICS

Rank	Characteristic	Average rank 2010	Average rank 2020
1.	Course goals and student duties are clearly defined	1,7 (1)	1,87
2.	Lectures where properly followed by exercises and seminars	2,0 (2)	1,97
3.	Course organization encourages student active participation	2,3 (3)	2,15

As in previous research, students emphasize that course should have clearly defined goals and student duties. This can be attributed to the fact that students are continuously evaluated during semester through series of projects and exercises on various courses so they prefer to be informed about their duties at the beginning of a semester so that they could plan their activities during semester. Last ranked course organization characteristic is that course encourages student active participation, that student perception can be attributed to fact that students have high workload during semester on various activities, so they prefer courses with lower demands.

Brint, Cantwell, and Hanneman [12] identified student's active participation in course. This research provide an overview of the student's academic engagement that is also related to the organization of the course. One of the characteristic is student-teacher interaction which encompasses the professional relationship that is realized between the teacher and the student in such a way that the teacher has a model that takes care of his students by providing them timely guidance and feedback. In addition, the level of academic response is very important, emphasizing the importance of high expectations, which is visible through the complexity of the course, such as seeking analysis and synthesis of reading material, critical thinking or applying theories in practice.

Research provided by Beran and Violato [13] shows that if the course is organized in such a way that discussions are expected and the application of learned work on examples, students will acquire higher levels of knowledge, and also they will have higher expectations in terms of assessment.

D. Teaching quality assessment management

Last set of questions explored students' attitudes about teaching quality management procedures and techniques.

First question asked students their attitudes whether all students should be required to complete a teaching quality survey. Two-thirds of students (66%) answered that it should be required and 34% think that it shouldn't be required.

Next question asked students whether they support the introduction of a compulsory quality survey of all courses taken during semester/year via student administration system. Majority of students (76%) support this mechanism of compulsory feedback on teaching quality.

Several studies ([2],[3],[6]) have highlighted the importance of providing feedback to students on implemented quality improvement procedures, so one question explored students' opinion whether the results of the quality of teaching surveys should be made public. Majority of students (61%) think that results should be published. Students were asked in another question do they think that improvement of quality of teaching could be influenced by the publication of the results of teaching quality survey and 79% answered positively. It would be interesting to further explore this discrepancy in results where 18% of students doesn't support publication of

results although they think that it could be beneficial for improvement.

Overwhelming majority (93%) of students think that they should be informed of the steps taken to improve the quality of teaching. This confirm results from several researches ([1],[4],[7]) which showed that good communication strategy represents crucial factor in establishment of good teaching quality management system. Good teaching quality management systems envision communication strategy which presents results of actions taken to improve quality of teaching to students and teachers by which they are motivated to participate in future quality management cycle.

Students were also asked to express what would be their preferred way of completion of teaching quality survey. For majority of them preferred way is on-line form (59%), followed by "on-line form on student administration system" with (28%) and only 13% of students preferred printed forms.

IV. CONCLUSION

The results of the research show that students' attitudes towards the characteristics of teaching quality have not changed significantly within ten years, except for the increase in the importance of practical applicability of knowledge, which can be explained by the reform activities in the past ten years. A greater change in attitudes is evident around rewarding good teachers and penalizing bad teachers. In 2010 survey, an equal number advocated penalizing poor teachers (89%) and rewarding good teachers (87%), while in this study there was a decrease in the number of students who would penalize poor teachers (80%) and an increase in the number of students who would reward good teachers (96%).

Teaching quality assessment surveys remain an important tool for collecting data on student attitudes about the teaching process, but they only represent the beginning of establishment of a teaching quality culture. Most commonly teaching quality control surveys are implemented as summative at the end of semester or academic year, but it is interesting to note that research by Winchester and Winchester [8] recommends that the utilization of weekly formative quality control survey helps teachers to improve their teaching and increase their summative score towards the end of academic year.

For conclusion, it is important to emphasize how the data collected by teaching quality survey is used, because the collected data only allow detection of problems, but doesn't automatically instigate their resolution, so procedures aimed at improving quality are crucial to achieve higher quality of teaching.

REFERENCES

[1] J. Brennan, T. Shah, *Quality assessment and institutional change*, Higher Education , Vol. 40, No. 3, pp. 331-349, 2000.

[2] J. Colford, *Student evaluation of modules – A student perspective*, Quality in Business Education, 2005.

[3] L. Harvey, *Student Feedback*, Quality in Higher Education, Vol. 9, No. 1,, pp. 3-20, 2003.

[4] C. T. Husbands, P. Fosh, *Students evaluation of teaching in higher education: experiences from four European countries and some implications of the practice*, Assessment and Evaluation in Higher Education, Vol. 18, Issue 2, pp. 95-114, 1993.

[5] K. Pavlina, M. Banek Zorica, A. Pongrac, *Student perception of teaching quality in higher education*, Procedia – Social and Behavioral sciences, vol. 15, pp. 2288-2292, 2011.

[6] J. Leckey, N. Neill, *Quantifying Quality: the importance of student feedback*, Quality in Higher Education, Vol. 7, No. 1, pp. 19-32, 2001.

[7] H. W. Marsh, *Students' evaluations of university teaching: research findings, methodological issues and directions for future research*, International Journal of Educational Research, Vol. 11, No. 3, pp. 253-388, 1987.

[8] T. M. Winchester, M. K. Winchester, A longitudinal investigation of the impact of faculty reflective practices on students' evaluations of teaching. British Journal of Educational Technology, 45, Issue 1, 112–124, 2014.

[10] S. Manca, M. Ranieri, *Facebook and the others. Potentials and obstacles of social media for teaching in higher education,* Computers & Education, Vol. 95, 2016, 216-230.

[11] L. Amhag, M. Stigmar, L. Hellstrom, *Teacher educators' use of digital tools and needs for digital competence in higher education,* Journal of Digital Learning in Teacher Education, 2019, DOI: 10.1080/21532974.2019.1646169

[12] S. Brint, A.M. Cantwell & R.A. Hanneman (2008). The two cultures of undergraduate academic engagement. Research in Higher Education, 49, 383–402.

[13] T. Beran, C. Violato (2009). Student ratings of teaching effectiveness: Student engagement and course characteristics. Canadian Journal of higher education, Vol. 39, No. 1, pp. 1-13.

[14] B. Malm (2009). Towards a new professionalism: enhancing personal and professional development in teacher education, Journal of Education for Teaching, Vol. 35, No. 1, 77-91, DOI: 10.1080/02607470802587160

[15] M. Blaškova, R. Blaško, A. Kucharčikova, (2014). Competences and competence model of university teachers. Procedia – Social and Behavioral Sciences, 159, pp. 457-467.

[16] Young, M. (1990). Characteristics of high potential and at-risk teachers. Action in Teacher Education. 11(4), 35-39.

[17] Roberts, G., Dyer, J. E. (2004). Characteristics of effective agriculture teachers. Journal of Agricultural Education, Vol. 45, Number 4, pp. 82-95.

Students' attitudes toward value-driven digital marketing

Ana-Marija Jadanec* and Tihana Babić*
*Algebra University College, Zagreb, Croatia
anamarija.jadanec@yahoo.com, tihana.babic@algebra.hr

Abstract - **Value-driven marketing has a goal to create a deep and meaningful value, not only in customers' business but also in customers' lives by appealing to their personal values and ethics. With the rise of a new wave of technology that enables interaction between individuals and groups, customers have become well informed about brands and their products, as well as their social responsibility. The aim of the research conducted among Millennials and Generation Z students on the Algebra University College was to examine how unprecedented access to information, social media and the rise of Marketing 3.0 has changed their minds and expectations about brands and whether they choose companies that care about the same causes they care about. In this paper, their views on the value-based advertising and efficacy of digital marketing efforts will be presented.**

Keywords – ***value-driven marketing, digital marketing, brands, social responsibility, students***

I. INTRODUCTION

In today's society, large corporations play an important role in the global economy. Consumers are looking for experiences and business models that touch their spiritual side. Richard Barrett discovered that human spiritual motivation can be adapted to the missions, visions, and values of a company [1].

Human beings are not indifferent to the world. As groups, they develop concepts of preferred and obligatory conduct and desirable and undesirable states of affairs. Values serve as criteria for selection in action, for judgment, preferences, and choices with grounds for the decision in behavior. In other words, values are principles that help people decide what is right and what is wrong and how to act accordingly [2].

Defining value starts from the theory of need proposed by A. Maslow in his work The Theory of Human Motivation (1943). "Needs lower down in the hierarchy must be satisfied before individuals can attend to needs higher up. From the bottom of the hierarchy upwards, the needs are *physiological, safety, love and belonging, esteem,* and *self-actualization* [3]. Unless the basic needs that ensure survival and security are met, there are no conditions for the emergence of complex, multi-value needs.

Using Maslow's theory, Ronald Ingehart (1977) coined the terms materialism and postmaterialism to explain how economic growth changes social values. Materialist values are based on economic and physical security needs that older generations lacked growing up during World War II and other crises. On the other hand, post-materialists have grown up in greater security, and their values are oriented toward social connection, self-esteem, and self-actualization. They attach more value to intangible goals such as *ideas, autonomy, freedom of speech, community, ecology,* and *gender equality* [2].

II. SOCIAL RESPONSIBILITY

If for an individual, values represent what is important to him in life, what motivates him and what sets his behavioral goals. For businesses, that isn't much different. Values represent what companies believe in, what drives and direct them in business. For that reason, companies need to identify, set, and communicate their core values and work by following them. It is a significant factor in determining the loyalty of employees and other stakeholders and as a basis for ethical and socially responsible business operations.

The definition of social responsibility demonstrates that individuals and companies have to be conscious of the impact they are having on the economic, social, and environmental aspects of society. Companies that base their organizational culture on corporate social responsibility and foster accountability to themselves their stakeholders and the public in return have a better image of themselves and attract more clients and investors [4].

In her article, K. Leonard recognizes "The four types of Corporate Social Responsibility: Philanthropy, Environment Conservation, Diversity and Labor practices, and Volunteerism" [5].

"*Philanthropy* – refers to the effort an individual or organization makes based on an altruistic desire to improve human well-being." [6]. Subjects of these donations can be environment and animal protection, health research, education, and art conservation.

Environmental Conservation - means companies focus on efforts to reduce their impact on the environment conducting business practices like waste management, the use of clean energy sources like the sun and wind, or supporting long-term issues such as global climate change and reducing carbon emissions.

Diversity and Labor practices - are important because the responsibility for employees affects the value of the brand. Turoń, K. in her paper on Corporate Social Responsibility to Employees, according to Kisil (2013), lists thirteen key actions which are associated with good labor practices: "*compliance with the principles of diversity; providing voluntary work; fair relationships between co-workers; opportunity to express views and*

ratings by employees; ensuring safety in the workplace; providing employees with information about any activities that are associated with them; providing medical care; ensuring an appropriate structure for work and fair wages; ensuring satisfactory conditions of employment; encouraging the development of employees; respecting private and family life of employees; supporting employees in unexpected circumstances; implementation of outplacement programs" [7].

Volunteerism - means putting aside time to contribute to your local community.

III. VALUE DRIVEN MARKETING

It started with product-centered Marketing 1.0, which revolved around making low production cost products that were affordable to the mass market. By the words of H. Ford: *"Any customer can have a car painted any color that he wants so long as it is black."* [1]. Information age brought informed customers who had different preferences and needs that needed to be addressed. In costumer-centric Marketing 2.0, "the customer is a king," but still a passive observer of marketing campaigns.

"In Marketing 3.0, we observed customers transforming into whole human beings with minds, hearts, and spirits" [1]. The new technology that connected the world led to changes in marketing. Consumers started actively giving feedback to companies and participating in product development and communication. As developed societies have collectively raised to the top of Maslov's pyramid, increasing prosperity made post-material values and self-actualization their primary aspirations. Consumers started to have anxieties about the globalized world and started looking for solutions to make it a better place to live. Companies who want to create connections need to identify their anxieties and desires and commit to working for social, economic, and environmental causes in ways that would touch customers' minds, hearts, and spirits.

Being value-based means that social, emotional, psychological or business values play a vital role, and profitability is balanced with corporate responsibility. Marketing 3.0 complements emotional marketing with the marketing of the human spirit.[1].

A. Cause-related marketing

"Cause-marketing is defined as a type of Corporate Social Responsibility (CSR) in which a company's promotional campaign has the dual purpose of increasing profitability while bettering society" [8].

In their book on Corporate Social Responsibility, P. Kotler and N. Lee distinguished six major types of corporate social initiatives, including marketing-related ones: *cause promotions, cause-related marketing,* and *corporate social marketing;* and ones that are outside the typical functions of marketing departments: *socially responsible business practices* and *volunteering.*

Corporate Philanthropy - a practice of making a direct contribution to a charity or cause, most often in the form of donations, and/or services; *Community Volunteering* - supporting local community organizations and causes and encouraging employees and partners to volunteer, and *Socially Responsible Business Practices* - conducting business practices that support social causes to improve community well-being and protecting the environment. These activities are usually not visible enough to impact consumer preference at scale." [9].

The remaining types of corporate social initiatives are forms of *cause marketing*, strategic actions that support business goals like increasing market share or building brand identity.

Cause-promotions support a cause by increasing awareness. Many companies link to specific causes that appeal more to their customers or support issues related to their core products, such as "cruelty-free" cosmetics brand promoting a ban on animal testing.

Cause-related marketing occurs when a company's donations are based on product sales. It has a goal of influencing a behavior change by giving a percentage of revenues to a cause. This offer is usually for a particular product, over some time and the company is often associated with a non-profit.

Corporate Social Marketing is a campaign intended to improve public safety, health, or the environment and change behavior, especially ones that could lead to loss of revenue. Mobile phone companies advise not to text while driving or brands of alcohol beverages advocate to "drink responsibly" [9].

B. Digital marketing

In the transition from traditional to digital marketing, the human-centric approach is even more important. As our communications become more digital, marketing became more human.

With intensified conversations with peers, social circles of family and friends are becoming more important and reliable sources of advice than any kind of media. To attract customers and build a connection, brands are beginning to have human properties. Value-driven brands have one important human characteristic - morality, they place social responsibility as their core differentiation and keep their promise, even if customers aren't keeping track. They come across as genuine and authentic, which are valuable assets in an increasingly transparent world. [10].

IV. MILLENNIALS AND GENERATION Z

Born between 1981 and 1996, Millenials are the last generation to grow up before the internet and technology and at the same time using all of its power. They joined the labor force in the height of the economic recession which dragged characteristics like freelancing work, longer stays in the parental home, postponing the marriage, and leaving the homeland for work [11].

Anyone born from 1997 onward is part of Generation Z [11]. This open-minded generation is known for celebrating differences and working for the future of equality regardless of race, gender, or sexual orientation. Nearly one-third of Generation Zs say people their age are more likely to think they are all equal, and 4 out of 10

believe their generation accepts differences in other people more than older generations. This is likely because many of the advances, such as gay marriage, from previous decades, seem obvious to them [12].

"The 2019 Millennial Impact Report, research on millennials and their interest in causes reveals that millennials engage with causes to help other people, not institutions; they support issues rather than organizations; they prefer to perform smaller actions before fully committing to a cause; are influenced by the decisions and behaviors of their peers; treat all their assets (time, money, network, etc.) as having equal value" [13].

"The 2019 Deloitte Global Millennial Survey that questioned Millennials and Generation Zs on their opinions about business' motivations and ethics" [13] showed they are most concerned about climate change/protecting the environment.

They put great importance on the societal impact and ethics of companies they support, and expect them to improve the livelihoods of their employees, society, and protect the environment. They started or deepened a relationship with a business because of the perception that the company's products or services have a positive impact on society and/or the environment, or believe a company was ethical. Despite a sharp decline in recent years, Millennials and Gen Zs believe that businesses in general, have a positive impact on society.

Perhaps the most worrying is respondents' opinions that traditional mass media negatively affects the world and is a slightly more reliable source of news than social media platforms. NGOs and nonprofits are ranked highest in terms of their impact on society and sources of reliable information even though it isn't their job to report the news [14].

V. THE RESEARCH METHODOLOGY

A. The Research Goals

The research aims to examine attitudes among Millennials and Generation Z students toward corporate social values and cause-related marketing:

1. *Which social causes matter to students in general?*
2. *Do students consider corporate social responsibility important?*
3. *Students' attitudes towards supporting causes in marketing communication?*

B. The Research Sample

The research was conducted among students on the Algebra University College.

The total number of population sample was 70 students (N = 70), on study programs: Software engineering (SE), Multimedia computing (MC), Digital marketing (DM), Visual communication design (VCD) and Design and communication management (DCM). The majority of students were from undergraduate study program Digital marketing; 28 of them (40 %), and from Visual communication, design 26 students (37.1 %). 10 students (14.3 %) were from Multimedia computing, 5 (7.1 %) from Design and communication management, and 1 (1.4 %) from Software engineering. The structure of all surveyed students according to the study program is presented in Table 1.

Study Program		SE	MC	DM	VCD	DCM
Total number of Students per Study Program (N)		1.42 %	14.2 %	40 %	37.1 %	7.14 %
Number of Students per Gender	Male	1.42 %	14.2 %	22.8 %	12.8 %	1.42 %
	Female	0 %	0 %	17.1 %	24.2 %	5.7 %
Number of Students per Study Level	Under-graduate	1.42 %	12.8 %	35.7 %	37.1 %	0 %
	Graduate	0 %	1.42 %	4.28 %	0 %	7.14 %

Table 1. The structure of the participants per Study Program in percentage, N = 70.

With 37 male students (52.8 %) and 33 females (47.1 %), the genders were divided into almost two equal groups.

Divided by age, 31 (44.2 %) students born until 1996. belong in the Millennials generation and 39 (55.7 %) of students born 1997. and later are Generation Z.

C. The Research Sample

The research was conducted through an anonymous voluntary survey during the winter semester of the 2019/2020 academic year.

A specially designed questionnaire in the Google Forms tool had 15 closed-ended questions, of which 3 were related to demographics. One of these questions included 11 statements for which respondents could indicate the degree of importance. For the rest, 8 questions answers were defined with the degrees of agreement, and two with degrees of importance according to the Likert scale.

At the beginning of the questionnaire, a descriptive definition of terms *Value-driven marketing* and *Corporate social responsibility* was determined with the aim of improving the understanding of the subject.

A quantitative method was used to analyze the research results. Data processing of survey results was performed via the Google Forms tool.

D. The Research Results

1. *Which social causes matter to students in general?*

The research results have shown that Millennials and Generation Z members (64.2 %) stated they value *labor rights, worker protection* the most, closely followed by *human rights*, and *gender equality* both chosen as the most important by 62.8 % of students. The next most

important values, chosen by half of the students (50 %) is *sustainable development (environmental protection)* and *active communication with consumers*, closely followed with *non-testing on animals* (48.5 %) and *tolerance* (47 % of students). *Supporting various non-profit organizations* and *different causes* matters to 38.4 % and *participating in the community* only to 30 % of surveyed students.

The structure that represents the students' degree of importance according to the Likert frequency scale is presented in Table 2. The statements represented by letters in the table are as following:

A. *Sustainable development and environmental protection.*
B. *Human rights.*
C. *Labor rights, worker protection.*
D. *Non-testing on animals.*
E. *Tolerance.*
F. *Gender equality.*
G. *Fairtrade (responsible relationship with suppliers).*
H. *Fighting poverty.*
I. *Participating in the community.*
J. *Active communication with consumers.*
K. *Supporting various non-profit organizations and different causes (fighting against various diseases, education, etc.)*

STATEMENT	Not important	Slightly important	Neutral	Important	Very important
A	1 %	1 %	14 %	34 %	50 %
B	1 %	1 %	7 %	27.1 %	62.8 %
C	3 %	0 %	7 %	24.2 %	64.2 %
D	3 %	4 %	13 %	40 %	48.5 %
E	1 %	1 %	13 %	37 %	47 %
F	3 %	2 %	9 %	24.2 %	62.8 %
G	1 %	1 %	11 %	35 %	50 %
H	1 %	6 %	19 %	31.4 %	42.8 %
I	0 %	3 %	26 %	41.4 %	30 %
J	1 %	1 %	13 %	35.7 %	48.5 %
K	3 %	3 %	21 %	34 %	38.5 %

Table 2. Distribution of the importance degree according to the Likert scale.

2. *Do students consider corporate social responsibility important?*

When asked about *corporate social responsibility (uniform consideration of the economic, social, and environmental impact),* 44.3 % of the total number of surveyed students considers it very important, 51.4 % consider important, with only 4 % neutral. On the importance of *contributing to the community in which, companies are present (supporting local charities, sponsorships, investing in development, etc.),* 32.9 % of students consider it very important, 58.6 % consider important, 7 % are neutral and 1 % consider it is only slightly important.

Distribution of the answers on the importance of corporate social responsibility and contribution to society is presented in Chart 1.

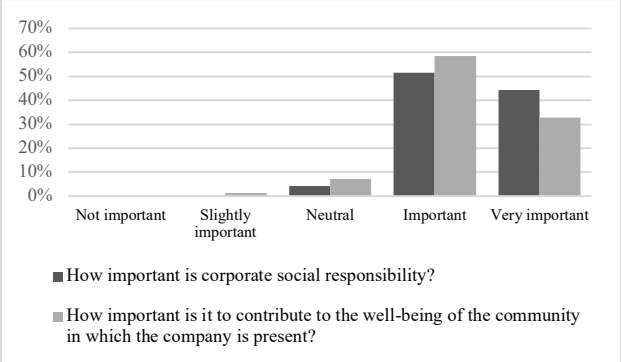


Chart 1. Distribution of the answers to the question: a. *How important is the corporate social responsibility of a company?* And b. *How important is to contribute to the community in which companies are present?*

To question does *corporate social responsibility* affect how they feel or behave toward a company, students were asked to grade how much being aware of one companies reputation affect their towards them, and do students attitudes differ for *small domestic companies* and *large corporations (Apple, Nike, McDonald's, Coca-Cola, etc.).*

a. Negative business practices

30 % of students are almost certain, and almost half (47.1 %) would change a relationship with a business (stop using company's products or services) if they found out about negative business practices *(worker exploitation, pollution, discrimination, etc.)* of *small domestic companies*. 19 % of students possibly would and 4 % are unlikely to be affected.

For *large corporations (Apple, Nike, McDonald's, Coca-Cola, etc.)* results are different.

Only 15.7 % of students are almost certain they would change their relationship with a business because of their negative business practices, which is 50% less than for local companies. 34.3 % would likely, 31 % possibly, 13 % unlikely, and 6 % rarely stop using the company's products or services.

Distribution of the answers about the effects of businesses' negative values on students' relationships with *small domestic companies* in comparison to *large corporations* is presented in Chart 2.

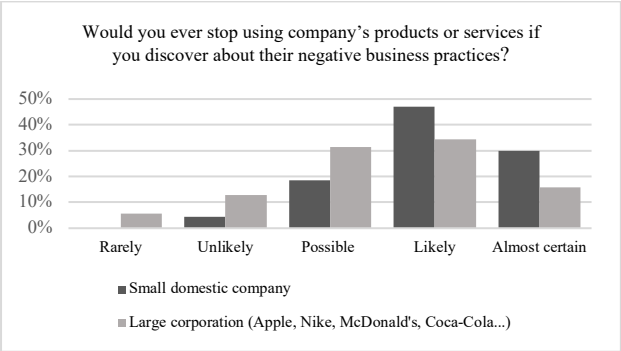


Chart 2. Distribution of the answers to the question: *As a consumer, would you stop using the company's products or services if you discover their negative business practices?* in a comparison between *small domestic companies and large corporations.*

b. Prominent positive values

On the other hand, if *a small domestic company* has prominent positive values *(corporate social responsibility, tolerance, protection of workers, sustainable development and environmental protection, fair trade, etc.)*, 25.7 % of participants would almost certainly change a relationship with a business (start using company's products or services), 42.9 % likely and 31 % possibly started using their products or services. 9 % of surveyed students would unlikely and 1 % would be rarely affected.

Large corporations (Apple, Nike, McDonald's, Coca-Cola, etc.) positive values would affect 27.1 % of students in a way that they would almost certainly start using the company's products or services, almost the same as for small domestic companies. 28.6 % of students would likely and 31 % possibly change their relationship with a business. 9 % is unlikely and 1 % would rarely be affected.

Distribution of the answers about the effects of businesses' positive and negative values on students' relationship with *local companies* in comparison to *large corporations* are presented in Chart 3.

In summary, the research results have shown that businesses' positive and negative values would more likely affect students' relationships with *small domestic companies* than *large corporations.*

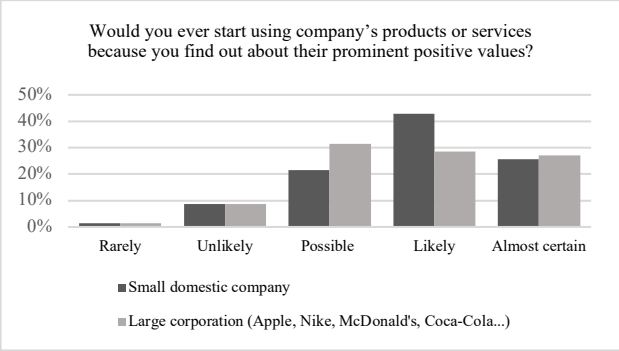


Chart 3. Distribution of the answers to the question: *As a consumer, would you start using the company's products or services because you find out about their prominent positive values?* in a comparison between *small domestic companies* and *large corporations.*

c. Socially responsible business practices

The research results have shown that 41.4 % of the total number of surveyed students strongly agree and 50 % agree that *incorporating socially responsible business practices* (*corporate social responsibility*, *fair trade, protection of workers, sustainable development, environmental protection, gender equality, tolerance, animal rights, etc.)* affects on company's popularity. Only 6 % neither agree neither disagree and 3 % disagree.

Overall, the majority of students agree that *corporate social responsibility* positively affects a company's image. Distribution of the answers if *incorporating socially responsible business practices affects a company's popularity* is presented in Chart 4.

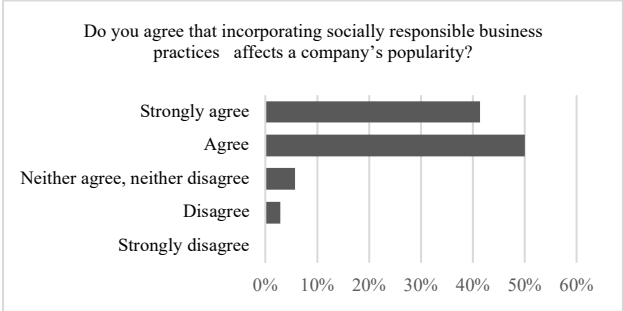


Chart 4. Distribution of the answers to the question: *Do you agree that value-driven marketing affects a company's popularity*?

3. *Students' attitudes towards supporting causes in marketing communication.*

a) *Cause-related marketing*

For *supporting causes* in marketing communication *(cause-related marketing),* 40 % of participants strongly agree and 47.1 % agree that it affects the company's popularity. 11 % of them neither agree neither disagree and 1 % strongly disagree with the statement.

The distribution of answers if *supporting causes effects* a company's popularity is presented in Chart 5.

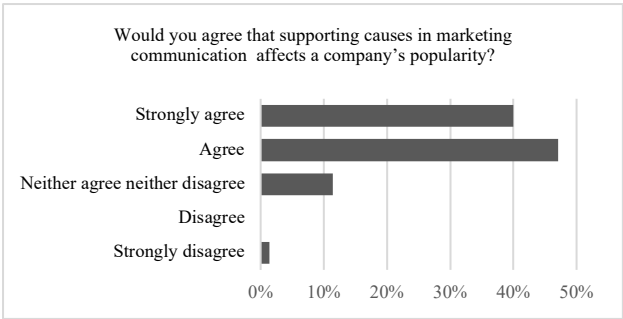


Chart 5. Distribution of the answers to the question: *Do you agree that supporting causes in marketing communication affects a company's popularity?*

b) *Supporting causes that students value*

42.9 % of the students strongly agree and 34.3 % agree they would rather use products and services from companies that in their marketing communication support causes they consider important. 23 % neither agree neither disagree and none of the students disagrees or strongly disagrees. The distribution of the answers is presented in Chart 6.

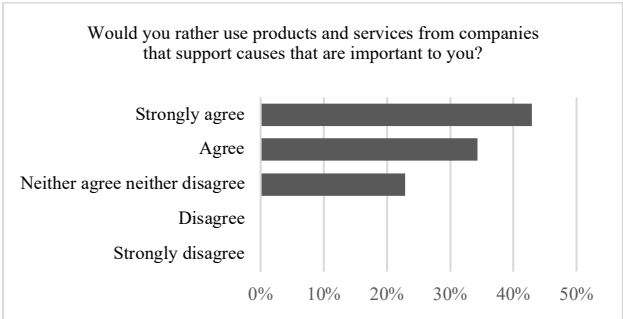


Chart 6. Distribution of the answers to the question: *Would you rather use products and services from companies that in their marketing communication support the same causes or represent values that are important to you?*

c) *Do students trust cause-related marketing?*

On the contrary, when asked if they trust that companies 17.1 % of students trust them very much, 37.1 % somewhat, and 36 % is undecided. 9 % not really with 1 % that does not trust at all. The distribution of the answers is presented in Chart 7.

In summary, a majority of students agree that cause-related marketing positively affects a company's image and would rather use products and services from companies that support causes they consider important. However, when asked about how much they trust that companies, a great part of students are undecided.

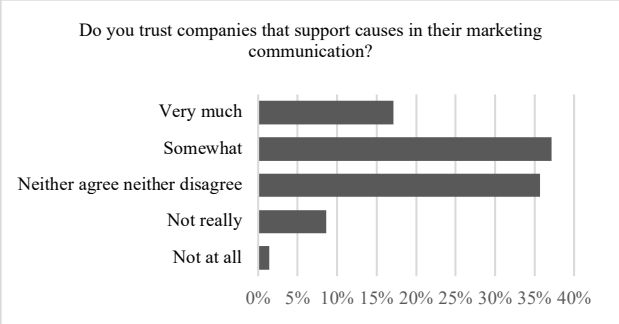


Chart 7. Distribution of the answers to the question: *Do you trust companies that support causes or values in their marketing communication (do you consider them sincere in their efforts)?*

E. *Limiting elements of the research and recommendations for further research*

The study was conducted on a small sample, it would be advisable to conduct it on a larger sample.

In future researches, the distinction and significant differences could be made between Millenials and Generation Z students. Also, it would be advisable to research which of Croatian brands or companies' social causes or platforms are Millennials and Generation Z familiar about.

VI. CONCLUSION

Millennials and Generation Z live in a society where businesses and brands have more power than ever and becoming responsible corporate citizens became important for a company's success. They expect companies to improve lives and provide livelihoods, but they don't see enough businesses standing up and filling the gap. They want to see meaningful actions and business leaders that serve as agents for positive change. Consumers expect companies to be *socially responsible* but they need to be convincing in their efforts to gain their trust. Although we see many examples of companies that put the values of good corporate citizenship into corporate statements of mission, vision, if those values aren't a major component of who they are as a company, the consumer will not trust them.

Millennials and Generation Z consumers have a better perception of a company doing something to make the world a better place. These generations are accustomed to advocating for the causes they are interested in through social media platforms. A value-based approach in marketing strategy, *cause-related marketing* can make a difference. Brands that successfully attach their core to some social cause, it stops being just something they support, it becomes who they are, and they become ambassadors. If brands want to appeal to these generations, they need to pay attention to their values.

REFERENCES

[1] P. Kotler, H. Kartajaya, I. Setiawan, "Marketing 3.0: From Products to Customers to Human spirit", John Wiley and Sons, Inc., Hoboken, New Jersey, 2010.

[2] M. Rokeach, "Understanding Human Values", The Free Press, New York, 1979.

[3] S. McLeod, "Maslow's Hierarchy of Needs", Simply Psychology, March 2020. Available at: https://www.simplypsychology.org/maslow.html/. [accessed: 20.03.2020].

[4] S. Vig, "Poslovna etika", CODUPO, Zagreb, 2019.

[5] L. Kimberlee. "Four Types of Corporate Social Responsibility." Small Business - Chron.com, http://smallbusiness.chron.com/four-types-corporate-social-responsibility-54662.html [accessed: 12.02.2020].

[6] A. Hayes, "Philanthropy Defined", Investopedia, January 2020. Available at: https://www.investopedia.com/terms/p/philanthropy.asp [accessed: 13.02.2020].

[7] K. Turoń, "Corporate Social Responsibility to Employees: The Best Labour Practices in Transport and Logistics Companies.", Journal of Corporate Responsibility and Leadership, 2017. Available at: https://www.researchgate.net/publication/312671862_Corporate_Social_Responsibility_to_Employees_The_Best_Labour_Practices_in_Transport_and_Logistics_Companies [accessed: 12.02.2020].

[8] M. Organ, "Cause Marketing - Definition", May 2017. Available at: https://causemarketing.com/research/cause-marketing-definition/ [accessed: 10.05.2020].

[9] P. Kotler, N. Lee, "Corporate Social Responsibility: Doing the most good for your company", John Wiley and Sons, Inc., Hoboken, New Jersey, 2005.

[10] P. Kotler, H. Kartajaya, I. Setiawan, "Marketing 4.0: Moving from Traditional to Digital", John Wiley and Sons, Inc., Hoboken, New Jersey, 2017.

[11] M. Dimock, "Defining generations: Where Millennials end and Generation Z begins", Pew Research Centre, January 2019. Available at: https://www.pewresearch.org/fact-tank/2019/01/17/where-millennials-end-and-generation-z-begins/ [accessed: 14.05.2020].

[12] "Gen Z: Audience you can't ignore", Awesomeness, 2017. Available at: https://awesomenesstvnetwork.files.wordpress.com/ [accessed: 20.03.2020].

[13] "The Millennial Impact Report", Achieve, The Case Foundation, 2017. Available at: http://www.themillennialimpact.com/latest-research [accessed: 14.05.2020].

[14] "The Deloitte Global Millennial Survey 2019", Deloitte, 2019. Available at: https://www2.deloitte.com/global/en/pages/about-deloitte/articles/millennialsurvey.html [accessed: 14.05.2020].

[15] S. Adkins, "The Big Book of Marketing, Cause-Related Marketing: Who Cares Wins", Butterworth-Heinmann, Burlington, Oxford, 2003.

The Web-based Lectures as Leverage for Developing the Sense of Belonging in the All-Russian Creative School-Contests

O.S. Fomichova* and V.A. Fomichov**

*Center of Social Competences "Dialogue of Sciences", State Budget Professional Educational Institution "Sparrow Hills", Universitetsky prospect 5, 119296 Moscow, Russia

**Department of Intelligent Monitoring Systems, Institute No. 3 "Control Systems, Informatics and Electric Power Engineering", Moscow Aviation Institute (National Research University), Orshanskaya street 3, 121552 Moscow, Russia

olga.s.fomichova@gmail.com; vfomichov@gmail.com

Abstract - **In modern Russia, the most gifted and motivated students of high schools usually enter the leading universities in biggest cities of the country, receive excellent education and usually never return to their native towns. A part of them finds jobs in abroad. As a consequence, the population of middle and small towns has been diminishing. In 2017, the authors launched an All-Russian Web-based creative school-contest for school children "Portrait of your town". It is a delicate opportunity for the child or an adolescent to tell about the beauty and originality of her/his town, to find her/his sophisticated look at the outstanding or usually not noticed facets of her/his town. In 2019, the track "Say thank you" was included into this school-contest. Its participants are to realize what persons from their native towns have made a difference in their life. The analysis of the sketches submitted to this track shows that it supports and improves emotional intelligence of the participant, analytical and critical thinking, creativity, figurative thinking, the ability to think and act in terms of public good. It may be conjectured that a new way of contributing to diminishing the acuteness of the mentioned social problem has been found.**

Keywords - creative school-contest; web-based lectures; emotional intelligence; student-self oriented learning model; soft skills

I. INTRODUCTION

In modern Russia, the most gifted and highly motivated students of high schools usually enter the leading universities in biggest cities of the country (mainly, in Moscow and Saint-Petersburg), receive an excellent education, and never return to their native towns and villages. A part of them finds jobs in abroad. As a consequence, the population of middle and small towns has been diminishing. This is one of the significant social problems in modern Russia,

In the beginning of 2017, the authors launched an All-Russian Web-based creative school-contest for school children "Portrait of your town"; it took place for the fourth time in January - June 2020. It is a delicate opportunity for the child at the age from 14 to 18 years to tell about the beauty and peculiarity of her/his town being for the child the place of living, to feel the own roots, to perceive herself/himself as a link in the chain of generations, to have a sophisticated look at her/his world and find her/his personal look at the outstanding or usually not noticed facets of her/his town - such facets without which, according to her/his opinion, the complete portrait of the country can't be formed. The technical support (Web-platform) was provided to our school-contest by the Association of Living Cities, supported by the Public Chamber of the Russian Parliament.

The most bright, beautiful thoughts from the essays of the winners of the school-contests 2017 - 2020 were presented at the first and second exhibitions "Country as an Orchestra" (March 2018 and May 2019) and at the exhibition "Portrait of the Country in the Hearts of Children" (July 2020) in the Federation Council of Russian Federation (RF) - the upper chamber of the RF Parliament ("it is called "State Douma"). The authors of the essays presented (partially) at these exhibitions were awarded by a Letter of Gratitude signed by a senator of RF (a member of the RF Federation Council).

In the beginning of 2019, a new track "Say thank you" was included into the school-contest "Portrait of your town". The diapason of ages was considerably expanded: it is from 6 years to 18 years. The participants are to realize what persons from their native towns have made a difference in their life, who has influenced them a lot in their choice of future occupation, who has given the model of behavior children would like to emulate. The first track "Say thank you" (February - June 2019) attracted the participants from 18 subjects of Russian Federation, including the subjects in Siberia, North and South of the country.

The participants submit the essays explicating their feeling of gratitude to the persons of the kind. Our **first working hypothesis** was that this track will be able to considerably contribute to developing in young people the sense of belonging to the native town. Then, after receiving an academic degree in a university, a part of young specialists will return to their native region or will regularly do something useful for their region.

The school-contest "Portrait of your town" has a methodical provision in the form of four written lectures posted on the Web. These lectures are prepared by the

first author of this paper and are oriented at the adult specialists (from a school or a library) helping children to prepare the essays for the school-contest. This interaction with an adult helps children and adolescents a lot to better explicate their feelings.

During last decade it has been broadly realized that education in knowledge society (or smart society) is to pay a particular attention to supporting and developing in the learners the significant, domain independent skills called soft skills [1], [2] or transversal skills [3], [4], first of all, analytical and critical thinking, creativity, out of box thinking, thinking and acting in terms of public good. The significance of creativity perceived by the world leading experts has been quickly increasing. In [5], the experts of the World Economic Forum (WEF) in Davos, Switzerland published the list of top 10 skills for the year 2015 and a forecast of the similar list for the year 2010. It is sufficient to say that the position of creativity shifted from the position No. 10 in 2015 to the position No. 3 in 2020. The forecast of the WEF list – 2020 includes also new skills in comparison with the WEF list 2015: emotional intelligence (the position No. 6) and cognitive flexibility (the position No. 10).

Our **second working hypothesis was** that the school-contest "Portrait of your town" and, in particular, the track "Say thank you" will contribute not only to developing the sense of belonging but also to supporting and developing a number of significant soft skills.

II. THE PECULIARITIES OF WEB-BASED LECTURES UNDERPINNED THE SCHOOL-CONTEST "PORTRAIT OF YOUR TOWN"

Four lectures written by the first author of this paper and posted on the Web-platform of the Living Cities Association (see http://xn--80addedeo5cat1j.xn--p1ai/shkola-konkurs-portret-tvoego-kraya/) are oriented mainly at the adults helping children to prepare home essays for the creative school-contest: teachers of primary, middle, and high school, the librarians, teachers of secondary education, and parents.

A. Choice of a Theme and Its Aspects

The first and principal discovery to be done by the participants of school-contest is to find a theme. A theme is something obvious or not obvious what individually defines a participant's direction of idea, the deepness of a feeling, the peculiarities of the character, of the look at the world, inner freedom and dignity.

A theme may be the feeling of space, a river, sky, a bridge, an old building, restoration, a thing in a town's museum, even the walk of the dogs across an embankment. To sum up, a theme may be everything what has left a bright semantic trace in their conceptual world's picture and forced them to have a new way of viewing themselves due to an impression from their unrepeatable and wonderful town.

A participant makes a try to answer the question about what enhanced her/his emotional impression from what she/he has seen and, in this way, makes brighter the semantic trace from this emotional impression (i.e., it is comprehension). There are possible numerous ways of viewing and aspects of viewing, in particular:

- the contours (because a contour gives a more general look at a thing, enables us to philosophically think over something);
- the details (garnish or not garnish);
- the light (sunny, during a rain, during a snowfall, under electrical lightening, inner luminescence of a thing, reflected light);
- the encirclement of a thing (a background that may provide absolutely different meanings and feelings);
- color;
- the sounds (of dribbling rain, children's voices, the singing of birds, the noise of waterfall, the violin of a street musician, etc.);
- what maximally attracted your attention and became a strong emotional experience starting the thought and images.

B. Developing Thought Producing Self of the School-contest's Participant

An impulse to the birth of a bright idea can be given by a word, image, sound, impression in case there is the inner readiness to this moment. The fireworks may give an analogy of this effect. Suddenly emerging stars, seemingly from nothing, grasp the imagination and generate new, unexpected stars which are fading, each following the own logic of combustion.

It is necessary to prepare children and adolescents to the birth of an idea, to understanding of the fact that the brain is able to produce the ideas being different from the ideas of other people. A comprehensive attitude to the work of the own brain, understanding that, by means of speculation, i.e. of analysis, generation, finding analogies, it is possible to solve a problem, to create something new, to find a figurative expression of the own feelings and to build own life in accordance with own objectives – all this is a final and necessary condition of comprehending by the child's the own thought producing self (this notion was introduced by the authors, in particular, in [6], [7]).

Successful children are thinking children being able to presume the existence of different world's conceptual pictures possessed by different people, the ability to effectively process symbolic information, to metaphorically think, to clearly express the ideas, being able to wonder and to see, as Denis Diderot said, the usual in unusual and the unusual in usual.

One old English story tells about three workers which were building the north wall of a cathedral. They fulfilled the same work: put the bricks one on another. A passer-by stopped and asked each of them what he was doing.

- I am putting the bricks, - answered the first one.

- I am building the north wall of the cathedral, - answered the second one. He is a good, qualified worker, because he sees that his work is coordinated with the general plan.

- I am praying God, - answered the third one, because he understands the correspondence between his work and the objective of building the cathedral.

It is highly desirable that the children and the adolescents, while acquiring knowledge, would be able to coordinate their every day work with a high goal.

C. The Significance of Using the Metaphors in the Text for Developing Children's Creativity

The skill to find an image which helps to clearly express an idea is a great art. The scientists drew attention to the fact that modern language is very figurative. Figurative (or metaphoric) reasoning means establishing an analogy between two different domains. It is the ability of the brain to join non similar (at the first look) entities, to discover a certain connection between them, to present the things by means of transporting a meaning from one domain to another one.

The experience of successful combination of two different disciplines, mathematics and poetry, in one course delivered in late 1990s – early 2000s in the Technological University of Rochester (the state New York, USA) can be considered as an example of developing figurative (or metaphoric) reasoning for stimulating mathematical reasoning [8], [9].

The scientists showed that such kinds of disciplines' combinations stimulate figurative reasoning in the process of solving mathematical tasks. As a result, the students demonstrate good results as concerns mastering mathematics and enjoy the learning of poetry.

The practice shows that the combination in one course of literature and painting and also of painting and psychology is successful too.

The use of metaphors in scientific papers, personal communication, presentations at international scientific conferences and other conferences helps to clarify the author's point of view and angle of look underpinning the stated material.

A broad use of metaphors in the process of writing a portrait of the native town or expressing gratitude to a person who was able to make a difference in the author's life will make the work for the contest more bright, interesting, and clear to the readers.

III. THE RESULTS OF THE FOURTH SCHOOL-CONTEST AND ITS TRACK "SAY THANK YOU"

A. General Information About the All-Russian Creative School-contest "Portrait of your town" and its Track "Say thank you"

The statistical results of the fourth creative school-contest "Portrait of your town" (January – June 2020) are as follows: the participants are from 24 subjects of Russian Federation, from West (Pskov) and North-West (Petrozavodsk) to North (Norilsk), Ural, Siberia, and Far East (including the island Sakhalin). 191 essays became the winners of the school-contest.

The exhibition "Portrait of the Country in the Hearts of Children" in the Federation Council of the RF Parliament (July 2020) consisted of 60 pictures containing the fragments with the brightest ideas from the works of the winners. 163 participants received a Letter of Gratitude from a senator of RF.

The analysis of the essays submitted to the track "Say thank you" has enabled us to construct an unexpectedly rich classification of the persons receiving "thank you" from the participants. This classification is as follows: 17% - friend; 15% - school teacher; 12% - mother; 10% - teacher of extra education; 10% - inspiring person; 7% - father; 7% - relatives; 7% - manager; 5% - grandmothers and grandfathers; 5% - unknown person; 3% - tutor; 2% - school employee. The constructed classification was used as a system of nominations.

B. Examples of Home Compositions

Example (analytical thinking). The 8 year old girl Kate writes: "One autumn day the parents led me to a ballet school. Yana K. became my teacher. She seemed to me being just, kind, and beautiful. She is teaching us good but rigorously. She wants everything to be OK in our life and that is why we are to undertake efforts. I've realized that it is necessary to make much efforts for achieving a result. I've realized that if one undertakes much efforts and works very much then the world and the people in this world will become better.

Thank you, my teacher, for evoking in me the belief that I will be able to cope with everything, and it means that I'll be able to do everything.

Example (empathy, mindfulness). A 11 years old girl Maria writes: "One spring day a young men opened a door of his car, intending to drive away. Suddenly he noticed a young girl who was crying, because her cat was unable to climb down to her from the top of a high tree. The young men reached the top of the tree, grasped the cat, and returned back with the cat. The young girl was happy".

Example (empathy, thinking and acting in terms of public good). A 8 years old boy Andrey writes: "I would like to say THANK YOU VERY MUCH to my teacher Maria K. She is a very kind person and a person who is ready to help. One autumn day we with my class had an excursion in Kremlin. It was November, and it was cold. I lost my hat in school. At the moment when all the classmates were dressed and were waiting for me, I was looking for my hat, but all my efforts failed. Then my classmates started to help me to find my hat, but we failed to do it. Then our teacher gave me her hat and went on excursion without hat. Everything was good, we successfully went on excursion, and nobody from us became ill".

IV.THE SCHOOL-CONTEST "PORTRAIT OF YOUR TOWN" AS THE THIRD IMPLEMENTATION OF THE STUDENT-SELF ORIENTED LEARNING MODEL

It is possible to indicate an interesting interpretation of the school-contest "Portrait of your town" in the context of the achievements of modern education theory. During last decade, the peculiarities of knowledge society have created the need for a new paradigm of educational process. One of the most significant peculiarities of the kind is the necessity to change a profession several times during the working period of life.

Proceeding from broadly accepted role of emotional intelligence (EI) in professional and personal life, the authors suggested a new learning model (LM) called Student-Self Oriented LM (SSOL-model). It is defined as the model being beneficial for self-cognition and self-construction through the prism of the acquired knowledge and life experience [10], [11], 12].

The principal distinguished features of the SSOL-model are as follows: (a) it takes into account and bases educational process on EI of the learners; (b) it aims at reaching cognitive engagement of the learners; for this, it proposes a conceptual learning environment instead of the environment based on mechanical remembering (it means addition of emotional color to the studied notions and, as a consequence, makes much easier the comprehension of these notions) [11], [23], [32].

The scientific literature describes two successful implementations of the SSOL-model. The first one is the *System of Emotional-Imaginative Teaching* (*the EIT-system*), based on the *Theory of Dynamic Conceptual Mappings (the DCM-theory)*. Both this theory and the EIT-system are developed by the authors of this paper. The EIT-system is aimed at systematic development of EI, reasoning skills, sound creativity, figurative (metaphoric) thinking, language skills, and communication culture at the lessons of language - mother tongue and second language, literature and poetry in two languages (on the example of Russian and English), symbolic languages of painting, sculpture, garden-park art, classic dance. We have accumulated the 30-year-long successful experience of using the EIT-system in extra education in Moscow, Russia. Many aspects of the DCM-theory and the EIT-system are described, in particular, in the authors' works [11], [13], [14], [15], [16], [17], [18], [19], [20], [21]. [31], [32].

The DCM-theory and the EIT-system belong to the constructive core of *cognitonics,* or *the science about the human being in the digital world* [11], [19], [20], [22], [23], [32]. The authors organized as the co-chairs six international conferences on cognitonics in conjunction with the international scientific multiconferences "Information Society" (Slovenia, Ljubljana, Jozef Stefan Institute, October 2009, 2011, 2013, 2015, 2017, 2019, see https://is.ijs.si/?page_id=903). Totally, the researchers form 25 counties participated with the papers in six international conferences on cognitonics.

The ideas and methods of cognitonics underpinned a number of educational projects in several countries, in particular, in Croatia and Macedonia [24], Croatia [25], United Kingdom [26], [27], and Mexico, United Kingdom and PR China [28].

In [29], the authors formulated the following expanded definition of cognitonics*: it is the science about the trajectories of raising the human being to such level of intellectual and spiritual height where the scale of his/her personality becomes proportional to the scale of the digital world.*

The second implementation of the SSOL-model is Art Cognitonics (AC) [11], [12], [23], [30], [32] - one of the principal branches of cognitonics. AC aims at tuning the EI of young children and adolescents with the help of well-known works of art. The goal is to create a bright semantic trace in the world's conceptual picture of the learner corresponding to an idea explaining or illustrating a moral value, communicative situation, a situation of making a decision, cognitive process itself, the process of self-cognition and consideration, the seething cocktail of emotions, a way of viewing the world around, etc.

AC establishes the links between the objects, situations, processes, views of a person (a beholder) and the work of art that becomes a metaphor or a vivid illustration (vivid mental representation) of something the beholder is considering about. That is why the consciousness of the beholder receives a considerable impulse to developing the ability of establishing diverse analogies and consequently to finding a new look at a situation.

Example. For enriching the color of their canvases, the impressionists made use of what is known as *division of color and optical blending*. E.g., to represent a green meadow, they put little dabs of blue and yellow on the canvas which are supposed to be combined to form green in the eye of the beholder – a far more intense green than one taken straight from the artist's palette. That is why it is impossible to understand the idea of a picture standing close to the canvas. We have to step aside and look at it from a certain distance to enjoy it and to have the desired effect.

We have the same situation in every-day life. "Multiple dabs, reflections" prevent us from grasping the sense of what is happening. As in case with impressionists' canvases, we have to have a look at the situation from a distance, and *distance in this case is equal to time distance.* We need some time to better understand what has happened, and this will help us to cope with the situation.

Let's consider the following principal factors providing the possibility to interpret the All-Russian creative school-contest "Portrait of your town" as the third implementation of the SSOL-model of educational process. As it was mentioned above, the main idea underpinned the creation of the SSOL-model was, while interacting with the learners, to take into account and to improve emotional intelligence of the learners.

We indicate a spectrum of the themes' facets but not the theme. The children and adolescents participating in the main track of the school-contest are to ask the question about the existence in their towns of such

precious peculiarities that the portrait of the whole country would be incomplete without describing these peculiarities. Every participant of the track “Say thank you” is to ask herself/himself whether there are such people in her/his surroundings that the interaction with them turned out to be a defining experience for the participant. Children discover the delighting traits of character, the manner of speech, the ability to attentively listen to for a long time, etc.

After having delighted, a child or an adolescent is to critically have a look inside herself/himself and ask herself/himself whether she/he possesses the traits of the character delighted her/him and whether she/he wants to improve her/his character. The participants of the track receive the possibility to understand and to appreciate what they do know. We create a thought provoking situation which stimulates children and adolescents to analyze facts, to think over, to develop critical thinking (when a school girl or boy understands that she/he doesn’t possess an attractive trait of the character). The transfer of an experience from one situation to another situation contributes in many cases to developing *cognitive flexibility*.

V. THE GROWTH OF SELF-ESTIMATION IN PEOPLE OF USUAL, MASS PROFESSIONS. GOODNET

In every country, a great majority of people have the professions which don’t attract public attention so much. This applies, in particular, to school teachers, hospital workers, nurses, bus drivers, postmen, gardeners. Such people usually believe that nothing in the country depends on them.

The victory of a school girl or boy in the track "Say thank you" becomes a thought provoking event in a small settlement or town. In many cases, the local newspaper tells about both the child and an adult being the hero of her/his sketch. As a result, many people in the country having usual, mass professions, have realized that they are able to make a difference for somebody at their working place. They become to be known in a small town or a village or in a district of a city. As a consequence, the self-assessment of these people increases. Children formulate new values; as a result, a new category of significant and distinguished people emerges.

Example (concierge). A 6 years old girl Vika writes: "It is monsieur Dima. He is a concierge in our home. Concierge is a man who defends us from bad people. Monsieur Dima is very kind. When I enter the home, I see him, and immediately my spirits are getting bright. Sometimes we are speaking, and sometimes he puts me on his shoulders and delivers me to the elevator, I am laughing. Monsieur Dima acts in this way, because he is kind hearted. Thank you, monsieur Dima, because you are kind and good".

The contest 'Say thank you" has shown the significance for children and adolescents of the notion "a hero of our time". Simple people from their surroundings thinking and acting in terms of public good become such heroes.

The significance of the track can be better grasped in case we compare it with the WWW. The great peculiarity of the Web is that it reaches every family, every person (being not a very young child). The track "Say thank you" can be interpreted as *a digital platform for explicating estimation and distributing good. Then the track can be called GoodNet* - it reaches every family, connects the generations and receives a response in every heart. The track "Say thank you" is a platform where the demand, significance, and reality of respectable attitude to each other bridges the gap between generations, explicates and enhances intrinsic interest towards other people.

VI. CONLUSION

The observations during several years are necessary in order to verify our first working hypothesis: the systematic participation in the school-contest “Portrait of your town” will improve in young people the sense of belonging to the native town, and it will contribute to increasing the proportion of young specialists returning to native town or native region after receiving higher education.

The analysis of the essays written by the winners of four All-Russian creative school-contests “Portrait of your town” convincingly confirms our second working hypothesis: this school-contest contributes to supporting and improving a number of significant soft skills: emotional intelligence of the participants, analytical and critical thinking, creativity, figurative thinking, out of box thinking, and the ability to think and act in terms of public good.

That is why there are the reasons to say about the creation of All-Russian digital platform helping both children and adolescents from various parts of the big country to develop a number of soft skills being precious both for professional success and personal life.

The planned next step consists in the attraction to the school-contest of the gifted children and adolescents from remote villages, in particular, from mountain and taiga villages. The purpose is to help gifted children and adolescents from the villages of the kind to increase self-assessment and to support and increase the faith in the own forces.

ACKNOWLEDGMENT

Many thanks to two anonymous referees of this paper for the remarks contributing to the quality of its text.

REFERENCES

[1] Soft skills. TVETipedia Glossary, 2019. Available at https://unevoc.unesco.org/go.php?q=TVETipedia+Glossary+A-Z&filt=all&id=602, [accessed: 21.12.2019].

[2] V. Buselic, “Information literacy and critical thinking freshman course experience”, in the Proceedings of the 41st International Convention MIPRO 2018 (Opatija, May 20-24, 2019), Conference “Computers in Education”, P. Biljanovic et al., Eds, Opatija, Croatia: Mipro and IEEE, 2019, pp. 920-925.

[3] Transversal skills. TVETipedia Glossary, 2019. Available at https://unevoc.unesco.org/go.php?q=TVETipedia+Glossary+A-Z&id=577, [accessed: 21.12.2019].

[4] R. M. Bottino**,** A. Chioccariello, L. Freina, and M. Travella, “Digital games in primary schools for the development of key transversal skills”, in Sustainable ICT, Education and Learning. IFIP WG 3.4 International Conference, SUZA 2019, Zanzibar, Tanzania, April 25–27, 2019, Revised Selected Papers, A. Tatnull and N. Mavengere, Eds. Cham: Springer, 2019, pp. 55-65.

[5] A. Gray, “The 10 skills you need to thrive in the fourth industrial revolution”, 2016. Available at

https://weform.org/agenda/2016/01/the-10-skills-you need-to-thrive-in-the-fourth-industrial-revolution/, [accessed: 21.12.2019].

[6] V. A. Fomichov and O. S. Fomichova, "The social responsibility of computer science specialists for the creative potential of the young generation". Intern. Journal of Artificial Intelligence in Education. vol. 11, 2000, pp. 208-219.

[7] O. S. Fomichova and V. A. Fomichov, "Computers and the thought-producing self of the young child", The British Journal of Educational Technology, vol. 31, pp. 213-220.

[8] M. Birken M. and A. Coon, "The pedagogical and epistemological uses of analogy in poetry and mathematics", International peer-reviewed on-line journal "Consciousness in Literature and the Arts", 2001, vol. 2.

[9] A. Coon and M. Birken, "Developing analogical thinking and creativity in university students", International Peer-Reviewed Online Journal "Consciousness, Literature, and the Arts", vol. 4, 2003.

[10] V. A. Fomichov and O. S. Fomichova, "The Student-Self Oriented Learning Model as a Paradigm for Supporting and De veloping Emotional Intelligence and Creativity", in Informacijska druzba - IS 2017. Proceedings of the 20th International Multiconference - IS 2017, Edited by V. A. Fomichov, O. S. Fomichova. Vol. Kognitonika/Cognitonics. October 9th-10th, 2017, Ljubljana, Slovenia . Jozef Stefan Institute, Ljubljana, pp. 11-16.

[11] V. A. Fomichov and O. S. Fomichova, Cognitonics and Its Significance for Education in the Digital Age. Moscow, Publishing House MAKS Press. 344 p. (in Russian).

[12] V. A. Fomichov and O. S. Fomichova, "The student-self oriented learning model as an effective paradigm for education in knowledge society". Informatica. An International Journal of Computing and Informatics, Vol. 43, No. 1, pp. 95-107, 2019

[13] V. A. Fomichov and O. S. Fomichova, "The theory of dynamic conceptual mappings and its significance for education, cognitive science, and artificial intelligence", Informatica. An International Journal of Computing and Informatics, vol. 8, pp. 31-148, 1994.

[14] V. A. Fomichov and O. S. Fomichova, "The artificial intelligence theory and highly effective methods of teaching young children foreign languages". Cybernetica, vol. XXXVIII, 1995, pp. 321-344.

[15] O. S. Fomichova and V. A. Fomichov, "Theoretical foundations of a new method of teaching children effective information processing". Informatica. An International Journal of Computing and Informatics, vol. 20, pp. 381-399, 1996.

[16] V. A. Fomichov and O. S. Fomichova, "An informational conception of developing the consciousness of the child", Informatica. An International Journal of Computing and Informatics (Slovenia), vol. 21, 1997, pp. 371-390.

[17] V. A. Fomichov and O. S. Fomichova, "A new theoretical and practical approach to early positive developing child's consciousness", in R. Trappl (Editor), Cybernetics and Systems'98. Proceedings of the 14th European Meeting on Cybernetics and Systems Research. Vol. 1, Vienna: Austrian Society for Cybernetic Studies, 1998, pp. 276-281.

[18] V. A. Fomichov and O. S. Fomichova, "Early development of natural-language-processing abilities as a key to up-bringing of children without conflicts", in G. Ritschard, A. Berchtold, F. Duc, D.A. Zighed, Eds. Apprentissage: des principes naturels aux methodes artificielles. Editions HERMES, Paris, 1998, pp. 67-81.

[19] O. S. Fomichova and V. A. Fomichov, "Cognitonics as an answer to the challenge of time", in Proceedings of the 12th International Multiconference Information Society - IS 2009, Slovenia, Ljubljana, 12 – 16 October 2009. The Conference Kognitonika/Cognitonics. Jozef Stefan Institute, Ljubljana, 2009, pp. 431-434.

[20] V. A. Fomichov and O. S. Fomichova, "A contribution of cognitonics to secure living in information society", Informatica. An International Journal of Computing and Informatics (Slovenia), vol. 36, 2012, pp. 121-130.

[21] V. A. Fomichov, "Conscious control during childhood, development of", in International Encyclopedia of the Social and Behavioral Sciences, Second Edition ,vol. 4. Oxford: Elsevier, 2015, pp. 666-672.

[22] V. A. Fomichov and O. S. Fomichova, "Cognitonics as a new science and its significance for informatics and information society", Informatica. An International Journal of Computing and Informatics (Slovenia), vol. 30, 2006, pp. 387-398.

[23] V. A. Fomichov and O. S. Fomichova, "An imperative of a poorly recognized existential risk: early socialization of smart young generation in information society", Informatica. An International Journal of Computing and Informatics, vol. 38, 2014, pp. 59-70.

[24] I. Panev, I. Pogarcic and L. Gjergjeska, "Goals of cognitonics in formal ICT education", in Informacijska druzba - IS 2017. Zvezek D. Proc. of the 20th Intern. Multiconference - IS 2017, Edited by V. A. Fomichov, O. S. Fomichova. Vol. D, Kognitonika/Cognitonics. October 9th-10th, 2017, Ljubljana, Slovenia. Jozef Stefan Institute, Ljubljana, 2017, pp. 61-65.

[25] I. Panev, "Creating multimedia systems according to principles of cognitonics", in Informacijska druzba - IS 2019. Zvezek E. Proc. of the 22nd Intern. Multiconference - IS 2019, Edited by V. A. Fomichov, O. S. Fomichova. Vol. E, Kognitonika. International Conference on Cognitonics. October 7th-8th, 2019, Ljubljana, Slovenia. Jozef Stefan Institute, Ljubljana, 2019, pp. 50-53.

[26] T. B. Kane, "Using cognitive tunnels in a new approach to building social elevators in the information society", Informatica. An International Journal of Computing and Informatics (Slovenia), vol. 38, p. 263-271, 2014.

[27] T. B. Kane, "A Cognitonics methodology for artificial persons", in Zbornik 20. mednarodne multikonference INFORMACIJSKA DRUZBA - IS 2017. Zvezek D. Proceedings of the 20th International Multiconference INFORMATION SOCIETY - IS 2017, Volume D. Kognitonika/Cognitonics. Uredila / Edited by Vladimir A. Fomichov, Olga S. Fomichova. 9.-13. oktober 2017/ 9th - 13th October 2017, Ljubljana, Slovenia. Ljubljana, Jozef Stefan Institute, 2017, pp. 30-33.

[28] P. Craig, N. Roa-Seiler, M. Martínez., and F. Lara Rosano, "A cognitonics approach to computer supported learning in the Mexican state of Oaxaca", Informatica. An International Journal of Computing and Informatics (Slovenia), vol. 38, 2014, pp. 241-248.

[29] V. A. Fomichov and O. S. Fomichova, "Predgovor/Foreword", in Informacijska druzba - IS 2019. Zvezek E. Proc. of the 22nd Intern. Multiconference - IS 2019, Edited by V. A. Fomichov, O. S. Fomichova. Vol. E, Kognitonika. International Conference on Cognitonics. October 7th-8th, 2019, Ljubljana, Slovenia. Jozef Stefan Institute, Ljubljana, pp. 3-4.

[30] V. A. Fomichov and O. S. Fomichova,, "Early Development of the Human Being Ideal Constituent and Art Cognitonics as the Answer to the Challenge of Knowledge Society". In: Informacijska druzba - IS 2015. Proc. of the 18th International Multiconference - IS 2015, Edited by V. A. Fomichov, O. S. Fomichova. Vol. F: Kognitonika/Cognitonics. Oct. 12th-13th, 2015, Ljubljana. Jozef Stefan Institute, 2015, pp. 27-32.

[31] O. S. Fomichova and V. A. Fomichov, "The pillars of higher personality development of net generation in smart society", in Zbornik 22. mednarodne multikonference INFORMACIJSKA DRUZBA - IS 2019. Zvezek E. Proceedings of the 22th International Multiconference INFORMATION SOCIETY - IS 2019, Volume E. Kognitonika/Cognitonics. Uredila/Edited by Vladimir A. Fomichov, Olga S. Fomichova. 7.-8. oktober 2019/ 7-8 October 2019, Ljubljana, Slovenia. Ljubljana : Jozef Stefan Institute, 2019, pp. 15-18 .

[32] O. S. Fomichova and V. A. Fomichov. Education with the Focus on Education in Knowledge Society. Moscow: MAKS Press, 2020 (in Russian).

STEAM Students and Their Expectations from Future Business Life: a Values-driven Workplace

Yelyzaveta Borysiuk* and Tihana Babić*
*Algebra University College, Zagreb, Croatia
yelyzaveta.borysiuk@racunarstvo.hr, tihana.babic@algebra.hr

***Abstract* - In the last few decades, the global economy has been witnessing a shift from manufacturing to a service economy. Affected by that structural transformation, the service component of products and goods is rising. Companies are highly in need of young talents who can provide high-quality service for their customers. While providing new services and experiences for their customers, employees raise their expectations and demand an encouraging place to work, better working conditions, equality, and engagement, thereby to get satisfaction and have a sense of meaning in everyday life. The world of technologies no longer allows organizations to neglect these employee's needs and requirements. People no longer come to work just because they have to, but because they want to make an impact and be useful while proceeding with personal development. This paper's purpose was to investigate the values and conditions that STEAM students expect in their workplaces in the future and identify the differences in regards to workplace demands that previous generations had sought.**

Keywords – workplace, value-driven workplace, service economy, STEAM students

I. INTRODUCTION

In the second half of the 20th century happened a shift from production to the service economy. This resulted in the appearance of highly skilled and highly paid producer service jobs in information systems, marketing, design, and finance. Together with globalization, the new opportunities and knowledge appeared. From working harder physically, people were getting to know how to work smarter [1]. When it comes to work conditions, in the early 1960s until the late 1970s employees were dealing with insecurity at work, frightened to be the least wanted generation, but they had realistic views about the economy and employment rates. As fast as the global economy was developing, they were ready to increase their expertise in technology, and were self-reliant, though, people did not expect lifetime employment. At that time, family and work started to be seen as separate institutions, and Generation X started to center their interest around their own homes and was less likely to advance and assume more responsibility in their jobs. That is why children born in 1981 – 1996 were surrounded not only by radical changes in the world economy and social norms and values but also by incredible support and care from their parents. Positive tolerance, feeling of having a common win, togetherness, open possibilities, best education, and technological revolution – these are the few things among dozen of changes, which impacted life choices and behaviors of new generations. Many employers can agree, that this generation has many more challenges, hence many more opportunities to become better and learn on the mistakes their parents and grandparents did. They understand the importance of having a balance between family and work while getting success, acknowledgment, and respect. From this arise the fact, that the current working-age generation is striving for meaning rather than a technical simple job, for value rather than a high salary, for happiness rather than everyday struggle. These are the notions that were not seen and spoken before, hence we can see it as a progressive mindset of modern people that will bring to a better future [2].

For the first time in the history of evolution, 4 the most progressive generations are working side by side: baby boomers, generation X, millennials, and generation Z. Each generation has its values, and unique experiences, coming from their critical life stages and historical changes in the world. Today, from the young generation we can hear completely different terms when it comes to values. The most valuable things among youth are truth, openness, freedom of speech and thoughts, personal impact, and sustainability. Even though none of the present generations is homogenous and amenable to diversity and adjustment, but the environment, social and cultural melting points, and global events – this is what should take into account when looking for a meaning and better quality of self-being in work ethic, expectations, and behavior [2].

Together with Millennials and Generation Z appeared a notion of STEAM. Science, Technology, Engineering, the Arts, and Mathematics – today progress is based on these fields. These students invent, lead, connect others, make things happen, and create orders of chaos. They figure out what to do and how to look at things differently, without a rule book. They are not following the path of their parents, though raising the plank of the quality of life, happiness from what they are doing, and satisfaction in all aspects of life.

Until recent times, the main task of employers was to ease the entry of the youngest generation in the workforce and helping them to engage in work, while providing the same amount of attention and respect to older generations. However, the time has changed, and the 21st century is a time for young and innovative generations, that show what

they are capable of while creating a happier life at home and work [3].

II. CHANGES IN WORKPLACE: TRANSFORMATIONS AND EXPECTATIONS

While a new generation is entering employment, companies are acting to attract and retain this new talent pool. The 'values revolution' was created and driven by the Millennials generation while others refer to a clash of values between different age groups. Business leaders are increasingly required to take these issues into account and develop tailored ways to encourage behaviors from each age group [4].

The main changes include, as Ester Perel highlighted:

- *the rise of expectations* - never before have young people expected so much from their career or their partner. They want flexibility, they want their workplace to be attentive to the wellbeing and jobs to help them find a sense of meaning and purpose;
- *people now bring emotional capital into the workplace* - people are encouraging emotions at the workplace. Authenticity, trust, belonging, transparency, and psychological safety are common workplace discussions[5];
- *balance of short term instead of long-term sustainability* - companies must carefully balance the short term, such as quarterly profits, versus long-term sustainability as a successful company. This requires recognizing the value of work that each person contributes and devising a fair, and sometimes creative, compensation plan. To have proactive rather than reactive behavior is the ethical thing to do [5].

That is why younger people have attributes to bring and expectations to be fulfilled, that affect the development of workplace relationships with team and organizational members from other generations [6]. This generation is not motivated by feelings of duty-working hard is not virtuous in itself, but it is worth it if they are singled out and recognized. They appreciate directness rather than an abstraction. They do not have automatic respect for authority and will feel free to make suggestions if they think it will improve things When it comes to wages, managers must balance issues of compensation equity, employee morale, motivation, and profits—all of which may have legal, ethical, and business elements [7].

Today young employees are very flexible and used to dealing with diversity. If you need someone to meet clients who are from a different culture or background, a young person-even if he or she is not from the same background-will adapt to this situation well. Young people today often have colleagues from different backgrounds and have a desire to work in multicultural and diverse companies. The knowledge of languages became one of the significant features of young generations. This became a huge benefit for both sides, and organizations broadened the horizons of communication and made new international connections. On the other hand, young employees may need some guidance on how to deal with older people. They may come off as disrespectful when they are merely being friendly and informal. Some young employees might need to be taught to "clean up" when talking to older folks, using "Mr." and "Mrs." and speaking more formally. Today's young employees also appreciate flexible schedules and independence. They don't respond well to micromanagement and will find rigid schedules stifling [8].

III. STEAM STUDENTS AND IMPACT OF THE 'UNEXPECTED' ON JOB-RELATED VALUES

One of the reasons, why Millennials and Generation Z are valuable workers during any circumstances is because most of them are STEAM students and postgraduates. These students invent, lead, connect others, make things happen, and create orders of chaos. They figure out what to do and how to look at things differently, without a rule book. They want to love their work and make the best out of it. In today's world, these young people are the ones who get the best jobs and the most freedom.

For years, employers were creating engagement structures without any differentiation between generations. As the millennial generation of STEAM students grows in the workforce and baby boomers retire, managers and human resources professionals will need to develop new, more ethical, and socially conscious, managing models.

Both the private and public sectors report that 21st-century workers require skills that many of today's graduates don't have. Students need more in-depth knowledge of math and science, plus the ability to integrate and apply that knowledge to solve the challenges facing our nation. Children who study STEM also develop a variety of skills that are essential for success: critical thinking and problem solving, creativity and innovation, communication, collaboration, and entrepreneurship [9].

However, by the rapid technological and social revolution, young people were lacking something, that will involve creativity. As a creative economy became one of the leading in the western part of the world, Millennials were more and more interested in this sector. Adding Art to the notion of STEM was a revolutionary act. It offered learners and their teachers another way for meaning-making, problem-solving and expressing understanding. With STEAM, students naturally see the endless possibilities and intersection points as they choose – even create – their career paths. That is one of the main distinctions between the older and younger generation – creativity and universality.

It is a democratization of thinking and an unstoppable flow of ideas that will bring any organization a huge development. The STEAM student learns to be a critical and creative thinker and to communicate well in person, on the page, and on the screen. They build cultural understanding so that they can make sense of the increasingly interrelated world, and act upon it in imaginative ways. And these skills and knowledge will help Millennials succeed at jobs in business, education, the not-for-profit sector, or government, that may not yet

exist [10]. The late Steve Jobs famously said, "It is in Apple's DNA that technology alone is not enough—it's technology married with liberal arts, married with the humanities, that yields us the results that make our heart sing" [11].

The modern world makes us understand that we live in an ever-changing, increasingly complex times and our citizens must possess the knowledge and skills to solve problems while creating more meaning and rise a plank of good life possibilities and values. Too many are falling for false Internet narratives and social media conspiracy theories daily. STEAM is the gateway way to 21st-century skills like critical thinking, communication, creativity, collaboration, and a better future. Students are mobile and willing to take new opportunities, regardless of the situation. Promoting accessibility and openness during a difficult time will prevent them from job hunting on the extra downtime they have. Innovative decisions are magnified in times like these, and their comfort with technology allows them to interview quickly — something prospective employers could take advantage of [12].

IV. THE RESEARCH METHODOLOGY

A. Purpose of the Research

The main purpose of research conducted among Algebra University College students was to investigate the values and conditions that STEAM students expect in their workplaces in the future and identify the differences in regards to workplace demands that previous generations had sought.

The main research objective was: *What values and conditions STEAM students expect in their workplaces in the future?*

To get a detailed insight the following sub-questions were defined:

1. *Are STEAM students already employed and in which type of organization, if so?*
2. *Which conditions STEAM students expect in their workplaces?*
3. *Which values STEAM students expect in their workplaces?*

B. The Research Population Sample

The research was conducted among the Algebra University College students during the winter semester of the academic year 2019/2020.

The total population sample was 83 participants (N=83) out of a total of 1393 students enrolled at the Algebra University College in the academic year 2019/2020 (almost 6 %). A sample of participants was pertinent. 47% (39) were females, and 53% (44) were males. Among 83 surveyed there are 32 (38,6%) full-time students, and 51 (61,4%) of part-time students. The majority of the participants (73%) were undergraduate students, which is 61 student, and 22 (26%) students of them are graduate students. Participants were students from 7 different STEAM study programs: 17 students of Software engineering, 6 students of System engineering, 6 students of Multimedia Computing, 13 of Data Science, 7 of Digital marketing, and 34 of Visual communications design. The structure of all surveyed according to the study programs is shown in Table 1.

TABLE 1. THE STRUCTURE OF THE PARTICIPANTS BY STUDY PROGRAMS IN PERCENTAGE, N=83.

Study Program	Percentage of Students per Study Program
Data Science	**16 %**
Design & Communications Management	**5 %**
Digital Marketing	**8 %**
Multimedia Computing	**2 %**
Software Engineering	**20 %**
System Engineering	**8 %**
Visual Communication Design	**41 %**

As the research paper mainly stresses the difference in generations, it was important to research the age of the participants. The division was between 3 groups. 6 participants were born between the years 1971-1979, 36 were born between 1982 and 1996, and 41 were born between the years 1997-2001. The structure of all surveyed according to the age group is shown in Table 2.

TABLE 2. THE STRUCTURE OF THE PARTICIPANTS BY AGE GROUP, N=83.

Age group	Participants in percentages
Generation X (1971-1979)	**7.2 %**
Millennials (1982-1996)	**43.4 %**
Generation Z (1997-2001)	**49.4 %**

C. The Research Methods

An anonymous voluntary survey was conducted through a specially designed questionnaire in the Google Forms tool. The questionnaire had 11 questions, of which 4 were related to demographics. For 4 questions answers were defined with the 5 degrees of frequency according to the Likert scale. 2 questions had 8 statements related to the expectations of the participants of conditions and values from the workplace.

A quantitative method was used for the analysis of the research results. The data processing of the survey results was made through the Google Forms tool.

D. The Research Results

1. *Are STEAM students already employed and in which type of organization, if so?*

The research results have shown, that almost half of the surveyed students, 38 of them, are currently working for an employer. While 23 students currently are unemployed but were employed in the past. 12 participants have never had an occupation before, and 10 are working for themselves, which means they have their own business. All the data is shown in Table 3.

TABLE 3. THE STRUCTURE OF THE PARTICIPANTS BY THE EMPLOYMENT RATE, N=83

Employed at past	Students in percentages
Self-employed	**12 %**
Work for employer	**45.8 %**
Employed at past	**27.7 %**
Not been employed yet	**14.5 %**

When it comes to the place of employment, 33 participants are working in a private company, and 3 of them working in the trade sector. 3 participants are working for start-up companies, and 27 are not employed at all. All the data is shown in Table 4. below.

TABLE 4. THE STRUCTURE OF THE PARTICIPANTS BY THE PLACE OF EMPLOYMENT, N=83.

Place of employment	Students in percentages
Private company	**40.2 %**
Trade sector	**4.9 %**
Start-up companies	**3.7 %**
Not employed	**32.9 %**
State company / agency	**8.5 %**
Freelancer	**6.1 %**
None of the mentioned	**3.7 %**

One of the most important objectives of the research was to understand the level of satisfaction of the participants from their current workplace. 8 surveyed are completely satisfied with their work, and 28 students are satisfied. While 12 of the participants are neither satisfied nor unsatisfied, only 5 people are not satisfied. All the data can be seen in Table 5.

TABLE 5. THE STRUCTURE OF THE PARTICIPANTS BY THE LEVEL OF SATISFACTION FROM WORK. N=83.

Level of satisfaction with the place of employment	Students in percentages
Completely satisfied	**9.8 %**
Satisfied	**34.1 %**
Neither satisfied, neither unsatisfied	**14.6 %**
Not satisfied	**6.1 %**
Completely unsatisfied	**0 %**
Not employed	**35.4 %**

2. *Which conditions STEAM students expect in their workplaces?*

Participants were offered to choose from several conditions, that should be satisfied while people are working. One of the most important conditions appeared to be a fair salary (68 participants strongly agree), respect for employees (70 participants strongly agree), and no harassment at a workplace (69 participants strongly agreed). Which can highlight, that people these days are demanding at least safety and respect and good payment for their work. The structure that represents the student's degree of preferences in conditions according to the Likert frequency scale is presented in Table 6. The statements represented by letters in the table are as following:

A. *A fair salary*
B. *Respect for employer and employee*
C. *A workplace where there is no harassment and abuse*
D. *Safe working conditions*
E. *Transparency*
F. *Freedom to make decisions concerning my scope of work*
G. *Education and development (that we are provided with additional education)*
H. *The right to privacy (eg. without employee supervision if the information does not explicitly relate to the work we do).*

TABLE 6. DISTRIBUTION OF THE DEGREE OF PREFERENCES IN WORKPLACE CONDITIONS ACCORDING TO THE LIKERT FREQUENCY SCALE.

STATEMENT	Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
A.	**0 %**	**0 %**	**0 %**	**18 %**	**82 %**
B.	**1 %**	**0 %**	**0 %**	**15 %**	**84 %**
C.	**0 %**	**0 %**	**0 %**	**17 %**	**83 %**
D.	**0 %**	**0 %**	**2 %**	**24 %**	**74 %**
E.	**0 %**	**1 %**	**2 %**	**29 %**	**68 %**
F.	**0 %**	**0 %**	**12 %**	**45 %**	**43 %**
G.	**1 %**	**1 %**	**5 %**	**46 %**	**47 %**
H.	**1 %**	**0 %**	**5 %**	**33 %**	**61 %**

3. *Which values STEAM students expect in their workplaces?*

According to the values chosen, the biggest amount of participants gave their preference to a sense of purpose at work (61 participants strongly agree), trust between employers and employees (53 participants strongly agree), emotional well-being (39 strongly agreed), respect for the customer (37 strongly agreed) and quality of service/ product of the company (56 strongly agreed). The

structure that represents the students' preferences in values according to the Likert frequency scale is presented in Table 7. The statements represented by letters are as following:

A. *Personal fulfillment*
B. *Ethical decision making*
C. *Sense of purpose (that our work makes sense)*
D. *Trust between employers and employees*
E. *Loyalty to the employer*
F. *Emotional well-being*
G. *Respect for the customer*
H. *Quality of service/product of the company*.

TABLE 7. DISTRIBUTION OF THE DEGREE OF PREFERENCES IN VALUES ACCORDING TO THE LIKERT FREQUENCY SCALE.

STATEMENT	Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
A.	**1 %**	**0 %**	**7 %**	**33 %**	**59 %**
B.	**0 %**	**2 %**	**5 %**	**41 %**	**52 %**
C.	**0 %**	**1 %**	**4 %**	**22 %**	**73 %**
D.	**0 %**	**1 %**	**4 %**	**31 %**	**64 %**
E.	**0 %**	**2 %**	**18 %**	**37 %**	**43 %**
F.	**1 %**	**2 %**	**10 %**	**40 %**	**47 %**
G.	**0 %**	**0 %**	**2 %**	**41 %**	**57 %**
H.	**0 %**	**0 %**	**4 %**	**29 %**	**67 %**

As the final question, participants were offered to vote for their decision, whether they would like to change jobs if they were not satisfied with any of the conditions or values. As a result, 51 of the surveyed would fist negotiate with their employers, while 16 students would stay at the workplace if their limits are not disrupted. And only 13 students would change their workplace straightaway. Detailed data in Table 8.

TABLE 8. THE STRUCTURE OF THE PARTICIPANTS BY A PROPENSITY TO CHANGE THE WORKPLACE. N=83.

Change of workplace	Students in percentages
Would rather negotiate with their employers.	**62.2 %**
Would stay at the workplace, if their limits are not disrupted.	**19.5 %**
Would change their workplace straightaway.	**15.9 %**
Would adjust because of no interest in the change of workplace.	**0 %**
Would stay because they have no other choice.	**2.4 %**

E. Limitations and recommendations for future researches

It is important to emphasize the difficulties encountered during the research and ultimately its limiting elements; the method of distributing the survey was limited to sampling, In future research, it would be advisable to conduct a research an equal sample of students of each study field so that more relevant data can be acquired.

Also, it would be advisable to conduct longitudinal research when every surveyed student will be employed to check whether their workplaces have met their expectations or have their expectations changed.

V. CONCLUSION

As a conclusion of the researches, it is visible that a proactive young generation is willing to be treated better, be more productive while keeping the borders of equality and fair treatment. They appreciate sustainability and are against harassment. Their main objectives are to make a workplace their life because they find an acknowledgment in it.

A meaningful and value-driven workplace means having and building trustful, ethical relationships with the work provider, while being interested and passionate about the process so in the end to receive powerful feedback, thus, create a balance between work-home relations. It is a completely new curve in business ethics and economic development, powered by a massive shift to the service economy and globalization. STEAM students no longer wait until the possibility will arise - they create these possibilities. It is a powerful movement of modern society that cannot be mistreated by work provide and participants in the professional world.

By research methodology, we could analyze what are the perceptions and believes modern students have about the situation in the professional worlds of their fields of occupation. We were able to see that they are aware of the challenges and morals inequalities that are present these days at work, as well as they are aware of the possibility of choice and advantages of education and knowledge they get in schools and universities. It is proving, that to become a leader it is not enough just to be able to manage and give orders. People of the modern young generation clearly define borders of their sense of importance, value, and respect, while striving to provide and build a good quality, trustful and ethical products or services. It is a very important thing to notice for employers while identifying values and main goals before creating a company and building a sustainable surrounding for the workers.

To reach a high level of future well-being and balance in life STEAM students these days are learning from the experience of previous generations, from up-to-day world economical and environmental crisis, technological progress, and diverse globalization. Surrounded by a huge flow of information and innovations, young people become more adaptable, flexible, experienced, and smarter than any other generation. However, under such a great pressure of world challenges, today young people are willing to find more than a job and a high salary, they want to have a life-long goal and purpose.

The Millennial workforce is already mobile and with that can bring positives income to an organization during difficult times. Now is the time to capitalize on fresh thinking and new ideas and not inadvertently leave the door open for your best employees to leave. The main benefit of having the younger generation engaged in work is that only this generation combines in themselves all possible knowledge of the past and present, their open-

mindedness and wide overlook on hard cases, that can take any organization on the new level.

REFERENCES

[1] "Living and working in the new economy", OpenLearn, 2020. Available at: https://www.open.edu/openlearn/people-politics-law/politics-policy-people/sociology/living-and-working-the-new-economy/content-section-2.2, [accessed: 31.08.2020].

[2] M. A. Paludi, C.A. Paludi, E. DeSouza, "Praeger Handbook on Understanding and Preventing Workplace Discrimination", Praeger, 2011.

[3] R. Stuart, "Developing the Next Generation", CIPD Learning to Work, Research report, June 2015. Available at: https://www.cipd.co.uk/Images/developing-next-generation_tcm18-10268.pdf [accessed: 18.05.2020].

[4] "Business Ethics across Generations", Business Ethics Briefing, IBE, Issue 48, 2015. Available at: https://static1.squarespace.com/static/518dd156e4b0829608e93ca3/t/5aa1c9aa53450ad3c962e6b6/1520552363963/B48+Ethics+Across+Generations.pdf [accessed: 15.05.2020].

[5] T. Lechner, "Overview of Ester Perel lecture on Workplace Dynamics", Positive Minds International, 2019. Available at: https://positivemindsinternational.com/esther-perels-solved-organizational-relationships/, [accessed: 10.05.2020].

[6] K. K. Myers, K. Sadaghiani, "Millennials in the Workplace: A Communication Perspective on Millennials' Organizational Relationships and Performance", Journal of Business and Psychology, 2010. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2868990/, [accessed: 18.05.2020].

[7] J. M. Twenge, "Generation Me: Why Today's Young Americans Are More Confident, Assertive, Entitled - and More Miserable Than Ever Before", Free Press, New York, 2006.

[8] E. Perel, "Relationship Dynamics in the Workplace", Esther Perel, 2020. Available at: https://estherperel.com/blog/letters-from-esther-4, [accessed: 10.05.2020].

[9] A. Jolly, "STEM vs. STEAM: Do the Arts Belong?", Education Week Teacher, 2014. Available at: https://www.edweek.org/tm/articles/2014/11/18/ctq-jolly-stem-vs-steam.html [accessed: 12.05.2020].

[10] The SHARE Team, "More Than a Trend: Crucial Ways in Which STEAM Shapes Student Thinking". Available at: https://blog.sharetolearn.com/leaders-link/steam-shapes-student-thinking/ [accessed: 12.05.2020].

[11] K. Kee, "From STEM to STEAM: The Future of the Liberal Arts", Faculty of Arts, Ottawa, 2016. Available at: https://arts.uottawa.ca/en/stem-steam-future-liberal-arts6, [accessed: 10.05.2020].

[12] C. Mackie, "STEM And Coronavirus: The Crisis Within The Crisis", Forbes, 2020. Available at: https://www.forbes.com/sites/calvinmackie/2020/04/21/stem-and-coronavirus-the-crisis-within-the-crisis/#19d689aee662 [accessed: 12.05.2020].

Cryptocurrency as the currency of the future: a case study among Algebra University College students

Antea Knežević*, Tihana Babić* and Zvonimir Musa*
*Algebra University College, Zagreb, Croatia
antea.knezevic@racunarstvo.hr, tihana.babic@algebra.hr, zvonimir.musa@racunarstvo.hr

Abstract - **The cryptocurrencies were designed to be a medium of exchange. The blockchain technology on which cryptocurrencies are based on offers many possibilities for computer science and all future businesses. For the past decade experts as well as laypeople have been experiencing cryptocurrencies in extremes. They either have a very positive attitude or a very negative attitude towards them. Experts who have very positive attitudes towards them believe that cryptocurrencies create new ways of conducting business and new ways of trust relationships are managed. Experts who have very negative attitudes towards them often emphasize the fact that they are often linked to negative connotations such as being a tool for criminal activities or skipping social responsibilities such as tax avoidance and corruption. They also emphasize the fact that it is a new, unexplored technology and an unstable market. The purpose of research conducted among students of Algebra University College was to explore their opinions and ways of using blockchain technologies in the present and their attitudes toward using cryptocurrencies in the future.**

Keywords – ***cryptocurrency, blockchain technology, case study, students***

I. INTRODUCTION

Representation of the European Commission in Croatia defines Cryptocurrencies as digital money, created in digital form as a means of digital exchange. Cryptocurrencies exist only on the Internet and have not been published by, or monitored by, a central bank or government. Precisely because cryptocurrencies are not controlled by the central bank, they are not formally money [1].

They are stored in a "digital wallet" on one of the websites that provide this service. Every transaction made is recorded. The general or public "ledger" in which all transactions and value changes of cryptocurrency units are recorded is called "blockchain".

A blockchain is a database that is shared across a network of computers. Once a record has been added to the chain it is very difficult to change. To ensure all the copies of the database are the same, the network makes constant checks [2].

The blockchain offers many possibilities for computer science and all future businesses. For the past decade experts as well as laypeople have been experiencing cryptocurrencies in extremes. They either have a very positive attitude or a very negative attitude towards them. Many kinds of research have been done, most of them showing that young people under the age of 35, mainly known as "millennials" are more prone to working with cryptocurrencies [3]. Because most researches on this topic are conducted in The United States of America, the authors of this paper decided to test this theory among students of Algebra University College.

II. ADVANTAGES OF CRYPTOCURRENCIES

Experts who have very positive attitudes towards cryptocurrencies believe that they create new ways of conducting business and new ways of how trust relationships are managed. The level of technology drives changes in the global economy and, accordingly, the emergence of virtual currencies is an important event in the world of business and finance.

The European Banking Authority has divided the benefits of using virtual currencies into two categories. Those are economic and individual benefits [4]. The economic benefits under the EBA relate to more facts. First, they relate to transaction costs. In principle transaction, costs should not exist because there are no intermediaries in the transaction. However, there are transaction costs, but they are much less compared to other forms of payment. The reason is the costs incurred in regulation, internal control, and similar. Secondly, approximately every ten minutes new blockchain entries are added. The geographic distance of the participants in the transaction does not matter. Regardless of their distance, the same timeline is required to complete the transaction. This characteristic is considered an exceptional advantage in opposition to other payment methods.

The basic individual benefit of cryptocurrencies is online transactions. The use of cryptocurrencies is beneficial to both buyer and seller. Because they are autonomous they have no issuer and no institution to control their circulation, which means the seller is granted full anonymity, which buyers mainly benefit from. A return possibility in a virtual currency transaction is not deductible. When making a payment for a selected product or we cannot get the virtual currency unit back on request. This is where the EBA highlighted the advantage by sellers because, with traditional forms of payment in the case of fraudulent defect reports, funds often would have to be refunded [5].

Except for spending money, Cryptocurrencies have been suggested to be a new and popular form of alternative investment. They hold several important advantages over traditional forms of currency such as independence, security, and price liquidity which can result in something like we experienced with Bitcoin, resulting in a massive surge of interest in cryptocurrency. Anyone who owned Bitcoin at the end of 2017 became very rich overnight. Bitcoin exploded in price, and needless to say, everyone who wasn't a part of that wanted to invest some money in case it happened again. And it did happen again in 2019 [6].

Many cryptocurrency users find it very important that it preserves their privacy. In order to be able to open an account and execute transactions through a bank, you are obliged to provide extensive personal information, but with cryptocurrency, you do not have to provide personal data and the transactions take place anonymously. Although the degree of privacy varies from cryptocurrency to cryptocurrency, most of them guarantee complete anonymity. And you can create as many account numbers or addresses as you want [7].

One of the most interesting and important features of cryptocurrencies is the blockchain technology that it's based on. Blockchain technology makes it possible to digitize every physical object or service and to put its value in the blockchain. For example, cars, houses, web page designs, domain names, etc. can be placed in the blockchain and traded via automatic smart contracts. As a result, the intervention of third parties such as banks, notaries, advisors, and accountants are no longer necessary and transactions can take place faster and cheaper. This can also apply to services such as insurance, loans, and similar. It can be used to manage your personal identity safely in the blockchain and use it, for example, to identify yourself when traveling, verifying the age when purchasing alcohol, etc. Finally, smart contracts can be used to fully automate government tasks, when collecting taxes, issuing permits, paying subsidies, and when holding local, regional, or national elections. In that way, everyone can see exactly what goes in and what comes out and the government can become a lot more transparent [8].

III. DISADVANTAGES OF CRYPTOCURRENCIES

As stated before, the advantage of cryptocurrencies is its anonymity but the downside of anonymity is its lack of confidence. The lack of a competent institution to control and guarantee creates a high risk of the counterparty. In electronic money, the issue of risk and trust is questionable, while in fiat money it is highly correlated with the national economy and politics [8].

While some experts applaud the anonymity feature of cryptocurrencies, many see them as a downside. The reason for that is their susceptibility to criminal purposes resulting from the anonymity of their use. The black market and the dark web are big users of cryptocurrencies. Criminals value their anonymity, as much as they value the ability to send vast sums of money around the world simply using their phone. This made cryptocurrencies relatable to the risk of money laundering, terrorist organizations, and other illegal activity, financing many negative connotations [9].

From a strictly economic view, many experts show a negative attitude towards cryptocurrencies because of their strong volatility and difficulty to predict. Since the beginning, cryptocurrencies had a highly volatile nature. This is one of the main reasons mass adoption is taking longer than it should. Many corporations don't want to deal with a form of money that is going through huge swings in volatility. Even though we had a positive experience of volatility at the end of 2017., when Bitcoin exploded in price, many other cryptocurrencies in use behave the exact opposite, and lose their value fast, creating a risk for potential investments[10].

Except for volatility is the number one risk of crypto investing, experts are also pointing out other potential problems. Regulatory and legal issues are two of the big obstacles facing the crypto sector. Because the technology is new, governments and banks have not yet formed a coherent fiscal policy for them. Therefore, there is a risk that their taxation status, trading rules, or even legality, could change and become more complicated. Also, a concern, although it's never been conclusively proved, it's widely assumed that insider trading, collusion, and market manipulation is present across the crypto sector[11].

And last, but not least, one big cryptocurrency issue is lack of awareness and understanding. It is a known fact that many people are still unaware of digital currencies. To be able to use and invest in cryptocurrencies people need to be educated to be able to apply it to their lives. Businesses are accepting bitcoins because of the advantages, but the list is relatively small compared to physical currencies. One of the reasons is that the workers need to be educated on Bitcoin so that they can help the customers. This will definitely take some time and effort. Otherwise, what is the benefit of such large companies accepting Bitcoin if its staff doesn't even know what digital currencies are[12].

IV. WORLD'S LARGEST CRYPTOCURRENCY BY MARKET CAP

At the time of October 8th, 2019 there were approximately 2,957 cryptocurrencies being traded with a total market capitalization of $221bn. This means that the top 10 cryptocurrencies represent roughly 85% of the total market value. Even though the list of top 10 cryptocurrencies changes very often, the first tree on the list is a constant. Those are Bitcoin, Ethereum, and Ripple.

Being the first one to launch in 2009., **Bitcoin** is the world's largest cryptocurrency by market cap. It is created by an anonymous developer Satoshi Nakamoto, whose true identity is never verified. Bitcoin offers the promise of lower transaction fees than traditional online payment mechanisms and is operated by a decentralized authority, unlike government-issued currencies [13].

Second on the list is **Ethereum**, a blockchain-based open public network platform that enables developers to build and deploy decentralized applications for use by businesses, as well as individual users. It acts as a platform

on which a whole virtual ecosystem can be developed, stored, executed, and used by users securely and anonymously. Unlike commission-based app stores demanding a cut for hosting and running various apps, the decentralized and autonomous nature of Ethereum keeps it zero- to low-cost[14].

Third, on the list is **XRP**. XRP is a currency that is part of the Ripple technology. Ripple is a technology mostly known for its digital payment network and protocol. Ripple's main specialty is payment settlement asset exchange that is similar to the SWIFT system for international money [15].

Over the last year, fourth and fifth place of the list has been exchanged by **Bitcoin Cash** and **Bitcoin SV**. Both of those coins are a result of a hard fork of Bitcoin. A hard fork is a radical change to the protocol of a blockchain network that makes previously invalid blocks/transactions valid or vice-versa[16]. This means both of those coins started as Bitcoin, but a group of miners decided they wanted to set new rules to the code, changing its initial idea and creating their own version of Bitcoin.

Some coins that are or have been in the top 10 cryptocurrencies by market cap are Tether, Litecoin, EOS, Stellar, and Monero.

Tether is a blockchain-based cryptocurrency stable coin whose crypto coins in circulation are backed by an equivalent amount of traditional fiat currencies, like the dollar, euro, or yen. It is always worth 1 American dollar[17]. **Litecoin** is a peer-to-peer cryptocurrency. It's a fully decentralized open-source, working as a global payment network developed with the aim to improve Bitcoin's shortcomings[18]. **EOS** is a decentralized system based on the blockchain that enables the development, hosting, and execution of decentralized applications on its platform, similar to Ethereum [19]. **Stellar** is an open-source payment technology, very similar to Ripple[20]. **Monero** is a digital currency that offers a high level of anonymity for users and their transactions. It has several privacy-enhancing features that improve upon Bitcoin [21].

One crypto token that hasn't made the top 10 list yet, but created a lot of attention over the last year is the **LINK** token with the Chainlink technology behind it. ChainLink is a platform that attempts to bridge the gap between smart contracts on the blockchain and real-world applications of the blockchain. It uses oracles, which find and verify real-world data and bring it on-chain to be integrated into smart contracts [22].

V. STUDENTS AND CRYPTOCURRENCIES

A survey conducted in the United States of America in August of 2019, questioned members of Millennials to question who had some information or knowledge on cryptocurrencies. Research showed 98.4 percent of them were familiar with Bitcoin, 77.7 percent with Ethereum, 71.6 with Litecoin, and 47.6 percent with Ripple. Most surprising of all, 21 percent of the respondents did not have access to traditional banking of any kind, which makes cryptocurrency a more likely and popular alternative to traditional banking among young adults and students. The reasons for such a huge popularity of crypto among young people are suspected to be technological advancements, and lack of faith in existing traditional banking systems [3].

VI. THE RESEARCH METHODOLOGY

A. Purpose of the Research

A survey conducted in the United States of America showed that cryptocurrencies were used at a high rate among young people (millennials). Testing that theory, the purpose of research conducted among students of Algebra University College was to explore their opinions and ways of using blockchain technologies in the present and their attitudes toward using cryptocurrencies in the future.

The research objective: *Do students from Algebra University College use cryptocurrencies?*

To answer the main research objective, the following research sub-questions were defined:

1. *Are students familiar with the concept and technology of cryptocurrency?*
2. *What characteristics of cryptocurrency do students consider to be their advantages?*
3. *What characteristics of cryptocurrency do students consider to be their disadvantages?*
4. *Do students use cryptocurrencies and / or plan to invest in them in the future?*

B. The Research Population Sample

The research was conducted among the Algebra University College students. A sample of participants was pertinent.

The total population sample was 64 participants (N = 64), which included students from 7 different study programs: Software engineering, System engineering, Digital marketing, Multimedia computing, Visual communications design, Design and communication management, and Data Science. The structure of all surveyed students according to the study program is presented in Chart 1.

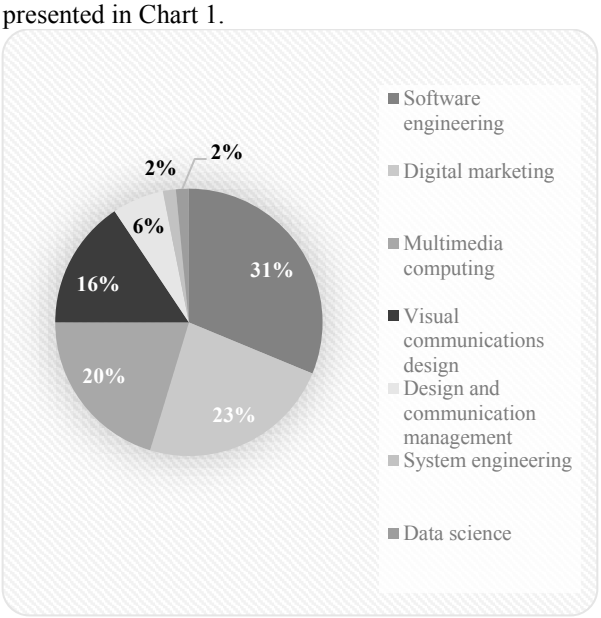


Chart 1. The structure of the participants per Study Program in percentage, N = 64.

The majority of participants were undergraduate students; 53 of them (82.8 %), while 11 (17.2 %) of participants were graduate students. 39 of students (60.9 %) were male and 25 (39.1 %) were female.

C. The Research Methods

The research was conducted through an anonymous voluntary survey during the winter semester of the 2019/2020 academic year.

A specially designed questionnaire in the Google Forms tool had 11 closed-ended questions, of which 4 were related to demographics. For 3 questions answers were defined with the degrees of frequency according to the Likert scale. 2 questions had predefined 10 statements related to the advantages and disadvantages of using cryptocurrencies. 3 questions were yes or no questions. 1 question offered the most famous cryptocurrencies for which students were able to indicate the degree of familiarity and agreement according to the Likert frequency scale.

To ensure a clear understanding of the terms, a descriptive definition of term cryptocurrency was specified at the beginning of the survey:

"A brief explanation of the terms:

Cryptocurrency is digital money, created in digital form as a means of digital exchange. It exists only on the Internet and has not been published, or monitored by, a central bank or government. Precisely because cryptocurrencies are not controlled by the central bank, they are not formally money.

They are stored in a "digital wallet" on one of the websites that provide this service. Every transaction that is made is a highly edited digital record. The general or public "ledger" in which all transactions and value changes of cryptocurrency units are recorded is called "blockchain"."

A quantitative method was used for the analysis of the research results. Through the Google Forms tool, the data processing of the survey results was made.

D. The Research Results

1. *Are students familiar with the concept and technology of cryptocurrency?*

The research results have shown that most of the total number of surveyed students (87.5%) are familiar with the concept of cryptocurrencies. But just over half of the students (51.6%) declared themselves familiar with the blockchain technologies on which cryptocurrencies are based on.

When it comes to particular cryptocurrencies, specifically the top 10 most valued, research results have shown that almost all of the total number of surveyed students (63) are familiar with Bitcoin. Less than half of the participants are familiar with Ethereum (28), Ripple (18) and Litecoin (17). The participants are poorly familiar with Monero (9), Chainlink (7), Tether (6), EOS (6), and Stellar (5).

The structure of student familiarity with cryptocurrencies is presented in Table 1.

CRYPTOCURRENCIES	Never	Rarely	Some-times	Often	Always
Bitcoin	0.0%	0.0%	1.6%	29.7%	68.8%
Ethereum	42.2%	3.1%	9.4%	17.2%	26.6%
Ripple	51.6%	4.7%	14.1%	15.6%	12.5%
Tether	57.8%	17.2%	12.5%	3.1%	6.6%
Litecoin	45.3%	21.9%	3.1%	10.9%	15.6%
EOS	56.3%	20.3%	9.4%	3.1%	6.6%
Stellar	60.9%	20.3%	6.6%	3.1%	4.7%
Chainlink	65.6%	17.2%	6.6%	3.1%	7.8%
Monero	62.5%	7.8%	12.5%	4.7%	9.4%

Table 1. Distribution of the students' familiarity with cryptocurrencies, N=64.

2. *What characteristics of cryptocurrency do students consider to be their advantages?*

In terms of advantages, 27 of the total number of students questioned has agreed that cryptocurrencies are a good investment, while 26 stayed impartial. A total of 19 students agreed that cryptocurrencies are a good way of saving (28 stayed impartial), but merely 14 agreed that cryptocurrencies are useful for protecting against the loss of value of traditional goods (27 impartial). Half of the students (32) agreed that cryptocurrencies are useful for online payments (24 impartial). Also, 28 agreed that cryptocurrencies are a safe way of paying (25 impartial). A total of 22 respondents agree that cryptocurrencies might replace money in the future (24 impartial). Half of the respondents (32) agree transaction transparency is a very important feature of cryptocurrencies (24 impartial) and 33 of them agreed that the strength of cryptocurrencies lies in the possibility to operate cryptocurrency transactions every day from 0 to 24 h, meaning that they provide better services than the banks that have this type of service limited (20 impartial). To the question of control, 27 respondents agree cryptocurrencies allow complete autonomy (25 impartial), but to the question of cost 33 neither agree nor disagree with the statement that cryptocurrencies offer lower transaction costs. And last, half of the respondents (32) agree there is a large selection of cryptocurrencies for different purposes (30 impartial).

The structure that represents the students' degree of agreement according to the Likert frequency scale is presented in Table 2.

The statements represented by letters in the table are as following:

A. *Cryptocurrencies are useful for investment purposes.*
B. *Cryptocurrencies are useful for saving purposes.*
C. *Cryptocurrencies are useful for protecting against the loss of value of traditional goods.*
D. *Cryptocurrencies are useful for buying online.*

E. *Cryptocurrencies are a secure payment method.*
F. *Cryptocurrency will someday replace the money.*
G. *Transaction transparency is a very important feature of cryptocurrency.*
H. *It is possible to operate cryptocurrency transactions every day from 0 to 24 h, which means that they provide better services than the banks that have this type of service limited.*
I. *Cryptocurrencies allow us to complete autonomy.*
J. *The cost of cryptocurrency transactions is minimal.*
K. *There is a large selection of cryptocurrencies for different purposes.*

STATEMENT	Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
A.	4.7%	9.4%	40.6%	31.3%	10.9%
B.	9.4%	14.1%	43.8%	20.3%	9.4%
C.	6.3%	26.6%	42.1%	10.9%	10.9%
D.	3.1%	7.8%	37.5%	23.4%	26.6%
E.	6.3%	7.8%	39.1%	23.4%	20.3%
F.	6.3%	18.8%	37.5%	21.9%	12.5%
G.	3.1%	7.8%	37.5%	26.6%	23.4%
H.	6.3%	6.3%	31.3%	26.6%	25.0%
I.	3.1%	14.1%	39.1%	25.0%	17.2%
J.	4.7%	4.7%	51,6%	14.1%	23.4%
K.	1.6%	3.1%	46.9%	21.9%	25.0%

Table 2. Distribution of the agreement degree according to the Likert frequency scale.

3. *What characteristics of cryptocurrency do students consider to be their disadvantages?*

In terms of disadvantages, 31 of the total number of students questioned has agreed that the technology behind cryptocurrencies is new and has not yet been fully developed (22 impartial). Most of the respondents (38) agree the cryptocurrency technology is difficult for laymen to understand (20 impartial). To the question of loss, 25 of the respondents neither agree nor disagree with the statement that there is no security in case of losing funds. 35 of respondents agree that the cryptocurrency market is unregulated, and even 47 agree cryptocurrencies have not yet been accepted as a payment method in most places, of which 24 strongly agree. Students agree (29) with the statement that cryptocurrencies make it easier to evade the law, and more than half (35) agree the cryptocurrency market is volatile. Most of the respondents (24) neither agree nor disagree with the statement that the cryptocurrency market enables online theft. And at last 27 of respondents agree cryptocurrencies support unrealistic human views about easy ways of making money without effort.

The structure that represents the students' degree of disagreement according to the Likert frequency scale is presented in Table 3.
The statements represented by letters in the table are as following:

A. *Cryptocurrency technology is new and has not yet been fully developed.*
B. *Cryptocurrency technology is difficult for laymen to understand.*
C. *There is no security in case of losing funds.*
D. *The cryptocurrency market is unregulated.*
E. *In most places, cryptocurrencies have not yet been accepted as a payment method.*
F. *Cryptocurrencies are useful for the avoidance of taxes and other government regulations.*
G. *Cryptocurrencies make it easier to evade the law.*
H. *The cryptocurrency market is volatile.*
I. *The cryptocurrency market enables online theft.*
J. *Cryptocurrencies support unrealistic human views about easy ways of making money without effort.*

STATEMENT	Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
A.	1.6%	14.1%	34.4%	35.9%	12.5%
B.	0.0%	7.8%	31.3%	32.8%	26.6%
C.	0.0%	7.8%	39.1%	28.3%	23.4%
D.	0.0%	7.8%	34.4%	34.4%	20.3%
E.	1,6%	1.6%	21.9%	35.9%	37.5%
F.	3.1%	1.6%	42.2%	29.7%	20.3%
G.	0.0%	1.6%	50.0%	28.1%	17.2%
H.	1.6%	4.7%	37.5%	32.8%	21.9%
I.	7.8%	12.5%	45.3%	17.2%	15.6%
J.	7.8%	7.8%	40.6%	21.9%	20.3%

Table 3. Distribution of the agreement degree according to the Likert frequency scale.

4. *Do students use cryptocurrencies and / or plan to invest in them in the future?*

The research results have shown that most of the total number of surveyed students (73.4 %) didn't own any kind of cryptocurrency at the moment of taking this survey and but almost half of the respondents (43.75%) answered that they will invest in cryptocurrencies in the future, of which 17.5% answered that they will surely invest. 25% of respondents answered that they probably won't invest in cryptocurrencies, of which only 6.3% answered they surely won't invest. 30.2% of respondents are not sure if they will invest.

The students' predictions on their future investing in cryptocurrencies are presented in Chart 2.

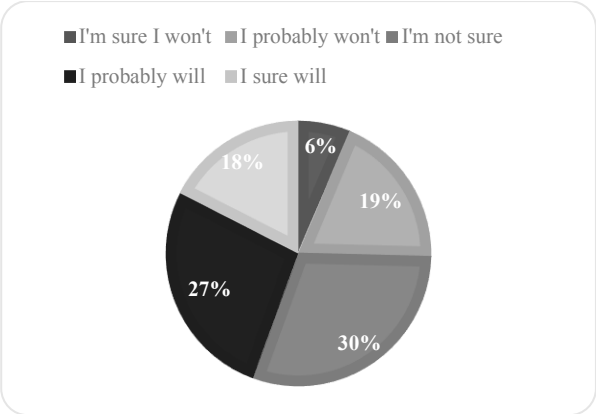


Chart 2. The structure of student's predictions in future cryptocurrency investing, N = 64.

VII. CONCLUSION

Most of the research done on young people under the age of 35, known as "millennials" has been done in the United States of America and shows that they are more prone to working with cryptocurrencies than the rest of the population. The research also shows that using cryptocurrencies is very normal for them and that most of them use it like regular payment options.

After conducting the same research in Croatia among students from one of Croatians leading IT studies, the research result shows that Croatian "millennials" are not there yet. Even though most of them are familiar with Bitcoin, only 26.6% of them own any kind of cryptocurrency. Also less than 50% of the total number of surveyed students planes to invest in it in the future as they find the technology new and hard to understand and the market unregulated and volatile.

This research was conducted before the outbreak of the Covid-19 pandemic which for Croatia escalated in the second half of March 2020. It would be interesting to repeat the research in one year to examine whether under the influence of new circumstances the students changed their opinions.

REFERENCES

[1] Europska komisija, Predstavništvo u Republici Hrvatskoj, "Kriptovalute i blockchain – sve što trebate znati", 2020. Available at: https://ec.europa.eu/croatia/cryptocurrencies_and_blockchain_all_you_need_to_know_hr [accessed: 01.02.2020].

[2] M. Murray, "Blockchain explained", Graphics Reuters, June 2018. Available at: http://graphics.reuters.com/TECHNOLOGY-BLOCKCHAIN/010070P11GN/index.html [accessed: 04.02.2020].

[3] Bitcoin warrior "Why is crypto more popular with young people?", February 2020. Available at: https://bitcoinwarrior.net/2020/02/why-is-crypto-more-popular-with-young-people/ [accessed: 08.02.2020].

[4] European Central Bank "Virtual Currency Schemes", October 2012. Available at: https://www.ecb.europa.eu/pub/pdf/other/virtualcurrencyschemes201210en.pdf [accessed: 09.02.2020].

[5] D. Vrkić, "Značaj i izazovi virtualnih valuta: studija slučaja Bitcoin", završni rad urn:nbn:hr:124:691787, September 2018. Available at: https://repozitorij.efst.unist.hr/islandora/object/efst%3A2282/datastream/PDF/view [accessed: 10.02.2020].

[6] J. Almeda, "Are Cryptocurrencies a Good Alternative Investment?" , Hackernoon, November 2019. Available at: ahttps://hackernoon.com/are-cryptocurrencies-a-good-alternative-investment-agr32a8 [accessed: 10.02.2020].

[7] Mobisun, "10 benefits of paying with cryptocurrencies", January 2019. Available at: https://www.mobisun.com/blog/10-benefits-of-paying-with-cryptocurrencies/ [accessed: 03.02.2020].

[8] G. Vora, "Cryptocurrencies: Are Disruptive Financial Innovations Here?" July 2015. Available at:http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1028.1690&rep=rep1&type=pdf [accessed: 12.02.2020].

[9] F. Bunjaku, O. Gjorgieva-Trajkovska, E. MitevaKacarski, "Cryptocurrencies – Advantages And Disadvantages" ISSN 1857-9973. Available at: http://eprints.ugd.edu.mk/18707/1/Cryptocurrencies.pdf [accessed: 13.02.2020].

[10] H. Hassani, X. Huang, E. Silva, "Big-Crypto: Big Data, Blockchain, and Cryptocurrency", Researchgate, October 2018. Available at: https://www.researchgate.net/publication/328407542_Big-Crypto_Big_Data_Blockchain_and_Cryptocurrency [accessed: 08.02.2020].

[11] S. Boukhalfa, "What are the disadvantages of cryptocurrencies?", Prescouter, November 2019. Available at: https://www.prescouter.com/2019/11/disadvantages-of-cryptocurrencies/ [accessed: 04.02.2020].

[12] Coinreport, "Advantages and disadvantages of bitcoin". Available at: https://coinreport.net/coin-101/advantages-and-disadvantages-of-bitcoin/ [accessed: 06.02.2020].

[13] J. Frankenfield, "Bitcoin", Investopedia, Available at: https://www.investopedia.com/terms/b/bitcoin.asp [accessed: 13.02.2020].

[14] S. Seth, "All about Ethereum" Investopedia, Available at: https://www.investopedia.com/tech/all-about-ethereum/ [accessed: 13.02.2020].

[15] J. Frankenfield, "Ripple (Cryptocurrency)", Investopedia, Available at: https://www.investopedia.com/terms/r/ripple-cryptocurrency.asp [accessed: 13.02.2020].

[16] J. Frankenfield, "Hard Fork (Blockchain)", Investopedia. Available at:https://www.investopedia.com/terms/h/hard-fork.asp [accessed: 13.02.2020].

[17] J. Frankenfield, "Tether (USDT)", Investopedia. Available at: https://www.investopedia.com/terms/t/tether-usdt.asp [accessed: 13.02.2020].

[18] J. Frankenfield, "Litecoin", Investopedia. Available at: https://www.investopedia.com/terms/l/litecoin.asp [accessed: 13.02.2020].

[19] S. Seth, "What Is EOS?", Investopedia. Available at: https://www.investopedia.com/tech/what-is-eos/ [accessed: 13.02.2020].

[20] R. Sharma, "What Is Stellar Blockchain?", Investopedia. Available at: https://www.investopedia.com/news/what-stellar/ [accessed: 13.02.2020].

[21] J. Frankenfield, "Monero", Investopedia. Available at: https://www.investopedia.com/terms/m/monero.asp [accessed: 13.02.2020].

[22] K. C. Tran, "What is Chainlink?", Decrypt, November 2019. Available at: https://decrypt.co/resources/what-is-chainlink [accessed: 13.02.2020].

New teaching methods in higher education - Management of Information Systems course

K. Aleksić-Maslać*,**, P. Vranešić * and B. Debić*

*Zagreb School of Economics and Management/Department of Management, Entrepreneurship and Digital transformation, Zagreb, Croatia
** Luxembourg School of Business, Luxembourg

kaleksic@zsem.hr
pvranesi@zsem.hr
boris@debic.net

***Abstract* - This paper presents the usage of modern teaching methods such as e-learning, gamification, case studies, teamwork and projects on the Management of Information Systems (MIS) course. As the course is held in the 7th semester of the undergraduate program on Zagreb School of Economics and Management (ZSEM), it is necessary to provide interesting and engaging materials and lectures for students that are soon entering the labor market. An important part of the course is covered by student projects which focus on information system (IS) or ERP (Enterprise Resource Planning system) implementation in companies. This aspect of the course develops positive competitive motivation, teamwork and presentational skills. Other examples of engaging elements implemented in this course are gamification, where students learn through playing, and e-learning component for distance learning, very useful for students that are on an exchange program abroad, athletes, students that work, etc. All course elements (project, discussion, cases, gamification, e-learning, etc.) present their own goals and learning outcomes. The aim is to present a course with important engaging elements in today's education sector with a combination of powerful learning methods and tools. The research will focus on the result analysis of the last 4 years of course practice.**

Keywords – Course, e-learning, gamification, project

I. INTRODUCTION

The usage of the new technologies embraces big possibilities for content development in order to motivate and engage students in class. By using e-learning, students are not bound to one place and time, but have the ability to choose their best suited time [1]-[3]. Learning Management Systems (LMS) helps students to reach lecture materials, assignments, quizzes for knowledge checking, discussions and other learning content [4]. A very important part of every e-learning system is discussions – which enable stakeholders to communicate with no distance problem [5]-[8]. In the paper "Important role of asynchronous discussion in e-learning system", authors have defined opened and closed discussions. [9] Opened discussions are optional between professors and students, students and professor, and with student to student, while closed discussions are related to lecture topics and are obligation. The emphasis goes to the student-student discussions as they are more relaxed when discussing something without the presence of the professor.

Different researches are dealing with e-learning quality system, whether from instructors' perspective [10,11] or users' perspective [12,13]. In "E-Learning Course Development – Quality Standards", authors have developed 11 standards by evaluating the quality of the e-learning system. [14]

Also, gamification has been used a lot more lately in order to motivate students to be more engaging in class. [15]-[19] Researches show that students like to use gamification, regardless of age, however, extrinsic motivation is more expressed with younger students as they have a bigger need for some kind of a reward. [20, 21] Gamification is very popular in both class learning (CL) and distance learning (DL). [22, 23]

Since students are learning how to prepare for the labour market, it is essential to help them develop the necessary skills, such as teamwork [24]-[26], presentational skills [27], and other.

This paper shows how different elements are used on Zagreb School of Economics and Management (ZSEM) course, part of the 7th semester – Management of Information Systems (MIS). [28] The course has a maximum score in the universities' e-learning evaluation process, gamification is used in various forms and it expects, promotes and develops teamwork skills among students.

The aim of this research is to analyze all course elements and explore how they affect student knowledge and the final grade by using modern teaching methods, whether CL or DL. The data set is made out of data collection of the last 4 years of MIS practice. Within the data set, there are grade records of 316 students, out of which 262 students were part of traditional CL, and 54 were part of the DL as they were either an athlete, exchange or worked at the same time as he/she studied. As

the grade elements are based on different learning outcomes, appropriate teaching methods are essential in order to affect student's engagement properly.

II. COURSE ELEMENTS

MIS is a course that is held in the 7th semester of the undergraduate program on Zagreb School of Economics and Management (ZSEM). The main focus of the course is to teach students how to plan and execute an Enterprise Resource Planning system (ERP) or any other information system (IS) implementation for a business. Students have a possibility to project fictive companies or real businesses.

Table I shows course elements of the first and last MIS generation. The first generation was in the academic year of 2005/2006, and the last generation was academic year of 2019/2020. The first noticeable change throughout the years is that the mid-term exams were reduced from 3 to 2 exams per semester. The reason for this action was to promote interactional activities, such as teamwork in case studies and especially in the project that grew from 15% of the grade to 30% of the grade. This means that students needed to do more practical assignments then just learning theory. Due to the tech development, other elements of the course were able to grow as well, such as gamification and highly developed online forum on the universities LMS called Moodle Loomen.

TABLE I. GRADE ELEMENTS COMPARISON OF THE FIRST AND LAST MIS GENERATION

Regular Syllabus Elements	2005/2006 (%)	2019/2020 (%)
Mid-term exam 1	12	13
Mid-term exam 2	12	13
Mid-term exam 3	12	
Case	10	14
Exercises	14	5
Project	15	30
Class Activity	5	5
Final Exam	20	20
Additional Activities		
Student Presentation	≤ 5	
Forum Activity	≤ 5%	
Forum 1 Activity		≤ 5%
Forum 2 Activity		≤ 5%
Gamification		
Online quiz		

A. Project on MIS

The main focus of the course lies in the project, the ERP or IS implementation of a fictive or real company.

The project is done by a team of 3-5 students and the development lasts 7-8 weeks within the course. Every team has a leader which is responsible for assigning everyone's role and for leading the project according to plan. The whole project is led within the system, such as Microsoft Project. Every week, students are expected to present their continuous work and each member of the team should present at least once. Weekly MS Project assignments are submitted through LMS Moodle Loomen where they are graded – or deducted 0,5% of the points if the deadline criteria are not met. Any student can be a member of any group – even the DL students which report through an online video call, Skype. At the end of the semester, a student conference is held and only the top five projects of the generation have the opportunity to present in front of other students, professors and business community representatives of that field, such as SAP, Oracle, etc. The last conference was the 11th Student MIS conference, held in the academic year of 19/20.

This aspect of the course develops positive competitive motivation, teamwork and presentational skills. The aspect of the team spirit is also addressed through the game of the Marshmallow challenge. [29]

B. Online activity on MIS

Online forums on Moodle Loomen enables students to discuss further interesting topics outside of classroom, which is an important part of every e-learning system. On MIS, there are three different forum discussions:

Professor-student – professor opens the topic of the discussion, students comment

Student-student – students are moderators and discuss among themselves

Quizzes – professor opens a question at any time (not scheduled) and the fastest student with the correct answer is rewarded with points – enabling motivation and engagement in the course.

At the beginning of the course, LMS used was WebCT that had developed forums for teams to communicate about the project. Nowadays, students are using different channels, such as WhatsApp, Messenger, etc. The number of discussions depends on the student activity within the semester. In the professor-student and student-student discussions the points depend on the quality of the content and receive a maximum of 5% of the whole grade percentage. The quiz is not limited.

C. Gamification on MIS

As part of the course, in order to revise the lectures with students, Kahoot is used as a gamification tool. [30] It is a effective tool to simply motivate students to learn by playing knowledge games – and having fun while doing it. When Kahoot was firstly implemented in the course, all students were given points, regardless of the knowledge shown, in order to motivate them towards this new teaching method. However, the Kahoot system has been upgraded in 19/20 by adding 2 criteria's – the average of the whole group needs to be above 50% with correctness of the answers; and because the first 5 students get extra points, the limit on incorrect answers is set, based on the given Kahoot.

In the last academic year, a new gamification tool was added to the course – badge. It was set on other elements such as the assignments and online discussions.

D. Other elements on MIS

In order to examine the students regarding the course lectures, two midterm exams are set on Moodle Loomen so student can immediately see their results. CL students have a scheduled exam in the computer labs of the university, while DL students do it online.

In addition to projects, teamwork is also encouraged through 4-5 case studies with different business cases which are provided during the semester. CL students do it in class, while DL students submit an assignment online.

Lab work is considered as an assignment that is uploaded on Moodle. The defined deadlines encourage responsibility in students for continuous learning. If the deadline is not met, points are deducted. If there is a certain error in the student work, feedback is sent after the submission assessment. Professors easily extract the data from the system, along with other elements.

Presentational skills are encouraged through weekly and final project presentations, through discussions on case studies and through individual presentations (related to the field of the course).

III. RESEARCH RESULTS

The aim of this research is to analyze all course elements and explore how they affect student knowledge and the final grade by using modern teaching methods, whether CL or DL. The data set is made out of data collection of the last 4 years of MIS practice. Within the data set, there are grade records of 316 students, out of which 262 students were part of traditional CL, and 54 were part of the DL as they were either an athlete, exchange or worked at the same time as he/she studied.

As the grade elements are based on different learning outcomes, appropriate teaching methods are essential in order to affect student's engagement properly. Table II shows the anova table between CL and DL students regarding student Project scores, as it shows that there is no significant difference between the groups.

TABLE II. ANOVA TABLE FOR PROJECTS - CL AND DL

SUMMARY

Groups	*N*	*Sum*	*Mean*	*Variance*
DL	54.00	1451.50	26.88	26.09
CL	262.00	7267.90	27.74	40.79

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	33.15	1.00	33.15	0.87	0.35	3.87
Within Groups	12028.02	314.00	38.31			
Total	12061.16	315.00				

In order to gain some further insight of the MIS data, correlation analysis within the data set was made.

As shown in the Table III, the correlation was made between 13 variables: Test_1, Test_2, Lab_work, Case_study, Project, Activity, Presentation, Forum_1, Forum_2, Kahoot, Quiz, Badge, Grade. The correlation results show us that every element has some kind of correlation with the final grade, meaning that every element plays its necessary role in forming the final grade as a result based on the engagement level of students.

The highest correlation with the grade is Test_1 with a correlation coefficient of 0,63, followed by Project with (0,57). Other correlations are also considered moderate, with Presentation (0,47), Test_2 (0,47), Case_study (0,46), Lab_work (0,43), Activity (0,39), Kahoot (0,38) and Badge (0,36). Only Quiz has shown low correlation coefficient of 0,23, and only Forum_2 has no correlation (0,08). To interpret this, Project was expected to be of high correlation as it is the main part of the course, however, other elements of online activity and gamification have shown that they play a moderate support role in the learning process. Forum_2 was also, as expected, of no correlation as students are admins in those discussions.

TABLE III. CORRELATION TABLE OF ALL MIS ELEMENTS

	Test_1	Test_2	Lab_work	Case_study	Project	Activity	Presentation	Forum_1	Forum_2	Kahoot	Quiz	Badge	Grade
Test_1	1												
Test_2	0.33	1											
Lab_work	0.36	0.25	1										
Case_study	0.41	0.14	0.28	1									
Project	0.30	0.21	0.36	0.20	1								
Activity	0.31	-0.02	0.33	0.47	0.26	1							
Presentation	0.12	-0.02	0.23	0.19	0.31	0.30	1						
Forum_1	0.13	0.14	0.18	0.17	0.18	0.09	0.22	1					
Forum_2	0.01	-0.04	0.07	0.02	0.04	0.08	0.02	0.23	1				
Kahoot	0.24	-0.05	0.23	0.31	0.36	0.52	0.31	0.05	0.08	1			
Quiz	0.09	0.08	0.08	0.06	0.12	0.15	0.24	0.12	0.01	0.23	1		
Badge	0.11	0.18	0.24	0.20	0.14	0.10	0.28	0.38	0.10	0.06	0.07	1	
Grade	0.63	0.47	0.43	0.46	0.57	0.39	0.47	0.30	0.08	0.38	0.23	0.36	1

A. *Project and Grade*

Further analysis is made to determine the quality of both CL and DL components. Figure 1 shows a scatterplot of two variables - student Project success and Grade, to display their relationship in more details. It shows a clear rising line as students that appreciate team-work, engagement and responsibility given to them, are rewarded for their accepted workload of the project – and vice versa. Students that exceled in the project had better grades at the end of the semester than the students who didn't.

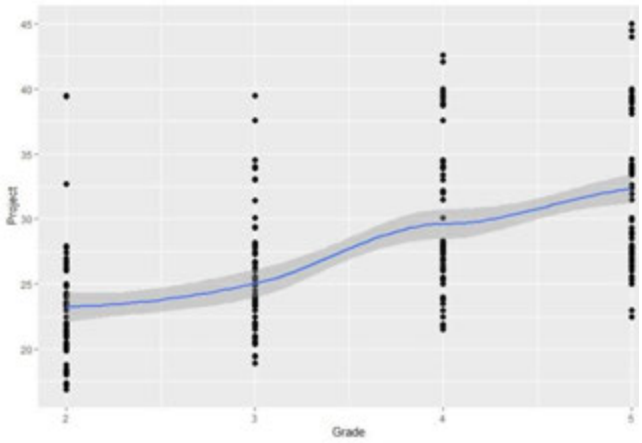


Figure 1. Scatterplot of Project and Grade

A boxplot, in Figure 2, is created to visualize Project results and to see if there are any differences in student points between CL and DL groups. As the anova table reveiled, students of both groups get the same student workload and grading condition. The mean of CL students is 27,74, while DL students have a mean of 26,88.

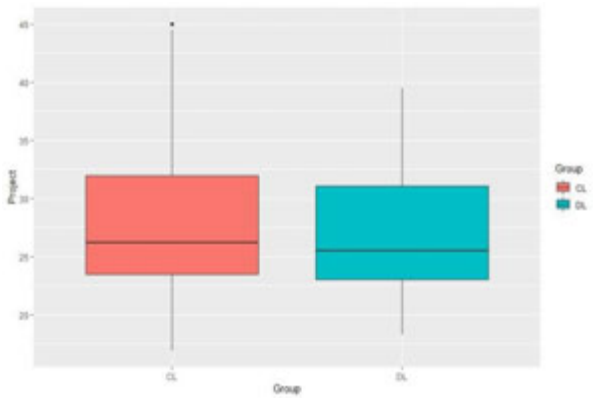


Figure 2. Boxplot of Project and Grade

B. *Gamification and Grade*

Kahoot and Badge scores are part of the gamification aspect. In Figure 3, the line is rising as it shows that students that are continuously engaged and prepared throughout the semester – are earning better knowledge and final grades.

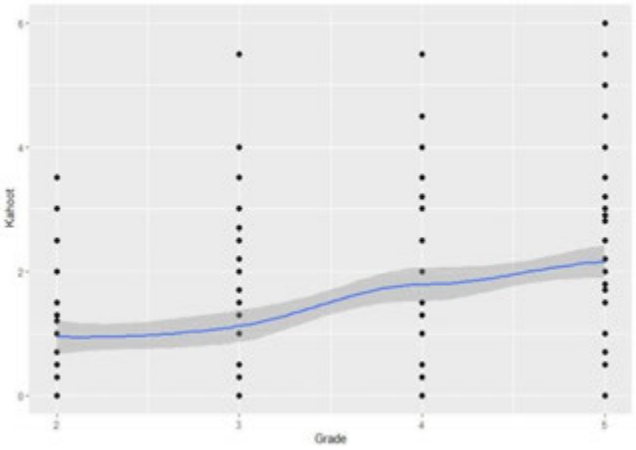


Figure 3. Scatterplot of Kahoot and Grade

As Badge was only used the last year of MIS, it doesn't provide as much as data regarding to other elements. However, in Figure 4, it is shown how couple of students that earned more points in badge have a better grade as well, meaning students that are engaged and are getting the best out of the course, are also using gamification tool for easier learning.

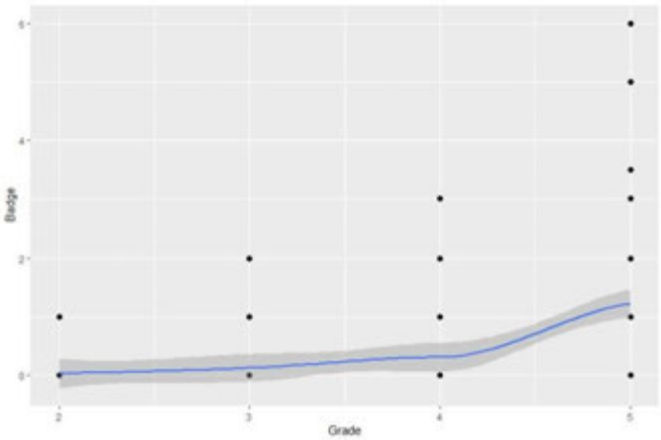


Figure 4. Scatterplot of Badge and Grade

C. *Online activity and Grade*

In the online activity aspect of the course, Activity, Forum 1 and Quiz elements show similar curve of the line - the points of students are rising as the better grade is formed. Figure 5 shows the scatterplot of student Activity, depending on the final grade. There are barley couple of students that got no points from any activity provided to them, and their best grade was good (3), as they were not continuously active and engaged during the semester.

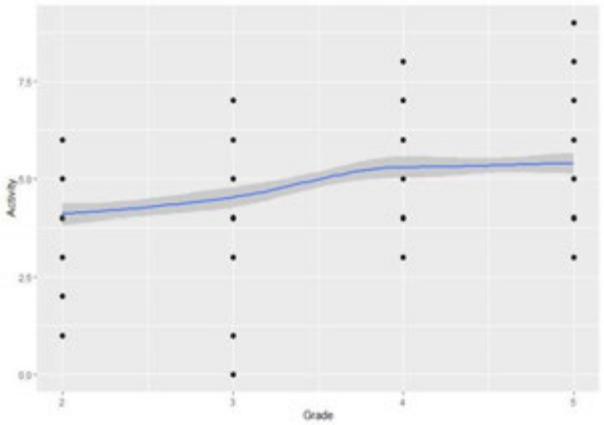


Figure 5. Scatterplot of Activity and Grade

Figure 6 shows a scatterplot for Forum_1 activity, where it is also visible how the line arises as the grades are higher.

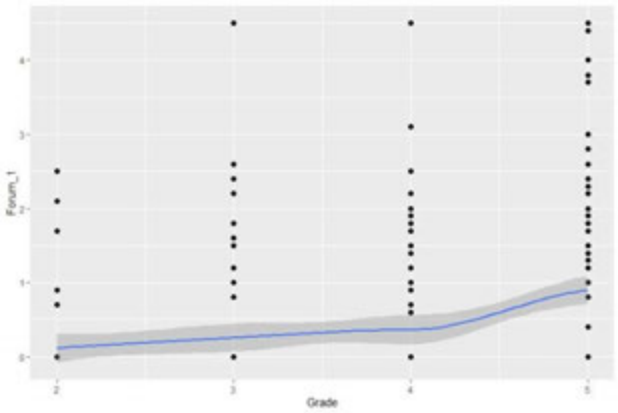


Figure 6. Scatterplot of Forum_1 and Grade

Figure 7 shows a lot less data points of the Quiz to Grade correlation scatterplot as only couple of students have the opportunity to get additional points by quickly providing the correct answer.

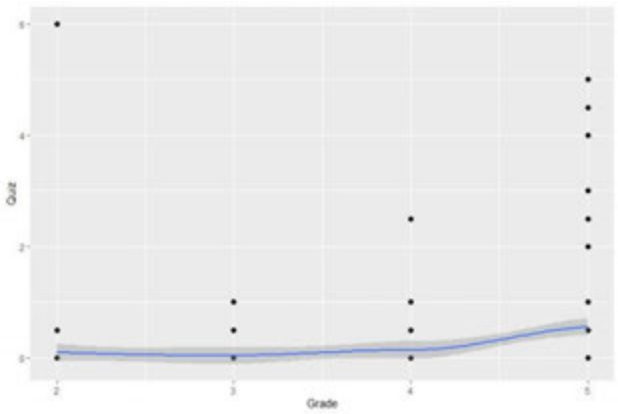


Figure 7. Scatterplot of Quiz and Grade

D. Other elements and Grade

The scatterplot for other elements has the same pattern as the correlations are positive with the grade.

It is authors constant development that adds value to the course with new teaching methods and tools that enable new elements for students to attack and gain the necessary knowledge and skills for the future labour market.

IV. CONCLUSION

The goal of the paper is to provide a course structure example with implications to the new teaching methods. It also shows different aspects of gamification usage, Elearning setup, team work assignments and case studies, projects and other elements to secure appropriate and meaningful student workload.

The results show that there is a moderate correlation of every element with the formed students' grade. The 4-year data set shows the quality of the grading system and student workload stability on MIS, whether CL or DL, enabling students to learn and progress at any point.

The rapid growth of technology has changed the educational system, it is only natural to change the course syllabus according to today's needs. Students are in need of interactive and constant learning to gain the necessary knowledge and skills for the future labour market.

For future research, authors will try to implement new LMS plugins for encouraging gamification, as well as continuously developing the badge system.

REFERENCES

[1] K. Kabassi, I. Dragonas, A. Ntouzevits, T. Pomonis, G. Papastathopoulos, and Y. Vozaitis, "Evaluating a learning management system for blended learning in Greek higher education, Springerplus, 2016.

[2] D. R. Garrison, "E-Learning in the 21st Century: A framework for research and practice", Taylor&Francis Group, 2011.

[3] D. Al-Fraihat, M. Joy, R. Masa'deh, and J. Sinclair, "Evaluating E-learning systems success: An empirical study", Computers in human behavior, volume 102, pp 67-86, 2020.

[4] K. Fertalj, N. Hoic-Bozic, and H. Jerkovic, "The Integration of Learning Object Repositories and Learning Management Systems", Computer Science and Information Systems, 7, 2010, pp. 387-407.

[5] M. Hammond, "A Review of Recent Papers on Online Discussion in Teaching and Learning in Higher Education", Journal of Asynchronous Learning Networks (JALN), Volume 9, Issue 3, October, 2005.

[6] K. Meyer, "Evaluating Online Discussions: Four Different Frames of Analysis", Journal of Asynchronous Learning Networks (JALN), Volume 8, Issue 2 – April, 2004.

[7] Y. Steimberg, J. Ram, R. Nachmia, and A. Eshel, "An online discussion for supporting students in preparation for a test", Journal of Asynchronous Learning Networks (JALN), Volume 10, Issue 4, December, 2006

[8] O. Calderon and C. Sood, "Evaluating learning outcomes of an asynchronous online discussion assignment: a post-priori content analysis", Interactive Learning Environments, Volume 28, Issue 1, pp 3-17, January 2020.

[9] K. Aleksic-Maslac, M. Korican, and D. Njavro, "Important Role of Asynchronous Discussion in E-Learning System", International Conference on Eng. Education and Research 2007 (ICEER 2007), Melbourne, Dec. 02-07, 2007.

[10] E. B. Cohen, and M. Nycz, “Learning Objects and E-Learning: an Informing Science Perspective”, Interdisciplinary Journal of Knowledge and Learning Objects, Vol. 2, 2006.

[11] J. Frydenberg, “Quality Standards in E-Learning: A matrix of analysis”, International Review of Research in Open and Distance Learning, Vol. 3, No. 2, 2002.

[12] U. Ehlers, “Quality in E-learning from a learner's perspective”, EURODL (European Journal of Open, Distance and E-Learn.), 2004.

[13] F. Concannon, A. Flynn and M. Campbell, “What campus-based students think about the quality and benefits of e-learning”, British Journal of Educational Technology, Vol. 36, No. 3, 2005.

[14] K. Aleksic-Maslac, M. Korican, and D. Njavro, “E-Learning Course Development – Quality Standards”, International Conference on Education and Information Systems, Technologies and Applications (EISTA 2008), Orlando, USA, June, 2008.

[15] S. Deterding, D. Dixon, R. Khaled & L. Nacke, “From Game Design Elements to Gamefulness: Defining Gamification”, Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, 2011.

[16] L. Pilar, S. Rojik, T. Balcarova and J. Polakova, “Gamification in education: current state”, 13th International Conference on Efficiency and Responsibility in Education, Volume 13, Prague, 2016.

[17] K. Aleksic-Maslac, B. Sinkovic, and P. Vranesic, “Influence of gamification on student engagement in education”, International Journal of Education and Learning Systems, Volume 2, 2017.

[18] D. Dicheva, C. Dichev, G. Agre and G. Angelova, “Gamification in Education: A Systematic Mapping Study”, Educational Technology & Society, 18(3), 75-88, 2015.

[19] P. Fotaris, T. Mastoras, R. Leinfellner and Y. Rosunally, “Climbing Up the Leaderboard: An Empirical Study of Applying Gamification Techniques to a Computer Programming Class”, Electronic Journal of e-Learning, May 2016.

[20] K. Aleksić-Maslać, B. Sinković, and P. Vranešić, "The Role of Competition and Reward Regarding student motivation in the gamification process of different age groups”, EDULEARN18 Proceedings , IATED Academy , 2018. Pp 1991-1999.

[21] K. Aleksić-Maslać, M. Rašić, and P. Vranešić, “Influence of gamification on student motivation in the educational process in courses of different fields“, 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018.

[22] M. Urh, G. Vukovic, E. Jereb, and R. Pintar, “The model for introduction of gamification into e-learning in higher education”, Procedia – Social and Behavioral Sciences, 197 (2015), pp. 388-397.

[23] A. Antonaci, R. Klemke, C. M. Stracke and M. Specht, “Gamification in MOOCs to enhance users' goal achievement”, Global Engineering Education Conference (EDUCON), 2017 IEEE (pp. 1654-1662).

[24] E. Salas, N. J. Cooke, and M. A. Rosen, “On teams, teamwork, and team performance: Discoveries and developments“, Human Factors, Volume 50, Issue , pp 540-547.

[25] V. P. Bossche,W. H. Gijselaers, M. Segers, and P. A. Kirschner, “Social and cognitive factors driving teamwork in collaborative learning environments - Team learning beliefs and behaviors”, Small Group Research, Volume 37, Issue 5, pp 490-521, 2006.

[26] M. Magzan, K. Aleksic-Maslac, and V. Juric, “Teamwork in Relation to Quality of E-Learning: Business Education Context”, Advances in Business-Related Scientific Research Conference (Olbia, Sardinia, Italy, Sep 8-10, 2010)

[27] I. Nugraha, and E. Eliyawati, “The use of video laboratory report to develop presentation skills in science teacher education students“, International Conference on Mathematics and Science Education (ICMSE 2018), Volume 1157, 2019.

[28] K. Aleksić-Maslać, D. Vasić, and M. Koričan, “Student Learning Contribution through E-learning Dimension at Course “Management Information Systems”“, WSEAS transactions on information science and applications, Volume 7, Issue 3, 2010.

[29] N. Suzuki, H. Shoda, M. Sakata, and K. Inada, „Essential Tips for Successful Collaboration - A Case Study of the "Marshmallow Challenge"“, Human Interface and the Management of Information: Applications and Services, Pt II (edited by: S. Yamamoto), Lecture Notes in Computer Science, Volume: 9735, pp. 81-89, 2016.

[30] Kahoot - https://kahoot.com (2020)

Non-determinism in nowadays computing and IT education

J. Henno*, H. Jaakkola** and J. Mäkelä***
* Tallinn University of Technology, Estonia
** Tampere University of Technology, Pori
*** University of Lapland, Rovaniemi, Finland
jaak@cc.ttu.ee

Abstract – **Once-upon-a time computers and computations they delivered were considered ultimately deterministic. But currently we encounter random events, i.e. non-determinism in many areas of computation practice: non-determinism introduced by network latency, inherently non-deterministic computations with 'big data' and 'deep learning' where the results are probability distributions with errors, which also are probability distributions and both depend on initial selection of samples etc. For IT education one of the most disturbing sources of non-determinism and non-repeatability of previous examples comes from massive use of libraries and API-s (Application Programming Interface), which has made common the 'top-down' style of programming and negligence to practical issues – finiteness of computer memory and speed. Significance of the classical source of knowledge - printed hard-cover books - is diminishing, since by the time books are out of print there are already new versions of programs, new protocols, new technologies and libraries, and these new versions often do not work with old ones. The most relevant source of information has become Internet. But Internet is full of useless sources, since 'Internet never forgets' – together with sources describing latest program versions, libraries, technologies etc. there are still around tens of publications which use some by now already outdated program versions, libraries, technologies. Students, who are eager to perform well in the next recruiting interview are spending many evenings trying to swim in this swamp of non-deterministic mess, where most of presented examples are not repeatable. And the situation is becoming worse, since many of authors e.g. YouTube videos do not want to teach, but to earn using Google AdSense.**

Keywords – computation; determinism; networks; deep learning; big data; Internet

I. INTRODUCTION

"Something automatic happened. Moodle is down" was the message from our Moodle manager; Windows 10 search does not find files in your computer – Microsoft proposes six sets of instructions [1] for you to correct the problem (involving editing the registry – always considered strongly no-no action for casual computer users); computer started, but for next half an hour screen shows only spinning dots – Microsoft consoles [2]: "It won't happen often", "it generally means that the updates are finalizing" – but this particular computer is not connected to network so updates are impossible; computer was connected, but Firefox refused for several minutes to update any tab; security cameras which are safe only when they are disconnected [3]. Everyone can add to this list his/her own experiences of "automatic happened".

In spite of all our achievements in IT technology such things happen and in order to decrease this "just happened" and to explain to our students how to decrease these happenings we have to understand, why these things happen.

The belief in determinism and repeatability of computations is one of cornerstones for most people dealing with computers. The repeatability principle is the basis of all science. Everyone can repeat the classical experiments of Newton or Galilei which became the base of classical laws of physics. This repeatability principle is also the cornerstone for our students – in order to believe in truths what we deliver those truths should be based on some repeatable evidence, students should be able to test them out themselves, to repeat the experiments, re-create programs and get the same results.

Unfortunately this principle of unconditional repeatability has started vacillate.

The first indications un-repeatability were voiced already in the beginning of the last decade in 'soft' fields - various subfields of medical, life and behavioral sciences, in fields, where all new discoveries are based on experiments and theoretical inferences could be stated only from these experiments. It turned out, that there is a "replication and reproducibility crisis" - many publications were based on experiments and data which nobody could repeat [4]. Poll organized in 2016 by the journal Nature show, that more than half of scientists considered this as a real crisis in science [5].

A similar "replication and reproducibility crisis", but for different reasons has appeared in programming or more generally, software production.

II. ESSENCE OF PROGRAMMING

In its very essence, programming is the art and ability of foreseeing future. We write something (the code), consisting of series of 'small' actions – move this value or those words to that variable, add or multiply this to that, repeat these actions until some condition becomes true (or vice versa) and we have to foresee, what actually happens, when all these small piece-vice actions are automatically executed, what will be the result – it should be exactly what we wanted it to be (computers are in principle deterministic!). But when our programs grow, become longer and longer (we want to do more and more!), we can not foresee everything – our

memory is limited. Thus our foreseeing ability weakens more and more and the result which appears after execution of our well-planned actions appears as non-foreseen, i.e. looks like nondeterminism.

It is sometimes estimated, that a chess grandmaster could foresee, i.e. to calculate the situation in the game for twenty moves ahead. For computers to achieve this is needed rapid processing power, e.g. the chess program X3D Fritz which was the first computer program to play on par with the world chess champion Garry Kasparov in 2003 run on four Intel Pentium 4 Xeon CPUs (Central Processing Unit – computer's processor) at 2.8 GHz – best PC processors at the time (today smartphones can run similar chess programs). Besides of processing speed computer chess programs have also other (for tasks from practice, e.g. chess – practically unlimited) resource – memory. For the chess programs which finally defeated the human chess master where from the very start of their development used databases of chess endgames [6]. The result – winning the human world champion – was based on these databases and a "brute force" calculation capacity – it run on custom VLSI chips to execute the alpha-beta search algorithm in parallel.

The chess program won the future with 'brute force', i.e. with more memory and quick processors. But in everyday programming practice programmers are constantly dealing with future and want to have for this formal concepts. Theoretical concept *future* together with closely related concepts *promise, delay* and *deferred future* were formalized already in the second half of seventies of the last century [7], [8]. Since the concepts *future, promise* etc. are very important in asynchronous web programming, these concepts are already introduced in several programming languages e.g. in the rather novel languages Dart [https://dart.dev/], Kotlin [https://kotlinlang.org/], but also in 'old hand' C++ [9], [10].

Formalism can only reduce programming errors, not to eliminate them. It is rather easy to err even in seemingly simple situation, as shows an example from Dart [11]). The following code contains an error – can you find it ?

```
String createOrderMessage () {
  var order = fetchUserOrder();
  return 'Your order is: $order';
}


Future<String> fetchUserOrder() {
  return
   Future.delayed(Duration(seconds:  4),
   () => 'Large Latte');
}
void main () {
  print(createOrderMessage());
}:
```

Non-determinism (and also non-repeatability of program results) comes from several sources.

Computers are finite devices, they represent e.g. real numbers (there are infinitely many of them!) using approximations, i.e. with some error; in mathematical operations these errors may accumulate, thus even the most basic laws of arithmetic's, associativity and commutativity of addition

$$a+(b+c)=(a+b)+c\,,\ a+b=b+a$$

may not hold. For instance, summing up a series of 10000 randomly generated real numbers first in forward and then – in backwards direction will usually produce different results:

Difference: 4.192202140984591e-13

Non-determinism in calculations appears constantly in calculations where the GPU (Graphic Processor Unit) is used [12], [13], e.g. in many Machine Learning areas. Most of Machine Learning algorithms are still local, i.e. designed and executed on one computer. But since machine learning algorithms employ massive computations, besides utilizing GPU is becoming popular also distributing these calculations over computer network and cloud [14], and communication on computer network inherently includes also non-determinism.

III. RISE OF INCERTAINTY AND NONDETERMINISM

The world is changing – rapidly; this has been repeated also rapidly and is one of few (the only?) things which is not changing, i.e. non-deterministic.

With the first stand-alone, i.e. not network-connected computers their determinism was nearly absolute. Computers are deterministic devices, if the produce random results then they are severely broken. Von Neumann, 'Father of the Information Theory' commented on this: "Anyone who attempts to generate random numbers by deterministic means is, of course, living in a state of sin."

Randomness appeared when computers were connected and appeared computer networks. If a computers communicate over network and one makes a request to another, then it is impossible to say when the answer arrives – this depends on all the other computers/servers on message's path. 'Classical' programming languages expect determinism – they want to get everything in expected time. In order to handle network's communication uncertainty and nondeterminism were in programming languages for functions which implemented network queries added a '*callback*' argument – a function, which was executed when query returned the answer (in previously unknown time) And answer may not arrive at all – browser shows the response "*404 – Page not found*", thus the queries begin with *try* and end with *catch(error)*. Web programming is asynchronous, programs which communicate with other computers over computer network should always be ready for arbitrary delays. This creates randomness, which is a natural part of communication in networked world.

Computers were invented for calculations and at that time the idea of calculation was that this is done in a single

computer using finite amount of data what is presented as an input at the start of the program. The first successful programming language for students, the BASIC (Beginners' All-purpose Symbolic Instruction Code), created for teaching in the Dartmouth College in 1964 even included program data at the end of program's text on separate lines beginning with "DATA …". Such a computation is totally closed – once it is started (with fixed input), nothing will disturb it and it even finishes or does not, i.e. goes into endless loop (Alain Turing understood this and more-or-less simultaneously with Austrian logician Kurt Gödel). But there are not so many such closed-form problems, i.e. problems, where data is presented locally at-once, at start of a program, especially taking into account computer's strongest feature – re-use of programs. The president of the mighty IBM (International Business Machines corporation - at the time of introduction of this name these 'business machines' were punched card tabulators) instinctively understood the situation when in 1943 he stated: "I think there is a world market for maybe five computers". And he was not alone, e.g. the grandson of the famous naturalist Sir Charles Darwin, at the time head of the Britain's National Physical Laboratory, where computers in 1946 were introduced considered one computer to suffice to solve all the problems that are demanded of it from the whole country [15].

And it seems that Watson, Darwin and several other prominent researchers of the time were right – the well-isolated problems with fixed amount of input data are already mostly solved and programs to solve these problems are already mostly written.

The programs for isolated problems with fixed data are in principle all similar to the classical programming example "Hello, World!", but there is little need for such "Hello World"-type programs any more (although many IT teachers in universities do not believe). Nowadays computers are not used for closed, finite sets of data – they are used for communication and big, open-ended sets of data. We use them for sending e-mail, watching TV or live 'reality' shows and computers are connected into large, complex networks, where messages are moved as small fixed-size packets. On networks messages 'hop' from server to server, their next destination is calculated 'on the fly'. Nobody can know beforehand the exact route they move or tame it takes, thus nobody can also not know the exact time of packet arrival or the rate of 'lost' packets – generally 1-2.5% is considered quite acceptable [16].

Packets are nearly independent, since the main Internet protocols: HTTP (Hypertext Transfer Protocol), UDP (User Datagram Protocol), DNS (Domain Name System) were designed stateless - each information packet travels on its own and does not know anything about any other packet. Each communication in Stateless Protocol is discrete and unrelated to those that precedes or follow. Stateless Protocols greatly simplify the design of Server, but increase traffic – client has to make a new request for every item. Modern web-pages contain hundreds of components, e.g. the newspaper "The New York Times" page for 25.02.2020 contains more than 40 separate items.

The 'step-wise' protocols were created in the 'innocence' era of the Internet. The current Internet, where nearly half of traffic is created by software bots – autonomous applications (apps) that execute tasks (scripts) over the Internet on connected computers. The all-expanding world of bots includes both good ones (e.g. Google crawlers, who constantly traverse Internet and store data for Google search, tools to improve programmers productivity [17] or Workshop [18] – a tool for freelancers witch scans the internet for freelance job offers, chatbots, appearing on many web sites – everyone can easily build one [19] etc.) and bad ones, who seek to carry out data theft, send spam, distributed denial-of-service (DDoS) attacks, account takeovers etc. In 2018 bad bots created 20,4% of all the Internet traffic [20].

One of the most notable good bot achievements was the in the Google's London-based AI company DeepMind developed Go-playing bot [21], which in 2016-2017 defeated top professional Go players. In chess it is estimated to be ca 10^{43} different positions [22]. Go is on several orders more difficult, in Go is estimated to be approximately $2.081681994 * 10^{170}$ legal positions [23]. Thus the Go-playing program had to use two neural networks and random search to prune the tree of future positions [24].

IV. MIMICKING HUMANS

Even more complex is the next great task (game) for computers – understanding natural language. Adult native English language speakers know 20,000–35,000 words [25] (Shakespeare – 35000 [26]), the number of possible sentences is infinite. Natural Language Processing (NLP) is done using neural networks and GPU-s and involve many random elements. But NLP programs are already on the verge of understanding natural language. Programs from the top 5-6 teams are progressing rapidly [27] and English language models allow already with very little experience in NLP to build a rather naturally behaving questions answering bot for website [28].

NLP systems have already proved their practical value: the Canadian NLP-based health monitoring platform BlueDot [29] scanned news reports and airline ticketing to predict the spread of diseases send first warning about the outbreak of coronavirus already on December 31, more than a week before the World Health Organization notified the public of virus outbreak in China [30].

Humans have tried to create systems which understand their own, natural/human languages for a long time and most approaches have been 'top-down', i.e. word's meaning was defined using some hierarchical system of concepts ('thing' –'living_thing'-'animal'-'mammal'-'human'-'John'). All such systems are subjective and do not agree with each other. NLP introduced a different, 'bottom-up' approach – word meaning is defined by word context, i.e. other words which occur often together with the word. When sufficiently large text corpus is analyzed for co-occurences appear concepts – groups of words with similar meaning [31], e.g. processing English-languge presentations from Mipre.CE 2017-2019 (25037 unique words) using the Tensorflow Python libraries

[32] revealed many closely-packed groups of words, i.e. concepts.

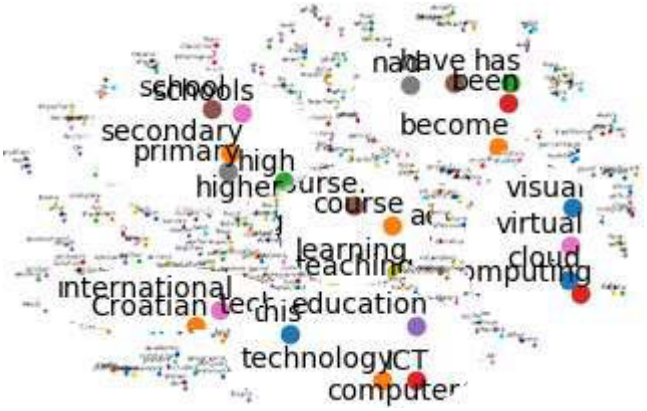


Figure 1. Visualization of some word groups from corpus created from Mipro.CE English-language presentations from 2017-2019

V. INTERNET IS WAKING

With AI-supported bots Internet is often already resembling living system, where your mobile phone may itself connect to your fridge in order to report to you what you should buy in order to entertain quests whose mobile just informed your mobile about their arrival etc. This is very different from 'classical' computing and requires massive layers of software to make it work and, most important – to keep it safe.

The first generation of computer safety tools – virus scanners - were based on scanning application's code for certain software patterns - this is like deciding about visitor by looking at his/her suite. Nowadays safety software is more and more based on inspecting actor's behavior – how does it differ from normal 'human' behavior? If a bad bot is constantly 'knocking your door', i.e. sending requests hoping to guess your password or simply cause your computer to fail (DDOS), then the number of such requests should be limited; if some application in your computer regularly emerges at midnight and tries to connect to somewhere then this looks like a bot (Windows can also wake computer up and start an update); if many customers want simultaneously order from a web shop, then also the number of requests should be limited (the web throttling [33]). Web applications (e.g. web shops) interface designers apply numerous safety guidelines [34], which are all the time becoming more stringent, e.g. the U.S. government Federal Information Processing Standard (FIPS) Publication 140-2 level 4 requires "the ability to be tamper-active, erasing the contents of the device if it detects various forms of environmental attack" [35] – the attacked computer should kill itself .

VI. REUSE – API-S

With growing complexity developers have to repeat good, already tested designs. Good, working designs are implemented as software libraries. Nowadays nobody starts to develop a (web) application with command-by-command style – all developers use 'library-by-library' style instead.

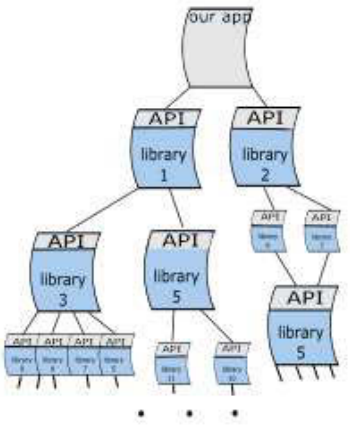


Figure 2. Modern programs are composed from many libraries

Libraries and their API-s (Application Programming Interface, the input-output functions which libraries reveal outside) have become the new programming language. There are thousands of libraries and API-s: the Python package index [36] lists over 200 thousand projects, the Javascripting index ("The definitive source of the best JavaScript libraries, frameworks, and plugins" [37]) lists over 1500 JavaScript libraries, PHP has over 200 thousand modules, Apache Maven project repository [38] lists for Java over 300 thousand packages and the node.js [39] – over 1.2 million packages.

The term API which originally was used only for interfaces of packages has nowadays much wider meaning, since there are many 'things', which also have programming interface – mobile phones, door locks and doorbells, smoke-detectors, fridges etc. Thus the number of API-s is accelerating. The web communication is mainly based on two web service API-s: the REST and the SOAP [40]. But besides these there many types of open-source API-s: for music, travel, beer and wine, machine learning, biometrix etc. [41], [42].

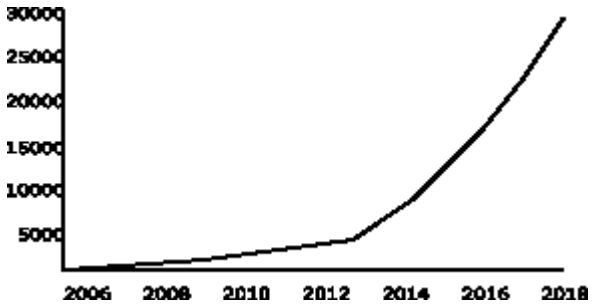


Figure 3. Growth of the number of API-s

Public API-s allow to use programs, i.e. functionality developed elsewhere over the web: explore and download NASA imaginary [43], add to you mobile phone a button for downloading stuff from Amazon [44] etc, And use of API-s has made prevailing the top-down style of application development.

VII. THE TOP-DOWN DEVELOPMENT STYLE

The classical programming style is 'bottom-up' – programmer writes code line-by-line and his productivity is (if needed) measured in loc (lines of code) units. This is the still the prevalent style in most software production textbooks and courses.

But those 'Hello World' programming snippets are usually all local, to be executed in local computer and they show the message 'Hello World' also on local computer screen. To print something to local computers screen – it were quicker to say the message directly using your own human voice. In our all-connected world based on computer networks the program should show the message in response to request from another computer on network. Such a program becomes far more complex and in our security-laden software creation environment its creation begins not from bottom, i.e. by writing programs code, but from top and involves several steps: initializing programming project, creating classes etc. For instance in Java language, using the *Spring* framework [45] these steps would be [46] (at least) following:

1. Starting the *Spring Initializr*; this enables including all dependencies (other libraries) needed for the application, here the *Spring Web dependency*. *Spring* does also a lot of project setup, for this user has to give values to project metadata variables *Group* and *Artifact* and to select the build tool: either *Maven* or *Gradle*; if user selects *Maven*, *Spring* creates the POM (Project Object Model) description xml-file (44 lines of xml); for *Gradle* will appear a bit shorter build description XML-file.

2. Next should be created the web service, which handles required interactions:

- first the GET request for greeting; the request body contains values of the request *id* and *message* parameters, optionally also user name parameter in the query string; the GET request should return a 200 OK response with JSON in the body that represents the greeting

- the corresponding JSON file in response is created by a resource representation class file *Greeting.java* (14 lines of Java code).

3. To handle HTTP requests in *Spring*'s framework for building RESTful web services is needed a *Resource Controller*, which is created with 11 lines of Java code.

The (more-or-less) understandable are two lines in file *Greeting.java*:

```
this.id = id;
this.content = content;
```

Although even here is difficult to understand numbering – when meeting a friend on street, we usually do not start the conversation in style:

"*First: Hello John! Second: how do you do?*"

All other more than 50 lines are web machinery, which programmer should use, but often does not understand, thus tries to build using the copy-paste programming method. Naturally, this is very error-prone method, thus greatly increases randomness in our digital word.

For students this takes lot of time to understand and many do not. The main thing many have learned is that there should be libraries, thus one student (experience of the first author) started a simple 2D browser (i.e. JavaScript) 'Flappy Bird' game [47] with linking 75 *node.js* classes. He was not able to use several on-line tutorials with code [48], since these tutorials do not use external libraries (not needed for such a simple task); he tried to use all 'bells and whistles', i.e. these classes and then composed totally senseless lines of code :

```
collides(us,them)
```

From here it follows, that in every frame the bird would collide with itself and thus could not fly.

VIII. COMPUTERS ARE FINITE

Top-down mindset usually does not support understanding of computer memory – such students often create programs, which 'eat memory' – they create new objects (use *malloc* or *new*) at run time; for new objects is needed also new memory, thus the memory used by program may start growing and finally the whole computer halts [49]; even if compiler manages to avoid crash with garbage collection, program still becomes slow, since new memory is obtained by request from Operating System (e.g. Windows).

With the top-down pattern of thinking students often do not understand that computer, computer's memory, network bandwidth etc. are all finite resources - they use them as if they were endless. Fortunately there are several tools which help to investigate low-level issues which may cause troubles and computer's unexpected behavior.

Many users drag all icons of their programs to desktop. This makes it very convenient to start these programs, but this also may make computer start (very) slow, since e.g. Windows loads all these icons at startup reading them from the file *desktop.ini*.

Somewhat similar attitude is prevailing among software producers. In order to make loading of their programs quick (every program wants to be quick), they make OS (Operating System) to load all these programs at computer start-up (booting) into computer's RAM (Random Access, i.e. run-time) memory, using for this every possible way, e.g. all possible registry settings. Microsoft utility *Autoruns* [50] reveals, that in an average Windows computer are on startup loaded hundreds of programs (besides Windows own modules). And many programs are so clever, that after un-enabling their entries in *Autoruns* after a while they re-appear, e.g. even if Microsoft Outlook or Skype are not installed, in *Autoruns* still appear 6 entries for Outlook (e.g. Outlook Social Connector, Access Outlook Data Collection Add-In), 5 – for Skype; 4 - for Google etc.



Figure 4. Small part (ca 1/5) of the list of autoruns entries

The unnecessarily loaded on startup programs all use computer memory and make booting slower, since most of them want on startup to connect to their servers for updates. Looking with the *TCPEye* [51] at the list of current network connections shows, that computer is actively keeping connection with tens of remote servers – this is the pause many computers exhibit at startup.

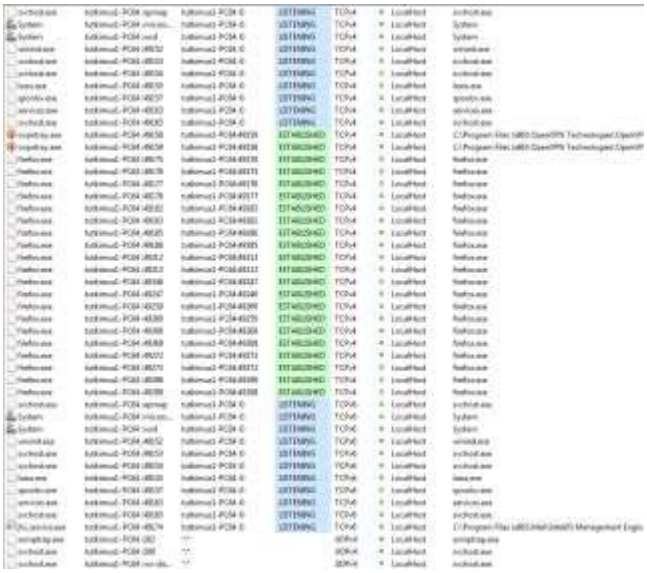


Figure 5. Part of network connections (ca ¼) active in the computer where this is written.

All these programs do something in computer; the *Process Monitor* [52] reveals, how many (often unnecessary) processes are started in your computer (a simplified information gives also *Windows Task Manager*: press Ctrl+Alt+Del). My computer was just started, but *Process Monitor* reveals, that Windows Explorer has already four times sent queries to a quite good free tool for creating CDs, DVDs, and Blu-ray disks CDBurnerXP [53], although the program has not been used.

All browsers (Firefox, Chrome, Edge etc – currently there are web 116 browsers [54]) boast with their speed. Browser is not a very big and complex program (e.g. Firefox – 555KB, Chrome – 1672 KB, Opera – 1493 KB) and they all have to use the same 'pipe', i.e. network connection, thus the main source of their speed (for opening a page) comes from cache – they save everything in order to open it on next time very quickly from the local hard disk cache, e.g. in this computer Firefoxe's cache is 1.2 GB.

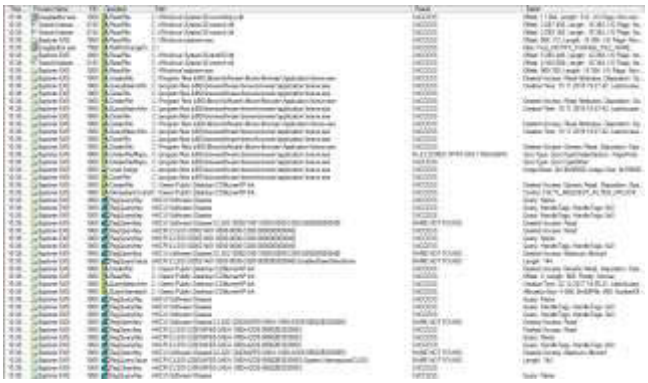


Figure 6. Processes (ca 1/5) in this computer.

Such strange entries in *Autoruns*, *TCPEye*, *ProcessMonitor* allow to discover viruses and bots, e.g. the site *BleepingComputer* [55] maintains a list of nearly 27 thousand automatically starting programs, divided to several categories: good, bad, unknown. Cleaning cache (or at least deleting browser's history – easy to do from browser's menu) deletes also all cookies, thus e.g. if an on-line newspaper first shows some pages for free and then announces "You have used all your free limit – start paying!", usually deleting the history re-enables free pages.

IX. AGE OF DATA

The world economy is changing. Humanity is overusing world's natural resources. We use ca 1.7 times more than nature is capable to re-produce – already by August 1st we have already consumed the resources nature can regenerate in a year [56], and the speed of our consumption of natural resources is increasing To satisfy our all-increasing needs and World's increasing population we have to invent something to replace natural resources, first of all – oil. The age of oil is replaced with the age of data, BIG DATA.

Data is not eatable, we can't drive our cars with data, however big. Data, BIG DATA allows us to reduce waste and losses in our complex world-wide economics, streamline production, distribution and consumption. And the effects are already present – according to the *United Nations Digital Economy Report 2019* [57], already in 2016 the worldwide digital sphere produced 15.5 percent of global GDP; by 2025 it is expected to grow to 24.3 percent of global GDP [58].

Just like oil, data also need processing. Processing of digital data happens in our digital devices – computers, mobile phones etc., their networks and many different communications and interactions which occur in these networks – computer-computer, human-human, human-computer-human. And waste and losses occur in many different steps of this process. Reducing or eliminating even small amount of losses, nondeterminism for message

receivers in communication would greatly increase efficiency of the whole mankind.

X. SOFTWARE ERROR

With complexity and constant updates increases possibility of errors. Updates are often created by programmers, who did not create the original program, do not know all the previous enhancements and corrections, especially security improvements. The results are often laughable, but they seriously affect mindset of computer users, who somehow have to deal with these errors or very difficult to understand messages.

Microsoft Word is probably the most essential tool for many users. With growing security standards its complexity grows. The "Trust Center" of Microsoft Word lists in "Advanced Options" (at least) 115 options, what user should be able to decide, whether to use or not and the "Trust Center" shows 17 file formats with different security options, some obviously already very old. "Word 2 and earlier" – Word 2 appeared more than 35 years ago.. For most (casual) users, all this is just a digital noise.

With growing requirements for security have increased also browser's errors in page delivery. Clicking in a search result in Google search may produce many responses:

There has been a critical error on your website.

Learn more about debugging in WordPress.

What is 'Worldpress' and why should I learn its debugging – this the task for site owners!

The page you are trying to view cannot be shown because the authenticity of the received data could not be verified. Please contact the website owners to inform them of this problem. –

If I can't open the site, then how should I 'contact the owners – please provide a phone number or an e-mail address !

The errors may seem insignificant, but for a person who is starting to learn programming they may be overwhelming.

XI. CONCLUSIONS

Complexity of software is constantly growing. In current software production overwhelming is the 'top-down' approach. Managers want everything to be done quicker and cheaper, software projects are presented as a hierarchical system of tasks and their implementation based on massive use of libraries and IDE-s. Low-level technological aspects – peculiarities of processes on computer networks, finiteness of all computer and network resources also grow. Although they often get less attention, these low-level technical details are important and not understanding them creates often "automatic happened", thus should be essential part of software production education.

The amount of practical knowledge what programmers currently should know seems to be already comparable with the most knowledge and practice-intense profession – medical. But the organization of education is different: medical students have more than 60 specialties and after graduation have to pass internship before they get a full doctor's license. It may be good to introduce something similar also to programming education. Some steps in this direction have been already taken, e.g. the Canadian e-commerce platform Shopify is offering the "Dev Degree": a 4-year, work-integrated learning program that combines 4,500+ hours of hands-on developer experience with 4,000+ hours of academic experience and produces accredited Computer Science degree from either Carleton University or York University [59].

[1] Windows Central. How to fix search problems in Windows 10. https://www.windowscentral.com/how-fix-taskbar-search-not-working-windows-10

[2] Microsoft Windows Support. Troubleshoot black screen or blank screen errors. https://support.microsoft.com/en-us/help/14106/windows-10-troubleshoot-black-screen-or-blank-screen-errors

[3] cnet. Smart vacuum's camera gives hackers a room with a view. https://www.cnet.com/news/hackers-can-peep-through-this-smart-vacuums-camera-research-shows/

[4] Pashler, H., Wagenmakers, E.J. Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? Perspectives on Psychological Science, 7(6) 2012, pp 528–530

[5] Baker, M. 1,500 Scientists Lift the Lid on Reproducibility. Nature, 533(7604) 2016, pp 452–454

[6] Feng-Hsiung Hsu. Behind Deep Blue: Building the Computer that Defeated the World Chess Champion. Princeton University Pres

[7] 7 Friedman, Daniel; David Wise. The Impact of Applicative Programming on Multiprocessing. International Conference on Parallel Processing 1976, pp. 263–272

[8] Hibbard, Peter. Parallel Processing Facilities. New Directions in Algorithmic Languages, (ed.) Stephen A. Schuman, IRIA, 1976

[9] http://www.cplusplus.com/reference/future/async/

[10] C++11 async tutorial. https://solarianprogrammer.com/2012/10/17/cpp-11-async-tutorial/

[11] Asynchronous programming: futures, async, await. https://dart.dev/codelabs/async-await

[12] M. Morin, M. Willetts Non-Determinism in TensorFlow ResNets. arXiv:2001.11396 [cs.LG]

[13] Tom B. Brown. ML Tidbits: Nondeterminism on the GPU. https://medium.com/tom-b-brown/tensorflow-nondeterminism-on-the-gpu-a0e86125fd06

[14] Mehdi El-Amine. NLP Research Lab Part 1: Distributed Representations. https://www.districtdatalabs.com/nlp-research-lab-part-1-distributed-representations

[15] Copeland, Jack (2006). Colossus: The Secrets of Bletchley Park's Codebreaking Computers. Oxford University Press. p. 109

[16] Mohammad Obaidat Faouzi Zarai Petros Nicopolitidis, editors. Modeling and Simulation of Computer Networks and Systems, Morgan Kaufmann 2015, pp 1-964 ISBN: 9780128008874, eBook ISBN: 9780128011584

[17] Best Productivity Bots Software. https://www.g2.com/categories/productivity-bots

[18] http://letsworkshop.com/

[19] 19 The European Chatbot Creation Platform. https://botnation.ai/en/

[20] Imperva. Bad Bot Report 2019: The Bot Arms Race Continues. https://www.imperva.com/resources/resource-library/reports/bad-bot-report-2019-the-bot-arms-race-continues/

[21] D. Silver1, A. Huang et al. Mastering the game of Go with deep neural networks and tree search.

https://storage.googleapis.com/deepmind-media/alphago/AlphaGoNaturePaper.pdf

[22] Claude Shannon. Programming a Computer for Playing Chess. Philosophical Magazine 1950, 41 p. 314

[23] Number of Possible Go Games. https://senseis.xmp.net/?NumberOfPossibleGoGames

[24] Neural Networks and Go. https://senseis.xmp.net/?NeuralNetworksAndGo

[25] The Economist. Lexical facts. https://www.economist.com/johnson/2013/05/29/lexical-facts

[26] B. Efron, R. Thisted. Estimating the number of unseen species: How many words did Shakespeare know? Biometrika (1976), 63:3, pp 435-447

[27] SuperGlue Leaderboard. https://super.gluebenchmark.com/leaderboard

[28] P. Dwivedi. NLP — Building a Question Answering model. https://towardsdatascience.com/nlp-building-a-question-answering-model-ed0529a68c54

[29] bluedot. https://bluedot.global/

[30] An AI algorithm detected the coronavirus outbreak a week before the CDC. https://www.beckershospitalreview.com/artificial-intelligence/an-ai-algorithm-detected-the-coronavirus-outbreak-a-week-before-the-cdc.html

[31] Hobson Lane, Cole Howard, Hannes Hapke. Natural Language Processing in Action. Manning, March 2019 ISBN 9781617294631 544 pages

[32] Tensorflow. https://www.tensorflow.org/

[33] Amazon Web Services. API Request Throttling. https://docs.aws.amazon.com/AWSEC2/latest/APIReference/throttling.html

[34] RedHat. API Security https://www.redhat.com/en/topics/security/api-security

[35] Thales. What is FIPS 140-2? https://www.thalesesecurity.com/faq/key-secrets-management/what-fips-140-2

[36] https://pypi.org/

[37] https://www.javascripting.com

[38] http://maven.apache.org/

[39] https://www.npmjs.com/]

[40] guru99. SOAP Vs. REST: Difference between Web API Services. https://www.guru99.com/comparison-between-web-services.html

[41] 8+ Biometrics APIs At Your Fingertips. https://nordicapis.com/8-biometrics-apis-at-your-fingertips/

[42] https://rapidapi.com/blog/types-of-apis/

[43] https://api.nasa.gov/

[44] https://aws.amazon.com/api-gateway/

[45] https://spring.io/

[46] spring. Building a RESTful Web Service. https://spring.io/guides/gs/rest-service/

[47] https://flappybird.io/

[48] w3shools. HTML game example. https://www.w3schools.com/graphics/game_intro.asp

[49] Pc.net. https://pc.net/glossary/definition/memoryleak

[50] Autoruns. https://docs.microsoft.com/en-us/sysinternals/downloads/autoruns

[51] TCPEye. https://tcpeye-network-tools.en.uptodown.com/windows

[52] Processmonitor. https://docs.microsoft.com/en-us/sysinternals/downloads/procmon

[53] https://cdburnerxp.se/

[54] Browser List. https://www.webdevelopersnotes.com/browsers-list

[55] https://www.bleepingcomputer.com/startups/

[56] Earth Overshoot Day: Press Release July 2018 English. Humanity has already exhausted Earth's natural resources for 2018. https://www.globalagriculture.org/whats-new/news/en/33330.html

[57] https://unctad.org/en/PublicationsLibrary/der2019_en.pdf

[58] Digital Spillover – Huawei. https://www.huawei.com/minisite/gci/en/digital-spillover/files/gci_digital_spillover.pdf

[59] Dev Degree. https://devdegree.ca/

An Experiential Learning Approach to Research Methods in Computer Science based on SMART Goals

Nicholas Caporusso
Department of Computer Science
Northern Kentucky University
Highland Height, United States
caporusson1@nku.edu

Abstract - **Research methods are perceived among the most challenging classes by students and faculty: in addition to theories, foundations, and technical aspects, they involve applying high-level critical-thinking, problem-solving, and communication skills. Simultaneously, methods courses play a significant and holistic role in contributing to the personal and professional development of students. Therefore, designing engaging and meaningful learning experiences is especially important, as research methods classes touch every dimension of student success. To this end, experiential learning (EL) approaches are among the most successful strategies adopted by faculty: in the last decade, the benefits of incorporating components, such as simulations, internships, service-learning, and fieldwork, have been demonstrated in several domains. Indeed, sharing diverse examples of applications of EL in methods classes in specific fields is beneficial for contributing to the discourse, generating new strategies, and identifying best practices.**

This paper presents the case study of a course in research methods in computer science entirely based on experiential learning. Specifically, the class was designed to match the critical steps of the scholarly publishing process. Throughout the semester, students generate and refine applied research projects that could eventually be published in actual academic conferences or scientific journals. The findings from three years of data demonstrate the effectiveness of the approach discussed in the paper.

Keywords - experiential learning, computer science

I. INTRODUCTION

Conducting research is a unique and fascinating activity that requires using a combination of diverse skills, from theoretical understanding to practical knowledge, and taking them to the next level to contribute to the scientific discourse. As a result, despite being particularly challenging and intellectually demanding, it is essential to expose students to research from the early stages of their education. Furthermore, this becomes especially valuable in college studies both at the undergraduate and graduate levels, where being involved in research helps students develop their higher-level abilities and contribute to their success. Whether they will choose to pursue a career in higher education or outside academia, critical thinking skills, and scientific approaches contribute to creating more robust foundations for their personal and professional development. To this end, methods courses are precious for introducing students to research and engaging them in broadening their theoretical horizons while simultaneously increasing their technical and practical knowledge within a particular domain and beyond the boundaries of a single discipline. While most courses focus on specific aspects of Bloom's taxonomy, research methods classes have the potential of thoroughly supporting students in mastering all the higher levels of the learning pyramid and becoming independent in their journey as a life-long learner. Moreover, in addition to enhancing their theoretical knowledge and technical skills, working with peers in collaborative research projects fosters the development of soft competencies, whereas documenting, reporting, and presenting the results of a research project help to become more proficient in communication. Due to their complexity and breadth, methods modules are often perceived as among the most challenging by both teachers and learners. Indeed, among the several aspects that have to be considered in the design of the course to make it relevant, the most prominent is student engagement. Specifically, when the same class is offered to different majors having diverse backgrounds and objectives, teachers have to address the additional challenge of fragmentation of methodological expertise, which makes it challenging to find a common theme that fits the interest of all students and to identify meaningful projects that help learners understand how to apply the contents learned in the class to their fields while keeping them engaged[1]. As outlined by several authors, this is particularly true for research methods courses in the field computer science[1]: on the one hand, models and theories can easily be applied to other domains; on the other hand, a large degree of interdisciplinarity, transferability, and meta-cognitive knowledge requires more accurate course design and planning to provide students with experiences that are not theoretically superficial and technically overwhelming, or vice versa. Previous literature highlighted the value of incorporating experiential learning (EL) [2] in methods courses and demonstrated their effectiveness in producing tangible outcomes that motivate students in developing skills that are transferrable to other domains [3]. Nevertheless, due to numerous intrinsic challenges, most courses are based on

simulated projects only. Although they can be a great learning environment for lower-level classes, they are perceived by more mature students as lacking practicality.

II. Related Work

Case studies and simulations are among the most widespread methods for introducing experiential learning in the classroom. Findings from several studies demonstrated the effectiveness of these approaches in enforcing theoretical concepts and, simultaneously, providing students with examples of how to implement them in a fictional project. They are particularly suitable in teaching younger student populations and introducing them to critical thinking, scientific reasoning, and concepts, such as hypothesis formulation, data collection, experimentation, and discussion of the analysis [5]. Also, simulation is particularly useful in training students in research tasks that involve performing specific routine procedures or potentially harmful interventions. In this context, games, 3d models, and virtual environments are very beneficial in supporting research without introducing any risk for animal or human subjects. Unfortunately, poor realism is the main factor affecting most programs based on simulations. Other experiential approaches based on real-world projects, such as the unClassroom methodology, have been proposed in research courses in communication and media [6]. Most of them involve connecting students with external partners and organizations (e.g., non-profits). Particularly, the value of incorporating a service-learning component in the classroom has been discussed extensively in the literature: projects that serve the community and help address major social themes result in the best outcome [7]. Although the introduction of an experiential component appears the key element in determining student success in a program [8], there is a lack of case studies in computer science compared to other disciplines.

III. Experiential Learning Methodology

This paper presents the case study of a graduate research methods class and describes an approach in redesigning its instructional content and delivery strategy with the aim of transforming the course in an environment where individuals can learn and realize actual research studies that have a tangible scholarly outcome. The case study involves a course offered to graduate students, which is part of a set of three classes: as a pre-requisite, students complete a pro-seminar in informatics, which introduces them to basic concepts of scientific research and to techniques for designing and conducting simulated studies; then, typically, the research methods course is taken at the beginning of the last year of the graduate program; specifically, the purpose of the course is to prepare students to complete their thesis project, which is realized during the culminating capstone.

The objective of the approach was to address some of the issues of current methods classes in computer science: specifically, the guiding design principles of the course aimed at overcoming the lack of realism of simulations in EL practices. Moreover, given the level of the course and the maturity of participants, its learning outcomes involved higher cognitive and meta-cognitive knowledge domains of Bloom's taxonomy. As a result, the primary goal was to leverage experiential learning as an enabler for reaching objectives that have an impact beyond the class, as it already happens with internships or service learning, which inherently involve real-world experiences. In this regard, the purpose of the course redesign was to bridge the disconnect between traditional methods classes and tangible outputs. To this end, the experiential learning component of the course was designed based on the actual process of scholarly research and academic publishing. As a result, lectures and assignments are structured in the form of a roadmap that allows learners to transition from an applied research idea to an accomplished study that can potentially be considered for publication in an actual academic conference or scientific journal. Students start by (1) identifying a research topic and (2) reviewing the recent scientific literature so that they can (3) design a project or study that could be (4) peer-reviewed in the form of a first draft and (5) reviewed by experts in a more mature format, in order to (6) produce a final version that is submitted to a scholarly outlet. This scheme of tasks and milestones serves as the backbone structure for designing assignments and their assessment strategy as well as for creating the instructional content, which is organized to provide students with the necessary theoretical and practical foundations for completing each step of the roadmap and achieving the final objective. The approach presented in this paper incorporates Kolb's experiential framework [2] consisting in a cycle of four activities: each milestone in the course is designed to foster opportunities for (1) having a concrete experience of the task, (2) observing the results by either reflecting on them individually or discussing in groups, (3) evaluating or reworking the output by incorporating feedback received by instructors, peers, or experts, and (4) identifying the next steps and iterations, which, in turns, enables experimenting with the process at the meta-cognitive level. As each milestone reproduces a step in the scholarly publishing workflow, course delivery and student supervision are crucial for making sure that the experiential learning cycles result in incremental steps towards accomplishing the final objective of obtaining quality research articles that can be potentially published. As a result, rather than evaluating the individual attempt, students are given unlimited submissions, each of which is the output of an iteration. By doing this, they can document their progress, receive feedback, incorporate it in their work, plan, and execute the next iteration. Simultaneously, each submission attempt receives a temporary grade that reflects the increasing quality of their deliverables until they are satisfied with the result. This, in turn, gives them the possibility of focusing on experimenting with different methods and learning from failure without feeling the pressure to deliver their best at the first attempt to receive the top grade. As some of the objectives are particularly ambitious for the duration of the course (i.e., sixteen weeks), they were defined according to the Specific, Measurable, Attainable, Relevant and Time-bound (SMART) framework, which enabled to divide each milestone down into goals that (1) have a clear and defined scope, (2) can be graded with criteria that are generic enough to be applicable to the different characteristics of the diverse types of projects and, simultaneously, appropriate for the specific task, (3)

can be achieved in the form of a very basic objective and further improvements, (4) are meaningful in terms of learning outcomes and student success, and (5) are compatible with the time available for the course and have a timeline that aligns with the content. By doing so, students have a clear overview of the general roadmap and its breakdown into individual tasks, so that they can evaluate their progress, identify strategies to stay on track, and decide on the next steps without feeling overwhelmed by the ambitious overarching goals of the course.

A. Milestone 1: project selection and proposal

Students are given complete freedom to select the research topic they are interested in exploring. To this end, they are advised to propose a project in line with their personal interests, professional experience, current or desired job; alternatively, they can choose based on their curiosity for a specific discipline or on their willingness to explore or learn a particular subject. The only requirement for the project is that it must involve computer science in its broader sense, from sorting algorithms to the use of social media. Moreover, the research topic requires the instructor's approval before students can start working on it. To this end, the submission of a one-page preliminary proposal is required, typically by the end of week two. This, in turn, has a three-fold objective: ensuring that the research project is feasible and sufficiently relevant to the scope and level of the course, refining the idea, and, simultaneously, enforcing a deadline for starting the work, which also helps reveal and address critical situations (e.g., paralysis by analysis). The preliminary proposal has to highlight the importance of the topic (e.g., citing recent statistics and reports that highlight the dimension of the problem), summarize the state of the art and the approaches taken by others in the recent developments, and outline potential directions for contributing to the scientific conversation. Students are given the freedom to steer the focus of their work later (i.e., as discussed in milestone 3) or to completely change the topic based on the outcome of the subsequent tasks (e.g., preliminary literature review). For instance, they could realize that the solution they had in mind already exists and, thus, they could switch to evaluating its performances in comparison to other approaches; alternatively, they could realize that their lack of technical knowledge prevents them from making a significant contribution to a field and, thus, they could choose to produce an extensive literature review with a reflection on the latest developments.

B. Milestone 2. Review of the state of the art

The second milestone, usually due after week four, requires students to become more familiar with the topic by identifying and reading recent peer-reviewed scholarly articles relevant to their research topic and by summarizing their findings in a literature review document that has the purpose of narrowing the focus of their project and outlining the final objectives of their work. To this end, students are advised to use the resources available on campus (e.g., the library and the writing center), search on dedicated scholarly engines, browse the digital repositories of academic societies and communities and explore the archive of scientific publishers and conference proceedings. Specifically, they are required to include at least the most significant 15 articles in their review and summarize and share them with the class in a short presentation in which they also discuss potential project directions based on their findings.

C. Milestone 3. Project/study design

In addition to rendering students more familiar with the domain chosen for their project and aware of the recent developments in the scientific discourse in a field, producing a review of the state of the art has the purpose of helping them finalize a decision about their project, that is, narrow the focus of their work, identify potential ways in which they can contribute, and choose the type of research approach. As for the latter, students are given complete freedom: they can realize descriptive, exploratory, correlational, explanatory, and analytical projects. As a result, their final article can review the state of the art, discuss a theory, detail the design of a model, implement and validate a new solution, analyze the performance of existing systems, evaluate users' perception of technology, or combine different approaches. Also, they can decide whether their work will be experimental or non-experimental, or if it will involve qualitative, quantitative, or mixed methods. Regardless of the type of contribution they choose to make, students are typically advised to prefer applied approaches rather than pure research projects, though this is not mandatory. This is for factors related to time constraints and to the heterogeneous background of the class. Regardless of the approach chosen for their research, the requirement is that their work must result in a little yet valuable contribution to or tangible improvement in a scientific field.

D. Milestone 4. Peer review

As the fourth step, students are required to produce the first draft of their paper and have it reviewed by their classmates. As this is due by week ten, the work is not expected to be in its final form yet: thus, students are asked to structure their first draft in two parts, that is, completed work and work in progress. Depending on the type of research, the maturity of the project, availability of data, and individual commitment, the former includes sections, such as abstract, introduction, related work, methodology, and data collection, in their final version. Conversely, the latter part, which typically features the results, analysis, and conclusion section, consists of a document in the form of a working draft that is still being completed. As a result, students are expected to keep working on this part of their paper, e.g., for adding more information, incorporating feedback from the instructor, their peers, or expert reviewers, and changing the structure and content of its sections accordingly. As a second goal for accomplishing this milestone, students are required to submit their article to three colleagues for peer review. Simultaneously, when they receive a review request, students are required to fulfill it within two weeks. The objective of this task is to have students collaborate on polishing their first drafts; separating the completed work from the work-in-progress helps them review each part with a different approach: in the former, they work more in-depth on removing typos, non-scholarly references, and wrong citation style, whereas in the work-in-progress section they mainly check for major issues in the design

and methodology. Moreover, as every project is different and the experiential nature of the course itself makes it difficult to define a specific rubric that fits every research type, the peer-review process is also utilized as a mechanism for grading students using a more general approach based on lexical correctness, adequateness of research design, consistency of methodology, the accuracy of results reporting, and narrative flow. Specifically, their work is evaluated as if the reviewer and the author were on the same team: after the review is completed, if the original draft included errors and the reviewer and the author were able to identify and correct them, the team receives the highest points. Otherwise, if the reviewer does not identify the issue or the author does not fix it, both students are deducted points.

E. Milestone 5. Expert review

In this step, students work with the objective of having their paper reviewed by at least three experts. To this end, they are asked to identify researchers in the field of their project, contact them, and collect their review using a standard form provided by the instructor. The review template was designed to facilitate students in interpreting experts' feedback, discussing it with the instructor, or sharing it with the class. As they know in advance about the task, students are advised to already pre-select the list of reviewers when they research the state-of-the-art and to contact the experts and ask for their availability well ahead of this milestone. However, students can choose to wait for the peer-review to be completed before sending their drafts to experts, or to initiate this review phase while waiting for their colleagues' feedback on their work. As a result, they can concurrently realize the peer-review and expert review phases. This is to provide students with more flexibility in case they are behind schedule and to help them catch up if they required extra time in the earlier phases (e.g., project selection). The assignment is graded based on the approach to the task and, specifically, on an appropriate selection of the experts, number of requests sent, and proactive communication with them: its completion does not depend on the number of responses received by the reviewers or by their feedback.

F. Milestone 6. Presentation and publication

As the last milestone, students have to accomplish three objectives: (1) identify at least three potential peer-reviewed scholarly outlets where their work would fit and select the one they think is more suitable for their paper, (2) finalize their article using the poster or paper template required by the editor, and (3) actually submit their contribution for consideration for publication. To this end, they are invited to utilize common search engines for calls for papers, upcoming conference calendars, and solicitations for contributions from publishers and editors; also, they are advised to consider the scholarly sources where they found the articles they cited as references and to ask their reviewers about potential conferences and journals where the article would be appropriate and more likely to be accepted. As for the second objective of this milestone, they can choose the preferred format of their submission based on the available choices for the specific outlet (e.g., short or full paper, poster, or oral presentation). Moreover, although peer-reviewed scholarly outlets are preferred, in case there are no suitable deadlines, students can send their contribution to other types of non-peer-reviewed conferences and journals. Submission typically occurs between weeks 14 and 16 of the course, when they also give a final presentation to the instructor and to the class. Students are graded based on the quality of their work, and they receive points for submitting their work to an appropriate scholarly outlet, regardless of the outcome of the review process.

IV. EXPERIMENTAL STUDY

In addition to evaluating the effectiveness of the proposed strategy compared to more standard experiential learning approaches based on simulations, the objective of the study was to verify whether, given the same complexity level of learning objectives in terms of Bloom's taxonomy, more realistic and ambitious goals result in improved learning outcomes and student satisfaction. To this end, data were acquired from groups of students enrolled in the research methods class over the course of three years (2017-2019), thus, resulting in a longitudinal experimental study. The content and the approach did not change significantly over the years, though, as discussed in the next Section, small improvements were introduced over time in order to provide students with more material and examples. Two sections of the course were offered simultaneously: in presence and online, with the latter consisting of approximately two-thirds of the total enrollment. Video lectures and face-to-face meetings via remote conferencing tools facilitated working with students on a one-to-one basis, to render the learning experiences in the two classes as close as possible. All the participants in the experimental study were graduate students from the group of individuals who enrolled over the course of three years. Subjects were given the option to withdraw from the study in case they did not want to contribute or in case they identified issues in complying with some of the assignments required for completing the class. For instance, as several individuals worked on projects involving their job, they were concerned about the sensitive nature of the information analyzed in their research in case of publication. Consequently, their data is not included in this paper. The study utilized a true-experimental design: at the beginning of the semester, students were administered a demographics questionnaire that enabled them to collect their gender, age, and type of program as criteria for stratified random assignment. The treatment for the experimental group consisted in the set of milestones described in the previous Section, whereas the control group was not expected to work on a real-world scientific project (i.e., they could replicate or simulate a study), and did not have any commitment in terms of expert review (Milestone 4) or submitting their contribution for publication (Milestone 6). A pre-course survey asked them questions about their motivation, objectives, and perception of their competences, whereas the post-course included additional items related to learning outcomes and perceived satisfaction that were utilized to evaluate the effectiveness of the treatment. Also, the turnaround time and the number of iterations required to accomplish a satisfactory result were acquired.

V. RESULTS AND DISCUSSION

The data analysis is based on the model presented in [9]. It evaluates the overall student satisfaction, learning outcomes, and perceived difficulty in the different phases based on the dimensions of (1) student's intrinsic and (2) extrinsic motivation, (3) student self-regulation, (4) student-instructor dialogue, (5) student-student dialogue, (6) instructor activities, and (7) course design. Specifically, intrinsic motivation includes factors such as personal interest and professional goals. In contrast, extrinsic motivation was associated with the primary intent of taking the required methods class to finish the program and graduate. Self-regulation was calculated in terms of time dedicated to the project compared to the available time in the semester; also, adherence to milestones and their timeline was taken into consideration. In addition to student-instructor dialogue and student-student dialogue, student-expert exchange was considered as a measure of interaction. Instructor activities were evaluated in terms of perceived quality of course delivery, whereas course design specifically referred to the overall structure of the class and, particularly for the experimental group, to milestones 1, 5, and 6. As reported in Figure 1, both groups were satisfied by the course, though they also perceived it as one of the most challenging classes in the program. However, while the experimental group shows a statistically significant difference (P=0.021) in terms of satisfaction compared to the control group (+13%), there is little difference between the groups in terms of perceived difficulty. Students' responses show that being able to work for tangible objectives and achieve them was a motivator. The difference between the experimental and the control groups can be particularly appreciated when considering the learning outcomes and, specifically, the perceived level of expertise at the beginning and at the end of the course (see Figure 2). In this regard, the average of the experimental group (61% and 95%, respectively) showed a 35% increase, whereas the control group (61% versus 89%) had a 28% improvement. In this regard, the experimental group's responses highlighted that the experience, and especially interaction with experts (as discussed later), helped them increase their proficiency. Small differences were found between the groups in terms of factors that had the most impact on the success of the initiatives (Figure 3). However, the control group is more polarized, whereas students in the experimental group realized the importance of all the learning experience components. Furthermore, each step in the process was analyzed from a quantitative and qualitative standpoint to evaluate student satisfaction and learning outcomes and to identify the activities that were perceived as having the most value for students (Figure 4). Surprisingly, in the control group, the value of the activity aligns with the effort required to complete it. As a result, the review of the literature (M2) and the project/study design account for 84% of the total milestones. Conversely, students in the experimental group realized the value of each milestone. Particularly, M5 (student-expert dialogue) and M6 (presentation and publication) scored 22% and 15%, respectively. This reflects the content of several responses in which students mentioned that interacting with an expert and realizing a real-world publication motivated them to learn more and improve the quality of their work.

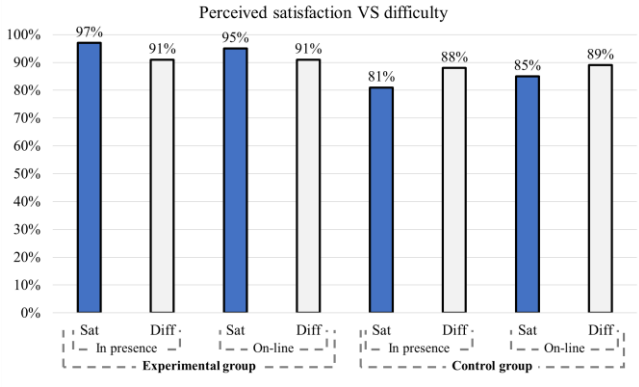


Figure 1. Perceived satisfaction (Sat) and difficulty (Dif) in the experimental and control groups in face-to-face and online classes.

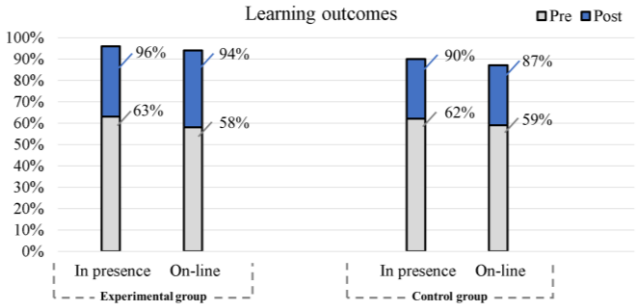


Figure 2. Perceived proficiency with research methods at the beginning (Pre) and at the end (Post) of the course.

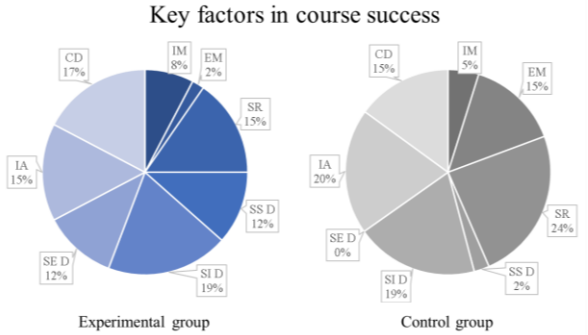


Figure 3. Perceived factors of success: intrinsic (IM) and extrinsic motivation (EM), self-regulation (SR), student-instructor dialogue (SI D), (5) student-student dialogue (SS D), student-expert dialogue (SE D), instructor activities (IA), and course design (CD).

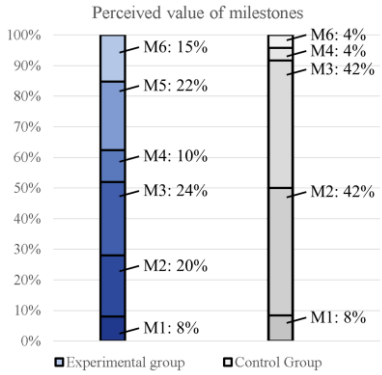


Figure 4. Perceived value of the milestones of the course in terms of learning outcomes: project selection and proposal (M1), review of the state of the art (M2), project/study design (M3), peer review (M4), expert review (M5), and presentation and publication (M6).

Students especially enjoyed the possibility of working on their proposals rather than being assigned to a topic [10] [11] [12] [13]. Most of them were able to independently select a project that was suitable for the course in terms of objectives and workload. Data about the initial selection show that this task was challenging for all the students, regardless of the group. This is in line with research on higher cognitive tasks in Bloom's taxonomy and reinforces the idea that aiming at more ambitious goals did not result in any additional overhead for the experimental group. Moreover, the results show that at the beginning of the class, students in both groups had poor confidence in their skills and lacked familiarity with scholarly research. Defining a direction for their research after reviewing the literature was perceived as the most difficult task. Nevertheless, the data do not show any significant difference between the experimental and the control groups in terms of perceived discomfort, which supports the argument that introducing more ambitious and specific goals that have the same complexity (from a Bloom's taxonomy standpoint) does not result in additional burden for students. The data also show an improvement in terms of satisfaction over the years: this is because students were provided with more examples of the deliveries expected as an output (articles written by alumni from previous years). Students were enthusiastic about the idea of peer review: the challenge of working in a team with others primarily resulted in an incentive to complete the task appropriately. Also, this gave them the opportunity to explore other topics and compare their work to the research done by others. Furthermore, this activity resulted in the highest engagement and was perceived as helpful to increase the overall quality of submissions. Figures 3 and 4 show that the selection of external reviewers was particularly interesting from a learning outcome standpoint. Also, students were able to perceive its value. Moreover, this task provides a unique opportunity to introduce topics that the course would not otherwise explore, such as conflict of interest: for instance, several students initially selected their relatives or coworkers as reviewers. Moreover, receiving opinions from experts other than the instructor was useful for having students understand how their work would be perceived in a professional environment, increasing their awareness of the quality of their deliverables, and having an external stakeholder hold them responsible. Fortunately, most experts agreed to act as reviewers, which, in turn, gratified students and increased their self-confidence. Also, student-expert communication resulted in the most significant improvement in the quality of submissions.

VI. CONCLUSION

This paper presented the case study of the redesign of a research method course in computer science using an experiential learning approach, primarily aimed at stimulating students to realize actual research projects. Specifically, the proposed strategy requires them to accomplish ambitious SMART goals that culminate in potentially publishing the results of their work in an academic conference or scientific journal. The paper outlined the motivations of the work, the teaching methodology, and the expected outcomes. Moreover, it presented the results of an experimental study that measured the effectiveness of the approach over several dimensions of student success. Based on the analysis of 3 years of data, the findings of this study show that exploring real-world projects results in better engagement with course content. Also, the findings support the argument that as research methods classes stimulate higher cognitive levels of Bloom's taxonomy, learning experiences can also incorporate ambitious real-world objectives that produce tangible outputs beyond the classroom.

REFERENCES

[1] H. Hassani, "Research methods in computer science: The challenges and issues". arXiv preprint arXiv:1703.04080. 2017.

[2] D.A. Kolb, R.E. Boyatzis, and C. Mainemelis, "Experiential learning theory: Previous research and new directions." Perspectives on thinking, learning, and cognitive styles, 1(8), pp.227-247. 2001.

[3] F. Alkan, "Experiential learning: Its effects on achievement and scientific process skills." Journal of Turkish Science Education, 13(2), pp.15-26. 2016.

[4] R.S. Rubin, Will the real SMART goals please stand up. The Industrial-Organizational Psychologist, 39(4), pp.26-27. 2002.

[5] G. Falloon, "Using simulations to teach young students science concepts: An Experiential Learning theoretical analysis". Computers & Education, 135, pp.138-159. 2019.

[6] L.H. Mandel, "Experiencing research firsthand: The "unClassroom" experiential learning approach to teaching research methods in a LIS master's program". Journal of Education for Library and Information Science, 58(4), pp.187-201. 2017.

[7] M. Welsh, "We are Humans First: Expanding Learning Outcomes in an Undergraduate Research Methods Course Through an Experiential Learning Project on Homelessness". Journal of Criminal Justice Education, 29(3), pp.414-435. 2018.

[8] S. Waite, and P. Waters, "Mobilising research methods: Sensory approaches to outdoor and experiential learning research". In Research Methods in Outdoor Studies (pp. 78-87). Routledge. 2019.

[9] S.B. Eom, and N. Ashill, "The determinants of students' perceived learning outcomes and satisfaction in university online education: An update". Decision Sciences Journal of Innovative Education, 14(2), pp.185-215. 2016.

[10] Caporusso, N., Zhang, K., Carlson, G., Jachetta, D., Patchin, D., Romeiser, S., Vaughn, N. and Walters, A., 2019, August. User Discrimination of Content Produced by Generative Adversarial Networks. In International Conference on Human Interaction and Emerging Technologies (pp. 725-730). Springer, Cham.

[11] Legleiter, A.M. and Caporusso, N., 2020, July. Flat-Design Icon Sets: A Case for Universal Meanings?. In International Conference on Applied Human Factors and Ergonomics (pp. 211-217). Springer, Cham.

[12] Santiago, K. and Caporusso, N., 2018, July. Addressing Mental Health Consequences of Social Networking from the Social Services Perspective. In International Conference on Applied Human Factors and Ergonomics (pp. 183-192). Springer, Cham.

[13] Caporusso, N., Chea, S. and Abukhaled, R., 2018, July. A game-theoretical model of ransomware. In International Conference on Applied Human Factors and Ergonomics (pp. 69-78). Springer, Cham.

Adaptive Drum Kit Learning System: Impact on Students' Motivation

Mladen Konecki
Faculty of Organization and Informatics, Varaždin, Croatia
mladen.konecki@foi.hr

Abstract – The use of computer technology in the process of learning to play musical instruments is still a relatively unexplored area. In this paper, a conceptual model of an adaptive system for learning to play musical instruments is presented, on the basis of which a prototype for learning to play the drum kit has been created. The results of a study in which the developed prototype has been compared with the alternative methods of self-learning to play a drum kit is also presented and elaborated. The results of conducted study have shown that the students who learned to play a drum kit using a developed system based on the created conceptual model were more motivated during the process of learning compared to the two control groups of the students who have been provided with video tutorial lessons or with an alternative computer system for learning to play a drum kit.

Keywords – automated and adaptive learning system, musical instrument, self-paced learning, drum kit playing, motivation

I. INTRODUCTION

Today, music is one of the most popular media, especially among young people. Research has shown that, among the younger population, 9 out of 10 children want to learn how to play musical instrument, but as they get older the interest is dwindling [1]. Research has also shown that in 54% of households in America someone plays a musical instrument [2].

Children, pupils and students are interested in digital media and it is necessary to have clear and effective methods of including technology and digital media in education in the best possible way so that both students and teachers benefit from it [3; 4]. This paper presents a conceptual model of an adaptive system for learning to play musical instruments.

Based on this conceptual model, a prototype for learning how to play drum kit has been developed. An evaluation was conducted to conclude about the impact of using the created prototype on the users and their motivation to learn. In this paper, the results of conducted research are presented and discussed.

II. BACKGROUND RESEARCH

The use of digital technology in education is not something new, but its application in the process of learning to play musical instruments has not yet reached its full potential. There are several different tools on the market that can be used for this purpose. There are good tools for learning to play the piano, which is the most popular instrument among young people [5]. Since it is relatively easy to connect the piano to a computer via the MIDI interface, it is not surprising that there are many quality programs in this domain: Playground sessions, Piano Marvel, Flowkey [6] and others.

As far as drums are concerned, electric drums can also be easily connected via a MIDI interface to a computer. But acoustic drums require additional hardware, so-called "triggers", and to properly setup this hardware requires additional effort. There are no adequate software solutions on the drum learning market other than the V-Drums Tutor DT-1 developed by Roland [7]. Recently, Roland has created a program "Melodics" which enables the users to buy lessons to learn to play the piano and drums [8]. There are also certain alternatives that come in the form of computer games but can also be considered educational software: for example, DTXMania [9], DrumMania [10], Rock Band [11] and others.

The main feature of all these computer programs is that they retrieve input from a musical instrument and can return feedback information - whether a note is played correctly or not. In addition to computer software, users who want to learn to play a musical instrument on their own also use video tutorials that can be found online or purchased [12].

Regarding scientific research in this field, there are certain tools that have been developed for this purpose: PianoFORTE [13], Piano Tutor [14], the i-Maestro project [15], Digital Violin Tutor [16], IMUTUS [17] and VEMUS projects [18], etc. Although there are several interesting solutions, what is missing in the development of these tools is the adequate evaluation of these tools on a sufficient number of users that would confirm their usability. There have been some evaluations of some of the mentioned tools but not on a larger scale.

Violin Tutor was tested by 12 students [19]. As a part of the IMUTUS project, the evaluation was conducted with participation of 12 students where 6 students used the created system and 6 students were in the control group [20]. Piano Tutor was tested by 18 users of whom 10 completed testing [21]. On the largest project, i-Maestro, the evaluation was never carried out. The same is the case with the PianoFORTE. Based on this, one of valid research directions would be to conduct a more extensive evaluation of mentioned tools.

III. CONCEPTUAL MODEL

In order to provide users with more than just the basic feedback functionality on accurately played notes, a conceptual model of an adaptable system has been created that can generate real-time lessons [22]. Also, this model incorporates the mechanics of advanced detection of various types of errors and generating lessons for faster elimination of these errors during playing [23]. Previous research has shown that there is a justification for developing a system that would have such functionalities [24]. Research has also been conducted to create a user interface that will adequately meet user needs [25]. The final, version of the conceptual model is shown in Fig. 1.

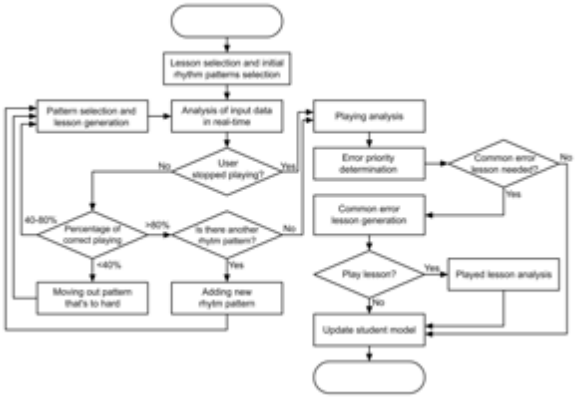


Figure 1. Adaptable drum kit learning system: conceptual model

The lesson that is generated for the particular user becomes easier or harder over time, depending on the percentage of correctly played notes. The model has been upgraded with frequent error analysis, and if it is detected that some type of error occurs frequently, a lesson is generated to help the user correct that type of error: a lesson is created containing rhythmic patterns in which this type of error occurs and the user can practice correcting that kind of mistake.

IV. DEVELOPED PROTOTYPE

Based on the created conceptual model, a prototype of a computer system for learning to play drums was created. The system consists of standard elements of intelligent teaching systems: domain model, user model, teaching model and user interface [26]. In the knowledge base, 144 rhythmic patterns of one beat were created, which were grouped within 9 categories. There are 4 types of playing learning activities in the system:

1. Video lessons - this is an activity where the user receives instructions in the form of videos
2. Exercise lessons - the activity of practicing playing a certain rhythmic pattern with a video lesson on how to play that specific rhythmic pattern
3. Jam Sessions - learning to play all patterns of a particular category of rhythmic patterns (dynamically generated lesson)
4. Review lessons - playing all rhythmic patterns of all categories so far (dynamically generated lesson)

Initially, the users need to create their own user profiles in which all the playing data will be stored. After that, the users come to the main screen of the prototype where they select the activity they want to activate (Fig. 2). There is a certain path that is suggested to the users, but the users have complete freedom over the choice of activity they want to play. Also, the users can repeat each activity as many times as they want.



Figure 2. Main screen of developed prototype

This screen also gives the feedback to the users on the performance of each activity in the form of a percentage. For each individual rhythmic pattern, the last 4 plays of that rhythmic pattern are memorized, and the percentage of correctly played notes is shown. The rhythmic pattern is declared learned when the users successfully play a single pattern 4 times in a row without error.

In the "Jam Session" activity (Fig. 3), the users play all the rhythmic patterns of a certain category and get the percentage of all the rhythmic patterns of that category that have been learned. When generating that lesson, patterns that are learned appear less frequently, and patterns that are in the process of learning appear more often to optimize learning time. In the "Review lesson" activity, there is also a higher chance of appearance of a pattern that has not been fully learned compared to the patterns that have been learned.

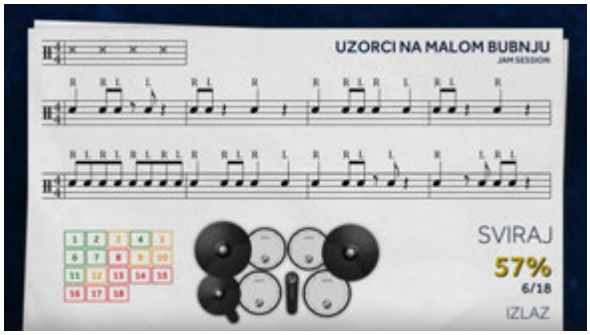


Figure 3. Jam Session screen

A special type of activity is a frequent error correction activity that becomes available to the users only if there is some type of playing error that is repeated frequently. Many types of errors that can be detected are supported [23].

V. CONDUCTED RESEARCH

In order to determine the efficiency of using a computer system in the teaching process based on the

newly created conceptual model and to determine whether the use of such system can positively affect users' motivation, an evaluation was conducted involving 90 students divided into 3 groups of 30 students:

1. The group that used the created prototype (experimental group)
2. The group that used alternative software DT-1 Drum Tutor (control group 1)
3. The group that used video lessons (control group 2)

To ensure the uniformity of all three groups, an exercise was created where students had to repeat 15 different rhythmic patterns by tapping their hands on the table. Based on the results, 3 uniform groups were created. In each group there were:

- 3 students who correctly played 14 patterns
- 5 students who correctly played 13 patterns
- 7 students who correctly played 12 patterns
- 6 students who correctly played 11 patterns
- 5 students who correctly played 10 patterns
- 4 students correctly played 9 patterns

The knowledge base for all 3 groups was the same: 144 rhythmic patterns in 9 sample categories. The first group used a created prototype where they had video lessons, feedback on accurately played notes through the user interface, and dynamic generation of lessons in real time based on playing precision. The group that used the DT-1 Drum Tutor had video lessons and feedback on accurately played notes via the program interface. The third learning group used only video lessons and had no feedback on playing accuracy. These are three typical scenarios of how individuals can learn to play drums on their own. Each student learned to play during 10 school hours by using a certain method.

In order to research the perception of students' motivation when learning to play drums by a certain method, two motivation theories were used: the "theory of flow" [27] and the theory of self-determination [28]. A standardized questionnaire developed by Jackson and Marsh [29] was used to measure flow state. A standardized measurement scale in the context of education was used to measure motivation by self-determination theory [30].

VI. Research Results

Since the data by which students' perception of their motivation was to be measured have been continuous numerical values and the groups were independent, the independent variable was categorical and the groups have been of equal size, the data was analyzed by ANOVA method.

In order for it to be used, two other assumptions had to be met: the normality of the distribution and the homogeneity of the distribution. The normality of distributions was tested by the Shapiro-Wilk test and the Kolmogorov-Smirnov test and the homogeneity of the distributions was tested by the Levene test. If the normality of the distribution is not satisfied, the Kruskal-Wallis H test is used and if the homogeneity of the distributions is not satisfied, the Welch test is used. The process of selection of statistical methods for analysis is shown in Fig. 4.

According to the theory of self-determination, intrinsic motivation is measured by a standardized questionnaire to measure the perception of motivation when performing an activity.

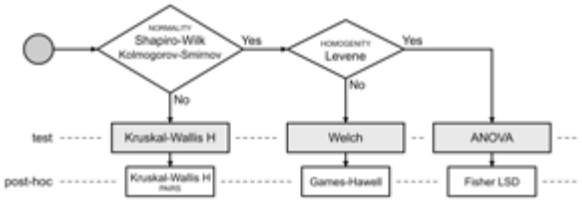


Figure 4. Selection of statistical methods

The questionnaire consists of 25 statements and the scale of agreement ranges from 1 to 7. Table I shows that the second group does not meet the normality of the distribution according to the Kolmogorov-Smirnov test and according to the Shapiro-Wilk test the first and second groups do not meet the normality of the distribution.

TABLE I. Kolmogorov-Smirnov test and Shapiro-Wilk test

	Kolmogorov-Smirnov			Shapiro-Wilk		
Group	Stat.	df	Sig.	stat.	df	Sig.
(1)	0.149	30	0.090	0.924	30	0.033
(2)	0.196	30	0.005	0.878	30	0.003
(3)	0.139	30	0.147	0.954	30	0.218

Given these results, a nonparametric Kruskal-Wallis test was performed for further analysis (Table II).

TABLE II. Kruskal-Wallis test

Group	N	Mean rank
(1)	30	51.65
(2)	30	49.10
(3)	30	35.75

	Motivation
Chi-square	6.431
df	2
Asymp. Sig.	0.040

Given that the significance is less than 0.05, this means that there is a statistically significant difference in the average value of motivation perception between the observed groups. After further analysis, it was found that there is a statistically significant difference between the first and third test groups, $X^2(1) = 5.823$, $p = 0.016$, with a rank average: 35.93 for the first group and 25.07 for the third group. The percentage to which the difference can be attributed to a different learning method is 9.87%. Regarding the first and second group, although there is a difference in the average rank (first group 34.82, second group 25.18), this difference did not prove statistically significant in this case ($p = 0.055$).

Flow state perception is measured by a standardized questionnaire consisting of 36 statements and scoring is performed by a standard Likert scale ranging from 1 to 5. In Table III the results of testing the normality of distributions are presented.

TABLE III. Kolmogorov-Smirnov test and Shapiro-Wilk test

	Kolmogorov-Smirnov			Shapiro-Wilk		
Group	Stat.	df	Sig.	stat.	df	Sig.
(1)	0.142	30	0.125	0.947	30	0.138
(2)	0.125	30	0.200	0.936	30	0.070
(3)	0.062	30	0.200	0.985	30	0.940

According to the obtained data, all three groups satisfy the normality of the distribution. The Levene test also

showed that the homogeneity of the distributions was met (Table IV).

TABLE IV. LEVENE TEST

LEVENE STATISTIC	DF1	DF2	SIG.
1.398	2	87	0.253

Given these data, a one-factor analysis of variances was performed (Table V).

TABLE V. ONE-FACTOR ANALYSIS OF VARIANCES

SOURCE OF VARIATION	SS	DF	MS	F	SIG.
BETWEEN GROUPS	3156.689	2	1578.344	3.429	0.037
WITHIN GROUPS	40049.133	87	460.335		
TOTAL	42205.822	89			

Given that the significance is less than 0.05, it can be concluded that there is a statistically significant difference between the observed test groups. As a post hoc test, Fisher's LSD test was performed (Table VI).

TABLE VI. FISHER'S LSD TEST

(I)GROUP	(J)GROUP	MEAN DIFF.	STD. ERROR	SIG.
(1)	(2)	1.633	5.540	0.768
	(3)	13.300	5.540	0.018
(2)	(1)	-1.633	5.540	0.769
	(3)	11.667	5.540	0.038
(3)	(1)	-13.300	5.540	0.018
	(2)	-11.667	5.540	0.038

Based on the obtained data, it is possible to conclude that there is a statistically significant difference in the level of flow status between the first and third test groups. There was no statistically significant difference between the first and second group ($p = 0.769$).

It is interesting to note that there is also a statistically significant difference in the level of flow status between the second and third groups ($p = 0.038$). This data indicates that the use of interactive computer software can further motivate the user compared to learning only with the help of video tutorials.

VII. CONCLUSION

In the process of education, motivation is very important. Lack of motivation can have a very negative impact on the quality of the teaching process as well as on the learning outcomes. Based on the obtained data, it can be concluded that the general use of computer software that provides real-time feedback on the accuracy of playing can positively affect the motivation of students. This conclusion can be made based on the results of the state of flow perception of test groups in the study.

Also, the theory of self-determination has shown that motivation can be significantly influenced by the method of learning. Given that there was no statistical difference between the second and third groups in the perception of intrinsic motivation of users and that difference exists between the first and the second group, it can be concluded that dynamically generated lessons in real time had a positive effect on students' motivation.

REFERENCES

[1] Classic FM, (February 25, 2020) Available at: https://www.classicfm.com/music-news/children-learning-musical-instruments-survey/

[2] C. Carlsbad. NAMM: "Gallup Organization Reveals Findings of "American Attitudes Toward Making Music" Survey," (February 23, 2020) Available at: https://www.namm.org/news/press-releases/gallup-organization-reveals-findings-american-atti

[3] G. L. Pierce, and P. F. Cleary, "The K-12 educational technology value chain: Apps for kids, tools for teachers and levers for reform," Education and Information Technologies, vol. 21, no. 4, 2016, pp. 863-880.

[4] P. G. Taylor, and B. S. Carpenter, "Mediating art education: Digital kids, art, and technology." Visual Arts Research, 2007, pp. 84-95.

[5] Creative Soul Music School. "What is the Most Popular Instrument to Play?" (February 23, 2020) Available at: https://www.creativesoulmusic.com/news/what-is-the-most-popular-instrument-to-play

[6] P. Guha. "Best Piano Learning Software – Top 5 Review and Picks" (February 23, 2020) Available at: https://www.mydigitalpianoreview.com/best-piano-learning-software.html

[7] V-Drums Tutor DT-1. (February 25, 2020) Available at: https://www.roland.com/global/products/dt-1/

[8] Melodics. (February 25, 2020) Available at: https://melodics.com/

[9] DTXMania. (February 26, 2020) Available at: https://osdn.net/projects/dtxmania/

[10] GuitarFreaks and DrumMania. (February 26, 2020) Available at: https://en.wikipedia.org/wiki/GuitarFreaks_and_DrumMania

[11] Rock Band (video game). (February 26, 2020) Available at: https://en.wikipedia.org/wiki/Rock_Band_(video_game)

[12] Electronic Drum Advisor, "The 15 Best Drum Lesson Resources Online – Free and Paid Options" (February 26, 2020) Available at: https://www.electronicdrumadvisor.com/the-best-drum-lessons-online/

[13] S. W. Smoliar, J. A. Waterworth, and P. R. Kellock, "pianoFORTE: a system for piano education beyond notation literacy." In Proceedings of the third ACM international conference on Multimedia, 1995, pp. 457-465.

[14] R. B. Dannenberg, M. Sanchez, A. Joseph, P. Capell, R. Joseph, and R. Saul, "A computer based multi-media tutor for beginning piano students." Journal of New Music Research. vol. 19, no. 2-3, 1990, pp. 155-173.

[15] K. Ng, and P. Nesi, "i-Maestro framework and interactive multimedia tools for technology-enhanced learning and teaching for music." In Automated solutions for Cross Media Content and Multi-channel Distribution, 2008, pp. 266-269.

[16] G. Percival, Y. Wang, and G. Tzanetakis, "Effective use of multimedia for computer-assisted musical instrument tutoring." In Proceedings of the international workshop on Educational multimedia and multimedia education, 2007, pp. 67-76.

[17] E. Schoonderwaldt, K. Hansen, and A. Askenfeld, "IMUTUS an interactive system for learning to play a musical instrument." In Proceedings of the International Conference of Interactive Computer Aided Learning, 2004, pp. 143-150.

[18] D. Fober, S. Letz, and Y. Orlarey, "VEMUS-Feedback and groupware technologies for music instrument learning." In Proceedings of the 4th Sound and Music Computing Conference SMC, 2007, pp. 117-123.

[19] J. Yin, Y. Wang, and D. Hsu, "Digital violin tutor: an integrated system for beginning violin learners." In Proceedings of the 13th annual ACM international conference on Multimedia, 2005, pp. 976-985.

[20] S. Raptis, A. Chalamandaris, A. Baxevanis, A. Askenfeld, E. Schoonderwaldt, K. F. Hansen, D. Fober, S. Letz, and Y. Orlarey, "Imutus-an effective practicing environment for music tuition." In Proceedings of the International Computer Music Conference, 2006, pp. 383-386.

[21] R. B. Dannenberg, M. Sanchez, A. Joseph, R. Joseph, R. Saul, and P. Capell, "Results from the piano tutor project." In Proceedings of the Fourth Biennial Arts and Technology Symposium, 1993, pp. 143-150.

[22] M. Konecki, "Learning to play musical instruments through dynamically generated lessons in real-time based on adaptive learning system." In Proceedings of the 25th Central European

Conference on Information and Intelligent Systems, 2014, pp. 124-129.

[23] M. Konecki, "Adaptive Drum Kit Learning System: Advanced Playing Errors Detection." In Proceedings of the 18th International Multiconference Information Society-IS, 2015, pp. 58-62.

[24] M. Konecki, "Self-paced computer aided learning of music instruments." In Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015, pp. 809-813.

[25] M. Konecki, "Adaptive drum kit learning system: User interface properties and features." In Proceedings of the 18th International Multiconference Information Society - IS 2015, VIVID 2015, 2015, pp. 228-236.

[26] S. Stankov, V. Glavinić, and M. Rosić, Intelligent Tutoring Systems in E-Learning Environments: Design, Implementation and Evaluation. New York, IGI Global, 2010.

[27] M. Csikszentmihalyi, Flow: The psychology of optimal experience. New York, HarperPerennial, 1990.

[28] R. M. Ryan, and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." American psychologist, vol. 55, no. 1, 2000, pp. 68-78.

[29] S. A. Jackson, and H. W. Marsh, "Development and validation of a scale to measure optimal experience: The Flow State Scale." Journal of sport and exercise psychology, vol. 18, no. 1, 1996, pp. 17-35.

[30] Self-determination theory. (February 2, 2020) Available at: http://www.selfdeterminationtheory.org/theory/

[31] S. S. Shapiro, and M. B. Wilk, "An analysis of variance test for normality (complete samples)." Biometrika, vol. 52, no. 3/4, 1965, pp. 591-611.

Impact of Distance Learning on Motivation and Success Rate of Students During the COVID-19 Pandemic

Mario Konecki
University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
mario.konecki@foi.hr

Abstract – COVID-19 pandemic has created a whole new set of challenges in all aspects and fields of education. Teachers and professors had to quickly implement online teaching procedures and use various tools for online communication and collaboration. Although this kind of teaching has been used at some universities, for many professors and teaching professionals this kind of approach was something completely new. But, it was not only new for the teachers, it was also new for the students. In this paper, an overview of challenges that are related to online education process is presented and elaborated. The research results about the impact of distance learning on motivation and success rate of students are also given and discussed.

Keywords **– COVID-19, online education, distance learning, teachers, students, challenges, motivation, success rate**

I. INTRODUCTION

COVID-19 pandemic has introduced a number of new challenges in the education system that teachers and professors had to deal with quickly and effectively. Distance learning has suddenly become the main way of educating students. And although distance learning was a common practice at some universities, for the majority of education professional this kind of teaching was something completely new and unfamiliar. Because of this, a quick and effective response was required. Universities have organized various forms of online support and education for their teachers, and teachers have made additional effort to react quickly.

However, many teachers have experienced a number of challenges. Some of these challenges have been dealt by the universities' support staff and some have been left to the teachers. In general, most of the teachers succeeded to quickly adapt and provide the students with a suitable distance learning materials and teaching methods. In this paper, an overview of challenges that have emerged regarding distance learning are presented and elaborated.

Another important aspect of distance learning is student motivation and success rate since students have found themselves in a completely new learning environment. Distance learning has made students somewhat less connected with their colleagues, but it has also enabled them to save traveling time and to work from the comfort of their homes. Students also had to do more self-management and had to organize their work. In this paper, the research results about the impact of distance learning on students' motivation and their success rate are presented and discussed.

II. DISTANCE LEARNING AND STUDENT MOTIVATION

COVID-19 pandemic has greatly affected the education system [1; 2; 3; 4]. A number of challenges have been put in front of universities, schools and teachers.

Existing research has shown a number of issues that have been identified [4]:

- Richer countries can access online education easily with the efforts of students and teachers but it is difficult for the developing or under developed countries to provide online education smoothly.
- Many students do not have laptop, mobile or Internet connectivity at home due to which it is difficult to attend online class.
- Spending long hours on laptops or mobile phones can lead to a cause of stress for teachers as well as students.
- Schools or colleges not only provide learning to the students but facilitate the students for other curricular activities and make them socially active.
- Long period disengagement of students with the schools and colleges can lead to increase in dropout rates.
- Classroom learning is preferred more over the online learning as classroom learning ensures more involvement of students.
- This pandemic has also effects on the employment rate as due to slow down of the economy there may be decrease in the rate of job offers.
- Students feel less motivated to attend online class as there may be no proper environment of study at home so only motivated students can benefit from the courses.
- Students are confused about their next steps as board exams, competitive exams, internships and trainings are cancelled during this time.

- Students are responsible for their learning during online learning.

In order to deal with the identified issues several steps have been proposed as a part of existing research [4]:

- Education ministries may understand the challenges faced by the students regarding Internet connectivity, computers, etc.
- T.V. and radio are the powerful tools which can be used to provide education to different sectors. The content may be delivered through these tools and teachers or students may contact with each other through social media such as WhatsApp, Facebook or SMS.
- Parents should support the children in studies as now parents have to play active role in the studies of the children.
- E-learning must be combined with classroom learning so as to take the best advantage of this combination.
- Measures should be taken to mitigate the effect of pandemic on the job offers and employment rate.

Online education is just one of several possible educational methodologies but it has practically overnight become the main way of providing education. The main educational methodologies are [5]:

- The traditional classroom education, where books and blackboards are used by the teacher as a teaching aid.
- Modern classroom education, where the classrooms are equipped with whiteboards, projectors or audio-visual display equipment and digital boards.
- Online education, where the information technologies and communications are used to help in the development and acquisition of knowledge from the different remote locations. It uses the Internet, video/audio and text communication, and software to create the learning environment.

When talking about online education then several types of online learning can be mentioned [5]:

- Knowledgebase
- Online support
- Asynchronous training
- Synchronous training
- Hybrid training

Knowledgebase type refers to lessons and other materials that are made available online. However, there is no online support from the teachers. Online support adds this kind of support to students via forums, e-mails or other communication channels. Asynchronous training refers to a form of education where lectures are prerecorded and made available online. Teachers also communicate with students on a regular basis. Synchronous training includes lectures in real time with live teachers support. Hybrid training combines online and live in-person education.

III. STUDENT SUCCESS RATE

Student motivation is a critical factor in overall student success rate [6]. Since students have to do more self-management in distance learning environment their motivation to learn has become even more important factor.

In order to support and increase students' motivation to learn in online learning environment it is very important to take appropriate steps and include necessary elements. One of the steps is to explain precisely and clearly what is the goal of a particular assignment. This includes the following [7; 8]:

- Explain to students why the task is important and interesting to them. It may be useful to link the task to practices that the students may use in their professional life.
- Define the learning objective of the task. Such objectives will identify the performance standards that a student needs to meet to reach the desired goal.
- Give advice in relation to the time required to complete the activity.
- Provide preliminary exercises that the students can practice, thereby building their confidence and boosting their motivation.

It is well known that intrinsic motivation is one of the crucial factors of success and this is also true in the case of online learning [9; 10; 11; 12; 13].

Another factor that can foster better focus of students is the feeling that other students are present in an online environment [14].

Research has shown that teachers' role is important in both offline and online education environment, especially regarding the motivation of students to learn and self-regulate their learning efforts [15]. In other words, the teachers' role does not end with supporting students in their assignments, course matter comprehension, and collaboration activities using appropriate technology, but continues with supporting and engaging students in order to boost their motivation to learn and achieve new goals.

Many research results have found that the quality and quantity of students' interaction and their satisfaction with the learning experience is promoted by teachers' presence and feedback, teaching efforts and encouraging collaboration [16; 17; 18; 19]. Students tend to value the flexibility of online education [19], however professor and teachers in many cases find challenging to address all learning styles and individual learning needs of each particular student [20]. One of more important demotivation factors is the isolation and anonymousness of each student [19].

There are many online communication tools that can be used for the purpose of distance learning. Some of the more know tools are:

- Zoom
- Webex
- BlueJeans

- Google Meet
- Microsoft Teams
- Adobe Connect
- BigBlueButton
- Jitsi

In order to select the most appropriate tool several factors have to be taken into consideration such as: the number of participants, audio and video quality, latency, complexity of the user interface, pricing, additional features, etc.

IV. METHODOLOGY AND RESEARCH RESULTS

In the continuation of this paper, the research results about the impact of distance learning on student motivation and success rate are presented and discussed. The research has been conducted among the first-year students of information technology study program during the distance education about programming (on the course Programming 1).

As soon as the COVID-19 pandemic began, the preparations for distance learning have started since it was the intention of the teachers to provide students with as similar experience compared to in-person education as possible. A number of tools have been analyzed and in the end Zoom has been selected as the tool that will be used in distance synchronous education. During the semester, Zoom has been used for lectures and individual or group consultations. The testing has been implemented by using own developed web system PTG (Programing Tasks Generator) [21; 22]. PTG is shown in Fig. 1.

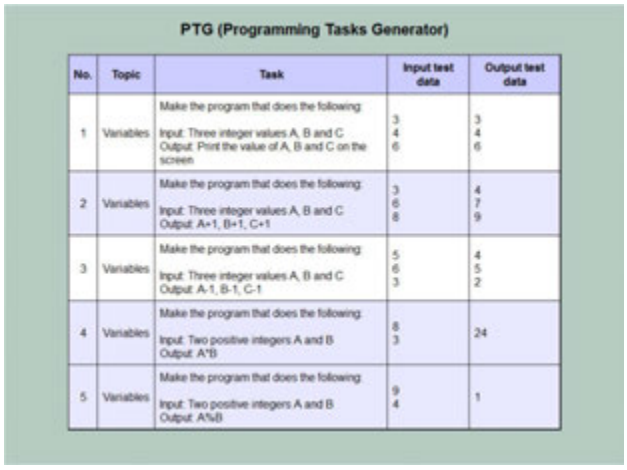


Figure 1. Web system PTG

Web system PTG enables individual generation of programming tasks for each student in a way that each student gets individual set of tasks. The system is available to students in discrete time slots and there is a limited time in which students have to submit the source code of solved programming tasks.

All asynchronous activities have been supported by the open-source learning management system Moodle (lecture materials, program examples, notifications and forum).

Theoretical tests have been created in the learning management system Moodle by following several principles that have been identified as important in online learning environment in order to get as objective results as possible:

- Limited testing time
- Limited time slots for each question
- Answering question by question without the possibility to go back
- Disabled copying of the content

Since part of the course was held in-person students were able to compare in-person lectures and course activities with distance learning education. To conclude about the quality of conducted distance learning students were asked to participate in the research via provided questionnaire (questionnaire items are shown in Table I).

More than half of first-year students of information technology study program taking the course about programming have participated in the research (132 students). The research was conducted using an online questionnaire in which all possible answers except the first one were based on the Likert scale with possible responses ranging from “Strongly disagree” (1) to “Strongly agree” (5). Research results are shown in Table I.

TABLE I. RESULTS OF CONDUCTED RESEARCH

Questionnaire item	Mean	Std. dev.
Compared to classical teaching, I consider distance learning: 1. To be of much lower quality and much less interesting 2. To be of lower quallity and less interesting 3. To be of the same quality and equally interesting 4. To be of higher quality and more interesting 5. To be of much higher quality and much more interesting	3,69	0,91
The course content presented to me via distance learning was understandable to me.	4,21	0,87
During the distance lectures, interaction with students was achieved.	4,62	0,62
I find the technology chosen for distance learning to be interesting and of high quality.	4,37	0,78
Distance learning has increased my motivation to learn the course content.	3,55	1,29
What grade would you give to the course teacher?	4,40	0,74
I want a part of the course to be organized via distance learning in the following academic years.	4,10	1,23

Research results have shown that students are satisfied with provided distance learning education and that they perceive it to be of similar and even slightly higher quality, as well as a bit more interesting, than classical in-person teaching. It was shown that presented course content was understandable to students, that necessary level of interaction was achieved and that used technology was chosen well. The teacher has been given high grades

by the students. This has shown that the way of teaching and interacting with students was organized and implemented well. Majority of students have stated that they would like to continue with distance learning in the following academic years.

The last aspect that was researched was the motivation of students. Research results have shown that the motivation of students to learn has increased in majority of the students. However, there are also students who simply prefer face-to-face communication and teaching. This is also reflected in standard deviation value.

All mentioned conclusions have also been confirmed by analysis of the comments that have been gathered from the students in the research process. Several student comments are given below:

- "Despite the unfortunate events that befell us, I believe that distance learning was the best possible solution that does not differ so much from classical teaching. The content is presented to us in the same way as if we were physically present in the lecture room, only we can follow the lectures from the comfort of our home."
- "I think that the lectures of the course Programming 1 are an example of what distance learning should look like. I also think that this way of teaching encourages greater activity of students in lectures."
- "The technology and applications that are used are great but I think there are too many different applications used across all the courses (I think there should only be one or two applications that are used on all courses). I absolutely agree that this kind of teaching increases my motivation to learn, although I don't exactly know why."
- "It is easier for me to understand the course materials via online teaching."
- "To begin with, I must commend the efforts to keep classes at a decent level despite the current situation. I think part of the teaching can be organized remotely without any problems. The current situation has forced me to work more for this course and I am currently more successful."
- "Distance learning is a very good way to replace the classical teaching, and over time it seems to me to be better than teaching in normal circumstances."
- "I am very satisfied with this way of teaching because I am much more concentrated than in a classic indoor lecture room. I absorb the course material more easily and I think it is a better way of teaching."
- "Much easier and more flexible way of following the lectures, equal quality and greater attendance rate that can sometimes raise the quality of teaching by enabling answers to additional questions or ambiguities that students may not take into account as much. All in all, better and easier :)"
- "I really like distance learning because those who don't want to follow the lectures don't bother me."
- "All is well! Classes are very close to their live version."
- "Excellent lectures, I hope that such way of teaching will be present in the future as well."

Regarding the student success rate in the distance learning environment, the results of the final practical and theoretical testing have been analyzed. The most demanding part of overall testing was the practical part that was conducted via web system PTG. Students have been given 4 programming tasks and they could be given from 0 to 20 points depending on their submitted programming solutions. Average number of points in the practical part of testing was 13,89 with standard deviation of 4,27. Distribution of points in the practical part of testing is shown in Fig. 2.

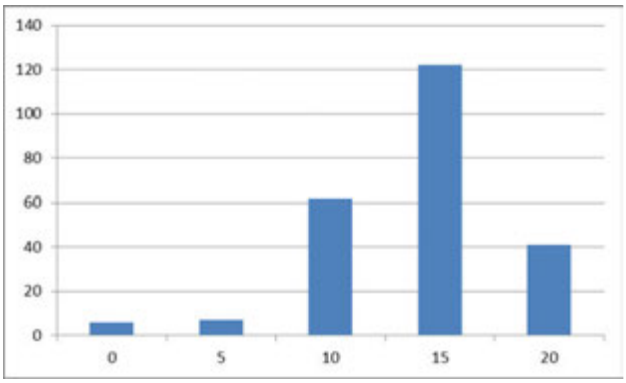


Figure 2. Distribution of points in the practical part of testing

In the theoretical part of testing students could be given from 0 to 14 points on the first test and from 0 to 16 points on the second test. The average number of points on the first test was 7,60 with standard deviation of 3,62 and 11,61 on the second test, with standard deviation of 3,37.

Based on the presented testing results it can be concluded that distance learning has been conducted successfully, that student success rate was good, and that distribution and averages of gained points were mostly in the expected or slightly better range.

Many students have shown slightly better knowledge and skills than expected and this shows that even the programming courses that are known to be hard to teach and learn [23; 24] can be taught via distance learning in high-quality and pedagogically sound manner.

V. CONCLUSION

COVID-19 pandemic has created a number of challenges for universities, schools and other educational institutions. By quick reaction, a number of available technological solutions, and by additional effort new teaching models have been implemented. Overview of mentioned challenges has been presented in this paper, along with the overview of factors that can affect student motivation and success rate.

Research results regarding programming course that has been conducted via distance learning have been presented and discussed. The results have shown that

students have perceived online programming course and distance learning as something interesting and that the course content has been presented in understandable and clear manner. Students' motivation to learn has also been slightly increased and they stated that distance learning is something they would like to be present in the future teaching activities. Students have graded their teacher with high grades and they have stated that the interaction was achieved in a way that is useful to students via interesting and easy-to-use communication tools.

Students' success rate has also been achieved at a level that would be expected, and even a slightly higher knowledge and skill level has been acquired and demonstrated by many students.

It can be concluded that distance learning can be conducted in an organized, effective and pedagogically sound manner, even in the case of programming courses that are known to be hard to teach and learn.

REFERENCES

[1] Sahu, P. (2020). Closure of universities due to Coronavirus Disease 2019 (COVID-19): impact on education and mental health of students and academic staff. Cureus, 12(4).

[2] Mian, A., & Khan, S. (2020). Medical education during pandemics: a UK perspective. BMC medicine, 18(1), 1-2.

[3] Moszkowicz, D., Duboc, H., Dubertret, C., Roux, D., & Bretagnol, F. (2020). Daily medical education for confined students during COVID‐19 pandemic: A simple videoconference solution. Clinical Anatomy.

[4] Sahib, S. M. (2020). A Study on The Perspective of Students Regarding Online Education During Covid-19 Pandemic. Purakala. ISSN 0971-2143. UGC CARE Journal, 31(42), 152-168.

[5] Basilaia, G., & Kvavadze, D. (2020). Transition to online education in schools during a SARS-CoV-2 coronavirus (COVID-19) pandemic in Georgia. Pedagogical Research, 5(4), 1-9.

[6] Ushida, E. (2005). The role of students' attitudes and motivation in second language learning in online language courses. CALICO journal, 49-78.

[7] Nehme, M. (2010). E-learning and students' motivation. Legal Educ. Rev., 20, 223.

[8] Laurillard, D. (2013). Rethinking university teaching: A conversational framework for the effective use of learning technologies. Routledge.

[9] Rakes, G. C., & Dunn, K. E. (2010). The Impact of Online Graduate Students' Motivation and Self-Regulation on Academic Procrastination. Journal of Interactive Online Learning, 9(1).

[10] Xie, K., & Ke, F. (2011). The role of students' motivation in peer‐moderated asynchronous online discussions. British Journal of Educational Technology, 42(6), 916-930.

[11] Xie, K., Durrington, V., & Yen, L. L. (2011). Relationship between students' motivation and their participation in asynchronous online discussions. MERLOT Journal of Online Learning and Teaching, 7(1), 17-29.

[12] Yang, Q. (2014). Students motivation in asynchronous online discussions with MOOC mode. American Journal of Educational Research, 2(5), 325-330.

[13] Beluce, A. C., & Oliveira, K. L. D. (2015). Students' motivation for learning in virtual learning environments. Paidéia (Ribeirão Preto), 25(60), 105-113.

[14] Deng, L., & Tavares, N. J. (2013). From Moodle to Facebook: Exploring students' motivation and experiences in online communities. Computers & Education, 68, 167-176.

[15] Fryer, L. K., & Bovee, H. N. (2016). Supporting students' motivation for e-learning: Teachers matter on and offline. The Internet and Higher Education, 30, 21-29.

[16] Dennen, V. (2005). From message posting to learning dialogues: Factors affecting learner participation in asynchronous discussion. Distance Education, 26(1), 127–148.

[17] McIsaac, M. S., Blocher, M. J., Mahes, V., & Vrasidas, C. (2006). Student and teacher perceptions of interaction in online computer-mediated communication. Educational Media International, 36(2), 121–131.

[18] McIntyre, S. (2011). Learning to teach online: Engaging and motivating students. http://www.academia.edu/612888/Engaging_and_motivating_students. Accessed May 1, 2020.

[19] Gedera, D., Williams, J., & Wright, N. (2015). Identifying factors influencing students' motivation and engagement in online courses. In Motivation, leadership and curriculum design (pp. 13-23). Springer, Singapore.

[20] Vanslambrouck, S., Zhu, C., Lombaerts, K., Philipsen, B., & Tondeur, J. (2018). Students' motivation and subjective task value of participating in online and blended learning environments. The Internet and Higher Education, 36, 33-40.

[21] Konecki, M. (2015). Improving Students'Focus in Introductory Programming Courses. Daaam International Scientific Book 2015, 165-172.

[22] Konecki, M., Lovrenčić, S., & Jervis, K. (2016). The use of assistive technology in education of programming. In 2016 ICBTS International Academic Research Conference Proceedings, Boston, USA (pp. 1-11).

[23] Konecki, M. (2014). Problems in programming education and means of their improvement. DAAAM international scientific book, 2014, 459-470.

[24] Konecki, M., & Petrlic, M. (2014). Main problems of programming novices and the right course of action. In Central European Conference on Information and Intelligent Systems (p. 116). Faculty of Organization and Informatics Varazdin.

Analysis and Development of a Robotic Disk Mounting Stand

Florian Koefler, Manuel Edlinger, Nikolas Jausz, Winfried Egger,
Wolfgang Werth and Lisa-Marie Faller
Engineering & IT, Systems Engineering,
Carinthia University of Applied Sciences (CUAS)
Europastraße 4, 9524 Villach

***Abstract*—Industry 4.0 and related automation concepts are nowadays vivid fields of research. Even more, such concepts are applied where humans have to work on tedious and dull or dangerous and dirty tasks. Additionally, small companies can also fill gaps where they do not find suitable personnel. Consequently, this industry is also reflected in our study program, the related automation and robotics lectures as well as in the Bachelor projects which carried out in our lab. In this work, we first present the teaching concept developed for robotics lectures and lab. Then, we present the successful application of the taught matter for the development of an automation concept for an industrial assembly task using a robot manipulator together with a programmable logic controller. In the target application a robot is used to assemble and glue 215 carbide blades into a steel disk to fabricate gearing components. In order to successfully implement such a concept, the process is consequently broken down in substeps where each one is analyzed individually.**

***Index Terms*—Industrial robotics, Robotics education, Disc mounting, Robot manipulator.**

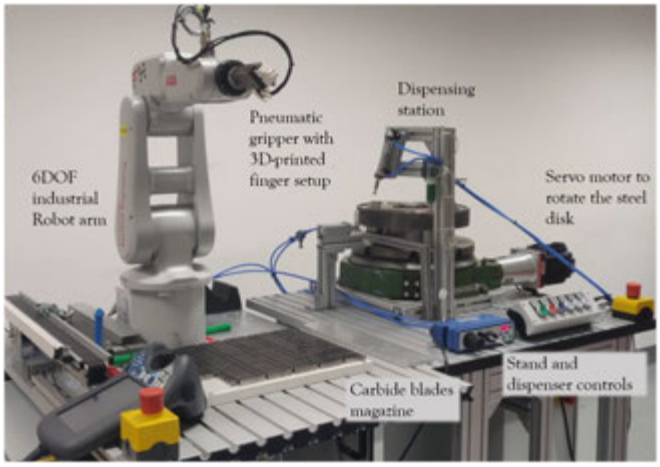


Fig. 1: Illustration of the complete assembly system.

I. Introduction

Industry 4.0 and process automation are prospering engineering fields with increasing importance [1]. Recent research targets the automation of field robotics and sensor development [2], but also manipulation strategies [3] and safety [4]. A specific trend is noticeable towards automation of dull, dirty and dangerous tasks in the context of industrial fabrication processes [5]. Consequently, robotic systems are more and more employed in applications where a human operator is in danger, has to carry out monotonic processes or is subject to excessive contamination. Together with all the advantages the co-working robots provide to our working structure, we still need to guarantee at the same time crucial the safe co-operation with human operators [6]. The application of robotic concepts to tedious industrial processes is an important part in engineering education. We consequently foster the work of students on these issues. One other main focus of our University is mechanical construction and 3D-printing in teaching as well as research [7]. So, we additionally support students in the design and fabrication of novel structural components using 3D-printing. In this context, a student project has been carried out in which a disc mounting process at an industry cooperation partner has been analyzed and implemented. The developed assembly system is shown in Fig. 1.

II. Task Related Robotics Lectures and Lab Exercises

Starting with the 4th semester of our Bachelor program in Systems Engineering, students have to choose one of three specifications. Additionally, they are also encouraged to start with their Bachelor project work. The presented project work is done by students who are partially enrolled in the specializations of automation technologies as well as robotics. As part of the specialization, students also take robotics lectures and lab exercises in order to prepare them properly for their future work in the industrial robotics field.

The robotics lectures and exercises are designed in a way that the material presented in the lectures theoretically backs up the exercises done in the respective lab. The number of students in each specialization is usually 10 at max. It is consequently comparably easy to form small students groups who can work jointly on the exercises. Additionally, also the available hardware can be used efficiently.

A. Lectures

The robotics lectures are basically subdivided into mainly theoretic parts on one hand and practical implementation of these theoretic concepts using the Robotics toolbox in Matlab [8]: after each four theoretic blocks one practical block with Matlab exercises is introduced. The content of

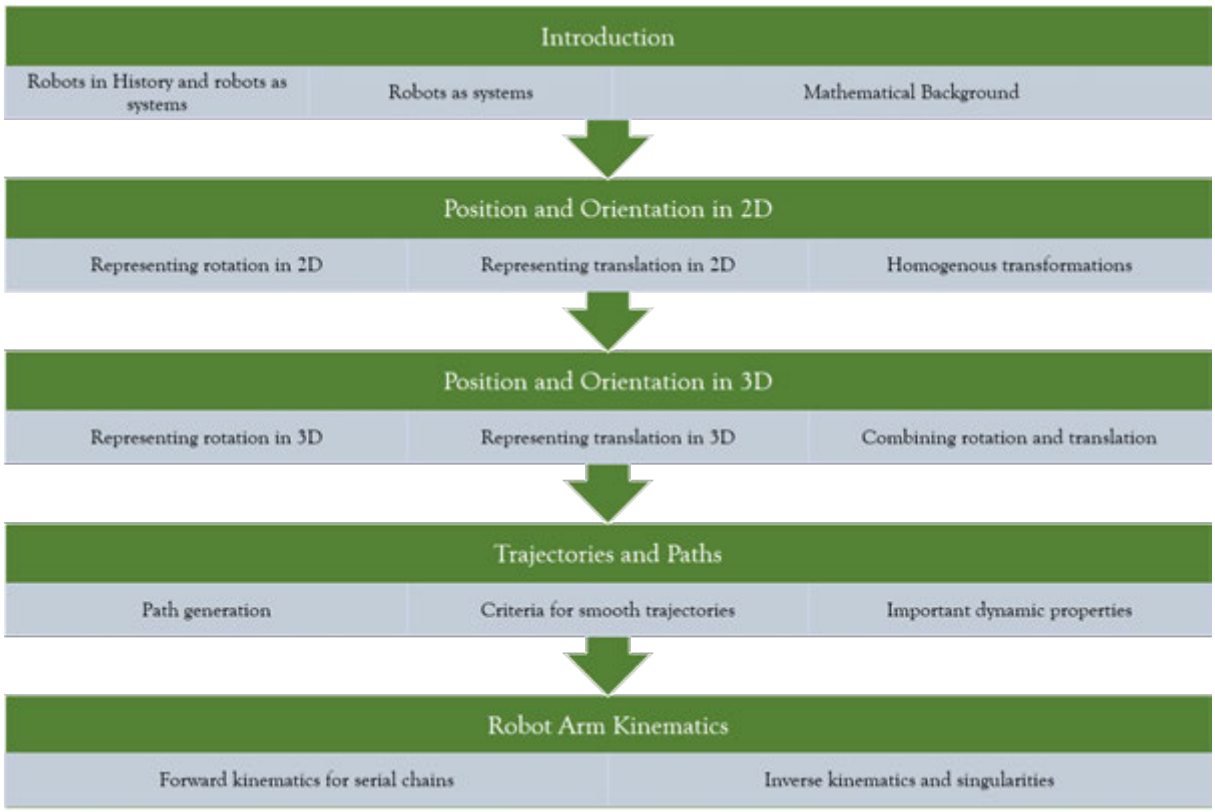


Fig. 2: Structure of the robotics lecture.

the lectures and topical coverage are shown in Fig. 2. As a start, general concepts of linear algebra are taught in order to generate a common understanding of terms as prerequisite. As a second block, the development of robots and robotic systems through history is illustrated and the main conceptual developments are pointed out. Afterwards, descriptions for position and orientation in two dimensional as well as three dimensional spaces are content of the lectures. Later, forward kinematics as means to describe manipulator geometries and the connection between end-effector position and joint angles are discussed. These concepts are shown in the lecture and later on exercised in the lab. Especially concepts of frames, homogenous transformations, base and tool transformations and the usage of different reference coordinate frames are important in this context. Additionally, inverse kinematics are introduced as a way to determine necessary joint angles in order to drive the robot manipulator end-effector to a certain position. In this context, we also discuss the problem of singularities and how these evolve dependent on different representations of rotations.

B. Lab Exercises

Previous research in engineering and robotics education has proven lab exercises as a valid means to improve learning outcomes [9]–[12] and enhance motivation of students [13]. Additionally, it has been argumented in [14] that group work forsters student collaboration. Consequently, lab exercises are done in small group with lecturer support. Three exercises are given and have to be fulfilled and documented accordingly. The first part is an introduction to RobotStudio, the software used for programming of the ABB robot arms. The first lab exercise is commonly used to familiarize the students with the handheld device for manual robot teaching, i.e. the control panel. Additionally, the different coordinate frames as used in the software are introduced and students can try moving the robot arm in each of these frames. Then, the students have to program the robot arm so that it would follow a labyrinthic path printed on an A3-sized paper. The respective path has to be programmed using the software and then stored to the robot controller. Additional exercises include using pneumatic and electric robot end-effectors which are available in a tool magazine.

III. Robotic Hardware

The robot arms used for the lectures and lab exercises are the ABB IRB120 (which is further also used for the presented project work) and the ABB IRB1200. The smaller robot weighs $w = 25\,\mathrm{kg}$ and has a payload of $l = 3\,\mathrm{kg}$, the larger one weighs $w = 52\,\mathrm{kg}$ and has a payload of $l = 7\,\mathrm{kg}$, both provide voltage and pneumatic connectors for gripper mounting. In Fig. 3 the tool changer mounted on the robot arm is shown. This is basically a connection device which allows to connect different end-effector tools and consequently also different grippers. It provides electrical and



Fig. 3: Robot arm with tool changer and pneumatic connectors.

pneumatic connection to the connected grippers. Additionally, it is possible to read sensor signals from the gripper through the robot controller and control software. This is important if students are assigned tasks where the gripper feedback is used to steer the end-effector movement. This is often essential in manipulation and learning tasks which are more and more state-of-the-art in novel robotic systems which allow for human-robot-collaboration and assistance scenarios.

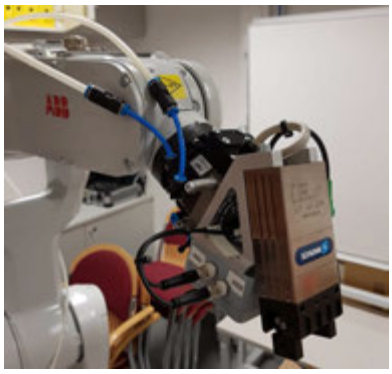


Fig. 4: Robot arm with mounted gripper.

IV. Task Description and System Analysis

At the industry partner, gearing parts are assembled manually by gluing 215 carbide blades into a steel disk (see Fig.6). To setup a suitable automation concept, this process is first broken down into distinct tasks which can be addressed step-by-step. First, the blades have to be picked up at a magazine using a 6 DOF robot arm and moved below a dispensing station to apply the proper amount of glue. Then the blades are inserted into cavities which are distributed equally spaced on the circumference of the steel disk. After inserting one such blade, the steel disk has to be rotated by a certain amount of radians in order to expose the subsequent cavity. The following distinct tasks have been identified:

- Introduction of carbide blades
- Application of glue using the dispensing system
- Taskcontrol
- Rotation of the steel disk during the assembly

Due to the high complexity of the overall process, it is necessary to distribute the tasks to various components.

V. Assembly System

For the introduction of the carbide blades, the ABB 6-Degrees-of-Freedom (6 DOF) robot manipulator arm IRB120-T is used. Programming of this arm is done in a proprietary software called RobotStudio. For the target task, two programming options were used: offline programming using RAPID (the proprietary programming language) as well as online programming using a teach-in method. When using the teach-in approach, it is necessary to move the robot end-effector to the desired position and store the coordinates to the robot controller. This method has been used to store the coordinates of the important via points for the pick up of the carbide blades, the position below the dispenser unit and the position where the blade is brought into the cavity of the steel disk. The control of the path taken between these points of interest is then controlled using offline programming. Additionally, necessary digital signals for the communication with the employed Programmable Logic Controller (PLC) as well as signal for opening and closing of the gripper are implemented. In order to assure safe operation of the assembly stand, retraction strategies for the robot arm have been developed and implemented. For the final adjustment of the carbide blades, two pneumatic cylinders are employed which move the blade into place after the robot arm drops it into the cavity. Then, the steel disk is turned by $\alpha_{rot} = 1.674°$ in order to expose the next cavity.

A. Dispensing System

The used glue is Loctite 620. This product has high viscosity which makes the use of a dedicated dispensing setup necessary. This system is mounted at a specified position at the front of the assembly stand (see also Fig.1), which can easily be reached by the robot arm. The dispensing system (Performus III) itself is available from Nordson. It is actuated pneumatically and controlled through the PLC. The application of the glue (see Fig. 5) can thus be achieved with controlled pressure and time.



Fig. 5: Illustration of dispensing glue on a carbide blade.

B. Main Control Unit and Profinet Communication

The control and timing of the whole assembly system was realized through a Siemens S7-1214C PLC. To program the PLC controller, the proprietary software TIA Portal was used. The main tasks of the PLC were starting, stopping and timing control as well as control of the servo motor movement. All components are connected using Profinet. One of two available Profinet setups is used here: Profinet IO. This is a standardized bus topology with one control unit and multiple decentralized field devices. It works according to a so-called 'producer-consumer-principle' where the data on the bus can be read by all connected devices. In order to be able to properly interpret the data, it is necessary for the devices to have the respective device-description-files (GSD-files). The devices are connected using Industrial Ethernet.

C. Disk and Servo Drive

The assembly stand is developed to be working is different types of steel disks (Fig.6). The difference is in the number of steel blades which have to be assembled (215, 170 or 150) as well as in the depth of the cavities. However, the diameter of the disks is constant. The weight of the disks is $w_e = 23\,\mathrm{kg}$ without carbide blades and $w_f = 32\,\mathrm{kg}$ fully assembled. In order to properly rotate the steel disk on the assembly stand, a hybrid stepper motor ES-M342120 by Leadshine is used in combination with a dividing head. In general, the stepper motor position is controlled by a servo drive module and the respective encoder. The motor offers up to $n_s = 300$ steps per revolution with a maximum stalling torque of $t_s = 12\,\mathrm{Nm}$. Additionally, it can be driven in microstepping mode which allows an angular resolution of $res_{ang} = 0.005\,^\circ$. However, the stepper motor torque would be significantly reduced in such a configuration. The dividing head can now be used to realize a gear reduction of $g_{red} = 120/1$. The dividing head is usually operated manually, it is constructed for high loads and is self-impeding, so that no additional braking system is necessary.

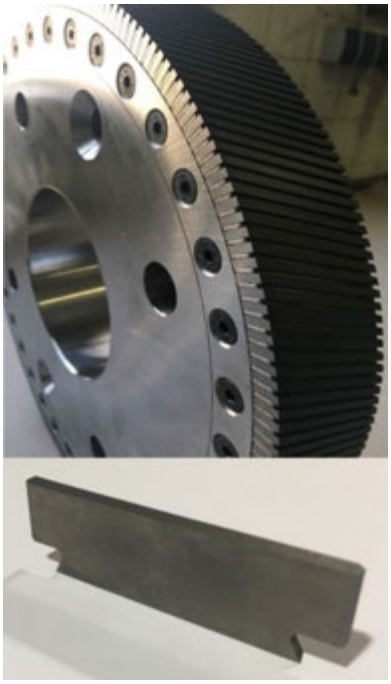


Fig. 6: Top: illustration of the steel disk, bottom: carbide blade glues into the cavities of the disk.

D. Construction Elements

In order to realize a proper assembly stand suitable for industry applications a variety of constructive elements are necessary:

- First of all, it is important to provide a low vibration table. This table was constructed from aluminium ITEM-profiles.
- For the connection of the dividing head and the stepper motor, a dedicated flange was constructed and milled.
- For the provision of the carbide blades a magazine is constructed in AutoCAD Inventor and milled from plastics.
- An operator terminal has been built and attached to realize start/stop buttons and various status lights.
- For the proper pick-up, moving and drop-down of the carbide blades, a specific gripper construction was necessary. The parts of the gripper have been designed, optimized and 3D-printed.

E. Software and Programming

The programming is realized in different software tools: the robot is programmed in RobotStudio and the PLC is programmed using TIA Portal. In TIA Portal, the main process control is realized and also the stepper drive control is implemented. To do so, various standardized programming languages are available (e.g., KOP, FUP, AWL). In this environment, different building blocks can be defined where each building block would control one element of the automation setup. The following building blocks are implemented:

- Main block: main routine for the task
- Motor block: motor control, homing
- Error block: for error handling
- SCL block: for control of status lights at the operator terminal

VI. Experimental Results and Conclusion

The PLC programming has been validated experimentally by driving the motor and dividing head as for a real assembly process. The result is illustrated in Fig.7. In the upper part of the graph, the motor velocity is illustrated in $\mathrm{mm/s}$ and in the lower graph, we see the position in mm. This graph is taken directly from the visualization at the PLC. Each of the peaks in the upper graph represents the placement of one of the carbide blades into the disk.

Various experiments are done using the PLC an respective programming in order to show limitations of the system in the lab. To initialize the system, a button is pressed which starts the initialization procedure. The curve which is recorded

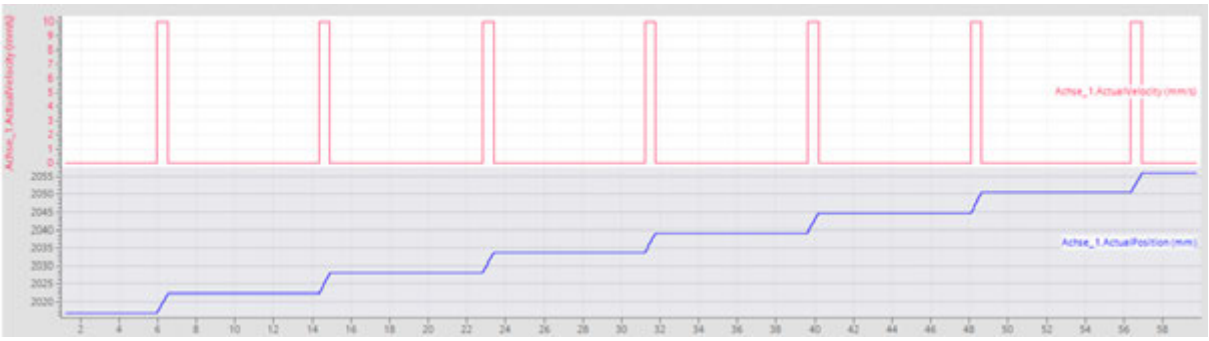


Fig. 7: Experimental results of stepper motor movement and position.

using the PLC visualization, however, indicates infinite motor velocity at the start of the process, this is an indication that the PLC input suffers from latency and is a fact which is further important for the students to be considered in latter industrial projects.

The concept of robotics lectures and lab exercises in our Bachelor program has been shown and the available hardware for the realization of industry related Bachelor projects has been presented. For the realization of the considered assembly task, the ABB IRB-120 robot arm has been employed together with a Siemens PLC system. Both components have been connected using the Profinet bus. To enable the rotation of the steel disk with sufficient accuracy, a servo drive has been employed in combination with an indexing head and an inductive sensor. Major challenges in this project were the robotic gripper which was built to accurately pick of the blades. Here, sufficient gripping capability and grip stability were considered in order to fabricate suitable components. As a result, a functional prototype of this automatic assembly system has been built and the feasibility has been demonstrated.

For the students, this implementation of an industrial concept showed important capabilities, as well as limitations, of the employed components. E.g., an inductive sensor was used to detect the position to which the steel disc should be rotated. During their work with the system, it was found that such a positioning approach could be improved by employing, e.g., a camera system.

Through the project work done in this context, the students get hands-on experience in the connection of industrial robotics and PLC systems. They gather experience in software programming as well as the design and fabrication of hardware components and go through a complete integration process covering all aspects of project work. Additionally, in the context of the teaching programs, the students learn how to setup and work with project plans and how to manage resources in order to successfully complete projects in larger teams.

Regarding the lecture content, they apply methods for forward and inverse kinematics and learn how to use those in a real-world scenario.

REFERENCES

[1] Y. Lu, "Industry 4.0: A survey on technologies, applications and open research issues," *J. of Industrial Information Integration*, no. 6, pp. 1–10, 2017.

[2] L.-M. Faller, T. Mitterer, S. Weiss, and H. Zangl, "3d- and inkjet-printed force sensors for tactile sensing," in *RoboTac: Workshop on Tactile Sensing and Perception, IEEE Int. Conf. on Intelligent Robots and Systems (IROS)*, 2018.

[3] Z. Xu, S. Kolev, and E. Todorov, "Design, Optimization, Calibration, and a Case Study of a 3D-Printed, Low-cost Fingertip Sensor for Robotic Manipulation," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2014.

[4] D. Hughes, J. Lammie, and N. Correll, "A Robotic Skin for Collision Avoidance and Affective Touch Recognition," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1386–1393, 2018.

[5] L.-M. Faller, C. Stetco, and H. Zangl, "Design of a Novel Gripper System with 3D and Inkjet-Printed Multimodal Sensors for Automated Grasping of a Forestry Robot," in *Proc. of the IEEE IROS 2019*, 2019.

[6] M. Geiger and C. Waldschmidt, "160-GHz Radar Proximity Sensor With Distributed and Flexible Antennas for Collaborative Robots," *IEEE Access*, vol. 7, pp. 14 977 – 14 984, 2019.

[7] L.-M. Faller, M. Krivec, A. Abram, and H. Zangl, "AM Metal Substrates for Inkjet-Printing of Smart Devices," *Meterials Characterization*, 2018.

[8] P. I. Corke, "Matlab toolboxes: robotics and vision for students and teachers," *IEEE Robotics and Automation Magazine*, vol. 14, pp. 16–17, 2017.

[9] J. M. Esposito, "The state of robotics education: Proposed goals for positively transforming robotics education at postsecondary institutions," *IEEE Robotics and Automation Magazine*, vol. 24, pp. 157–164, 2017.

[10] L.-M. Faller and H. Zangl, "An Electromagnetic Force Balance Experiment for Measurement Science Education," in *IMEKO Joint Symposium TC1- TC7- T13*, 2017.

[11] L.-M. Faller, H. Zangl, and J. P. Leitzke, "Fault Injection Software Tools and Robust Design Principles for Reliability and Safety in Measurement Science Education," in *IMEKO Joint Symposium TC 1 - TC 7 - TC 13*, 2016.

[12] F. Zamora, "Work in Progress: Novel Didactic Training Platform for Transistor Devices and Applications," in *Proc. of the IEEE Global Engineering Education Conference (EDUCON)*, 2019.

[13] A. Khanlari, "Robotics integration to create an authentic learning environment in engineering education," in *Proc. of the IEEE Frontiers in Education Conf.*, 2016.

[14] K. F. Frykedal and E. H. Chririac, "Re-engineering the curriculum for the 21st century," *Internat. J. of Disability, Development and Education*, vol. 65, pp. 183–198, 2017.

SJSU GO

Rod Fatoohi, Sheryl Ehrman, Lisa Francesca, Chanip Chong, Joseph Khoury, Anthony Minaise, Dan Pham, Manas Abhyankar, Jia Chen, Raymond Chen, Antonio Corona-Sanchez

Charles W. Davidson College of Engineering, San Jose State University, San Jose, California 95192, USA

***Abstract*—The Davidson College of Engineering at San Jose State University (SJSU), San Jose, California, offers a rewards program for students to get points for attending professional and academic events called GO. Currently students submit events for the program through a Google form with many limitations. In this project we have developed mobile applications (for both iOS and Android systems) and a website to submit events, to view points and prizes, and to navigate the system. The goal is to increase student participation in these events and ease of use of the system.**

Keywords—mobile applications, software development, professional events.

I. INTRODUCTION

The Davidson College of Engineering (CoE) at San Jose State University (SJSU) recently initiated a program called GO [1]. It promotes networking, career-building, building skills through industry contacts, and hosting events at the school. Through these events, students learn important skills for their future career and meet with industry leaders. Students are encouraged to attend these events and can submit proof of their attendeance to earn points, which can be converted to prizes awarded by the college. However, the program seems to lack participation partly of the current submission process, which is based on a Google form, and sometimes students are unware of these events. Also, verifying and managing a large body of students' submissions in such a system is quite a challenge. Currently, there are over 200 students using the GO program in a college with over 6000 students.

Just as technology is becoming more integrated into our daily lives, researchers are finding ways to use technology to help promote student involvement and learning [2]. Byron and Khazanchi [3] show that rewards help stimulate student creativity and motivation in school. They analyzed whether and how rewards affect students to be more creative. Their design was to test the extent to which rewards are given to see students' creative outcomes. When students are satisfied with their work, they are more likely to be motivated by rewards [4] and more likely to stay active in the reward program [5]. There is also a link between rewards in the workplace and performance management. Lawler [6] shows that employees who were given rewards after a positive review report tended to perform better in the workplace.

The development of the GO Program mobile app in iOS and Android as well as the overhaul of the desktop website is primarily focused on increasing the participation of the college students in campus events to better prepare them for their future careers. Based on the work by Wong [7], a desktop application will be necessary as not all users will be inclined to use the mobile app. For mobile applications, both Android and iOS implementations might be needed since it is hard to determine which mobile OS the user will use. As found by Götz [8], there are negligible personality trait differences between iOS and Android users, making it hard to draw any conclusions about which OS the user's phone will operate on.

II. SYSTEM DESIGN

Fig. 1 shows a top-level view of the design architecture. A similar architecture is used for all three applications: Android, iOS and the web. For the current implementation, we used Firebase, a Google developed backend database solution that provides cross-platform support for the three applications. Users, as students, and admins, as program administrators, interact with the frontend to request and submit data. The backend serves data to the frontend based on client requests and stores client data. Fig. 2 shows the database entry diagram of the application.

Fig. 3 shows a flow diagram of the overall application. A client accesses the GO program by navigating to the main page and either logging in or signing up if the client does not already have an account. After successful client login, the client is directed to their dashboard where aggregate information such as total points accumulated, recent event activity, and available prizes is shown.

Fig. 4 shows the system functional requirements. A user, a qualified student, should be able to login or sign up for an account. Also, a user should be to choose an event type, from a drop-down menu, or insert an event being attended if it is not listed in the menu. In addition, a user should be able to upload a document of proof or a photo of an attended event. Moreover, a user should be able to view their accumulated points and claim a prize. On the other hand, an administrator should be able to add and revise possible events in the list, to review the proof that a user has uploaded and assign corresponding points, and to pull statistical data in different formats (such as percentage of departments that actively engaged, highest scores of individuals, number of participated students, eligibility, and student feedback).

III. IMPLEMENTATION

For the iOS implementation, we designed and created the Minimum Viable Product (MVP) as an iOS app. We also used the programming language Swift and the Integrated Development Environment (IDE) Xcode as well as Firebase (as it was used in all our implementations). We used both

the pre-installed simulators available in Xcode and physical iPhones to run the app for testing. To keep our codebase organized, we used the Model-View-Controller (MVC) as a software design pattern. This meant grouping files based on the category they fit in: Model files would consist of files that handled data, View files would consist of files that contained anything the user would see, and Controller files would consist of files that controlled the data that would be passed between the Model and the View. In addition, we used Object-Oriented Programming to organize our code and to create objects, such as a User object and an Event object, in order to follow good coding design.

In order to build and design the Android implementation and website, we used Android Studio, Github, Firebase, and Android phones. Android Studio is an IDE that supports Android operating system and provides an emulator, code templates, and GitHub integration. GitHub is a code hosting platform for collaboration and version control between developers to allow them to share files, and make changes to code without affecting the work of others. We built the frontend with Android Studio and using Kotlin, a native development language for Android, for all EventListeners, onClick actions, and interface communications.

For the web design, we used the Visual Studio Code to develop the webpage. We also used HTML and CSS, a language that describes the style of an HTML document, for the frontend of the website. In addition, we used Javascript to communicate with the backend.

IV. TESTING & RESULTS

For the iOS app, we used both white-box and black-box testing (internal structures of application known to tester in white-box vs unknown in black-box). We successfully tested many functionalities using iPhone Xs Max and running iOS 12.4 including sign up, login, review prizes, and claim a prize. Figs. 5 and 6 show UI mockups of the iOS app.

For the Android app, Android Studio also provides two methods of testing an app natively: 1) using the built-in Android Virtual Device (AVD) to emulate an Android phone, and 2) packaging and installing an APK onto your own Android device.

For the web application, we conducted functionality testing, usability testing, and compatibility testing. We also focused on specific unit tests, regression testing, and integration testing in order to validate the entire application. For example, in regression testing, we tested many functionalities including sign up, login, event submission, claim prize, and log out.

V. CONCLUSION & FUTURE WORK

In this project, we overhauled the GO program website and submission form by creating two mobile applications (for iOS and Android) and a new website that would enable more students to benefit from the program. This is only the first phase of the project where we were able to develop, implement and test prototypes of the three applications. More debugging and testing are needed as well as adding more features.

Next steps for this project are to properly deploy these applications so students can use them. Currently, we are working to deploy the iOS app in Apple App store and use the Apple Developer TestFlight for beta testing. Additionally, we plan to integrate these applications with the San Jose State University database to properly assess student emails for verification.

REFERENCES

[1] GO Program, URL: https://www.sjsu.edu/engineering/student-success/ease/go/index.php

[2] P. Treuer and L. Belote, "Current and Emerging Applications of Technology to Promote Student Involvement and Learning," New Directions for Student Services, (78), 17-30, 1997.

[3] K. Byron and S. Khazanchi, "Rewards and creative performance: A meta-analytic test of theoretically derived hypotheses," Psychological Bulletin, 138 (4), 809–830, 2012.

[4] R. Q. Danish and A. Usman, "Impact of Reward and Recognition on Job Satisfaction and Motivation: An Empirical Study from Pakistan", International Journal of Business and Management, 5(2), 159-167, 2010.

[5] H. T. Keh and Y. H. Lee, "Do reward programs build loyalty for services? The moderating effect of satisfaction on type and timing of rewards.," Journal of Retailing, 82 (2), 127-136, 2006.

[6] E. J. Lawler. "Reward practices and performance management system effectiveness," *Organizational Dynamics* 32 (4): 396-404, 2003

[7] S. H. R. Wong, Shun, "Which platform do our user prefer: website or mobile app?" Emerable Group Publishing Limited, 40 (1), pp. 103-115, 2012.

[8] F. M. Götz, S. Stieger, and U-D. Reips U-D, "Users of the main smartphone operating systems (iOS, Android) differ only little in personality," PLoS ONE 12(5): e0176921, 2017

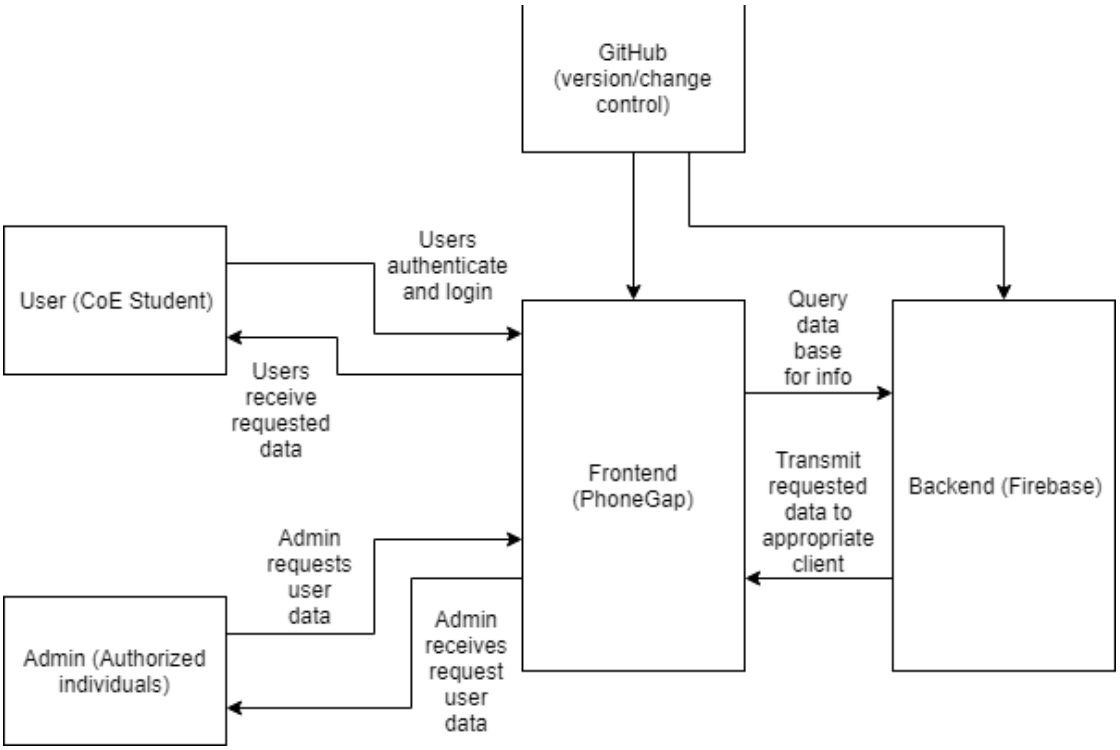


Figure 1. Design architecture

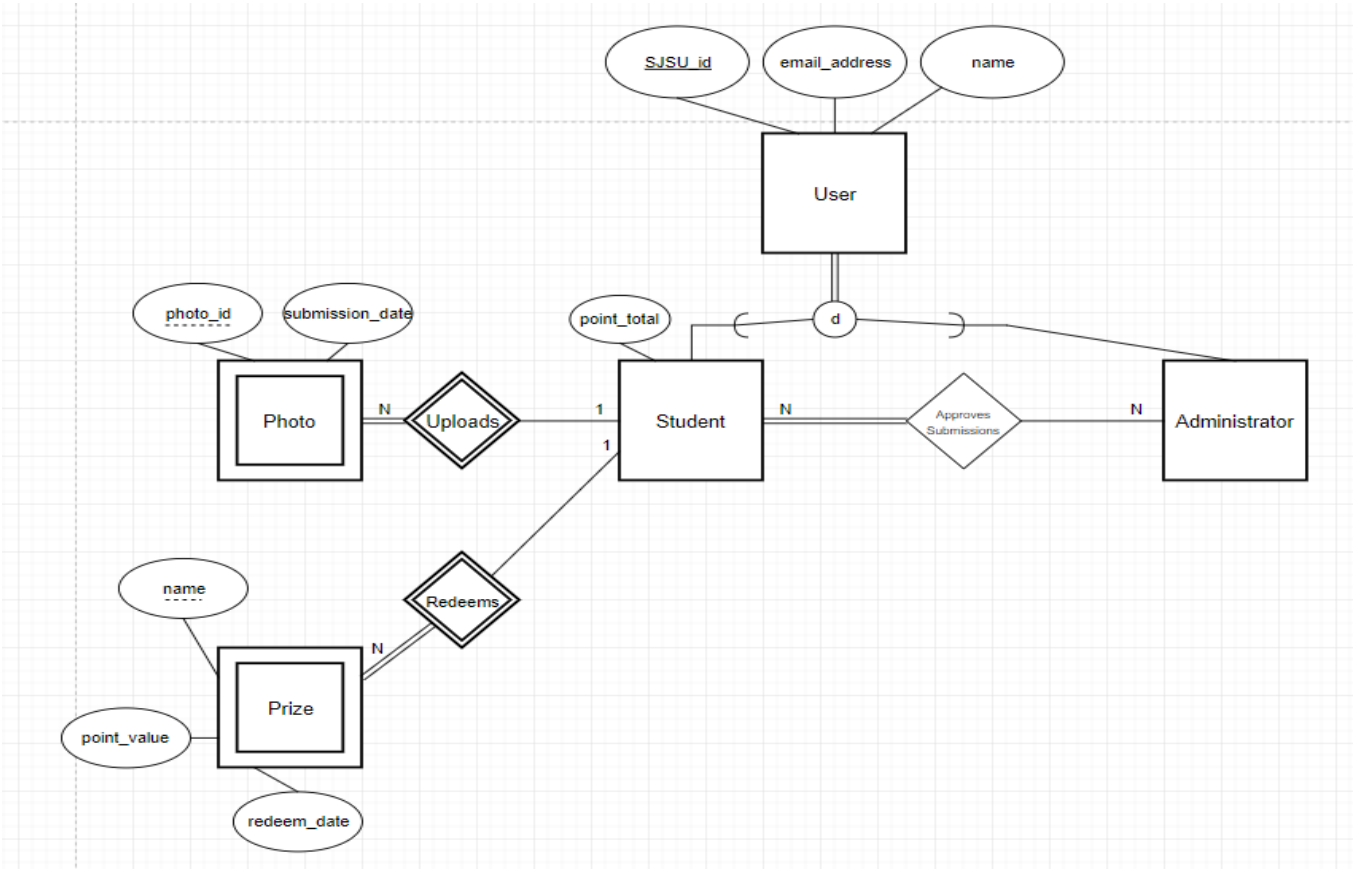


Figure 2. Database entry diagram of the application

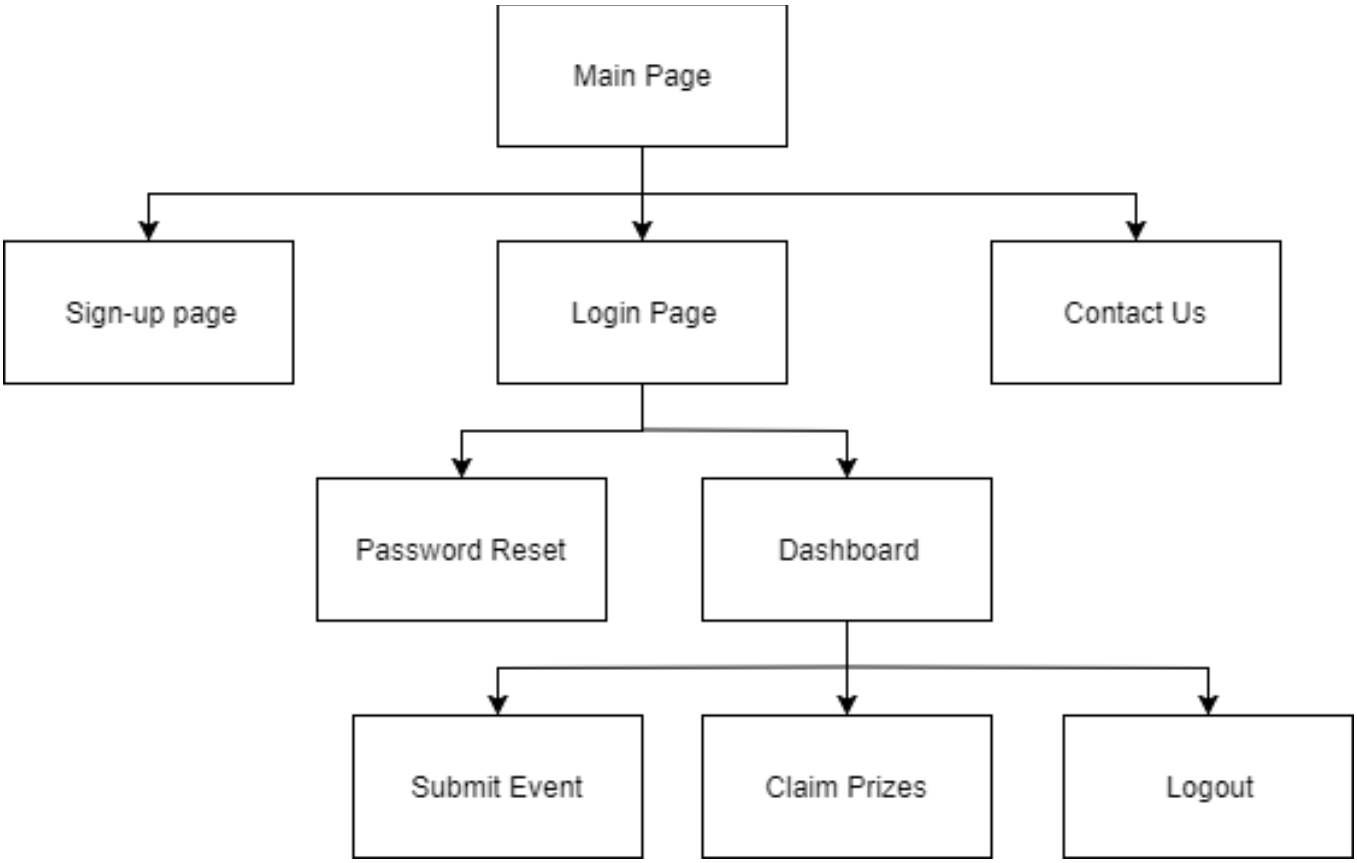


Figure 3. Flow diagram of the application

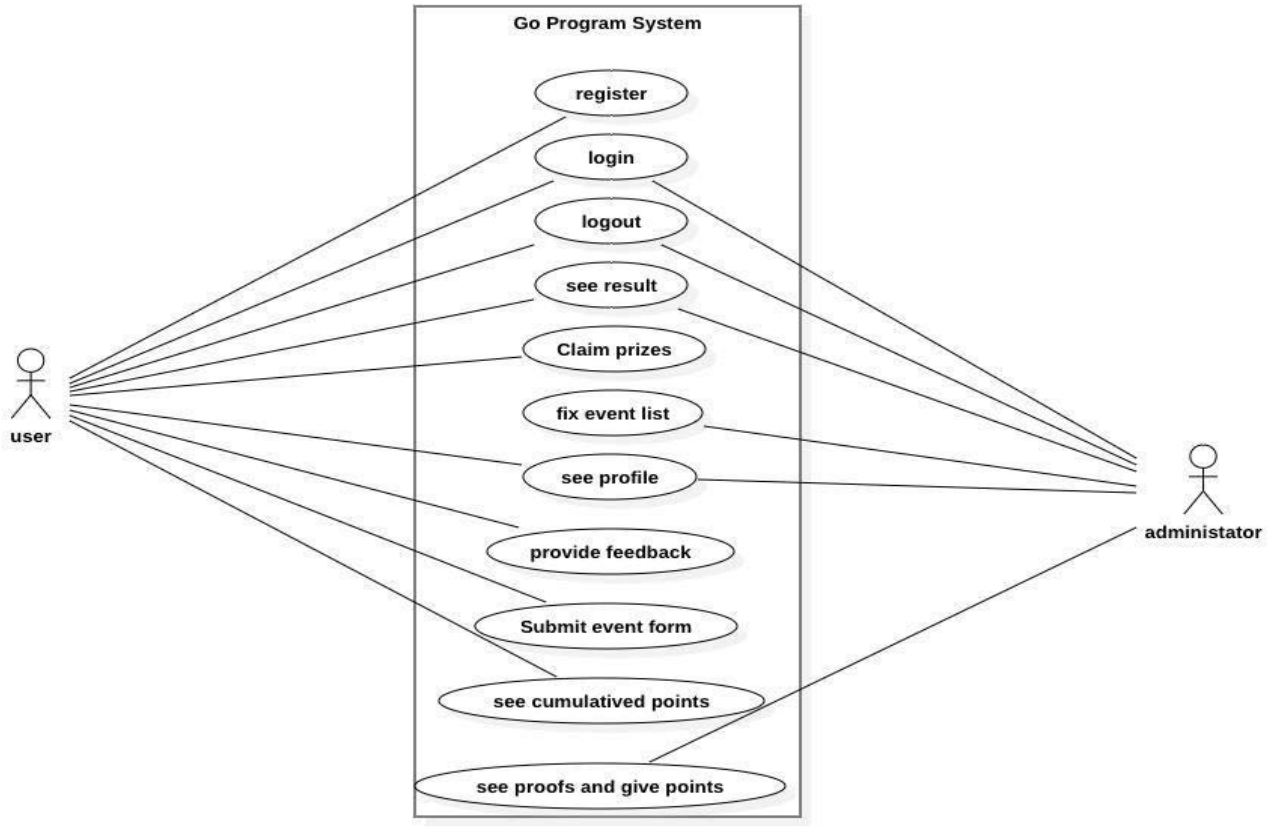


Figure 4. User case diagram of requirements.



Figures 5 and 6. UI mockup of iOS app login and dashboard screens

Online and in-class computer science teacher training - Oracle Academy program experiences

F. Urem*, D. Jureković ** and E. Ban ***
* Polytechnic of Šibenik, Šibenik, Croatia, frane.urem@vus.hr
** Oracle Hrvatska d.o.o., Zagreb, Croatia, darko.jurekovic@oracle.com
*** Education and Teacher Training Agency, Rijeka, Croatia, eban@azoo.hr

***Abstract* - Computer science teachers constantly need to improve their competencies as information technology evolves rapidly. Today, most such trainings are holding online because of the obvious advantages over in-classroom training such as lower costs and a comfortable environment. Therefore, it is important to investigate the results achieved by computer science teachers in online training with in-class educations when the training has the same learning outcomes and maintained by the same instructor. As part of this research, computer science teacher's training conducted within the Oracle Academy program in the Central and Eastern Europe region were analyzed as they are held online and in-class, with the same teaching materials, with equal knowledge tests and maintained by the same instructors. The analysis found that teachers enrolled in traditional education received similar grades as those who participated in online training. However, the dropout rate for midterm and final exams was higher with online training. Teachers' feedback on the advantages and disadvantages of both types of training has recorded, with the major conclusion that they prefer online learning for their training.**

Key words – online; in-class; training; Oracle Academy

I. INTRODUCTION

Teaching in the field of information and communication technologies (ICT) is becoming one of the most challenging professions today due to the accelerated increase of knowledge in the field of information science and computer science teachers have directed to continuous professional development and improvement of their knowledge. To this end, teachers are trained in organized learning in workshops or seminars (i.e. in-class learning), as well as in various forms of online education and learning. For example, the experience of online training of computer science teachers by the Singapore Ministry of Education in the late 1990s is one of the first global examples of online professional development [1].

According to a source from [2], in-class learning has defined as "*training conducted by a trainer, supervisor or another suitably qualified person in any training center, conference room, crib room or office"*. In-class learning today often includes blended (or hybrid) learning as an experience that combines classic lectures with the use of e-learning systems on which teaching materials and tests can be found.

Different definitions for online learning can be found today, but in the context of this paper, a definition from the Oracle Academy (OA) program may be used that says [3] "*virtual, instructor-led, collaborative, web-based training*". In this type of learning, students watch a virtual lecture, in a real-time with a real lecturer, through online meetings platforms while working through their digital lessons and assessments. Due to several advantages (resource-efficient, flexible, more accessible, self-paced learning), online learning is becoming the most significant form of professional development today.

Due to the recent COVID-19 pandemic, all European education systems have introduced substitute online learning in mainstream education. This process has proven to be very successful with the only commonly known weakness that students can more easily cheat on an exam where they use digital technologies. The latter will present a challenge in the future of education because of the unstoppable process of digitalization of teaching and the increasing use of online learning.

This paper explores the experience of OA programs during the implementation of activities aimed at improving the competences of computer science teachers. Particular emphasis has placed on exploring teachers' attitudes to online and in-class learning based on their previous experience in the OA training courses.

II. ORACLE ACADEMY TEACHER TRAINING

The OA program has founded in 2013 by Oracle Corporation and currently brings together more than 16,000 educational institutions worldwide [4] to improve teacher skills, but also connecting teachers to share best practices in computer science teaching. In Central and Eastern Europe, the OA program has been operating since 2015, wherein the meantime, dozens of different online and in-class training organized for teacher education. Teachers can develop their skills through specially designed training in the fields of computer programming (Java Foundations and Java Programming) and databases (Database Foundations, Database Design, and Programming with SQL) [4]. Training is also organized to prepare teachers to teach kids and beginners in computer programming (e.g., Java Fundamentals, Getting Started with Java Using Alice, Creating Java Programs with Greenfoot) [5]. In addition, the OA program offers teachers a range of self-learning training (e.g. Artificial Intelligence with Machine Learning in Java, Programming

with PL / SQL, APEX - Application Development Foundations).

OA instructors (teachers recruited from institutions participating in the OA program) conduct online and in-class training using specially developed digital curricula. Each OA course is delivered over several meeting dates (referred to as "*sessions*") and each session is taught by in-class or online instructors using the online meeting platform (Cisco Webex [6], Zoom [7]). Upon the close of the course, and no more than 13 days after, the instructor-trainer is required to provide each participant with completed status and final grade. A course has considered to have completed if teacher attends all sessions, meet the learning objectives of each lesson within the course, review lesson materials and complete homework before class, participate and contribute actively in each lesson. If the course is completed, then the participant can take the midterm and final exam on a separate e-learning system (minimum 70%, for grade- Passed).

III. STRUCTURE OF RESPONDENTS WITH INDICATORS OF PREVIOUS PROFESSIONAL DEVELOPMENT

This paper presents the results of a survey of online learning attitudes of a group of computer science teachers from Central and Eastern European countries with similar educational systems (Croatia, Bosnia and Herzegovina, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Slovakia, and Slovenia) who have previously participated in OA educations. Approximately 30% of teachers (76 out of 256) responded to the submitted questionnaire. Initially, data have collected on the type of institution in which they work their gender, age, the number, and type of professional development they have attended over the last two years.

When considering the type of educational institution in which teachers work, it is evident that the majority of surveyed teachers work in secondary schools (shown in Table I).

The predominant age of the surveyed teachers is from 40 to 50 years (shown in Table II).

The gender distribution (shown in Table III) is even with minor discrepancies in elementary education (female predominance) and higher education (male predominance).

According to the data collected, the teachers have mostly trained in online training (shown in Table IV). Quantitative indicators of teacher professional development for the last 2 years have shown in Tables V and VI. Most teachers have participated in at least three in-class training, or at least four online training in the last two years, which is satisfactory.

In addition to the mentioned questionnaire, internal data on OA courses in which the surveyed teachers participated have additionally collected and processed. Using such data, the drop-off rates and the average grade of the passed exams in both types of training courses have compared (shown in Table VII). Drop-off rate is the ratio of teachers who did not complete the course (failure to take the exam or drop out of teaching) and the total number of students. The average grade includes the grades of all participants who have successfully passed the midterm and final exams in a certain group of courses.

TABLE I DISTRIBUTION BY TYPE OF EDUCATIONAL INSTITUTION

Type of educational institution	Total share
Elementary education	25,00%
Secondary education	62,50%
Higher education	12,50%

TABLE II AGE DISTRIBUTION

Type of educational institution	< 40	from 41 to 50	> 50
Elementary education	23,81%	47,62%	28,57%
Secondary education	32,00%	40,00%	28,00%
Higher education	40,00%	50,00%	10,00%
Total share	**30,86%**	**43,21%**	**25,93%**

TABLE III GENDER DISTRIBUTION

Type of educational institution	Male	Female
Elementary education	23,81%	76,19%
Secondary education	50,00%	50,00%
Higher education	70,00%	30,00%
Total share	**45,68%**	**54,32%**

TABLE IV DISTRIBUTION OF TEACHER TRAINING IN THE LAST TWO YEARS

Type of educational institution	Online	In-Class
Elementary education	66,74%	33,26%
Secondary education	61,78%	38,22%
Higher education	62,28%	37,72%
Total share	**63,35%**	**36,65%**

TABLE V NUMBER OF ONLINE TRAINING COURSES TAKEN IN THE LAST TWO YEARS

Type of educational institution	< 4	from 4 to 10	> 10
Elementary education	42,86%	23,81%	33,33%
Secondary education	26,00%	40,00%	34,00%
Higher education	40,00%	40,00%	20,00%
Total share	**24,69%**	**43,21%**	**32,10%**

TABLE VI NUMBER OF IN-CLASS TRAINING COURSES TAKEN IN THE LAST TWO YEARS

Type of educational institution	< 3	from 3 to 10	> 10
Elementary education	23,81%	66,67%	9,52%
Secondary education	24,00%	58,00%	18,00%
Higher education	40,00%	40,00%	20,00%
Total share	**33,33%**	**50,62%**	**16,05%**

TABLE VII DROP-OFF RATES AND AVERAGE EXAM GRADES AT ANALYZED OA TRAINING

Delivery type	Drop off (%)	Average grade
Online only	40,54%	86,80%
In-class	28,28%	82,62%

IV. TEACHERS' ATTITUDE TOWARDS OA ONLINE AND IN-CLASS TRAINING

The teacher survey further explored possible reasons for a slightly higher dropout rate on OA online courses. Two presumed reasons have investigated reduced motivation and the need to spend more time and energy on learning. According to the survey results, teachers did not have a problem with motivation to learn in online training, regardless of the type of education system they come from (shown in Table VIII). In addition, teachers expressed the view that they did not have to invest more energy and time in learning in online OA training compared to OA in-class training (shown in Table IX).

As part of the teacher survey, it was checked whether teachers were more inclined to in-class teaching before the experience of OA online courses, i.e. whether their attitude towards online learning subsequently changed. According to the results from Table X, it may have said that teachers initially did not prefer OA in-class teaching. After the experience of OA online courses, the affirmative attitude towards OA online learning is more pronounced, and the change in attitude has especially pronounced in teachers coming from primary schools (shown in Table XI).

Finally, the questionnaire examined the impact of the experience of teacher participation in OA online courses on the holding of online classes for their students. It has found that the experience of participating in online OA training helped them to hold online classes for their students, and especially such an attitude has expressed for primary school teachers (shown in Table XII).

TABLE VIII LEARNING MOTIVATION IN TEACHERS WHO PARTICIPATED IN ONLINE OA TRAINING

	Question/Answers: I had a problem with learning motivation while attending online Oracle Academy training.				
Type of educational institution	**Disagree**	**Rather disagree**	**Neutral**	**Rather agree**	**Fully agree**
Elementary education	28,57%	38,10%	23,81%	4,76%	4,76%
Secondary education	26,00%	40,00%	28,00%	4,00%	2,00%
Higher education	30,00%	60,00%	0,00%	0,00%	10,00%
Total share	**27,16%**	**41,98%**	**23,46%**	**3,70%**	**3,70%**

TABLE IX WORKLOAD IN TEACHERS WHO PARTICIPATED IN OA TRAINING

	Question/Answers: I had to spend more energy and time learning in Oracle Academy online training than in-class training				
Type of educational institution	**Disagree**	**Rather disagree**	**Neutral**	**Rather agree**	**Fully agree**
Elementary education	33,33%	47,62%	9,52%	4,76%	4,76%
Secondary education	30,00%	56,00%	12,00%	2,00%	0,00%
Higher education	20,00%	60,00%	10,00%	10,00%	0,00%
Total share	**29,63%**	**54,32%**	**11,11%**	**3,70%**	**1,23%**

TABLE X TEACHERS 'ATTITUDE ABOUT ONLINE LEARNING, BEFORE PARTICIPATING IN OA TRAINING

	Question/Answers: In case you were able to choose before taking the course, you would rather choose to attend Oracle Academy in-class training.				
Type of educational institution	**Disagree**	**Rather disagree**	**Neutral**	**Rather agree**	**Fully agree**
Elementary education	23,81%	14,29%	23,81%	23,81%	14,29%
Secondary education	20,00%	38,00%	26,00%	10,00%	6,00%
Higher education	20,00%	30,00%	30,00%	10,00%	10,00%
Total share	**20,99%**	**30,86%**	**25,93%**	**13,58%**	**8,64%**

TABLE XI TEACHERS 'ATTITUDE ABOUT ONLINE LEARNING, AFTER PARTICIPATING IN OA TRAINING

	Question/Answers: After the experience of participating in online Oracle Academy training, you would rather choose to attend Oracle Academy in-class training.				
Type of educational institution	**Disagree**	**Rather disagree**	**Neutral**	**Rather agree**	**Fully agree**
Elementary education	38,10%	28,57%	9,52%	23,81%	14,29%
Secondary education	26,00%	36,00%	24,00%	10,00%	6,00%
Higher education	30,00%	20,00%	30,00%	10,00%	10,00%
Total share	**29,63%**	**32,10%**	**20,99%**	**13,58%**	**3,70%**

TABLE XII TEACHERS 'ATTITUDE ABOUT THE BENEFITS OF OA TRAINING FOR THEIR ONLINE TEACHING

	Question/Answers: The experience of participating in online Oracle Academy training has helped me keep online classes for my students				
Type of educational institution	**Disagree**	**Rather disagree**	**Neutral**	**Rather agree**	**Fully agree**
Elementary education	0,00%	0,00%	23,81%	52,38%	23,81%
Secondary education	2,00%	10,00%	22,00%	36,00%	30,00%
Higher education	0,00%	20,00%	30,00%	30,00%	20,00%
Total share	**1,23%**	**8,64%**	**23,46%**	**39,51%**	**27,16%**

V. CONCLUSION

Online education is becoming the prevailing standard of professional development for computer science teachers due to several advantages. Based on the research conducted in this paper, it has confirmed that computer science teachers increasingly prefer online professional development, which is understandable due to their need for continuous improvement and connection to technology in everyday work. In the context of OA training, they have the view that online learning is not more demanding compared to comparable in-class learning (with the same learning outcomes and teaching materials) and have no problem with motivation to learn in an online environment. Teachers who complete OA training achieve equal success during the online and in-class training exams. They also think that the OA online training experience has helped them in online teaching for their students. Although not particularly significant in the context of the observed training, a slightly higher dropout rate in online training compared to comparable in-class training remains a topic for future research.

REFERENCES

[1] Industrial Registrar, New South Wales Industrial Gazette Vol. 349, Part 2 18 March 2005 Pages 265 - 543, 2005., ISSN 0028-677

[2] Oracle Corporation, Oracle Academy Training Opportunities, 2020., available at: https://academy.oracle.com/en/training-summary.html, [April 2020.]

[3] K. Thiam Seng (Ministry of Education, Singapore) and L. Sai Choo (Ministry of Education, Singapore), Information Communication Technology in Education - Singapore's ICT Masterplans 1997–2008, 2008, ISBN-13 978-981-281-848-5.

[4] F.Urem, D.Jureković, E. Ban, „Enhancing digital competences of Croatian teachers of informatics - Oracle Academy use case", 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018

[5] F.Urem, D.Jureković, E. Ban, „Engaging students in computer science with extracurricular projects – Oracle Academy Workshops in a Box use case", 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2019

[6] Cisco Systems, Inc, Cisco Webex, 2020., available at: https:// https://www.webex.com, [April 2020.]

[7] Zoom Video Communications, Inc, Zoom Solutions, 2020., available at: https://zoom.us, [April 2020.].

A Novel System for Automatic, Configurable and Partial Assessment of Student SQL Queries

Mario Fabijanić*, Goran Đambić* and Bojan Fulanović*
* Algebra University College/Software Engineering, Zagreb, Croatia
mario.fabijanic@algebra.hr, goran.dambic@algebra.hr, bojan.fulanovic@algebra.hr

***Abstract* - At some universities, students are required to write SQL queries to pass database-related courses. When it comes to assessing students' queries, three options are available: manual assessment (all work is done by teachers), semi-automated assessment (computer system provides parts of the final score, teachers give final mark), and fully-automated assessment (all work is done by a computer system). There are many currently available computer systems that can be used for both semi-automated and fully-automated assessments. While evaluating them, it became clear that their focus was not primarily on providing partial marks for substandard students' solutions and also not primarily on teachers' influence on the assessment process. Since these two requirements are very important for some universities, this paper proposes a novel fully-automated assessment system that is focused on satisfying these requirements. The early version of the system is built and evaluated against manual assessments of student work, resulting in the conclusion that the system is indeed very promising. Furthermore, the system discovered some inconsistencies in marks awarded manually by teachers, giving another perspective on the usefulness of the proposed system.**

Keywords – SQL; assessment; fully-automated; partial marks; configurable assessment process

I. INTRODUCTION

These years, there are more and more expectations for database technologies in general. For instance, the appearance of social network systems brought new demands for databases and systems that we knew earlier, which lead to the emergence of new database systems, like Not Only SQL Databases (NoSQL) [1]. For a while, traditional systems like RDBMS seemed to be losing their great market share, but the majority of business applications still use them, and RDBMS vendors are simultaneously finding a way to adapt for more and more data generated these years. An example is the availability of big data in educational contexts, which changes the way how the educational process can be monitored and evaluated [2]. Similar situations happened also earlier, in the 1990s when Object-Oriented Database Systems appeared with an idea to be a substitute for Relational Database Systems, but Object-Oriented databases still haven't found widespread acceptance like relational databases [3]. It is obvious that Relational Databases as technology survived challenges from the past, and not just that, Relational Databases can even perform as good as other proposed systems, if used properly [4]. When using up to the medium amounts of data that could be organized in row and column format, relational databases perform better [5]. So, each system could be the solution for particular use cases [6]. Business applications built for business operations will often be the best served by relational databases and may even be served exclusively by relational databases [6]. Also, even though the relational database idea is many decades old, innovative methods of teaching relational database normalization are invented these years [7].

Today's workplace is certainly dominated by enterprise-level DBMSs from vendors such as Oracle, IBM, and Microsoft. Having skill and experience with one of these enterprise-class DBMSs can greatly enhance a student's chances of a successful entry into the world of IT careers [8].

At the Algebra University College, database-related courses are called Introduction to Databases (obligatory for Software Engineering students and System Engineering students) and Database Development (obligatory only for Software Engineering students). The objectives of Introduction to Databases course are to learn how to design simple relational databases and to learn how to use SQL language to retrieve, insert, update, and delete data from a relational database. It means that students learn about relational databases from scratch, which turned out that it is very important since a lot of students are not even familiar with a need to organize data in a way to, for instance, avoid data inconsistency and redundancy. The objectives of Database Development course are to broaden knowledge of relational databases and SQL language and to learn how to prepare a database for access from programming code. More precise learning objectives are divided onto learning outcomes defined for both courses so that students are aware of what exactly they need to know to accomplish a positive evaluation of their knowledge per each learning outcome. So, knowledge evaluation tests are organized in a manner that students must provide their own SQL code for each test assignment. No single or multi-choice questions are used (or any other type of questions that would leave a possibility that students don't actually know how to write appropriate SQL code) because the idea is that students should build their solutions, test and debug them on sample database, which is the way of work that waits for them in the industry. Being said, it is the best possible way of knowledge evaluation.

Despite the quality of that kind of knowledge evaluation, there are also some challenges. First of all, around 240 students are currently enrolled in the course Introduction to Databases. Since teaching is held by two teachers, who also mark all exams, students are inevitably waiting for their results for a couple of days. After the process of evaluating, insights into the exams are held for anyone who wants to know where he or she made a mistake. From the experience, around 15% of students use this opportunity, which also prolongs the time when students' grades are finalized. All of that causes dissatisfaction for students and also their teachers, which are under time pressure, and more importantly, they become prone to make mistakes during assessments.

Instead, there are many papers that explore the use of computers for semi-automated or fully-automated assessment [10][11][12][13][14][15][16][17][18]. The semi-automated assessment provides parts of the final score, and teachers do the final grading. In the fully automated assessment, teachers do not have to grade each student's solution individually, but they need to prepare for setting up the grading process [9]. Also, semi-automatic assessment gives students grades with immediate feedback about imperfections, mistakes, and errors in their solutions, to help them understand both the mistakes and the given grade, and in that way they support the learning process [9].

Some of the existing solutions are:

SQL Tester [10] is a tool that gives full points to the student only if the student solution exactly matches the reference answer. That's not optimal because a student doesn't get any comments regarding his or her mistakes made in the solution. Instead, a student gets only an error message from the RDBMS.

SQLFE System [11] is an automatic grading system and system which implement peer review. The main reason for including peer review is to offer the students richer learning experience, but that makes it more complex since it employs more reviewers.

ASQLAG System [12] consists of two main parts: assignment management and automated SQL grader. The automated SQL grader is designed to evaluate only the results of the SQL statement.

SQLg [13], or SQL-Statement Grader, has a high-quality evaluation process that gives student feedback concerning syntax, helping them to identify their mistakes, but the tool only focuses on helping student practice the SQL statements, it cannot be used as for exam assessment.

SQL-KnoT [14], or Knowledge Tester, is a tool that also provides feedback to the student, but it compares the solution of the student with the correct result stored earlier. The student solution must be the same as the correct solution stored earlier.

SQLify [15] is evaluating the queries by the determining value of Conjunctive queries. Different reviewers make possible solutions, possible with different criteria. No feedback is provided.

ActiveSQL [16] supports online assessment and gives students immediate feedback. It also gives student grade defined as a percentage, comparing their results with the correct results, as the proportion of correct cells against the higher of either the total cell count of the correct solution of the total cell count of the student answer. It does not identify the student's understanding level and feedback given to the student is poor. There is also a lack of additional grading criteria.

SQLator [17] is a web-based tool that provides immediate binary feedback to the learners, so there is no useful feedback to students. SQLator does not check or analyze the SQL syntax at all, it provides the estimation: either a correct or incorrect statement.

AsseSQL [18] provides immediate feedback for the student solutions in a way to guide them to what corrections they need to apply compared to the results of the correct solution. If the submitted answer is syntactically incorrect, an SQL error message is displayed, without further explanation.

To summarize, none of those tools implement all of the following criteria for the student solutions:

- Is there any syntax error in the student solution?
- Is there any prohibited elements in the student solution?
- A tool should match student solution with the correct solution.
- A tool should give the result of the evaluation.

Also, it seems that all those solutions don't give the teachers the needed flexibility to define his or her own criteria for evaluating students' solutions.

Therefore, the motivation for this paper is to define a model and propose a tool for teachers to give them the ability to define exam question, to define one of the possible correct solutions, and to define the criteria for evaluation, for instance:

- Does the solution compile? If not, the percentage of point reduction might by 100%
- Is the FROM clause valid? If not, the percentage of point reduction might be 100%
- Is the WHERE clause valid?
    - If not, the percentage of point reduction is 50%
    - If yes, but with wrong columns included, the percentage of point reduction is 25% per the wrong column
    - If yes, with correct columns, but wrong values included, the percentage of point reduction is 25% per wrong value
- Is the SELECT clause valid?
    - If not, the percentage of point reduction is 10% per column that is missing

Also, the implementation of the model should be web-based and easy to use for students and teachers. It should provide feedback for students in the form of PDF report,

generated immediately after students upload their solutions, being able to download and analyze it. The report should use the above evaluation criteria, pointing at the parts where mistakes are being made, that way helping students to understand given points.

It needs to be said that the proposed model, in its initial version will be able to evaluate fundamental clauses of the SELECT statement (SELECT, FROM, WHERE, ORDER BY), fundamental clauses of the INSERT statement (VALUES), and fundamental clauses of UPDATE and DELETE statements. In future versions, additional clauses of SELECT statements will be added (GROUP BY, HAVING, all types of JOIN), but also SUBQUERIES as a querying method will be supported, additional INSERT possibilities will be possible to evaluate (like the combination of INSERT and SELECT statements), the composition of statements like when SELECT, INSERT, UPDATE and/or DELETE will also be possible to evaluate.

II. METHODOLOGY

The main focus when proposing our model for automatic assessment of SQL statements was the need of students and teachers; students' goals are to receive results as soon as possible, together with correct and helpful feedback about mistakes they made. Teachers' goals are to minimize efforts in marking student solutions and providing detailed feedback and also to maintain a consistent marking process for all students.

In the proposed model, when it comes to defining an exam, the teacher is required to define three inputs for every question they want to define: free form text of the question, one possible correct SQL answer (called a model) and which marking rules they want to apply to the process of marking student's solution (called a student answer, or answer for short). The system is designed to be extensible: currently, there are 15+ defined and implemented marking rules, but it is very easy to define additional. The main idea is that authors plan to receive teachers' feedback on their needs and then implement additional marking rules so each marking style can be accommodated in this system.

As the main forte of this proposed system is to allow awarding partial marks to student solution, we believe this approach of fine-grained marking rules will allow exactly that. Each marking rule consists of a condition, percentage of points to add or subtract if the condition is (not) met and whether marking can continue if the condition is not met or not. Fig. 1 shows marking rules for marking SELECT statement and a similar approach is taken for marking INSERT, UPDATE, and DELETE statements.

The proposed system comes with a few templates for marking each statement. The teacher can start with a template that is the closest match for his marking style, adapt it to fit her style perfectly, and then save it as a template for future use. Fig. 1 shows a template that is used at our home university for marking a basic SELECT statement. The first condition is that an RDBMS engine can compile the student's query, which is checked by sending the query to the engine and asking it to be compiled. If it does not compile, student feedback contains actual RDBMS error message, all points are subtracted and the process of marking is finished. FROM correctness is check next and if the student used the wrong data source, the process is also finished. Then, validation rules are checked against the main parts of SELECT statements. For each missing column in the SELECT list, 10% of points are subtracted. Optionally, the teacher can choose to subtract a certain percentage of points if the SELECT list is not exactly correct, without checking each individual column. Currently, there are two marking conditions that mark the WHERE part. The first one is used to check for a correct result of the WHERE filter, and the second one is used to check if correct columns were used for filtering. It is possible that the result of WHERE is wrong, but that the student used correct columns just with wrong operators, function, or literals. In this case, partial points are awarded. ORDER BY part is marked in a similar way: if the rows are not ordered as expected, a check is made to see if they are correctly ordered based on the first specified column in ORDER BY so partial points can be awarded.

To check marking rules, two protocols were defined, dynamic, and static. In the first, dynamic protocol, the marking rule check is done by sending the original or modified query to the RDBMS engine and deducing the marking result from RDBMS response. The second protocol is based on insight gotten from analyzing student query. The query is tokenized, converted into canonical form, semantics is given to sequences of tokens and then used to decide on the marking rule results. For example, if a model answer to a question is this:

```
SELECT Naziv, Boja
FROM Proizvod
WHERE Boja = 'Crvena'
```

And student answer is this:

```
SELECT p.Naziv, p.Boja
FROM dbo.Proizvod AS p
WHERE p.Boja = 'Crveni'
```

Marking rules 4 and 5 from Fig. 1 will be assessed by combining dynamic and static approaches. First, student answer will be transformed into this:

```
SELECT ProizvodId
FROM Proizvod
WHERE p.Boja = 'Crveni'
ORDER BY ProizvodId
```

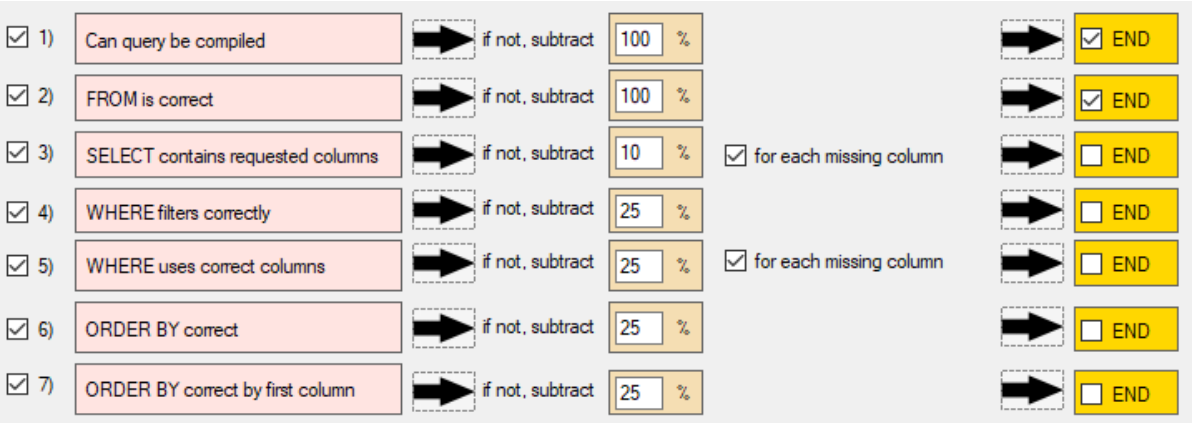


Figure 1. Marking rules for marking SELECT statement

After canonicalization and transformation, model and student queries are sent to the RDBMS engine. Returned result sets are checked for the exact match – if any value in any row is different in these two result sets, marking rule 4 is considered as failed and points are subtracted. Of course, by looking at model and student queries it is obvious that their results will not be equal, so marking rule 5 is the teacher's attempt to give partial points. This is done by static analysis in which a set of model filters is found to be { Column: "Boja", Operator: Equals, Value: "Crvena" } and a set of answer filters is { Column: "Boja", Operator: Equals, Value: "Crveni" }. The second protocol observes that both Column and Operators are the same and that Levenshtein's distance [19] between two Values is 1, so marking rule 5 is considered as passed and the student's answer is awarded partial points.

Some additional characteristics if the proposed model will be presented next. In order for this model to work, a previously prepared database should be made available. It should be the same database student used at the exam when writing her queries. It is also worth noting that marking SELECT statements, the proposed system will deduce the correct columns, operators, and values based on the model query, leaving the teacher free of any additional configurations for each question. When it comes to marking INSERT, UPDATE, and DELETE statements, the system uses transactions at serializable isolation level to make sure that the only state changes made to the database are the ones made by student queries. At the end of each question marking process, the transaction is rolled back, leaving the database at the original state and ready for the next marking. Certain questions are asked by the proposed system to be able to understand the context of both the model and answer queries, two examples being: what is the primary key of the FROM table and what is the complete set of columns available for the FROM table. Although the proposed model was developed for SQL Server, these questions and all other where possible were asked in the form of queries based on INFORMATION_SCHEMA schema, making it possible to work the same in most currently available RDBMSes on the market.

III. RESULTS

System for automatic assessment of SQL statements contains import feature which uses predefined well-formatted CSV file. The teacher fulfills CSV file with student solutions and imports them into the system. The teacher then needs to define three inputs for every question: free form text of the question, one possible correct SQL answer (called a model), and which marking rules they want to apply to the process of marking student's solution. After that, the system process the data and generate the report, also in the CSV file.

To validate the capabilities of the proposed model for automatic assessment of SQL statements, we used 135 grades given by teachers that manually evaluated those SELECT statement solutions at some earlier exam period and compared them with the grades that were given by the system for automatic assessment of SQL statements. Analyzed SQL statement assessments are restricted to simple SELECT statements, without JOINs, GROUP BY and HAVING clauses, aggregate functions, and subqueries. The Maximum for each grade is 1 point, and Fig. 2 shows that total points (for 135 evaluated assignments) given by the proposed system are close to total points given by teachers.

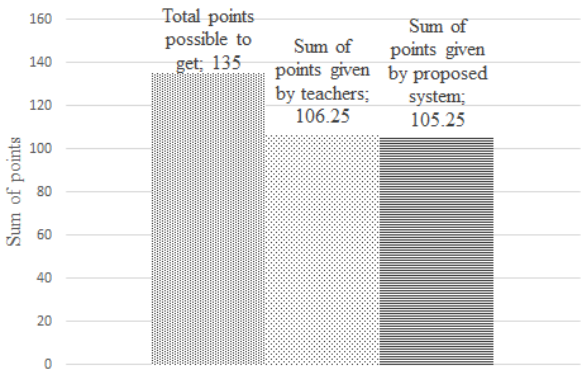


Figure 2. Total points possible compared with the sum of points given by teachers and by the proposed system

Also if we analyze an average of given points, we can see that are similar values of average points given by teachers and average points given by the proposed system (Fig. 3), which is an early indication that the proposed system could be appropriate for its purpose.

When analyzing the number of assignments and their distribution by given points (Fig. 4), we can see that number of assignments evaluated with a minimum (0) and maximum (1) points is bigger when the teacher evaluated assignments. When the proposed system evaluated assignments, more assignments are evaluated with partial points. It means mean that teachers more often evaluate on a binary basis, and the proposed system more often evaluate with partial points, which can be explained as more precise, therefore fairer.

Further on, we conducted statistical hypothesis testing using the Chi-square test ($\chi 2$ test) based on a sample of 135 grades. Our null hypothesis is that there is no significant difference between the grades given by the system and grades given by teachers. Table 1 shows observed frequencies of grades, given by the teacher and proposed system, and Table 2 shows expected frequencies of grades, needed to calculate degrees of freedom and $\chi 2$ value.

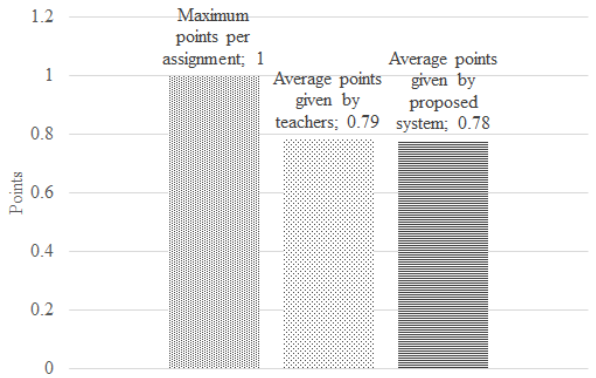


Figure 3. Maximum points per assignment compared with average points given by teachers and proposed system

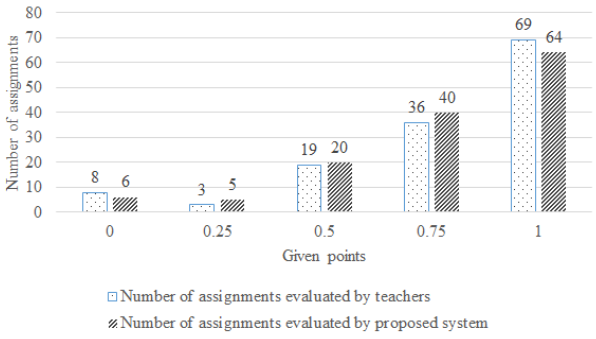


Figure 4. Number of evaluated assignments and their distribution by given points, by teachers and proposed system

TABLE I. OBSERVED FREQUENCIES OF GRADES

Grades	0	0.25	0.5	0.75	1	Total
Proportion of grades assigned by teachers	8	3	19	36	69	135
Proportion of grades assigned by system	6	5	20	40	64	135
Total	14	8	39	76	133	270

TABLE II. EXPECTED FREQUENCIES OF GRADES

Grades	0	0.25	0.5	0.75	1	Total
Proportion of grades assigned by teachers	7.0	4.0	19.5	38.0	66.5	135
Proportion of grades assigned by system	7.0	4.0	19.5	38.0	66.5	135
Total	14	8	39	76	133	270

The number of degrees of freedom is:

$$(Number\ of\ rows - 1) * (Number\ of\ columns - 1) = = (2 - 1) * (5 - 1) = 1 * 4 = 4 \quad (1)$$

$\chi 2$ value is calculated using the formula:

$$\chi 2 = \sum_{n=0}^{t} \frac{(f0 - ft)2}{ft} = 1.210 \quad (2)$$

For a significance level of 5% and 4 degrees of freedom (1), critical value $\chi 2$ is 9.488. Compared with the contingency table, $\chi 2$ value is 1.210 (2), so we concluded there is no statistical significance between grades given by the system compared with grades given by teachers, and our null hypothesis is not rejected.

IV. CONCLUSION

This paper proposed a novel model for automatic assessment of SQL statements which should help students and teachers; students will receive their grades quicker, but they will also get helpful feedback about mistakes they made, which is important for their learning process. As opposed to most existing solutions, this model is strongly focused on the partial marking of student solutions and giving teachers a tool to configure the exact marking procedures.

The system was configured by a teacher who also conducts manual marking of students' solutions and the system was used to mark 135 student solutions. Conducted statistical hypothesis testing using Chi-square showed that there is no statistical significance between grades given by the system compared with grades given by teachers, so the possibility of using this system in the university environment looks very promising. To do so, current limitations of the system need to be addressed: the system is currently restricted to simple SELECT statements, without JOINs, GROUP BY and HAVING clauses, aggregate functions, and subqueries.

Additionally, the system discovered cases with minimal, but existing nonconsistency mistakes made by teachers during manual grading resulting in different

numbers on awarded points for the same student solutions. This situation does not happen with our proposed system, resulting in additional positive effects on the grading process. From the work done in this paper, it would seem that this proposed system might be a good starting point in building a more comprehensive solution, with additional features like this (but not limited to): implementation remaining clauses in the SELECT statement, INSERT, UPDATE and DELETE statements, complex tasks that should be resolved with the combination of statements, etc.

REFERENCES

[1] Malik, A., Burney, A. and Ahmed, F. (2020) A Comparative Study of Unstructured Data with SQL and NO-SQL Database Management Systems. Journal of Computer and Communications, 8, 59-71.

[2] FISCHER, Christian, et al. Mining Big Data in Education: Affordances and Challenges. Review of Research in Education, 2020, 44.1: 130-160.

[3] ALZAHRANI, Hibatullah. Evolution of Object-Oriented Database Systems. Global Journal of Computer Science and Technology, 2016.

[4] Hohenstein, U. and Jergler, M. Database Performance Comparisons: An Inspection of Fairness. DOI: 10.5220/0007926602430250 In Proceedings of the 8th International Conference on Data Science, Technology and Applications (DATA 2019), pages 243-250.

[5] Aghi, Rajat, et al. "A comprehensive comparison of SQL and MongoDB databases." International Journal of Scientific and Research Publications 5.2 (2015): 1-3.

[6] NANCE, Cory, et al. Nosql vs rdbms-why there is room for both. 2013.

[7] AMIN, Mohammad, et al. Teaching Relational Database Normalization in an Innovative Way. The Journal of Computing Sciences in Colleges, 2019, 48.

[8] WAGNER, William P.; PANT, Vik. Using virtual servers to teach the implementation of enterprise-level dbmss: A teaching note. Journal of Information Systems Education, 2019, 21.4: 1.

[9] AL-SALMI, Aisha. Semi-automatic assessment of basic SQL statements. 2019. PhD Thesis. Loughborough University.

[10] 10. KLEEREKOPER, Anthony; SCHOFIELD, Andrew. SQL tester: an online SQL assessment tool and its impact. In: Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education. 2018. p. 87-92.

[11] WAGNER, Paul J. The SQL File Evaluation (SQLFE) Tool: A Flexible and Extendible System for Evaluation of SQL Queries. In: Proceedings of the 51st ACM Technical Symposium on Computer Science Education. 2020. p. 1334-1334.

[12] Singporn, Pornwipha, Pakpoom Vichianroj and Areerat Trongratsameethong. "ASQLAG - Automated SQL Assignment Grading System for Multiple DBMSs." (2018).

[13] KLEINER, Carsten; TEBBE, Christopher; HEINE, Felix. Automated grading and tutoring of SQL statements to improve student learning. In: Proceedings of the 13th Koli Calling International Conference on Computing Education Research. 2013. p. 161-168.

[14] BRUSILOVSKY, Peter, et al. An open integrated exploratorium for database courses. ACM SIGCSE Bulletin, 2008, 40.3: 22-26.

[15] DE RAADT, Michael; DEKEYSER, Stijn; LEE, Tien Yu. Do students SQLify? improving learning outcomes with peer review and enhanced computer assisted assessment of querying skills. In: Proceedings of the 6th Baltic Sea conference on Computing education research: Koli Calling 2006. 2006. p. 101-108.

[16] RUSSELL, Gordon; CUMMING, A. Automatic checking of sql: Computerised grading. In: Proceedings of the Twelfth International Conference on Learning, Granada, Spain. Published in the International Journal of Learning. 2005. p. 127-134.

[17] SADIQ, Shazia, et al. SQLator: an online SQL learning workbench. In: Proceedings of the 9th annual SIGCSE conference on Innovation and technology in computer science education. 2004. p. 223-227.

[18] PRIOR, Julia Coleman. Online assessment of SQL query formulation skills. In: Proceedings of the fifth Australasian conference on Computing education-Volume 20. Australian Computer Society, Inc., 2003. p. 247-256.

[19] LI, Yujian; LIU Bo. A Normalized Levenshtein Distance Metric. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 29, Issue: 6, 2007. p. 1091-1095.

Experience with e-learning in Teaching Combinatorics and Data Processing

M. Pokorný*
* Trnava University, Faculty of Education, Trnava, Slovak Republic
mpokorny@truni.sk

Abstract – **Young people widely use their mobile phones, tablets and notebooks. So it is natural to use modern technologies also in teaching mathematics. Blended learning represents a combination of e-learning and face-to-face instruction. In the ISET 2019 conference in Hradec Králové, we presented that blended learning had improved the results of our students in Combinatorics and Data Processing. In this paper, we continue the research and compare the results of students from three academic years. Moreover, we compare the results of external students who have to be taught only by e-learning due to restrictions caused by COVID-19.**

Keywords – e-learning; blended learning; mathematics teaching; ICT in education; interactive teaching materials

I. INTRODUCTION

Teachers have integrated modern technologies into education for more than forty years. During that long period of time, methods of modern technologies integration into education have significantly developed. Forty years ago, innovative teachers used isolated computer programs to reach some teaching aim. Nowadays, teachers use learning management systems, e-learning, blended learning, mobile learning, etc. Nowadays, all developed countries spend a huge amount of money to support proper integration of ICT into teaching process.

There are many studies focused on efficiency of modern technologies integration into education. Majority of them have indicated a significant improvement of the level of students' knowledge, if they are taught by e-learning or blended learning. On the other hand, there are studies that state that an improvement of the knowledge is not significant. In our opinion, modern technologies could be more efficient than a traditional face-to face instruction only if they utilize the advantages that cannot be reached in a traditional education. For example, placing a pdf document into learning management system could hardly be more efficient than learning from a printed book. We fully agree with Žilková [7], who states that the quality of electronic education is primarily determined by the accurate e-content.

It is generally known that mathematics could not be learned by passive reading of printed or electronic materials. In mathematics learning, an active approach of the learner is really important. Thus, if we want to integrate modern technologies into mathematics teaching, we have to use methods that activate students. It is important to change students from passive listeners or readers to active learners. It could be done for example by an integration of interactive applications into e-learning materials. In a summarizing study about mathematics teaching using ICT, Žilková [8] underlines the need for creation of interactive environment and for utilization of dynamic activities in mathematics teaching. Thanks to the interactive applications, students are much more involved in their learning process, which have a positive impact on their learning outcomes. The successful integration of modern technologies into mathematics teaching, which properly uses interactivity, can be found for example in [1], [5], [6]. The study of Malatinská, Pokorný, and Híc [2] proved the improvement of secondary school students' knowledge caused by proper integration of interactive applications into mathematics teaching, as well as the improvement of students' attitudes towards mathematics.

In this paper, we focus on the efficiency of an e-learning course that contains a lot of interactive applications.

II. DESCRIPTION OF THE RESEARCH

In this part of the paper, we analyze the results of the research focused on the efficiency of e-learning in teaching mathematics at Trnava University. The research builds on the findings of the research of Pokorný [3]. In this research, Pokorný compared the results of the students from the subject Combinatorics and Data Processing. In the academic year 2017/2018, students were taught by a combination of ten face-to-face lessons with a teacher and study materials from the teacher. In the following academic year, the subject was taught by blended learning, which consisted of ten face-to-face lessons with a teacher and the e-learning course, which contains 73 interactive applications. At the face-to-face lessons, a teacher explained the theory, presented a systematic way of solving combinatorial problems, explained how to efficiently study the corresponding part of an e-learning course, and underlined the need for an active approach to learning by solving problems and working with interactive applications.

A comparison of the results of students from both academic years proved a significant difference in the results of the final test in favor of the group taught by blended learning in the academic year 2018/2019. In our research, we expand the findings from the above mentioned research. We focus on the results of the following four groups of students:

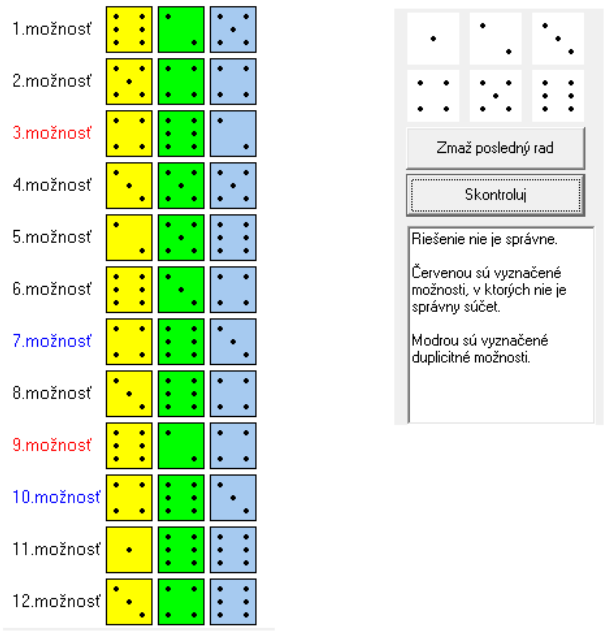


Figure 1. Example of a feedback after an incorrect solution.

Group A: Students taught by a combination of ten face-to-face lessons with a teacher and study materials from the teacher in academic year 2017/2018.

Group B: Students taught by a combination of ten face-to-face lessons with a teacher and an e-learning course with interactive applications in academic year 2018/2019.

Group C: Students taught by a combination of ten face-to-face lessons with a teacher and an e-learning course with interactive applications in academic year 2019/2020.

Group D: External students taught only by an e-learning course with interactive applications in a summer term of academic year 2019/2020. These students could not have face-to-face lessons with a teacher due to restrictions caused by COVID-19 in the Slovak Republic.

As we have mentioned above, all groups had electronic study materials. However, the main difference is in the level of their interactivity. The students of the group A had materials from the teacher, but the feedback provided to the students about correctness of their solutions of the tasks was limited. They had instructions from the teacher how to study actively, but if they had problems with any task, the printed materials could not warn them about their mistake or give them an appropriate advice. Thanks to the interactive applications, the students of the groups B, C, and D were offered more detailed feedback. They were encouraged to study combinatorics and data processing by an active way, and the interactive applications checked the correctness and completeness of their solutions, as well as give an advice in case of wrong solution. Thus, groups B, C, and D were made to learn mathematics by their active approach to learning more than the students of the group A. As we will prove later, the provided feedback had positive impact not only on the level of knowledge of students, but also on their ability to evaluate their level of knowledge and their satisfaction with the final assessment from the subject.

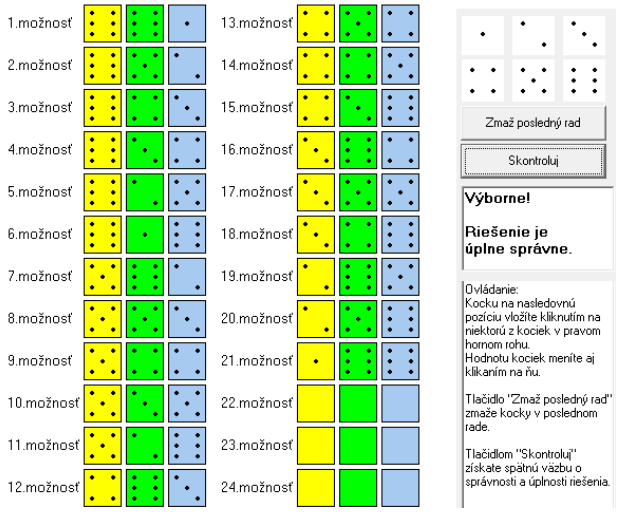


Figure 2. Example of an interactive application – a correct solution.

As we have already mentioned, we consider blended learning as an ideal way how to learn combinatorics and data processing at our faculty. However, the situation at Slovak universities in the summer term of the academic year 2019/2020 was influenced by COVID-19. Since it was not possible to teach face-to-face lessons, the group D obtained only detailed written instruction from the teacher how to study and an e-learning course. We utilized this situation as a possibility to compare the efficiency of blended learning and e-learning of the subject *Combinatorics and Data Processing*.

The groups B, C, and D could benefit from 73 interactive applications, which were added into the e-learning course *Combinatorics and Data Processing*. These applications enable to fully utilize the advantages of modern technologies, which cannot be reached by traditional printed materials. An example of the application is depicted in Figures 1 and 2. In the application, the task is to find all possible different ways how we can obtain sum 13 when rolling three different dice. During working with the application, students can anytime check the correctness and completeness of the solution. In Figure 1, a feedback after the wrong solution is depicted. The duplicate records are highlighted by a blue color, while the records with different sum from 13 are highlighted by a red color. Naturally, the students are allowed to continue their solution until it is complete and correct. Working with applications teaches them to choose a suitable and reasonable procedure to solve combinatorial tasks.

TABLE I. RESULTS OF THE FINAL TEST

Group	***Numbet of students***	***Average score***	***W***	***Null hypothesis***
A	128	64.20	0.973	Rejected
B	137	73.07	0.985	Accepted
C	127	72.12	0.978	Rejected
D	44	73.66	0.984	Accepted

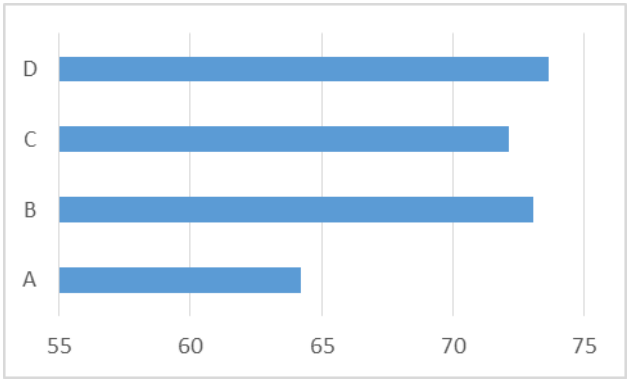


Figure 3. Average score in the final test

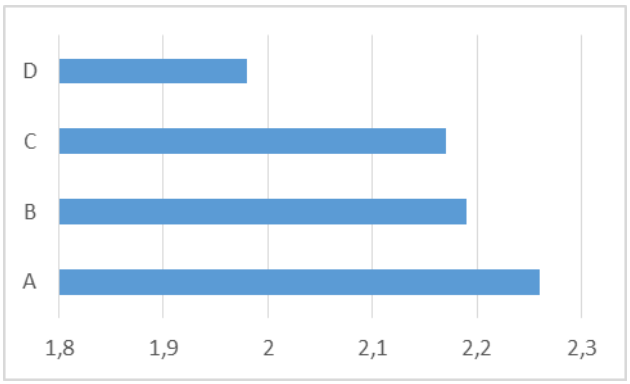


Figure 4. Average study mean of the students

The active approach to learning for groups B, C, and D was also supported by integration of three tests, which provided students with an immediate feedback about their level of knowledge from combinatorics, combinatorial thinking in probability, and data processing. Thus, students were able to prepare for the final test better.

As we have already mentioned, in [3] a significant improvement of the results was proved between groups A and B. To evaluate the efficiency of our e-learning course and interactive applications, we expanded the research by another two groups, C and D. Moreover, the group D was taught only by e-learning course without any face-to-face instruction, which enables to study the importance of face-to-face instruction.

At the end of the term, all students had to pass a final test. Let us notice that the level of difficulty of the final test was very similar for all groups. All tests consisted of ten tasks, which focused on a specific problems. For example, the first task was focused on creation of all ten permutations with repetition from the set of five objects, from which two and three are the same. The second task was focused on creation of a subset of the permutations of four different objects, etc.

To compare the knowledge of the groups, we analyze the results of the groups in the final test. The summary of the results can be found in Table I. The average score of the groups in the final test are depicted in Figure 3.

We can notice that the results of the group A are much worse than the results of groups B, C, and D, which are quite similar. To be able to verify the significance of the differences, we used the statistical tests. Firstly, we used the Shapiro-Wilk test to test the normality of the results of each group. The values of the test statistic *W* are given in Table I, as well as the result about the rejection of the null hypothesis in the Shapiro-Wilk normality test. Since we reject the null hypothesis for groups A and C, we do not compare the results by a parametric t-test, but by a non-parametric Mann-Whitney U test. The values of the test statistic, critical value accepted range, and a conclusion about rejection of the null hypothesis are given in Table III. From data in Table III it follows that there is a significant difference between the results of groups A and B, A and C, and A and D. In all cases, the group A is significantly worse. Moreover, no significant difference between the results B, C, and D was proved by Mann-Whitney U test. Thus, we can conclude that all groups taught by our e-learning course with interactive applications are better than the students of the group, which did not have the e-learning course.

TABLE II. AVERAGE STUDY MEAN OF THE STUDENTS

Group	***Numbet of students***	***Average study mean***	***W***	***Null hypothesis***
A	128	2.26	0.945	Rejected
B	137	2.19	0.981	Rejected
C	127	2.17	0.978	Rejected
D	44	1.98	0.935	Rejected

TABLE III. COMPARISON OF THE RESULTS OF THE FINAL TEST

Compared groups	***Value of the test statistic***	***Critical value accepted range***	***Null hypothesis***
A – B	4.43	[-1.96:1.96]	Rejected
A – C	3.93	[-1.96:1.96]	Rejected
A – D	3.26	[-1.96:1.96]	Rejected
B – C	0.72	[-1.96:1.96]	Accepted
B – D	0.15	[-1.96:1.96]	Accepted
C – D	0.64	[-1.96:1.96]	Accepted

To exclude the possibility that the results of group A are significantly worse because the students are worse in general, we compare the results of students in five different subjects at our faculty. The summary of the average study mean of all groups in these subjects can be found in Table II. The average study mean of the groups is depicted in Figure 4.

We can notice that the average study mean is the best in group D and the worse in group A, while the means of groups B and C are quite similar. To be able to verify the significance of the differences, we again used the statistical tests. Firstly, we used the Shapiro-Wilk test to test the normality of the average study means of each group. The values of the test statistic *W* are given in Table II, as well as the result about the rejection of the null hypothesis in the Shapiro-Wilk normality test. Since we reject the null hypothesis for all groups, we do not compare the results by a parametric t-test, but by a non-parametric Mann-Whitney U test. The values of the test

statistic, critical value accepted range, and a conclusion about rejection of the null hypothesis are given in Table IV. From data in Table IV it follows that there is a significant difference between the results of groups A and D, as well as between groups B and D. In both cases, the group D is significantly better. However, we can notice that no significant difference between the results of groups A, B, and C was proved by Mann-Whitney U test. Thus, we can conclude that we can consider the groups A, B, and C to be equivalent, while the group D is generally better.

Further, we pay attention to the analysis of the results of questionnaire, which was focused on opinions of students on combination of e-learning and face-to-face instruction. The questionnaire was administered in groups B and C.

Firstly, we compare the opinions of students on the objectiveness of their assessment, and their satisfaction with the assessment. The students expressed their opinion on the statements:
Q1: My final evaluation from *Combinatorics and Data Processing* is objective.
Q2: My final evaluation from the other subjects is objective.
Q3: I am satisfied with my final assessment from *Combinatorics and Data Processing*.
Q4: I am satisfied with my final assessment from the other subjects.
The results are depicted in Figure 5. As we can see, the majority of our students consider their final evaluation from *Combinatorics and Data Processing* to be objective. Moreover, the majority of our students are satisfied with their final evaluation. We think that it is caused by a wide feedback of e-learning course about a level of students' knowledge. However, the responses of students to the same questions about other subjects are significantly different. About a quarter of students do not consider their evaluation from the other subjects to be objective and they are not satisfied with their evaluation. Thus, we conclude that our e-learning course has a positive impact on the ability of students to judge their level of knowledge. Let us notice that there is no significant difference between the average assessment from *Combinatorics and Data Processing* and from the other subjects.

Secondly, we focus on the opinions on the statements:
Q5: E-learning course was really helpful.

TABLE IV. COMPARISON OF THE AVERAGE STUDY MEANS

Compared groups	***Value of the test statistic***	***Critical value accepted range***	***Null hypothesis***
A – B	0.08	[-1.96:1.96]	Accepted
A – C	0.36	[-1.96:1.96]	Accepted
A – D	1.98	[-1.96:1.96]	Rejected
B – C	0.21	[-1.96:1.96]	Accepted
B – D	2.19	[-1.96:1.96]	Rejected
C – D	1.94	[-1.96:1.96]	Accepted

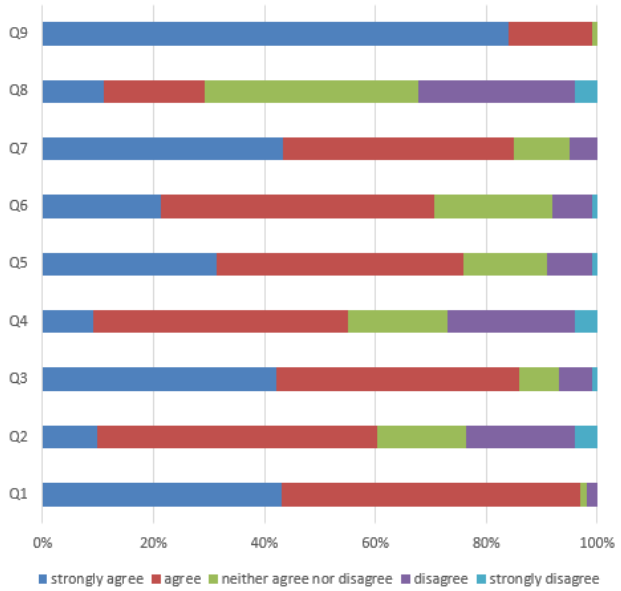


Figure 5. Opinions on the statements Q1 – Q9.

Q6: Interactive applications were really helpful.
Q7: I would like to have similar study materials for all subjects.
Q8: E-learning suits me better than classical face-to-face instruction.
Q9: In mathematics, I prefer written exams to oral ones.
As we can see, 75 per cent of students consider our e-learning course to be really helpful, and 70 per cent of them consider interactive applications to be helpful. 84 per cent of our students would like to have similar study materials for all subjects. However, it is not possible to state that our students prefer e-learning to a classical face-to-face instruction. Thus, blended learning seems to be an ideal combination for them. Finally, 99 per cent of our students prefer written exams in mathematics to the oral ones.

Thirdly, we asked the students whether they had printed the texts from the e-learning course. 30 per cent of our students had printed all the texts from the course, 26 per cent had printed the majority of the texts, and 23 per cent had printed some texts. Thus, there is a significant number of students, who prefer reading texts in a printed form.

III. CONCLUSION

Although a preparation of interactive e-learning courses takes a lot of time, it is clear that it is worth to invest time and money to properly designed interactive e-learning courses that make students to learn mathematics by active solving of tasks.

The analysis of the results of our research revealed that blended learning, which combines ten face-to-face lessons with a teacher and an interactive e-learning course, is the best alternative how to teach the subject *Combinatorics and Data Processing* at our faculty. Moreover, the research proved that thanks to the interactive e-learning course, the students have not problems to master the content of the course although without face-to-face instruction, which could not take place due to restrictions

caused by COVID-19. Although the results of group D were equivalent, we could not state that it is possible to replace blended learning of *Combinatorics and Data Processing* at our faculty by e-learning, because the students from group D are better in other subjects.

To end up, we agree with Thorne [4], who states that blended learning could become one of the most significant developments of the 21st century.

ACKNOWLEDGMENT

The paper was supported by the KEGA 001UMB-4/2020 project entitled *Implementation of Blended Learning into Preparation of Future Mathematics Teachers and Future Computer Science Teachers*.

REFERENCES

[1] J. Fialová, "On a Mobile Application for Learning Positive and Negative Integers," 2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA), Starý Smokovec, Slovakia, 2019, pp. 179-184.

[2] S. Malatinská, M. Pokorný, and P. Híc, "Efficiency of Blended Learning in Teaching Mathematics at Primary School." Information, Communication and Education Application, Advances in Education Research, vol. 85, pp. 6-11, 2015.

[3] M. Pokorný, "Blended Learning Can Improve the Results of Students in Combinatorics and Data Processing," 2019 International Symposium on Educational Technology (ISET), Hradec Kralove, Czech Republic, 2019, pp. 207-210,

[4] K. Thorne, Blended learning: how to integrate online and traditional learning. London: Kogan page, 2003.

[5] P. Voštinár, "GeoGebra Applets for Graph Theory," EDULEARN17 Conference, Barcelona, 2017, pp. 10142-10148.

[6] P. Voštinár, "Pythagorean Theorem – Mobile Application," INTED 2018, pp. 0800-0805.

[7] K. Žilková, "Manipulating geometry course and its advantages and risks in e-learning education," / "Prednosti a riziká vzdelávania prostredníctvom e-learningového kurzu manipulačná geometria," XXVI. Didmattech 2013: New Technologies in Science and Education, pp. 222-227, 2014.

[8] K. Žilková, Školská matematika v prostredí IKT. Bratislava: Pedagogická fakulta UK Bratislava, 2009.

Assessment of The Role of Technology in the Process of Tolerance

E. Gürbüzer*, G. Dağlı**, F. Altinay* and Z. Altinay*
* Graduate School of Educational Sciences, Near East University, Nicosia, Cyprus
** Faculty of Education, University of Kyrenia, Kyrenia, Cyprus
ergengurbuzer@gmail.com

Abstract – Technology is the most important communication tool that enables people to reach agreement, show tolerance and get closer to each other. Tolerance is the process of individuals' empathizing with the feelings, thoughts, and behaviors of the people around them, and at the same time, respecting and trusting them in an unrequited way. The aim of the present study is to assess the role of technology in the process of tolerance according to computer science teachers. In the study, qualitative research approach was used. The aim was to obtain significant and detailed data regarding the role of technology in the process of tolerance by using case study. The research was carried out with 20 computer science teachers chosen with maximum variation sampling method. In the research, semi-structured interview was used as a type of data collection technique. The data of the research were analyzed through content analysis. In conclusion, technology use helps to establish healthy communication between teachers. Also, technology helps to close communicative gaps in the process of tolerance. At the same time, technology plays a major role in communicating to increase tolerance because technology is a factor that helps bring teachers closer.

Keywords – technology; role of technology; tolerance; computer science teacher; school

I. INTRODUCTION

One of the issues that still remains on the agenda today is the importance of human rights. Human rights are the rules that ensure people to maintain their lives in the most efficient way and aim to protect humanitarian values [1, 2, 3]. What is meant by human rights is essentially the existence of universal rights. Universal rights are all of the rights that are worthy of human dignity and are of vital importance [1, 3, 4].

There are many elements within the concept of human rights. Some of these are concepts such as freedom, democracy, education, health, peace and justice [1, 2, 3]. Undoubtedly, one of the most important of these concepts is tolerance. Tolerance is one of the most important concepts that differentiates human from other living things. This is because tolerance is one of the most important values that make human beings human [4].

Today, with the effect of globalization, cultures are intertwined or different cultures are living together [5, 6, 7, 8, 9]. Therefore, it is important that all people on earth learn to live in peace [10]. The creation of this peaceful environment is, of course, linked to tolerance. It is important that people embrace everyone around them without discrimination, and respect the differences by always giving priority to their tolerant behavior. It is an undeniable fact that life will be easier for everyone if individuals show tolerance in every aspect of life [11].

It is important to exhibit tolerant behaviors in all living spaces where people and other living things take place [12, 13]. It is an undeniable fact that the climate and culture of the school will be positively affected if teachers, administrators, students and parents under the roof of the school mutually tolerate each other. It can be argued that as a result of tolerant behaviors, the efficiency and quality of the school in educational and instructional areas will increase [14, 15, 16].

Technology is the most important communication tool that enables people to reach agreement and get closer to each other. Without technology, people cannot express themselves and cannot meet on a common ground in today's context [17, 18].

When the previous studies in this field were examined, it was observed that predominantly the studies on the role of technology were conducted. From this point of view, it can be argued that the role of technology in the process of tolerance is not studied sufficiently. It is anticipated that this research is important for the above reasons and will advance the previous studies by filling the gaps in the related field.

The purpose of the present study is to assess the role of technology in the process of tolerance according to computer science teachers. After this study, it is anticipated that a more democratic school dynamics will emerge among teachers and tolerant teacher behaviors that will support education management processes will be encouraged. In addition, it is thought that technology will be used more effectively among teachers and the role of technology in displaying tolerant behaviors will be considered more. At the same time, it is anticipated that the efficiency of education management processes can be increased in line with the results and recommendations obtained from this study. Answers to the following questions were sought in line with the purpose of the study:

i) What do computer science teachers think about the importance of technology use in the process of tolerance?

ii) What are the opinions of the computer science teachers about the role of technology in the process of tolerance?

II. METHOD

A. Research Model

In the present study, qualitative research approach was used. In this research, the aim was to obtain significant and detailed data regarding the role of technology in the process of tolerance by using case study. Case study is a research method that includes in-depth and detailed examination of a study subject as well as contextual conditions related to a subject in social sciences [19, 20].

B. Study Group

The study group of the research consisted of 20 computer science teachers. Maximum variation sampling method, which is one of the purposeful sampling methods was used in the research.

Purposeful sampling is widely used to identify and select knowledge intensive cases in qualitative researches [21]. In this research, the aim of using purposeful sampling is to form a data-rich group of participants. Maximum variation is a purposeful sampling strategy and its primary objective is to exemplify the heterogeneous structure [22]. The reason for using maximum variation in this research is to explain the problem more thoroughly by forming a wide range of participants.

While determining the computer science teachers who form the study group of the research, variables such as the genders of computer science teachers, their years of seniority in the profession, the district of the school they work at are, and the socio-economic structure in the region where the schools they work at are, were considered and the aim was to keep the variety of the study group at maximum level.

C. Data Collection Tools and Techniques

In the research, interview was used as qualitative data collection technique. Interview is a widely used data collection technique in qualitative researches [23]. The reason for using interview technique in the research is that this technique is superior in terms of revealing the experiences, emotions, thoughts and opinions of participants and is based on speech. Interview technique is divided into three, being unstructured interview, structured interview and semi-structured interview [24].

In this research, semi-structured interview was used as a type of data collection technique. Semi-structured interview is the most commonly used research method in social sciences. Whereas there are solid questions, which do not enable the orientation of someone in structured interview, it is possible that new ideas come in sight with the answers of the interview, by adding flexibility to the questions, thanks to the semi-structured interview [21]. Semi-structured interview technique was used in this study, both because it draws a road map for the researchers and because it brings in flexibility regarding the questions.

Before the interview questions prepared by the researchers were prepared, a detailed literature review was conducted on the concept of technology and tolerance. The interview questions were prepared by taking the opinions of three experts in the fields, two qualitative research experts and one language expert, following which 2 open ended questions were formed in total. Pilot scheme was performed; the interview form was tested on two computer science teachers and it was finalized in line with the answers and reactions of the computer science teachers interviewed and entitled "The role of technology in the process of tolerance – Interview Form". The purpose of this interview form is to reveal and evaluate the role of technology in the process of tolerance.

D. Data Collection Process

An official permission letter was received from Northern Cyprus Ministry of Education for the interview form. Interviews were conducted with the computer science teachers in the sample between March 2019 and May 2019. Interviews were conducted within the working hours, in the teachers' own schools, on a voluntary basis and with an official permission letter.

Before the interview, participants were reminded of the ethical rules. Within the framework of ethical rules, it was reminded that participation in the research was purely based on voluntariness, that identity information would not be shared with anyone other than the researcher and that the collected data would not be used for a purpose other than academic researches.

Since the participants stated that they were disturbed by the voice recording during the interviews, written record was performed instead of voice recording. At the end of the interview, each participant was asked if they had any opinions they wanted to add or remove, and the interviews were ended according to their wishes. It usually took 15 to 20 minutes for the participants to answer the questions.

E. Data Analysis

The data of the research were analyzed through content analysis. Content analysis is a research method that examines and analyzes different forms of documents such as text, image, audio, video and communication tools. Social scientists in particular use content analysis to systematically examine patterns of communication [25]. In other words, content analysis is an analysis technique that can collect and interpret similar data together in order to reach expressions and connections that can make sense of the data [23, 26]. Since the aim is to study the similarity relations of the data obtained and to gather and systematically analyze similar data, content analysis was utilized in the research.

In the first stage of the analysis of this study, the researchers conducted literature review in the field and since there were no explanatory categories, they formed their on categories based on the data obtained. Accordingly, the researchers adopted the induction approach and codings were made. In the second stage, themes were revealed. In the third stage, the data obtained from the participants were gathered and grouped on the basis of themes. In the last step, the present findings were explained in detail by utilizing numbers and tables.

Throughout the analysis of the interviews, abbreviation codes were formed for each participant,

being T1, T2, T3, ... and T20. For instance, T2 abbreviation code refers to the second computer science teacher with whom the interview was conducted. The purpose of these abbreviation codes is both to keep the names of the participants confidential and to provide convenience during the study.

The findings obtained as a result of data analysis were presented in tables in parallel with the sub-objectives of the research and all necessary explanations were made.

F. Validity and Reliability

Regarding internal validity, the data were written in detail and the way the results were obtained was explained clearly. During the explanations, the views of the participants were frequently quoted directly and the data were supported by these views. It was also considered important that the research findings were both consistent and significant in themselves and the data revealed had the quality to form a whole. Regarding external validity, the necessary explanations for the obtained findings to be tested in other studies are included in the research. For the results in the research to be generalized and transferred to similar environments and groups, the readers were informed in detail about all stages of the study.

In order to ensure the internal reliability of the analysis, the proposed formula (Reliability = Consensus / (Consensus + Disagreement)) was used in the reliability calculation [27]. According to [27], the fact that the result of reliability calculations for the researches is over 70% makes the research reliable. Since 95% consistency is obtained in the research, it can be argued that the interview records in the study are reliable in terms of the identified categories. The characteristics of participants such as their gender, professional seniority and field of study were identified for external reliability. In this way, other researchers conducting similar studies may consider these definitions while forming their samples. At the same time, the data of the study were stored so that they could be reviewed later by others on demand.

III. FINDINGS

A. The Importance of Technology Use

The opinions of computer science teachers regarding the importance of technology use in the process of tolerance were inquired and the related findings were presented in Table 1.

TABLE I. THE THOUGHTS OF COMPUTER SCIENCE TEACHERS ON THE IMPORTANCE OF TECHNOLOGY USE IN THE PROCESS OF TOLERANCE

Themes	**Number of opinions**	**Percentage (%)**
Technology use helps to establish healthy communication between individuals	12	60
Technology use helps to share ideas	8	40

As can be seen in Table 1, 12 computer science teachers think that technology use helps to establish healthy communication between individuals. This opinion is the most expressed by the teachers. Participant T6 expressed the opinion by stating that *"Communication and interaction between individuals increase with technology use. In this way, the climate of the individuals' environment is positive. Therefore, the importance of technology use should never be underestimated."*

8 computer science teachers expressed their opinions by saying that technology use helps to share ideas. Participant T19 expressed his opinion on this issue in the words *"People have different ideas. People can only meet on a common ground if they share their ideas with each other. This is only possible with the correct technology use."*

B. The Role of Technology in the Process of Tolerance

The opinions of computer science teachers about the role of technology in the process of tolerance were inquired and the relevant findings were presented in Table 2.

TABLE II. THE THOUGHTS OF COMPUTER SCIENCE TEACHERS ON THE ROLE OF TECHNOLOGY IN THE PROCESS OF TOLERANCE

Themes	**Number of opinions**	**Percentage (%)**
Technology is a key to understanding	7	35
Technology helps to close communicative gaps in the process of tolerance	7	35
Technology helps teachers to express their tolerance behaviors and expectations	6	30

When Table 2 is examined, 7 computer science teachers are observed to have the opinion that technology is a key to understanding. Regarding this issue, T13 conveyed her opinion by stating that *"The role of technology is at the forefront in a multicultural school. Because of technology, teachers learn to respect each other while they are together. At the same time, they learn to develop tolerance for differences."*

7 computer science teachers stated that technology helps to close communicative gaps in the process of tolerance. T9 shared his thought on this issue by stating that *"Technology helps to spread tolerance among teachers. With the use of technology, you gain a broader understanding of a particular group of people, thus closing the gap between your differences."*

6 computer science teachers stated that technology helps teachers to express their tolerance behaviors and expectations. T1 conveyed his thoughts on the issue by stating that *"With the help of technology, teachers can reflect tolerance behaviors towards their colleagues. In summary, through technology, teachers can express how they attribute tolerance and what their expectations are."*

IV. DISCUSSION AND SUGGESTIONS

The main purpose of this research is to assess the role of technology in the process of tolerance according to computer science teachers. In the study conducted in accordance with this purpose, it was found that 60% of

teachers stated that technology use helps to establish healthy communication between individuals. As [17] stated, technology use allows the generated messages to be transmitted to the other party. Thus, technology use prepares the ground for the realization of communication. Reference [28] obtained similar results in their research by emphasizing that technology use reminds people of their existence, socializes them, and makes communication effective by offering them common living spaces. The remaining 40% of teachers stated that technology use helps to share ideas. Reference [28], who conducted research on this subject, emphasized that technology use draws a conscious path to communication and it provides the best transmission of thoughts. Based on these results, it can be said that technology use is the primary requirement of communication. In this context, school administrators have a big duty. School administrators are suggested to encourage teachers to use technology more.

Regarding the role of technology in the process of tolerance, 35% of teachers stated that technology is a key to understanding. As [17] stated, technology is a priority tool for understanding. In this context, it can be argued that in order to work in a cultural school, it is important to bring together teachers with different ideologies, behaviors, qualifications and policies. In this way, the country can produce a better result and create a successful, harmonious, pragmatic and tolerant society. At the same time, it can be argued that technology is a factor that helps bring teachers closer. The remaining 35% of teachers stated that technology helps to close communicative gaps in the process of tolerance. As [28] stated, technology is transmitted with the help of its constituent components. Technology is responsible for the formation and transmission of the messages as desired. At this point, it can be argued that technology is very important in eliminating communicative errors and gaps in the process of tolerance. Based on these results, it can be said that technology is a great tool for raising awareness and understanding of spreading tolerance among teachers. The remaining 30% of teachers stated that technology helps teachers to express their tolerance behaviors and expectations. In the light of this result, it can be said that technology is one of the most effective tools by which teachers express their tolerant behaviors and expectations about tolerance. In this point, it can be suggested to transform technology literacy into a state policy and to organize in-service trainings for teachers.

V. CONLUSION

In conclusion, technology use helps to establish healthy communication between teachers. The role of technology in the process of tolerance should never be overlooked because technology is an indispensable tool to understanding. At the same time, technology helps to close communicative gaps in the process of tolerance. In addition, technology plays a major role in communicating to increase tolerance because technology is a factor that helps bring teachers closer.

REFERENCES

[1] Y. K. Kepenekçi and C. Aslan, ‘‘An analysis on human rights in high school turkish literature and technology and expression textbooks,’’ International Periodical for The Languages, Literature and History of Turkish and Turkic, vol. 6, pp. 476-494, 2011.

[2] M. Bajaj, Human Rights Education: Theory, Research, Praxis. Philadelphia, PA: University of Pennsylvania Press, 2017.

[3] J. Donnelly and D. Whelan, International Human Rights, 5th ed., New York, NY: Routledge, 2018.

[4] M. A. Özdilekler, A. Günsel, G. Dağlı, and E. Gürbüzer, ‘‘How symbols and rituals affect school culture and management,’’ in Open and Equal Access for Learning in School Management, F. Altınay, Ed. London: IntechOpen, 2018, pp. 77-93.

[5] B. Almond, ‘‘Education for tolerance: cultural difference and family values,’’ Journal of Moral Education, vol. 39, pp. 131-143, 2010.

[6] O. H. B. Hansen, ‘‘Teaching tolerance in public education: organizing the exposure to religious and life-stance diversity,’’ Religion & Education, vol. 38, pp. 111-127, 2011.

[7] P. Zgaga, ‘‘Education for ‘‘a better world’’: is it still possible?,’’ Education Inquiry, vol. 2, pp. 331-343, 2011.

[8] N. Davids and Y. Waghid, Tolerance and Dissent within Education, 1st ed., London: Palgrave Macmillan, 2017.

[9] K. Fleming, ‘‘Who is ‘‘diverse’’?: (in) tolerance, education, and race in Hong Kong,’’ in Education, Ethnicity and Equity in the Multilingual Asian Context, J. Gube and F. Gao, Eds. Singapore: Springer, 2019, pp. 89-103.

[10] M. S. Alzyoud, A. F. Al Khaddam, and A. S. Al-Ali, ‘‘The impact of teaching tolerance on students in jordanian schools,’’ British Journal of Humanities and Social Sciences, vol. 15, pp. 33-52, 2016.

[11] E. Gürbüzer, F. Altınay, Z. Altınay, and G. Dağlı, ‘‘An evaluation of educational and administrative decision making process in tolerance school context for job motivation,’’ Quality & Quantity, vol. 52, pp. 1059-1076, 2018.

[12] F. Luthans, B. C. Luthans, and K. W. Luthans, Organizational Behaviour: An Evidence-Based Approach, 13th ed., Charlotte, NC: IAP-Information Age Publishing, Inc., 2015.

[13] İ. Yaraş, ‘‘The evaluation of the primary school teachers having the value of tolerance and fair in which take place in fifth class students’ social science schedule, unpublished master’s thesis,’’ Mustafa Kemal University, Institute of Social Sciences, Hatay, Turkey, 2015.

[14] H. Türe and A. F. Ersoy, ‘‘Social studies teachers’ perceptions of tolerance,’’ Journal of Social Studies Education Research, vol. 5, pp. 31-56, 2014.

[15] F. Altinay, G. Dagli, and Z. Altınay, ‘‘Role of technology and management in tolerance and reconciliation education,’’ Quality & Quantity, vol. 51, pp. 2725-2736, 2017.

[16] K. Ozberk and F. Altinay, ‘‘Opinions of teachers and school administrators about organizational citizenship behaviors and disability services,’’ International Journal of Disability, Development and Education, vol. 66, pp. 551-564, 2019.

[17] T. M. Holtgraves, Technology as Aocial Action: Social Psychology and Technology Use. Mahwah, NJ: Lawrance Erlbaum Associates, Publishers, 2002.

[18] V. K. Bhatia, Analysing Genre: Technology Use in Professional Settings. New York, NY: Routledge, 2013.

[19] R. K. Yin, Case Study Research and Applications: Design and Methods, 6th ed., Thousand Oaks, CA: SAGE Publications, Inc., 2018.

[20] S. J. Tracy, Qualitative Research Methods: Collecting Evidence, Crafting Analysis, Communicating Impact, 2nd ed., Hoboken, NJ: Wiley-Blackwell, 2019.

[21] D. Silverman, Doing Qualitative Research, 5th ed., London, UK: SAGE Publications, Inc., 2017.

[22] W. Ellet, The Case Study Handbook: A Student’s Guide. Boston, MA: Harvard Business Review Press, 2018.

[23] M. Q. Patton, Qualitative Research & Evaluation Methods: Integrating Theory and Practice, 4th ed., Thousand Oaks, CA: SAGE Publications, Inc., 2015.

[24] N. K. Denzin and Y. S. Lincoln, The SAGE Handbook of Qualitative Research, 5th ed., Thousand Oaks, CA: SAGE Publications, Inc., 2018.

[25] V. V. D. Ham, Analyzing A Case Study: Pocket Study Skills, 1st ed., London, UK: Palgrave Macmillan, 2016.

[26] D. Silverman, Interpreting Qualitative Data, 5th ed., Oliver's Yard, London: SAGE Publications, Inc., 2015.

[27] M. B. Miles, A. M. Huberman, and J. Saldana, Qualitative Data Analysis: A Methods Sourcebook, 4th ed., Thousand Oaks, CA: SAGE Publications, Inc., 2019.

[28] S. K. Yalçın and M. Şengül, ''The role and function of the technology in the process of communication,'' Turkish Studies, vol. 2, pp. 749-769, 2007.

Distance learning: examples of good practice, analysis and experience

Maja Mačinko*, Ana Sović Kržić ** and Ivan Mudri***
* OŠ Ivana Cankara/Cankareva 10, Zagreb, Croatia
** University of Zagreb Faculty of Electrical Engineering and Computing, Unska 3, Zagreb, Croatia
***OŠ Ivana Cankara/Cankareva 10, Zagreb, Croatia
* maja,macinko@skole.hr
** ana.sovic.krzic@fer.hr
*** ivan.mudri@skole.hr

Abstract - Due to the COVID-19 pandemic, schools in Croatia are closed for students. Teachers had to develop some new ways of teaching/learning. In this paper, we present how subjects of STEM (Technical culture and Informatics) can be learned online and what do students think about it. Additionally, to the regular STEM classes, we will describe an online competition Robotic Simulation League and its goals.

Keywords - STEM; technology, information technology; robotics; distance learning; educational robots

I. INTRODUCTION

Distance learning has become a necessary type of teaching today. The possibilities it offers are endless. The learning process is placed in the hands of the students, and it is the teacher who creates the content tailored to them. The student determines the pace of work and is often very encouraging for students because of the possibility of repeating the process and additional sources of knowledge that they can search on the Internet. Feedback is extremely important, both from teachers and students. Further, a great advantage of this type of teaching is that students learn from each other.

Education needed to be adapted to the current situation in the world, so distance learning has emerged as the best learning option. The combination of synchronous learning where students had live instruction through tools such as Teams video calls in Office 365 [1] and Zoom [2] where the teacher explained while students followed and asked questions, and asynchronous learning where the teacher sent self-study materials resulted in a successful end of the school year.

CARNet and Microsoft provided great support to Croatian schools with their Teams [3] and Yammer [3] classrooms, which are free for educational institutions, and provided an online classroom that combined virtual face-to-face connections, files, tasks and conversations all on the same platform that students accessed via cell phones, tablets and computers according to recommendations of the Ministry of Science and Education [5]. Additionally, Robotic Simulation League for elementary school students is launched. The League is based on the Miranda software and also provided all participants with free use of the software. In this paper, we will show some of the examples of good practice that have been used in Croatian schools and what students think about this way of schooling.

II. DISTANCE LEARNING EXAMPLES

A. Technical culture

The current situation in the world has led us all to teach students online. Technical culture as a subject performed in cabinets and specialized classrooms equipped with tools has faced special challenges. How to make interest in practical work, make sure that they can complete tasks on their own without injuries at work, and think about the materials they have at home so that they do not have to go shopping.

The fifth graders made a useful geometric object in order to repeat the technical drawing and some basic knowledge (Fig. 1). The task was fun and creative for the students and they showed diversity and creativity.

Sixth graders had two practical assignments. The first task was to make a rocket and design a propellant for it, and also to record a video for the above. They were given the task to connect it with the topic of fuel energy. Another task they had was to build a lattice bridge [6] (Fig. 2). Spaghetti or skewer sticks were used as material. After completing the task, they had to test the load capacity of the bridge, photograph everything and send it to their virtual classroom.

Seventh grade students covered the topic of metal and its processing with the material. After learning all about metals, their task was to make a useful object out of metal. They made jewelry, a cube for the Man Don't Get Angry, and a Criss-Cross game (Fig. 3).



Figure 1. Geometric body made of plastic bottle.

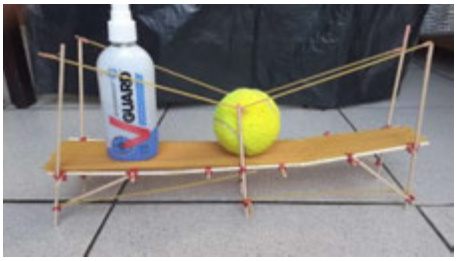


Figure 2. Bridge made of spaghetti and skewer sticks.



Figure 3. Recycling of metal – Criss-Cross game.

In addition to this type of work, with the topic of electrical installations in the house, their task was to monitor the electricity consumption in their house according to the given table and record the situation for seven days. In this way, they learned about the rational consumption of electricity.

Eighth graders repeated their knowledge of electronics and robotics and were finally given the practical task of making a car that can drive. Some used a balloon and air, others used water. It turned out that the students were willing to cooperate and could hardly wait for assignments to do something useful.

B. Informatics

Switching to the online classes was not a problem in Informatics class. Students got used to using a computer in class, and all they needed at home was a computer or tablet with Office 365 for schools package and internet. In order not to turn the teaching of Informatics into mere copying of the answers from the textbook, the students were given the tasks they had to complete in order to get feedback on their work. At the same time, this was a great opportunity to use the full range of applications that can easily be used for evaluation.

Second-grade students created characters, backgrounds, and programs in Scratch [7]. For each created character, they had to make an accompanying story e.g. why they had chosen that character. One of the stories goes like this (Fig. 4): “The picture shows a ruined city destroyed by a villain and fled, so the Hulk came. The picture shows destroyed buildings and roads." The task was a lot of fun for the students, and they were able to show their creative side in it and at the same time get the frustration out of themselves.



Figure 4. Scratch Program – Hulk.

Sixth graders were given the role of an editor at the publishing house, and the resulting text had to be edited, reviewed, and approved for publication. With this type of task students skills in MS Word were tested and at the same time their creativity was encouraged. Students were active seekers of information who revises and updates their knowledge through the process of gathering new information [8]. The students really liked such a role, but what fascinated them was the Kahoot quiz after each lesson [9].

In seventh grade Kahoot also proved to be a great tool, but while teaching HTML lesson, a video lesson was recorded for them, which was posted on the popular Youtube platform (Fig. 5). The combination of a new lesson, working on HTML and the Youtube platform proved to be effective, and the students expressed that it would be much easier and more fun for them to have similar lectures. Incorporating interactive video into a distance learning environment may not always result with improved learning, but this kind of tool may lead to better learning outcomes and higher learner achievement [10].

Eighth-grade students were given a project to create a website on Wix [11] (Fig. 6), and had to follow the rules of working on project tasks and teamwork. The students were competitive and expressed how this type of work also draws them from the monotony of other subjects.



Figure 5. Video lesson on Youtube platform.



Figure 6. Students website on Wix.com.

C. RoSiL – Robotics Simulation League

Robotics is an interdisciplinary research area between computer, mechanical, electronic and information engineering. Therefore, robots are ideal tools for teaching Informatics and Technical culture in primary schools. During the regular lessons, students build robots and learn the basics of mechanics, connect wires to sensors and actuators and learn the basics of electrical engineering and at the end they make programs for the robots learning basics of computer science and informatics. In distance learning most of those activities are not possible to perform: students do not have real educational robotics kits at home. In order to counteract this problem, Croatian robotic association [12] started an online robotics simulation league (RoSiL) for primary school students (fourth to eighth grade). Each student performs individually, from her / his home. They do not need robots, only a computer or a tablet with Internet connection. The simulation software that is used is Miranda [13] and it is translated to Croatian language. Students program in Scratch-like language. New written and video lectures and new assignments are given every week. Each lecture introduces new commands that should be used in the assignments. Students have the whole week to solve the assignments. Their last solution will be visible to judges. The judges will grade performance of the robot (efficacy, efficiency and duration) as well as a code that students wrote. 145 students participated this year as well as seven students from elementary school Ivana Cankara.

An example of the assignment is given in Fig. 7. The robot needs to reach Goal as accurately as possible. The robot is not perfect and has a random noise during the drive to make it more realistic and more difficult at the same time. Students need to include that in their solutions. At the beginning of the League assignments are easy, so even beginners can program the robot. Each week assignments are more difficult to keep students in League, to inspire them and help them to grow in programming.



Figure 7. Robotic Simulation League RoSiL uses Miranda Software [13].

D. Survey and analysis

The aim of the survey was to determine students' experiences in distance learning and to gain insight into the issues that students encounter. The data were collected using an anonymous online survey, which is created with Google Forms service. The survey was distributed through the education social network Yammer. The students of the elementary school Ivan Cankar Zagreb, from 5th to 8th grade, participated in the research, who were also the target group. A total of 181 students participated in this survey. Table I. shows the distribution of students by grade.

TABLE I. DISTRIBUTION OF STUDENTS BY GRADE.

Grade	Number of participants	Percentage
5th	45	24,86
6th	47	25,97
7th	47	25,97
8th	42	23,20
Total	181	100,00

In the survey, students assessed their own experience in distance learning. The first question was the assessment of dealing with distance learning, ranging from 1 (I can't deal with it) to 5 (I can completely deal with it). The arithmetic mean of the respondents on this question was 3.93, with a standard deviation of 0.90 and a mode of 4. This means that the students gave themselves a very good grade in managing with distance learning.

Furthermore, students were asked if they agreed with the following statements on a scale from 1 (strongly disagree) to 5 (strongly agree). Table II shows the results.

Results are showing that students prefer classical learning, they do not think that the tasks are more complicated than what they are used to. The vast majority do not think that they have difficulties in meeting the set deadlines and that their concentration at home is the same as in class.

Furthermore, we see that the vast majority of students are considered completely independent when working in distance learning, and at the same time that some teachers assign too many tasks. Most students think that teachers are well prepared for leading online lectures and that students spend too much time solving assignments.

Table III. shows the results of how much time students spend daily on assignments. In the table, we see that most students spend 3 to 6 hours solving assignments, which roughly corresponds to the daily schedule of a student spends in school.

TABLE II. STATEMENTS ABOUT DISTANCE LEARNING AND ANSWERS OF STUDENTS IN PERCENTAGE.

Statment	Percentage of scale				
	1	2	3	4	5
Distance learning is better than traditional teaching.	35,36	29,83	15,47	9,94	9,39
Assigments in online teaching are more complicated.	9,94	22,10	22,10	29,28	16,57
It's hard for me to meet deadlines.	24,86	29,28	17,13	16,02	12,71
I don't have the concentration to work at home.	34,81	23,76	15,47	18,23	7,73
I am completely independent in my work.	3,87	19,89	14,36	31,49	30,39
Some teachers send too many assignments.	10,50	9,94	14,92	26,52	38,12
Some teachers did't hadle online enviroment well.	24,31	18,23	24,86	17,13	15,47
I spend too much time on assignments.	14,92	14,92	17,13	24,86	28,18

TABLE III. TIME SPEND ON ASSIGNMENTS

Hours spend on assignments	Number of participants	Percentage
1-2	11	6,08
3-4	71	39,23
5-6	59	32,60
7-8	30	16,57
>8	10	5,52

In the last part of the survey, students were asked about the most common problems they encounter in online classes from 1 (never) to 5 (often). Figures 8, 9, 10, 11 and 12 show the distributions of the responses obtained with respect to a particular problem.

About 25% of students stated that they regularly, and 38% of them often encounter technical difficulties (Fig. 8). Only 3% stated that they had never encountered technical difficulties in distance learning.

36% of students said they never encountered a lack of technical support from teachers (Fig. 9). And about 6% said they regularly face a lack of technical support.

Fig. 10 shows students' responses to the frequency of self-organization problems. 16% of them stated that they never encounter problems in self-organization, and about 7% of them regularly encounter problems in self-organization.

12% of students said they had never encountered a problem with handing over assignments. And at the same time, 12% of them regularly have problems with the assignment deadlines (Fig. 11).

Finally, 25% of students had never encountered a problem with teacher's feedback (Fig. 12). And only 7% of them regularly encounter a problem around feedback.

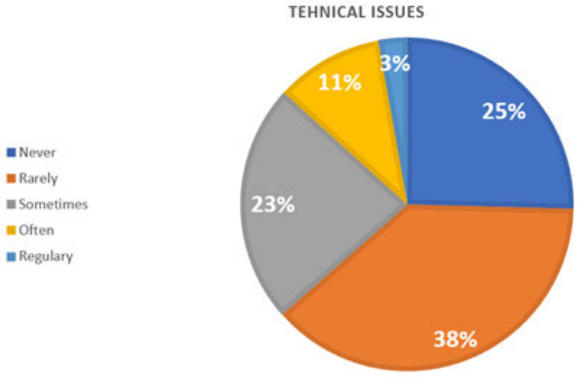


Figure 8. Problems related to the technical difficulties.

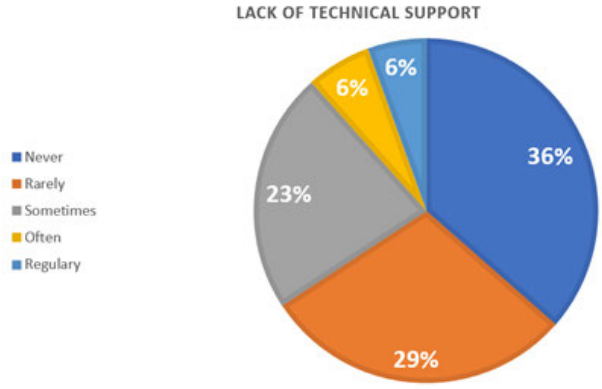


Figure 9. Problems related to the lack of technical support.

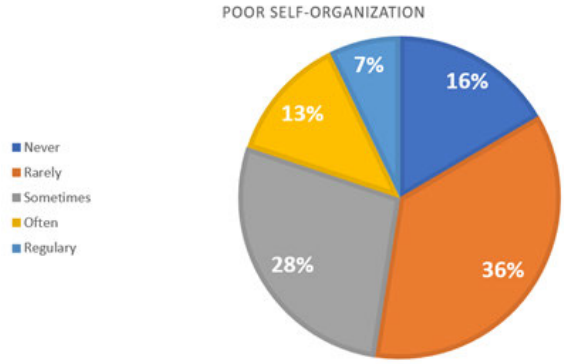


Figure 10. Problems related to the poor self-organisation.

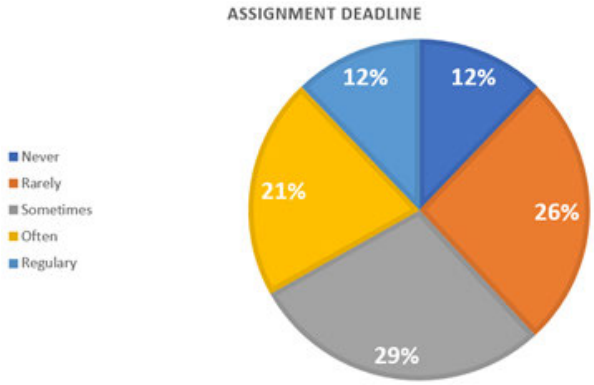


Figure 11. Problems related to assignment deadline.

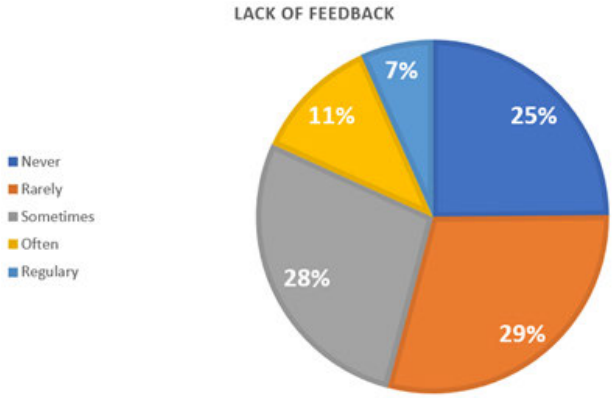


Figure 12. Problems related to the lack of feedback.

Problems which occurred during distance learning are easily solvable. First, we need to enable all students to have equal working conditions, which would solve problems with technical difficulties. With distance learning, students learn what it means to have a deadline, and at the same time learn to organize better. They said that in the beginning of distance learning everything was in chaos, and they were very desperate. But with time, they gained experience and they started to cope with the situation.

One of the problems is lack of feedback, which sometimes happens, but the majority of students stated that distance learning gives them more feedback on a personal level, which they usually do not get in classical learning.

III. CONCLUSION

In this paper we described the online lessons for two STEM subjects: Technical culture and Informatics. For both subjects, distance learning was a great opportunity for students to apply their theoretical knowledge into practical tasks. Results of their work showed that they like this kind of learning more than just learning theory. Although we do the same thing in classical classes, maybe we should orient the learning a little more on practical work and less on theoretical tasks.

One of the phenomena we have noticed is that students who are shy and distant from others are much better at presented form of learning.

The transition to distance learning was overnight, but the results of the survey show that distance learning is possible if we provide all students with equal (technical) working conditions, give them valid feedback and give them adequate support and guidance.

ACKNOWLEDGMENT

This work has been supported by Croatian Science Foundation under the project UIP-2017-05-5917 HRZZ-TRES and Croatian robotic association.

REFERENCES

[1] Office 365 za škole, Retrived from,https://office365.skole.hr
[2] Zoom video comunications, Retrieved from https://zoom.us/
[3] Teams - Online virtual Classroom, Retrieved from https://teams.microsoft.com, 2020.
[4] Yammer. Retrieved from https://www.yammer.com/o365.skole.hr/#/home 2020.
[5] Škola za život, "Preporuke o organizaciji radnog dana učenika tijekom održavanja nastave na daljinu", Retrieved from https://skolazazivot.hr/ , March 2020.
[6] The James Dyson Foundation North America. Challenge Cards. Retrieved from https://www.jamesdysonfoundation.com/resources/challenge-cards.html. 2020.
[7] Scratch, retrieved from https://scratch.mit.edu/about. 2020.
[8] Niederhauser, D.S. & Stoddart, "Teachers' instructional perspectives and use of technological software", Teaching and Teacher Education 17, 15-31, 2001.
[9] Kahoot, retrieved from https://kahoot.com/schools-u/. 2020.
[10] Vural, O.F., "The impact of a question-embedded video-based learning tool on e-learning", Educational Sciences: Theory and Practice, 13(2). 1315-1323. 2013.
[11] Wix.com, retrieved from https://www.wix.com/. 2020.
[12] Croatian robotic association. Robotics Simulation League. Retrieved from http://ar.hrobos.hr. 2020.
[13] Irai. Miranda - Universal software for STEM robots simulation. Retrieved from https://www.miranda.software. 2020.

Understanding the Factors that Influence Secondary School Teachers' Intention to Use e-Learning Technologies for Teaching After the COVID-19 Pandemic

S. Babić*, S. Križan Sučić**, G. Sinković**
* Polytechnic of Rijeka, Business Department, Croatia
snjezana.babic@veleri.hr
** Faculty of Informatics, Juraj Dobrila University of Pula, Croatia
sanja.krizan.sucic@student.unipu.hr
giorgio.sinkovic@unipu.hr

Abstract - The development and efficacy of use of e-learning in secondary schools after the COVID-19 pandemic depends on the teachers' intention to use e-learning technologies for teaching and should not be generalized since understanding of said intention depends on educational contexts. This study investigates the factors that predict Croatian secondary school teachers' intention to use e-learning technologies in their future classrooms after the COVID-19 pandemic. The results of our study show that attitudes and educational values, computer anxiety, self-efficacy, course characteristics, prior experience with e-learning technologies, education for use of ICT and pedagogical use of ICT are related to the perceived Croatian secondary school teachers' intention to use e-learning technologies in their teaching after the COVID-19 pandemic. Findings reported in this paper can be a foundation for future research in the field and help explain the barrier to acceptance of e-learning technologies, as well as a framework for the development of e-learning strategy in secondary schools after the COVID-19 pandemic.

Keywords – secondary school teachers, intention to use, e-learning technologies, online teaching, COVID-19 pandemic

I. INTRODUCTION

The education system is considered the foundation for the development of the digital society. The use of e-learning technologies in teaching by secondary school teachers has an important role in the development of the digital competencies of their students, which are required for the continuation of their studies and/or their entrance to the labor market. For those students that use Information and communications technology (ICT) more during their classes, the use of the same technology in their future work will be a natural continuation of this process. Therefore, the use of e-learning technologies in teaching in secondary schools, which before the COVID-19 pandemic was largely dependent on the acceptance of said technologies by each individual teacher, is extremely important.

The implementation of e-learning technologies in teaching requires certain e-learning competencies by the teachers. According to the model developed by Babić and Bubaš [1], the educational context is very important for the understanding of teachers' competencies for online teaching and its acceptance. A teacher can be motivated both intrinsically and extrinsically to implement e-learning technologies in teaching.

In secondary education system, the period during the COVID-19 pandemic is particularly interesting seeing as all teachers have had to use certain e-learning technologies in their teaching. This means that secondary school teachers, like all other teachers, have found themselves in a position where they have had to use e-learning technologies for teaching, regardless of the level of their e-learning competence.

With the assumption that once the COVID-19 pandemic ends, the teachers are going to be able to choose the ways in which to integrate e-learning technologies into their classes, the following research questions arise: (1) To what extent do secondary school teachers intend to use e-learning technologies in their teaching? (2) What factors influence the intention to use e-learning in teaching by secondary school teachers?

Since the acceptance of e-learning technologies by secondary school teachers depends on the narrower and broader educational environment, the understanding of factors that influence the acceptance of e-learning technologies is being discussed once again. On the other hand, according to the UTAUT model [2] the intention of future behavior related to technology directly influences the actual use of technology in an environment.

Therefore, the main research objective of this paper was to determine the influence of selected factors that influence the intention to use e-learning technologies by Croatian secondary school teachers in teaching after the COVID-19 pandemic.

The following section contains a short description of the educational environment during the COVID-19 pandemic, followed by factors influencing secondary school teachers' intention to use e-learning technologies for teaching before

the COVID-19 pandemic proposed in the relevant literature. A description of empirical research and findings is provided in section three. A short conclusion is given in the last section.

II. BACKGROUND TO THE RESEARCH

A. Educational environment during the COVID-19 pandemic

Some of the main factors that impact the acceptance of e-learning technologies for teaching by teachers are institutional factors, which according to the conceptual model developed by Babić and Bubaš [1] include: technical and social support for e-learning, ICT infrastructure, strategies and policies and social influence.

Before the COVID-19 pandemic, the use of e-learning technologies in secondary schools by the teachers was voluntary. However, following the closure of education institutions in many countries, including Croatia, due to the COVID-19 virus, the process of teaching and learning continued remotely [3].

This means that almost all countries have changed their strategies and policies as far as in-person classes at many education institutions are concerned. Teachers are being encouraged to use e-learning technologies for teaching. According to Bates [4], "e-learning is not synonymous with distance learning" and "e-learning can be a component of both classroom and distance learning".

Distance learning in Croatian secondary schools began on 16 March 2020. Since then, Croatian secondary school teachers have had different support channels and tools at their disposal to assist them in the implementation of e-learning technologies in teaching. For example, CARNet [5] has provided technical (free tools and services) and pedagogical support; Ministry of Science and Education [6] has provided guidelines and instructions for the preparation of distance learning; teaching and learning materials have been made available in digital format (for example, Agency for Vocational Education and Training and Adult Education (www.asoo.hr/en/) has made available materials for vocational secondary schools); and also, teachers have been able to use various different platforms for teaching: Moodle, Microsoft Office 365 (for example, Teams, Yammer), Google Classroom, Edmodo.

B. Factors influencing secondary school teachers' intention to use e-learning technology before the COVID -19 pandemic

Based on the theoretical foundation of the TAM [7], TPB [8], and UTAUT [2] model, Teo [9] proposed a model to explain the students' and teachers' intention to use technology for teaching. Teo [9] has found perceived usefulness, perceived ease of use, attitude toward use, facilitating conditions, and computer self-efficacy to be instrumental in explaining teachers' intention to use technology.

Eickelmann, and Vennemann [10] emphasize that teachers' beliefs of the pedagogical advantages of ICT differ between countries.

The research results by Ghavifekr et al. [11] confirm that performance expectancy, effort expectancy, social influence and education policy have significant effects on teachers' behavioral intention to use information and communication (ICT) tools in the classroom. They have also indicated the following factors as significant in using ICT tools by teachers: limited accessibility and network connection, limited technical support, lack of effective training, limited time and lack of teachers' competency. Furthermore, the results show that the use of ICT tools by female teachers in the classroom is lower compared to male teachers.

Syvänen et al. [12] have found that ICT competence, the concordance of the educational use of ICT with the teaching style, school support, and attitudes to the educational use of ICT are the key predictors of the technostress by teachers.

The results by Ifinedo et al. [13] have indicated that perceived technological knowledge, teachers' knowledge and perceived knowledge for integrating technology, teaching experience, and class size influence the technology integration into teaching.

Jung et al. [14] have confirmed that the supportive culture, teachers' self-efficacy, and knowledge have significant impact on teachers' technology integration and that the knowledge is the most significant variable.

Based on the theory of planned behavior (TPB), Saleh [15] has found that perceived behavioral control and attitude toward the technology are main predictors of teachers' intention to use e-learning technologies for teaching.

For example, in Croatian higher education institutions by teachers who are non-users of e-learning, Babić et al. [16] relying on the Conceptual Model of Competencies of Teachers in Hybrid Learning Environment of HEI [1], have found a positive correlation between intention to apply e-learning and self-efficacy, and a negative correlation with computer anxiety. Furthermore, attitude and educational values, characteristics of the university course and social influence are significant factors that impact the intention to use e-learning by non-users of e-learning.

III. RESEARCH

The aim of this study was to identify potential predictors of Croatian secondary school teachers' intention to use e-learning technologies for teaching after the COVID-19 pandemic.

A. Participants

One hundred and thirty eight secondary school teachers in Croatia participated in this research, among whom 119 (86.2%) were women and 19 (13.8%) men. Most teachers were under 40 years of age (81.9%).

It is important to emphasize that all participants in this study have had to use e-learning technologies in their teaching during the COVID-19 pandemic.

In addition, the majority of teachers work in vocational four-year secondary schools (53.6%), followed by teachers working in grammar schools (35.5%), teachers working in vocational three-year secondary schools (6.52%), while the

lowest percentage of teachers works in art high schools (2.17%) and others (2.17%).

Furthermore, the majority of secondary school teachers who participated in this study work in the field of natural sciences and mathematics (27.5%), followed by teachers of social sciences and humanities (21.7%), vocational subjects (20.29%), language and communication area (15.9%), information technology (10.1%), other (3.6%), while the smallest percentage of non - users work in the field of art (0.7%).

The analysis of the results shows that the sample of practicality of Croatian secondary school teachers mostly (34.9%) consisted of teachers with 4 to 10 years of teaching experience, while 26.1% of teachers have taught from 1 to 3 years, 20.3% of them have had 11 up to 20 years of teaching experience, 13.77% of them have had less than one year of teaching experience, and only 5.8% of them have had more than 21 years of teaching experience.

Less than half (49.3%) of the participants in this research have been using e-learning technologies for more than 10 years, 16.7% of them have been using technology between 6 and 10 years, between 2 and 3 years (12.3%), between 4 and 5 years (10.9%), less than two years (5.8%), while 5.1% of them have stated that they had not used e-learning technologies before the COVID-19 pandemic.

More than half (52.9%) of the participants in this study did not use e-learning technologies in their teaching before the COVID-19 pandemic. Furthermore, 41.3% of teachers did not acquire pedagogical knowledge to use e-learning technologies in teaching. The majority (66.7%) of teachers were educated for the technical functionality of e-learning technology and tools (see Table I.)

TABLE I. PERCENTAGE OF SECONDARY SCHOOL TEACHERS IN RELATION TO THE USE OF E-LEARNING TECHNOLOGIES IN THEIR TEACHING BEFORE COVID-19 PANDEMIC (N=138)

Before COVID-19 pandemic	**Percentage of participants in relation**	
	Yes (%)	**No (%)**
Used e-learning technologies in their teaching	47.1	52.9
Educated for pedagogical knowledge to use e-learning technologies in teaching	58.7	41.3
Educated for the technical functionality of e-learning technology and tools	66.7	33.3

The results shown in Table II suggest that the majority (44.1%) of teachers in this study have used cloud service Office 365 (Yammer, Teams and others tools), Google Classroom are used by 23.2% of teachers, other e-learning technologies (9.4%), Edmodo (5.1%), Merlin (1.5%), while the smallest percentage of teachers have used Zoom (0.7%) for their teaching during COVID-19 pandemic.

TABLE II. PERCENTAGE OF SECONDARY SCHOOL TEACHERS IN RELATION TO THE USE OF E-LEARNING TECHNOLOGIES IN THEIR TEACHING DURING COVID-19 PANDEMIC (N=138)

E-learning technologies	**Percentage of participants in relation to use (%)**
Google Classroom	23.2
Yammer	12.3
Office 365 (other tools)	18.8
Loomen	16.0
MS Teams	13.0
Zoom	0.7
Edmodo	5.1
Merlin	1.5
Others	9.4

B. Instruments and Procedure

Some factors selected for this study are an integral part of the Conceptual Model Dimensions of Teacher's Competence for E-Learning Development by Babić and Bubaš [1] and have been adapted for this study. The selected factors are [1]: ease of use of e-learning technology in teaching, attitude toward the use of e-learning technology for teaching, educational value of application of e-learning, computer anxiety, self-efficacy. Intention to use e-learning after the COVID-19 pandemic is a newly developed scale based on UTAUT model [2].

The research was conducted through an online questionnaire (Google Forms) and official Facebook groups by Croatian secondary school teachers in April and May 2020 during the COVID-19 pandemic. Participation was voluntary and anonymous.

The questionnaire contained items developed for the assessment scale(s), for which we used the Likert scale of 1-5 points (1 - completely untrue; 2 - mostly untrue; 3 - neither true nor untrue; 4 - mostly true; 5– completely true).

The results Cronbach's alpha coefficients (see Table III) show that all scales had a good internal consistency (from 0.83 to 0.93).

TABLE III. NUMBER OF ITEMS AND INTERNAL CONSISTENCY OF THE SCALE (CRONBACH ALPHA COEFFICIENTS; N=138)

Assessment scale	**Number of items**	**Cronbach alpha**
Ease of use of e-learning technology in teaching	5	**.93**
Attitude toward the use of e-learning technology for teaching	5	**.87**
Educational value of application of e-learning	11	**.93**
Computer anxiety	5	**.83**
Self-efficacy	4	**.86**
Intention to use e-learning technologies for teaching after COVID-19 pandemic	5	**.91**

C. Findings

1) Intention to use e-learning technologies after the COVID-19 pandemic by Croatian secondary school teachers

The Intention to use e-learning technologies by secondary school teachers after the COVID-19 pandemic scale was measured by a 1-5 point Likert scale for responses (1 – totally untrue; 2 – mostly untrue; 3 – neither true nor untrue; 4 – mostly true; 5– totally true).

The average score of Croatian secondary school teachers in this study regarding their responses to the items of the assessment scale *Intention to use e-learning technologies after the COVID-19 pandemic* is 18.93 (σ=5.4). This result demonstrates that the participants in this research indicate their intention to use e-learning technologies for teaching after the COVID-19 pandemic.

On average, the participants' responses to the items of the assessment scale *Intention to use e-learning technologies after COVID-19 pandemic* vary from 3.35 to 4.12. The arithmetic means (see Table IV) show that secondary school teachers intend to use a combination of e-learning and traditional classroom environment following the end of the COVID-19 pandemic (M=4.12), 11.5% of the participants do not agree with this approach and 28.3% remain undecided.

Furthermore, in the following school year after the COVID-19 pandemic, Croatian secondary school teachers intend to use e-learning for online communication and collaboration with students in at least one of their classes (M=3.88), as well as create new online content or adapt an existing one for at least one of their classes (M=3.82).

TABLE IV. ITEMS OF MEASUREMENT SCALES INTENTION TO USE E-LEARNING TECHNOLOGIES FOR TEACHING AFTER COVID-19 PANDEMIC BY CROATIAN SECONDARY SCHOOL TEACHERS (M-MEAN, σ-STANDARD DEVIATION, FREQ. AND %; N=138)

Intention to use e-learning technologies after COVID-19 pandemic	***Freq. (%)***					***M (σ)***
	1	***2***	***3***	***4***	***5***	
for teaching	9 (6.5)	7 (5.0)	39 (28.3)	36 (26.1)	47 (34.1)	**3.76** (1.17)
combined with traditional classroom environment	7 (5.0)	10 (7.3)	12 (8.7)	39 (28.3)	70 (50.7)	**4.12** (1.16)
to create online content	16 (11.6)	6 (4.3)	21 (15.2)	39 (28.3)	56 (40.6)	**3.82** (1.32)
for online communication and collaboration with their students	8 (5.8)	9 (6.5)	31 (22.5)	34 (24.6)	56 (40.6)	**3.88** (1.19)
for creating online tests for students	24 (17.4)	15 (10.9)	28 (20.3)	31 (22.4)	40 (29.0)	**3.35** (1.44)

Most of the participants (60.2%) agree almost entirely with the claim that indicates their intention to use e-learning technologies or an e-learning tool (e.g. Yammer and other tools Office 365, Edmodo, Google Classroom, Merlin, Loomen, etc.) in their teaching in the future. Only a little over half of the participants (51.4%) strongly agree with the claim that indicates that in the following school year after the COVID-19 pandemic, they will use e-learning for creating online tests for students in at least one of their classes, while 28.3% of the participants strongly disagree with this, and 20.3% remain undecided (see Table IV).

2) The relationship between intention to use e-learning technologies after the COVID-19 pandemic and selected variables

Statistical analysis was performed using t-tests for small independent samples and one way analysis of variance (ANOVA). Our results revealed no statistically significant correlation between teachers' intention to use e-learning technologies after the COVID-19 pandemic and gender, teachers' education for use of ICT, age, type of secondary school domain, the overall duration of use of e-learning technologies in teaching before the COVID-19 pandemic, type of e-learning technologies used during the COVID-19 pandemic, teaching experience.

However, the results of t-tests confirm that the intention to use e-learning technologies after the COVID-19 pandemic among the participants in this research that have been using e-learning technologies in teaching before the COVID-19 pandemic (M=20.54, SD=4.40) is significantly greater (t=3.44, df=136, p=0.001) than among those that haven't been using them (M=17.49, SD=5.81). Also, the intention to use e-learning technologies for teaching after the COVID-19 pandemic among the participants that have been educated for pedagogical use of ICT (M=19.95, SD =4.36) is significantly greater (t=2.72; df= 136; p=0.007) in relation to those that haven't been educated for pedagogical use of ICT in teaching (M=17.47, SD= 6.36).

A Pearson correlation coefficient was calculated to test relationships between mean values of scales and mean value of the scale *Intention to use e-learning technologies for teaching after COVID-19 pandemic* by Croatian secondary school teachers (see Table V).

As shown in Table V, strong positive correlations were found between the intention to use e-learning technologies after the COVID-19 pandemic by secondary school teachers, their attitude toward the use of e-learning technologies in teaching (r=.77, p<0.00), and their perception of educational value of application of e-learning in teaching (r=.77, p<0.00).

Furthermore, the results show (Table V) that the intention to use has a low negative correlation with computer anxiety (r=-.48, p<0.01) and has a low positive correlation with ease of use of e-learning technology in teaching (r=.43, p<0.00). Lastly, self-efficacy had a small positive correlation with intention to use e-learning technologies after COVID-19 pandemic by secondary school teachers (r=.34, p<0.00).

TABLE V. CORELATIONS BETWEEN MEASUREMENT SCALES AND THE SCALE INTENTION TO USE OF E-LEARNING TECHNOLOGIES FOR TEACHING AFTER COVID-19 PANDEMIC BY CROATIAN SECONDARY SCHOOL TEACHERS (N=138)

Assessment scale	***Users***
Ease of use of e-learning technology in teaching	.43*
Attitude toward the use of e-learning technologies for teaching	.77*
Educational value of application of e-learning	.77*
Computer anxiety	-.48*
Self-efficacy	.34*

Note: * $p < 0.01$.

These results (Table V) show that by developing a positive attitude, by better understanding of educational value of application of e-learning technologies in teaching, by decreasing apprehension of use of e-learning technologies in teaching, by increasing perception of ease of use of e-learning technologies in teaching, as well as by increasing self-efficacy in application of e-learning in teaching; the intention to use e-learning technologies in teaching by secondary school teachers in the future also increases.

3) Predictors of Intention to use of e-learning technologies in teaching after the COVID-19 pandemic by Croatian secondary school teachers

The results of multiple regression analysis showed that the educational value of application of e-learning (β=0.44, p=0.000), attitude toward the use of e-learning technologies for teaching (β=0.37, p=0.000), and computer anxiety (β=-0.13, p=0.034) were the significant predictors of Croatian secondary school teachers' intention to use e-learning technologies for teaching after the COVID-19 pandemic (Table VI). These predictors explained 68% of the variance of teachers' intention. The regression model obtained is statistically significant (F=55,56; p=.000), while the multiple correlation coefficient (R=0.82) suggests moderate correlation.

TABLE VI. RESULTS OF REGRESSION ANALYSIS ON MEASUREMENT SCALES IN RELATION TO CROATIAN SECONDARY SCHOOL TEACHERS' INTENTION TO USE E-LEARNING TECHNOLOGIES FOR TEACHING AFTER COVID-19 PANDEMIC

Variables	***Beta***	***t***	***Sig.***
Self-efficacy	-,01	-,09	.928
Ease of use of e-learning technology in teaching	-.02	-,21	.836
Educational value of application of e-learning	**.44****	5.42	**.000**
Attitude toward the use of e-learning technologies for teaching	**.37****	3.94	**.000**
Computer anxiety	**-.13***	-2.14	**.034**

Note: ** $p < 0.01$, * $p<0.05$

We can conclude that Croatian secondary school teachers who understand educational values of e-learning use, have a positive attitude toward e-learning technologies for teaching and have a lower level of computer anxiety; exhibit intention to use e-learning technologies after the COVID-19 pandemic to a greater extent.

IV. CONCLUSION

The findings of this study show that, on average, Croatian secondary school teachers have expressed their intention to use e-learning technologies after the COVID-19 pandemic.

Most of the participants (79%) have indicated that, after the COVID-19 pandemic, they intend to use e-learning combined with traditional classroom environment, 68.9% have the intention to create teaching and learning materials in digital format, while 65.2% have the intention to use e-learning technologies for communication and collaboration with students after the COVID-19 pandemic. Only a little more than half of the participants (51.4%) have indicated their intention to create online tests for their classes in the following school year after the COVID-19 pandemic. This offers the conclusion that secondary school teachers believe online tests are not easy to create and/or use in teaching, and that therefore, half of them has not indicated the intention to use online tests in their teaching after the COVID-19 pandemic.

It was determined that selected demographic characteristics (gender, teachers' education for use of ICT, age, type of secondary school domain, the overall duration of use of e-learning technologies in teaching before the COVID-19 pandemic, type of e-learning technologies used during the COVID-19 pandemic, teaching experience) did not have a direct influence on behavioral intention of using e-learning technologies for teaching by Croatian secondary school teachers. However, results of relationship tests confirm that, among the participants in this research, there is a statistically significant correlation between the intention to use e-learning technologies after the COVID-19 pandemic, the use of e-learning technologies in teaching before COVID-19, and teachers' education for pedagogical use of ICT. These findings are similar to those of prior studies [10], [11], [12], [13], [14], [16].

The results of the descriptive analysis show that more than half (52.9%) of the participants in this study did not use e-learning technologies in their teaching before the COVID-19 pandemic, 41.3% of secondary school teachers did not acquire pedagogical knowledge to use e-learning technologies in teaching. The majority (66.7%) of teachers were educated for the technical functionality of e-learning technology and tools. Therefore, almost half of the teachers did not have prior experience with e-learning technologies for teaching, which is an important factor in acceptance of e-learning in teaching (for example, [13], [15], [16]).

This study has found a positive correlation between intention to use of e-learning technologies after the COVID-19 pandemic and attitude toward the use of e-learning technologies for teaching, educational value of application of e-learning, ease of use of e-learning technology in teaching, self-efficacy, and a negative correlation with computer anxiety. Similar results have been obtained by [9], [14], [16].

The attitude toward the use of e-learning technologies for teaching, educational value of application of e-learning, and computer anxiety can explain 68% of variance of intention to use e-learning technologies after the COVID-19 pandemic by Croatian secondary school teachers.

Based on the results obtained in this research, it can be concluded that almost half of Croatian secondary school teachers have not had an adequate level of e-learning competence during the COVID-19 pandemic and that they have been dependent on the technical and pedagogical support both inside and outside of their primary institutions. Those teachers who have shown less computer anxiety, who have developed positive attitude toward the use of e-learning technologies for teaching, and who have understood the educational values of e-learning during the COVID-19 pandemic; have shown greater intention to use e-learning in their teaching in the future after the COVID-19 pandemic.

Our study's findings have implications for both researchers and practitioners. This research has shown that the adapted factors from the Conceptual Model Dimensions of Teacher's Competence for E-Learning Implementation [1], can be used to understand the factors influencing the intention to use e-learning by secondary school teachers, and the researchers can use all factors of this model in future research. Practitioners can use the results of this study for the purpose of developing strategies and policies related to organizational learning in the context of e-learning at secondary schools.

REFERENCES

[1] S. Babić, and G. Bubaš, "Assessment of competencies of online teachers: pilot study and survey evaluation." eLearning 2015-6th International Conference on e-Learning, pp. 164-171, 2015.

[2] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: toward a unified view", MIS Quarterly, vol. 27(3), pp. 425-47, 2003.

[3] School portal, Školska knjiga, "Operation of education systems in Europe", available at: https://www.skolskiportal.hr/nastava-na-daljinu/rad-obrazovnih-sustava-u-europi/ [accessed 12 June 2020].

[4] A. W. Bates, "Technology, e-learning and distance education", Psychology Press, 2005.

[5] Distance learning, CARNet, available at: www.carnet.hr/udaljenoucenje [accessed 12 June 2020].

[6] Ministry of Science and Education of the Republic of Croatia, "Guidelines regarding assessment and grading in a virtual environment", available at: https://mzo.gov.hr/UserDocsImages /dokumenti/Engleski/ 7-4-2020/Guidelines%20for%20a sse ssme nt%20and%20grading%20in%20a%20virtual%20environment.pdf [accessed 12 June 2020]

[7] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology", MIS Quarterly, vol. 13(3), pp. 319-340, 1989.

[8] I. Ajzen, "The theory of planned behavior. Organizational behavior and human decision processes, vol. 50(2), pp. 179-211, 1991.

[9] T. Teo, "Students and teachers' intention to use technology: Assessing their measurement equivalence and structural invariance. Journal of Educational Computing Research, 57(1), pp. 201-225, 2019.

[10] B. Eickelmann, and M. Vennemann, "Teachers 'attitudes and beliefs regarding ICT in teaching and learning in European countries", European Educational Research Journal, 16(6), pp. 733-761, 2017.

[11] S. Ghavifekr, T. Kunjappan, L. Ramasamy, and A. Anthony, "Teaching and Learning with ICT Tools: Issues and Challenges from Teachers' Perceptions", Malaysian Online Journal of Educational Technology, 4(2), pp. 38-57, 2016.

[12] A. Syvänen, J. P. Mäkiniemi, S. Syrjä, K. Heikkilä-Tammi, and J. Viteli, "When does the educational use of ICT become a source of technostress for Finnish teachers?", Seminar.net - International journal of media, technology and lifelong learning, 12 (2), pp. 95-109, 2016.

[13] E. Ifinedo, J. Rikala, and T. Hämäläinen, "Factors affecting Nigerian teacher educators' technology integration: Considering characteristics, knowledge constructs, ICT practices and beliefs", Computers & Education, 146, 103760, pp. 1-17, 2020.

[14] Y. J. Jung, K. Cho, and W. S. Shin, "Revisiting critical factors on teachers' technology integration: the differences between elementary and secondary teachers", Asia Pacific Journal of Education, 39(4), pp. 548-561, 2019.

[15] S. Salleh, "Examining the influence of teachers' beliefs towards technology integration in classroom", The International Journal of Information and Learning Technology, 33(1), pp.17-35, 2016.

[16] S. Babić, M. Čičin-Šain, and G. Bubaš, "A study of factors influencing higher education teachers' intention to use e-learning in hybrid environments", In 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 998-1003, IEEE, 2016.

Distance learning caused by the COVID-19 pandemic in Croatia: What do newspaper portals actually deliver to readers?

M. Kuzelj and K. Šamija
Faculty of Humanities and Social Sciences, Zagreb, Republic of Croatia
majakuzelj89@gmail.com

Abstract - **Distance learning has been introduced to all students in Croatia to ensure uninterrupted education delivery during the ongoing COVID-19 pandemic. Digital platforms, online applications, and communication tools bridge the gap to traditional education. The forced transition to virtual classrooms is a huge challenge for all parties involved and the way media reports can both, inform and calm, but also add fuel to the overall stressful times. This research paper aims to underline the power of mass media and the potential influence on public perception of distance learning, as well as the responsibility to contribute to a stable education system once the pandemic is over. The conducted quantitative content analysis of three newspaper portals arrived at the conclusion that the majority of newspaper portals reported about distance learning poorly and one-sided. A neutral and negative interpretive framework dominates mostly due to portals' unfamiliarity with the topic. None of the analyzed newspaper portals published complete information lacking the drive for a solution-oriented way of reporting. At the end of this research paper, recommendations are given on how newspaper portals might improve their way of reporting about educational issues during a crisis thus contribute to the development of a better informed public debate.**

Keywords – distance learning; COVID-19 pandemic; newspaper portals; public perception; content analysis

I. INTRODUCTION

In March 2020, the World Health Organization (WHO) declared the COVID-19 outbreak a pandemic [1]. Countries focused their epidemic measures on ensuring social distance in different areas and in education it led to the closure of educational institutions. So far 1.37 billion students and affected in 138 countries and around 60.2 million teachers are no longer in the classroom [2]. The first part of this research paper is dedicated to the transition to distance learning to explain the theory and research context.

Teaching and learning from a distance are fundamentally different from a regular classroom, so student expectations and satisfaction cannot be the same. The lack of human contact, technical or network related issues, stress caused by social isolation, or homework overload might lead to frustration. Therefore, it is crucial to communicate openly, raise issues, provide support to teachers, and encourage students and their families through mass media. The pandemic has again highlighted the role and responsibility of media during this crisis, which is in the focus of the second part of this paper.

However, it seems that newspaper portals in Croatia are more likely to poorly report about distance learning during the COVID-19 pandemic in a sensationalist way. Therefore this research paper aims to understand the way of reporting and what newspaper portals in Croatia are actually reporting about. In the third part of the paper key research findings from the conducted quantitative content analysis are presented. All online articles about distance learning from three main Croatian newspaper portals made the sample for this research.

II. DISTANCE LEARNING CAUSED BY THE COVID-19 PANDEMIC

Distance learning is identified as a contributor to the development of educational systems. It creates additional education opportunities, contributes to increasing the variety of education technologies, offers easy access, saves time, and shows a positive effect on saving costs. Therefore studies and research related to this way of learning have significantly increased and gained importance over the last years. However, to fully utilize its benefits, distance learning should be carried out in certain circumstances including planned time, structured space, dedicated support, the right motivation, participants' preparedness. Additionally, the benefits apply in normal times, but during a crisis the fast roll-out of distance learning might reveal challenges even beyond technologies used. Distance learning can be used for a wide range of reasons, while in the case of the COVID-19 pandemic the reason is the maintenance of the education system. The COVID-19 pandemic is an ongoing example of a fast transition to distance learning and both, benefits and shortcomings, will be explained in this part of the paper followed by the case of the Republic of Croatia.

A. Distance learning, its benefits, and shortcomings

Distance learning is "the process of creating and providing access to learning when the source of information and the learners are separated by time and distance, or both" [3]. There are two main methods of learning used in distance learning: synchronous and asynchronous learning [4]. Synchronous learning requires all participants to be present at the same time, while asynchronous refers to learning from defined material based on an individual schedule. Distance learning can be

a combination of both methods, while many technologies enable remote teaching and communication. Some of the technologies are print, television and radio, online content, E-mail, audio- and videoconference, open education platforms and applications, and digital communication apps.

Distance learning enables a flexible approach for teachers and students, while students can adjust learning technologies to their specific needs. Distance learning is generally not as effective as in-person classes, yet there is some evidence that learning from a distance based on online lessons shows a higher level of effectiveness in some cases and under controlled circumstances [5] because e-learning can be adjusted to students' needs, re-listened or accelerated compared to face-to-face classroom lessons. The level of effectiveness varies amongst age groups and students in lower grades need a structured learning environment, as well as organizational effort from their families at home [6]. If students do not perceive the technology as useful, they will struggle with distance education [7]. Also, students who usually struggle will likely struggle even more online, so distance learning requires a "strong curriculum and strong pedagogical practices" [6]. Other obstacles identified during the COVID-19 pandemic are lack of experience in distance learning technologies before the crisis among teachers, communication struggles between teachers and parents, too high expectations from parents in supporting children, digital inequity across national regions (access and devices), emotional issues in keeping up with the regular schedule while learning from home, and students overload [8].

B. The case of Republic of Croatia

The Ministry of Science and Education of the Republic of Croatia started the curricular reform of education in 2016,. The umbrella name of the project is "School for life" (Croatian "Škola za život") and had a spotlight on digitalizing schools appropriate for the age and the need of students, providing teachers with skills and tools to ensure teachers' digital independence, encouraging the use of digital content and multimedia, and working in a virtual environment [9]. Some of the enablers making the project possible were hiring of additional teachers, Internet access, free access to learning content, equipping classrooms with projectors and smartboards, digitalization of textbooks, providing every teacher with a laptop, creating digital content, digital competencies training for more than 50.000 teachers, tablets for students. In 2018 "Information Technology" has become a compulsory subject in the 5th and 6th grades of primary school. What had started as an initial pilot in 2015 on 150 schools, was planned to cover all schools national-wide in 2020. Unfortunately, the COVID-19 pandemic speeded up all activities to be deployed in just two weeks.

Distance learning is technically supported by the national research and education network of Croatia – CARNet. Every teacher is assigned a unique user name for free access to learning platforms. Also, guidelines, video instructions, and support are provided for teachers, students, and their parents. Universities are given full freedom, while primary and secondary schools are given the possibility of choosing among supported platforms: *Office356* (*Microsoft Teams, Yammer, OneNote Class*), *Google G Suite for Education* (*Google Classroom, Hangouts*), *Cisco Webex Meetings*, *CARNet* (*Loomen, Meduza*), *Adobe Connect*, and others (*Edutorij, e-Lektire, portal Nikola Tesla*) [10]. Once a platform has been selected on the school level, teachers can create virtual classrooms and subjects and invite participants. Teachers assign independently learning material to their classes (TV, audio, video, slides, or other) and all communication is carried out in virtual environments. The Ministry of Science and Education has also prepared a national agenda for each grade, video lessons, TV lessons for lower grades, helpful material, and other media for teachers to be used if suitable.

After a few weeks of distance education the Ministry of Science and Education has conducted a survey among more than 4.000 teachers and professors [11], about how they perceive their own as well as student's adoption to the new ways of teaching and learning. Based on the results 93 % of them are satisfied with the equipment and technology used in distance learning. At the same time 95 % of them are satisfied with their approach of delivering their presentation to students, while 89% believe their students are doing well. The results show a high level of satisfaction with new technologies as well as tools distance learning is based on.

III. MEDIA DURING THE COVID-19 PANDEMIC

While in the first phase of national-wide distance learning the main focus has been on technical solutions and providing digital content, in later stages ways of supporting teachers as well as students' families in isolation were questioned, the spotlight has shifted to emotional matters and the media has become crucial. Media reporting plays a key role in perception, management, and even the creation of a crisis [12]. The goal of media reporting should, therefore, be focused on calming the public [13], as well as warning [14], and informing according to crisis communication recommendations. In this part of the paper crisis communication will be further explained, followed by how media influences readers' perception.

A. Crisis communication

Mass media presents information in a way to attract as many readers and viewers as possible, often without correctness, which might have serious consequences especially in times of a crisis [15]. Therefore, over the last decades media reporting in times of a crisis has become very important and scholars of crisis management emphasize the importance of working with mass media in situations of a crisis [16]. Some rules pointed out are focusing on key messages, grouping negative news in one general message, openness for feedback, and a fact-driven approach without assumptions [17]. In order to mitigate emerging diseases, media reporting should be focused on how to influence people's behavioral changes, which has proven critical for limiting the spread of disease [18] or related stress and fear. At the same time scholars emphasize the importance of media influencing public

perception and beliefs during a crisis, but also shed light on the other side and raise concerns about press freedom in parallel with social relevancy [19].

B. How the media influences readers' perception

The central role of the media in forming public perception and opinion on diverse social events has been a research topic among many scholars. Especially in cases in which the audience does not possess direct experience, they become reliant upon the media to inform them. What people know, think, and believe about events beyond their personal first-hand experience is orchestrated by how these events are reported by the media [20]. Two theories explain how the media influences peoples' perception: the cultivation theory and the agenda-setting theory. Based on the cultivation theory the media has significant power to form public beliefs and is "primarily responsible for the perceptions of day-to-day norms and reality" [21]. On the other side, the agenda-setting theory explains the influence beyond what to think and forming opinions but telling people what to think about, based on topics covered by the media [22]. The media decides what to report about and how and "these stories, in whole or in part, are assimilated and accommodated into the emotional fabric and cognitive structures of individual readers and viewers" [20].

When it comes to covering educational issues in media, some scholars have provided a critical view of media reporting about education systems, especially in times of changes and reforms. A negative influence on peoples' perceptions due to focus on negative more than positive news was pointed out as one of the major concerns [23].

IV. CROATIAN NEWSPAPER PORTALS REPORTING ABOUT DISTANCE LEARNING DURING COVID-19 PANDEMIC

A. Research design and methodology

This research aims to understand the way newspaper portals in Croatia are reporting about distance learning in times of the COVID-19 pandemic and the study is built around the main hypotheses: (H1) Newspaper portals in Croatia are more likely to poorly report about distance learning during the COVID-19 pandemic, and (H2) Newspaper portals in Croatia are characterized by one-sidedness when reporting about distance learning during the COVID-19 pandemic.

The research activity carried out to achieve this aim is the quantitative content analysis of data collected from analyzing articles on Croatian newspaper portals published in a 2-month time frame starting with March 1st, 2020 until April 30th, 2020, as this article is being written. Content analysis is a research method for the objective, systematic and quantitative examination of communication content [24]. Conclusions drawn from ten categories will be presented.

The most relevant newspaper portals in Croatia based on [25] are *24sata.hr*, *jutarnji.hr*, and *vecernji.hr* and they are included in the research. Articles were selected based on search results in the news section using the keyword "distance learning" (Croatian "škola na daljinu") and in case they were connected to the research topic.

B. Key findings

Our analysis has identified a total of 169 articles about distance learning on three main newspaper portals. Most articles were published on *jutarnji.hr* (98), followed by *24sata.hr* (46) and *vecernji.hr* (25). The size of the research sample indicates the importance given to this topic in the newspaper portals media channel.

The majority of newspaper portals have reported about distance learning as a process with an accent on its long-term character, development, and change in time. Examples are articles about equipment behind distance learning, processes in other countries modeled on the Croatian solution, schedules (*vecernji.hr*), foreign examples, online thesis presentations (*24sata.hr*), explanation of how CARNet works, supporting apps, the announcement of distance learning manuals (*jutarnji.hr*). The second subject was about issues within and certain obstacles were discussed. Some examples of reported issues are students' open letters to the Minister, students in risk groups, unions about the shortcomings of distance learning, inadequate usage of online channels (*vecernji.hr*), how to ensure distance between students at school, salary decrease for teachers (*24sata.hr*), opposite points of view regarding the reopening of schools, students being overwhelmed, questions regarding how students will be evaluated during distance learning (*jutarnji.hr*). When it comes to the human aspect, reports mostly include testimonials and thoughts from teachers, parents, or students.

TABLE I. MAIN SUBJECT [A]

Newspaper portal	*Person / people*	*Process*	*Issue*	*Conflict*
jutarnji.hr	14 (14,29 %)	69 (70,41 %)	13 (13,26 %)	2 (2,04 %)
vecernji.hr	1 (4,00 %)	15 (60,00 %)	7 (28,00 %)	2 (8,00 %)
24sata.hr	2 (4,35 %)	26 (56,52 %)	17 (36,96 %)	1 (2,17 %)

A. Some of the subcategories included: mentioned person, actions, statements, development of distance learning, development stages, key problem in focus, obstacles, conflicting thoughts, conflicting statements.

Analysis shows that *jutarnji.hr* mostly reported about distance learning in terms of describing it as a new way of schooling (e.g. how final papers are defended online). However, none of the analyzed articles provides all information, rather only fragments. *24sata.hr* and *jutarnji.hr* tend to repeat already published articles, statements, and stories, especially regarding the topic of closure/reopening of schools. Among the three analyzed portals *jutarnji.hr* pays the most attention to how students cope with distance learning, while *24sata.hr* mostly reported about the equipment for distance learning.

TABLE II. TOPICS [B]

Newspaper portal	*jutarnji.hr*	*vecernji.hr*	*24sata.hr*
Closure/reopening of schools	30 (30,62 %)	5 (20,00 %)	26 (56,52 %)

Newspaper portal	*jutarnji.hr*	*vecernji.hr*	*24sata.hr*
Protection measures	1 (1,02 %)	1 (4,00 %)	na
Description of distance learning	41 (41,84 %)	3 (12,00 %)	5 (10,88 %)
Distance learning via TV	4 (4,08 %)	4 (16,00 %)	1 (2,17 %)
How students cope with distance learning	9 (9,18 %)	1 (4,00 %)	2 (4,35 %)
How teachers cope with distance learning	4 (4,08 %)	4 (16,00 %)	2 (4,35 %)
Journalist comment	1 (1,02%)	na	na
Equipment for distance learning	3 (3,06 %)	4 (16,00 %)	8 (17,40 %)
Advice for students and teachers	3 (3,06 %)	3 (12,00 %)	1 (2,17 %)
Foreign examples	2 (2,04 %)	na	1 (2,17 %)

B. Some of the subcategories included: governmental decisions, epidemiological measures, technical review, instructions, schedule of school hours, skills, grading, subjective opinion of journalists, needed equipment, lack of equipment, general advice, specific advice, comparison to other countries, positive examples, negative examples.

All three portals have demonstrated strong one-sidedness when it comes to the way topics were presented in the articles. One side of the story is presented and only one opinion is included without being challenged from an opposing side. Our analysis discovered that in most of the articles statements of the Minister of Science and Education Blaženka Divjak were presented, while at the same time no other opinions (parents, students, unions, etc.) were included. Some examples are presenting problems without evidence, pointing out how teachers were surveyed while students not, pointing out different results of surveys (*vecernji.hr*), cities asking for instructions, while the Minister responds in a neutral tone (24sata.hr), attack on statements and political decisions (*jutarnji.hr*).

TABLE III. TOPIC PRESENTATION[C]

Newspaper portal	*One-sided*	*Many-sided*	*Cannot be determined*
jutarnji.hr	65 (66,33 %)	11 (11,22 %)	22 (22, 45%)
vecernji.hr	10 (40,00 %)	2 (8,00 %)	13 (52,00 %)
24sata.hr	28 (60,86 %)	9 (19,57 %)	9 (19,57 %)

C. Some of the subcategories included: opinion mentioned, reporter comment, objective transmission of information.

Our analysis has confirmed that the majority of the headlines correspond to the related article texts. Regarding the article examples where the headlines do not correspond in most of the cases the headlines were made of statements taken out of context. It was mostly not specified who the statements belong to and the goal was to achieve sensationalism and attract readers.

TABLE IV. THE HEADLINE CORRESPONDS TO THE TEXT[D]

Newspaper portal	*Yes*	*No*
jutarnji.hr	97 (98,98 %)	1 (1,02 %)
vecernji.hr	17 (68,00 %)	8 (32,00 %)
24sata.hr	37 (80,43 %)	9 (19,57 %)

D. Some of the subcategories included: information from the title is presented in the text, sensationalism elements.

Statements are included in the majority of articles on all three newspaper portals.

TABLE V. STATEMENTS INCLUDED

Newspaper portal	*Yes*	*No*
jutarnji.hr	89 (90,82 %)	9 (9,18 %)
vecernji.hr	20 (80,00 %)	5 (20,00 %)
24sata.hr	41 (89,13 %)	5 (10,87 %)

The profile of the most often mentioned person on all three newspaper portals is a politician, often Minister Blaženka Divjak. Unlikely to the other two portals *jutarnji.hr* often included statements from the educational environment (students, teachers, principals, deans, rectors), as well as statements from the Croatian Institute of public health, psychologist, celebrities, and a director.

TABLE VI. PROFILE OF THE PERSON WHOSE STATEMENT IS INCLUDED

Newspaper portal	*jutarnji.hr*	*vecernji.hr*	*24sata.hr*
Politician	58 (59,18 %)	12 (48,00 %)	20 (43,48 %)
MoES / civil servant / public institution representative	22 (22,45 %)	4 (16,00 %)	16 (34,78 %)
Trade union / association representative	13 (13,27 %)	3 (12,00 %)	7 (15,22 %)
Parent	11 (11,22 %)	na	16 (34,78 %)
Croatian Institute of public health	6 (6,12 %)	na	na
NGO (e.g. “Hrabri telefon”), psychologist	11 (11,22 %)	na	na
CARNet	2 (2,04 %)	1 (4,00 %)	4 (8,70 %)
Student / teacher / dean / rector	36 (36,73 %)	3 (12,00 %)	14 (30,43 %)
Media representative	5 (5,10 %)	1 (4,00 %)	2 (4,35 %)
Celebrities	4 (4,08 %)	na	na
Director	1 (1,02 %)	na	na

In case statements are included in articles, the purpose is mostly to include expert opinions (mostly by Minister Blaženka Divjak) or comments made by others. It is noticeable that the solutions to the presented problem are almost not present.

TABLE VII. PURPOSE OF THE STATEMENT

Newspaper portal	*To inform*	*To comment*	*To provide experts opinions / prognosis*	*To present solutions*
jutarnji.hr	48 (48,98 %)	68 (69,39 %)	53 (54,08 %)	na
vecernji.hr	7 (28,00 %)	7 (28,00 %)	10 (40,00 %)	na
24sata.hr	16 (34,78 %)	43 (93,48 %)	16 (34,78 %)	4 (8,70 %)

In the majority of articles the interpretive framework is neutral, then followed by a negative approach.

TABLE VIII. INTERPRETIVE FRAMEWORK[E]

Newspaper portal	*Positive*	*Negative*	*Neutral*
jutarnji.hr	13 (13,26 %)	18 (18,37 %)	67 (68,37 %)
vecernji.hr	1 (4,00 %)	7 (28,00 %)	17 (68,00 %)
24sata.hr	2 (4,35 %)	15 (32,61 %)	29 (63,04 %)

E. Some of the subcategories included: advantages of distance learning, good examples, disadvantages of distance, bad examples, mere information.

The analyzed articles on all three portals are mostly not based on sensationalism (scandals, exaggeration, dramatization, one-sidedness, infotainment, and violation of human privacy). It is noticeable that *jutarnji.hr* has exposed the highest number of sensationalist articles. Sensationalism is not visible in the content of the analyzed articles, rather in the headlines. Intriguing statements, words that cause confusion and fear, clickbait titles are chosen to be present in headlines.

TABLE IX. SENSATIONALISM[F]

Newspaper portal	*Yes*	*No*
jutarnji.hr	38 (38,78 %)	60 (61,22 %)
vecernji.hr	10 (40,00 %)	15 (60,00 %)
24sata.hr	20 (43,48 %)	26 (56,52 %)

F. Some of the subcategories included: scandal, exaggeration, dramatization, infotainment, violation of human privacy.

Our analysis has shown that in the majority of articles no elements causing moral panic (deviance, social problems as a threat to generally accepted values, collective behavior, government reaction, or the initiation of social resistance) were detected. Nevertheless, *jutarnji.hr* has revealed to have published the largest number of articles with such elements. Some examples are articles about the stress of returning to school and risk of infections, overload of work for teachers (*vecernji.hr*), headlines about parents resisting to send their kids to school, headlines with intriguing statements (*24sata.hr*), powerless parents and teachers, and responsibility shifted from the Ministry to parents (*jutarnji.hr*).

TABLE X. ELEMENTS CAUSING MORAL PANIC[G]

Newspaper portal	*Yes*	*No*
jutarnji.hr	16 (16,33 %)	82 (83,67 %)
vecernji.hr	2 (8,00 %)	23 (92,00 %)
24sata.hr	11 (23,91 %)	35 (76,09 %)

G. Some of the subcategories included: deviance, social problems as a threat to generally accepted values, collective behavior, government reaction, initiation of social resistance.

V. CONCLUSION

It seems that the COVID-19 pandemic is occurring at a time in history characterized by high-speed internet, high level of digital literacy starting at a young age, and mass media spreading information in the speed of light, thus has to maintain teaching and learning activities. However, the way distance learning has been pushed to be executed overnight it might leave a bad memory in the future and be associated with lockdowns, new platforms, and technologies that didn't work, or parent's frustrations and stress. The pandemic is causing much more than just forced shift to distance learning. The influence goes beyond technology exposing deep inequities.

As a direct consequence of anti-pandemic measures we are currently experiencing a boom in distance learning, but what remains open is what will happen after the crisis. After the crisis the development of distance learning in Croatia could be taken further and use the opportunity of learnings gained from times of pandemic to establish itself as a strong alternative to learning as we knew it before. To make such long term development possible a well-informed public debate is crucial and media is the strongest enabler. A significant number of articles about distance learning have been published within a short time frame (March, 1st – April, 30th). In terms of the number of articles *jutarnji.hr* and *24sata.hr* especially stood out. Importance and media space has therefore clearly been given to the topic. What the media decides to report about and in what way significantly influences readers' perceptions and beliefs. Especially in times of crisis responsible crisis communication calls for subjectivity, well-informed reporting, and a multi-sided approach. Sensationalism, a negative interpretative framework, and elements causing moral panic should be avoided.

We began this article with two hypotheses regarding the way newspaper portals in Croatia are reporting about distance learning during the COVID-19 pandemic and we assumed it to be poor and one-sided. The results of this study confirm that distance learning on all three analyzed portals is mostly presented in a one-sided way, only one side of the stories was shown, opinions and statements were asked from one side, in the majority of cases politicians, therefore public debate was not encouraged.

Our research arrived at the general conclusion that the majority of newspaper portals wrote about distance learning processes and detected issues and faced obstacles that show that they find distance learning as such as quite new, unfamiliar, and unclear. In the majority of articles the interpretive framework is neutral, then followed by a negative approach. A neutral and negative interpretive framework dominates mostly due to newspaper portals' unfamiliarity with the topic. Another conclusion derived from the findings of this study is that none of the three analyzed newspaper portals published complete information in their articles. Articles were focused on segments, information delivery was partial and poor. Involving more parties would foster public debate and

could significantly contribute to problem-solving. At the same time articles have not shown solution-oriented approaches. When it comes to sensationalism, it has been identified mostly only in headlines.

Based on the theory review at the beginning of this article the goal of media reporting during a crisis should be focused on calming the public, warning, and informing. As a result of the learnings gained through the quantitative content analysis of 169 articles on three main newspaper portals in Croatia the following recommendations are proposed for future reporting about educational issues in times of a crisis: (1) focus on a well-informed way of reporting based on facts and knowledge about the topic, (2) provide readers with full information, not just segments, to achieve transparency, (3) seek for a multi-sided reporting approach including more parties to foster dialogue towards a problem-solving approach, and 4) clickbait headlines should completely be avoided in the context of this social topic during a crisis without exception. The recommendations might help in the development of a better informed public debate.

Early-stage researches and articles about distance learning during the COVID-19 outbreak can be found, as well as quantitative approaches to evaluate the effects of media in times of a crisis, yet the relationship between them remains a potential for future research. When it comes to the limitations of the study, this research paper presents insights gained through the analysis of one single case and within a certain time frame leaving space for deeper research after the pandemic. Additionally, further research of the public could help to understand the influence on the reader's perceptions. This paper, therefore, presents a fragment contribution to the overall goal to better understand the importance of responsible, objective, and multi-sided reporting in times of a crisis. Answering the question of how media can successfully intertwine its economic needs on one side and the need for clean objectivism on the other is far less understood. Therefore this is an area where insights and learnings gained from specific cases are much needed.

REFERENCES

[1] World Health Organization. (2020, March 11). *WHO Director-General's opening remarks at the media briefing on COVID-19.* [Online] Retrieved from https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (Accessed: 2020, April 25)

[2] UNESCO. (2020, March 24). *1.37 billion students now home as COVID-19 school closures expand, ministers scale up multimedia approaches to ensure learning continuity*. [Online] Retrieved from https://en.unesco.org/news/137-billion-students-now-home-covid-19-school-closures-expand-ministers-scale-multimedia (Accessed: 2020, April 23)

[3] Honeyman, M. and Miller, G. (1993). *Agriculture distance education: A valid alternative for higher education?* Proceedings of the 20th Annual National Agricultural Education Research Meeting. pp. 67-73.

[4] Bušelić, M. (2012). *Distance Learning – concepts and contributions*. Oeconomica Jadertina, pp. 23-43.

[5] Gutierrez, K. (2016, April 7). *Facts and Stats That Reveal The Power Of eLearning* [Online] Retrieved from https://www.shiftelearning.com/blog/bid/301248/15-facts-and-stats-that-reveal-the-power-of-elearning (Accessed: 2020, April 25)

[6] Loeb, S. (2020). *How Effective Is Online Learning?* Education Week. Vol. 39(28).

[7] Christensen, E. W., Anakwe, U. P. and Kessler E. H. (2001). *Receptivity to distance learning: the effect of technology, reputation, constraints, and learning preferences.* Journal of Research on Computing in Education. Vol. 33(3), pp. 263-279.

[8] Snelling, J and Fingal, D. (2020, March 16). 10 *strategies for online learning during a coronavirus outbreak.* [Online] Retrieved from https://www.iste.org/explore/learning-during-covid-19/10-strategies-online-learning-during-coronavirus-outbreak (Accessed: 2020, April 26)

[9] Ministry of Science and Education Republic of Croatia. (2020). *How we have introduced distance learning* [Online]. Retrieved from https://skolazazivot.hr/croatia-how-have-we-introduced-distance-learning/ (Accessed: 2020, April 26)

[10] CARNET. (2020). *Udaljeno učenje* [Online]. Retrieved from https://www.carnet.hr/usluga/udaljenoucenje/ (Accessed: 2020, April 26)

[11] N1info. (2020, April 6). *Gotovo 90% učitelja misli da se učenici dobro snalaze u nastavi na daljinu* [Online]. Retrieved from http://hr.n1info.com/Vijesti/a497943/Online-nastava-u-Hrvatskoj.html (Accessed: 2020, April 25)

[12] Ma, R. (2005). *Media, Crisis, and SARS: An Introduction.* Asian Journal of Communication, Vol. 15, pp. 241-246.

[13] Kuzelj, M., Dokman T. and Katalinić, J. (2018). *Kibernetički napadi i krizno komuniciranje - izvještavanje novinskih portala.* S. MIPRO Proceedings 2018. Opatija: Hrvatska udruga za informacijsku i komunikacijsku tehnologiju, elektroniku i mikroelektroniku – MIPRO, pp. 455-460.

[14] Quarantelli, E. L. (1991). *Lessons From Research: Findings on Mass Communications.* Newark : Disaster Research.

[15] Reissová, A., Žambochová, M. and Sláma, V. (2013). *The role of mass media in crisis communication.* Faculty of Economics and Administration, Scientific Papers of the University of Pardubice, Series D, Vol. 25, pp. 189-199.

[16] Williams, D. E. and Olaniran, B. A. (1998). *Expanding the crisis planning function: Introducing elements of risk communication to crisis communication practice.* Public Relations Review, pp. 387-400.

[17] Tafra-Vlahović, M. (2011). *Upravljanje krizom. Procjene, planovi, komunikacija.* Zaprešić: Visoka škola za poslovanje i upravljanje „Baltazar Adam Krčelić".

[18] Yan, Q. et al. (2016). *Media coverage and hospital notifications: Correlation analysis and optimal media impact duration to manage a pandemic.* Journal of Theoretical Biology. Vol. 390, pp. 1-13.

[19] Lee, A.Y.L. (2005). *Between Global and Local: The Glocalization of Online News Coverage on the Trans-regional Crisis of SARS.* Asian Journal of Communication. Vol. 15 (3).

[20] Fields, B.A. (2006). *School discipline coverage in Australian newspapers: impact on public perceptions, educational decisions and policy.* International Conference of the Australian Association for Research - AARE.

[21] Infante, D., Rancer, A. and Womack, D. (1997). *Building communication theory*. 3rd ed. Prospect Heigts: Waveland Press.

[22] Dearing, J. and Rogers, E. (1996). *Communication concepts 6: Agenda-setting.* Thousand Oaks, CA: Sage.

[23] Berliner, D. and Biddle, B. (1995). *The manufactured crisis*. Reading, MA: Addison Wesley.

[24] Berelson, B. (1952). *Content Analysis in Communication Research.* New York : Free Press.

[25] Digital News Report. (2020). *Croatia 2019* [Online]. Retrieved from http://www.digitalnewsreport.org/survey/2019/croatia-2019/ (Accessed: 2020, April 25)

The significance of transformational team members

Tihana Babić*, Zvonimir Musa*
*Algebra University College, Zagreb, Croatia
tihana.babic@algebra.hr, zvonimir.musa@racunarstvo.hr

Transformation is a new reality in the global economic system and thus every organization. Companies and consequently educational institutions face the challenges of constant knowledge sharing and continuous improvement of competences, skills, and expertise whilst trying to achieve performance levels that just a few decades ago would have been unimaginable. A need for learning while doing and accepting new knowledge from our team members has become indispensable. Only when a team has a shared belief in 'psychological safety', team members can create, increase efficiency, and make a transformational positive culture. Performing at the highest level seeks transformational positive team members who possess a transformational characteristic. Transformational characteristics provide their owners not only the ability to make data-driven decisions but to work through challenges in a way that they make others better in the process. Because of their qualities to increase morale, to promote cooperation and harmony, which directly and indirectly increases job performance, they not only do their job excellently, they help other team members do their jobs better.

Keywords – digital transformation, transformational leadership, transformational team, transformational team member, STEAM

I. INTRODUCTION

Today's economy, based on creating new services, new solutions, new inventions and fine-tune of product or service to the needs of customers and clients, is moved into the Conceptual Age, which demands to be ready to understand high concepts, coupled with high touch, as stated by Daniel Pink. High concepts thinking relate to identifying patterns and opportunities, appreciating artistic and emotional beauty, and combining what appear to be unrelated ideas into strong innovative actionable ideas. High touch deals with being empathetic, understanding the nuances of human interaction, experiencing the joy in others and one's self; and stretching beyond our everyday approach to doing things with a stronger focus on our purpose or meaning for life[1].

To be successful, specialists, scientists, engineers, and all others need to be practically creative and put their knowledge and experience into creating new things, services, or products continuously during their time and work. The majority of businesses, institutions, and academic teams need to be able to function 'outside of the box', implement new creative solutions, and very often create solutions and products that have never been conceived before.

This is not only common to everyday business, but also academic institutions, public service, and all other parts of society. Continuously changing the way we do business, the way we teach, constantly challenging the existing, and trying to think of a better way of doing things. High level of creativity and holistic approach to everything that is done – the relationships, the work, the passions[1] requires more than just adequate educational background. It requires a way of functioning on an everyday level that enables an individual to continuously produce new solutions to problems and situations that have never happened before, but moreover "to live a more abundant life, helping others do the same"[1].

Therefore, teamwork becomes our inevitability[2], and team members who can contribute to their teams through their work, influence, and impact, and create value not only in a form that they successfully deliver what their team is set to deliver, but can help other team members to bring more value and bring the whole team closer to its goal, are of inestimable importance.

II. TRANSFORMATION IN BUSINESS

A. Digital transformation as a business reality

The term 'transformation' became widely regarded in the business community as a term encompassing a necessity to survive in constantly changing circumstances of business reality. Application of digital technologies that impact all aspects of society and business often referred to as digital transformation (DX), can be defined as the remaking of every aspect in a way that everything is being rethought, simplified, and improved[3]. Companies, institutions, and other forms of entities, often need to make adjustments to their processes and organization of everyday work. These adjustments were usually regarded as reorganizations, which very often resulted in several organizational units being compacted into one or some bigger ones being divided into smaller ones. These changes provided the organization short-term agility, necessary to respond to a certain situation that came along.

Companies and businesses needed to drastically change a significant array of their internal and external processes, bring new products to the market and implement new business solutions that all together require such a significant change of current setup, which resulted in the fact that the newly setup company only resembled the pre-transformation one. With a completely reorganized structure, new processes and fundamentally changed the way the business is done. This meant that the company was reorganized, but in nature, it was transformed into a new

one. These kinds of endeavors required a high level of creativity from employees on all levels and especially the leadership structure. Very often, it is done with help from outside experts and implementation of solutions that were gathered from other similar situations.

"The New Reality" study, based on a series of 50+ interviews with leading experts from the charity, third sector, commercial, and digital technology worlds, predominantly aimed at established organizations[4] summed way to achieve digital transformation, and stand out as the main four aspects to digital transformation: *mindset, people, process* and *tools*[5].

B. Transformational leadership

Style of leadership in which the leader identifies the needed change, creates a vision to guide the change through inspiration, and executes the change with the commitment of the members of the group[6], and involves a process by which a leader appeals to followers' values, needs and aspirations to elevate their morality and motivation to achieve their fullest potential is the transformational form of leadership[7]. According to Burns, who first introduced the concept of transforming leadership, transforming leadership is a process in which "leaders and followers help each other to advance to a higher level of morale and motivation"[7]. In contrast, Burns suggested that the transactional form of leadership is based on an exchange relationship between leaders and their followers, such that each party's needs are met via workplace transactions (e.g., followers help to achieve the leader's objectives; leaders reward followers with a salary, bonuses, or other benefits)[8].

Bass extended the work of Burns, by helping to explain how transformational leadership can be measured and how it impacts follower motivation and performance. Also, he used the term "transformational" instead of "transforming"[9].

The transformational form of leadership is superior to the transactional form as it is positively related to followers' satisfaction with their leader overall job satisfaction, organizational commitment, work motivation, and supervisory ratings of leader effectiveness[10]. In its ideal form, it creates valuable and positive change in the followers with the end goal of developing followers into leaders[11].

Sosik and Jung assert inactive leadership is the most passive and ineffective form of leadership, transactional leadership is somewhat active and effective, and transformational leadership is the most active and effective form of leadership[12]. Howell and Avolio[13] noted that the difference between management by exception—active and management by exception—passive is in the timing of the leader's intervention:

- *active leaders* - monitor followers' behavior, anticipate problems, and take corrective actions before the behavior creates serious difficulties;
- *passive leaders* - wait until the behavior has created problems before taking;
- *leaders who score high on laissez-faire leadership* avoid making decisions, hesitate in taking action, and are absent when needed action, laissez-faire leadership is the avoidance or absence of leadership, actually nonleadership[14].

The Full Range Leadership Development Model concerning taken actions and achieved effects, according to Lindsay and Woycheshin, adapted from Bass and Avolio, is presented in Figure 1[8].

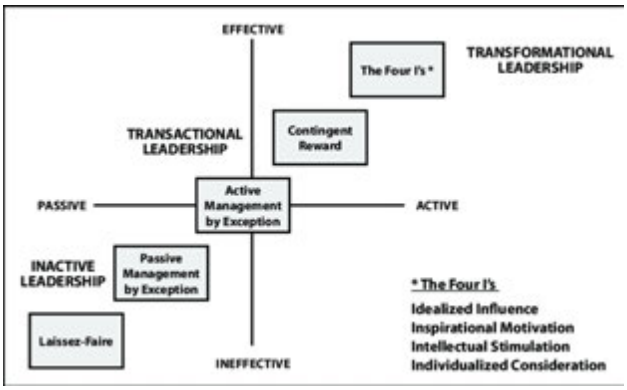


Figure 1. The Full Range Leadership Development Model concerning taken actions and achieved effects (source: Lindsay and Woycheshin, 2015 adapted from Bass and Avolio, 1994).

Transformational leaders (TL) are leaders who "transform" groups or organizations, focus on followers, motivating them to high levels of performance, and in the process, help followers develop their leadership potential. There are four elements of transformational leadership, also known as the 4 I's: *charisma or idealized influence (II), inspirational motivation (IM), individualized consideration (IC), intellectual stimulation (IS)* [9][15][16][17].

According to Sosik and Jung[18], who further elaborated those elements, while contingent reward may be effective in motivating followers to reach targeted goals, that approach is limited in achieving success in the innovation-based economy of today, because people need to have both extrinsic and intrinsic motivation to succeed. They stated five needed components (5 I's):

- *Idealized influences (behavior)* – TL often seek authenticity (as Mohandas Gandhi advised his followers to *'always aim for complete harmony of thought, word and deed'*), by displaying high levels of moral behavior, virtues, and character strengths, as well as strong work ethics.
- *Idealized influences (attributes)* – the idealized leader's role modeling of high levels of performance and ethics, adding to the idealized influence followers attribute to a TL. The result is followers strongly identification with a leader and their high levels of trust and commitment to the leader.
- *Inspirational motivation* – a leader behavior that involves developing and articulating a vision, an example of positive vision Dr. Martin Luther King's *"I Have a Dream"* speech. When articulating a clear vision of the future, with confidence in their team synergy and collective success, inspiring leaders increase followers' willingness to excel.
- *Intellectual stimulation* – while inspirational motivation triggers the followers' emotions,

intellectual stimulation increases the rationality and intellect of followers, who become more creative, innovative, and more willing to change for the better.

- *Individualized consideration* – TLs' ultimate aim is to develop followers into leaders themselves by spending time listening, coaching, and teaching for followers' development. When leaders treat followers as individuals with different needs, abilities, and aspirations, they are displaying individualized consideration.

The Full Range Leadership Development Model concerning leaders influence followers according to Stafford, adapted from Burns, 1998 and Sosik, Jung, 2010, is presented in Figure 2[19].

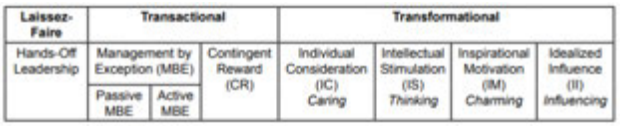


Figure 2. The Full Range Leadership Model according to leaders influences followers (source: Stafford, 2015 adapted from Burns, 1998 and Sosik, Jung, 2010).

C. Transformational characteristics

According to Bass[15], transformational leaders encompasses the following aspects:

- *raise awareness of moral standards;*
- *highlight important priorities;*
- *increase followers' need for achievement;*
- *foster higher moral maturity in followers;*
- *create an ethical climate (shared values, high ethical standards);*
- *encourage followers to look beyond self-interest to the common good;*
- *promote cooperation and harmony;*
- *use authentic, consistent means;*
- *use persuasive appeals based on reason;*
- *provide individual coaching and mentoring for followers;*
- *appeal to the ideals of followers;*
- *allow freedom of choice for followers*[21].

Transformational leadership is value-centered; leaders and followers share visions and values, mutual trust and respect, and unity in diversity[22]. Bass and Steidlmeier, while questioned the morality of transformational leadership in paper *Ethics, character, and authentic transformational leadership behavior*, stand out that "ethically, values may be imparted directly to followers by authorities whom they respect and trust, and from whom they want guidance: priests, physicians, parents, and teachers. If the values espoused are immoral, then the authorities are pseudo-transformational"[23].

Bass and Steidlmeier as evidently stated that to be truly transformational, leadership must be based on moral foundations. The four components of authentic transformational leadership (4 I's) stand in contrast with pseudo-transformational leadership based on:

(1) the moral character of the leaders and their concerns for self and others;
(2) the ethical values embedded in the leaders' vision, articulation, and program, which followers can embrace or reject; and
(3) the morality of the processes of social ethical choices and action in which the leaders and followers engage and collectively pursue[23].

D. The authentic, integrative and virtuous person as transformational

Bass and Steidlmeier, concluded that in leadership, as in life, character matters, and that from the literature on transformational leadership, it is clear that there are many points of congruence between the "authentic moral sage" and the "authentic transformational leader". Critics attribute manipulative, deceptive, and other such devious behaviors to so-called transformational leaders. But it is pseudo-transformational leaders who are deceptive and manipulative. They stated that self-aggrandizing, fantasizing, pseudo-transformational leaders can be branded as immoral, but authentic transformational leaders, as moral agents, expand the domain of effective freedom, the horizon of conscience, and the scope for altruistic intention. Their actions aim toward noble ends, legitimate means, and fair consequences.

"Leaders are authentically transformational when they increase awareness of what is right, good, important, and beautiful when they help to elevate followers' needs for achievement and self-actualization, when they foster in followers higher moral maturity, and when they move followers to go beyond their self-interests… Pseudotransformational leaders may also motivate and transform their followers, but…they will encourage 'we-they' competitiveness and the pursuit of the leaders' self-interest instead of the common good."[15]. Bass and Steidlmeier concluded that engaged as authentic transformational leaders are in the moral uplifting of their followers, in the sharing of mutually rewarding visions of success, and in enabling and empowering them to convert the visions into realities, they should be applauded, not chastised[23].

Integrity presents a mature person who *"behaves in accordance with his beliefs, does not say one thing and does another; publicly expresses his moral beliefs, even when those beliefs are not popular; takes care of others and their needs; speaks the truth and takes over responsibility for what he thinks, feels and does; does not act in front of others; is honest and has an inner sense of moral consistency; lives honestly and authentically and does not pretend to be something he is not"*[24]. As Vig highlighted, values such as *courage, prudence, optimism (faith), perseverance, humility,* and *modesty* make a person a complete or complete person, ie a person of integrity[25].

III. TRANSFORMATION IN TEAMS

A. Transformational organization culture

Before discussing the conditions needed for the team to be transformational, it is important to understand the transformational environment. Kerzner stated factors that

influence the transformational environment include: *work in a multinational setting and interfacing with different cultures; alignment to broader business objectives; an environment conducive to change; overcoming employees resistance to change; understanding (co)worker sensitivity to the change; winning (co)worker support for changes; assess to what helps the need to achieve the changes; making people believe that they are not being manipulated; making sure that people can cope with the changes; diffusing emotional feelings; face-to-face meetings to verify that people understand the need for the change; workshops when necessary; providing training for the changes; managing bad news effectively; managing unprecedented scrutiny by stakeholders; and managing scrutiny by the media*[26].

In an environment where *'change'* is an everyday reality and projects are tended to be performed in a transformational way, the skills needed can be significantly different than traditional project management skills.

To increase the overall business performance of a company, both financial and non-financial, its business should be based on building ethical organizational cultures[25]. Organizations with positive, virtuous ethical cultures enjoy bottom-line and top-line benefits, including *higher employee job satisfaction, increased legal compliance, and rule-following, increased organizational, commitment, increased cooperation, increased change management success, increased attraction of high potential talent, lower turnover, lower health care costs*, and *lower legal risk*[27].

B. Transformational teams

"Teamwork" means different things to different people[28], and some of the key components of highly successful teams are:

- *Work importance* – Pink[29] identified that the three main things that drive people in work are *autonomy*, *mastery*, and *purpose.* There is a deeper meaning behind the work that people do and it should go beyond profit and making money. It also should be communicated to the team members.
- *Clear personal values* - values are the things that matter the most in the way people live and work. Team members must know what their values are and how they connect to the larger story of their company.
- *Clear goals and defined role on the team* – goals can be short-term or long-term, but each team member should be clear what their role is on the team or the project.
- *Trust* - is a foundation element for highly effective teams. Every positive interaction helps in creating that foundation of trust.
- *Confidence to share a new idea or take a risk* – team members mustn't embarrass, reject, or punish other team members. The concept of *'psychological safety'*[30] is the belief that the team environment is a safe place to share ideas and take risks. This requires practice and courage and has an immensely positive impact on teams when achieved.

As contemporary companies have a constant need to transform themselves and the way they operate as a whole; similar is required from specific organizational units and teams. Not only leaders, but all internal participants need to be able to adjust and upgrade the way they operate, the way they produce results, and eventually change the product or service they provide. These changes require all team members to implement new practices, acquire new knowledge, and develop new skills. One of the consequences of this situation are phenomena like lifelong learning and need to acquire new knowledge every day, increase the span of expertise, improve competences, and sharpen present skills.

This situation is even more tangible in teams where a high level of creativity and expertise is needed for adequate performance. To put creativity into production, team members have to, alongside their educational background, be able to receive new knowledge, challenge the existing methods of working and put in practice the newly acquired cognitions in the most efficient way. Therefore, learning and the ability to learn is becoming more and more crucial for the everyday performance of many individuals. The dynamics and speed of changes that happen during our new working lives are of such intensity, that usual educational methods are not proving to be sufficient, which brings up the importance of on the job training and learning by doing.

C. Transformation to an adequate team member in practice

As a consequence of all mentioned, the importance of team members who can critically assess the performance and practices of other team members, define new knowledge that needs to be acquired by them, and communicate effectively those elements to their colleagues, has become a foundation of modern teams. Those team members who can help their colleagues to acquire new knowledge and the new practices and effectively implement them in the everyday process can be defined as '*transformational positive team members'*.

Transformational positive team members are those team members who have mastered the skills and knowledge specific for the operation of the team they work in. Also, they maintain a high level of their performance and, above average, communication skills in combination with emotional intelligence level, usually found in teachers and trainers. As the complexity of processes and operations performed in a team rises, the significance of the above-mentioned elements increases. For example, this highly comes into effect with teams operating in research and development areas, medical services, IT, or mission-critical engineering like space exploration and other areas.

One of the crucial issues that all organizations, and especially the above-mentioned ones, face is the induction of new team members into the existing teams and processes. This issue is even more tangible in situations where new team members are new graduates coming out of the education system and entering the professional environment. The induction period, consisting of activities that enable to do their work in a new professional role, represents a significant investment for the organization because the costs of new employees' salaries and working infrastructure are immediately present from the first day of

their engagement, while their full productivity is something expected to happen in the future. The challenges can be even deeper if the newly induced team members cause damage due to their inexperience and lack of on the job training. Even if the new team members are experienced professionals, in a new professional environment they are still not fully acquainted with accustomed processes and practices that prevail in the existing team and organization. Therefore, shortening the induction period is of great importance for all teams and organizations where the cost off mistakes can be extremely high or even measured in human lives.

Military teams are an example of training and constant on-the-job-learning teams, especially those who operate on complex technical systems or special tactics. The complexity of these systems and the potential, unwanted damage they can cause to human lives and infrastructure, makes them extremely sensitive to human error. Thus, every member, before joining those teams, undergoes an intensive training program followed by constant practices and drills. These activities are imposed to ensure that every team member is highly acquainted with all technology systems and correct procedures for doing the job the right way. A crucial element of training programs are team members who can see and recognize unwanted behaviors, lack of technical knowledge, or even a personal attitude that has consequences on performance of the individual and maybe the team in general. The transformational team member has a valuable influence to transform their current contribution to the team from insufficient to one that can bring value to the team, or just prevent the damage.

Also, in medical teams, a positive and transformational role of senior team members, who can help new colleagues in gaining extensive practical knowledge span in the shortest period, can make a difference between life and death. For example, a senior surgeon inducting a group of residents surgeons can, by fulfilling his transformational role, shorten the time of the induction, transfer a vast array of experience and make the new doctors ready to perform more complex surgeries, with a higher level of expertise, and subsequentially, avoid usual beginners mistakes in a larger scale.

D. Transformational positive team members

Contemporary transformational leadership theories and practices are concentrated mainly towards team leaders and exploring their abilities to lead their teams and make them able to achieve more and to reach higher levels of potential that they have originally conceived. In light of the immense changes in the everyday workplace on all levels, transformational leadership as such is not always enough to bridge the gap between current and required level of performance of each team member individually, and at the team altogether. Especially in types of teams mentioned earlier where lives can be in danger.

Therefore, transformational team members are necessary to transfer the knowledge and improve not only professional skills but also the "life skills" of team members, without assuming a team leadership position. By creating a team environment in which a team leader is leading a team through sharing knowledge and improving capabilities of his team members and at the same time, having other team members improving the performance and skill levels of their colleagues, especially as moral agents uplifting the entire team, the team can develop more quickly and in a shorter period can reach its full operational potential, and even surpass it in near future.

This is not limited only to practical performance improvement but also on elevating personal values and moral standards. Without morality, people are doomed to the consequences of the misuse of freedom, but also of knowing the consequences of the misuse of freedom, responsibility and ultimately even the knowledge and skills they possess, which Dostoevsky portrayed through the words of his hero: *"You are guilty in front of others because you could have been the only lighthouse in the darkness, and you were not"*[31]. The importance of high personal values and moral standards of team members is essential for establishing a positive transformational culture, where trust and understanding among team members are enabling the knowledge transfer and its implementation.

IV. TRANSFORMATION IN EDUCATION

Educational systems face 'new reality' challenges, as they represent a link between students and the labor market, preparing them to be full shareholders of modern business. Thus, the knowledge that is being transferred to students is changing constantly in order to keep up with contemporary trends in areas for which the students are being prepared. More and more the old fashion ex-cathedra lectures are being replaced with practical case studies and assignments, which require active students' participation and very often, within a team.

Not only that students acquire a more adequate level of expertise, but at the same time they practice the teamwork, which they will encounter in real-life situations once they embark the professional life. The majority of expertise that is being taught within the current education system, especially for IT, design, engineering, and business, learning methods that are being used, like case studies and practical hands-on learning in teams, is giving new significance to transformational team members.

Under the influence of digital transformation, students benefit from STEAM skills (Science, Technology, Engineering, Art, and Math content areas) if learning includes problem-solving, critical thinking, creativity, as well as the most important features of teamwork; communication, collaboration, and conflict management. With the rise of the STEAM teams, the practical implementation and development of transformational team members have taken one of the most important roles in team learning and acquiring knowledge of complexed materia, positive personal values, and high moral standards.

It is of crucial importance not to ignore, as Žitinski pointed out, that "*Education is a Moral Concept*" because it is above instrumental values. The specific goal of education is the expansion of human knowledge and cognition of what is 'right', and therefore correct. "As a rule, A MAN WITH THE RIGHT EDUCATION BECOMES GOOD[32].

V. CONCLUSION

Although leaders are primarily expected to be role models, congruence is 'conditio sine qua non' and this requirement to '*always aim for complete harmony of thought, word and deed'* stands in front of each team member. Thus, not only leaders should have transformational characteristics, but every team member should strive to achieve, develop, perfect them. In real life team situations, team members very often have a huge impact on other team members. That impact can be positive or negative, it is rarely indifferent or status quo.

Adjustments that support a transformational environment should not include only readiness for change, and the meaning of education is not only to provide a solution for long-lasting and structured learning.

It is of particular importance to strengthening students in becoming transformational positive team members, as morally aware members of society, who aim actions to noble directions of the future.

REFERENCES

[1] D. H. Pink, „A Whole New Mind; Moving from the Information Age to the Conceptual Age", Riverhead Books, New York, 2005.

[2] G. Tudor, V. Srića, „Menadžer i pobjednički tim", M.E.P. Consult d.o.o., Zagreb, 1996.

[3] T. Hunt, "How to Make Digital Transformation a Business Reality?", 2018. Available at https://medium.com/@techhunt2195/how-to-make-digital-transformation-a-business-reality-b841c75e18b3, [accessed: 07.05.2020].

[4] J. Dood, "The New Reality". Available at https://thenewreality.info/about-new-reality, [accessed: 07.05.2020].

[5] J. Dood, "Poster: 20 Ways to achieve Digital Transformation", The New Reality. Available at https://thenewreality.info/poster, [accessed: 07.05.2020].

[6] Business Dictionary, "Transformatonal Leadership", 2020. Available at http://www.businessdictionary.com/definition/transformational-leadership.html, [accessed: 07.05.2020].

[7] J. M. Burns, „Leadership", Harper and Row, New York, 1978.

[8] D. R. Lindsay, D. Woycheshin, "Overcoming Leadership Challenges: International perspectives, Canadian Defence Academy Press, Kingston, Ontario, 2015.

[9] B. M. Bass, „Leadership and Performance", Free Press, New York, 1985.

[10] T. A. Judge, J. E. Bono, "Five-Factor Model of Personality and Transformational Leadership", Journal of Applied Psychology, Vol. 85, No. 5, 751-765, 2000. http://www.timothy-judge.com/Judge%20&%20Bono%20JAP%202000.pdf, [accessed: 07.05.2020].

[11] N. Ch. Roberts, "Transforming Leadership: A Process of Collective Action", Reserchgate, 2018. DOI: 10.1177/001872678503801103. Available at https://www.researchgate.net/publication/248150557_Transforming_Leadership_A_Process_of_Collective_Actiol, [accessed: 07.05.2020].

[12] J. J. Sosik, D. I. Jung, "Full Range Leadership Development: Pathways for People, Profit, and Planet", Psychology Press, New York, 2010.

[13] J. M. Howell, B. J. Avolio, "Transformational leadership, transactional leadership, locus of control, and support for innovation: Key predictors of consolidated business-unit performance", Journal of Applied Psychology, 78, 891–902, 1993.

[14] T. A. Judge, R. F. Piccolo, "Transformational and Transactional Leadership: A Meta_Analytic Test of Their Relative Validity, Journal of Applied Psychology, Vol. 89, No. 5, 755-768, 2004. http://www.panglossinc.com/TA-TF%20Paper--JAP%20published.pdf, [accessed: 07.05.2020].

[15] B. M. Bass, „Transformational leadership: Industrial, military, and educational impact", Mahwah, Lawrence Erlbaum Associates, New York, 1998.

[16] B. M. Bass, B. J. Avolio „Transformational leadership response to critiques", in M.M. Chemers, R. Ayman (Ed.), Leadership theory and research: Perspectives and directions, pp. 49-80, Free Press, New York, 1998.

[17] R. E. Riggio, "Are You a Transformational Leader?", Psychology Today, 2009. Available at https://www.psychologytoday.com/intl/blog/cutting-edge-leadership/200903/are-you-transformational-leader, [accessed: 07.05.2020].

[18] J. J. Sosik, D. I. Jung, "Full-Range Leadership Development: Pathways for people, profit, and planet", Second edition, Routledge, New York, 2018.

[19] Stafford "Full-Range Leadership", Holm Center. Available at https://www3.nd.edu/~jthomp19/AS300/1_Fall%20Semester/27%20Sep/Full_Range_Leadership_V1.pdf, [accessed: 08.05.2020].

[20] B. M. Bass, "The ethics of transformational leadership" in J.B. Ciulla (Ed.), Ethics, the heart of leadership, pp. 169-192, CT: Praegar, Westport, 1998.

[21] C. E. Johnson, "Chapter 7: Normative Leadership Theories" in Meeting the Ethical Challenges of Leadership, SAGE Publications. Available at https://www.sagepub.com/sites/default/files/upm-binaries/59330_Chapter_7.pdf, [accessed: 07.05.2020].

[22] G. W. Fairholm, „Values leadership: Towards a new philosophy of leadership", Praeger, New York, 1991.

[23] B. M. Bass, P. Steidlmeier, "Ethics, character and authentic transformational leadership", Leadership Quarterly, 10(2), p.p. 181-217, 1999.

[24] D. Miljković, M. Rijavec (2006), "Tko su dobir ljudi", IEP, Zagreb.

[25] S. Vig, "Poslovna etika", Codupo d.o.o., Zagreb, 2019.

[26] H. Kerzner, Project management 2.0", Leveraging tools, distributed collaboration, and metrics for project success, International Institute for Learning, Inc., New York, 2015.

[27] S. D. Olson, "Shaping an Ethical Workplace Culture", SHRM Foundation's Effective Practice Guidelines Series, Society for Human Resource Management, Alexandria, 2013. Available at https://www.shrm.org/hr-today/trends-and-forecasting/special-reports-and-expert-views/Documents/Ethical-Workplace-Culture.pdf, [accessed: 17.05.2020].

[28] J. Banner, "Is Teamwork a Part of Your Team's Work?", Team Effectiveness, Flash Point Leadership Consulting, 2016. Available at https://www.flashpointleadership.com/blog/is-teamwork-a-part-of-your-teams-work, [accessed: 07.05.2020].

[29] D. Pink, "The surprising Truth About What Motivates Us". Available at https://www.youtube.com/watch?v=KgGhSOAtAyQ, [accessed: 07.05.2020].

[30] A. Edmondson, "Building a psychological safe workplace", TEDx Talks, 2014. Available at https://www.youtube.com/watch?v=LhoLuui9gX8, [accessed: 07.05.2020].

[31] M. Mihajlov, "Domovina je sloboda", Radio B92, Beograd, 1994.

[32] M. Žitinski, "Education is a Moral Concept", Naše more 53(3-4), 140-147, 2006.

Digital transformation as a process of using digital technologies for monitoring and designing the user experience

Ivan Radoš* and Tihana Babić*
*Algebra University College, Zagreb, Croatia
ivan.rados@racunarstvo.hr, tihana.babic@algebra.hr

Abstract - **Digitization made possible businesses to monitor consumer habits to adjust their business activities to increase sales and competitiveness of their company. The focus of this research is the influence of age and gender on the transfer of personal vehicle ownership in the Republic of Croatia in 2019 and the representation of brands in these transfers. The first part of the paper discusses the theoretical aspects of digital business transformation and the possibility of using digital technologies in monitoring consumer behavior and improving customer experience. The central part of the paper presents the results of the research. The results show that the transfer of personal vehicles is mostly made by men and by age group 30 to 39 years. In total, the most represented brand in the transfer is Volkswagen, both for men and women. In the concluding part of the paper, the key research results are summarized, which companies dealing with motor vehicle transfer can use to improve the digital transformation of their companies by designing improved customer experience of their users.**

Keywords – ***consumer behavior, consumption, digitization, property transfer, personal vehicles***

I. INTRODUCTION

In the digital age, business processes and customer experience should be reimagined. Customers and how organizations engage with their customers are in the center of any successful sales, marketing, and market requirements. Recording a business on physical document papers is no longer sustainable. With mass computer use, digitalization established simpler and more efficient ways of working. But mostly, it wasn't about changing the ways how to do business or create new ones, but to offer faster and easier the same customer services as before. As digital technology started to evolve, the new ways of doing things and business became possible. With this potential of new technologies started also a reshaping of companies' approach to customers. Customers' service through advertising, marketing, and sales offerings began to adapt to customers' preferences. These actions from possibilities became a necessity, and the digital transformation was made possible.

II. DIGITAL TRANSFORMATION

Process of using digital technologies to create new — or modify existing — business processes, culture, and customer experiences to meet changing business and market requirements are called digital transformation[1].

In the process of digital transformation, technology enables users to search for what they want, whenever and however they want. The development of new technologies, software, and applications has enabled organizations to connect with customers at various levels - from purchasing information, decision-making assistance, and the purchasing process to the exchange of satisfaction information. Therefore, digital transformation also means a change in all forms of business and processes in which employees of the organization interact with its customers[2].

The digital transformation process made companies take a step back and revise everything they do, from internal systems to customer interaction, both online and in person. The old model was to wait for customers to find and come to you, whether in person or by calling. But in the contemporary digital environment, progressive companies embrace the advantages of new technologies. Digital transformation in action is to extend service offerings of the company by meeting customers on their platforms of choice[1].

A. Digital transformation in EU

As the European Commission highlights digital transformation in every organization is individual, so the development path of each organization is different. Significantly, the digital transformation of business and society has a huge growth potential, which is recognized by the European Commission as a potential for the whole of Europe.

Therefore, the European Commission's comprehensive response to the complexity of the digital transformation process is the Digital Europe program, a central element of the 2021-2027 Multiannual Financial Framework (MFF) proposal. The aim is to provide a spending instrument tailored to the operational requirements for digital capacity building[2].

The five priority areas within the Digital Europe Program for 2021-2027 are:

1. *supercomputing*
2. *artificial intelligence*
3. *cybersecurity and trust*
4. *advanced digital skills*
5. *ensuring the widespread use of digital technologies across the economy and society*[3].

B. *Digital transformation in the Republic of Croatia*

The Digital Economy and Society Index (DESI) is a composite index that summarises relevant indicators on Europe's digital performance and tracks the progress of EU Member States in digital competitiveness. The five dimensions of the DESI are:

1. *Connectivity – fixed broadband, mobile broadband, fast and ultrafast broadband and prices;*
2. *Human capital – internet user skills and advanced skills;*
3. *Use of the Internet – citizens' use of internet services and online transactions;*
4. *Integration of digital technology – business digitization and e-commerce;*
5. *Digital public services – e-Government and e-health*[4].

Digital Economy and Society Index (DESI), ranking for 2019, set up Croatia in 21st place out of 29 EU member states as a country with low scores on the index, as shown in Chart 1[4].

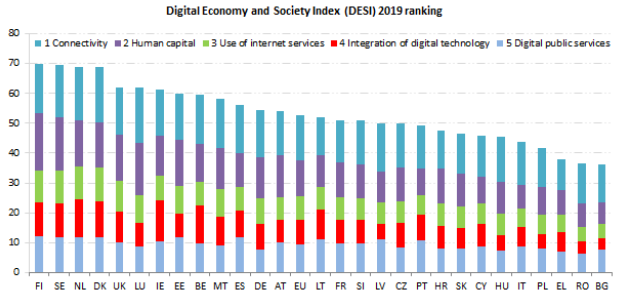


Chart 1. Digital Economy and Society Index (DESI) 2019, source: European Commission.

The same analysis shows, as stated by the Representative Office of the European Commission in Croatia[2], that Croatia has made progress in the categories of internet use and digital public services. Croatians are among the largest readers of online news in the EU, and Croatian companies are taking advantage of social media, Big Data technology, and e-commerce. However, despite some progress, Croatia has generally not made significant improvements in the area of internet connectivity compared to 2018 and has remained at the bottom of the scale, compared to the other Member States, and every fifth citizen of Croatia still do not use the Internet at all. In the category of digital transformation, ie integration of digital technology within companies, Croatia ranks 18th among EU member states, and in the category of digitalization of public services Croatia ranks 22nd among 28 EU member states, with an average lower than the EU average.

C. *Digitization of economic sectors*

Integration of Digital Technology, as represented in DESI 2019, regarding the digitization of economic sectors is progressing at a different pace, according to their own specific needs and starting points[4]. For example, in the trade of motor vehicles and motorcycles sector around 25 % percent of the enterprises have a high or very high Digital Intensity Index (DII).

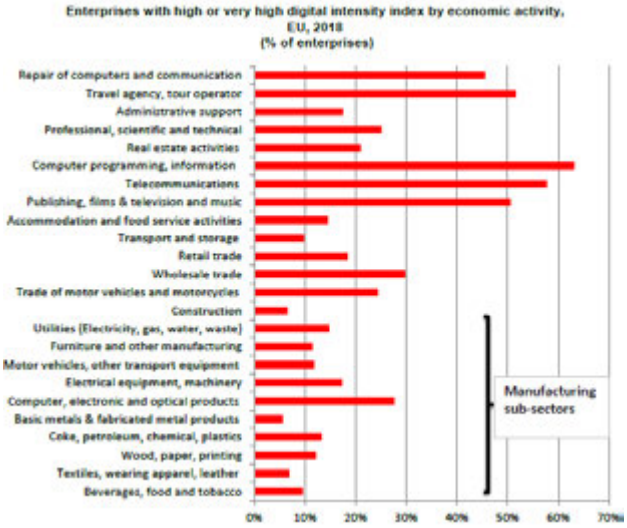


Chart 2. Enterprises with high and very high Digital Intensity Index by economic activity, EU, 2018 (% of enterprises), source: European Commission.

III. CONSUMER BEHAVIOR

The consumer behavior study examines how emotions, attitudes, and preferences of consumers affect buying behavior. Elements of consumer behavior are the subject of study in many scientific disciplines such as psychology, sociology, anthropology, marketing, and especially behavioral economics. Research has shown that purchasing decisions and consumption activities are not so easy to predict.

Customer relationship management (CRM) databases are an asset that helps to manage the company's interaction with current and potential customers. CRM technology provides companies to collect and manage large customer data and then carry out strategies based on that information. Data collected through CRM programs can help companies to solve problems from initial targeting, generating more reliable sales forecasts, increasing customer support effectiveness, increasing sales, improving customer retention to design effective customer service programs[5].

"Customers have always had a relationship with brands, but sophisticated tools for analyzing customer data are finally allowing marketing organizations to personalize and manage those relationships"[6]. Once the customer's relationships with the brand are profiled, advanced strategic goals can be created based on customers' expectations. If a brand meets those expectations, customers will respond by becoming passionate about it, but any failure to meet the expectations will be interpreted as a violation of the relationship and the relationship will be undermined[6].

A. *Consumer habits of individuals when buying personal motor vehicles*

Looking through the context of personal car ownership, the transfer of ownership between two stakeholders can be

driven by different needs. Most of the participants who find themselves in the ownership transfer process are satisfying to own needs to buy or sell the vehicle. Sometimes this need is triggered by the desire to give someone their vehicle as a gift. When it comes to personal car ownership, a common reason for property transfer is the age of the vehicle. The reason for such a decision is in fact that older the vehicle, the easier it is to transfer as its property value falls. For new vehicles that are still under warranty, this transfer is less frequent.

Considering that new vehicles always come with a higher price than used cars, it is expected that the owner will want to keep ownership until he can return investment. It can be concluded that persons who will buy used cars have lower expectations from the used vehicle then it would have from new, due to the fact vehicle will most likely have an outdated warranty and its no longer serviced in the official car repair shop. All things consider used cars shift ownership more often then used cars. Fact that new cars have lower servicing costs is also affecting the need for ownership transfer.

B. *The Automobile Consumer Studies*

Some previous research shows that demographic factors, such as age or generational affiliation and gender can have a great impact on consumer's behavior in buying vehicles for personal use.

In 2015, Deloitte fielded a survey in 19 countries to explore consumers'mobility choices and transportation decisions. In total, more than 23,000 individuals representing a broad range of cross-generational— Baby Boomers, Generation X, and Generation Y—automotive consumers responded to the survey. According to generational views, individuals today have many transportation options and increase their transportation decisions are differing across Generation Y. It is interesting to notice that, for example, 64 % of Generation Y consumers love their car, but they are 3 times more likely to abandon their vehicles if cost increase than other generations[7].

The worldwide research of the influence of gender on travel and transportation systems still show that women are somewhat less likely than men to hold drivers' licenses and to own cars, but differences between men and women in vehicle ownership and miles driven are far greater among older age cohorts and are diminishing over time as younger cohorts mature and become the majority of the population. But despite the lessening of those differences over time, some differences remain very persistent over time. For example, women continue to work closer to home than men, are more likely to use public transit for work trips than men[8].

IV. THE RESEARCH METHODOLOGY

A. *The Research Purpose /the Research Goals*

The general objective of this research was to examine the impact of individual demographic variables gender and age and variable brand on the ownership transfer of personal vehicles, which reflect the pattern of customer behavior on the purchase of personal car ownership, in terms:

- ***The personal vehicle*** belongs to a category whose maximum permissible mass does not exceed 3,500 kg and which are designed and constructed for the carriage of not more than 8 passengers, not counting the driver's seat. Personal vehicles may be combined with a trailer with a maximum authorized mass not exceeding 750 kg.
- ***Ownership transfer*** is a process between two individuals which is conducted by the station for technical inspection within the registration area according to the residence or the owner. After the process is done, the new owner of the vehicle can fully legally control the vehicle as is own property. In a current context, ownership transfer is viewed as a used car selling record.

Main research questions are defined as follows:

1. *Does gender affect consumer behavior regarding vehicle transfers?*
2. *Does age affect consumer behavior regarding vehicle transfers?*
3. *Does the brand affect consumer behavior regarding vehicle transfers?*

B. *The Research Sample and Method*

Data were collected from vehicle technical inspection stations in the Republic of Croatia in 2019, transferred to the databases of the Ministry of the Interior of the Republic of Croatia (abbreviated in Croatian as MUP). For data processing and cleaning (wrong entries, incomplete entries) software solutions were used based on the tools developed in the agency PROMOCIJA plus d.o.o. (Marketing Analysis and Statistics of Croatian Car Market).

For this study, data has been grouped by gender, age, the brand of vehicle, and ownership transfer frequency. Processing by gender, age, and vehicle brands was done through the C# and MSSQL.

The total number of samples was 271.108 transfers, which included 181.842 males (67%) and 89.266 females (33%). The average age of males is 43.45 and for females 41.91.

C. *The Research Results*

The results of the research are presented according to the research questions, as follows:

1. *Does the gender of the vehicle affect the distribution of ownership transfers?*

Data based on 271.108 personal vehicle ownership transfers in 2019 in the Republic of Croatia show that more than two-thirds of all vehicle ownership transfers, 181.842 of them, were made by men, while less than one-third 89.266 were made by women. The distribution of ownership transfers of personal motor vehicles per gender in 2019 in Croatia in percentages is presented in Chart 3.

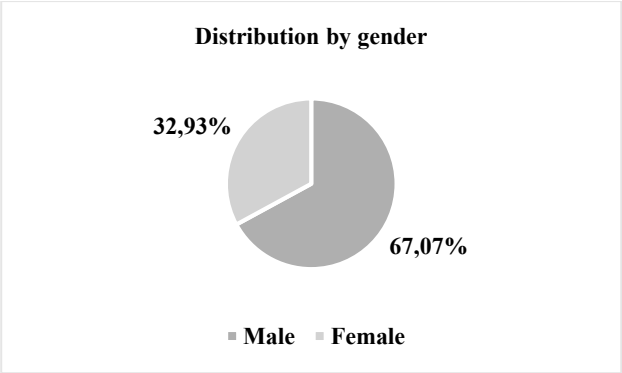


Chart 3. Distribution of ownership transfers of personal motor vehicles per gender in 2019 in Croatia; *N = 271.108.*

2. *Does age affect consumer behavior regarding vehicle transfers?*

Records of participants, for this research, are divided by age groups: 18-29 years old, 30-39 years old, 40-49 years old, 50-59 years old, 60 years, and more (60+). Data shows that:

- in the age group of 18 to 29 years, twice as many men than a woman made the transfer of personal vehicle ownership;
- in the age group of 30 to 39 years, that difference in transfer of personal vehicles ownership between men and women is somewhat smaller, as well as
- in the age group of 40 to 49, while
- in the age group of 50 to 59 years difference in transfer of personal vehicle ownership between men and women increases.
- In the age group of 60+ in the biggest difference the transfer of ownership made by men and women can be noticed; almost three-quarters personal vehicle ownership transfers are made by men.

The distribution of ownership transfers of personal motor vehicles per age group in 2019 in Croatia in percentages is presented in Table 1.

Age group	Male	Female
18-29	66,81 %	33,19 %
30-39	63,63 %	36,37 %
40-49	64,90 %	35,10 %
50-59	69,13 %	30,87 %
60+	73,15 %	26,85 %

Table 1. Distribution of ownership transfers of personal motor vehicles per gender and year in 2019 in Croatia; *N = 271.108*

3. *Does the brand affect consumer behavior regarding vehicle transfers?*

a. *The relation between gender and vehicle brand*

In the Republic of Croatia, during 2019, in a total number of all personal vehicle ownership transfer, the most represented brands of personal motor vehicles were Volkswagen, Opel, Renault, Fiat, Peugeot, Mercedes, and Citroen. In total, the top 3 the most represented brands during the transfer for males were Volkswagen, Opel, Renault, while the least represented brand was Citroen. Among women, in total, the top 3 the most represented brands in the transfer of personal motor vehicles in 2019 were Volkswagen, Renault, and Opel, while the smallest representation of the brand is Mercedes.

Gender distribution by the 8 top brand preferences in Croatia during 2019 is presented in Table 2.

Brand	Male	Female
Volkswagen	21,42 %	8,90 %
Opel	10,66 %	5,62 %
Renault	10,61 %	6,08 %
Fiat	6,35 %	3,61 %
Peugeot	6,31 %	4,06 %
Mercedes	6,12 %	2,52 %
Citroen	4,87 %	2,89 %

Table 2. Gender distribution per brand preferences in 2019 in Croatia, *N = 271.108*

b. *Relation between age and vehicle brand preference*

Regarding the relation between age and brand of personal motor vehicles in the transfer of ownership preferences, total data has shown that Volkswagen is the most represented brand for all age groups.

In the age group of 18 to 29 years, the second most represented brand is Opel and the third is Renault, as for age group 60+. In the age groups of 30 to 39, 40 to 49, and 50 to 59, the second most represented brand is Renault and third is Opel.

The selection of the top three most common brands by the age group in ownership transfers of personal motor vehicles in 2019 in Croatia is shown in Table 3.

Age	Brand	Vehicle ownership transfers
18-29	Volkswagen	54,35 %
	Opel	23,31 %
	Renault	22,34 %
30-39	Volkswagen	48,06 %
	Renault	26,89 %
	Opel	25,05 %
40-49	Volkswagen	46,39 %
	Renault	28,57 %
	Opel	25,05 %
50-59	Volkswagen	46,76 %
	Renault	26,99 %
	Opel	26,25 %
60+	Volkswagen	45,11 %
	Opel	28,29 %
	Renault	26,60 %

Table 3. Gender distribution per brand preferences, *N = 271.108*

c. The relation between age, gender and vehicle brand

Regarding the relation of all three variables *age, gender,* and *vehicle brand*, data has shown that:

- in the age group of 18 to 29 years in the Republic of Croatia, during 2019, the transfer of personal vehicle ownership was more often made by men (40.174), while women do it in twice as rare (19.962). Within the male group, the least represented a brand of the 8 included in the selection is Citroen. Within the female group, the least represented a brand of the 8 included in the selection is BMW.
- In the age group of 30 to 39 years, the transfer of personal vehicle ownership is more often made by men (40.013), while women do it in a significantly smaller number, almost twice as rare (22.872). Within the male group, the least represented a brand of the 8 included in the selection is Citroen, as for the previous age group. Within the female group, the least represented a brand of the 8 included in the selection is BMW, the same as for the previous age group od women.
- In the age group of 40 to 49, the transfer of personal vehicle ownership is more often made by men (37.258), while women make it (20.149). Within the male group, the least represented a brand of the 8 included in the selection is BMW. Within the female group, it is the same, the least represented a brand of the 8 included in the selection is also BMW, as well as for the previous age group, same as for the previous two age groups of women.
- In the age group of 50 to 59 years, the transfer of personal vehicle ownership is more often made by men (32.581), while women do it more than twice less (14.546). Within the male group, the least represented a brand of the 8 included in the selection is also BMW, the same as within the female group.
- In the age group of 60+, the transfer of personal vehicle ownership is more often made by men (33.565), while women make up barely a third of the total number (11.584). Within the male group, the least represented a brand of the 8 included in the selection is BMW, the same as within the female group.

In summary, as the results have shown, the most preferenced brand from the 8 most presented brands in personal vehicle transfer ownership in Croatia during 2019, both for men and women, for all age groups, is brand Volkswagen. The least represented brand is for men, ages 18 to 39, is Citroen, and ages from 40 to 60+ are BMW, while for women for all age groups, the least represented brand during personal vehicle ownership transfer is also BMW.

The selection of the top eight most common brand by the age group in ownership transfers of personal motor vehicles in 2019 in Croatia is presented in Table 4.

Age	Brand	Male	Female
18-29	Volkswagen	32,73 %	27,76 %
	Bmw	13,69 %	5,76 %
	Opel	12,35 %	15,35 %
	Renault	11,23 %	15,93 %
	Audi	10,99 %	6,29 %
	Peugeot	6,91 %	11,95 %
	Fiat	7,91 %	10,20 %
	Citroen	4,17 %	6,76 %
30-39	Volkswagen	29,08 %	25,34 %
	Bmw	9,81 %	6,65 %
	Opel	13,93 %	15,36 %
	Renault	14,61 %	17,12 %
	Audi	8,90 %	7,13 %
	Peugeot	8,68 %	10,80 %
	Fiat	8,17 %	9,35 %
	Citroen	6,81 %	8,23 %
40-49	Volkswagen	28,55 %	24,16 %
	Bmw	7,52 %	5,95 %
	Opel	14,48 %	14,77 %
	Renault	16,03 %	17,75 %
	Audi	7,99 %	6,79 %
	Peugeot	9,36 %	11,81 %
	Fiat	8,40 %	10,11 %
	Citroen	7,66 %	8,67 %
50-59	Volkswagen	29,84 %	24,40 %
	Bmw	5,42 %	5,09 %
	Opel	15,30 %	16,98 %
	Renault	15,67 %	17,61 %
	Audi	7,03 %	5,22 %
	Peugeot	9,59 %	11,25 %
	Fiat	9,68 %	10,49 %
	Citroen	7,47 %	8,96 %
60+	Volkswagen	26,80 %	22,34 %
	Bmw	3,31 %	3,90 %
	Opel	19,08 %	18,55 %
	Renault	17,15 %	18,37 %
	Audi	5,11 %	4,76 %
	Peugeot	9,74 %	11,57 %
	Fiat	10,40 %	11,68 %
	Citroen	8,42 %	8,84 %

Table 4. Gender distribution per age and brand preferences, *N = 271.108*

V. CONCLUSION

During the research, several valuable insights were gained. Research has shown that gender is a parameter that needs to be calculated when forecasting personal vehicle ownership transfer behavior. Both genders do not fluctuate on the same level given, and some of this can be attributed to the historical relationship of women and driving, which is present today, especially for older age groups.

Also, the preferences of certain brands of vehicles are common to all age groups, as well as for gender, and the results show a similar trend for the least preferred. That can be explained by the fact that most brands have different marketing targets. Brands tend to profile themselves within a demographic niche. But if that isn't so, more detailed research is needed.

Although, digitization has made it possible to move from keeping data in paper form to the ability to monitor data about customers and transfers that occur during the business process, for digital transformation to be fully brought to life it is necessary to research what customers want and fully adapt the company's efforts to provide the best possible user experience.

The results of this research can be used to advance the digital transformation of companies through two approaches. The first approach is to strengthen existing customer behaviors, and the second approach is to predict the behavior of existing customers, which through advertising, marketing, and other sales activities will influence the increase of customers who are currently less represented.

Business process planning following the possibilities of digital transformation is still in its early age, so companies must determine their capabilities, systematically use the benefits of digital tools, and in accordance with the knowledge gained through digital transformation tools, adapt their promotional activities to increase sales in a way that they design a user experience that will bring them the desired business results, and the users the desired user experience.

REFERENCES

[1] Salesforce, „What is Digital Transformation?“, 2020. Available at: https://www.salesforce.com/products/platform/what-is-digital-transformation/, [accessed: 16.05.2020].

[2] Europska komisija, „Što je zapravo digitana transformacija i kakve nas promjene očekuju“, 2020. Available at: https://ec.europa.eu/croatia/what_is_digital_transformation_changing_hr/, [accessed: 16.05.2020].

[3] Europska komisija, „Ulaganje u budućnost; Digitalna transformacija u razdoblju 2021.2027.“, EU BUDGET FOR THE FUTURE, 2018. Available at: https://ec.europa.eu/commission/sites/beta-political/files/budget-june2018-digital-transformation_hr.pdf/, [accessed: 16.05.2020].

[4] European Commission,"Digital Economy and Society Index Report 2019- Integration of Digital Technology", 2020. Available at: https://ec.europa.eu/digital-single-market/en/integration-digital-technology, [accessed: 16.05.2020].

[5] Bain & Company, “Customer Relationship Management”, Management Tools, 2018. Available at: https://www.bain.com/insights/management-tools-customer-relationship-managementy, [accessed: 17.05.2020].

[6] J. Avery, S. Fournier, J. Wittenbraker, “Unlock the Mysteries of Your Custome Relationships”, Harvard Business Review, Marketing, 2014. Available at: https://hbr.org/2014/07/unlock-the-mysteries-of-your-customer-relationships, [accessed: 17.05.2020].

[7] Deloitte, “2014 Global Automotive Consumer Study; Exploring consumers'mobility choices and transportation decisions”, 2014. Available at: https://www.autonews.com/assets/PDF/CA92618116.PDF, [accessed: 17.05.2020].

[8] M. Wachs, “The Automobile and Gender: An Historical Perspective”, University of California, Berkeley. Available at: https://www.fhwa.dot.gov/ohim/womens/chap6.pdf, [accessed: 17.05.2020].

The immigrant integration online training program in Finland

A. Hartikainen*, M. Ahola**, M. Apiola* and E. Sutinen*
* University of Turku, Turku, Finland
** LAB University of Applied Sciences, Lahti, Finland
anihar@utu.fi

Abstract - **Finnish integration training for immigrants aims to improve social inclusion into Finnish society and to promote employment. The first fully online integration training and the still ongoing program has conducted since 2015. During the training activities, the education provider has collected varieties of data from 2015 to 2019. Utilizing design science theory, authors demonstrate that integration training succeeds online at least as well as in the classroom when the design is implemented in a customer-oriented way lingually, pedagogically, and technically accessible. When teachers and the group learn collaboratively, and with ensured interaction, the training supports integration and finding career pathways. This study describes frugal and inclusive educational innovation. One main finding is that the teacher's ability to facilitate learning situations in distance-based synchronous e-learning differs from classroom instruction and requires detailed planning in advance. The training needs to prepare immigrants for the requirement of Finnish society. The desires of the students shape the training from the beginning.**

Keywords – e-learning; immigrant integration; distance learning; design science; education innovation

I. INTRODUCTION

A. Background of the design research

This study is design research of a fully online immigrant integration training program, which is the first in Finland, and one of the first globally, which is designed from beginner level and organized fulltime. This design is the first Finnish fully online integration training program aimed at training immigrants in Finland with language skills and promoting their occupational skills, and career planning. The design was conducted in Arffman Finland Ltd. (formerly Arffman Consulting Oy), an immigrant training company in Finland, by pedagogical coordinator Marja Ahola with team members Raisa Haikala and Ulla-Riitta Mikkonen. Arffman Ltd. provides online integration training since 2015 in Finland. This research utilizes data collected from 2015-2019 from 329 participants, who started and finished their integration training before July 2019. The findings suggest that the implementation of online integration training is as well as classroom-based integration training. The sample is large enough to exclude the possibility that participants have a great study or digital competence before selected for online training. Students' background and education varied significantly, with some participants with low literacy skills and low digital literacy. In the developed online training model, interaction is on the focus. Every day at the training begins with the synchronous guidance of groups to ensure that each participant can learn. Gradually, as the teacher learns more about the student's desire and competences, the learning and guidance process is personally tailored.

The factors of frugality guided the design of the training. The curriculum is broad, as are the target outcomes of integration training and inclusion [1][2]. The design of the implementation took place during the training and rested on the requirements of students. This approach is ideal for learners who do not have a target language proficiency or whose digital competence is low.

The research questions of this study are as follows:

1. How to design fully online integration training based on the Finnish Act on the Integration of Immigrants?
2. How well does the fully online integration training succeed in terms of participants' language level?
3. How well does the fully online integration training succeed in terms of participants' employment?
4. How well does the fully online integration training succeed in terms of participants' career pathways?

B. Integration Training in FINLAND

In 2019, 4.52% of the people living in Finland had different citizenship than Finnish. The past 20 years showed an increase in the percentage of immigrants in Finland; however, these numbers remain one of the lowest compared to other EU countries [3][4][5].

Out of the majority of immigrants coming to Finland, 21.8% came on a family basis, 8.4% as refugees or asylum seekers, 31.2% of the people seeking residence permits came to study or non-seasonal work, and 38.6% for other reasons [6][3]. The unemployed immigrants have a right to study in government-paying integration training. Employed or studying immigrants are omitted from intensive language learning, as no permanent body exists to support their studies.

Less than 35% of participants who take integration training in Finland reach the target language level B1.1 [7], which considered to be the threshold level in the European Framework of Languages [8] and also one of the requirements for applying for Finnish citizenship [9].

Finland is a country with a huge area and long distances. The population density in 2016 was 18 inhabitants/km² [10]. At the same time, Finland is a technology-based country with speedy and reliable

internet connections. In 2018, 97% of homes could get a stable and quick internet connection throughout the country. 92% of Finnish inhabitants use the internet; therefore, the digital competence of citizens is crucial for the Finnish economy. Information technology (IT) workers constitute a major and rapidly growing share of the overall Finnish workforce. Public services are mostly digital, and the government invests considerable effort into citizens' digital skills [11][12][13].

The funder of integration training, the Centre for Economic Development, Transport and Environment (ELY) monitors the learning outcomes of these integration training. The continuity of training depends on the achievement of objectives. The integration training of immigrants is organized in Finland as labor policy training within the Integration Act framework in all major municipalities. Each unemployed immigrant arriving in Finland receives an integration plan. Unemployed immigrants have the right and obligation to participate in free integration training, which lasts from six to twelve months or participate in other activities to support integration or employment during an integration period of three years [14].

The training aims to promote the integration and employment of immigrants and to support their inclusion into Finnish society. The training is full-time. In 2015, the length of the study day was seven hours. New models of integration training were introduced in 2017. The length of the study day changed to five hours in both classroom and distance-based integration training and diversified implementations.

With the planning of online integration training, both the Finnish Ministry of Labor and the funder were interested in whether integration training could start at the beginner level with low digital competence. The premise of the training is that participation is straightforward and accessible with the three dimensions of accessibility [15] [16], and participation is achievable for everyone, despite language proficiency or digital competence [17][18].

During the implementation of this research, language skills and learning requirements differentiated them, bringing together peers from different regions. Today, almost everyone, also those with low literacy, has a smartphone, and they can use it to stay connected to their networks by instant messaging, voice messaging, video calls, and email. The training is open to all types of learners. A student may have low literacy and low digital competence, but almost everyone has a smartphone and instant messaging applications, which they can use from the beginning of the training.

II. RELATED RESEARCH

Digital competence is part of citizenship [19]. Digital habits are the key to e-learning. However, many have low digital literacy, such as searching for jobs and other opportunities online, even though job search today is almost entirely online, as was shown in a research conducted in South Africa [20].

As was shown in a comparative study of Spanish adult students' attitudes toward ICT in the classroom, blended and distance language learning modes, the students, understand their digital competence. It can be utilized in course planning. E-learning should be designed based on an individual's desire and competence [21].

In Lagier's research [22], non-Western learners preferred group activities, interaction, relationship building rather than competition. Those who studied actively in the group received higher test scores. Research on Educational technology in transnational higher education in South East Asia showed the cultural politics of flexible learning [23]. Online learning is suitable for all types of learners but requires co-design with students depending on their cultural barriers, conditions, and competences.

Scaffolding in learning is defined to mean support, the process of support, interaction in a zone of proximal development (ZPD), or cooperation between the student and instructor [24]. Scaffolding facilitates students to achieve their goals in learning. Responsibility is shifting gradually to students as they develop their understanding and skills [25]. Students achieved communicative competence in a social context and through interaction [26] [27].

In Flanders, Dutch L2 teaching has shown that e-learning is well suited for low-literacy learners. The findings obtained challenge the previous notion that e-learning would only be suitable for learners with a higher level of education [18].

According to De Paepe [18], the main constraints in the design and implementation of an online curriculum are negative beliefs about the effectiveness of online language learning, costs, lack of technical and pedagogical support, and lack of teachers' digital skills. The most critical success factors were course design, technical, financial, and pedagogical support, development of learners 'competencies and skills, changes in attitudes, and development of guidance designers [18].

As was shown in the research of learner's perceptions from a Turkish distance language education context and sustainability in lifelong learning, students believe they learn language skills in the distance program. When productive skills as spoken interaction and writing in various genres are concerned, they believe that they cannot achieve these without the existence and guidance of a teacher explicitly teaching in a face to face context. On the other hand, students give good feedback on synchronized e-learning, where they meet with the group and the teacher, go through things together, have the opportunity to ask questions and discuss. Real-time interaction guided by a language teacher has proven to be a very functional solution to solve particular grammar and vocabulary issues, as well as to practice accurate pronunciation and spelling problems using text chatting [28].

A study of L2 e-learning in Portugal mentions that bringing together students of the same level from different schools is a benefit of e-learning [29].

High-quality online training requires planning, as also observed in other researches [30][21][17]. Topics and the learning functions are defined and distributed to teachers

according to the factors of linguistic and pedagogical accessibility [15][16] has also shown that lingual accessibility plays a significant role in e-learning [16].

Design is an innovation in inclusive education [23] [31]. Design is a frugal innovation [32] formed by linguistic, pedagogical, and technical implementation. Also, that e-learning does not require expensive investment or licensing from the education provider, and implementation is sustainable [16] [21] and equitable [21] as it requires high quality and effective interactive, inclusive education into the home of students economically and stably.

III. RESEARCH DESIGN

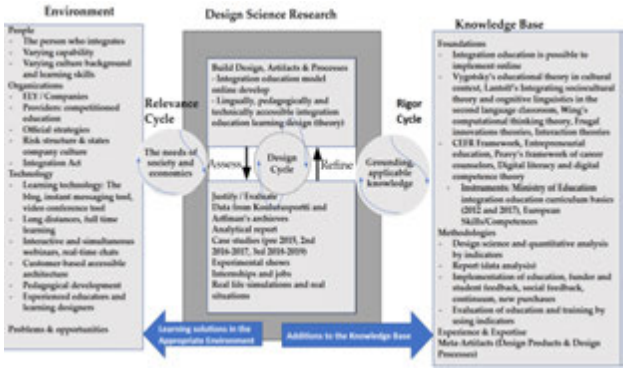


Figure 1: Design science cycles and significant artifacts in the Finnish Online Integration Training 2015-2019, source: M. Ahola, 2020

This research follows the design science research (DSR) method. DSR studies support a pragmatic research paradigm that calls for the creation of innovative artifacts to solve real-world problems and enhance educational processes [32]. (See Figure 1): The process divided into three cycles: The Relevance Cycle, the Design Cycle, and the Rigor Cycle.

Relevance Cycle inputs requirements from the contextual environment into the research and introduces the research artifacts into environmental field testing [32]. In this research, the environment is the requirements of Finnish society and economy and consists of three primary artifacts. 1. The people who are integrating with varying capability, cultural background, and learning skills, 2. the organizations: The conditions of the funder, companies, competition training, official strategies, risk structure, state's company culture, and Integration Act, 3. Technological solutions: learning technology, e.g., the blog, instant messaging and video conference tools, long distances and full-time learning, interactive and simultaneous webinars, real-time chats, customer-based available architecture, pedagogical development, experienced educators and learning designers.

The Relevance Cycle outputs the main problem to the Design Cycle. The Design Cycle supports a research activity for the construction and evaluation of design artifacts and processes. This research developed a fully online integration training as an ongoing process. The evaluation and development of the design followed a lingually, pedagogically, and technically accessible learning design. The evaluation is justified and grounded on feedback which was collected electrically from students during online integration training, and at the end of the training. Data was collected by employment services, which analyze and made statistics for training and influences on the new purchases.

In regards to immigrants who participated in the program during the years 2015-2019, 64% were female, while 36% were male. Students live in a wide area. Some moved during the training. In most cases, students were allowed to continue their integration course despite relocating. The students lived in 130 municipalities.

42% of students were born in the 1980s, so the majority of students were in their thirties during the training. Some 21% of the participants were born in the 1990s, while some 20% were born in the 1970s.

In regards to language, 24% of the students spoke Russian as their mother tongue. The second most common mother tongue was Arabic, which was spoken by 15% of participants. The third most common language was Thai, spoken by 11% of participants. Some 7% of the participants were native English speakers. 5% spoke Spanish, 3% Chinese, and 3% spoke the Austronesian language of Tagalog. The rest 32% spoke other languages. In total, the students had 48 different mother tongues.

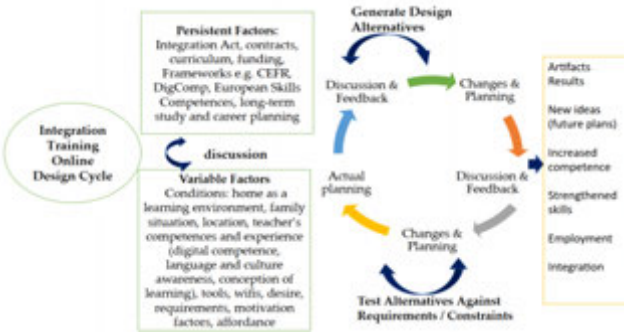


Figure 2: Description of making Immigrant Integration Training Online design (persistent and varying factors), source: M. Ahola, 2020

Cooperation between funders, educators, companies, and students is essential. The activities and feedback report to the official systems of the state administration, where the teacher forwards the outcomes and further pathways for the students. The feedback from the funder, collaborate companies, participants, and teachers were also collected. A report of the training to the employment services and the funder is also provided. Also, teachers analyzed the feedback. The concept followed a continuous design and development iterations based on feedback.

Design is an inclusive innovation formulated by utilizing students' resources, tools, and competencies, environments that create affordances, home as a learning environment, and free digital services. The approach avoided multiple logins and expensive online platforms. Learning content has been modified into everyday language and formatted into functions. Persistent factors, such as curriculum content, can be pre-processed into a format suitable for the web. Interactive virtual learning situations are created to the simulations in virtual classroom or facilitated in authentic locations. The daily

study programs and acquired skills informed the blogs supported by the curriculum aims.

The varying factors of institutional arrangements forming after the teacher have contacted the student, found out his or her digital and communicative competences, and conditions. Digital habits vary a lot between the students. As was observed, mainly, interaction is usually the strength of each participant, as almost all those entering training interacts with their relatives and friends in their home countries via mobile or network. They are familiar with mobile practices, although they are low-educated, or have low literacy and low digital competence. Similar experiences are described by Matli & al [20] in Latin America.

Online integration training connects participants across regions. Thus, the opportunity is to have sufficient groups with a similar target and progress, as was observed by Lagier [22]. Forming groups based on language proficiency, learning competence, a particular demand, requirement, interest, or profession, requires determined planning and ongoing collaboration between teachers. At the beginning of the training, the emphasis is guiding and building trust and empathy.

The Rigor Cycle provides grounding theories and methods and adds the new knowledge generated by the research to the growing knowledge base. Applicable knowledge for the development of design, grounded for Vygotsky's [33] educational theory in cultural context, Lantolf's [34] integrating sociocultural theory and cognitive linguistics in the second language classroom, form the conception of learning. Computational thinking [35] is needed to recognize digital competence and to define the demand for scaffolding. Frugal innovations theories [31][36] define design in an economical and user-oriented way. The design has required frameworks that take into account language learning, digital literacy, and career competencies. Methodologies, which enable this study are design science and quantitative analysis by indicators, report for data analysis, implementation of training, feedback, continuum, and new purchases, which enable evaluation and development of training.

IV. RESULTS

The language level is one of the most important indicators of successful integration education. Of all those who completed the immigrant integration online training, 36% achieved CEFR proficiency Levels [8] target level B1 in the Finnish language. Approximately every second, 52%, of those who studied at least 200 days, reached the target level B1 of integration training.

Regarding placement after the training, 43% of students applied for a job or internship at the end of their course. Some 17% had a job and 16% a study place. 3% of students found employment in a family business or started their own business. 9% started a work trial. 9% moved out of the labor force. The majority of those who moved out of the labor force started maternity leave or care leave, a few completed their courses and were referred for an occupational health survey.

88% of students went through online guidance and learning processes by described design. Without career paths, 9% moved out of the labor force and 3% which dropout.

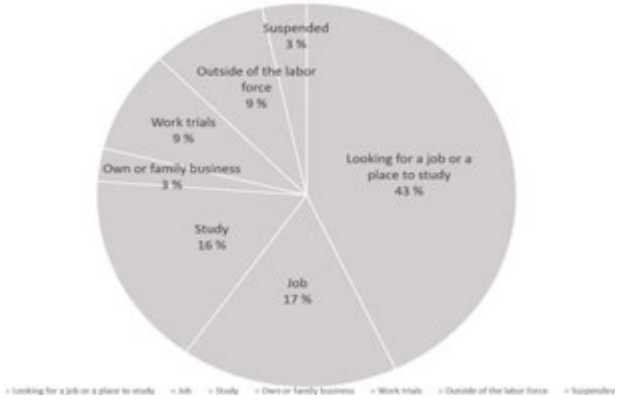


Figure 3: Placement after Online Immigrant Integration Training, source: M. Ahola, 2020

The training emphasizes daily interaction with the group and the teacher, as well as an individual guidance process that progresses from the learner's requirements.

Learning of productive skills is arranged online, although some participants have low digital competence. At the beginning of language learning, it is sufficient that each learner can open a communication application, a microphone, and a video connection. E-learning with synchronous interaction is intensive and gradually scaffolds the student to active participation and inclusion. A multi-channel network environment utilizes all the senses simultaneously. Our results show that in the online integration training, context-oriented and synchronized, e-learning promotes linguistic competence and enhances language learning. E-learning produces content that can retake. Slow-progressing students repeated what they had learned, e.g., by watching webinar recordings.

Our findings show that differences between learning cultures confuse many students at first. For many, the conception of e-learning is entirely different from the implementation of our online training. Students with low-literacy and low digital competence tend to have a cynical and skeptical attitude towards e-learning with distance before the first learning experience. After getting used to online learning systems and routines, attitudes became neutral or positive.

Findings from the research show that online scaffolding supports the learner's development into a self-effective actor in the digital society. A detailed understanding of objectives, task completions, and teacher feedback increases commitment and achievement of targeted learning outcomes. Our results also show that the teacher's role changed to that of facilitator and counselor. Scaffolding in the e-learning process keeps learners' attention on the essentials, adjusting the requirements of the exercises, maintaining motivation, and relieving learner's frustration.

A gradual transition to the Finnish operating environment initially requires a lot of guidance and understanding of cultural differences. During the training, language and study skills, culture, social practices, and

working life skills are learned in parallel. In the online immigrant integration training, students are engaged in joint designing. E-learning is developed together with students. Participants were asked about the tools available, the availability of the internet connection (if necessary, the institution provides them), the language level, whether they already have Finnish language skills or another auxiliary language with the teacher. Additionally, issues to be clarified before starting the training are conditions, study skills, study history, occupation, motivation, social skills, responsibility, and restrictions on participation. The information influences the effects of implementation designing and tool selection.

V. DISCUSSION

The first research question asked, *"How to design fully online integration training based on the Finnish Act on the Integration of Immigrants?"* The results and analysis have shown the following.

First, the implementation must be effortless to students and teachers. The integration training relies on three dimensions of accessibility: lingual, pedagogical, and technical. In addition to learners with higher education and digital competence, participants with low literacy and low digital competencies successfully participate in immigrant integration online training. Second, the teacher's role is advanced to the facilitator and counselor. Learning outcomes are better if the guidance and counseling process is carefully defined. The training requires teachers who know precisely how to manage and scaffold learning processes in a digital environment. Third, interaction defines the whole design to encourage the desire of students to interact with teachers and peers. Interaction design has contributed to the formation of interactive learning situations and networking at training by providing unexpected interactions.

Our second, third and fourth research questions asked: *"How well does the fully online integration training succeed in terms of participants' language level, employment, and career pathways?"* The results show the following.

First, contrary to previous expectations, the targeted learning outcomes of online integration training, especially productive skills, developed at least as well as, or even slightly better than in classroom-based training. Second, the learners benefit from their previous competence and know-how in learning tools, and because learning reaches them over long distances from areas where education is not available. Third, the majority of the students has reached the target outcomes on language proficiency and had a planned career path and placement at the end of the training.

VI. CONCLUSIONS

Based on the findings, the authors have a new understanding about arranging of online learning, interaction and design process.

A fully online integration training is a way to learn the Finnish language and successfully integrate into the digital society. Success requires a redefinition of the pedagogy. Further research for intercultural online guidance and the framework for integration is needed. The teachers indicated some surprising situations for further research. Teachers are required to take risks and approve that there are no answers to everything. The final form of design may not be what the teacher is used to or expected.

Learners need to adapt to the new learning culture, which might aim e.g., towards entrepreneurial learning. Synchronous, accessible e-learning strengthens digital competence, inclusion, equality, and employment opportunities. Therefore, learner-centered e-learning is suitable for many immigrants, which are gradually scaffolded from a teacher-centered to a self-effective culture and to participate in the digital society.

Both the Flemish experience [17][18] and the data from the study showed that low-educated learners benefit from the diversity of e-learning, such as retaking lessons, studying at their own pace, or from the personal attention of the online educator.

Once the varying factors have identified, it is possible to guide the students to a group that meets their desire and to select optimally accessible, group-specific technological tools and pedagogical, lingual, and cultural-conscious choices. Observations in this research are similar, as Ilin describes [28]. Instant messaging has taught writing skills in a real context.

The strength of online training has been the flexibility, which makes it possible to provide an appropriate level of training for everyone by differentiation and combining groups. Similar findings were reported in the research of Lourenço and Martins [29]. In integration training, the group usually has 15-25 participants. In online training, the teacher needs to set aside time for speaking exercises with a peer and in a group because there are no spontaneous speaking situations. Therefore, learning arrangements need to support ways to learn productive skills. At the same time, the teacher receives confirmation that the skill has been learned.

By forming skills to be learned as a function of real-life situations, e.g., "I can book an appointment with the doctor" or "I can order in a restaurant" make it possible to learn practical communicative competence [24][26][33] in everyday life and working life situations. The pragmatic approach encourages students to expand and search for more information in a guided way, using the language they are learning. The next steps in our research will be to pay attention to collaborative learning networks and entrepreneurial learning.

REFERENCES

[1] National Board of Education. Fundamentals of the Integration Plan for Adult Immigrants. Principles of the Teaching Plan for Adult Migrants' Integration Education 2012 Regulations and Guidelines 2012: 1. In Finnish. Juvenes Print - Tampereen Yliopistopaino Oy. 2012.

[2] National Board of Education. Updated implementation models for integration training 2017. In Finnish. 2017. Available at: https://www.oph.fi/sites/default/files/documents/188626_koto_koulutusmalleja_2017_final.pdf, [accesed: 10.05.2020]

[3] Statistics Finland. Immigrants in the population. In Finnish. 2019. Available at: https://www.stat.fi/tup/maahanmuutto/maahanmuuttajat-vaestossa.html, [accesed: 10.05.2020]

[4] The Family Federation in Finland. Number of immigrants. In Finnish. 2019. Available at: https://www.vaestoliitto.fi/ tieto_ja_tutkimus/vaestontutkimuslaitos/tilastoja/maahanmuuttajat /maahanmuuttajien-maara/, [accesed: 10.05.2020]

[5] Eurostat. Statistics Explained. 2019. Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Migration_and_migrant_population statistics/fi/finland_fi.pdf, [accesed: 10.05.2020]

[6] Finnish Government. Regional development prospects. In Finnish. 2019. Available at: http://julkaisut.valtioneuvosto.fi/ bitstream/handle/10024/161538/TEM_28_19_Alueelliset_kehitysnaky mat_Kevat_2019.pdf, [accesed: 10.05.2020]

[7] The National Audit Office of Finland. Performance Audit Report. Integration training. Audit reports of the National Audit Office of Finland, vol 15/2018. Helsinki. Grano. In Finnish. 2018. Available at: https://www.vtv.fi/app/uploads/2018/10/15152336/VTV-Tarkastuskertomus-15-2018-Kotoutumiskoulutus.pdf, [accesed: 10.05.2020]

[8] Council of Europe. The CEFR Levels. 2001. Available at: https://www.coe.int/en/web/common- european-framework-reference-languages/level-descriptions, [accesed: 10.05.2020]

[9] M. Ala-Kauhaluoma, S. Pitkänen, S., J. Ohtonen, F. Ramadan, L. Hautamäki, M. Vuorento, and H. Rinne, Multimodal study on the effectiveness of integration measures. In Finnish. Publication of the The National Audit Office of Finland., 1, 2018.

[10] Official Statistics of Finland (SVT), Demographic structure. Helsinki. In Finnish. 2020. Available at: http://www.stat.fi/til/vaerak/index.html, [accesed: 10.05.2020]

[11] European Commission. Finland Country Report. 2019. Available at: https://ec.europa.eu/info/sites/info/files/file_import/2019-european-semester-country-report-finland_en.pdf, [accesed: 10.05.2020]

[12] Ministry of Employment and the Economy, Finland's artificial intelligence period Finland as a leading country in the application of artificial intelligence: Objective AND recommendations for action. Publications of the Ministry of Employment and the Economy 41/2017. In Finnish. 2017. Available at: http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/80849/TEMrap_41_2017_Suomen_teko%C3 %A4lyaika.pdf, [accesed: 10.05.2020]

[13] European Commission. Index of the digital economy and society (DESI) 2018: Finland. In Finnish. 2018. Available at: http://ec.europa.eu/information_society/newsroom/image/document/2018-20/fi-desi_2018-country- profile-lang_4AA45BA7-D6C7-68CD-07425D441D640D06_52337.pdf, [accesed: 10.05.2020]

[14] Finlex. Act on the Promotion of Integration. 2010. Available at: https://www.finlex.fi/fi/laki/ajantasa/2010/20101386, [accesed: 10.05.2020]

[15] M. Ahola, R. Haikala and A. Zafar, Accessible Language Learning Online. ICERI2019 Proceedings, 1. 2019, pp. 2432–2439. Available at: https://doi.org/10.21125/iceri.2019.0645, [accesed: 10.05.2020]

[16] A. Downer, A.Shapoval, O. Vysotska, I. Yuryeva and T. Bairachna. US e-learning course adaptation to the Ukrainian context: lessons learned and way forward. BMC medical education, 18vol 1. 2018, pp- 1-10.

[17] L. De Paepe, C. Zhu and K. Depryck. Learner Characteristics, Learner Achievement and Time Investment in Online Courses for Dutch L2 in Adult Education. Turkish Online Journal of Educational Technology-TOJET, 17(1), 2018, pp. 101-112.

[18] L. De Paepe, C. Zhu and K. Depryck. Development and implementation of online Dutch L2 courses in adult education: educators' and providers' perceptions of constraints and critical success factors. Innovation in Language Learning and Teaching, 13(3). 2019, pp. 277-291. Available at: https://doi.org/10.1080/17501229.2018.1462369, [accesed: 10.05.2020]

[19] S. Carretero, R. Vuorikari and Y. Punie. The Digital Competence Framework for Citizens. Publications Office of the European Union. 2017.

[20] W. Matli and M. Ngoepe, Capitalizing on digital literacy skills for capacity development of people who are not in education, employment or training in South Africa. African Journal of Science, Technology, Innovation and Development, 12(2), 2020, pp. 129-139. Available at: https://doi.org/10.1080/20421338.2019.1624008, [accesed: 10.05.2020]

[21] M. Arrosagaray, M. González-Peiteado, M. Pino-Juste and B. Rodríguez-López. A comparative study of Spanish adult students' attitudes to ICT in classroom, blended and distance language learning modes. Computers & Education, 134, 2019, pp. 31-40. Available at: https://doi.org/10.1016/j.compedu.2019.01.016, [accesed: 10.05.2020]

[22] J. Lagier, Distance learning and the minority student: special needs and opportunities. The Internet and Higher Education, 6(2), 2003, pp. 179-184. Available at: https://doi.org/10.1016/S1096-7516(03)00023-X, [accesed: 10.05.2020]

[23] C. Ziguras, Educational technology in transnational higher education in South East Asia: the cultural politics of flexible learning. Journal of Educational Technology & Society, 4(4), 2001, pp. 8-18.

[24] M. Suni, Second language in interaction: Sharing linguistic resources in the early stages of second language acquisition (No. 94). In Finnish. University of Jyväskylä, 2008.

[25] N. F. Jumaat and Z, Tasir, Instructional scaffolding in online learning environment: A meta-analysis. In 2014 International Conference on Teaching and Learning in Computing and Engineering, 2014, pp. 74-77.

[26] M. M. Bakhtin, The dialogic imagination: Four essays, ed. Michael Holquist, trans. Caryl Emerson and Michael Holquist (Austin: University of Texas Press, 1981), 84(8), 80-2. 1981.

[27] H. Dufva, M. Suni, M. Aro and O. P. Salo, Languages as objects of learning: language learning as a case of multilingualism. Apples-Journal of Applied Language Studies. 2011.

[28] G. Ilin, Sustainability in Lifelong Learning: Learners' Perceptions from a Turkish Distance Language Education Context. Sustainability, 11(19), 5284, 2019.

[29] C. Barros Lourenço and A. Sousa Martins, Online Courses of Portuguese as a Second Language. In Proceedings of the 7th International Conference on Computer Supported Education-Volume 1, 2015, pp. 380-385.

[30] G. Osman and S. C. Herring, Interaction, facilitation, and deep learning in cross-cultural chat: A case study. The Internet and Higher Education, 10(2), 125-141. 10(2), 2007, pp. 125–141. Available at: https://doi.org/10.1016/j.iheduc.2007.03.004, [accesed: 10.05.2020]

[31] J. Chataway, R. Hanlin and R. Kaplinsky, "Inclusive innovation: an architecture for policy development," Innovation and Development, 4(1), 2014, pp. 33-54.

[32] A. R. Hevner, S. T. March, J. Park and S. Ram, S. Design science in information systems research. MIS quarterly, 2004, pp. 75-105.

[33] L- S. Vygotsky, Thinking and speech. In The collected works of L.S. Vygotsky: Vol 1. Problems of general psychology. Transl. N. Minick, eds. R. W. Rieber & A. S. Carton. New York: Plenum Press, 1987, 251.

[34] J. P. Lantolf, Integrating sociocultural theory and cognitive linguistics in the second language classroom. Handbook of research in second language teaching and learning, E. Hinkel, Eds, vol II, New York: Routledge 2011, pp. 303-318.

[35] J. Wing, Research notebook: Computational thinking—what and why? The Link Magazine, 2011.

[36] M. Hyypiä and R. Khan, R. Overcoming Barriers to Frugal Innovation: Emerging Opportunities for Finnish SMEs in Brazilian Markets. Technology Innovation Management Review, 8(4), 2018, pp. 38-48. National Board of Education. 2012. Fundamentals of the Integration Plan for Adult Immigrants. Principles of the Teaching Plan for Adult Migrants' Integration Education 2012 Regulations and Guidelines 2012: 1. Juvenes Print - Tampereen Yliopistopaino Oy.

Assessing the Impact of Mobile Educational Games on Student's Success within Mathematics Subject in Primary Schools

Majlinda Fetaji*, Ermira Kajtazi*, Bekim Fetaji**, Halil Snopce* and Marika Apostolova*
* South East European University/Computer Science, Tetova, North Macedonia
** Mother Teresa University/Computer Science, Skopje, North Macedonia
m.fetaji@seeu.edu.mk, ek21776@seeu.edu.mk, bekim.fetaji@unt.edu.mk, h.snopce@seeu.edu.mk

***Abstract* - The research focus is set in investigating the impact of educational games as instructional strategy in student's success within the course of Mathematics in High School. In order to assess the impact we have developed a mathematical mobile Game for Android Devices, to be used as a case study, and assessed the students' achievements. We have used quantitative research methodology and as research method used a questionnaire to collect the data. We collected the data before and after playing the game, so we can assess the difference of the time spent on educational games and the level of transferred knowledge. The system is expected to help high school students to easily orient and engage during lecture classes and easily find and use the learning content. The system has been evaluated on its functionalities and the usability, on two groups of users, one group with computer science background and the second group of users as novices with different field background. Afterwards we used regression analyses to determine the impact. The research study tries to contribute with identified impacting factors for such systems, analyses of the improvement in user friendliness, and usability. Insights and recommendations are provided.**

Keywords - mobile devices, mobile learning, mathematics, game based learning, evaluation

I. INTRODUCTION

Mobile devices nowadays have made a huge impact in the society, especially in areas that include social life, business, health and education, industry etc [3]. According to [2] the adoption and use of IT in teacher Education have a positive impact on teaching, learning, and research. It will increase flexibility so that teacher educators and teachers can access the information regardless of time and geographical barriers. According to [4] educators today are using technology to engage pupils and encourage them to learn by putting the education in educational apps. Pupils find traditional way of learning hard and boring. Numerous studies [2],[3],[6] have shown evidence of increased interest to learn when mobile devices are incorporated into the learning environment. Over 80,000 apps are classified as education and learning based [4]. The purpose of this research is to find out how mobile game applications will impact the process of learning and assess the results in the subject of Mathematics. Following the guidelines from [5] we expect that learning subject of Mathematics by a mobile game application to be developed will be more interactive and will have positive effect by increasing pupil's interest in learning.

II. RESEARCH METHODOLOGY

We have used quantitative research methodology and as research method used a questionnaire to collect the data. This research had as target the fifth grade pupils in primary school of Mitrovica, Shaban Idrizi. In the research two classes have been involved, a control and experimental group. The study used the an experimental approach. Participants consisted of (50) students who were registered in a course called "Strategies of Teaching and Learning" offered by the Department of Curriculum and Instruction at the College of Education to students majoring in English at the Department of English Language at the College of Science and Arts. As usual participant students were divided into experimental and control groups of (25) students in each. The course content was uploaded to the University Google Classroom System for students in the experimental group. Another hardcopy of the course content was delivered to students in the control group by hand in the first meeting. Data collection tools included an academic achievement test and conversational skills rating scale. We collected the data before and after playing the game, so we can assess the difference of the time spent on educational games and the level of transfered knowledge. Another purpose of this research is to find out how pupils and teachers will adopt the mobile learning methodology in learning process. Since children nowadays have access to various entrainments and games, we expect this app not to be difficult for them to use. By using a new and entertaining learning method, we have find out whether there will be a significant increase in the success and the results of achievement in the process of learning. Another aim of the research is to increase the motivation and interest of learning by not wasting their time playing but learning and playing in the same time. Our study is built on numerous studies involving teaching novice programming using mobile technology in Mathematic subject. This study is an action research. Participants of this research will be divided in two groups. One of the classes will be assigned as a control group, and the other one will be assigned as the experimental groups. A preliminary survey from all groups will be conducted.

This survey questions will aim to realize the student's background knowledge of the topics, usage of the internet and other technological devices. One class will use the mobile application Play and learn that we have created and the other will use the traditional method of learning. The application contains a lot of options for pupils to learn. The basic option is the Mash mole game, in five categories with five different Mathematical subjects whom pupils have already learned. The other option is quizzes with the same subjects and also the learning materials option. Pupils have the option to see their height score for each category so they can track their progress. They have also the help option, and about application option. This action research was 5 weeks long. First of all, the two groups take tests for the first two subjects. They had two weeks to learn and prepare, one of them with the application and the other group in the traditional way. After two weeks they took another test for the same subjects. The same procedure was practiced for the 3 other subjects. The tests and observations were collected and recorded. Furthermore, questionnaires have been recorded and transcribed for analysis. After action research we have done students' needs analysis, experimental and empirical research where we measure parameters and comparative research. After action research we will do experimental and empirical research where we measure parameters and comparative research.

III. HYPOTHESIS AND RESEARCH OBJECTIVES

This research study objective is to investigate some important research questions as follows:

- What is the impact of Mobile Educational Games on increasing the attention and interest in the learning process?
- What is the impact of Mobile Educational Games on improving student's success?
- What is the aspect of adaption of the game to the pupils and teachers?
- How Mobile Educational Games affects students, their motivation and will to learn?
- Mobile Games have improved my thorough understanding of the course content?
- Does Mobile Educational Games encourage competition among pupils?

The research study has these hypothesis:

H1: "Educational games have a positive impact on improving student success and achievement"

H2: "Educational games have an impact on increasing students' attention and interest in learning".

IV. BACKGROUND RESEARCH

According to [6], [7] nowadays pupils use mobiles to play different games. The subject of the research in this paper is the educational application of mobile games for teaching and learning using mobile devices or smartphones, specifically applying it in teaching the subject of Mathematics with an interface in Albanian language; by the application of mobile educational games, in order for pupils to gain knowledge in easy and entertaining way. According to [4] mobile learning in Saudi Arabia, for instance it has been encouraged to be used in higher education institutions because of a set of factors such as the availability of mobile phones, their ability to motivate students, and the freedom and privacy they provide to share information. According to [2] Mobile Learning is defined as E-learning that uses mobile devices or learning connected to a mobile device. It is mainly based on mobility of technology, mobility of learners and mobility of learning [3]. According to [1] incorporating interactive tools one methodology will positively influence to make their class more interesting and motivate the pupils to learn more. There are a few educational applications in Albanian language, so a mobile game application will make a small contribution to the educational system in Albanian language. It will be an example of a new method for teachers to teach and pupils to learn.

V. RESULTS AND DATA ANALYSIS

In order to find out how the educational game responds to the user requirements a questionnaire has been given to in total 64 pupils one before they used the game, and afterwards they used the game. The questionnaire in total had 9 question three to capture the general data about gender, academic achievement and the others to capture the research questions input. It has questions about the system and one before they used the game and one after they used the game.

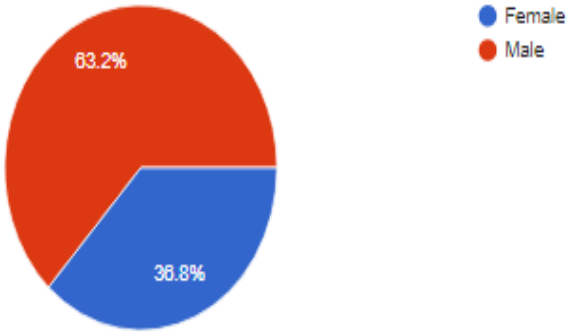


Figure 1. Gender Analyses of respondents

Elementary school that has been used as Case Study has more male respondents 63.2% then female 38.8%

Q2: Do you have access to digital educational resources?

As shown in the figure below, 25% of pupils have excellent access to educational resources while 10% have very good access and 25% have good access. Overall we can see that they have good access to digital educational resources.

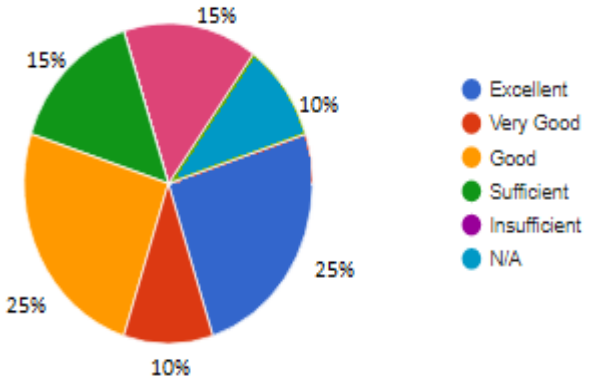


Figure 2. Access to digital educationa resources

Q3: Can Educational Games can help you learn better?

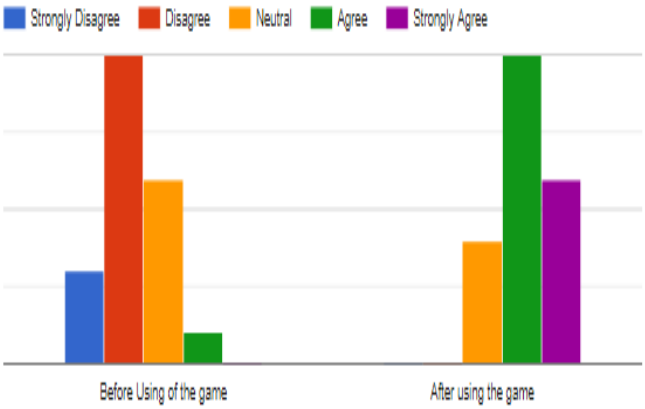


Figure 3. Do you think Mobile Games can help you learn better?

From the feedback from the respondents we can see that they initially where neutral or disagree that educational games can help them learn better, but after they tried the educational game and experience it practically they changed their mind and consider that they can learn better.

Q4: The impact of Mobile Educational Games on increasing the attention and interest in the learning process is high?

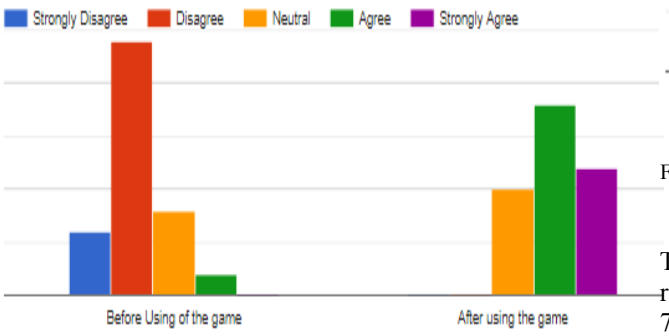


Figure 4. The impact of Mobile Educational Games on increasing the attention and interest in the learning process is high?

As shown in the figure above, we can see that pupils initially where skeptical about the use of digital educational games but after they have used it most of them agree that their interest and attention has been increased.

Q5: The impact of Mobile Educational Games on improving student's success is high?

From the answers of the respondents we can see that large number of pupil initially disagreed that educational games can improve their success, but after they used it they consider it can improve their success.

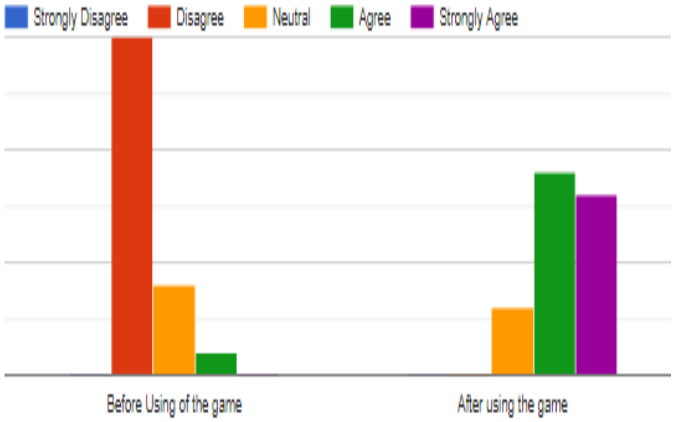


Figure 5. The impact of Mobile Educational Games on improving students success is high?

Q6: Mobile Games have improved my thorough understanding of the course content?

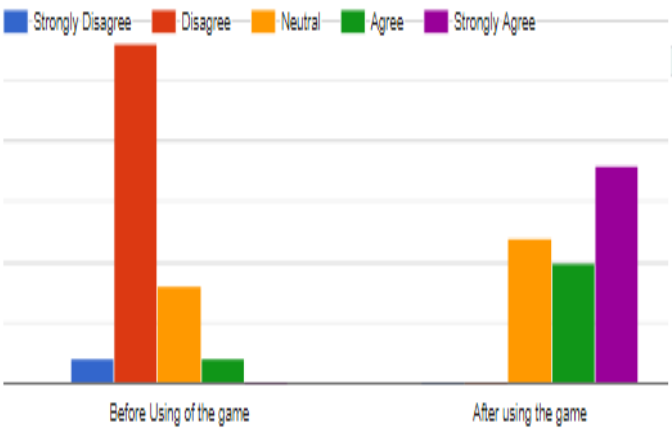


Figure 6. Mobile Games have improved my thorough understanding of the course content?

The results from these questionnaires showed a positive relationship between the post-test and pre-test, $T(1,62) = 79.989$, $p < 0.001$, and there were no significant differences between genders $T(1,62) = 3.076$, $p = 0.079$.

TABLE I. ANOVA RESULTS FOR CONDITION FROM POST- TEST

Variation	The sum of squares	diff	Mean Square	T Value	S
all	48.538	3	48.538	1.396	0.319
gender	460.588	4	603.014	9.006	0.007
all * gender	51.587	7	63.078	0.976	0.326
errors	2686.247	27	37.658		
Total	5228.472	41			

TABLE II. ANCOVA RESULTS FROM POST- TEST

Variation	The sum of squares	diff	Mean Square	T Value	S
all	463.276	2	463.276	17.051	0.001
gender	87.911	3	84.917	2.925	0.074
Post-Test	2307.206	4	2207.212	79.968	0.001
all post-Test	137.964	6	134.964	4.849	0.038
error	1699.849	26	32.017		
Total	6071.508	41			

*P<0.05, **P<0.001 (THESE ARE STANDARD VALUES FOR THRESHOLD LEVELS)

VI. CONLUSION

In this research study, the impacts of game based learning instruction method has been examined. We particularly aimed at investigating the opinion and attitude of pupils and therefore used two questionnaires, one before they used the game in learning mathematics used the game. Results showed that mobile learning had quite significant effect on both students' academic achievement and conversational skills. In light of these findings faculty members were recommended to adapt the use of Mobile Learning in their classes. Within the research study concluded that there is quite increase in positive attitude and interest after the game has been used from pupils and it increased their attention and level of transfer of knowledge.

Educational games can be viable help in providing learners with additional learning resources and increasing their knowledge. The most important issues identified are: lack of exposure to new teaching methodologies, non-existing educational games in local languages, especially in Albanian language in the school used as case study, lack of internet access in school, and have been the most important ones identified. Addressing this issues will increase substantially the positive usage of educational games in increasing the level of knowledge by including the game in study curricula.

This research study aimed to contribute with investigating the impact at pupils in gaining knowledge involving educational games in their curricula's. From the pupils feedback we can see that they have increased knowledge of the subject matter and the anova analysis confirmed this. We expect this insights to help other researchers to further investigate and use game based learning in liaison with fleeped learning and analyses the advantages of this approach.

REFERENCES

[1] Ekaterina Pechenkina, Daniel Laurence, Grainne Oates, Daniel Eldridge, & Dan Hunter . (2017). Using a gamified mobile app to increase student engagement, retention and academic achievement. International Journal of Educational Technology in Higher Education. Retrieved from https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-017-0069-7

[2] Iván Cantador, & José M. Conde. (2018) (n.d.). EFFECTS OF COMPETITION IN EDUCATION: A CASE Sudy in e-larning environment. Retrieved from https://pdfs.semanticscholar.org/95a0/4babb8841f3f644e2d7d497c98807eac3595.pdf

[3] Josh Cuevas, & Hannah B Miller. (2017). Mobile Learning and its Effects on Academic Achievement and Student Motivation in Middle Grades Students. Retrieved from https://www.researchgate.net/publication/318969235_Mobile_Learning_and_its_Effects_on_Academic_Achievement_and_Student_Motivation_in_Middle_Grades_Students

[4] Kathy Hirsh-Pasek, Jennifer M. Zosh, Roberta Michnick Golinkoff, James H Gray, Michael B Robb, & Jordy Kaufman. (2015). Putting Education in "Educational" Apps:Lessons From the Science of Learning. Psychological Science in the Public Interest. Retrieved from https://www.researchgate.net/publication/275410459_Putting_Education_in_Educational_Apps_Lessons_From_the_Science_of_Learning

[5] Keengwe, J. (2018). Handbook of Researchon Mobile Technology,Constructivism, andMeaningful Learning. (J. Keengwe, Ed.) IGI Global. Retrieved from https://www.academia.edu/35129614/Handbook_of_Research_on_Mobile_Technology_Constructivism_and_Meaningful_Learning_A_volume_in_the_Advances_in_Educational_Technologies_and_Instructional_Design_AETID_Book_Series

[6] Royce Kimmons, Brant G. Miller, Julie Amador, Christopher David Desjardins, & Cassidy Hall. (2015). Educational Technology Research and Development. Technology integration coursework and finding meaning in pre-service teachers' reflective practice. Retrieved from https://www.researchgate.net/publication/281068124_Technology_integration_coursework_and_finding_meaning_in_pre-service_teachers'_reflective_practice/references

[7] Swan, K., Mark van 't Hooft , & Annette Kratkosk. (2015). Teaching and learning with mobile computing devices: Closing the gap. Mobile Technology: The future of learning in your hands. Retrieved from https://www.academia.edu/2738439/Teaching_and_learning_with_mobile_computing_devices_Closing_the_gap

The teacher's role in discovery, preparation, and development of gifted students in the field of informatics

G. Atanasova, P. Hristova *
* University of Rousse, IIT Department, Rousse, Bulgaria
gatanasova@uni-ruse.bg

Abstract - **The national development strategies of many countries prioritize the stronger orientation to an economy based on the knowledge, on the development of research work, and technological development and innovation. Bulgaria is one of the few countries in the world that offers young students the opportunity to study computer science from a very early age in extracurricular classes and also organizes contests in computer science for students of secondary school classes. As a result, Bulgarian students can be recognized on average as the priorities of international information competitions. The role of teachers in discovering and preparing gifted children for such competitions is an irrevocable and challenging task. Teachers need to use focused and well-planned approaches. The general didactic principles for creating a learning process are presented. Essential for the successful preparation of future medalists is the early start of learning the basic principles of programming. Some of the fundamental practical situations that show the significant role of the teacher are presented. Proof of the applicability of the proposed approach is the performance by years of students from the school in Rousse.**

Keywords - teachers' support, gifted trainee, student's competitions in programming, pedagogical approaches

I. INTRODUCTION

The necessity for a modern knowledge-based information society development places serious demands on educational institutions. They are called upon to prepare both computer literate citizens and highly qualified software specialists. This problem has been identified as essential within the European Union (EU). In the Europe 2020 strategy, special attention is given to EU countries "to ensure sufficient availability of cadres with natural, mathematical and engineering education and to focus school curricula on creative, innovative and entrepreneurial thinking" [1]. This EU Directive is in line with the approach taken in Bulgaria and the multiannual efforts of teachers and university lecturers to involve early adolescents in towards new technologies and informatics and build a lasting interest. Bulgaria has traditions and achievements in preparing little students to participate in competitions in informatics from an early age. Not accidentally, the country was the initiator and host of the first European Junior Olympiad in Informatics (EJOI) in the year 2017 [2]. Organizing out-of-the-classroom informatics curricula enables gradual and consistent learning of algorithms and programming skills using natural children's interest in computers from computer games. The positive results achievement is closely related to the answers to the following two important questions: What should be empowered teaching children and are there any teachers to train them? The work on the discovery, preparation, and development of young talents in the field of informatics has its specificities that need to be taken into account. Recently, more and more scientists have been drawing attention to the means used for little students' preparation in out-of-the-classroom computer science schools. In Bulgaria, many authors discuss mentioned above questions in their works. Some authors make an exhaustive analysis of the informatics and information technologies curricula in Bulgarian schools [3], [4]. They also outline the history and evolution of the competitions and the Olympiads in Informatics. Some of them suggest approaches for managing the building of the participants in the Olympiads of Informatics and IT. The suggested models, more often, are presented as a system of basic tasks for prominent students' preparation for participation in the Olympiads in Informatics. Other Bulgarian author offers a spiraling approach to train 10 - 11-year-old children in informatics [5]. In her previous work, Hristova offers an objective approach for organizing the extra-curricular work in Informatics for pupils up to the 5th grade [6].

According to N. Wirth, "Programming is a constructive art" the question is: What approaches and methods should be appropriate to prepare the beginner students to master this creative activity so to ensure high quality and feasibility (methodical, cognitive, and organizational) of this training? [4].

The role of the teacher as a person, a specialist in informatics, and a scientist to achieve the ultimate goal is irrevocable. Still, the research related to the preparation and requirements of lecturers which preparing students for contests in Informatics is relatively small. There is a necessity from scientifically based approaches to work for both the pedagogic and informatics. The purpose of the present work is to offer a scientific ground approach to the work of the teachers in the discovering and preparation of gifted children for programming competitions. The hypothesis is the application of the pedagogical principles in the learning process will give a chance of a teacher to increase the number of successful competitors.

II. METHODOLOGY

A. Didactic Principles In The Programming Training Process

A scientific approach has been based on established scientific considerations. In this regard, based on an analysis of general didactic principles for building a learning sequence, their applicability was sought to prepare gifted children for programming. Prof. Petrov defines the principles of education as a system of basic positions, leading ideas, or general requirements. He outlines the overall activity of the trainees and the teachers in the learning process [7]. They are the theoretical basis for determining the content, so they have an essential place in the pedagogical science that determines the proper planning and implementation of the learning process. It follows each principle description in the context of the specific work of teachers in the preparation of competitors in informatics:

1) Principle of scientific excellence. It relies on the normative link between science and the subject matter. This principle requires the content to comprise objective scientific facts, concepts, laws, theories of the field concerned, and to reveal contemporary scientific achievements. Focused on the components of the learning process in programming training, this principle defines socially determined to learn programming objectives and scientifically refined algorithmic tasks addressing contemporary scientific problems, using scientifically proven methods and approaches to constructing algorithms, forming concepts, laws, theoretical formulations and practical realization in a specific programming language.

2) Principle of visibility. The visibility provides the primary perception of the subjects studied, included in the curriculum, and the construction of the concept. This principle describes how it is ensured the transition from the concrete to the abstract through sensory support. This principle of visibility, in general, can be denoted by the requirement to use visual aids in the educational process. The main factor that directly stimulates the activity and consciousness of learners in the learning process. This principle is particularly important for quality education programming and has a high value for the practical training of future contestants. Each algorithm operator, the change of the value of the variables, and the path for transforming the input data into the desired results are visual perception. In the programming training, conditional visual and dynamic visualization is applicable, the first being provided by the graphical representation of the algorithms and the second using animation environments or visualization of algorithms. The result is an increased observance and abstract thinking in small contestants.

3) Principle of consciousness. This principle is the unity between the learner's learning attitude and the understanding of the learning material. Consciousness is an essential prerequisite for overcoming the mechanical learning of learning information. Consciousness as personality quality is expressed as unity between intellectual, emotional, and behavioral activity.

4) Principle of activity. The essence of this principle is the conscious and active perception of the studied subject, its rationalization, creative processing, and application. It is aware of the bilateral relationship between the subjects in the learning process - the active and conscious position grows in awareness of the goals and activation of the cognitive activity of the trainees [8].

Objectives in the preparation of students for participation in programming competitions can be decomposed into intermediate and specific learning objectives so to provide a motivating pedagogical impact in planning and organizing the learning process. How the learning process in the preparation part is carried out is essential for increasing the motivation for adopting an active and conscious role of the students in education. The stimulation of mental activity and consciousness in the programming training is directly dependent on the methods and techniques used with proven activation value such as problem situations, practical tasks, situations suggestive of creative decisions. Activity and consciousness are the basis for creating a prerequisite for successful programming competitors.

5) Accessibility principle. Accessibility is the guiding principle for the didactic transformation of scientific information into the learning content of a subject. Accessibility requires full coordination between objectives, learning content; age, and mental characteristics of learners; methods and means of learning which enable knowledge, skills, and habits to be controlled by tensions of the forces for a regulated time [9]. The principle requires the presentation of scientific knowledge to be carried out systematically and consistently, well-reasoned, based on the planned schooling time and the psycho-physiological characteristics of the trainees. A specific problem with the application of this principle is the difficulty measure how to determine the optimal option. [9].

The following didactic rules formulated by Comenius in the Great Didactics as early as the 17th century are applied to successfully developing talented children in programming. [10]:

- from easy to difficult - a gradual transition from well-absorbed knowledge to learning knowledge that causes difficulty and requires more intellectual and emotional effort;
- from the known to the unknown - the learning of the new knowledge is based on the previous life experience theoretical knowledge and practical habits and skills;
- from the simple to the complex - begins with the absorption of more elementary concepts, facts, regularities and a smooth transition to more complex ones;
- from the close to the distant - the cognitive horizons are gradually expanding.

6) Principle of System and Consistency. The basis of this principle lies in the systematic nature of science. The scientific logic of the subjects studied, and the regularities in the development of the learners' thinking. Comenius postulates to go from facts to conclusions, from examples

to rules to expose the teaching material as a system. "This principle requires knowledge, skills, and habits to form a particular order in which each element of the material has relation to others" [9]. The principle of orderliness also manifests itself in all components of the training process in the field of informatics: orderliness towards learning objectives means a complete project of the professional qualities of the personality that is in line with the curriculum, fundamental and advanced in the knowledge and skills in the field of algorithms and specific programming language; establishing links between individual sub algorithms and methods for their creation to create a holistic view of the subject studied and to create a practical opportunity to solve tasks of a different nature; the educational technologies, forms, and methods designed in a rational and efficient system in a consistent logical order; the activity of the trainees is motivated and planned rationally, systematically and consistently to earn the knowledge and skills.

7) Principle of permanent knowledge and skills. The essence of this principle is to provide in-depth knowledge, skills, and habits. The goal is to create an opportunity for learners to use learning as a basis for future intellectual activity and solving practical tasks. "Permanent is that knowledge that retained for a long time in the students' minds and can be reproduced without much effort and applied in new different situations". [7].

8) The principle of individual approach. This principle concerns the fact that the process of acquiring knowledge and skills is strictly personal. Its application implies the differentiation of the teaching activity according to the individual characteristics of gifted children. Individualization is achieved toward approaches, methods, forms, and means of differentiating the teaching work: personal curriculum, cognitive complexity tasks, and more.

B. The Role of Teachers in the Extracurricular Schools of Informatics

Training in programming for students aged 9-14 is specific. It demands compliance with the age characteristics of children. The term "trainer" is more appropriate and more precise when it comes to extracurricular training. A trainer in extracurricular computer science can be everyone who has the task of transferring his or her knowledge and skills in informatics in useful for the trainee.

The proposed approach combines the described didactic principles and consists of the following: the trainer explains, demonstrates various algorithmic problems. The learners earn and train until they master the knowledge at their rate. Students with high interest in the field of informatics are involved in the training. They need to expand their knowledge in the field of informatics, acquire practical programming skills, and deeply study basic algorithms for working with complex data structures and gain experience in solving non-standard tasks. In this connection, the role of the teacher is crucial for both beginners and established contestants.

The systematic approach is needed to be expanded and deepened, taking into account the specifics of the informatics for young students. The focus is on rapidly basic programming knowledge acquiring. Pupils have to understand it by developing as many algorithms as possible to solve. The teacher has to select and arrange learning tasks. The end goal is competitors to master tasks close to the practice. It is not enough for them to reproduce known algorithms. Often students have to combine and modifying a set of already known algorithmic steps, as well as compiling new ones. Every programming competition requires a successful deal with complex formal objects. So the teacher has to train students to develop mathematical knowledge, logical, systemic, and algorithmic thinking. The initial preparation of small talents involves the acquisition of knowledge and skills of diverse aspects. Essential knowledge units are purely programming skills, mathematics, information technologies, and logic. Insufficient absorption of some knowledge leads to unwanted failures and disappointments at a later stage. The complex nature of the preparation can initially stress and overwhelm the students. For this reason, the teacher should give new knowledge very carefully for enough time. Figure 1 depicts the training steps in ascending order [11].

The first step is beginners to acquire excellent computer literacy. Students have to adhere to working with a personal computer and learn some commands of the operating system through game tasks. Early age is a period of functional literacy of children. Computer literacy has significant importance for future motivation. In the beginning, the teacher has the responsible task to show the students that the computer is used not only for games but also for work.

The second step is training in algorithms. Learners need to learn to create algorithms to solve the problem. An appropriate description of the algorithms will help them in writing the right programs. The teacher has the strategic role of selecting algorithmic problems. He has to arrange them in ascending cognitive sequence.

The third step is to create a relevant logical model of the program task. At this stage, the new knowledge should be presented in small portions, supported by appropriate examples.

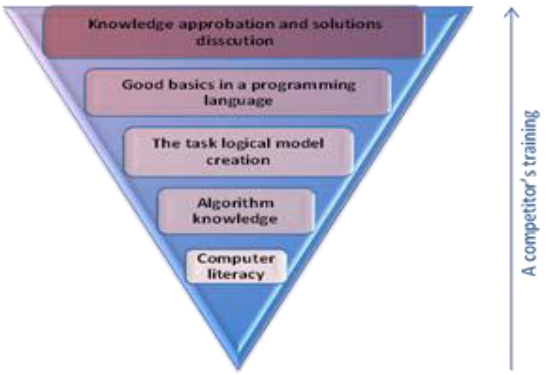


Figure 1. The training steps ascending order [11]

The analysis of the condition is essential here. The teacher explains that behind the words in the task description of each programming assignment, there is a logical model that they have to reach. The difficulty at this stage of learning is the building skills in the abstract thinking of students. Abstraction in programming appears to be at a higher level for young students. The teacher takes care of overcoming this problem by gradually increasing the level of abstraction. Answering the following questions helps students: What is there in the task description?, What is the required result in the task?, How could be obtained the result from the given data?

The next step in teaching informatics is to learn a programming language and the basic principles of programming. The proposed approach does not depend on the applied programming language.

Currently, the training focuses on the C++ language. The students have to master the syntax and semantics. They have to deal very well with the programming environment and the sequential steps to prepare a working program. The choice of the programming environment is determined by this, which is used in competitions in all age groups.

The teacher has the responsibility for learning a programming language to start in parallel with learning algorithms. This correlation helps the acquired algorithms to be programmed and immediately implemented and tested on a computer.

The teacher has an essential role in training future participants in informatics competitions to be able to detect and correct errors in their programs themselves. They have also to be able to assess the various possible solutions to a problem to choose the best option. For this reason, the last and most extensive step in the learning process is knowledge approbation and solution discussion.

C. Results

The new and different in the proposed approach for discovery, preparation, and development of gifted students in the field of informatics is the usage of didactic principles in determining the activities of the teacher and the division of work into stages. This definition ensures consistency in building skills and stability in the knowledge of competitors. There is a consistency in the results achieved by avoiding extreme loads and stress for the young students.

The proposed approach was developed, piloted, and implemented dynamically in a real learning process conducted in the School of Informatics at Center for Student, Technical and Scientific Creativity in the city Ruse. An important criterion for its qualities is the presentation of the students in the national computing competitions, in which they traditionally take prizes.

There are ten schools in Bulgaria for training competitors in informatics. The composition of the national team is formed by the first ten students in the maintained national rank list. They receive the right to participate in the National Campus of Informatics, organized under the auspices of the Ministry of Education and Science and the Union of Mathematicians in Bulgaria. Table 1 presents the number of participants from Ruse in these camp schools.

There is a quota of four participants for each country for the international competitions European Junior Olympiad in Informatics (EJOI) and Junior Balkan Olympiad in Informatics – JBOI. Since 2016, 25% of the participants in the national team are from the school in Ruse [12].

III. CONLUSION

There are different approaches to discovering, preparation and developing gifted students in the field of informatics. Unfortunately, programming is a wide-ranging area, often requiring complex intellectual efforts. Therefore, the organization of this type of training cannot be exposed in a condensed form as some pure "recipe training". Efficient success depends heavily on the teachers' activities in discovering and developing trainees. Teachers build up competitors' profiles - from his diligence and intuition to learning based on basic examples, from the creation of positive motivation, the rapid formation of lasting skills, and learning both for group work and self-study.

The results show that in the presence of ten schools, training in the proposed approach provides 25% of national competitors. The percentage of medals won by international competitions in the period 2017 - 2019 by competitors from the Ruse school compared to the total number of awards for Bulgaria is 16.67%. Our graduates have won 8 out of a total of 48 medals.

Each of the ten schools applies its approach to discovering, preparing, and developing of competitors in informatics. For this reason, the proposed approach cannot be compared with a similar one. Summaries about its effectiveness could be made from the results of the students.

The achievements of the competitors in informatics from Ruse prove the hypothesis that the application of the didactical principles in the learning process will give a chance of a teacher to increase the number of successful competitors.

Developing new approaches to training for beginner-child programming, verifying them in real-life conditions, and sharing experience gained to contribute to the successful realization of the tasks set out by the EU strategy.

The study of the role of the teacher in the discovery, preparation, and development of gifted students for competitions in informatics will be in-depth by detailing the teachers' actions in the selection of learning tasks and the evaluation of their solutions.

TABLE 1. Number of participants from Ruse in schools of MES and UMB

Year	6 grade	7 grade
2013	1	1
2014	0	2
2015	2	1
2016	1	2
2017	1	1
2018	1	1

REFERENCES

[1] Europe 2020 strategy, 2016. Available at: https://www.dges.gov.pt/en/pagina/education-and-training-2020-strategic-framework, [accesed: 07.05.2020]

[2] European Junior Olympiad in Informatics, 2017. Available at: https://ejoi.org/, [accesed: 07.05.2020]

[3] Garov, K. "Theory and practice of the preparation of prominent and talented students for participation in the Olympiads and competitions in Informatics and Information Technologies", Doctoral thesis, Sofia, 2008

[4] Garov K., Aneva St. "For the study of Informatics and IT in Elementary School, Mathematics and Mathematical Education", Proceedings of the XXXIV Spring Conference of UMB, Borovets, 2005

[5] Yovcheva, B. "Spiraling Approach in Programming Training for 10-11 Year Old Children", Doctoral Thesis, Shumen, 2009.

[6] Hristova, P. "Object approach for organizing extra-curricular work in informatics for students up to V grade", Doctoral thesis, Rousse, 2010.

[7] Petrov, P. Didactica, Veda Slovena - ZG, Sofia, 2002.

[8] Todorina, D and collective, Pedagogy, part III, Didactics, Blagoevgrad, IM Neofit Rilski, 2003.

[9] Stavreva, G. Methods of Teaching Biology, Plovdiv, MP Publishing, 2002.

[10] Comenius, J. (1907). The great didactic. London: A. and C. Black. Retrieved from https://archive.org/details/cu31924031053709/page/n12/mode/2u.

[11] Atanasova G., Hristova P. "Methodological aspects of the initial training of students for participation in programming contests", Proceedings BCI-AICT 2015 Balkan Conference on Informatics, Craiova, Romania, No Vol-1427, pp. 1-9.

[12] Web portal of the Team of Bulgarian Experts in Informatics, Available at: http://math.bas.bg/infos, [accesed: 15.05.2020]

Word Cloud Analytics of the Computer Science Research Publications' Titles over the Past Half Century

L. Abazi-Bexheti*, A.Kadriu * and M.Apostolova*
* South East European University/Faculty of Contemporary Sciences and Technologies, Tetovo, Macedonia
l.abazi@seeu.edu.mk, a.kadriu@seeu.edu.mk, m.apostolova@seeu.edu.mk

***Abstract* – The field of computer science is becoming increasingly important as almost every aspect of our lives is more and more dependent on computers and the Internet. As a result, computer hardware and software technology are rapidly increasing in power and sophistication. Consequently, the data and the research topics and publications in this field are also evolving rapidly.**

In this study we used the data form the Computer Science Bibliography DBLP database for the research publications in the field of Computer Science the past 50 years. Further on, we analyze the articles (journal publications) and proceedings as categories with larger numbers of entries and that provide more significant outcomes. The overall dataset is split in decades and for better demonstration of the information and the evolving topics, text visualization tool was applied.

The results generated by using word clouds provide a much faster and easier way to detect the topics and other changing details about the publications in each decade of the past half century.

Keywords - CS, word clouds, DBLP, Computer Science Research Topics

I. INTRODUCTION

The field of Computer Science is one of the most changing and evolving research fields. This change is empowered by even more changing technology supported with exponential growth of hardware capabilities and significant growth in software productivity as well. Consequently, almost every aspect of our lives is progressively influenced and enriched using technology.

On the other side, the data generated by systems, sensors, and other computers is increasing exponentially and the number of research papers published in Computer Science conferences and journals has increased constantly every year [3].

Usually these research papers are in the form of publications in conference proceedings or journal articles and they provide a starting point for many young scholars and researchers. In this regard, there are many publications that deal with the content analysis of a certain number of papers in a given field in a certain period of time and identify gaps or further paths for research [10][11][12]. However, due to the large number of data and publications it is difficult and time consuming to do an overall study of the evolution of the key topics in computer science research. The idea of this study was to track and show how the landscape of Computer Science research topics changed over time by analyzing 'the titles of research publications' in the past half century.

Based on our previous research where the Computer Science Bibliography DBLP data was used for quantitative structured literature review of research on e-Learning from its appearance till nowadays [6], in this study we broadened our research on the overall research publications in Computer Science [8].

In addition to this, for better demonstration of the information and the evolving topics, word cloud visualization tool was employed.

We considered that this representation will give to the reader a much faster and easier way to detect the topics and other changing details about the publications. By visualization of information we can see the connections and patterns that matter. Additionally, the information can be designed so that it makes more sense and gives the possibility to concentrate only on the important information. Our sense of sight is the fastest of the five senses and the eye is exquisitely sensitive to patterns in variations in color, shape and pattern. Combining "the language of the eye" and the "language of the mind" provide a powerful combination to gather insights from data [4].

The data for this study was gained through the DBLP. This service provides open bibliographic information on major computer science journals and proceedings. DBLP indexes more than 4 million articles and contains titles of articles, their authors, years of publication etc.

Word clouds are interesting to use as a visual aid with possibility to underscore the main topics. A word cloud is built from the words used in a text or subject, in which the size of each word indicates its frequency or importance. Lately, they are very popular, and the users find them very appealing. They provide a visualization form for text that is recognized for its aesthetic, social, and analytical values. On the other hand, there are several researchers that argue on the inadequacy and limitations of the word clouds. They mainly argue that there is a lack of natural reading order in how words are laid out; the font size is not appropriate for communicating quantitative information; the difference in word size might be due to word length rather than value and similar [2][9].

Taking into consideration all the ‘pros’ and ‘cons’ for the word cloud, for a dataset with over 4 million entries, we found word count as a very suitable visualization tool that provides a very interesting fast and informative venture through the key research topics in Computer Science. The readers will easily notice the larger, bold words and understand the main topics in certain periods.

II. METHODOLOGY

The DBLP dataset is downloadable and represents a very interesting resource for investigating the trends in terms of research fields, most productive authors as well as evolution analysis of various aspects of research publications [1][7].

In our analysis, we have created three datasets from the DBLP database, in Json format as shown in Figure 1, Figure 2 and Figure 3.

{"title":["Robotics - basic analysis and design."],"author":["William A. Wolovich"],"publisher":["Holt, Rinehard and Winston"],"isbn":["978-0-03-006119-6"],"year":["1987"],"pages":["393"]}

Figure 1. Part of Json file about books

{"title":["Parallel Integer Sorting and Simulation Amongst CRCW Models."],"author":["Sanjeev Saxena"],"year":["1996"],"journal":["Acta Inf."],"pages":["13"]}

Figure 2. Part of Json file about articles

{"title":["Concurrency and Concurrent Constraint Programming"],"author":["Ugo Montanari","Francesca Rossi"],"year":["1994"],"pages":["22"],"booktitle":["Constraint Programming"]}

These datasets contain data for three categories based on types of publications [8]:

- articles (journal publications) - counting 1,952,444 papers.
- conference proceedings - counting 2,361,456 papers.
- books - counting 15,617 books.

Since the focus of our analysis is identifying research trends in the field of Computer Science, we have continued our analysis only on the first two categories: articles (journal publications) and proceedings, which have a much larger number of entries and hence would provide a more significant outcome.

Further on, for each category separately, a word cloud visualization method is used. From the dataset records, for each paper, only the title and the year of publication is extracted. In this way a list of paper titles for each decade is created. From the words used in the titles further on word cloud is created where each word is associated with a frequency value. Also, visual representation of the information is created. We used the representation in which words are positioned in an unordered layout, to optimize the use of space, and frequency values are mapped to font size. Also, different colors are assigned to the words in order to create more visual clouds which turn to be more acceptable by the users.

III. ANALYSIS

A. Analysis of the journal articles dataset

In this part is presented the study of the dataset of articles (journal publications). As mentioned above, this dataset contains 1952444 records distributed from 1950 to 2019. For each decade the word cloud technique is applied and below are presented the results for the past seven decades.



Figure 4. Word cloud of titles of journal articles through the time period 1950-1959 (19 records)

As it can be seen in Figure 4, in the initial stage of journal articles in computer science dominates the term analog, computers and computational methods. As we move toward the next decade 1960-1969, we can notice a significant increase in the number of publications. The number is doubled and the word ‘programming’ and ‘programming languages’ gain focus (Figure 5).

Also, the word ‘contribution’ appears in the title of the publications more and more, which indicates the application and the contribution of computer science studies in various fields.



Figure 5. Word cloud of titles of journal articles through the time period 1960-1969 (40 records)

In the next decade, 1970-1979, the number of research papers was four times larger compared to the previous decade and almost 10 times more compared with the initial decade (Figure 6). These numbers speak a lot of the rapid and exponential growth of computer science research and for evolving topics of research. In some decades there is a concrete distinction of the main words (topics) while in some there is a variety of topics with similar frequency but still contribution and applications remain the focus of the research. Another interesting aspect of the word cloud presentation is that besides the key main topics, one could identify some new significant trends that are appearing in the research such as the word ‘digital’ or ‘systems’.



Figure 6. Word cloud of titles of journal articles through the time period 1970-1979 (186 records)

As we continue the journey through the next decade (1980-1989), the number of the records reaches 897 (Figure 7). In this decade, we see the continuation of the ‘programming’ topic as the focus of the research but at the same time additional topics such as ‘systems’ and ‘design’ are gaining weight as well.



Figure 7. Word cloud of titles of journal articles through the time period 1980-1989 (897 records)

In the next decade (1990-1999), with 3709 number of records we see the topic ‘systems’ as the main topic but also topics such as: ‘software’, ‘engineering’, ‘networks’ and ‘design’ are gaining importance in the research journals (Figure 8). Another interesting aspect that can be noticed is that the publications in German language are taking a notable place in the field of computer science research.



Figure 8. Word cloud of titles of journal articles through the time period 1990-1999 (3709 records)

Further on, in the first decade of this century (2000-2009), the number of records is 3999 (Figure 9). The research topics are more concrete and narrowed; ‘java’, ‘linux’, ‘windows’, ‘oracle’ and similar. Also, topics such as ‘web’, ‘data’ and ‘learning’ are gaining attention of the researchers. Once again, through the word clouds we see the initiation of the research topics that are

becoming the main topics of research in the next years to come.



Figure 9. Word cloud of titles of journal articles through the time period 2000-2009 (3999 records)

In the last decade (2010-2019) are included the records form the DBLP till the date that it was downloaded and further on analyzed (Figure 10). We see that the topic 'systems' is the most frequent word on the research articles titles, showing the growth of interdisciplinary dimension of computer science and its applications. Also, we see that topics such as 'web', 'models, 'design' continuing to be important but at the same time new topics such as 'intelligence', 'social', 'security' and many others are also entering the stage of research in computer science articles.



Figure 10. Word cloud of titles of journal articles through the time period 2010-2019 (3752 records)

B. Analysis of the conference proceedings dataset

Another aspect of our research was to analyze the part with conference proceedings, which is a much larger dataset and can also lead to interesting results.

As it can be seen in the following word clouds the conference proceedings publications provide a step ahead in terms of research topics compared with journal articles. This is mainly due to the fact that the review and publication time of conference publication is usually much shorter compared to the journal publications.

The first decade dataset for conference publications starts in 1960-1969 and as it can be seen it has main 'topics' of the research such as 'systems' appearing much earlier than in journal articles.



Figure 11. Word cloud of titles of conference proceedings through the time period 1960-1969

In the next decade (1970-79), in addition to the topic 'system', the topics such as "database', 'data model', 'design' and 'simulation' are gaining weight while other topics are fading down (Figure 12). For instance, 'language', 'automatic' and other topics are lighter compared to the previous decade.

.



Figure 12. Word cloud of titles of conference proceedings through the time period 1970-1979

In the 1980-1989 decade, 'algorithm design' is a new tendency as well as the 'usability' is becoming a new keyword in the field (Figure 13) and continues as such in forthcoming decades (Figure 14, Figure 15). From the last two figures the perception is that in the latter two decades the keyword 'system' starts to lose weight. Usability continues to be an important part of the research, together with 'based' and 'model', which indicate that nowadays CS research is focusing more on applicability in interdisciplinary fields, with solutions

based on concrete models. Also, we can notice a variety of topics, with a heterogeneous frequency of appearance which once again show that the computer science research is entering in many aspects and dimensions. In the last decade fields 'neural networks' and 'evaluation' increase their relevance.



Figure 13. Word cloud of titles of conference proceedings through the time period 1980-1989



Figure 14. Word cloud of titles of conference proceedings through the time period 1990-1999



Figure 15. Word cloud of titles of conference proceedings through the time period 2000-2009



Figure 16. Word cloud of titles of conference proceedings through the time period 2010-2019

IV. Conclusion

Taking into consideration the rapid growth of the number of publications in one hand and the rapid change of the computer science field on the other hand, the aim of this study was to provide a journey through the changing topics in the research publications in computer science.

Through the word cloud visualization, the study brings to the user an easy and fast way to get an insight of the key topics in the field for a certain decade in thousands of papers.

The results summarized above lead to several conclusions as presented above for each decade separately.

An appealing outcome for instance was that although the journals in terms of publications are more relevant, in terms of identifying the trends the conference proceedings seem to offer a step ahead. This is a very important result and guide especially for young researchers who are in the stage of identifying their research objectives and current research gap.

With the provided analysis it can be also concluded that computer science research is becoming very wide and diverse, giving solutions to different problems in many other disciplines.

References

[1] A. Kadriu, L. Abazi-Bexheti, H. Abazi-Alili, and V. Ramadani, "Investigating trends in learning programming using YouTube tutorials", International Journal of Learning and Change, 12(2), pp. 190-208, 2020.

[2] D. McCandless. "The beauty of data visualization", 2020. Available at: https://www.youtube.com/watch?v=5Zg-C8AAIGg, [accesed 02.03.2020].

[3] E.Y. Gorodov and V.V. Gubarev, "Analytical Review of Data Visualization Methods in Application to Big Data", Journal of Electrical and Computer Engineering, 2013.

[4] Introworks, "Language of the eye", 2020. Available at: https://intro.works/news-insights/language-of-the-eye-2/, [accesed 20.02.2020].

[5] J. Heer and B. Shneiderman, "Interactive dynamics for visual analysis", Communications of the ACM, vol.55, no.4, 2012.

[6] L. Abazi-Bexheti, A. Kadriu and M. Apostolova, "Quantitative structural literature review in e-Learning", 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2017.

[7] L. Abazi-Bexheti, A. Kadriu and M. Apostolova, “Self learning trends in the field of introductory programming”, In 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 0715-0719, IEEE, 2018.

[8] L. Abazi-Bexheti, A. Kadriu and M. Apostolova, “Investigating Gender Gap in Computer Science”, Research Community, MIPRO 2019, Opatija, Croatia, 2019.

[9] L. Combs and P. Roman, “Beyond the word cloud visualizing text with Python”, District Data Labs. Available at: https://www.districtdatalabs.com/beyond-the-word-cloud, [accesed 11.03.2020].

[10] P.S. Lee, J.D. West, and B. Howe, “Viziometrics: Analyzing Visual Information in the Scientific Literature”, IEEE Transactions on Big Data 4(1), pp. 117–129, 2018.

[11] R. Agrawal, A. Kadadi, X. Dai, and F. Andres, “Challenges and opportunities with big data visualization”, 7th International Conference on Management of computational and collective intelligence in Digital EcoSystems, ACM, pp. 169-173, 2015.

[12] S. Julie and N. Iliinsky, “Beautiful Visualization: Looking at Data Through the Eyes of Experts”, Cambridge: O’Reilly, 2010.

Using WebIDE as a distance learning tool for high school programming

Mirza Mesihović* and Vedran Ljubović* and Ibrahim Muharemović*
*Faculty of Electrical Engineering, University of Sarajevo, BIH

***Abstract*—During COVID-19 pandemics, not unlike their peers in other parts of the world, the high-school students in Bosnia and Herzegovina needed to get accustomed to learning from their homes. This proved to be particularly difficult for those who were scheduled to learn basic C++ programming, not only the in-person support by their teachers is invaluable in these early stages of learning, but many of them did not have access to adequate hardware or software. A web-based integrated development environment (WebIDE) was deployed, and a number of sample assignments were created. Altogether, over 130 students from 5 schools participated actively, solving a total of 65 assignments (as of June 1st). Based on questionnaire results, overall the experiment was deemed successful, and the continued use of WebIDE after life returns to normal was justified.**

***Keywords* -** distance learning; webide; introductory programming; K-12 education; COVID-19

I. Introduction

Before the COVID-19 pandemics, Bosnia and Herzegovina introduced digital technologies in education fairly slowly, especially in the secondary education. In the recent years there has been a considerable push to introduce the ICT to children in elementary schools (which in B-H lasts 9 years, typically including children between 6 and 14 years of age), especially in the Sarajevo Canton. Starting in 2016, the subject titled "Informatics" was included in the elementary school curriculum throughout the nine years. This subject is mostly focused on the appropriate use of computers, mobile devices and the Internet, various applications, covering topics such as critical evaluation of online content, cyber-bullying, programming etc. [1]

Since 2014, a number of elementary schools were equipped with educational tablets through "The School of Future" program. Starting in 2017, the relevant legislation of the Sarajevo Canton require that all schools that have adequate equipment should perform one week per school-year as an "online week" during which students stay at home and learn using the Internet. A variety of online platforms were used in schools, including Office 365, Google Classroom, and Moodle. The success of this pioneer project in the first year varied from school to school [2].

As expected, higher education institutions (HEI) were the most advanced in adopting digital technologies, and several schools have approved Distance Learning (DL) curricula, so several of them offered assistance with DL technology to elementary and high schools during the COVID-19 pandemics. One of the greatest problems that emerged during the pandemics was related to a number of students not having access to computers with adequate performances to use the prescribed online tools. Various organizations, groups and individuals scrambled to organize donations of PCs to school children.

In Canton Sarajevo, most high schools also have a subject titled "Informatics", although it has only one lesson per week allotted, which in many schools includes C++ programming as part of its syllabus. Gimnazija – the comprehensive education high-school offers a broad education in a number of subjects, preparing students for enrolling into an university, and its second-grade curriculum includes C++ programming. In addition, "Information technology" programme has been offered by several schools starting in 2017 with a curriculum that includes more lessons allotted for programming. Schools are plagued with shortage of equipment, inadequate and outdated textbooks, and a lack of qualified teaching staff necessary to fulfill this curriculum.

Web-based integrated development environment (WebIDE) is a technology that has become mature in the last 5-10 years. It is well known that WebIDE offers certain advantages to educators. Installation and maintenance of computer labs is simplified, the only software that is required is a modern web browser.

Traditional integrated development environments (IDE) are fairly resource intensive, which presents a problem for students who can't afford powerful computers, while WebIDEs can even be used (although not comfortably) from mobile devices such as tablets or phones. Students can access the exact same environment from home, school, or a cafe. Most importantly, while IDEs are usually professional tools that are not easy to learn or friendly towards novice programmers, WebIDE can be customized with various extensions for novice programmers in a way that is easy to develop, evaluate and deploy to a large number of students.

All this considered, when the in-class education was suspended due to the COVID-19 pandemics, a WebIDE platform previously in use at a Sarajevo HEI was offered to high-school students. The immediate goal was to assist those schools where C++ programming curriculum was interrupted due to students inability to use computer labs. Schools were not instructed on how to use the platform, but most used it as a tool for distance learning, as opposed to remote learning. Due to administrative procedures, this platform became available only on May 1st and was used until June 1st. A total of 137 students from five schools logged into the platform and used it to complete a number of programming assignments prepared by their professors.

There is some existing research on the use of WebIDE in higher education, however usage of WebIDE in K-12 education is insufficiently studied. This paper documents the experiences of using a WebIDE platform in high schools in

Sarajevo and attempts to fill this void in research.

II. Related work

Learning by doing is an integral part of modern CS education [3], [4]. This effort is thwarted by professional programming environments, which are invaluable tools in the hands of professionals, but to novice programmers they can often be overwhelming. A multitude of features and options can often present an excessive cognitive load for students, while the information about what is actually happening in the code is either missing or difficult to interpret. This was the rationale behind the decision to try and develop an educational IDE. Seminal work in this regard are BlueJ [4] and GILD [5], while more recent efforts include greenfoot [6].

In the recent years, improvements in web standards, web development technology and consumer hardware have made it possible to create an integrated development environment as a web application executing in a web browser. This technology was developed primarily to simplify application development in scenarios that would otherwise require complex cross-compiling and deployment strategies, such as embedded development (Raspberry Pi) and cloud development (Amazon AWS Lambda) [7].

However, the research in the area of computer-assisted learning technologies shows that WebIDE can be useful in programming education, and in particular to enforce Test-Driven Learning (TDL), where regular programming assignments are supplemented with tests to verify if student solution is correct. Research shows that groups of students who were taught using TDL show better results on tests and quizzes, and a more positive attitude towards the course [8], [9], [10].

Another topic that was studied in literature is use of the WebIDE to promote formal reasoning and specification in the process of software development [11]. The software development process is inherently iterative, and WebIDEs give an opportunity to provide instructive feedback in the form of verification conditions (VCs) at each iteration [11] which is infact also closely related to TDL.

Using a web-based environment opens various opportunities for collaboration. Computer-Supported Collaborative Learning (CSCL) is an active field of research, however in the past it was mostly focused on education in social studies. Early attempts to use CSCL tools such as wiki, forum and chat in programming education have shown limited success [12], [13]. An interesting idea that was well-received is usage of an audio discussion feature [14], [13].

A more recent approach is a shared editing workspace where two or more students work on the same code or can see solutions by other students in real-time [15], [16]. While it is shown in literature that student satisfaction and attitudes towards such systems are generally positive, and the resulting code is of higher quality on average, it is difficult to show an improvement in programming performance of individual students [17]. Students who are used to working in a collaborative environment might show a decreased performance in individual tests.

On the other hand, there were other innovative ideas for collaboration. Codeopticon tool allows a tutor to see in real-time the coding workspace of many students and help them using a text-based chat [18]. These workspaces are read-only to tutor. Another useful feature is logged playback (also read-only) where the tutor can see how student progressed towards the finished task and which errors they made [19]. A common problem with text-based collaboration tools such as chat is difficulty in expressing the exact location in code where the problem exists. This, too, can be addressed in WebIDE [19].

A very common problem among novice programmers is their difficulty to understand various error conditions, either because error messages are confusing [20], [8], or due to inability to visualize the current program state. Some degree of visualisation can be provided by debugger tools, however they are usually very difficult to use. Our experience indicate that a low percent of students use debugger despite receiving a detailed instruction on its use. There is a wealth of research on code visualization techniques in programming education, and even collaborative visualization [21], however it would be beneficial if such tools were integrated with the development environment.

There is some anecdotal evidence that web-based environments are more popular among female students [10].

Škorić et al. [22] evaluate several available WebIDEs for use in an introductory programming course. CodeBoard was found to be the best. It should be noted that this research was published in 2016, several tools have emerged since that provide a much higher level of interaction.

III. Description of implemented system

Starting in 2015, a WebIDE platform was implemented at a course in introductory C and C++ programming delivered on a HEI in Bosnia and Herzegovina. To reduce the development costs, the institution opted to implement an existing open-source solution as oposed to devloping its own. The initial research showed that none of the available WebIDEs provided all of the necessary features. However, Cloud9 IDE was selected as the most full-featured solution, in particular in regards for its C/C++ support.

The open-source edition of Cloud9 was a single-user software, so development of a new platform was started providing the missing support for multiple users, as well as an interface for teachers. This infrastructural software was developed in PHP, and its features include:

- User management (integrated with institutional LDAP server)
- Sandboxing – limiting users so they can't access others environment without permission
- User monitoring and remote assistance
- Statistical information on usage
- Course management

Unfortunately, in the intervening time the Cloud9 project was acquired by Amazon, close-sourced, and further development on the project all but seized. For these reasons, current

work is mostly focused on porting the existing work to Theia project (http://theia-ide.org).

A. System architecture

The system currently has over 2000 registered users. At any moment, up to 300 users can be active simultaneously (note that users are automatically logged out after a certain period of inactivity). For each user, an instance of (single-user) Cloud9 environment is executed in a sandbox.

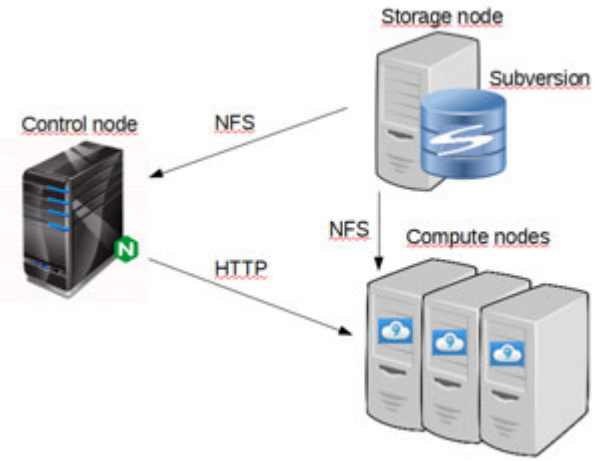


Fig. 1. WebIDE distributed architecture.

To facilitate this, a distributed architecture was developed specifically for WebIDE (Fig. 1). WebIDE cluster consists of one control node, one storage node and a number of compute nodes. Control node is the only publicly accessible node. It contains an nginx server which serves the dashboard and teacher interface (using php-fpm module), and also serves as a proxy server towards compute nodes using HTTP.

Compute nodes execute user Cloud9 instances which are served by nodejs server. By increasing the number of compute nodes, one can scale the system for a larger number of users.

Finally, all user files and configuration reside on storage node. These files are shared with compute nodes and the control node using NFS protocol. Storage node also holds monitoring service which tracks every file change and commits it to Subversion server. Cloud9 is configured to autosave all changes as they are typed, meaning that the system creates detailed usage traces, sometimes down to individual keypresses. Users also have Git repositories at their disposal where files are commited once each day.

B. Student user interface

Upon logging in, users are greeted by a dashboard showing system operational status and news. This also provides a distraction while user Cloud9 instance starts in the background. This dashboard includes a one-on-one chat using a free "website support" tool. For the duration of COVID-19 experiment, volunteer university students provided assistance to high-schoolers using the platform. In addition, various schools used tools such as Microsoft Teams and Google Hangouts to create chatrooms where students could discuss assignments with their teachers.

A number of Cloud9 plugins were developed using its native JavaScript API:

- Assignment plugin
- Send homework plugin
- Processing compiler output to provide a more user-friendly display
- Considerable work was done on improving debugger support and making it more user-friendly
- Integration with Valgrind profiler, offering dynamic (run-time) code analysis and detection of errors such as uninitialized variables or memory leaks
- A plugin for unit testing using our custom software testing platform
- Help plugin, which integrates with C and C++ documentation



Fig. 2. Student view of the assignment plugin.

Of greatest interest is the assignment plugin (Fig. 2), which provides a task-oriented view for the students as an abstraction of the actual filesystem. The assignments are easier to navigate since they are hierarchically organized by subjects, and then chronologically. Upon opening the assignment for the first time, a default code is generated from template, allowing half-finished assignments where students are required only to provide a portion of code or code snippet.

C. Teacher user interface

Teachers are presented with a list of courses, followed by a list of student groups within the course. For each group a window with detailed statistics is presented (Fig. 3). These statistics include: estimated time that each student spent solving an assignment, number of compilations and number of tests.

By clicking on the cell corresponding to a student and an assignment, teachers can see the exact code that a student is working on currently, with real-time updates [18]. Teachers can also replay the log and see how students' codes evolved through time [19], and there is a textual log which is color-coded by types of actions performed by students (Fig. 4).

Teachers can create assignments by providing an assignment text, template code and tests. All of these fields are optional.

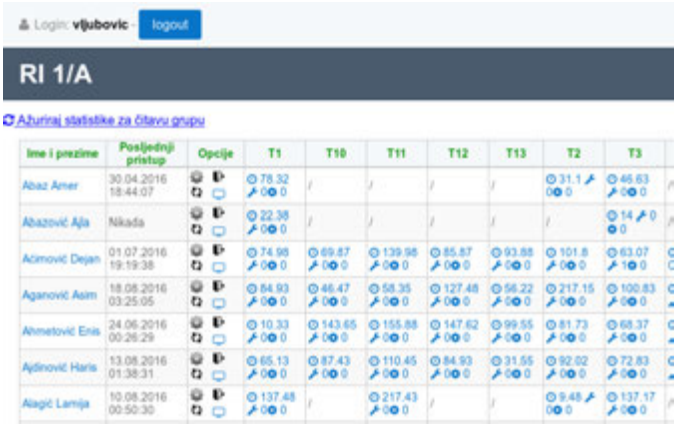


Fig. 3. Teacher user interface showing statistics per student and assignment.

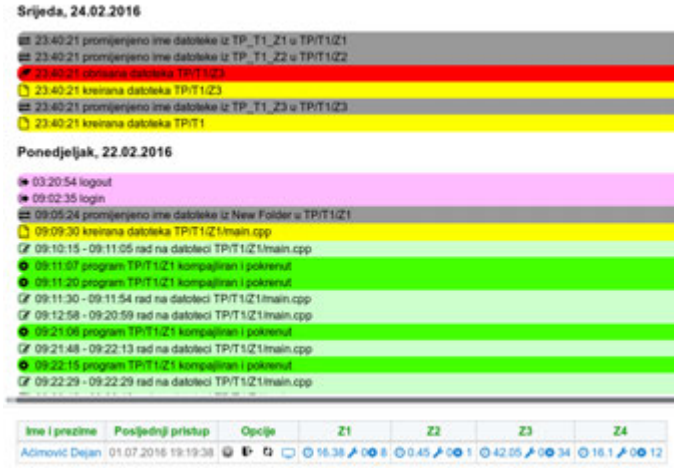


Fig. 4. Textual log of student activity on a file.

IV. Evaluation of the System and Analysis

Between May 1st and June 1st, 2020, the system was used by 137 students from five schools, most of whom are in the second grade (15-16 years old) and units taught introductory C++ programming. It should be noted that the whole idea was not implemented earlier due to a number of administrative obstacles. The organization of online education based on the traditional IDEs for those who had the technical capacity to use them was already underway in most schools.

That was why it was impossible to assess the knowledge of students and their progress in programming performance. Instead, students and teachers were asked to complete a questionnaire that included a selection of questions providing an opportunity for self-assessment, and also to gauge their attitudes towards the system. A total of 89 students and five teachers completed this questionnaire.

The questionnaire includes seven questions for students and seben for teachers, each with answers on the Likert scale, as well as a textual field for comments. The questions facilitate both summative and formative evaluation [23], but also emphasize that the system could be improved in the future.

Before the analysis of the results, it should be said that the platform was offered as-is, with no particular instructions on how it should be used. The teachers opted to use it mostly in a distance learning scenario (as opposed to remote learning), that is, for the duration of an online class all of the students were expected to log in and work on assignments, and the students could continue working at a later convenient time.

Teachers were given some instruction on how to create assignments, but the time was too short to instruct them on how to create tests. There were a total of 65 teacher-created assignments. In addition, the platform was preloaded with a selection of material from university teaching experience (only those assignments that were adequate for high-school level), giving another 40 assignments with tests. Since this second group of assignments was not mandatory, less than 10% of students attempted to solve them.

The attitudes of students towards the system were mildly positive. Most students agree (but not strongly agree) that the assignments were easier to solve through the platform than they would be without it (Fig. 5), although over 40% strongly agree that the provided instruction materials and videos were sufficient for them to learn how to use the system (Fig. 6). Likewise, there is a very small bias of students agreeing to questions “The platform helped me learn more about programming” and “The platform motivated me to independently research and study”.

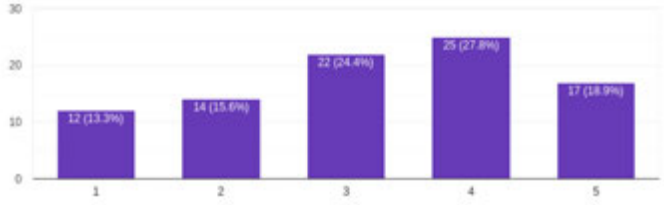


Fig. 5. Answers to the question “It was easier to solve the assignments using the platform than it would be without it” (student questionnaire).

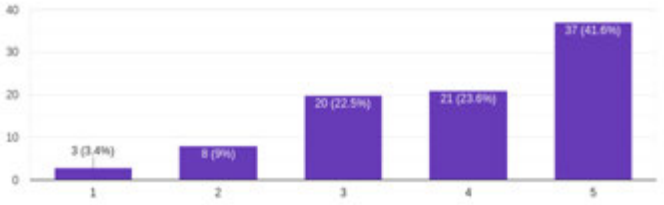


Fig. 6. Answers to the question “Provided instructions and videos were sufficient for me to learn how to use the platform” (student questionnaire).

It should be understood that a large percent of students saw the new platform introduced this late in the semester as an additional burden and not as a tool to help them learn.

Only seven students entered text into the free-form text field. Their biggest complaint was the confusing error messages,

which corresponds well to literature. Apparently, the C++ compiler on the server is more strict in some aspects than the one they used in school, so some language constructs that they were taught did not work.

Since only five teachers completed the questionnaire, it is difficult to analyze it statistically. Nevertheless, 4 out of 5 teachers strongly agree that the use of platform such as this could facilitate improvements in curriculum (the remaining teacher agrees). Teachers were divided on the issue of whether the platform helped the students learn more about programming.

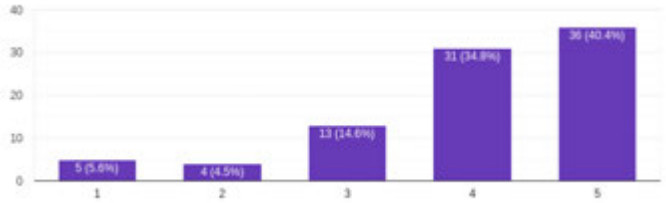


Fig. 7. Answers to the question "Automated testing is useful for learning programming" (student questionnaire).

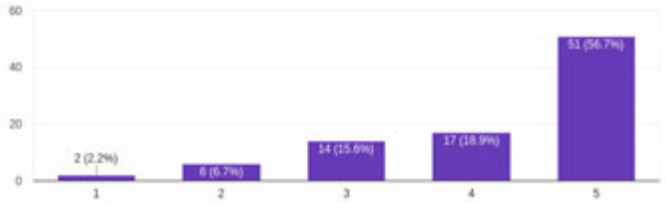


Fig. 8. Answers to the question "It was helpful to have someone available for questions" (student questionnaire).

Most students and teachers strongly agree that automated testing is useful learning and teaching in programming (Fig. 7). In addition, students agree very strongly that online help via chat tool was useful (Fig. 8) which somewhat contradicts the literature. A student and a teacher wrote in their comments that it would be better to have classical chat rooms than one-on-one chat.

Both students and teachers agree that it would be good to have such a tool in informatics education. Two of the teachers provided a detailed list of suggestions for teacher interface that will certainly help to improve its usability.

V. Conclusion and Future Work

During one month of lockdown caused by COVID-19 pandemics, a WebIDE tool was used in high-school education, for introductory C++ programming. Results were analyzed from the perspective of system use patterns as well as a questionnaire for self-assessment.

The results indicate that WebIDEs are highly useful in elementary programming education. Test-driven learning (TDL) approach was welcomed by students and teachers alike, although the time was insufficient to use it to full extent. Online textual chat as a form of help is likewise useful. The few negative attitudes can be attributed to the overall depression of the epidemics as well as the novelty of the tool being introduced so late during school-year.

Teachers and students alike believe that a tool such as this should be used throughout the school-year, allowing them to get used to it.

Future work should be focused on improvements in two areas: clearer and more helpful error messages [8], and better visualization of program state (possibly by improving the debugger [20]). A community-based repository of content should, over time, provide a wealth of assignments with tests [10]. The work should continue on the chat tool, with focus on the ability to share code and point to specific locations in code [19]. Usage of eye-tracking is another promising research direction [24].

References

1 "Nastavni planovi i programi za predmet Informatika za osnovnu školu (I, II, III, IV, V, VI, VII, VIII i IX razred)," 8 2019. [Online]. Available: https://mon.ks.gov.ba/sites/mon.ks.gov.ba/files/npip_informatika_za_os_2019.pdf

2 Ražanica, M., "Eksperimentalna online konfuzija," *Školegijum*, 2017. [Online]. Available: http://www.skolegijum.ba/tekst/index/1409/eksperimentalna-online-konfuzija

3 Lahtinen, E., Ala-Mutka, K., and Järvinen, H.-M., "A study of the difficulties of novice programmers," *Acm sigcse bulletin*, vol. 37, no. 3, pp. 14–18, 2005.

4 Van Haaster, K. and Hagan, D., "Teaching and learning with BlueJ: an evaluation of a pedagogical tool." *Issues in Informing Science & Information Technology*, vol. 1, 2004.

5 Storey, M.-A., Damian, D., Michaud, J., Myers, D., Mindel, M., German, D., Sanseverino, M., and Hargreaves, E., "Improving the usability of Eclipse for novice programmers," in *Proceedings of the 2003 OOPSLA workshop on eclipse technology eXchange*, 2003, pp. 35–39.

6 Kölling, M., "The greenfoot programming environment," *ACM Transactions on Computing Education (TOCE)*, vol. 10, no. 4, pp. 1–21, 2010.

7 de Farias, C. M., Júnior, P. G., Pereira, M. V., Brito, I. C., dos Santos, I. L., Pirmez, L., Delicato, F. C., and Carmo, L. F., "Web2Compile-CoT: A Web IDE for the Cloud of Things," in *International Conference on Internet and Distributed Computing Systems*. Springer, 2015, pp. 19–30.

8 Dvornik, T., Janzen, D. S., Clements, J., and Dekhtyar, O., "Supporting introductory test-driven labs with WebIDE," in *2011 24th IEEE-CS Conference on Software Engineering Education and Training (CSEE&T)*. IEEE, 2011, pp. 51–60.

9 Hilton, M. and Janzen, D. S., "On teaching arrays with test-driven learning in WebIDE," in *Proceedings of the 17th ACM annual conference on Innovation and technology in computer science education*, 2012, pp. 93–98.

10 Janzen, D. S., Clements, J., and Hilton, M., "An evaluation of interactive test-driven labs with WebIDE in CS0," in *2013 35th International Conference on Software Engineering (ICSE)*. IEEE, 2013, pp. 1090–1098.

11 Kabbani, N. M., Welch, D., Priester, C., Schaub, S., Durkee, B., Sun, Y.-S., and Sitaraman, M., "Formal reasoning using an iterative approach with an integrated web IDE," in *Proceedings of the 2nd International Workshop on Formal Integrated Development Environment (F-IDE 2015)*, 2015.

12 Guzdial, M., Ludovice, P., Realff, M., Morley, T., and Carroll, K., "When collaboration doesn't work," in *Proceedings of the International Conference of the Learning Sciences*, 2002, pp. 125–130.

13 Bravo, C., Duque, R., and Gallardo, J., "A groupware system to support collaborative programming: Design and experiences," *Journal of Systems and Software*, vol. 86, no. 7, pp. 1759–1771, 2013.

14 Ghorashi, S. and Jensen, C., "Integrating collaborative and live coding for distance education," *Computer*, vol. 50, no. 5, pp. 27–35, 2017.

15 Goldman, M., Little, G., and Miller, R. C., "Real-time collaborative coding in a web IDE," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 2011, pp. 155–164.

16 Warner, J. and Guo, P. J., "CodePilot: Scaffolding end-to-end collaborative software development for novice programmers," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 1136–1141.
17 Nguyen, V., Dang, H. H., Do, N.-K., and Tran, D.-T., "Enhancing team collaboration through integrating social interactions in a web-based development environment," *Computer Applications in Engineering Education*, vol. 24, no. 4, pp. 529–545, 2016.
18 Guo, P. J., "Codeopticon: Real-time, one-to-many human tutoring for computer programming," in *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, 2015, pp. 599–608.
19 Vandeventer, J. and Barbour, B., "CodeWave: a real-time, collaborative IDE for enhanced learning in computer science," in *Proceedings of the 43rd ACM technical symposium on Computer Science Education*, 2012, pp. 75–80.
20 Efopoulos, V., Dagdilelis, V., Evangelidis, G., and Satratzemi, M., "WIPE: a programming environment for novices," in *Proceedings of the 10th annual SIGCSE conference on Innovation and technology in computer science education*, 2005, pp. 113–117.
21 Moreno, A., Myller, N., and Sutinen, E., "JeCo, a collaborative learning tool for programming," in *2004 IEEE Symposium on Visual Languages-Human Centric Computing*. IEEE, 2004, pp. 261–263.
22 Škorić, I., Pein, B., and Orehovački, T., "Selecting the most appropriate web IDE for learning programming using AHP," in *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2016, pp. 877–882.
23 Mark, M. A., Greer, J. E. *et al.*, "Evaluation methodologies for intelligent tutoring systems," *Journal of Artificial Intelligence in Education*, vol. 4, pp. 129–129, 1993.
24 Obaidellah, U., Al Haek, M., and Cheng, P. C.-H., "A survey on the usage of eye-tracking in computer programming," *ACM Computing Surveys (CSUR)*, vol. 51, no. 1, pp. 1–58, 2018.

Ability of the Information Science teachers to teach programing in the lower grades of primary school

Janko Žufić* and Aleksandra Žufić**
* Juraj Dobrila University of Pula, Faculty of Educational Sciences, Pula, Croatia
** Elementary school Veli Vrh Pula, Pula, Croatia
jzufic@unipu.hr, aleksandra.zufic1@skole.hr

***Abstract* - In anticipation of the introduction of the elective subject of Information Sciences for the lower grades of primary school, a question of the professional teaching competences of students - future teachers and current teachers of Information Sciences (with an emphasis on programing), comes into the focus. This article discusses their ability from professional (knowledge of programing) and pedagogical aspects. We present the results of a research conducted among students of the Faculty of Education Sciences with a module in Information Sciences and Informatics Teachers in the Class Teaching study group. Respondents, through self-evaluation of their knowledge and competencies in the field of programing, gave answers from which it can be concluded whether they are ready to teach programing. The article also presents a case study in which students of the final year of the Faculty of Education Sciences with a module of Information Sciences, made a video clip of processing new contents for a programing class, as a methodical exercise. The programing languages used were Scratch, Python and the micro:bit simulator. Project assignments were correlated with the subjects of Croatian language, Physical and Health Culture, Music Culture, Mathematics and Information Sciences. The results of the research indicate that the respondents are generally ready, but they would need additional education in programing domain.**

Keywords: programming, Scratch, classroom teaching, computer thinking and programming

I. INTRODUCTION

The decision to adopt the curriculum for the subject of Information Sciences for primary and secondary schools in the Republic of Croatia was published in the Narodne novine on March 6, 2018 [1]. This decision has already been in effect for the school year 2018/2019 and has been applied to the students from the 5th grade of primary schools through the 4th grade of secondary schools. For the students of the lower grades of primary schools, the curriculum will be applied in the school year 2020/2021. However, the introduction of the initial curricula for the subject of Information Sciences as an elective subject in the lower grades of primary schools raises a question of the professional competences of our teachers of Information Sciences, students of the final year of the Teacher education - Faculty of Education Sciences with the module of Information Sciences and students of Information Sciences – Class Teaching study group.

Today's children are often called the *digital natives*. They have no fear of new technologies or getting information from online sources, and they like to learn "harder, happier and more serious things" [2]. Previously, teachers of Information Sciences mostly taught in the upper grades of primary schools or in secondary schools, and only few of them have some experience in working with younger students. In this paper, we were interested in the opinion of teachers of Information Sciences about their own pedagogical and professional competencies that they have acquired so far. Do they have sufficient competencies, or do they need additional education?

Information Technology should have a positive effect on the quality of life and not have a purpose of exclusively teaching specific skills. [3] The Curriculum sets teaching and learning outcomes of the subject of Information Sciences as development of critical thinking, creativity and innovation through the use of information and communication technology, and development of computational thinking, problem-solving skills, and programing skills [1]. At the same time, programing is seen in a broader context than the creation of computer programs, i.e. as a way of solving problems that are practiced with the help of computers, and can be transferred to areas outside the boundaries of computing. Therefore, children should have the opportunity to learn programing, not necessarily to become programmers, but to practice problem-solving. [4]

The activities and contents used should correspond to the cognitive stage of the child's development. The younger the students, the more appropriate it is to approach them with concrete examples and tangible contents, such as computer games without the use of computers, the so-called unplugged activities. In the next phase of learning, it is appropriate to move on to managing images of objects. It is appropriate to use visual programing languages in which commands are used to manage events in the micro-world. Coding in abstract programing languages such as Python, C++ and others is considered the final stage of learning programing.

We distinguish two types of visual programing languages in which images of objects are managed in the micro-world. Some are visual-textual, like Logo, and others are visual block-based, like Scratch, Alice, Game Maker, Blockly and the likes. The basic advantage of block-based programing languages is the significantly reduced possibility of syntax errors, which in text languages often result in frustration and reduced motivation to learn programing. The focus has shifted from the syntax of a particular language to logical

thinking. Such languages are visually appealing, commands are displayed in blocks that are arranged like a jigsaw puzzle, and the result of running the program is immediately visible on the screen.

In addition, a feature of block-based programing languages is to shift the programing context from solving mathematical problems to programing games, animations, stories, and similar contents. Programing context is thus moved away from activities that are not attractive to digital natives, and placed in an environment that allows not only the development of mathematical abilities, but encourages creativity in other fields, thus taking into account other types of intelligence, and not exclusively mathematical-logical type of intelligence. [4]

Scratch has recently become one of the most popular block-based programing languages. It is a free program developed at the famous American university, the Massachusetts Institute of Technology (MIT), in a group called Lifelong Kindergarten. The idea is to encourage children to explore, come up with, and realize their ideas, which they can then show to others over the Internet, while adopting the basics of programing, math, visual and interactive design. Programing in Scratch is completely visual. Children write programs by stacking ready-made blocks into larger constructions. By using Scratch, a child can effortlessly learn to create simple computer games and programs on their own, and in the same way they can turn their creativity into interactive stories and multimedia animations. [5]

As stated before, the aim of this paper is to determine whether current or future teachers of Information Sciences in primary school education, possess professional competencies to teach Information Science in the domain of Computational Thinking and Programming. The authors also wanted to find out what the situation in other countries was. They made a brief overview of the existing literature; unfortunately, this topic has not been adequately researched. It seems that there is a notion that teachers are competent enough, or that conditions differ from those in Croatia.

Sáez-López and Miyata [6] state that in Spain, where they work, there is no ICT in primary schools and that on the Scratch Day, teachers had to make great efforts in preparation for working with students, and that results of the students were not satisfactory.

López-Escribano and Sánchez [7] state that their experience with teachers is that they generally do not want to learn programming languages, including Scratch. They state that mastering programming languages, including Scratch, requires a high level of abstraction and the complexity of the concepts they have to learn, and that this is a serious problem for teachers. Although Scratch has tried to overcome these obstacles, many teachers do not feel comfortable using programming languages.

A study conducted by Jimogiannis and Komis [8] focuses on teachers 'attitudes towards ICT. Students primarily emphasized feelings such as: 1. Anxiety and stress, 2. Self-efficacy, 3. Desire and pleasure, 4. Perceptions about ICT value and usefulness etc and the study concluded that Effective teachers' training is considered among the most important factors for the formulation of positive attitude towards ICT, as in the previous research conducted by Yildirim [9].

Fesakis and Serafeim [10] reached the similar conclusion as López-Escribano and Sánchez [7]. They conducted research with prospective students at the University of the Aegean in Greece. The research focused on students' opinions on the effect of a one-semester introduction to the Scratch environment for computer programming in primary education. Students had basic ICT skills, but no computer programming skills. Nevertheless, students - future teachers find Scratch useful and getting acquainted with it has significantly reduced the stress and anxiety associated with the ability to use ICT in educational practice.

II. CASE STUDY

In this case study we present the Information Sciences (IS) projects assigned to the students of the fifth and final year of the Teacher education - Faculty of Education Sciences with the module of Information Sciences.

Students are adequately trained for in-person teaching of the IS subjects in primary schools (1st to 8th grade). As a part of the curriculum, each student usually gives one or two IS presentations to the primary school students. In the context of the COVID 19 pandemic, classes could not be conducted in classrooms, and students had to prepare and present one teaching unit in the form of a video lesson that will be used for online teaching. Preparation of the video lesson was to be the final assignment for the subject "Informatics teaching methodology III", and students had to show that they have acquired sufficient professional and pedagogical competencies to work with primary school students.

Professional competencies imply an average knowledge of programing (Computational Thinking and Programing - which is quite demanding for most students of the Teacher education - Faculty of Education Sciences). Pedagogical competencies imply that students have learned to transfer knowledge and skills, i.e. to teach primary school students from the pedagogical-didactic-psychological-methodological aspect.

Nine students were given assignments with a very broad framework on how to prepare their presentation. Each student could choose a method of presentation. They were supposed to record the preparation of the process by using *ActivePresenter 8* (All-in-One Screen Recorder, Video Editor & eLearning Authoring Software [11]) or similar application for audio and visual recording of the lecturer. They were supposed to make a video lesson (teaching material) from which the students will be able to learn independently. Video lesson should have lasted from 10 to a maximum of 45 minutes. All assignments were to be done in one of the programing languages, and the following were used: a) Scratch - programing language and development environment for creating games (block-based programing language); b) Standard structural / object-oriented Python programing language, which is often defined as a language for teaching an algorithmic approach to problem solving [12] and c) The MakeCode

platform for programming micro: bits (block-based programing language).

Goal of selected tasks was to achieve the learning outcomes in the subject Informatics domain of *Computational Thinking and Programing*. They correlated with other subjects: Croatian language, Music culture, Mathematics, Physical and health culture, and with the interdisciplinary topic Health.

Students were not particularly enthusiastic about the assignments because they had to prepare lessons in a way they were not used to. They had knowledge of programing in the Python programing language, but not in the Scratch. They encountered screen and monitor recording programs, and video editing programs, but not the ActivePresenter. Therefore, students had to use online versions of textbooks by Croatian publishers (primarily [13]), as well as the instructions for using ActivePresenter [11]. The eighth student was also supposed to use the MakeCode platform for programming micro:bits [14] with accompanying educational materials [15].

The first student had to prepare an IT lesson for the subject of Music Culture, for the 3rd grade of the primary school. The assignment was: "*We play in Scratch - Creating a quiz to recognize a definite number of songs*." The video lesson was made in Scratch and it can be viewed at https://youtu.be/h5clpKyygL0.

The second student had to prepare an IT lesson for the subject of Music Culture for the 3rd grade of the primary school. The assignment was *We play in Scratch - programing a song.* The student chose the song *And so ticks the little clock*. The video lesson was made in Scratch, and it can be viewed at https://youtu.be/7ZPzGtk5m0s

The third student had to prepare an IT lesson for the subject of Music Culture for the 3rd grade of the primary school. The assignment was *We play in Scratch - Music quiz to recognize several musical instruments.* The video lesson was made in Scratch and it can be viewed at https://youtu.be/BqJZQIUJYSs.

The fourth student had to prepare an IT lesson for the subject of Mathematics for the 4th grade of the primary school. The assignment was *Mathematical quiz in Scratch - Guessing numbers from 1 to 64 - introduction of variables*. The video lesson was made in Scratch and it can be viewed at https://youtu.be/vUCYzpFNIwM.

The fifth student had to prepare an IT lesson for the subject of Mathematics for the 4th grade of the primary school. The assignment was *Final project in Scratch - Creating a math quiz: multiplication with numbers from - 12 to 12*. The video lesson was made in Scratch and it can be viewed at https://youtu.be/GOvMuqknbdg

The sixth student had to prepare an IT lesson for the subject of Croatian Language for the 3rd grade of the primary school. The assignment was *We create stories by programing - a story: 'Little Red Riding Hood'*. The video lesson was made in Scratch and it can be viewed at https://youtu.be/ENq9rQOwsDk.

The seventh student had to prepare an IT lesson for the subject Physical and Health Culture for the 4th grade of the primary school. The assignment was *Programing the presentation of several physical exercises in Scratch*. The video lesson was made in Scratch and it can be viewed at https://youtu.be/2m5q0eWUWq4.

The eighth student had to prepare an IT lesson for the interdisciplinary subject Health for the 4th grade of the primary school. The assignment was *Making a pedometer with additional options on the micro:bit*. The video lesson was made in the MakeCode platform for micro:bit programing, and it can be viewed at (https://youtu.be/xI2N7mNBnJg.

The ninth student had to prepare an IT lesson for the 7th grade of the primary school. The assignment was *Processing loops and functions through turtle graphics*. The video lesson was made in the Python programing language, and it can be viewed at (https://youtu.be/THEJYvEv2mw.

Learning outcomes from the subject Informatics domain Computational Thinking and Programming (B.3.1., B.3.2., B.4.1., B.4.2., B.7.1. and B.7.3) were to be met through "the Curriculum of the subject Information Science for primary and secondary schools" [1]. All Scratch programing solutions were supposed to have a 'Scene' and all the programs were supposed to be interactive. Mentor wanted to know how much time students needed to complete the assignments. Students reported that it took them between 15 and 25 hours because they needed to master Scratch, record applications, prepare auxiliary materials (music, sketches, drawings, pictures) for the process of recording teaching content, multiple recording of individual parts due to errors observed by students, multiple recording individual parts due to finishing indicated by other students (peer assessment), multiple recording of individual parts due to finishing pointed out by the teacher and post-productions, audio processing, insertion of audio and video transitions, text insertion, insertion of introductory and checkout tips, image acceleration etc.

Students evaluated each other's papers and marked with a grade of 5 (on a scale of 1 to 5), and the teacher-mentor also gave them excellent grades (from 87 to 98% of possible points) and expressed satisfaction with the work done by students. Students were also very satisfied with the final work and the level of acquired knowledge and experience.

Video (lessons) materials were also presented to the members of the *Association - Society of Informatics Teachers Pula* that consists mostly of the primary school teachers of Information Sciences in Pula and the surrounding area. Members of this society (16 members) have been very active since its inception, some 15 years ago and they were quite satisfied with the video materials.

The conclusion of the case study is that when teachers make an effort and use the resources available to them, it is possible to create excellent online teaching materials that can facilitate online lecturing. Since it takes a long time to make such material, it is recommended that teachers share the recorded lessons with each other.

III. RESEARCH

The goal of this study is to determine whether current teachers of Information Science, current fifth year students Teacher education - Faculty of Education Sciences with the module of Information Sciences and students of Information Sciences – Class Teaching study group, have professional competencies for teaching Information Sciences in the classroom, with the domain on the Computational Thinking and Programing.

For this purpose, we compiled a questionnaire with approximately 20 questions. The questionnaire was filled out by the Information Sciences teachers who work as teachers in primary and secondary schools, as well as fifth-year students of the Faculty of Education Sciences with the module Information Sciences and students of Information Sciences – Class Teaching study group. The survey was anonymous, filled out via a Google Form, and the completion time was, on average, 4 minutes. Respondents did not receive any benefits for completing the survey, other than being able to obtain survey results a week after its completion.

The survey group consisted of 58 respondents. 46 respondents of the survey group were female (79%) and 12 male (21%).

17 respondents were under the age of 30, 28 were between the ages of 31 and 45, and 13 were over the age of 45.

Two thirds of the respondents, 39 of them (67%) came from the County of Istria, and 1-3 respondents came from 13 other counties (mostly evenly represented throughout Croatia - from the Vukovar-Srijemska County to Split-Dalmatia County).

In terms of the respondents' teaching experience in primary and secondary schools, 28 of them (48.3%) had over 10 years of experience, 12 (20.7%) had 5-10 years of work experience, 8 (13.8%) had 0- 2 years of teaching experience, 6 of them (10.3%) were students and had no work experience, and 4 (6.9%) had from 2 to 5 years of work experience.

Majority of the respondents, 41 of them (70.7%) had no experience in teaching the lower grades in primary schools, 16 of them (27.6%) taught the lower grades in primary schools for up to 5 years, and 3 of them (5.2%) taught the lower grades in primary schools from 5-10 and over 10 years.

When asked if they taught Information Sciences in primary and secondary schools, 27 (46.6%) respondents taught IS for over 10 years, 7 (12.1%) of them from 5 to 10 years, 5 of them (8.6%) from 2 to 5 years, 6 of them (10.3%) up to 2 years, 6 of them were students (10.3%), and 7 (12.1%) were primary school teachers studying the module Information Sciences had no experience as teachers of Information Sciences.

The next question was *How long have you worked as an Information Sciences teacher in the lower grades of primary school?* As the teaching of Information Sciences in lower grades was discontinued a few years ago (at least in the County of Istria), we assumed there would be no teachers of Information Sciences with working experience in lower grades of primary schools. Our assumption proved correct. As many as 41 of them (70.7%) had no experience, 10 of them (17.2%) had up to 2 years, 3 (5.2%) from 2 to 5 years and from 5 to 10 years, and only one (1.7%) respondent had more than 10 years of work experience with teaching Information Sciences in the lower grades of primary schools.

Conducting IT workshops in clubs or similar in or out of school premises can be of great help in acquiring both pedagogical and professional competencies. Twenty nine (50%) respondents stated that they did not hold any IT workshop, 13 (22.4%) respondents stated that they held one or two, while 8 (13.8%) respondents held from 3 to 5 or more than 5 workshops.

Pedagogical competencies were acquired by Information Sciences teachers mainly, 40 (70.0% of them) during their regular faculty education. Fifteen of them (24.8%) acquired it by enrolling in *Pedagogical Competences* class after completing studies that did not have pedagogical competencies in its program (technical or economic studies), and 3 of them (5.2%) are currently in the process of acquiring competencies.

The next question was *How many times have you participated in programing training workshops, lasting at least 10 hours?* Nineteen respondents answered "never", 11 respondents participated in 1-2 workshops, 9 in 3 or 4 of them, and 19 respondents participated in 5 or more.

How many times have you held an Information Sciences workshop at a professional conference? Forty six of them replied "never", 6 of them held 1 or 2 workshops, 4 of them held 3 or 4 workshops, and two respondents held 5 or more workshops.

Although we expected better results in response to the question *How many times have you held an Information Sciences workshop at the County Expert Council?,* the improvement was insignificant: 35 (60.3%) respondents never held a workshop, 11 (19.0%) held 1 or 2 workshops, 3 (5.2%) held 3 or 4 workshops, and 9 (15.5%) respondents held 5 or more workshops.

The most interesting part of this research contains the following five questions in which respondents were asked to assess their knowledge of the Scratch programing language. Questions were associated with a measuring scale from 5 to 1, where 5 indicates excellent knowledge and 1 indicates insufficient knowledge. Each question was scored with an estimated value that the respondent has chosen. The minimum number of points for 5 questions is 5, and the highest is 25. The higher the value, the more competent the respondent is. However, the first step in the analysis of the initial set of statements was to assess the reliability of the analyzed measurement scales. This was achieved by Cronbach's alpha coefficients for each statement. The Cronbach's alpha coefficient is a measure of the internal consistency of a set of statements, and can assume a value between 0 and 1; the closer to the value of 1, the more reliable the measurement scale. The values of the Cronbach's alpha coefficient for all items are given in Table 1.

TABLE 1. ITEMS WITH CRONBACH'S ALPHA COEFFICIENT

Item	Cronbach alpha coefficient
Q1 [To what extent do you know and can explain to the students] *how to create an interactive quiz*	0.9322
Q2 [To what extent do you know and can explain to the students] *how to program a song by following musical notes*	0.9370
Q3 [To what extent do you know and can explain to the students] *how to make an animation of an object*	0.9206
Q4 [To what extent do you know and can explain to the students] *how to create a menu*	0.9259
Q5 [To what extent do you know and can explain to the students] *how to set a Scene*	0.9316

Kline [16] states the reliability criteria of the measurement scales: if the reliability coefficient takes a value of about 0.9, the reliability can be considered excellent, for a value of about 0.8, the reliability can be considered very good, for values around 0.7, the reliability can be considered acceptable. A reliability coefficient of less than 0.5 indicates that more than half of the observed variance could be due to an error.

Based on the results shown in the previous table, and the reference values stated in the literature, it can be concluded that the developed measurement scales have an excellent level of reliability.

Table 2 shows the results of the self-assessment of knowing Scratch and it compares the obtained values in different groups of respondents. The results (column *Average*) are graded and interpreted as follows: the respondents with up to 7 points have insufficient knowledge; respondents with 8-12 points have acceptable knowledge; respondents with 13-17 exceeds the acceptable knowledge respondents with 18-22 points have very good knowledge, and respondents with 23-25 points have excellent knowledge. Standard deviation and Median refer to the individual number of respondents in each group item. In the column *I have professional competencies*, the scale of points goes from 5 to 1, where the mark 5 indicates that respondents have excellent professional competencies, 4 - that respondents generally have professional competencies, 3 – that the respondents have average competencies; 2 - that respondents generally don't have sufficient professional competence, and mark 1 - that respondents don't have the necessary professional competence. The column *I need additional professional training for Scratch* shows the number of respondents who think that they need or do not need additional professional training in Scratch.

TABLE II. COMPARISON OF OBTAINED VALUES BY SELF-ASSESSMENT OF RESPONDENTS

Group	Respondents	No of Respondents	Average	Stdev	Median	I have professional competencies		I need additional professional training	
						Average	Stdev	No	Yes
I.	Students	6	18,5	5,24	20,5	4,17	0,983	1	5
	Informatics teachers working up to 5 years	11	13,5	8,41	15	4,18	0,874	3	7
	Informatics teachers working over 5 years	34	15,6	6,16	16	4,10	0,923	5	23
	Informatics teachers who do not teach Informatics	7	17,9	8,00	19	3,90	0,831	1	8
II.	Respondents who taught Informatics in lower grades of primary school	17	15,8	6,65	17	4,18	0,809	2	14
	Respondents who did not teach Informatics in lower grades of primary school	41	15,8	6,90	17	4,05	0,921	8	28
III.	Respondents who held programing workshops	29	17,6	6,50	19	4,45	0,736	7	20
	Respondents who did not hold programing workshops	29	14,0	6,65	14	3,72	0,882	3	22

The obtained values are classified into 3 groups: Group I. – Comparison between students and working teachers of Information Science based on the length of employment; Group II. – Comparison between the respondents with teaching experience of Information Science in lower grades of primary school with those without it; Group III. – Comparison between respondents with experience in holding workshops in programing in extracurricular and / or out of school activities. The same respondents participated in all three groups.

The results obtained from Group I indicate that final year students are introduced to Scratch in course of their studying and can respond with professional competencies to the requirements of teaching computational thinking and programing in the lower grades of primary schools. Slightly worse results are achieved by Information Sciences teachers who have recently completed their studies but are not currently working as Information Sciences teachers.

Information Sciences teachers with up to 5 years of teaching experience scored the worst results in the Average section (on the border of sufficient and good). There is no statistically significant difference between the obtained mean values at the significance level of $p < 0.05$. In determining whether there is a difference between students and unemployed IT teachers on one hand and employed IT teachers, in the section *I need professional training in Scratch*, using the χ^2 test it was determined that there is no statistically significant difference in significance level of $p < 0,05$ ($\chi^2 = 2.455$, and the cut-off frequency for df = 1 is $\chi^2 = 3.843$). The results obtained were as expected. Students had an opportunity to work with Scratch during their studies, while teachers who are already working have not had an opportunity to learn

Scratch while studying, and have not had an opportunity to work with Scratch in schools, because Scratch was commonly used in teaching the lower grades of primary schools, while Logo or Python programing languages are used in the higher grades.

Results obtained for Group II. indicate there is no statistically significant difference between the average values between the respondents who taught Information Sciences in lower grades and those who did not. In the answers to the question *I have professional competencies*

and in the answers to the question *I need additional professional training in Scratch* between the respondents who taught Information Sciences and those who did not, there is also no statistically significant difference in the level of significance of $p < 0.05$. Such results were expected.

Results for the Group III show that the achieved average values, between respondents who held programing workshops outside of regular classes and those who did not, are statistically different (for two-sided t test, $t = 2{,}117$ $N1 = N2 = 29$ with $p < 0{,}05$) and we can conclude that respondents who held programing workshops have better average results than those who did not. Furthermore, there is a statistically significant difference among those who conducted programing workshops as extracurricular activity in the section "I have professional competencies" at the level of significance $p < 0.05$ (two-sided t-test, $t=3.394$, $N1=N2=29$), but there is no statistical significant difference between the answers given by the respondents regarding the question *I need additional education* at the level of significance $p < 0.05$ ($\chi^2 = 1.6207$, and the limit value is 3.843 for $df = 1$). Such results were expected. Those who held programing workshops have many more hours of programing work with students than those who taught Information Sciences in schools. Programing is performed in only one of the four domains and the number of programing hours is relatively small. With a larger number of hours of programing lessons, teachers gain more self-confidence, and it is possible that they have encountered another block-based programing language in their work.

In general, a conclusion can be drawn that is valid for all three groups. Although most of respondents believe that they have very good professional competencies, they also need additional professional training in Scratch.

IV. Conclusion

Teachers and students who are the future teachers of Information Sciences feel adequately professionally competent to conduct the teaching of Computational Thinking and Programing (one of the domains of the subject of Information Sciences) in the lower grades of primary school. Almost everyone thinks that they would need additional training in the block-based programing language Scratch. The article also presents a case study that shows how, through a practical project, students of the Teacher education - Faculty of Education Sciences at the Juraj Dobrila University of Pula, in the course of Methodology of Teaching Information Sciences are professionally trained for the most demanding part of teaching the Information Sciences.

References

[1] Odluka o donošenju kurikuluma za nastavni predmet Informatike za osnovne škole i gimnazije u Republici Hrvatskoj, Narodne novine, službeni list Republike Hrvatske NN 22/2018-436, 06.03.2018. https://narodne-novine.nn.hr/clanci/sluzbeni/2018_03_22_436.html visible on 30.05.2020.

[2] Ž. Knezović "Nastava informatike u drugom razredu osnovne škole" - Methodical workshop, Intelligent e-learning systems, Šibenik, Croatia, 2005. https://bib.irb.hr/datoteka/209637.KnezovicZ.pdf visible on 30.05.2020.

[3] CS Fundamentals Curriculum Guide https://docs.google.com/document/d/1UqCgO06NzB1L6y83fnwnUcYdKr3MooJAaUZajj48DnI/preview visible on 30.05.2020.

[4] Što je Scratch? [November 24, 2015] Programerko, http://www.programerko.hr/2015/11/24/sto-je-scratch/ visible on 30.05.2020.

[5] M. Mladenović "Poučavanje početnog programiranja oblikovanjem računalnih igara", Doctoral thesis, Faculty of Science, University of Split, Split, 2019.

[6] J.M. Sáez López and Y. Miyata "Integration Scratch in primary education",Scratch conference, 2013 Barcelona https://www.researchgate.net/publication/301590669_INTEGRATING_SCRATCH_IN_PRIMARY_EDUCATION visible on 30.05.2020.

[7] C. López-Escribano and R. Sánchez-Montoya "Scratch y Necesidades Educativas - Especiales:Programación para todos" RED. Revista de Educación a Distancia. Número 34 (2012.) http://www.um.es/ead/red/34 visible on 30.05.2020.

[8] G. Fesakis and K. Serafeim, "Influence of the familiarization with "scratch" on future teachers' opinions and attitudes about programming and ICT in education" ACM SIGCSE Bulletin 41(3):258, August 2009, DOI: 10.1145/1595496.1562957

[9] A. Jimogiannis. and V. Komis "Attitudes and opinions of secondary education teachers against the application of ICT in instruction", In proceedings of the 4th Pan-Hellenic Conference of HICTE on "ICT in Education" (Athens, Greece, Sep 29- Oct 3, 2004)

[10] S. Yildirim "Effects of an educational computing course on preservice and inservice teacher: a discussion and analysis of attitudes and use", Journal of Research on Computing in Education, 32(4), pages 479-496 (2000.)

[11] ActivePresenter 8, Video Editor & eLearningAuthoring Software https://atomisystems.com/ - visible on 30.05.2020

[12] P. Brođanac, L. Budin, Z. Markučič, S. Perić "Python – jezik za podučavanje algoritamskog pristupa rješavanju problema" Proceedings of the 33rd International convention on information and communication technology, electronics and microelectronics : MIPRO 2010; Proceedings Vol IV; Computers in education: CE, Rijeka

[13] E-sfera - textbook platform Školska knjiga, Zagreb, https://www.e-sfera.hr/ visible on 30.05.2020.

[14] MakeCode platform for micro: bit programming https://izradi.croatianmakers.hr/ visible on 30.05.2020.

[15] IRIM - Institut za razvoj i inovativnost mladih – Educational materials https://izradi.croatianmakers.hr/bbc-microbit-uvodna-stranica/ visible on 30.05.2020.

[16] R.B. Kline "Principles and Practice of Structural Equation Modeling", The Guilford Press, New York, 1998

Perception of Using VAR Technology in Football After Completion of Training and Education and Experiences of Croatian Video Assistant Referees (VARs) and Assistant VARs (AVARs)

I. Lucić*, S. Babić*, D. Vučkov**
* Polytechnic of Rijeka, Business Department, Croatia
ivan.lucic@veleri.hr
snjezana.babic@veleri.hr
** Judge at County court in Rijeka, Croatian football federation – Head of VAR education, Croatia
dodovuckov@gmail.com

Abstract - By introducing VAR technology in football the process of decision-making during the game has been made easier for referees on the field, where they coordinate with a team of experts that are inside the VOR room. Additional education and practical experience of all participants involved in this process is required for the application of VAR technology in football in order to ensure the accuracy and adequate speed of the decision-making process. This paper shows the results of the research on the perception(s) of using VAR technology in football of Croatian Video Assistant Referees (VARs) and Assistant VARs (AVARs) that received training and education regarding the use of VAR technology on the football field. The results and findings of this paper can help all those who want to contribute to the advancement of application and functionality of VAR technology in football.

Keywords – VAR technology; video assistant referees; football; VARs; AVARs

I. INTRODUCTION

Football is one of the most popular sports nowadays and it attracts large masses of people that can watch a large number of games on a daily basis. Among the key participants in football games are football referees who make the final decisions that can influence the outcome of the game.

During football matches, disputable situations happen that need to be resolved fairly and correctly by the referees. The practice has shown that there is a problem in detecting certain irregularities that happen during matches due to the impossibility of reviewing said irregularities within a given (actual) time frame. Additional problems arose when the viewers watching matches on TV, and instant replays, could see from other angles that the referees had made incorrect ~~calls~~ decisions based on their poor assessment of certain situations during the matches they were refereeing [1].

Therefore, the introduction of Video Assistant Referee (VAR) technology in football has had a clear and defined goal: help the referees mitigate their mistakes during football matches that have a significant impact on the outcome(s) and quality of said matches; provided it doesn't undermine their authority. This allows for bigger transparency and fairness of outcomes of matches, which has an impact on the satisfaction of all participants, referees, players and supporters [1].

However, for a group sport with such a rich tradition, introducing technology and changing fundamental rules of the game presents a large change, which has been met with wide opposition from players and coaches alike [1].

For the application of VAR technology during football matches a team of trained experts is needed, and it includes: Video Assistant Referees (VARs) and Assistant VARs (AVARs) and Replay Operator (RO) that work in the Video Operation Room (VOR) [2, 3, 4].

Practice has confirmed the advantages of introducing VAR technology in football, but it has also indicated the possibilities for improvement of its application. As an example, Spitz et al. [5] have found that the accuracy of correct decision after VAR intervention in relation to the initial referee's decision significantly increases from 92.1% to 98.3%. On the other hand, based on the analysis of data from Spanish LaLiga during 2018-19 season, Errekagorri et al. [6] have found that the use of VAR system in matches hardly has any impact on the game in elite football when playing time and technical-tactical and physical performances are taken into consideration.

Samuel et al. [7] emphasize that the critical aspect of refereeing football matches is the decision-making process by referees, and have concluded based on results of research (sample of 20 elite referees and assistants), that the factors that influence the referees' decision-making process are: physical fitness and fatigue, field positioning, visual attention, contextual factors, game management, expertise and cognitive processing, psychological factors, and team factors (including Video Assistant Referees).

In order to improve VAR and its use further, it is imperative to investigate the opinions and attitudes of VARs and AVARs on the matter.

This paper puts a special emphasis on the introduction of VAR technology in Croatian football. The main aim of this paper is to examine the perception(s) of use of VAR technology in football after completion of training and education and experiences of Croatian Video Assistant Referees (VARs) and Assistant VARs (AVARs).

The following section contains the most important characteristics and results of previous research regarding the application of VAR technology in football, followed by a description of the research methodology, research findings and discussion which was conducted on the sample of 24 Croatian VARs and AVARs. A short conclusion is given in the last section.

II. RELATED WORK

A. VAR technology in football

Video Assistant Referee System was used in football for the first time in a big tournament at the 2018 FIFA World Cup in Russia; however, the International Football Association Board (IFAB) approved the use of Video Assistant Referee (VAR) in March 2016.

The use of Video Assistant Referee System in football matches is only permitted if competition organizer has received written permission from The IFAB and FIFA, i.e. if they meet the requirements (as set out in the VAR Handbook) and all the VAR protocol. The main principles on which the use of VAR system is based is the assistance to the referee on the field, but only in case of "clear and obvious error" or "serious missed incident" in relation to: Goal/no goal, Penalty/no penalty, Direct red card (not second yellow card/caution), mistaken identity of player [3, 4].

Considering the complexity of the activities performed by VAR team members and referees during the refereeing of football matches with the help of VAR, additional education and training is required beforehand; in terms of both theoretical and practical application of VAR technology.

Ruiz et al. [8] have introduced the concept of "digital referees" and have emphasized that football referees must have technical, management, and fitness skills but that they also must develop new digital skills and knowledge about the new digital rules of match(es). Therefore, the choice of didactic teaching strategies and correct program to train video assistant referees is very important. For example, in other contexts, the results of research conducted by Armenteros et al. [9] confirm that referees and assistant referees are willing to use multimedia teaching materials for education for their activities, which is a good indicator of acceptance of innovative technologies in future work.

Results of a study by Carlos et al. [10] show that after using of VAR system in elite football, there was a decrease in the number of the off side situations, fouls and yellow cards and an increase in the number of minutes added to the playing time.

Successful application of VAR technology in football depends on the following factors: the accuracy of reviewed decisions, number of reviews and their duration, and the impact of reviews on the match and its emotional aspect.

Moreover, insufficient number and poor positioning of cameras can have an influence on the transparency of the referee's decision, which can reflect on the quality of the match. Also, the communication between VAR team members, both amongst themselves and with the referee on the field is of extreme importance, and it has to be clear and understandable [1].

Also, Ugondo and Tsokwa [11] have explored communicative frame reference for Video Assistant Referee and field-based referees and found that major factors that impact the decisions of the field-based referee in contestable situations are: angular lens, camera perspective and the rules of the game alongside referee's prejudice.

B. Application of VAR technology in Croatian football

The preparations for the introduction of VAR technology in Croatian football had lasted one year, and VAR system was used for the first time on 31st January 2020, in a match of Croatian First Division Football League.

There are 16 referees with the education necessary to work with VAR; 12 VARs, 14 AVARs, and 5 technicians, i.e. ROs (Reply Operators) [12]. It is important to take into consideration that one VAR person is an instructor and that one AVAR is an instructor's assistant and due to this fact they are neither operating as VAR/AVAR in the competition, nor taking part in this research.

VAR system in Croatia is implemented at every stadium in a specially designated room. The Croatian Football Federation (HNS) has opted for a plan with 6 or 8 cameras, depending on the requirements of each specific game [13].

Croatel Company, as a technical partner of the Croatian Football Federation, is in charge of setting up VAR system in Croatian football. It sets up cameras on stadiums and it is in charge of TV coverage of matches. The VOR (Video Operation Room) also has access to said coverage. Croatel operates using a system called Xeebra, designed to offer video-assistance to referees. The system was developed by EVS Company.

III. EMPIRICAL STUDY

The aim of this study was to examine perceptions of using VAR Technology in Football after completion of training and education and practical experience by Croatian Video Assistant Referees (VARs) and Assistant VARs (AVARs).

A. Procedure

The research in this study was conducted through a Google Forms online questionnaire from 20 to 23 June 2020. The research was conducted through an email request to fill out the online questionnaire sent to all VARs and AVARs of the Croatian Football Federation (HNS), 24 in total, by the Federation's main VAR instructor. Participation was voluntary and anonymous.

B. Apparatus

For the purposes of this research, a new questionnaire was developed containing 35 closed-ended questions and 1 open-ended question. 9 questions were related to general characteristics of the participants, 26 questions were related to perceptions of use of VAR technology in football. Responses to the questionnaire items were modulated on a five-point Likert scale (1=totally untrue; 2=mostly untrue; 3=neither true nor untrue; 4=mostly true; 5= totally true).

Some items in this questionnaire were adapted from two technology acceptance models: Technology Acceptance Model - TAM [14], and Unified Theory of Acceptance and Use of Technology – UTAUT [15], and some items were newly developed.

The last question in this questionnaire was an open-ended question regarding the drawbacks of VAR technology from the technical (IT) aspect.

After the initial form of the questionnaire was created, it was sent to an expert of VAR technology application in football, i.e. VAR instructor of the Croatian Football Federation.

C. Participants

Twenty-four Video Assistant Referees (VARs) and Assistant VARs (AVARs) participated in this research. The results (see Table I.) show that all of the participants were male (24; 100%). All of the participants were of three age groups; most of them were between 31 and 40 years of age (15; 62.5%), 6 were between 41 and 45 years of age (25.0%), while the rest were between 46 and 50 years of age (3; 12.5%). A little over half of the participants were assistant referees (14; 58.3%), while the rest were referees (10; 41.7%). In terms of VAR technology officials, more than half of the participants were AVARs (14; 58.3%), while the rest were VARs (10; 41.7%) (see Table I).

As shown in Table 1, most of the participants in this research have had 5 to 10 years of refereeing experience in the Croatian First Division Football League (10; 41.7%); 8 of them (33.3%) have had 10 to 15 years of refereeing experience; 5 (20.8%) have had more than 5 years of refereeing experience in the Croatian First Division Football League; while only 1 participant (4.2%) has had less than five years of refereeing experience. Among VARs and AVARs participating in this research, most have active (officiating) referee status (20; 83.3%), while 4 (16.7%) are retired and are only VARs or AVARs.

The data analysis (see Table I.) shows that most of the participants, 11 of them (45.8%), have refereed more than 10 matches with the help of VAR technology, while 10 of them (41.7%) have refereed between 5 and 10 football matches with the help of VAR technology. Furthermore, only 3 participants have refereed with the help of VAR technology for less than 5 years.

The aforementioned data analysis shows that Croatian VARs and AVARs have enough practical experience to assess their satisfaction in terms of refereeing football matches assisted by VAR technology.

TABLE I. BASIC CHARACTERISTICS OF CROATIAN VARS AND AVARS IN THE STUDY (N=24)

Variables		Percentage of participants in relation	
		Freq.	%
Gender	Male	24	100.0
	Female	0	0
Age	Less than 20 years	0	0
	21-30 years	0	0
	31-40 years	15	62.5
	41-45 years	6	25.0
	46-50 years	3	12.5
	51-55 years	0	0
	More than 55 years	0	0
Type of referee	Referee	10	41.7
	Assistant referee	14	58.3
Type of VAR technology referee	VAR	10	41.7
	AVAR	14	58,3
Years of refereeing experience in Croatian First Division Football League	Less than 5 years	1	4.2
	5-10 years	10	41.7
	10-15 years	8	33.3
	15-20 years	3	12.5
	More than 20 years	2	8.3
Status	I am an active referee	20	83.3
	I am only VAR/AVAR	4	16.7
Refereeing experience with the help of VAR technology	Less than 5 matches	3	12.5
	Between 5 and 10 matches	10	41.7
	More than 10 matches	11	45.8

With regard to the question: "How would you assess your knowledge in the field of application of VAR technology in football?" all of the participants believe they are able to apply VAR technology in football more than well. 15 of the participants (62.5%) consider their knowledge to be very good, while 9 (37.5%) consider their knowledge to be excellent, as shown in Table II.

Furthermore, the results of the data analysis show that je 16.67% of the participants (four of them) are mostly satisfied with the education and training conducted for the purpose of using VAR technology, while 83.3% of the participants are completely satisfied with the conducted education and training.

D. Findings

The items were measured by a 1-5 point Likert scale for responses (1 – totally untrue; 2 – mostly untrue; 3 – neither true nor untrue; 4 – mostly true; 5– totally true).

The following table shows the results of the data analysis, collected for the purpose of this research, regarding the use and attitudes of VARs and AVARs toward the application of VAR system as means of assistance in refereeing football matches. The items in Table II. were measured by a 1-5 point Likert scale for responses (1 – totally untrue; 2 – mostly untrue; 3 – neither true nor untrue; 4 – mostly true; 5– totally true).

The results in Table II. indicate that, on average, VARs and AVARs responses to the items vary from 1.38 ("totally untrue") to 4.88 ("totally true").

According to the data gathered (see Table II.), all participants (24; 100.00%), VARs and AVARs of the Croatian Football Federation, have agreed with the items

that they *find it easy to use VAR system in football matches since they have the necessary resources, knowledge and skills;* also that they *consider the introduction of VAR system in football a very good idea,* and that they are *satisfied with the quality of the results the use of VAR in football has yielded.*

Furthermore, all the participants (24; 100.00%) agree that *communication between VARs and AVARs and ROs (Reply Operators) during the matches is clear and understandable;* they *would recommend the application of this technology to everyone else*, and they *wish to continuously improve their skills in the application of VAR system in football.*

The arithmetic means (see Table II.) indicate that around 96% of the participants consider that *the use of VAR technology in football increases the transparency of the refereeing and the effectiveness i.e. correctness of the decisions made by the referees on the field*, as well as that *using VAR during football matches has become a normal occurrence* for them. Moreover, most of the participants (87%) consider that *the players believe that using VAR technology in football is a very good idea*; and also that they *have no problem "forgetting" about VAR technology during the matches*, and that *the functionalities that VAR has at its disposal provide enough support to the referees on the field (e.g. screen size, image clarity, slow motion, etc.)*.

Results in Table II show that a smaller number of participants (3; 12.5%) believe that VAR system doesn't fit well enough with the existing culture of football matches, and 1 participant (4.17%) is undecided; while most of the participants (20; 83.34%) believe that *VAR technology fits well with the traditional football environment.*

Most of the participants (75%) agree with the item: "Considering the results yielded by VAR system, I find it simple to referee football matches", while 25% remain undecided. Also, 75% of the participants agree with the item: "VAR system improves the quality of football"; 4 participants (16.67%) remain undecided on the matter, and only 2 participants (8.33%) disagree.

The mean (see Table II.) indicates that, on average, the participants (M=3.42, SD=1.19; Mdn=3.5) find the item: *quality decisions made by the referees with the help of VAR are for the most part dependent on the functionality of the video technology itself* to be mostly true. Thus, the results show that 16.67% of the participants disagree with this item, while 29.17% of the participants remain undecided, and a little over half of the participants (54.17%) find this item to be "mostly true" and/or "totally true". These results show that almost half of the Croatian VARs and AVARs believe that making a quality decision as a football referee is still dependent on the human factor and that the functionality of video technology isn't enough in and of itself. Most of the participants (70%) are of the opinion that the spectators/viewers believe that football refereeing wouldn't be sufficiently modern (contemporary) without VAR technology; while only 6 (25%) of the participants remain undecided, and 1 (4.11%) participant disagrees with this item.

The mean of 3.13 (SD=0.83; Mdn=3) indicates that, on average, the participants are undecided with regard to the item: *VAR system is reliable and there are no software glitches*. A little more than half of the participants (13; 54.17%) remain undecided and 29.17% of them are of the opinion that VAR technology has no software glitches; while 4 (16.67%) of the participants disagree with the item. Since satisfaction with the use of software is one of the most important factors in the acceptance of technology, generally; the results of the data analysis in this research show that Croatian VARs and AVARs have indicated the need need for the development of even higher quality VAR technology software.

Furthermore, most of the participants (10; 41.67%) couldn't decide with regard to their opinion on the *simplicity of acquisition of skills required for using VAR*, while 1/3 of the participants (8; 33.33%) have disagree with this item. These results are confirmed by the mean which indicates that, on average, the participants (M=2.71, SD=1.05; Mdn=3) are undecided. These results show that the application of VAR technology isn't simple and that it requires certain user competencies which are acquired through both theory and practice; i.e. by refereeing football matches.

A little over half of the VARs and AVARs of the Croatian Football Federation participating in this research, 54,17% of them, disagree with the item: "Cameras used by VAR in Croatia are sufficient to cover all the necessary angles during a match"; while 8 (33.33%) are undecided, and only 3 (12.5%) of the participants agree. This is confirmed by the mean (M=2.46, SD=0.87; Mdn=2.5) which indicates that, on average, the participants disagree with this item (Table II). This result is expected, seeing as internationally a larger number of cameras is used, thus enabling better field coverage and higher transparency in decision-making. With a larger number of cameras available, difficulty of the decision-making task for the referee is lower because more quality information is provided to the referee by VARs and AVARs.

Moreover, the mean (see Table II.) indicates that in this research, on average, VARs and AVARs (M=1,71, SD=0.89; Mdn=1.5), have found the item: "I believe that the use of VAR technology could be done successfully without ROs (Reply Operators), provided that I attend the course on using VAR technology within the scope of ROs' functions" to be mostly untrue. It is interesting to note that 2 of the referees (8.33%) believe they could successfully implement the aforementioned; while one referee (4.17%) has remained undecided. These results indicate that the RO (Reply Operator) is an integral member within the VOR room that performs important and specific activities. The data shown in Table II., indicate that, on average, the participants (M=1.38, SD=0.63; Mdn=1), have found the item: "I am apprehensive about the possibility of losing control of the match due to the use of VAR technology" to be totally untrue; and only 2 participants (8.33%) are undecided on this matter. This offers the conclusion that VARs and AVARs have sufficient level of competence to use VAR system during football matches, and that they believe in its functionality.

TABLE II. THE AVERAGE SCORE AND MEDIAN VALUES OF VARS AND AVARS (N=24)

Items	*M (σ)*	*Mdn*
Taking into consideration resources, knowledge and skills required to use VAR system, I find using VAR technology easy during the refereeing of football matches.	4,29 (0,45)	4
Considering the results yielded by VAR system, I find it simple to referee football matches.	4,04 (0,73)	4
I consider the use of VAR system in football a very good idea.	4,88 (0,33)	5
I am apprehensive about the possibility of losing control of the match due to the use of VAR technology	1,38 (0,63)	1
Quality decisions made by referees with the help of VAR system in football are, for the most part, dependent on the clarity of communication between referees and VARs	4,21 (0,87)	4
Quality decisions made by referees with the help of VAR are for the most part dependent on the functionality of the video technology itself	3,42 (1,19)	3.5
Communication between VARs/AVARs and ROs (Reply Operators) during the matches is clear and understandable.	4,42 (0,49)	4
Functionalities that VAR has at its disposal provide enough support to the referees on the field (e.g. screen size, image clarity, slow motion, etc.).	4,17 (0,62)	4
Cameras used by VAR in Croatia are sufficient to cover all the necessary angles during a match.	2,46 (0,87)	2.5
It is simple to acquire the skill(s) for using VAR system.	2,71 (1,06)	3
I believe that the use of VAR technology could be done successfully without ROs (Reply Operators), provided that I attend the course on using VAR technology within the scope of ROs' functions	1,71 (0,89)	1.5
I have no problem "forgetting" about VAR before making a decision on the field, i.e. VAR doesn't represent a burden for me.	4,33 (0,94)	5
With the help of VAR technology, referees are able to make more factually correct decisions on the field.	4,5 (0,65)	5
VAR system in football makes the refereeing process easier for referees.	4,46 (0,76)	5
By using VAR system in football, transparency of refereeing is increased.	4,71 (0,54)	5
Generally, using VAR system in football improves the quality of football.	4,13 (0,97)	4
Use of VAR system in football increases the effectiveness i.e. correctness of the decisions made by the referees on the field.	4,63 (0,56)	5
I am satisfied with the quality of results yielded by the use of VAR in football.	4,46 (0,5)	4
VAR system is reliable and doesn't have glitches.	3,13 (0,83)	3
My experience in using VAR has been better than expected.	4,00 (0,76)	4
It has become normal (natural) for me to use VAR during football matches.	4,58 (0,57)	5
I would recommend VAR system to everyone for refereeing football matches.	4,75 (0,43)	5
I believe VAR technology fits well with the traditional football environment.	3,96 (1,02)	4
I am of the opinion that players believe that using VAR technology in football is a very good idea.	4,21 (0,64)	4
I am of the opinion that spectators/viewers believe that, nowadays, football refereeing without the use of VAR technology isn't sufficiently modern (contemporary).	3,83 (0,75)	4
I intend to continuously improve my skills in the application of VAR system in football.	4,96 (0,2)	5

The participants were given the following open-ended question: "In your opinion, what is the key drawback of VAR technology; if you look at it from a technical aspect (as an application, software)?"

In this respect, most of the participants, 11 of them (45.83%) indicated *"insufficient number of cameras for full quality coverage of all the necessary angles of the football field"*; 4 (16.67%) of the participants indicated it was *"a VAR system crash during match"*, while 1 (4.17%) participant indicated "*The Virtual Offside Line (VOL)*" as a drawback of using VAR technology in football.

1) Difference between VARs' and AVARs' perceptions of using VAR technology after education and experience

The Mann-Whitney U test was used to compare differences between two independent groups, VARs and AVARs, regarding their perceptions of using VAR technology in football after experience and education. All items shown in Table II. were analyzed.

In the continuation of this paper, only the results that have revealed statistically significant difference between VARs and AVARs are shown. The analysis of the collected data has revealed that VARs (Mdn=5) recognize significantly more (U=26.00, p=.004) that with the help of VAR technology the referees on the field can make more factually correct decisions in relation to AVARs (Mdn=4). Results of the analysis indicate that VARs (Mdn=3.5) recognize significantly less (U=38.50, p=.049) that the use of VAR improves the quality of football in relation to AVARs (Mdn=5). VARs (Mdn=4) believe significantly less the use of VAR technology in football matches to be "better than expected" (U=37.00, p=.039) in relation to AVARs (Mdn=4.5). Finally, significantly more (U=30.00, p=.006) AVARs (Mdn=5) consider it has become normal (natural) for them to use VAR during football matches in relation to VARs (Mdn=4).

Aside from the aforementioned items, other items (see Table II.) bear no statistically significant difference between VARs and AVARs.

IV. DISCUSSION

VAR technology in football can be viewed as an innovation that has its course of acceptance by all participants in its process of implementation. The main task of this video technology is to help and assist the referees on the field in their decision-making process during the matches, and thereby increase the transparency and fairness of the outcome of the game, as well as the quality of football in general.

Some national football leagues have already accepted, implemented and are using VAR during their matches, while others are still in the acceptance phase. The implementation of VAR system isn't simple, seeing as the quality of its use depends on a larger number of participants in this process (VARs, AVARs, ROs, referees and others).

For the implementation of VAR system certain skills, knowledge and competencies are needed, and these are acquired through both theory and practice.

Satisfaction with the use of technology is crucial for its acceptance and development. VARs and AVARs help and assist the referees on the field, thus their competencies and perception(s) of use of VAR system are of great importance. Seeing as the outcome of the quality use of VAR is dependent on the environment, the matter of the

perceptions of those taking part in this process is always important and discussed anew.

The research results in this paper, which are based on the data collected from 24 Croatian VARs and AVARs of the Croatian Football Federation, indicate that they find it easy and simple to use VAR system and that they believe that its introduction in Croatian football has been a good idea.

They have indicated their complete agreement with the main characteristics of the use of VAR technology, i.e. pointed out that it enables them to:

- make more factually correct decisions,
- make the refereeing process easier for the referees,
- increase the transparency and efficiency of the refereeing, i.e. the accuracy of the decision-making process on the field.

Also, Croatian VARs and AVARs that have participated in this research have indicated that the application of VAR technology in football matches doesn't represent a burden for them when it comes to making decisions and that they are able to forget about VAR during the match without any problem; as well as that they are not apprehensive about the possibility of losing control of matches due to the use of VAR technology; and they have also indicated that they believe that a referee's decision doesn't depend solely on the functionality of VAR system. The obtained results have confirmed that after completion of training and education and practical experience, Croatian VARs and AVARs have recognized and acknowladged the basic advantages of introducing VAR technology defined by [1]. As the drawback of the application of VAR technology Croatian VARs and AVARs have indicated the problem of insufficient number of cameras to cover all the necessary angles during matches, while only one of the participants has indicated the problem of "*The Virtual Offside Line (VOL)*", which are necessary to make the correct decisions. Similar results have been obtained in previous research [e.g. 1, 10, 11].

Also, the participants in this research have pointed out that VAR system isn't 100% reliable because it has been known to crash in some cases, which is a glitch in technology that requires additional time to repair; which can be a challenge for the refereeing process and the continuation of the game.

The data analysis indicates slight but statistically significant differences between VARs and AVARs in relation to their perceptions of use of VAR technology in football. More AVARs than VARs believe that the use VAR system in football increases quality and that said quality is "better than expected", as well as that it has become normal (natural) for them to use VAR during football matches; while more VARs than AVARs believe that VAR system enables them to make more correct and fair refereeing decisions. These results are in correlation with the difference in their activities; it is believed, however, that the obtained results require additional investigation, seeing as there are other factors that could have an influence on the confirmed difference in their perceptions. It is especially important to point out that there is a need for additional technical and software solutions in the future in order improve VAR even further.

V. CONCLUSION

Based on the results obtained in this paper, it can be concluded that VAR technology has largely met the expectations of VARs and AVARs of the Croatian Football Federation, based on the satisfaction shown with their education and practical experience, as well as their intention to continuously improve their skills in the application of VAR system in football. Therefore, it is necessary to invest in lifelong learning programs for further development of VARs and AVARs, as well as to introduce new technology in education programs at all levels of formal education. [8, 9].

Our findings can help researches as a foundation for future research, as well as all practitioners who aim to further improve and develop VAR technology in football.

REFERENCES

[1] M. Armenteros, A. J. Benitez, M. A. Betancor, "The Use of Video Technologies in Refereeing Football and Other Sports", Routledge, New York, 2019.

[2] Fédération Internationale de Football Association (FIFA), Video assistant referee (VAR), available at: https://football-technology.fifa.com/en/media-tiles/video-assistant-referee-var/ [Accessed 12. 6. 2020.]

[3] The International Football Association Board (IFAB), VAR protocol, available at: https://www.the-afc.com/documents/var-protocol [accessed 12 June 2020]

[4] The International Football Association Board (IFAB), Video assistant referee (VARs), Implementation handbook for competitions, available at: https://www.docdroid.net/W2cATO8/var-handbook-v8-final-pdf [accessed 12 June 2020]

[5] J. Spitz, J. Wagemans, D. Memmert, A. M. Williams, & W. F. Helsen, "Video assistant referees (VAR): The impact of technology on decision making in association football referees. *Journal of Sports Sciences*, pp. 1-7, 2020.

[6] I. Errekagorri, J. Castellano, I. Echeazarra, & C. Lago-Peñas, "The effects of the Video Assistant Referee system (VAR) on the playing time, technical-tactical and physical performance in elite soccer", *International Journal of Performance Analysis in Sport*, 1-10, 2020.

[7] R. D. Samuel, G. Tenenbaum, & Y. Galily, "An integrated conceptual framework of decision-making in soccer refereeing", *International Journal of Sport and Exercise Psychology*, pp. 1-23, 2020.

[8] M. F. Ruiz, M. A. Betancor & M. Armenteros, "Didactic Teaching Strategies for the VAR", *The Use of Video Technologies in Refereeing Football and Other Sports*, Routledge, New York, 2019.

[9] M. Armenteros, S. S. Liaw, M. J: Sánchez-Franco, M. Fernández & R. A. Sánchez, "Analysis of FIFA referees and assistant referees' motivational factors towards the Multimedia Teaching Materials", *Education and Information Technologies*, *22*(3), pp. 841-872, 2011.

[10] L. P. Carlos, R. Ezequiel & K. Anton, "How does Video Assistant Referee (VAR) modify the game in elite soccer?", *International Journal of Performance Analysis in Sport*, *19*(4), pp. 646-653, 2019.

[11] P. I. Ugondo, M. Tsokwa, "Interpreting Video Assistant Referee and Goal-Line Technology Communication: The Pitch-Based Referees Perspective", International Journal of Trend in Scientific Research and Development, 3(4), pp. 1058-1062, 2019.

[12] The Croatian Football Federation (HNS), available at:: https://hns-cff.hr/news/21021/hns-dobio-dozvolu-ifab-a-za-koristenje-var-a/ [accessed 12 June 2020]

[13] The Croatian Football Federation (HNS), VAR handbook, available at: https://hns-cff.hr/files/documents/16518/VAR_prirucnik-2018-2019.pdf [accessed 12 June 2020]

[14] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology", MIS Quarterly, vol. 13(3), pp. 319-340, 1989.

[15] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: toward a unified view", MIS Quarterly, vol. 27(3), pp. 425-47, 2003.

Implementacija sustava za automatsko praćenje prisutnosti na Visokom učilištu Algebra

Bojan Fulanović, Andrej Lacković i Daniel Bele
Visoko učilište Algebra, Zagreb, Hrvatska
bojan.fulanovic@algebra.hr, andrej.lackovic@algebra.hr, daniel.bele@racunarstvo.hr

Sažetak - Ovaj rad opisuje implementaciju automatskog praćenja prisutnosti na nastavi unutar Visokog učilišta Algebra. Pojedinačno prozivanje velikog broja studenata oduzima puno vremena zbog čega se gubi efektivno vrijeme održavanja nastave. Kako bi izbjegla navedene probleme Visoko učilište Algebra uvela je automatsku prijavu studenata i profesora putem NFC čitača i pripadajućih kartica. Sinergija postojećeg sustava za evidentiranje studenata i sustava čitača kartica postignuta je putem web servisa.

Ključne riječi - NFC, čitači kartica, web servisi, sinergija, SQL, React, Entity Framework, .NET

I. UVOD

Ideja vodilja ovog rada je prikazat opravdanost korištenja informacijskih komunikacijskih tehnologija (ICT) u svrhu evidentiranja prisutnosti na nastavi. Broj studenata se iz godine u godinu povećava i trenutno su kapaciteti učionica u pravilu popunjeni. Učionice u kojima se odvija nastava mogu primiti do 100 studenata. Evidentiranje tolikog broja studenata od strane nastavnika nerijetko uzme i do 10 minuta vremena što iznosi gotovo 10% vremena koje je predviđeno za odvijanje nastave.

Kako bi riješila navedeni problem Visoko učilište Algebra uvela je automatske čitače kartica u svoje učionice. Student prilikom ulaska u učionicu prislanja svoju karticu na NFC čitač koji taj broj kartice evidentira u vlastitoj bazi podataka. Postojeći sustav za evidenciju studenata, Infoeduka, prilikom otvaranja termina konkretnog kolegija od strane nastavnika, komunicira sa web servisom koji joj isporučuje podatke evidentiranih kartica u definiranom vremenskom intervalu, tj. u terminu trajanja određenog kolegija. Sustav Infoeduka na osnovu primljenih podataka ažurira prisutnost svih studenata koji su se prijavili nakon čega je dovoljan jedan klik kako bi se evidencija pohranila u bazu podataka. Napravljena je i interna React web aplikacija za provjeru očitanja kartica gdje svaki nastavnik može provjeriti sve prijavljene kartice studenata i nastavnika, te uvidom u zapise ukazati na eventualne pogreške ili pokušaje malverzacije od strane studenata. Naime studenti ponekad znaju zlorabit navedeni sustav na način da svoju karticu daju nekom od kolega koji bi osim vlastitog evidentiranja, evidentirao i kolegu studenta koji nije prisutan na nastavi. Kako bi to spriječili, prije same pohrane podataka nastavnici su dužni provjeriti broj prijava u sustavu sa brojem prisutnih studenata u učionici te u slučaju nepravilnosti ručno ukloniti iz sustava studenta koji nije prisutan. Kako bi se spriječilo takvo ponašanje na slijedećim terminima predavanja, neodlaznošću se kazni i student koji je očitao karticu studenta kojega nema. Također u slučaju ometanja nastave od strane studenata, nastavnici su ih dužni uputiti da napuste učionicu kako bi ostali studenti mogli neometano pratiti nastavu. Tada se ručno kroz sustav može poništiti dolaznost navedenim studentima.

II. NFC

NFC (Near Field Communication) odnosno komunikacija bliskog polja je tehnologija za prijenos podataka putem radio frekvencije na udaljenostima od svega par centimetara. Jedna od najčešće korištenih mogućnosti NFC tehnologije danas je plaćanje računa putem pametnog telefona, npr. iOS operativni sustav



Slika 1. Ilustracija komunikacije NFC uređaja

koristi Apple Wallet aplikaciju koja putem NFC tehnologije omogućuje beskontaktno plaćanje računa, također koristi se kod izrade „pametnih kuća“ gdje je omogućeno otključavanje vrata putem NFC kartica ili pametnih telefona, moguće ju je koristit za dijeljenje kontakata, slika i video zapisa između dva telefona, također koristi se i u kontekstu marketinga te u autoindustriji.

Prednost u praktičnom smislu naspram trenutno popularne bluetooth tehnologije je u tome što NFC ne zahtjeva uparivanje pa je vrijeme ostvarivanja kontakta tj. prijenosa između dva uređaja puno brže. Sigurnosni aspekt je pokriven uvjetovanom blizinom između 2 uređaja, a aplikacije koje rade sa osjetljivim podacima tipa Apple Wallet zahtijevaju identificiranje korisnika prije samog plaćanja.

NFC tehnologija radi na frekvenciji od 13.56Mhz brzinom od 424kbps što je puno sporije od bluetooth tehnologije ali načelno to ograničenje ne stvara problem obzirom da je količina podataka koja se putem NFC uređaja prenosi iznimno mala.

Razlikujemo dva tipa NFC uređaja, aktivni predajnik i pasivni prijamnik. Karakteristika pasivnog prijamnika je što ne treba imati napajanje a to mu daje mogućnost male građe u obliku kartica ili naljepnica (Slika 2). Za vrijeme proizvodnje unaprijed su programirani i sadrže statički sadržaj kojim se putem aktivnog predajnika koji ima spoj na Internet može dinamički upravljati. Razlikuju se po količini podataka koji se na njih može pohraniti te po brzini prijenosa. Tip 1 je najsporiji (106 kbps) i najmanjeg kapaciteta (96B) dok je tip 4 najbrži (424 kbps) i najvećeg kapaciteta (32KB).



Slika 2. Pasivni prijamnik

Predajnik također može raditi u pasivnom modu što je razlog moguće komunikacije između 2 telefona.

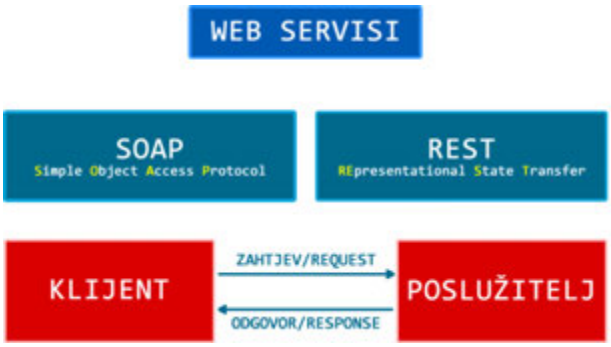
Predajnik i prijamnik rade po principu magnetske indukcije putem koje se napajaju pasivni prijamnici i prenose podaci..

III. WEB SERVISI

Web servisom se smatra svaki servis koji je dostupan putem interneta, neovisan je o platformi i programskom jeziku te koristi standardiziran protokol i format slanja poruka. Web servisi su smješteni na poslužiteljima (engl. *service provider*) i služe za razmjenu podataka između klijenta i poslužitelja. Klijent i poslužitelj komuniciraju putem HTTP (*Hyper Text Transfer Protocol*) protokola slanjem sadržaja u formatu kojeg razumiju i klijent i poslužitelj (XML/JSON).

Poslužitelji uz web servise objavljuju i opis servisa gdje je točno specificirano kako se web servis konzumira i koje funkcionalnosti podržava. Opisi su u XML formatu i nazivaju se WSDL (*Web Services Decscription Language*) dokumenti. Servis za objavljivanje i pronalaženje web servisa naziva se UDDI (*Universal Description, Discovery and Integration*). Može se smatrat online repozitorijem WSDL datoteka koje korisnici mogu pretraživati.

Dva su načina implementacije web servisa, SOAP i REST.



Slika 3.Implementacija web servisa

A. SOAP

SOAP (*Simple Object Access Protocol*) predstavlja protokol, skup strogo definiranih pravila i definicija kojih se moraju držati dva računala koja međusobno komuniciraju putem interneta. Sve poruke se izmjenjuju putem SOAP poruka koje su u XML formatu strukturirane u tri dijela Envelope, Header i Body. Envelope objekt predstavlja ishodišni element koji objedinjuje Header i Body elemente. Header je opcionalan i sadrži stavke koje opisuju kompleksne tipove te sadrži parametre za autentifikaciju, a Body element sadrži konkretne podatke.

B. REST

REST (*REpresentational State Transfer*) danas predstavlja primarnu tehnologiju za izradu web servisa koji je zbog svoje jednostavnosti potisnuo SOAP. Ne predstavlja protokol već način za razvoj web servisa.

Svaki REST servis treba imati slijedeće značajke:

- Zahtjevi bez stanja (*Stateless*) – ne čuvaju se podaci o klijentu prilikom zahtjeva
- Svaki resurs ima jedinstvenu adresu URI (*Uniform Resource Identifier*)
- Resursi se mogu spremiti u keš memoriju
- Može se izraditi u slojevima

Podatak koji se traži se može isporučiti u različitim formatima tipa JSON, XML, JPEG, PDF…

Za pristup resursu koristi se uvijek isti URL, a pomoću HTTP metoda određujemo što sa resursom treba napraviti, npr. metoda može imati vrijednost GET (dohvaćanje podataka), POST (kreiranje podataka), PUT (ažuriranje podataka), DELETE (brisanje podataka)

IV. WEB APLIKACIJA

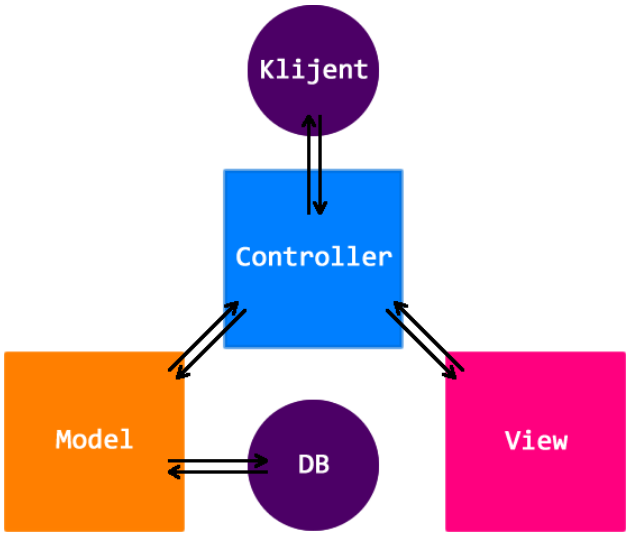
A. MVC

ASP.NET je razvijen od strane Microsofta 2002 godine kao nasljednik ASP tehnologije. Koristi se za izradu dinamičkih web aplikacija te zbog korištenja CLR

(*Common Language Runtime*) omogućava programerima da koriste bilo koji .NET podržan jezik.

MVC je razvojni okvir za izradu web aplikacija koji koristi obrazac Model-View-Controller (Slika 4). Spomenuti obrazac dijeli aplikaciju u 3 dijela:

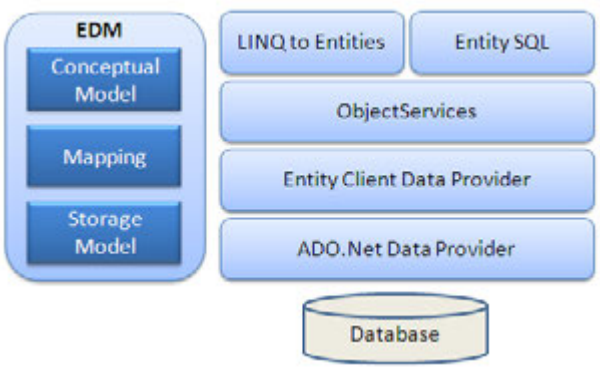
1. Model – predstavlja skup klasa koje predstavljaju podatke s kojima u aplikaciji radimo. Ne implementiraju nikakvu programsku logiku već služe isključivo za prijenos podataka
2. View (Pogled) – predstavlja izgled aplikacije. Obično prikazuje podatke (*Model*) koje mu isporučuje upravljač (*Controller*)
3. Controller (Kontroler/Upravljač) – predstavlja sponu između modela i pogleda, skup klasa koje upravljaju programskom logikom.



Slika 4.MVC arhitektura

B. Entity Framework

Jedna od vodećih tehnologija za pristup podacima iz baze podataka je Entity Framework. Razlog popularnosti leži u tome što iznimno olakšava razvoj aplikacija jer omogućava da se tablicama i zapisima pristupa u obliku programskih objekata i njihovih svojstava. Alat djeluje kao ORM (*Object Relational Mapper*) čiji je cilj podatke i aplikaciju objedinit u jednu cjelinu.



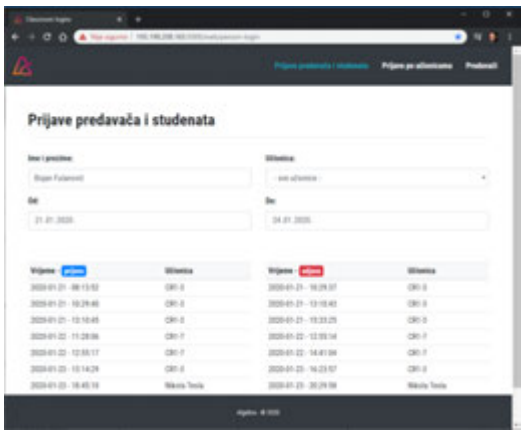
Slika 5.Entity Framework arhitektura

Arhitektura Entity Framework-a može se podijeliti u 3 cjeline (Slika 5):

- Konceptualni model – predstavlja skup klasa i pripadajućih relacija na strani aplikacije
- Model pohrane – predstavlja bazu podataka s pripadajućim tablicama, relacijama, i storanim procedurama
- Mapiranje – skup informacija koje povezuju tablice iz modela pohrane s objektima iz konceptualnog modela

C. React biblioteka

Razvijena je od strane Facebooka 2011 godine predstavlja trenutno najpopularniju biblioteku za razvoj korisničkih sučelja u web aplikacijama. React reagira na promjenu stanja (state) u podacima koje korisničko sučelje prikazuje te automatski ažurira sučelje prema ažuriranim podacima. Ažuriranje sučelja se odvija iznimno brzo zbog manipulacije virtualnog DOM-a. Za razliku od klasičnog DOM-a, virtualni dom predstavlja skup elemenata od kojih se sastoji pojedina komponenta što čini pretragu i ažuriranje elemenata neusporedivo bržim. Mogućnost raščlanjivanja kompleksnog sučelja na manje cjeline (komponente) te fleksibilnost kod izbora tehnologije za obavljanje ostalih aspekta aplikacije koji se ne vežu usko uz sami pogled, olakšava razvoj i održavanje kompleksnih aplikacija što React čini savršenim izborom. Koristeći React biblioteku napravili smo internu aplikaciju putem koje svaki nastavnik može provjeriti da li je uspješno očitao svoju karticu na čitaču (Slika 6). Aplikacija omogućuje filtriranje podataka u zadanom vremenskom razdoblju za određenog predavača ili studenta u određenoj učionici. Koristi se pripremljenim web servisom koji podatke izvlači direktno iz sustava čitača.



Slika 6.Interna aplikacija za pregled prijava



Slika 7.Prikaz prijava u sustavu Infoeduka

V. IMPLEMENTACIJA SPOMENUTIH TEHNOLOGIJA

A. *Ugradnja čitača*

U svaku učionicu ugrađena su po 2 NFC čitača. Jedan za prijavu nastavnika i studenata i drugi za odjavu nastavnika. Oba čitača su spojena na modul koji je zadužen za komunikaciju sa bazom podataka. Prislanjanjem kartice na čitač evidentira se broj kartice u bazi podataka. Baza sadrži ogroman skup predefiniranih tablica od kojih nam je zanimljiva ona koja sadrži zapise o korisnicima koji posjeduju kartice (USERINFO) i tablica u kojoj se prati očitavanje kartica (acc_monitor_log)

B. *Kreiranje WEB-API servisa*

Sustav čitača koristi internu bazu podataka i nema mogućnost komunikacije sa vanjskim servisima. U tu svrhu izradili smo Web-Api servis koji ima pristup bazi podataka čitača. Web servis nudi pristupnu točku (access point - AP) kojoj pristupa sustav za evidenciju polaznika web aplikacija Infoeduka i prije spomenuti interni sustav za evidenciju prijava. Prilikom unosa evidencije Infoeduka komunicira sa web servisom koji joj vraća sve prijave studenata i nastavnika iz svih učionica u zadanom periodu nakon čega se radi filtracija podataka kako bi se izdvojio prijave za konkretnu učionicu. Na osnovu primljenih podataka u sustavu Infoeduke svakom studentu je dodijeljena dolaznost i na predavaču je samo da klikom na gumb pohrani evidenciju u sustav. U slučaju oštećenja ili gubitka kartice, nastavnik ima mogućnost ručnog unošenja dolaznosti. Takav način unosa je evidentiran u sustavu i jasno se vidi kojim studentima je prisutnost unesena putem čitača, a kojima ručno (Slika 7).

VI. ZAKLJUČAK

Sustav očitavanja kartica je implementiran prije nešto manje od godinu dana i možemo slobodno zaključiti da je opravdao sva uložena sredstva. Izbačeni su papiri i prozivke napisanih studenata čime se obzirom na broj studenata po svakom predavanju uštedi oko 10 minuta vremena što ako usporedimo sa ukupnim trajanjem predavanja od 90 minuta iznosi gotovo 10% vremena. Također mogućnost manipulacije s evidencijom prisutnosti je sveden na minimum, a podaci su uredno upisani u bazu podataka što nam daje mogućnost naknadnog pisanja izvještaja dolaznosti ako za time bude potrebe. Ovaj sustav je pokazao da se uvođenjem informacijskih i komunikacijskih tehnologija (ICT) postojeće poslovanje itekako može unaprijediti i biti svrsishodno.

LITERATURA

[1] H. Chang - Everyday NFC Third Edition: Near Field Communication Explained ISBN-13: 978-0982434031, Studeni 2017.

[2] A.Gaurav - Building RESTful Web Services with .NET Core: Developing Distributed Web Services to improve scalability with .NET Core 2.0 and ASP.NET Core 2.0, ISBN-13: 978-1788291576, Svibanj 2018

[3] C. Kirupa - Learning React: A Hands-On Guide to Building Web Applications Using React and Redux (2nd Edition), ISBN-13: 978-0134843551, Svibanj 2018

[4] M. Price - C# 8.0 and .NET Core 3.0 – Modern Cross-Platform Development: Build applications with C#, .NET Core, Entity Framework Core, ASP.NET Core, and ML.NET using Visual Studio Code, 4th Edition, ISBN-13: 978-1788478120, Listopad 2019

[5] https://www.service-architecture.com/articles/web-services/web_services_explained.html (pristupljeno 12.12.2019)

[6] https://www.entityframeworktutorial.net/what-is-entityframework.aspx (pristupljeno 21.12.2019)

Nauči me igrati - igrifikacijom do boljih rezultata

Tamara Ređep*, Tea Pavičić Zajec* i Bogomir Marčinković**
* I. Osnovna škola Varaždin/ Varaždin, Hrvatska
** Osnovna šola Bistrica ob Sotli, Slovenija
tamara.redep@skole.hr; tea232pavi@hotmail.com; bogomir.marcinkovic@guest.arnes.si

***Sažetak* - Aktivna uloga učenika jedna je od glavnih smjernica suvremene nastave pa se zbog toga u nastavi potiče upotreba različitih vrsta igara. Primjena mehanizama, dizajna i elemenata igre u nastavi naziva se igrifikacija. Cilj istraživanja ovog rada bio je ispitati utječe li korištenje edukativne društvene igre potpomognute IKT elementima *„Angles and the Dragon“* na motivaciju za učenje predmeta geografije i povijesti kod učenika sedmih i osmih razreda I. osnovne škole Varaždin. Igra je proizašla kao rezultat Erasmus KA2 projekta *„Teach me how to play“*. Iako nisu potvrđene sve tri postavljene hipoteze, dobiveni rezultati u ovome istraživanju jasan su pokazatelj pozitivnog utjecaja primjene igrifikacije u nastavi kao jedne od mogućih metoda poučavanja koja pridonosi motivaciji učenika i osuvremenjivanju nastavnog procesa.**

Ključne riječi – Erasmus projekt, igrifikacija; motivacaija; osnovna škola; povijest; geografija

I. UVOD

Angažman i uključenost učenika u proces učenja od velike su važnosti za ostvarivanje ciljeva odgojno-obrazovnog sustava, jer utjecajem na motivaciju učenika utječe se i na njihov angažman u ostvarenju cilja [1]. Međutim, obrazovni stručnjaci moraju biti svjesni da se promjenom generacija učenika, mijenjaju i stilovi učenja i trendovi u nastavi [2]. Iz tog razloga njihova uloga se značajno mijenja te umjesto tradicionalnog, didaktičkog podučavanja, učitelj priprema suvremeno okruženje i zadatke za samostalno učenje kroz koje vodi učenike i daje im konstruktivne povratne informacije [3].

Prema priručniku za Primjenu scenarija poučavanja, digitalnih alata i obrazovnih trendova, suvremene pedagoške metode usmjerene su na učenika, njegove potrebe, interese i motive, pri čemu potiču kreativnost, komunikaciju, samostalnost i kritičko mišljenje te koriste elemente zabave i igre u poučavanju [4]. Zahvaljujući elementima igre koji su preuzeti iz različitih vrsta igara i adaptirani u nezabavan kontekst nastave, vjeruje se da će učenici gledati na odgojno – obrazovne ciljeve kao nešto dinamično, zabavno i proaktivno [1].

U pedagogiji se igra smatra sredstvom odgoja i obrazovanja, „slobodna, samomotivirajuća, nesvrhovita djelatnost u kojoj sredstva dominiraju nad ciljem“ [5].

Neke od vrsta igara:

- *igra s pravilima* (engl. game) - regulirana je pravilima igranja i pojedinci se natječu kako bi ostvarili određeni cilj ili ciljeve uz ograničene resurse (npr. kartanje, šah i sl. društvene igre), a primarna joj je svrha zabava.
- *ozbiljna igra* (engl. serious game) - koristi se u ozbiljne svrhe, npr. obrazovne i poslovne svrhe, a ponajprije se odnosi na korištenje računalnih igara koje su u potpunosti izrađene u svrhu učenja i stjecanja vještina, a ne u svrhu zabave (npr. igra koja osigurava trening zaposlenika) [5].

Korištenje igre ili elemenata igre u nastavnom procesu s ciljem ostvarivanja unaprijed zadanih odgojno-obrazovnih ishoda naziva se igrifikacija (engl. Gamification) [4].

II. IGRIFIKACIJA U OBRAZOVANJU

Igrifikacija je od velikog značaja u ostvarivanju ciljeva i zadaća u budućnosti cjelokupnog sustava odgoja i obrazovanja [2]. Primjenjuje se kako bi se unaprijedilo učenje i rješavanje problema te utjecalo na aktivnosti i motivaciju učenika. Time bi učenje za učenika predstavljalo ugodan proces, što bi u konačnici rezultiralo usvajanjem znanja [6]. To se može odnositi na korištenje općepoznatih igara u nastavi, na osmišljavanje igara namijenjenih učenju točno određenih sadržaja ili razvijanju određenih vještina te na uporabu edukativnih digitalnih igara [4].

Primjena igrifikacije u odgojno-obrazovnom sustavu treba biti pomno promišljena, a aktivnosti dobro analizirane i pripremljene kako njena primjena ne bi izazvala negativan učinak, te kako bi implementacija bila što kvalitetnija [1]. Svaka igra koja se koristi u poučavanju i učenju uvijek mora biti u funkciji ostvarivanja postavljenog odgojno-obrazovnog cilja i u dogovorenim vremenskim okvirima [4]. Za što učinkovitiji rezultat potrebno je voditi računa o pravilnom balansu učenja i elementa zabave tijekom nastavnog procesa, ali i usklađenosti pedagoških i psiholoških načela [1].

Proces integracije načela dizajniranja igara unutar različitih obrazovnih iskustava predstavlja izazov za nastavnike i potrebne su praktične smjernice kako to učiniti skladno i na učinkovit način [7]. Guido M. [8] predlaže pet koraka za uspješnu primjenu učenja temeljenog na igrama u razredu.

To su:

1. utvrditi razloge korištenja igre u nastavi,
2. prethodno testirati igru i provjeriti je li u skladu s obrazovnim ciljevima,
3. osigurati da ispunjava očekivanja roditelja,
4. odvojiti dovoljno vremena za igranje,
5. procijeniti napredak učenika kroz igru.

Vodeći se ovim koracima, prije nego odluče koristiti igru u nastavi, nastavnici moraju utvrditi za što je žele koristiti, kako bi zadovoljila i njihove potrebe i potrebe učenika. Svrha igre tako može biti olakšavanje učenja učenicima koji postižu slabije uspjehe u ostvarenju ishoda učenja, obogaćivanje nastavnih sadržaja kroz integraciju različitih vrsta medija u igrama (postavljanje pitanja putem teksta, zvuka ili slike), ili pak poticanje učenika na grupni rad i međusobnu suradnju unutar razreda.

Preporuka je da nastavnici prije primjene i sami odigraju igru te tako provjere je li ona doista usklađena s ishodima učenja. Igrajući igru potrebno je obratiti pažnju na mogućnost kontrole sadržaja od strane nastavnika, prilagodbe sadržaja obzirom na sposobnost i dob svakog igrača, jednostavnost upotrebe igre i jasnoću njenih uputa te mogućnost implementacije različitih vrsta sadržaja kako bi nastava učenicima bila što zanimljivija.

Roditelji bi također trebali biti informirani o planovima učenja i uvođenja metode igrifikacije u nastavu. Iskustva pokazuju da angažman roditelja i učenika u planiranju nastavnog procesa od strane škola značajno utječe na povećanje učinka ostvarenja ishoda učenja, kao i aktivnosti učenika u školi. Na taj način, roditelji će biti upoznati sa pravom svrhom igre u obrazovnom procesu, kao i dobrobitima koje ona donosi.

Nastavnici bi trebali unaprijed planirati i odrediti vrijeme igranja igre te u tome trebaju biti dosljedni. Na taj će način lakše ostvariti definirani cilj nastave, a učenici će imati mogućnost ostvariti ishode učenja.

Tijekom igranja igre potrebno je prikupljati informacije o tijeku igre te tako procijeniti napredak učenika, kako bi se na vrijeme otkrile poteškoće na koje učenik nailazi. Prikupljanje podataka može se ostvariti kroz metode izvještaja u igri, samoprocjene učenika, rasprave na satu o poteškoćama, napretku i postignućima. Sve to pomaže u oblikovanju nastave u razredu jer pruža informacije potrebne za prilagodbu lekcija i aktivnosti te rješavanje problema i mogućnost nadogradnje na novo znanje [8]. Također, to može biti jedan od primjera moguće primjene igre za provedbu formativnog vrednovanja.

A. Prethodna istraživanja i rezulutati

Koncept igrifikacije izaziva zanimanje znanstvenika zbog svoje prirode kojom se nastoji nešto ozbiljno učiniti zanimljivim te se provode mnoga istraživanja [1].

Izučavajući koncept igrifikacije u obrazovanju Medica Ružić i Dumančić u opisu povijesti igrifikacije u obrazovanju iznose rezultate nekoliko relevantnih istraživanja. Spominju rezultat istraživanja Randelovog tima[1] u kojem 36% ispitanika više vole igre, dok 5% preferira klasičnu nastavu. Nadalje, iznose rezultate istraživanja Wolfa[2] koji ukazuju na to da je pristup zasnovan na igrama stvorio značajan porast razine znanja u odnosu na konvencionalne metode poučavanja [2].

Plantak Vukovac, Škara i Hajdin u sklopu svog provedenog istraživanja o korištenju i stavovima nastavnika o igrifikaciji u osnovnim i srednjim školama u analizi rezultata uočili su da većina učitelja nije upoznata s konceptom igrifikacije. Glavni razlog tome je nedostatak izobrazbe koji bi obuhvaćao i upoznavanje sa navedenim konceptom. Zaključili su i da je većina nastavnika zainteresirana za primjenu metoda igrifikacije primarno radi povećanja pažnje i aktivnosti učenika te unapređenja kvalitete nastavnog procesa [6].

Sánchez-Rivas, Ruiz-Palmero i Sánchez-Rodríguez [9] u svojoj su studiji produbili percepciju nastavnika o dva modela ocjenjivanja koji se koriste u nastavi u osnovnom obrazovanju: tradicionalnom ispitu i ispitu temeljenom na igrifikaciji. Potvrdili su da provjera znanja zasnovana na igri, iz perspektive znanstvene didaktike, ima značajne prednosti u odnosu na tradicionalno ispitivanje provjere znanja.

Ortiz, Chiluiza, i Valcke [10] fokusirali su se na korištenje igrifikacije u visokom obrazovanju u STEM području. Analizom 30 znanstvenih radova utvrdili su da se igrifikacija većinom primjenjuje u polju računalnih znanosti, najčešće kao kombinacija elemenata igre (bodovi, bedževi i rang liste) te da pozitivno utječe na studentsku aktivnost.

No, Taspinar, Schmidt i Schuhbauer u svom istraživanju [11] rezultatima ankete otkrili su da većina učenika, njih 67%, svrstavaju igre na ploči kao broj jedan među igrama koje igraju i općenito im pridodaju veću popularnost u odnosu na računalne igre. Stoga, suprotno velikoj većini projekata digitalne igrifikacije njihova je ideja bila stvoriti igru na ploči koja se temelji na tipičnoj igrifikacijskoj mehanici. Stvorena igra na ploči je bila zamišljen kao nastavni elemenat koji će se uklopiti u interaktivni nastavni koncept. Po završetku primjene igre, i učenici i nastavnik potvrdili su njezin pozitivan učinak podizanja motivacije za učenje.

III. PROJEKT „NAUČI ME IGRATI“

Erasmus+ projekt „Nauči me igrati“ („*Teach me how to play*“)[3] usmjeren je na razvoj novih metoda poučavanja temeljenih na igrifikaciji kako bi kroz izazove igre učenje postalo zabavno, a kod učenika se potaknuo razvoj kreativnosti i znatiželje [12].

Projekt je, 2018. godine bio ocijenjen od strane Agencije za mobilnost i programe EU s najviše bodova od svih prijavljenih projekata u Hrvatskoj. Ukupnu vrijednost

[1] Randel, J.M., Morris, B.A., Wetzel, C.D. i Whitehill, B.V. (1992). The effectivness of games for educational purposes: a review of recent research. Simulation and gaming, 23(3), 261-276.

[2] Wolfe, J. (1997). The effectiveness of bussiness game sin strategic management course work. Simulation & Gaming, 28(4), 360-376.

[3] http://os-prva-vz.skole.hr/erasmus_kutak/erasmus_ka2_hrvatska

projekta od 89.280,00 € financira Europska Unija, a uz I. OŠ Varaždin koja je nositelj ovog dvogodišnjeg projekta, u njemu sudjeluju učenici i učitelji OŠ Bistrice ob Sotli iz Slovenije i CEIP Bernardino Perez iz Španjolske. Projektni timovi svih škola sudionica zajedno rade na realizaciji projektnih aktivnosti čija je glavna svrha osmišljavanje, izrada i testiranje društvenih edukativnih igara potpomognutih IKT tehnologijom u odgojno – obrazovno procesu.

Cilj svake igre je osuvremeniti nastavni proces iz područja geografije, povijesti te kulture i umjetnosti kako bi učenici na taj način lakše i brže savladali nastavne sadržaje iz predmeta koji su im teški i nezanimljivi zbog mnoštva informacija i činjenica, a što bi u konačnici rezultiralo postizanjem boljih obrazovnih rezultata. Istraživačkim načinom rada, učenici razvijaju svoju kreativnost te tako postaju kreatori edukativnih digitalnih sadržaja, dok su učitelji samo u ulozi mentora. Ujedno, tijekom mobilnosti učenici i učitelji upoznaju igru projektnog partnera domaćina te tako na zanimljiv način uče o povijesnim, geografskim i kulturološkim obilježjima zemlje svojih partnerskih škola, pri čemu ujedno rade i na usavršavanju jezičnih kompetencija.

Kako bi svi materijali, fotografije i upute vezane uz igre bile dostupne što većem broju učitelja otvoren je eTwinning projekt, a svi materijali biti će postavljeni na Twinspaceu[4]. Proces diseminacije rezultata vršiti će se i na web stranici I. OŠ Varaždin, diseminacijskoj platformi te putem medija.

A. Igra „Angels and the Dragon“

U sklopu projekta „Nauči me igrati“ („*Teach me how to play“)* ukupno će biti izrađeno 6 društveno edukativnih igara, po dvije igre od svake partnerske škole sudionice projekta, a „*Angels and the Dragon*“[5] je prva igra izrađena od strane hrvatskog tima.

Kao što je i definirano postavljenim ciljevima projekta, igra obuhvaća obrazovne nastavne sadržaje iz predmeta povijest, geografija, ali i matematike i kemije za 7. i 8. razred osnovne škole, a potpomognuta je IKT – om (upotreba web 2.0. alata, QR kodova, proširene stvarnosti).

Svi sastavni dijelovi igre izrađeni su ručno i rezultat su marljivog timskog rada 12 učenika sedmih i osmih razreda te 8 članova projektnog tima. Vanjski izgled igre u obliku je drvenog, ručno izrađenog i oslikanog sanduka (Slika 1.) koji u svojoj unutrašnjosti skriva igraču ploču sa zemljopisnom kartom Republike Hrvatske i mjeračem vremena. S gornje strane poklopca sanduka nalaze se elementi potrebni za igranje igre (polazna priča i pravila igre, 3D naočale, slijepe karte, blok za pisanje i računanje, igrače figure, kartice sa zadacima) (Slika 2.). Obzirom da je službeni jezik projekta engleski i pitanja na karticama napisana su na engleskom jeziku, kako bi ih svi sudionici projekta mogli razumjeti i riješiti (Slika 3.)



Slika 1. Vanjski izgled sanduka s igrom



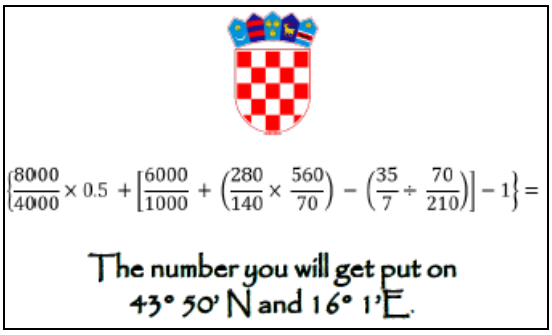
Slika 2. Unutrašnjost sanduka – ploča za igranje s pripadajućim elementima



Slika 3. Primjer pitanja na kartici – integracija IKT-a

[4] Twinspace – virtualni platforma za suradnju registriranog eTwinning projekta; Izvor: IGI Global, Dictionary https://www.igi-global.com/dictionary/learning-one-another-etwinning-project/30750

[5] http://os-prva-vz.skole.hr/erasmus_kutak/erasmus_ka2_hrvatska/game



Slika 4. Primjer pitanja na kartici – korelacija matematike i geografije

Igra je predviđena je za najmanje 2, a najviše 6 igrača starijih od 12 godina. Igrači postaju anđeli kojima je glavna misija zauvijek protjerati zmaja te spasiti Varaždin od požara i potpunog uništenja. Rađena je na principu „*room escape*" pri čemu učenici timski rješavaju probleme i pronalaze tragove, a za ostvarenje svoje misije moraju u roku od 30 minuta uspješno riješiti 10 zadataka i isključiti 10 žarulja koje su upaljene oko Varaždina.

IV. METODOLOGIJA ISTRAŽIVANJA

Obzirom na rezultate prethodnih istraživanja o pozitivnim učincima igrifikacije u nastavi, te prema preporuci priručnika za Primjenu scenarija poučavanja, digitalnih alata i obrazovnih trendova [4] da se povežu i integriraju sadržaji učenja sa svakodnevnim životnim okolnostima te da se poučavanje i učenje odvija uz primjenu IKT-a, cilj istraživanja opisanog u ovome radu bio je ispitati postoji li utjecaj korištenja izrađene edukativne društvene igre potpomognute IKT elementima na motivaciju za učenje predmeta geografije i povijesti kod učenika u osnovnoj školi.

Istraživanje je provedeno na početku školske godine 2019./2020. na uzorku od 86 učenika (44,2% ženskih i 55,8% muških ispitanika) I. Osnovne škole Varaždin. Ispitanici su bili učenici sedmih (65,1%) i osmih razreda (34,9%) obzirom da su oni prethodno obradili nastavne sadržaje zastupljene u igri.

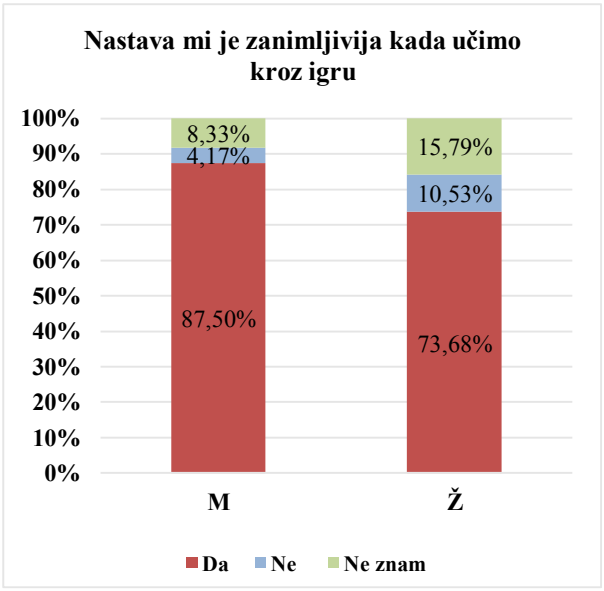
Za potrebe istraživanja kreiran je online anketni upitnik koji se sastojao od pitanja zatvorenog tipa, a kojim se željelo ispitati podržavaju li još uvijek učenici korištenje tradicionalnih načina poučavanja u školi ili bi željeli uvesti neke novine u nastavu kao što je metoda igrifikacije, te postoji li razlika u interesu za uvođenje igre kod djevojčica u odnosnu na dječake. Ujedno, željelo se utvrditi i mišljenje učenika o lakoći usvajanja nastavnih sadržaja iz povijesti i geografije uz pomoć edukativnih igara te ispitati utječe li njihovo korištenje na zanimljivost nastave te samim time i na stupanj motivacije učenika za sudjelovanje na nastavi i učenje.

U radu se polazi od sljedećih hipoteza:

1. (H1) Kada uče kroz igru nastava je podjednako zanimljiva i dječacima i djevojčicama.
2. (H2) Nastavne sadržaje iz povijesti lakše je usvojiti uz pomoć edukativnih igara.
3. (H3) Učenici koji nastavne sadržaje iz geografije uče pomoću igre, motiviraniji su od učenika koji uče tradicionalnim načinima poučavanja.

V. REZULTATI ISTRAŽIVANJA

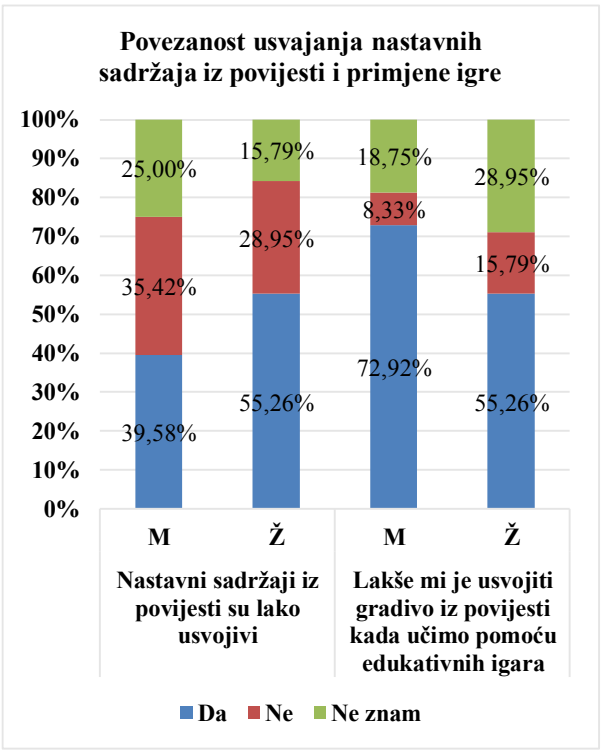
Pri odgovaranju na pitanje je li im nastava zanimljivija kada uče kroz igru 81,4% učenika odgovorilo je potvrdno, 11,6% njih ne zna odgovor, dok je 7% ispitanika odgovorilo negativno. Detaljnijom analizom po spolu utvrđeno je da se od ukupnog broja ispitanih dječaka njih 87,5% slaže s tom tvrdnjom, a od ukupnog broja ispitanih djevojčica tvrdnju potvrđuje 73,68% (Slika 5.) Obzirom da ne postoji statistički značajna razlika između muških i ženskih ispitanika ovi podaci potvrđuju prvu hipotezu (H1) koja govori da je nastava podjednako zanimljiva i dječacima i djevojčicama kada uče kroz igru.



Slika 5. Postotak učenika prema spolu kojima je nastava zanimljivija uz igru N=86

Nastavni sadržaji iz povijesti lako su usvojivi za 46,5% ispitanika, teško usvojivi za njih 32,6%, dok 20,9% učenika ne zna odgovor na to pitanje. Da im je lakše usvojiti gradivo iz povijesti kada uče pomoću edukativnih igara, potvrdno se izjasnilo 65,1% učenika, 11,6% je negiralo tu tvrdnju, dok 23,3% njih ne zna odgovor. Obzirom na to da se većina učenika izjasnila potvrdno, dobiveni rezultati potvrđuju drugu hipotezu (H2). Ako pak analiziramo podatke prema spolu tada je gradivo iz povijesti lako usvojivo za 55,26% djevojčica i za samo 39,58% dječaka. Stavljajući ovu tvrdnju u odnos s tvrdnjom da im je lakše usvojiti gradivo iz povijesti

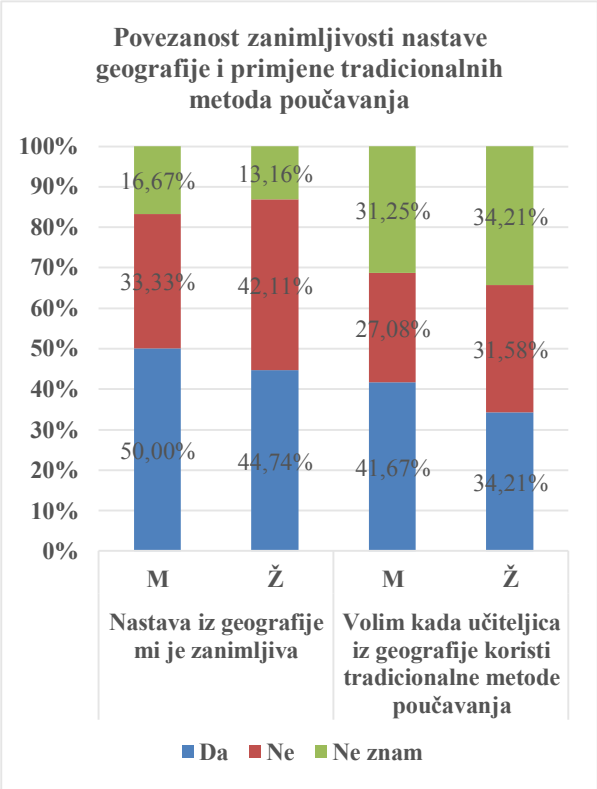
kada uče uz pomoć igara, vidljivo je da to želi više dječaka, njih čak 72,92% u odnosu na 55,26% djevojčica (Slika 6.). Ovaj odnos možemo tumačiti na način da su dječaci općenito skloniji igranju igara te stoga smatraju da će im igre pomoći u savladavanju nastavnih sadržaja iz povijesti.



Slika 6. Povezanost usvajanja nastavnih sadražaja iz povijesti i primjene igre, obzirm na spol, N=86

Nastava iz geografije zanimljiva je 47,7% učenika, 37,2% to negira, dok je 15,1% ispitanika neodlučno. Kada učiteljica iz geografije koristi tradicionalne metode poučavanja još uvijek voli 38,4% učenika, 29,1% to ne voli, a 32,6% njih ne zna. Iako je prethodno utvrđeno kako je učenicima nastavne sadržaje iz povijesti lakše usvojiti uz pomoć edukativnih igara, ovi dobiveni podatci upućuju nas na činjenicu da učenici još uvijek vole kad na nastavi geografije učiteljica koristi tradicionalne metode poučavanja. To znači da primjena igre na nastavi geografije nema statistički značajan utjecaj na motiviranost učenika čime je opovrgnuta treća hipoteza (H3).

Detaljnija analiza po spolu pokazuje nam da je nastava iz geografije podjednako zanimljiva i dječacima i djevojčicama (50,0% dječaka i 44,74% djevojčica), te gotovo isto tako podjednako još uvijek vole primjenu tradicionalnog načina poučavanja (41,67% dječaka i 34,21% djevojčica) (Slika 7.).



Slika 7. Povezanost zanimljivosti nastave geografije i primjene tradicionalnih metoda poučavanja, obzirom na spol, N=86

VI. ZAKLJUČAK

Kako bi se ostvario koncept suvremene škole, pokušavaju se pronaći nove tehnike i pristupi u svrhu povećanja aktivnosti i samostalnosti učenika u procesu učenja [1]. Suvremene pedagoške metode koriste elemente zabave i igre u poučavanju [4], a nove obrazovne tehnologije u obrazovanju potiču i stvaraju obrazovno okruženje unutar kojeg je moguće koristiti igrifikaciju [2].

Igrifikacija predstavlja upotrebu mehanizama, dizajna i elemenata igre u ne zabavnom okruženju, a sve to u svrhu promicanja motivacije, truda i lojalnosti [6]. Koncept igrifikacije kao novitet primjenjuje se ovisno o volji i želji nastavnika koji će, prema vlastitom mišljenju, igrificirati svoje nastavne sate [1] uz svrhovito i strukturirano planiranje.

Pregledom literature uočeno je da igre poboljšavaju široku paletu vještina, uključujući socijalne, emocionalne, kulturne, intelektualne i kritičke vještine [8], a to je ujedno bio poticaj za provedbu istraživanja opisanog u ovome radu.

Cilj istraživanja bio je ispitati postoji li utjecaj korištenja edukativne društvene igre „*Angles and the Dragon*“ potpomognute IKT elementima na motivaciju za učenje predmeta geografije i povijesti kod učenika sedmih i osmih razreda I. osnovne škole Varaždin. Igra je proizašla kao rezultat projektnih aktivnosti Erasmus KA2 projekta „*Teach me how to play*“.

Obzirom da je 87,5% dječaka i 73,68% djevojčica odgovorilo potvrdno na tvrdnju da im je nastava zanimljivija kada uče kroz igru, dobiveni rezultati istraživanja potvrđuju prvu hipotezu (H1 – „*Kada uče kroz igru nastava je podjednako zanimljiva i dječacima i djevojčicama"*). Isto tako potvrđena je i druga hipoteza (H2 – „*Nastavne sadržaje iz povijesti lakše je usvojiti uz pomoć edukativnih igara*") jer je većina učenika (65,1%) potvrdila izjavu da je nastavne sadržaje iz povijesti lakše usvojiti uz pomoć edukativnih igara, a uz dodatnu analizu po spolu to je izjavilo više dječaka (72,92%) nego djevojčica (55,26%). Ako to stavimo u odnos s činjenicom da je dječacima gradivo iz povijesti teže usvojivo nego djevojčicama, tada to možemo tumačiti na način da dječaci smatraju da će im igre pomoći u savladavanju nastavnih sadržaja iz povijesti. Iako potvrđenost prvih dviju hipoteza logički nameće pretpostavku da će potvrđena biti i treća hipoteza (H3 – „*Učenici koji nastavne sadržaje iz geografije uče pomoću igre, motiviraniji su od učenika koji uče tradicionalnim načinima poučavanja"*) ona je ipak opovrgnuta. Naime, 38,4% učenika još uvijek voli kada učiteljica iz geografije koristi tradicionalne metode poučavanja, 29,1% njih to ne voli, a 32,6% ispitanika ostalo je neodlučno.

Iako nisu potvrđene sve tri postavljene hipoteze, dobiveni rezultati u ovome istraživanju jasan su pokazatelj pozitivnog utjecaja primjene igrifikacije u nastavi kao jedne od mogućih metoda poučavanja koja pridonosi motivaciji učenika i osuvremenjivanju nastavnog procesa.

LITERATURA

[1] M. Škara, "Primjena igrifikacije u obrazovanju", Diplomski rad, Sveučilište u Zagrebu, Fakultet organizacije i informatike, Varaždin, 2017. [29.09.2019.]. Dostupno na:
https://urn.nsk.hr/urn:nbn:hr:211:239578

[2] I. Medica Ružić i M. Dumančić, "GAMIFICATION IN EDUCATION", Informatologia, vol.48, br. 3-4, str. 198-204, 2015. [19.09.2019.]. Dostupno na:
https://hrcak.srce.hr/151675.

[3] Franković, "Učenje temeljeno na didaktičkim računalnim igrama". Rijeka: Sveučilište u Rijeci, Odjel za informatiku. 2016. [29.09.2019.]. Dostupno na:
https://www.inf.uniri.hr/files/studiji/poslijediplomski/kvalifikacijski/Frankoivc_Ivona_Kvalifikacijski_rad.pdf

[4] I. Mišurac, „Priručnik: Primjena scenarija poučavanja, digitalnih alata i obrazovnih trendova", Zagreb, CARNet, 2017. [01.10.2019.]. Dostupno na:
https://pilot.e-skole.hr/wp-content/uploads/2016/12/Prirucnik_Scenarij-poucavanja.pdf

[5] S. Lovrenčić,, D.P. Vukovac., B. Šilbar, B. Nahod, D. Andročec, M. Šestak, Z. Stapić, Igrifikacija: prema sistematizaciji termina na hrvatskom jeziku, 2018. [17. 09.2019.]. Dostupno na:
https://www.researchgate.net/publication/327792169_Igrifikacija_prema_sistematizaciji_termina_na_hrvatskom_jeziku

[6] D. Plantak Vukovac, M. Škara i G. Hajdin, "KORIŠTENJE I STAVOVI NASTAVNIKA O IGRIFIKACIJI U OSNOVNIM I SREDNJIM ŠKOLAMA", Zbornik Veleučilišta u Rijeci, vol.6, br. 1, str. 181-196, 2018. [20.10.2019.]. Dostupno na:
https://doi.org/10.31784/zvr.6.1.14

[7] C. Dichev, D. Dicheva, "Gamifying education: what is known, what is believed and what remains uncertain: a critical review. International Journal of Educational Technology in Higher Education.", 2017. [03.10.2019.]. Dostupno na:
https://educationaltechnologyjournal.springeropen.com/articles/10.1186/s41239-017-0042-5

[8] M. Guido, "5 Steps to Implementing Game-Based Learning in the Classroom [+ Game Examples & Downloadable Guide], 2016. [05.10.2019.]. Dostupno na:
https://www.prodigygame.com/blog/implementinggame-based-learning-in-the-classroom-examples/

[9] E. Sánchez-Rivas, J. Ruiz-Palmero, J. Sánchez-Rodríguez, "Gamification of Assessments in the Natural Sciences Subject of Primary Education", Educational Sciences: Theory & Practice, 2019., 19(1), 95-111. [27.09.2019.]. Dostupno na:
http://dx.doi.org/10.12738/estp.2019.1.0296

[10] M. Ortiz, K. Chiluiza, M. Valcke, "GAMIFICATION IN HIGHER EDUCATION AND STEM: A SYSTEMATIC REVIEW OF LITERATURE, Proceedings of EDULEARN16 Conference, p. 6548-6558., 2016. [27.09.2019.]. Dostupno na:
https://core.ac.uk/download/pdf/147027732.pdf

[11] B. Taspinar, W. Schmidt, H. Schuhbauer, "Gamification in Education: A Board Game Approach to Knowledge Acquisition", Procedia Computer Science, Volume 99, 2016, Pages 101-116, 2016. [01.10.2019.]. Dostupno na:
https://doi.org/10.1016/j.procs.2016.09.104

[12] T. Pavičić Zajec, T. Ređep, "Angels and the Dragon", Pogled kroz prozor, br.113, svibanj 2019. [19.09.2019.]. Dostupno na:
https://pogledkrozprozor.wordpress.com/2019/05/31/angels-and-the-dragon/

Vrednovanje digitalnih kompetencija putem CRISS projekta

Marina Mirković
Tehnička škola, Požega, Republika Hrvatska
marina.mirkovic@skole.hr

Sažetak

Tehnička škola u Požegi u školskoj godini 2018./19. bila je uključena u međunarodni CRISS projekt (Demonstration of a scalable and cost-effective cloud-based digital learning infrastructure through the Certification of digital competences in primary and secondary schools). Škola je u projekt uključena posredstvom CARNET-a koji je partner na projektu CRISS zadužen za pilotiranje CRISS modela digitalnih kompetencija i platforme na razini Hrvatske.
Svrha je projekta razvoj i implementacija digitalne platforme za stjecanje, vrednovanje i certificiranje digitalnih kompetencija učenika u osnovnim i srednjim školama. Testiranje CRISS platforme provodilo se je kroz masovni pilot u nekoliko europskih zemalja. U Hrvatskoj je sudjelovala 21 škola s preko 1000 učenika i 92 nastavnika.
Rješavanjem zadataka iz različitih scenarija učenja, učenici postižu određene podkompetencije.
CRISS-ov okvir za vrednovanje digitalnih kompetencija detaljno je razrađen i na jednostavan način pomaže nastavnicima u analizi i procjeni učeničkih kompetencija.
Projekt je olakšao nastavni proces i učenje, kao i vrednovanje. Uz primjenu multimedije i vanjskih resursa potaknut je interes učenika, pojednostavljeno učenje i razvijeni suradnički odnosi.

Ključne riječi – digitalne kompetencije, CRISS projekt, CRISS platforma, scenarij učenja, podkompetencije

I. UVOD

Tijekom godina, infrastruktura i društvo pretrpjeli su mnoge promjene, pa tako i kompetencije koje današnji učitelj mora razviti kod učenika. Sve smo prisutniji na internetu, te je naša odgovornost da budemo svjesni važnosti te „prisutnosti“, kako građana tako i profesionalaca, u skladu s ulogom koju imamo u zajednici.

Prema Organizaciji za ekonomsku suradnju i razvoj (OECD), kompetencija je više od samog znanja i vještina. Ona uključuje sposobnost suočavanja i rješavanja složenih zahtjeva nadograđivanjem psihosocijalnih resursa (uključujući vještine i stavove) u određenom kontekstu [1].

Digitalne kompetencije ključne su za učenje, rad i aktivno sudjelovanje u društvu. Europska komisija je 2013. godine objavila Okvir za razvoj i razumijevanje digitalnih kompetencija u Europi (DIGCOMP), koji je obnovljen Okvirom digitalnih kompetencija za građane (DigComp 2.1) [2]. Digitalne kompetencije su jedna od osam kompetencija koje je potrebno usvojiti završetkom obveznog obrazovanja. One će omogućiti učenicima uspješan početak rada u njihovu zanimanju te nastavak cjeloživotnog obrazovanja [3].

Zajednički okvir sa standardima i indikatorima koji omogućuju vrednovanje i certifikaciju digitalnih kompetencija ne postoji. Stoga su i različiti planovi za njihov razvoj u školama. Razvojem digitalne platforme u CRISS projektu nastoji se riješiti ovaj problem. Ona omogućuje implementaciju metodološkog okvira za certifikaciju digitalnih kompetencija u Europi.

II. CRISS PROJEKT

Svrha je projekta razvoj i implementacija digitalne platforme za stjecanje, vrednovanje i certificiranje digitalnih kompetencija učenika u osnovnim i srednjim školama. Testiranje CRISS platforme provodilo se je kroz masovni pilot s više od 490 škola, uključujući 25 400 učenika i 2290 nastavnika u nekoliko europskih zemalja [4]. U Hrvatskoj je sudjelovala 21 škola s preko 1000 učenika i 92 nastavnika. [5] U Tehničkoj školi u Požegi u projekt je bilo uključeno 53 učenika i dva nastavnika.

CRISS projekt počeo je 1. siječnja 2017. godine, a trajao do 31. prosinca 2019. godine. Ukupna vrijednost projekta je oko 6 milijuna eura, od čega je Europska komisija financirala 4,87 milijuna eura. Koordinator CRISS projekta je britanska kompanija EXUS Software LTD. U projektu je sudjelovalo četrnaest partnera među kojima su javne ustanove, sveučilišta, mala i srednja, te velika poduzeća iz osam europskih zemalja. Hrvatski partneri na projektu su Fakultet organizacije i informatike Sveučilišta u Zagrebu, koji ima ključnu ulogu u pripremi i provedbi projekta te Hrvatska akademska i istraživačka mreža – CARNET. Projekt se financira iz okvirnog programa Europske unije za istraživanja i inovacije Obzor 2020 [5].

III. CRISS PLATFORMA

Putem CRISS platforme učenici razvijaju digitalne kompetencije kroz nastavu različitih predmeta. Nastavnici praćenjem razvoja učeničkih kompetencija potiču učenike na daljnji razvoj tijekom učenja.

CRISS-ovoj online platformi može se pristupiti putem postojećih preglednika bez potrebe za instalacijom vanjske programske podrške. Učenici mogu pristupiti platformi putem interneta kako bi stekli digitalne vještine tijekom prolaska predmetnog gradiva. Istovremeno sustav

nastavnicima omogućuje prikupljanje dokaza i vrednovanje učeničkih radova.

Scenarij procjene kompetencije (CAS - Competence Assessment Scenario) je scenarij u kojem je učenik središnji element kod rješavanja problema, razvoja projekata ili traženja rješenja u realnim kontekstima i smislenim situacijama. U CRISS-u CAS predlaže skup aktivnosti i zadataka koji omogućuju ocjenu jednog ili više kriterija uspješnosti. Integracijske aktivnosti omogućuju ocjenu različitih kompetencija i predmeta u određenom scenariju učenja. Specifičan zadatak omogućuje zaključivanje dijela aktivnosti (npr. izračunati, kopirati, stvoriti, preuzeti, urediti, razraditi, podijeliti, prenijeti itd.) To je jasno definirano djelo, ponekad kratkog ili ograničenog trajanja, dodijeljeno ili očekivano od pojedinca.

CRISS platforma (slika 1.) omogućuje tijek rada potreban za planiranje, postavljanje i rad CAS-a. Nastavnik i učenici prolaze slijedeće korake:

• Nastavnik planira, dodjeljuje sadržaj i evaluacije.

• Učenik prima obavijest, čita zadatak i stvara dokaz/ priču. Šalje dokaz na procjenu.

• Nastavnik prima obavijest i procjenjuje učenikove dokaze prema konfiguraciji rubrike koju je unaprijed definirao u planiranju [6].



Slika 1. Alati CRISS platforme

Učenicima i nastavnicima u projektu bilo je zanimljivo i korisno iskustvo sudjelovanja u testiranju digitalne platforme. Koristili su 10 scenarija poučavanja (slika 2.) s raznim aktivnostima koje nastavnik može prilagođavati svojim potrebama, a povezane su s različitim školskim predmetima (povijest, geografija, informatika, matematika, biologija).



Slika 2. CRISS scenariji poučavanja

Sa stranice „Zadatak za rješavanje“, učenici prilikom klika mišem pristupaju zadatku kako bi ga mogli analizirati i izvršiti putem dokaza. Učenici dokaz mogu izmijeniti sve dok ga nisu poslali na ocjenjivanje.

Na dnu kartice dokaza nalazi se izbornik za slanje dokaza učitelju na ocjenu. U svakom zadatku koji mu dodijeli učitelj, učenik pronalazi pridruženi gumb za Izradu dokaza /priče [7].

Priča ili dokaz je djelo koje je učenik izradio prema zadatku koji mu je dodijelio učitelj, a koje je potrebno predati kako bi se procijenio proces njegova učenja.

U sustavu postoji set IKT alata. To je potpuno integrirani skup IKT alata i resursa za jednostavno stvaranje, obradu i ponovnu upotrebu različitih oblika digitalnog sadržaja, usluga, aplikacija i kontekstualnih podataka za interaktivno učenje. Omogućuje nastavnicima da ih uključe u svoje planirane tečajeve i lekcije. Učenici ih koriste za samostalne ili zajedničke radove vezane uz zadatke koje im zadaju nastavnici.

Pri rješavanju zadataka učenici su koristili i upoznali različite digitalne alate pri čemu su izrađivali umne mape, vremenske trake, stripove (slika 3.), plakate te zajedno u grupama izrađivali prezentacije i dokumente. Učenici su aktivno sudjelovali u izradi zadataka te pokazali iznimnu zainteresiranost za prikazane radove. Osim toga, sudjelovali su u samorefleksiji i vršnjačkoj refleksiji vrednovanjem vlastitih digitalnih kompetencija i radova svojih razrednih kolega.



Slika 3. Učenički rad/dokaz – strip u Pixtonu

U javnoj galeriji učenici mogu pristupiti i svim javnim sadržajima, kako iz osobnog središta (škole), tako i iz cijele zajednice CRISS projekta, koji su prikazani prema razini privatnosti sadržaja [7].

Svi sadržaji imaju 4 razine privatnosti:

• Privatno: Nitko, osim autora i njegovih nastavnika - pod uvjetom da je autor učenik te škole - ne može vidjeti radove.

• Dijeljeno s razrednom skupinom:

• Dijeljeno s članovima centra: mogu ih vidjeti učenici iste škole

• Javni: Na internetu ih može pronaći bilo tko, ako ima poveznicu.

Pomoću javne galerije učenici su imali priliku naučiti više, proučiti primjere i ideje dobre prakse.

Struktura scenarija (zadaci i aktivnosti) osmišljena je na način koji je učenicima omogućio razumijevanje svakog modula i provođenje aktivnosti.

Učenici su u timskom radu imali uloge: koordinator, tajnik, supervizor, govornik i pomoćnik. Svaki od njih imao je svoje obveze i odgovornosti.

Nastavnik je tijekom projekta pratio upravljanje zajedničkom mapom te vodio bilješke na temelju kojih je, na kraju projekta, izvršio završno vrednovanje.

Svaka grupa je u obliku dnevnika rada morala evaluirati sve što su kao grupa ali i kao pojedinci odradili. Na početku svake sesije grupa provjerava dogovore od prošle sesije. Na kraju svake sesije grupa i svaki učenik provjeravaju što su sve napravili, kako su to napravili, identificiraju dobre i loše strane svog upravljanja projektom i zajednički se dogovaraju oko onih stavki s kojima su imali problema te na koji način ih mogu izbjeći u sljedećoj sesiji.

Isto tako, svaki učenik izradio je vlastitu procjenu svog rada, putem eseja i rubrike.

Učenici su vrednovali svoje kolege uzimajući u obzir završni proizvod koji su napravili te njihovo prezentiranje.

Kontinuirana i dvosmjerna komunikacija te dobra suradnja između nastavnika i učenika rezultirali su jačanju samouvjerenosti učenika u napredak aktivnosti, odnosno vlastiti napredak. Bez poteškoća su istraživali zadatke i aktivnosti koje im je dodijelio nastavnik uz pretraživanje web mjesta, poveznica i softvera za vježbanje te pristupanje podacima o aktivnostima. Omogućen je prijenos datoteka, videozapisa, veza koje se koriste u aktivnostima.

Podaci o učenicima povezani su s nadzornom pločom nastavnika koji može pratiti njihov napredak. Dostupan je napredak u radu za svako područje (slika 4.).



Slika 4. Analitika učenja na platformi CRISS

Svaka potkompetencija ima pridruženu značku koja odražava njezino postignuće. Značka je povezana s posebnim opisnim izvješćima. Primjenjujući pravilo 2/3, konačna ocjena certifikacije uzima u obzir dva najbolja rezultata od tri. Dakle, procjena potkompetentnosti jednaka je srednjoj vrijednosti najbolja dva pokušaja [8]. Dostignuće potkompetencije utvrđuje se kad je srednja vrijednost jednaka ili veća od 70%. Zbirka značaka (slika 5.) uvjet je za konačnu potvrdu.

Svi učenici uključeni u projekt u Tehničkoj školi, Požega stekli su određene potkompetencije. Među njima, najviše stečenih digitalnih potkompetencija istaknuli su se učenici Luka Blažević (2.f) i Emanuel Grbac (1.f).



Slika 5. Analitika učenja za učenika - Moje značke

Kroz zadatke i aktivnosti u scenarijima poučavanja učenici su proučili brojne teme:

- Upravljanje provjerom privatnosti.
- Zaštita od krađe identiteta i prijevara kojima se pokušava pristupiti privatnim podacima na mreži.
- Poznavanje opasnosti pri korištenju interneta, koje mogu utjecati na uređaje i digitalne sustave.
- Korištenje strategija za zaštitu uređaja i digitalnih sustava od vanjskih prijetnji.
- Učinkovito i skladno upravljanje alatima, uređajima i resursima.
- Na zanimljiv i zabavan način upoznali su svoj rodni grad.
- Naučili su kako riješiti svakodnevne probleme u radu računala.
- Stekli su znanje o autorskim pravima.
- Naučili su kako određeni programi pomažu osobama s invaliditetom da koriste računala i komuniciraju putem njih.

S obzirom na važnost CRISS projekta, kao jednog od vodećih programa za znanost i inovacije Europske unije, digitalna platforma nije se testirala samo kroz masovni pilot, već razvoj platforme kontinuirano su pratili vanjski recenzenti osiguravajući kvalitetu primjene i razvoja platforme te implementacije projekta.

IV. CRISS-OV OKVIR ZA VREDNOVANJE DIGITALNIH KOMPETENCIJA

CRISS-ov okvir za vrednovanje digitalnih kompetencija detaljno je razrađen i na jednostavan način pomaže nastavnicima u analizi i procjeni učeničkih kompetencija.

Pet je područja digitalnih kompetencija:

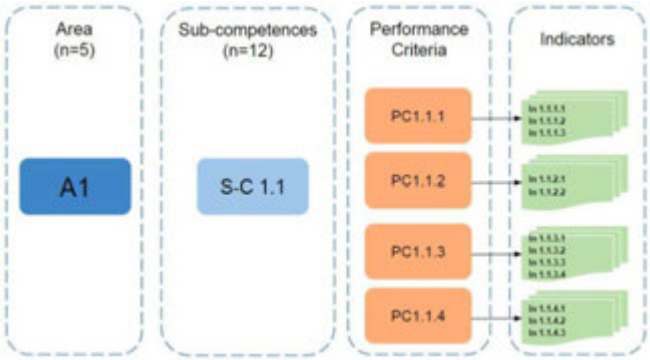
- digitalno građanstvo,
- digitalna komunikacija i suradnja,
- pretraživanje i upravljanje digitalnim informacijama,
- kreiranje digitalnog sadržaja
- digitalno rješavanje problema [8].

Za razliku od okvira za digitalne kompetencije Europske unije na koji se CRISS-ov okvir oslanja, on predstavlja nadogradnju u smislu da je pažljivo razrađen u smjeru razvijanja učeničkih digitalnih kompetencija te njihova usvajanja i vrednovanja od strane nastavnika, ali i u smjeru vršnjačkog vrednovanja i samovrednovanja (slika 6.).



Slika 6. CRISSov okvir za vrednovanje digitalnh kompetencija [9]

Operativni koncept CRISS predlaže pet područja koja grupiraju 12 potkompetencija. Svaka se potkompetencija sastoji od niza kriterija uspješnosti koji kompetencije prevode u konkretnije elemente onoga što bi učenici trebali pokazati. Svaki kriterij uspješnosti procjenjuje se prema skupu pokazatelja (slika 7.). Ovi su pokazatelji uočljive karakteristike kriterija uspješnosti i sastoje se od unaprijed definiranih mjera ili drugih vrsta kvalitativnih podataka s kojima se ispituje učenikov dokaz ili uz koje se vrednuje [9].



Slika 7. Hijerarhijska struktura CRISS sustava procjene kompetencija [9]

CRISS metodološki pristup temelji se na pedagogiji integracije. Umjesto da se samo testira digitalno znanje i vještine učenika, model zasnovan na pedagogiji integracije omogućava nastavnicima da procjenjuju digitalnu kompetenciju učenika ugrađenom u predmetne ili međupredmetne problemske situacije. Pri tome dolazi do izražaja primjena učenja u praktičnim situacijama, učenje učenika kroz povezivanje naučenog sadržaja s vrijednostima poput postajanja odgovornim građaninom, kompetentnim radnikom, neovisnom osobom. Učenici se ocjenjuju u složenom scenariju. Pridruživanje naučenih elemenata pruža učeniku sposobnost da mobilizira svoje znanje i vještine za učinkovito suočavanje sa svakodnevnim situacijama [9].

CRISS strategija procjene digitalne kompetencije predviđa da se vrijednost svakog područja, potkompetencija i kriterij uspješnosti dodjeljuju u skladu s njegovom relevantnošću za školsko obrazovanje.

Kako bi se osigurala valjanost, pouzdanost i pravičnost postupka ocjenjivanja, CRISS predlaže razvoj scenarija procjene sposobnosti (CAS) gdje se digitalna kompetencija razvija i vrednuje u kontekstu u skladu s nizom pravila. U tom smislu CAS-ovi su integrirani u školski kurikulum.

Zadaci ocjenjivanja su dovoljno složeni i pružaju potreban izazov. Procjene sadržavaju otvorene zadatke koji uspostavljaju smislen i/ili autentičan kontekst problema u stvarnom svijetu.

Za nastavnike uključene u projektu, razvijen je MOOC koji omogućuje upoznavanje nastavnika s okvirom digitalnih kompetencija i platformom CRISS. Time je omogućeno bolje razumijevanje alata, te poboljšana interakcija s njim kod izvođenja različitih aktivnosti [8].

V. ZAKLJUČAK

CRISS projekt pruža okvir za digitalnu kompetenciju učenika kao i alat namijenjen razvoju, procjeni i potvrđivanju digitalne kompetencije. CRISS alat pomaže u prikupljanju i prikazivanju svih vrsta dokaza povezanih s digitalnom kompetencijom, te omogućava njegovu procjenu kroz različite scenarije poučavanja.

Razvoj CRISS konstrukcije digitalnog sustava za kompetenciju temelji se na operativnom konceptu i najboljim praksama te preporukama u vezi s istraživanjima i općenitim preporukama za ocjenu opće kompetencije. Predloženo rješenje je dovoljno fleksibilno da integrira razvoj i procjenu digitalne kompetencije u različite discipline i sadržaje, prilagođavajući se različitim nastavnim planovima i programima. Koristi scenarije procjene kompetencije koji pomažu u kontekstualizaciji i učvršćivanju procjene u realnim i smislenim situacijama. Konstrukcija pruža više mogućnosti za obavljanje i provjeru kompetencija, a oslanja se na više metoda i instrumenata za prikupljanje dokaza u prilog postupku procjene i certificiranja.

CRISS platforma pruža sigurno radno okruženje putem kojeg su učenici poboljšali svoje digitalne vještine, stekli nova znanja i iskustva, upoznali različita okruženja učenja, postali kreativniji, koristili softverske alate koje su već poznavali te naučili nove i primili izvještaj o napretku digitalne kompetencije. U timskom radu (4-5 učenika po grupi) razmjenjivali su ideje, surađivali i donosili odluke. Uz suradničke odnose razvili su i kreativni pristup domaćim zadaćama.

Kroz izradu vlastitih priča/dokaza učenici su koristili mobilne telefone i računala u obrazovne svrhe, te razvili suradnju s razrednicima i nastavnicima.

Nastavnici su također razvili svoje digitalne kompetencije kroz upoznavanje s novim programima i alatima, kao i suradnju s kolegama i učenicima, kako u školi tako i na daljinu. Primjena rubrika, smorefleksije i vršnjačkog vrednovanja pomogla je i olakšala postupak vrednovanja prema kurikularnoj reformi, a scenariji poučavanja omogućili su međupredmetnu povezanost nastavnih sadržaja.

Sustav ocjenjivanja primijenjen u projektu CRISS, koji vrednuje i potvrđuje digitalne kompetencije, jedinstveno je i inovativno rješenje koje se može primijeniti i na druge vrste kompetencija.

LITERATURA

[1] OECD, "Definiranje i odabir ključnih kommpetencija", dostupno na https://pisa.ncvvo.hr/wp-content/uploads/2018/05/Definiranje-i-odabir-klju%C4%8Dnih-kompetencija_Final-version.pdf

[2] DigComp 2.1: The digital competence framework for citizens with eight proficiency levels and examples of use, dostupno na https://publications.jrc.ec.europa.eu/repository/bitstream/JRC106281/web-digcomp2.1pdf_(online).pdf

[3] European Commission (2006), "Recommendation 2006/962/EC of the European Parliament and the Council of 18 december 2006 on key competences for lifelong learning", Brussels: European Commission., dostupno na https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:394:0010:0018:en:PDF

[4] CRISS H2020, "Acquisition, assessment and certification of students' digital competence in primary and secondary schools", dostupno na https://www.crissh2020.eu/

[5] CRISS projekt, CARNET, dostupno na https://www.carnet.hr/projekt/criss/

[6] Teacher Eportfolio Structure, "User guide CRISS platform for teacher" , dostupno na http://www.ktonycia.com/CRISS/Eportfolio_Structure_Teacher.pdf

[7] Student Eportfolio Structure, "User guide CRISS platform for student", dostupno na http://www.ktonycia.com/CRISS/Eportfolio_Structure_Student.pdf

[8] CRISS MOOC: "Acquisition, assessment and certification of digital competence", dostupno na https://www.crissh2020.eu/user-guide/

[9] L. Guàrdia, M. Maina, A. Julià, "Digital competence assessment system: Supporting teachers with the CRISS platform", dostupno na https://www.crissh2020.eu/research/, prihvaćen za tisak

Implementacija različitih robotskih modela u nastavi tehničke kulture i informatike

Petar Dobrić *, Maja Mačinko** i Dalia Kager***
* *OŠ Julija Klovića/Nova cesta 133, Zagreb, Hrvatska
** OŠ Ivana Cankara/Ivana Cankara 10, Zagreb, Hrvatska
***OŠ Eugena Kvaternika, Školska 4, Velika Gorica, Hrvatska
*e-mail: dobripero@gmail.com
**e-mail: xjamax@gmail.com
***e-mail: dalia.kager@skole.hr

Današnji način obrazovanja novih generacija od učitelja sve više zahtijeva da budu inovativni i kreativni te da prate nove trendove i tehnologije. Nastava predmeta informatike i tehničke kulture je isprepletena i međusobno povezana. Ovim radom želimo prikazati tri konkretna primjera poučavanja sadržaja kroz različite robotske modele u nastavi. Ti modeli prate aktualne kurikulume, međupredmetne teme te procese učenja koji stavljaju učenika u fokus. Primjerima korištenja Arduino platforme, Fischertechnik setova te Boson setova prikazat ćemo kako je djeci na zabavan način moguće prikazati situacije i primjere iz stvarnog života.

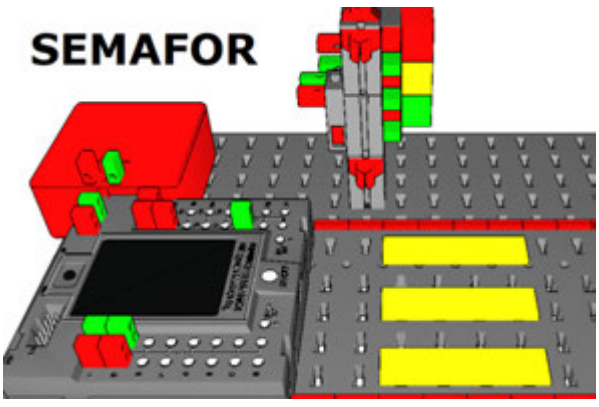
Ključne riječi - Fischertechnik; Arduino; Boson; micro:bit; mikrokontroleri; robotski sustavi

I. UVOD

Kao jedna od netradicionalnih metoda učenja elektronike, elektrotehnike i programiranja je učenje pomoću robotskih sustava. U posljednjih nekoliko godina vidljivo je ubrzano širenje uređaja koji sakupljaju i upravljaju informacijama u realnom vremenu putem nekog sučelja i komponente među njima, kao što su sonde i roveri u Svemiru, ABS sustavi, umjetnih zamjenski udovi i mnogi drugi. Kako ti uređaju postaju sveprisutni, digitalno kompetentna osoba trebala bi biti upoznata s načinima kako takvi uređaji rade, kao i mogućnostima i ograničenjima koja oni nude. Korištenjem nekih od ovih modela kod djece razvija se logičko razmišljanje i zaključivanje popraćeno primjerima iz stvarnoga života što je najbolji način učenja kroz igru.
Danas je učenje popriličan izazov ako se vodimo činjenicom da djeca često imaju problema pri usvajanju osnovnih koncepata i rješavanju problemskih zadataka.
Objektno orijentirani princip učenja je odličan način usvajanja znanja i logičkog razmišljanja tijekom rješavanja problemskih zadataka, te kao takav olakšava proces učenja kod učenika jer odmah vide rezultat svog programa. Izvođenje zadanog problemskog zadatka odvija se pomoću određenog sučelja, koje je povezano s računalom i ulazno-izlaznim elementima (žaruljice, motori, različiti senzori...). Time se kod učenika poboljšavaju njihove psiho motoričke vještine i pobuđuje se interes za istraživanjem.

A. Fischertechnik konstrukcijski setovi i robotski modeli u tehničkoj kulturi, informatici i robotici

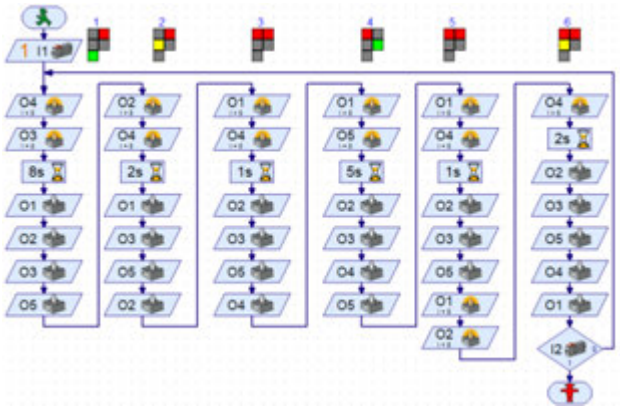
Suvremeni koncepti učenja i programski kurikulumi u nastavi tehničke kulture i informatike omogućavaju aktivan i zabavan pristup usvajanja znanstvenih koncepata pomoću automatiziranih robotskih modela. Učenje kroz igru je izvrstan oblik usvajanja očekivanih ishoda, gdje je učenik aktivan sudionik i kreator robotskih modela. Razumijevanje tehničkih pojmova i principa funkcioniranja tehničkih tvorevina pospješuje razvoj inženjerskog načina razmišljanja i rješavanje različitih problemskih zadataka. Multidisciplinarni pristup omogućuje usvajanje različitih znanja i praktičnih vještina koje su nužne za uspješno obavljanje zanimanja koja danas ne postoje. Najveći izazov za učitelje je kako osigurati učenicima napredovanje, osobni rast i razvoj. Fischertechnik robotski modeli omogućuju usvajanje znanja iz više područja tehnike: strojarskih konstrukcija, elektrotehnike, elektronike, automatike, matematike i informacijskih i komunikacijskih tehnologija (IKT).



Slika1. Semafor - model

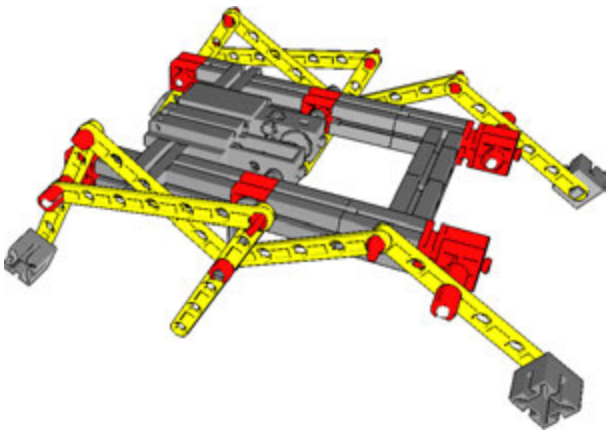
Fischertechnik slagalice omogućuju izradu raznovrsnih automatiziranih robotskih konstrukcija uz mnoštvo kombinacija. Povezivanje gradivnih elemenata u funkcionalni model robota uz postavljanje ulaznih i izlaznih elemenata olakšava ožičenje istih.

Elektronički sklop – sučelje upravlja robotom povezujući računalo i robotski model uz algoritam napravljen u RoboPro programu. Napisani program prenosimo putem međusklopa - sučelja koje je spojeno vodičima s robotskim modelom.



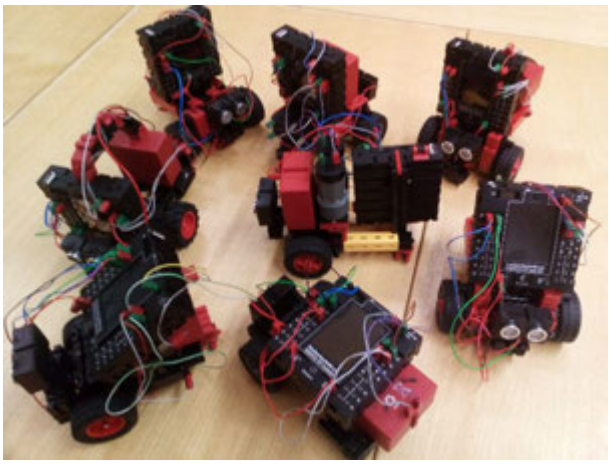
Slika 2. RoboPro semafor - algoritam

Algoritamske naredbe sučelje prevodi u strujne impulse koji pokreću električne elemente robotskog modela. Sučelje na računalo spajamo pomoću USB kabela ili Bluetootha koji prenosi programe s računala u memoriju robota.



Slika 3. Robo hodač - model

RoboPro je programski jezik s grafičkim sučeljem koji je intuitivan za uporabu i jednostavan za upravljanje robotskim modelima. Usvajanje i savladavanje logike programiranja kroz kontrolne programe osigurano je uporabom dijagrama tijeka. Grafičko korisničko sučelje omogućava odabir naredaba i procedura pomoću miša. Izrada programskog algoritma odvija se povezivanjem grafičkih ikona u cjelinu koja predstavlja programsko rješenje prikazano dijagramom tijeka. Dijagram tijeka je građen od programskih elemenata koji se umeću i povezuju unutar prozora programa pomoću miša (tehnika povuci – pusti). Grafički prikaz tijeka programa olakšava praćenje tijeka programa i učenje logičkog pristupa programiranju. Usvajanje osnovnih vještina rada s programom je intuitivno i idealno za učenje računalnog razmišljanja i programskog koda.



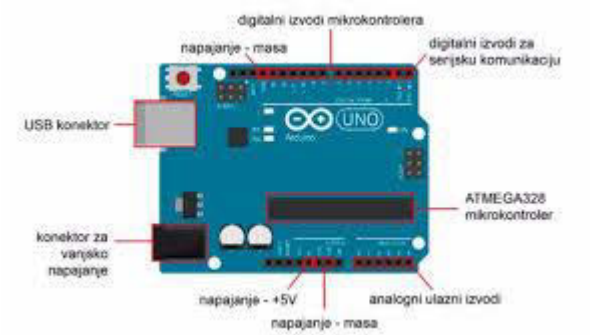
Slika 4. Robotska vozila - modeli

Primjena robotike u izvannastavnim kurikulumima uvodi učenike u svijet sadašnjosti i priprema ih za izazove u budućnosti. Kvaliteta nastavnog procesa osigurana je kontinuiranim provođenjem kurikuluma u višim razredima osnovne škole kroz nastavu tehničke kulture i informatike. Izrada različitih konstrukcija robota od petog do osmog razreda u nastavi tehničke kulture i njihovo pokretanje spajanjem električnih elemenata u strujnim krugovima, priprema je za izradu algoritamskih rješenja na nastavi informatike.

B. Arduino mikrokontroler kroz Internet stvari i sučelje za izradu aplikacija

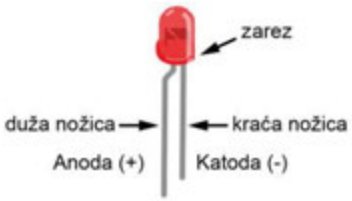
Igra s mikrokontrolerima odličan je način da učenici dosegnu nova znanja koja su korak u budućnost. Mikrokontroler je zapravo maleno računalo smješteno na integrirani sklop. Da bi mikrokontroler znao kako raditi, potrebno je napisati program te ga na njega učitati. Važno je napomenuti da je Arduino platforma, kao i sve Arduino pločice, otvoreni sustav i svi se programi i promjeri mogu preuzeti sa službene stranice www.arduino.cc. Time se učiteljima olakšava rad jer nema potrebe za plaćanjem programa za koje u školama često nema mogućnosti.
Arduino je platforma nastala prije desetak godina u Italiji, a kako su ga razvijali, prilagodili su ga novim izazovima čime je postao jedna od popularnijih platformi za izradu aplikacija. I, vrlo bitno, u uređenju prostora i okoliša. Svojim jednostavnim korištenjem tijekom godina poslužio je kao oruđe u izradi mnogih projekata, a danas je projektna nastava srž samog kurikuluma nastave tehničke kulture i informatike.
Često se u kontekstu spominje i Internet Stvari (eng. *Internet of Things*). To je pojam koji označava povezivanje malih sustava ugrađenih u stvari iz okoline u mrežu preko koje će se razmjenjivati podatci. Arduino, kao savršen primjer jednostavne upravljačke komponente, može poslužiti kao „jezgra“ koja upravlja senzorima, motorima i sl. u Internetu Stvari.

Na Arduino pločici smješteni su izvodi mikrokontrolera koji služe za njegovo povezivanje s elementima koje spajamo na njega. Dijele se na digitalne i analogne.

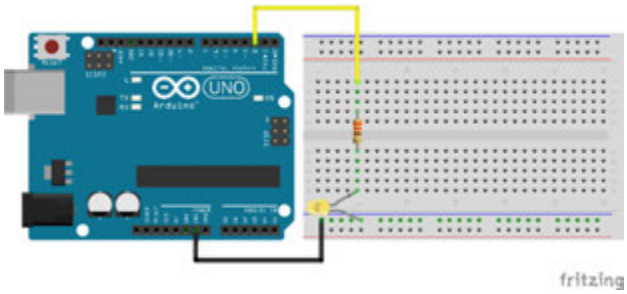


Slika 5. Arduino Uno dijelovi

Digitalni izvodi mikrokontrolera imaju dva moguća stanja: nisko logičko stanje (kada je napon nula), te visoko logičko stanje (kada je napon +5V). Analogni izvodi očitavaju napon u vrijednostima od 0V do +5V. Napajanje (+5V) često se koristi za spajanje senzora, aktuatora i daje malu struju. Napajanje GND negativni je pol istosmjernog napona, a napajanje +3V daje vrlo malenu struju. Pri spajanju bilo koje elektroničke komponente moramo paziti da na svakom izvodu (pinu) mikrokontrolera maksimalna struja iznosi 40mA, a napon +5V.
Kako svjetleća dioda koja je najčešće korišten elektronički element u vježbama radi otprilike na +2V, a napajanje mikrokontrolera je +5V, potrebno je ispred nje serijski spojiti odgovarajući predotpornik kako bismo ju zaštitili. Svjetleće ili LED diode su elektroničke komponente su diode koje emitiraju svjetlost.

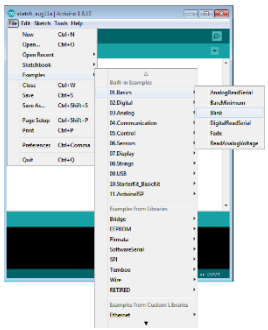


Slika 6. Svjetleća dioda



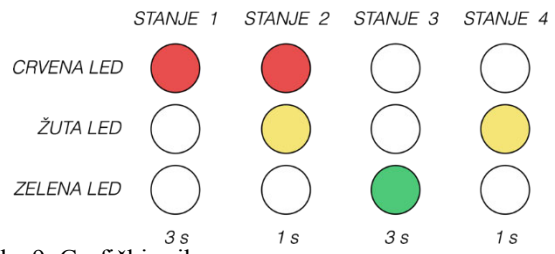
Slika 7. Svjetleća dioda u strujnom krugu sa predotpornikom

Arduino programski jezik ima sintaksu i jezična pravila identična C++-u, no ima ugrađene i standardne Arduino biblioteke (primjere) koje nije potrebno dodavati u program. One pružaju temeljne funkcionalnosti za rad s periferijom, pristup serijskom sučelju i konstante vezane za logičke razine, analogne priključke i slično. Najjednostavniji je program *Blink* koji otvaramo klikom na izbornik *File* i odaberemo *Examples* pri čemu se pojavljuje kategorija dostupnih primjera.



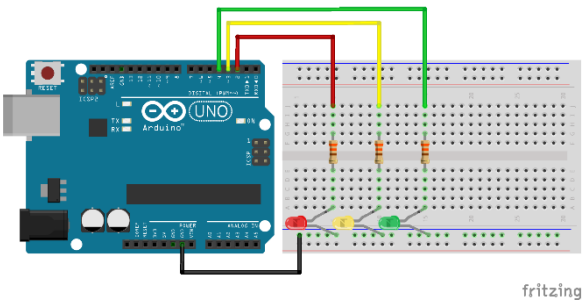
Slika 8. Korištenje gotovih primjera u biblioteci

Jedan od najčešće upotrebljavanih primjera pri korištenju Arduino mikrokontrolera jest izrada semafora za automobile, te kasnije naprednije verzija modela semafora za automobile i pješake. Primjeri iz svakodnevnog života su osnova nastavnog kurikuluma i savršeno pokazuju konkretne primjere kako nešto doista radi. Ponajviše kada se pojavi „Heureka efekt“ kod djece kada im sve proradi i osjete da su uspjeli.
Na Arduino pločicu spojimo tri svjetleće diode – crvenu, žutu i zelenu. Na početku na semaforu svijetli samo crveno svjetlo tri sekunde. Nakon toga svijetle crveno i žuto svjetlo zajedno jednu sekundu. Potom svijetli samo zeleno svjetlo tri sekunde i na kraju samo žuto svjetlo jednu sekundu. Nakon toga proces kreće ispočetka.



Slika 9. Grafički prikaz programa

Prikaz spajanja elemenata na eksperimentalnoj pločici i povezivanja s Arduino pločicom vidljiv je na sljedećoj slici.



Slika 10. Spajanje semafora na eksperimentalnu pločicu

Kod izvršavanja programskog kôda mikrokontroler izvodi jednu po jednu naredbu. Kada bi usporili vrijeme, vidjeli bismo da bi se kod prijelaza semafora iz stanja u stanje diode uključivale jedna po jedna, a ne više njih odjednom kako nam se čini kada gledamo naš model. Srećom,

mikrokontroler svoj zadatak obavlja vrlo brzo, pa se našem oku čini kao da se crveno i žuto svjetlo isključilo u isto vrijeme kad se uključilo zeleno svjetlo.

```
void setup() {
  pinMode(11, OUTPUT);
  pinMode(12, OUTPUT);
  pinMode(13, OUTPUT);
}

void loop() {
  digitalWrite(13, 1); //crveno
  digitalWrite(12, 0);
  digitalWrite(11, 0);
  delay(3000);          // 3 sekunde
  digitalWrite(12, 1); //crveno i žuto
  delay(1000);          // 1 sekund
  digitalWrite(13, 0);
  digitalWrite(12, 0);
  digitalWrite(11, 1); //zeleno
  delay(3000);
  digitalWrite(12, 1); //žuto
  digitalWrite(11, 0);
  delay(1000);
}
```

Slika 11. Programski kod semafora

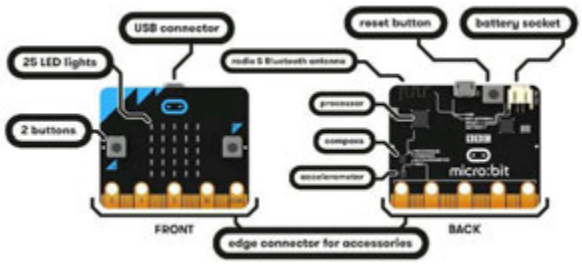
Na Arduino mikrokontroler mogu se spajati i mnogi dodatni senzori i uređaji pomoću ploča za proširenje funkcionalnosti koje nazivamo štitovi (eng. *Shield*). Štitovi se spajaju direktno na Arduino i najčešće imaju dodatne ulazne ili izlazne priključke kojima se sklopovski dodaju nove funkcionalnosti. S donje strane štita nalaze se nožice postavljene u jednak raspored kao i priključci na Arduino pločici, a s gornje strane štita su priključci, također u istom rasporedu. Štit se spaja na Arduino pločicu umetanjem svih njegovih nožica u odgovarajući priključak. Moguće je čak i spajanje više štitova jednog na drugi. Važno je paziti da štitovi ne koriste iste priključke niti protokole.



Slika 12. Arduino spojen sa štitom

C. *BBC micro:bit i BOSON Starter Kit u redovnoj i izbornoj nastavi informatike te izvannastavnim aktivnostima*

Učenje programiranja u nastavi informatike učenicima je često teško i demotivirajuće. Sukladno s novim kurikulumom nastavnog predmeta informatika koji prednost daje ishodima učenja, a ne sadržaju, učenike možemo aktivirati i motivirati za učenje osnovnih koncepata programiranja, za razvoj vještina rješavanja problema i algoritamskog mišljenja upravo uporabom dostupnih robotskih modela i mikroračunala. S obzirom da su micro:bit računala od 2017. godine dostupna u velikom broju škola u Republici Hrvatskoj kroz donacije u projektima *STEM revolucija* i *ProMikro* Instituta za razvoj i inovativnost mladih, a uzevši u obzir činjenicu mogućnosti povezivanja s raznim senzorima i drugim modulima te da ga je moguće programirati s pomoću raznih popularnih programskih jezika (*Python, JavaScript, Scratch*) na različitim digitalnim uređajima i operacijskim sustavima, micro:bit se pokazao kao idealan posrednik za ostvarivanje ishoda učenja kurikulumske domene *Računalno razmišljanje i programiranje* u svim razredima osnovne škole kao i za stjecanje digitalnih kompetencija i digitalnih vještina 21. stoljeća.



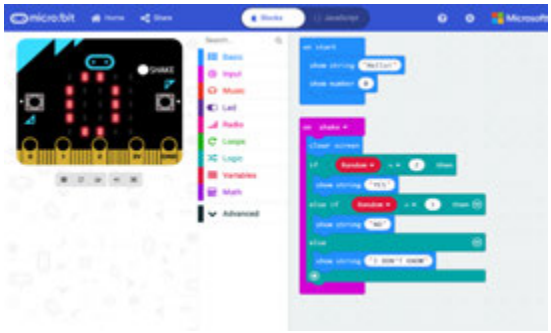
Slika 13. Fizičke značajke micro:bit računala (Izvor: https://microbit.org/images/microbit-hardware-access.jpg)

Micro:bit ima brojne značajke – 25 LED dioda koje se mogu zasebno programirati, dvije programibilne tipke, senzore za svjetlo i temperaturu, senzore pokreta (mjerač ubrzanja i kompas), bežičnu radio i Bluetooth komunikaciju te izvode (pinove) koji omogućuju povezivanje raznih električnih komponenti i senzora s micro:bitom te njihovo programiranje.

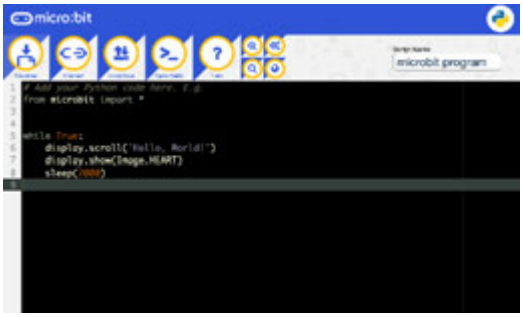
Micro:bit je moguće vizualno programirati *online* i *offline* u Makeblock editoru pomoću blokova naredbi, što ga čini prikladnim za početno učenje programiranja, a u samom editoru postoji i opcija koja blokove naredbi prikazuje u JavaScript programskom jeziku što pak omogućuje programiranje naprednim korisnicima. Vizualni programski jezik Scratch ima dodatak s naredbama za programiranje micro:bita pa ga na taj način mogu programirati i mlađi učenici osnovne škole. Micro:bit se može programirati i u programskom jeziku Python s pomoću Python editora. Programiranje micro:bita podržano je na različitim operacijskim sustavima (Windows, Linux, MAC OS), kao i na pametnim uređajima putem dostupne aplikacije (za Android i iOS). Sve su to značajke koje ovo mikroračunalo, uz dostupnost u školama i malu nabavnu cijenu, čine idealnim za implementaciju u nastavi informatike.



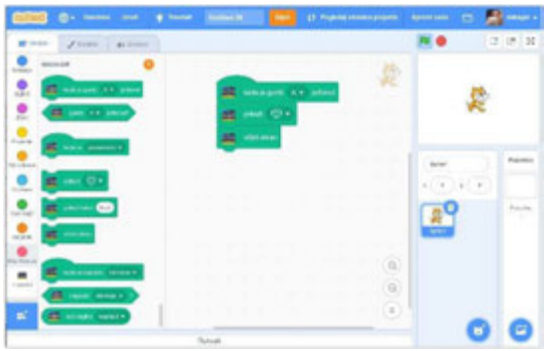
Slika 14. Micro:bit MakeCode editor – JavaScript jezik



Slika 15. Micro:bit MakeCode editor za vizualno programiranje



Slika 16. Micro:bit Python editor



Slika 17. Micro:bit dodatak za Scratch

DFRobot Boson Starter Kit za micro:bit sadrži 8 modula (ventilator, LED žaruljica, senzor pokreta, senzor zvuka, potenciometar, RGB LED traka, servo motor i tipkalo), koji pokrivaju najpopularnije digitalne i analogne senzore i motore, podržavaju interakciju zvuka, svjetla i pokreta. Povezivanje micro:bita s modulima olakšano je s pomoću pločice za spajanje elektroničkih modula, a moduli komuniciraju s micro:bitom preko 3-pin sučelja.



Slika 18. Dijelovi BOSON Starter Kita

BOSON Kit osmišljen je za praktično STEM obrazovanje, funkcionalan je i bez kodiranja, a moguće ga je kombinirati s Lego setom, ili pomoću magneta, vijaka i ljepljive vrpce graditi razne konstrukcije, pa je prikladan i u STEM i STEAM obrazovanju (uključuje i umjetnost – eng. *Arts*). Osim s micro:bit računalom, kompatibilan je i s Arduino tehnologijom. Dostupan je u četiri različita seta: *Starter Kit, Science Kit, Coding Kit* i *Inventor Kit.*

Učenici s pomoću micro:bita i BOSON Starter Kita uče osnovne, ali i napredne koncepte programiranja kroz primjere iz stvarnog života koji su im poznati, surađuju u timskom radu te suradnički uče, upoznaju projektni način rada, razvijaju komunikacijske i prezentacijske vještine te poduzetništvo [1] . Osim ostvarivanja ishoda kurikuluma međupredmetnih tema (*Poduzetništvo, Osobni i socijalni razvoj, Uporaba IKT-a, Učiti kako učiti*), informatika se povezuje i s ishodima učenja drugih predmeta (Matematika, Fizika, Tehnička kultura, Likovna kultura, Hrvatski i Engleski jezik itd.), što ovaj model čini interdisciplinarnim. Implementacija ovog modela obuhvaća suvremene metode poučavanja koje stavljaju učenika u fokus i potiču njegovu aktivnost – projektni rad, timski rad, učenje istraživanjem, problemsko učenje kao i *tinkering* pristup učenju koji sve više dobiva na popularnosti u neformalnim postavkama učenja, a gradi se na pedagogiji temeljenoj na istraživanjima i koristi neke od najzanimljivijih i motivirajućih elemenata učenja, usredotočenog na imerzivno, osobno učenje. *Tinkering* se također sve više promatra kao snažno sredstvo za razvijanje vještina dvadesetog stoljeća poput kritičkog razmišljanja, kreativnosti, suradnje, rješavanja problema, komunikacije, odgovornosti, samopouzdanja, digitalne pismenosti i poduzetništva, što odgovara izazovima koji su postavljeni današnjim generacijama učenika, a vezani su uz njihov budući profesionalni razvoj[2].

[1] R. Burbaite, V. Stuikys and R. Damasevicius, "Educational robots as collaborative learning objects for teaching Computer Science," 2013 International Conference on System Science and Engineering (ICSSE), Budapest, 2013, pp. 211-216.
doi: 10.1109/ICSSE.2013.6614661

[2] Harris, Emily & Winterbottom, Mark & Xanthoudaki, Maria & Calcagnini, Sara & Pijer, Inka. (2016). A PRACTITIONER GUIDE FOR DEVELOPING AND IMPLEMENTING TINKERING ACTIVITIES.



Slika 19. Implementacija BOSON Starter Kita i micro:bita u nastavi informatike



Slika 20. Implementacija BOSON Starter Kita i micro:bita u nastavi informatike – timski projektni rad



Slika 21. Implementacija BOSON Starter Kita i micro:bita u nastavi informatike – učenička prezentacija projektnog rada

II. ZAKLJUČAK

Učenjem kroz praktične primjere iz svakodnevnog života postižemo povećani interes kod djece. Osim što dobiju povratnu informaciju o realizaciji procesa na licu mjesta, uče koristiti primjere iz svakodnevice koja ih okružuje te dobivaju povratnu informaciju i ideje za poboljšanje rada. Robotika je svuda oko nas i učenje kroz robotske modele kojih je danas na tržištu mnogo može biti odličan način realizacije novog kurikuluma i postizanja ishoda učenja.

Suvremen pristup učenju putem projektne nastave

Nataša Boj *, Aleksandra Tonković **, Edgar Glavaš***
*III. osnovna škola Čakovec, Čakovec, Republika Hrvatska
**III. osnovna škola Čakovec, Čakovec, Republika Hrvatska
***Županijska bolnica Čakovec, Republika Hrvatska

natasa.boj@skole.hr
alex.tonkovic@gmail.com
eglavas@myself.com

Projektna nastava omogućuje učeniku da postane aktivan sudionik u odgojno–obrazovnom procesu. Takvu ulogu učenika omogućava, između ostalog, i primjena informacijsko-komunikacijske tehnologije. Projektna nastava uključuje iskustvenu, istraživačku i interdisciplinarnu nastavu usmjerenu na učenika. Učenika se stavlja u prvi plan prema njegovim interesima i sposobnostima. Projektna nastava kao iskustvena nastava, omogućuje učenicima usvajanje trajnih znanja.

Učitelj dobiva novu ulogu savjetnika, medijatora, koordinatora i mentora. Njeguje se timski rad i partnerski odnos učenika i učitelja kao i ostalih sudionika nastavnog procesa.

Cilj ovog rada bio je istaknuti važnost projektne nastave u suvremenom konceptu obrazovanja. Ispitani su stavovi učenika o projektnoj nastavi; o zanimljivosti projektne nastave u odnosu na tradicionalnu, o zastupljenost projektne nastave u redovnoj i izbornoj nastavi, o motiviranosti učenika za projektnu nastavu.

U tu svrhu je provedeno istraživanje u jednoj gradskoj školi u Čakovcu. Istraživanje je provedeno 2019. godine. Uzorak je obuhvatio 90 % učenika sedmog i osmog razreda. Rezultati istraživanja su pokazali da je projektna nastava učenicima zanimljivija u odnosu na tradicionalnu. Nadalje, da je projektna nastava više zastupljena u izbornoj nastavi, da postoji razlika u motiviranosti za projektnu nastavu s obzirom na spol, ali i s obzirom na postignuti uspjeh učenika u prethodnoj školskoj godini.

Ključne riječi **– projektna nastava, tradicionalna nastava, informacijsko-komunikacijska tehnologija, stavovi učenika, motivacija, uspjeh**

I. UVOD

Globalizacijski procesi i informacijska revolucija kao imperativ društva postavili su nove izazove u odgojno-obrazovnoj praksi. Ubrzani život i stalne promjene u društvu zahtijevaju od pojedinca cjeloživotno učenje. Pojedinac svojim znanjem i vještinama doprinosi napretku suvremenog društva. Suvremeni svijet donosi niz informacija, pristup medijima i suvremenim tehnologijama. Informacije su dostupne s različitih izvora, stoga je temeljna uloga odgoja i obrazovanja u suvremenom društvu omogućiti učenicima potpuni rast i razvoj u cjelovitu osobu. Naglasak suvremene nastave je na inovativnosti, kreativnosti, snalaženju u novim situacijama, samostalnosti, istraživanju, aktivnom učenju s ciljem samoostvarenja. Vrlo je važno tehničko obrazovanje učenika za razvoj sposobnosti potrebnih za rješavanje problema. Projektna nastava učenicima pruža mogućnost međusobne suradnje. Učenici su vješti u primjeni informacijsko-komunikacijske tehnologije, određena znanja i vještine su stekli kroz formalno obrazovanje na nastavi informatike. Upravo zato, oni na jednostavan način komuniciraju međusobno kao i sa učiteljem te kvalitetno odrađuju projektne zadatke, bilo samostalno ili radom u timu. Na taj način stječu znanja i vještine koje će im biti potrebne na tržištu rada.

Didaktičko-metodička paradigma suvremene nastave u tom kontekstu polazi od razmatranja i ostvarenja odgojno-obrazovne prakse putem projektne nastave uzimajući u obzir interdisciplinarni pristup, uvažavajući kurikulum, strategije i metode rada, individualne potrebe učenika, poticajno okruženje. Pojmovno određenje projekta je višeznačno. Pojmovi koji se vežu uz naziv projekt u literaturi su: vremenski ograničena aktivnost, vremenski skup aktivnosti, privremeni pothvat. Meyer definira projekt kao pokušaj nastavnika i učenika da se društveno značajan problem poveže s interesima sudionika, obradi i dovede do rezultata. Projekt je moguće ostvariti u odgojno-obrazovnoj praksi putem projektne nastave.

Uz projektnu nastavu vežu se pojmovi u literaturi: podrška nastavnika, suradničko učenje, izvanučionička nastava, interdisciplinarnost, usmjerenost na učenika, suvremena nastava, istraživačka nastava, moderni kurikulum, najsloženiji oblik rada, istraživačka nastava. Krsnik navodi oblike interaktivne nastave: grupni rad, otkrivačka nastava, problemska nastava, problemski usmjerena nastava, istraživačka nastava i konstruktivistička nastava. Svi ti oblici koriste se u projektnoj nastavi.

Projektna nastava je model organiziran oko projekta koja se temelji na problemskoj nastavi. Učenici uče i istražuju uz pomoć interaktivnog učenja u okviru pojmovnog konstrukta nastavne strategije s ciljem da dođu do spoznaja putem istraživanja. Nastava je aktivno usmjerena na učenika s ciljem usvajanja znanja i vještine potrebne za cjeloživotno učenje. Projektna nastava je nastava u kojoj učenici rade na istraživačkim ili radnim projektima. Istraživačka i projektna nastavna metoda omogućuje učenicima aktivno iskustveno učenje. Projektna nastava uključuje problem koja zahtijeva od učenika da u određenom vremenskom razdoblju prema određenim etapama riješi taj problem i prezentira rješenje problema. Završna etapa pruža mogućnost vrjednovanja postignuća rada u svrhu unaprjeđivanja procesa učenja te bolje motivacije učenika i ostalih sudionika za istraživački i stvaralački rad.

Sami počeci projektne nastave sežu u doba stare Grčke. Putem metode razgovora, učitelji su nastojali aktivno uključiti učenike u proces učenja postavljajući problem i tražeći način da ga učenici i riješe međusobnim dijalogom. Platon je postavio temelje kritičkog mišljenja. Aristotel se zalagao za međusobnu interakciju odgojitelja i odgojenika, a cilj odgoja je bio harmonijska razvijena osoba. U srednjem vijeku, J. A. Komensky je zagovarao zornost i aktivnosti učenika. J. J. Rousseau je širio ideju kako treba voditi računa o interesu učenika, poticati njegovu samostalnost, zalaganje i opažanje. Zalagao se za individualnost i (samo)kritičnost učenika. J. H. Pestalozzi je isticao povezivanje proizvodnog rada i nastave. Odgojne ciljeve je podijelio na odgoj glave (uma), odgoj srca (morala) i odgoj ruke (radni odgoj). J.F.Herbert ističe važnost interesa učenika, pomaganje drugima. Uvodi početak i završetak nastave. J. Dewey je začetnik učenja radom. Nastava pruža učenicima fleksibilnost, slobodu u rasporedu sati. Učenici uče prema vlastitim interesima i sposobnostima. Pažnja je usmjerena na učenika. Uvodi osnovne metode učenja koje su danas iznimno važne za projektnu nastavu. To su aktivnost, istraživanje, eksperiment, promatranje i zaključak. Učitelj ima ulogu koordinatora, a učenici sami dolaze do rješenja problema, što je ujedno priprema i za njihov daljnji život. Krajem 19. st. i početkom 20.st. u škole se uvodi ručni rad, što omogućuje povezanost intelektualnog i fizičkog.

Današnja projektna nastava uključuje aktivnost učenika, potiče kritičko razmišljanje, pruža mogućnost stvaralačkog istraživanja, inovativnost i proširivanje znanja uvažavajući suvremene trendove i dostignuća informacijsko-komunikacijskih tehnologija. Prethodne ideje početaka projektne nastave su poticaj za daljnje unapređenje i uspješno provođenje projektne nastave putem ohrabrivanja sudionika, permanentnim usavršavanjem učitelja, poticanjem na suradnički rad. Postoji potreba za osuvremenjivanjem nastavnog procesa i korištenjem različitih metoda rada.

Očekuje se da će rezultati ovog rada doprinijeti afirmaciji projektne nastave u području suvremenog odgoja i obrazovanja u osnovnoj školi danas.

II. REZULTATI RANIJIH ISTRAŽIVANJA

Postoje od ranije rezultati istraživanja vezani za tematiku projektne nastave u svijetu i u Republici Hrvatskoj.

U Japanu je provedeno istraživanje o motiviranosti učitelja i učenika za projektnu nastavu. U istraživanju je sudjelovalo 126 učitelja i 631 učenik srednjih škola u Hong Kongu. Rezultati su pokazali visoku učiteljevu intrinzičnu motivaciju. Veća motivacija učitelja povezana je s većom motivacijom učenika. Istraživanje su proveli Shui-fong Lam i dr .

Provedeno je istraživanje u Turskoj. Isik i Gucum su ispitivali motivaciju učenika putem projektnog pristupa učenju u znanosti i tehnologiji. Uzorak je obuhvatio 75 učenika sedmih razreda osnovne škole. Učenici su bili podijeljeni u dvije skupine, kontrolnu i eksperimentalnu. U eksperimentnoj skupini se primjenjivao suvremen pristup učenju putem projektne nastave, dok je kontrolna skupina obuhvaćala tradicionalan pristup učenju. Primijenjen je motivacijski upitnik. Rezultati su pokazali statističku značajnost u organizaciji, korištenju kognitivne strategije, unutarnjoj motivaciji.

U Republici Hrvatskoj je provedeno istraživanje utjecaja projektne nastave na promjenu stava kod učenika prema nastavnim sadržajima iz biologije. Istraživanje je provedeno u Velikoj Gorici u gimnaziji. Uzorak je obuhvatio 108 učenika. Učenici su bili podijeljeni u eksperimentalnu i kontrolnu skupinu. Kod eksperimentalne skupine se stav prema botanici promijenio u pozitivnom smjeru nakon izvođenja projektne nastave, dok je kod kontrolne skupine ostao isti. Učenici koji imaju pozitivniji stav prema botanici, pokazuju i bolje znanje u tom području. Bolji učenici prema školskom uspjehu ne pokazuju pozitivniji stav prema projektnoj nastavi. Nadalje, provedeno je istraživanje koje je ispitivalo učestalost primjene metoda i postupaka poučavanja u nastavi prirode i društva. Utvrđeno je da se rijetko

primjenjuje informacijsko-komunikacijska tehnologija, projektna metoda i rad sa stručnim tekstovima u nastavi. Nedovoljno su zastupljene metode/postupci koji potiču istraživačko i iskustveno učenje.

S obzirom na nedovoljna istraživanja o projektnoj nastavi i njenoj zastupljenosti u osnovnoj školi u Republici Hrvatskoj, u tu svrhu provedeno je istraživanje u Međimurskoj županiji.

III. ISTRAŽIVANJE U OSNOVNOJ ŠKOLI MEĐIMURSKE ŽUPANIJE

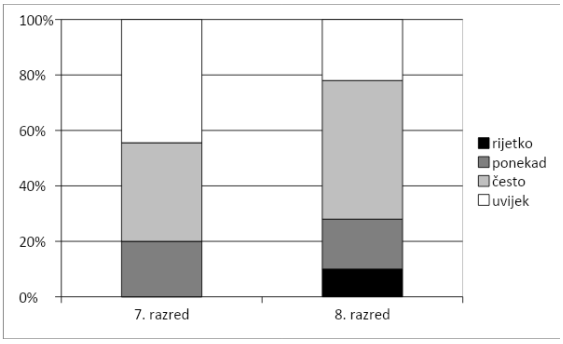
Istraživanje je provedeno kako bi se utvrdila motiviranost učenika za projektnu nastavu, zastupljenost projektne nastave u ukupnoj nastavi s naglaskom na razliku između redovne (obrazovnih i odgojnih predmeta) i izborne nastave. Uzorak je obuhvatio sve dostupne učenike sedmog i osmog razreda, ukupno 95 učenika.

Istraživanje je provedeno u drugom polugodištu školske godine 2018./2019. u osnovnoj školi na području Međimurske županije. Anketiranje učenika provedeno je za vrijeme satova informatike u osnovnoj školi, a termini provođenja istraživanja unaprijed su dogovoreni. Na početku anketiranja, učenicima su dane upute o postupku i trajanju anketiranja. Učenici su ispunjavali upitnik u prosjeku 15 minuta. Učenici su upute dobili i pisanim putem. Objašnjen im je detaljno cilj istraživanja, anonimnost upitnika i svrha istraživanja. Učenici su savjesno popunili upitnik. Nakon anketiranja i prikupljanja podataka, izvršena je obrada i interpretacija istih. Ispitanici su izražavali stavove o zastupljenosti projektne nastave prema vrsti nastavnog predmeta (redovni ili izborni). Ispitana je motiviranost učenika za projektnu nastavu prema spolu, i prema zaključnoj/završnoj ocjeni na kraju prošle godine. Provedeno je ispitivanje stavova učenika sedmih i osmih razreda o zanimljivosti projektne nastave u odnosu na tradicionalnu. Analizirani su i prikupljeni podaci statistički obrađeni kvalitativno i kvantitativno, hi kvadrat testom, za pojedine uzorke Kruskal-Walisovim testom, te prikazani grafički histogramima.

U svrhu istraživanja, konstruiran je anketni upitnik i skala procjene za učenike. Prvi dio anketnog upitnika odnosio se na sociodemografska obilježja učenika. U anketnom upitniku sociodemografska obilježja uključivala su spol, dob, ocjene iz pojedinih predmeta. Upitnik i skala procjene su konstruirani u svrhu ovog istraživanja. Obuhvaćali su tvrdnje: veseli me učenje putem projektne nastave i volim sudjelovati u njoj, više naučim putem projektne nastave, projektna nastava je zanimljivija od tradicionalne nastave. Na tvrdnje su učenici odgovarali s nikad, rijetko, često ili uvijek.

IV. REZULTATI ISTRAŽIVANJA

Slijede grafički prikaz rezultata anketnog upitnika i skale procjene koje su ispunjavali učenici predmetne nastave 7. i 8. razreda.



Slika 1.Projektna nastava je zanimljivija od tradicionalne prema razredu. (p=0,025)

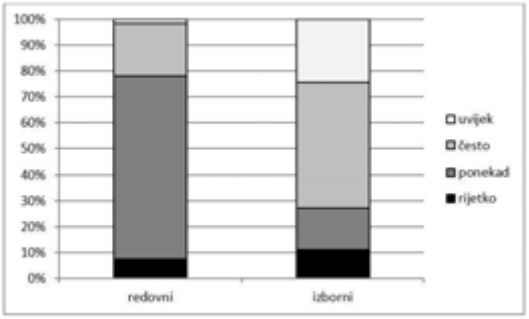
Uspoređeni su stavovi učenika s obzirom na razred prema tradicionalnoj i projektnoj nastavi (slika 1). Uzorak je obuhvatio učenike sedmih i osmih razreda.

Rezultati učenika sedmih razreda: 20% učenika smatra da je projektna nastava ponekad zanimljivija od tradicionalne, 35% njih smatra da je često zanimljivija dok 45% učenika smatra da je projektna nastava uvijek zanimljivija od tradicionalne.

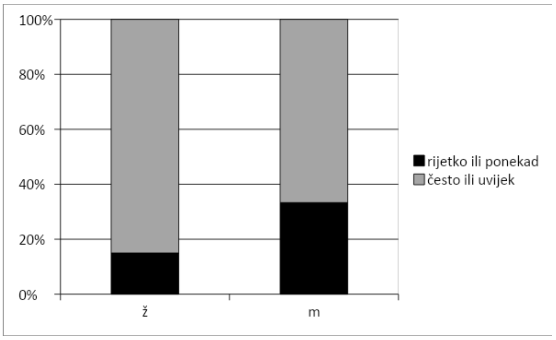
Rezultati učenika osmih razreda: 28% učenika smatra da je projektna nastava rijetko ili ponekad zanimljivija od tradicionalne, 50% njih smatra da je često zanimljivija dok 22% učenika smatra da je projektna nastava uvijek zanimljivija od tradicionalne.

Ukupno uspoređujući rezultate učenika sedmih i osmih razreda, nema velike razlike što se tiče njihovih stavova o zanimljivosti projektne nastave. Projektna nastava.je ipak nešto zanimljivija učenicima sedmih razreda.

Prikazana je procjena učenika o zastupljenosti projektne nastave prema tipu predmeta (slika 2); redovni i izborni. U izborne predmete spadaju Informatika, Njemački jezik, Vjeronauk i Građanski odgoj. Svi ostali predmeti su redovni i kao takvi obvezni.



Slika 2. Zastupljenost projektne nastave prema tipu predmeta



Slika 3. Motiviranost za projektnu nastavu prema spolu ispitanika (p=0,044)

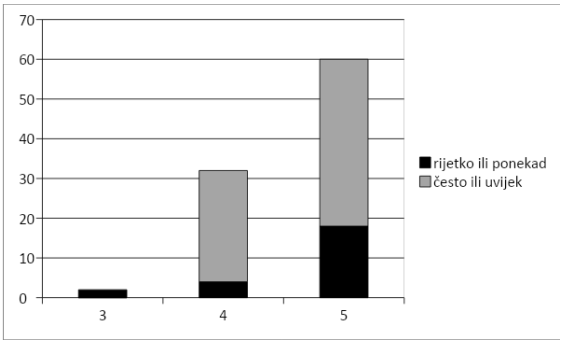
Rezultati zastupljenosti projektne nastave prema tipu predmeta (redovni i izborni) su slijedeći: Redovna nastava: Najveći postotak učenika, njih 69% je reklo da je projektna nastava ponekad zastupljena, 22% je reklo da je to često, 7% - rijetko.

Izborna nastava: Najveći postotak učenika, njih 53% je reklo da je projektna nastava često zastupljena, 19% je reklo da je to ponekad, 16% - uvijek i 12% - rijetko.

Ukupno uspoređujući rezultate zaključujemo da je projektna nastava, prema stavovima učenika sedmih i osmih razreda, znatno više zastupljena u izbornoj u odnosu na redovnu nastavu.

Prikazana je procjena motiviranosti učenika za projektnu nastavu prema spolu (slika 3). 15% učenica smatra da su rijetko ili ponekad motivirane za projektnu nastavu, dok 85% njih smatra da su motivirane često ili uvijek. 33% učenika smatra da su rijetko ili ponekad motivirani za projektnu nastavu, dok 67% njih smatra da je to često ili uvijek. Postoje razlike između učenika i učenica.

Grafikon na slici 4 prikazuje procjenu motiviranosti učenika za projektnu nastavu prema zaključnoj ocjeni na kraju nastavne godine. Ocjene su dobar, vrlo dobar i odličan.



Slika 4. Motiviranost za projektnu nastavu prema završnoj ocjeni ispitanika (p=0,55)

Iz grafičkog prikaza je vidljivo da učenici koji su u prethodnoj školskoj godini imali zaključnu ocjenu dobar, nisu motivirani za projektnu nastavu. Tek 2% njih se izjasnilo da interes postoji rijetko ili ponekad. Nadalje, vidljivo je da vrlo dobri učenici pokazuju veću motiviranost za projektnom nastavom. Njih 28% se izjasnilo da su često ili uvijek motivirani za taj oblik nastave, dok je 4% reklo da je to rijetko ili ponekad. Grafički prikaz pokazuje da su najveći interes za projektnom nastavom pokazali učenici s odličnom zaključnom ocjenom. 18% njih smatra da je rijetko ili ponekad motivirano za projektnu nastavu, a 42% se izjasnilo da je često ili uvijek. Dakle, vide se razlike u procjeni motivacije učenika s obzirom na zaključnu ocjenu u prethodnoj školskoj godini. Učenici koji su imali zaključnu ocjenu odličan, pokazuju najveću motiviranost za projektnu nastavu.

V. RASPRAVA

Cilj ovog rada je bio ispitati stavove učenika o projektnoj nastavi; o zanimljivosti projektne nastave u odnosu na tradicionalnu, o zastupljenost projektne nastave u ukupnoj nastavi, o motiviranosti učenika za projektnu nastavu, gledano prema spolu i uspješnosti učenika s obzirom na brojčano vrednovanje u prethodnoj školskoj godini.

U tu svrhu provedeno je istraživanje. Istraživanje je dalo odgovor na pitanje o zastupljenosti projektne nastave u osnovnoj školi iz perspektive učenika. U tome su se uvažavala sociodemografska obilježja ispitanika. U zastupljenosti projektne nastave prema tipu predmeta (redovni i izborni), uočena je razlika.

Vidljivo je da je projektna nastava više zastupljena u izbornim u odnosu na redovne predmete.. Naša je pretpostavka da ima više razloga za to. Jedan od njih je taj što kod učenika postoji veća motiviranost za izborne u odnosu na redovne predmete, stoga samim time postoji i veća motiviranost za projektnom nastavom. Drugi razlog može biti manji broj učenika u razredu/grupi kod izbornih u odnosu na redovne predmete, što učitelju može olakšati izvođenje projektne nastave. Nadalje, u izbornim predmetima, učenicima se daje djelomična mogućnost sudjelovanja u kreiranju kurikuluma prema njihovim željama i potrebama.

Ne postoji znatna razlika u zastupljenosti projektne nastave kod učenika sedmog i osmog razreda.

Postavlja se pitanje zašto su učenice više motiviranje za projektnu nastavu u odnosu na učenike? O tome bi se dalo raspravljati. Nekako je logično odgovoriti na pitanje zašto su učenici koji postižu bolji uspjeh u učenju, više motivirani za projektnu nastavu. Vjerojatno zato što oni s većim uspjehom mogu odraditi i takvu vrstu nastave, kao i to da im je ona veći izazov u odnosu na tradicionalnu nastavu.

Provođenje projektne nastave poželjno je provoditi i njegovati kontinuirano u svim nastavnim

predmetima, no postoje određene prepreke za realizaciju takve nastave. Prisutnost projektne nastave vezana je uz promjene u kurikulumu, ograničenim nastavnim sadržajima i temama, nedostatkom vremena u redovnoj nastavi. Isto tako, što je viši razred, učenicima je zahtjevniji kurikulum što daje manje prostora i vremena za realizaciju projektne nastave. Ipak, dosta učenika smatra da je uvelike prisutna projektna nastava u ukupnoj nastavi. Stoga je naglasak na individualiziranom pristupu kurikuluma prema pojedinim razredima i učenicima. Poželjno bi bilo povećati satnicu u kurikulumu i omogućiti slobodne teme za provođenje projektne nastave. Time se pruža mogućnost novog učenja na drukčiji, zanimljiviji način koji je učenicima ujedno i motivirajući.

Projektna nastava omogućuje timski rad, potiče socijalizaciju učenika i mogućnost usvajanja trajnog znanja. Projektna nastava uključuje metodu učenja otkrivanjem i istraživanjem i kao takva je nužna želimo li da učenici budu, po završetku školovanja, konkurentni na zahtjevnom tržištu rada.

VI. ZAKLJUČAK

Promjene u društvu, a koje su izazvane ubrzanim razvojem tehnologije, nužno nameću i potrebu promjene obrazovanja.

Tradicionalna nastava više ne može zadovoljiti potrebe današnjeg tržišta rada. Poslodavcima je u interesu zaposliti osobu koja uz stručna znanja posjeduje i mnoga druga znanja i vještine. Ta dodatna znanja i vještine su: timski rad, istraživanje, kritičko promišljanje, komunikativnost, asertivnost, svrhovito i vješto korištenje informacijsko-komunikacijske tehnologije, odgovornost, vrednovanje, prezentacijske vještine i dr. Jasno je da učenici ta znanja i vještine ne mogu steći tradicionalnom, već projektnom nastavom.

Projektna nastava nije nova, njezini počeci sežu u prošlost, kao i njezina primjena. Međutim, u našem obrazovnom sustavu je, nažalost, bila sustavno zanemarivana. No, ni danas ona nije pravilo, već je njezina realizacija vezana, većinom, za dobru volju pojedinog učitelja, iako trenutna reforma obrazovanja poznata pod nazivom „Škola za život“ također zagovara takvu vrstu nastave.

Projektna nastava je izazov, kako za učenika, tako i za učitelja. Pred učitelja se postavljaju novi izazovi konstantnog stručnog usavršavanja, bilo u primjeni informacijsko-komunikacijske tehnologije, novih digitalnih alata, načina planiranja nastave, načina vrednovanja i dr. Učenicima se, projektnom nastavom daje prilika da primjene stečena znanja i vještine.

Ovaj rad je dokazao da su promjene obrazovanja vidljive i u praksi.

VII. POPIS LITERATURE

[1] Bognar, L. i Matijević, M. (2002). Didaktika. Zagreb: Školska knjiga.

[2] Borić, E. i Škugor, A. (2014). Ostvarivanje kompetencija učenika istraživačkom izvanučioničkom nastavom prirode i društva. Croatian Journal of Education, 16 (1), 149-164.

[3] Braičić, Z., Đuranović, M., i Klasnić, I. (2015). Metode i postupci poučavanja i učenja u nastavi prirode i društva. Hrvatski časopis za odgoj i obrazovanje 17 (1), 83-95

[4] Cindrić, M., Miljković, D. i Strugar, V. (2010). Didaktika i kurikulum. Zagreb: IEP-D2.

[5] Cindrić, M., Miljković, D. i Strugar, V. (2010). Didaktika i kurikulum. Zagreb: IEP-D2.

[6] Fabijanić, V. (2014). Projektna nastava: primjena u izradi istraživačkih radova učenika. Educ.biol. 1:89-96.

[7] Isik, O. i Gucum, G. (2013). The Effect of Project Based Learning Approach on Elementary School Students' Motivation Toward Science and Technology Course. Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 28(3), 206-218.

[8] Kelly, T. i Kellam, N. (2009). A Theoretical Framework to Guide the Re-Engineering of Technology Education. Journal of Technology Education, 20(2), 37-49.

[9] Krsnik, R. (2003.). Problemski usmjerena nastava, deklarativno ili stvarno? U: Šesti hrvatski simpozij o nastavi fizike: Problemski i istraživački usmjerena nastava fizike. Split: HFD, str. 6-14, preuzeto 27.8.2019. sa http://nastava.hfd.hr/simpozij/2003/2003-Krsnik.pdf.

[10] Lam, S.; Cheng, R.W., i Ma, W.Y.K. (2009). Teacher and student intrinsic motivation in project-based learning. Instructional Science, 37(6):565-578.

[11] Matijević, M. i Radovanović, D. (2011) Nastava usmjerena na učenika. Zagreb: Školske novine.

[12] Meyer, H. (2002). Didaktika razredne kvake. Zagreb: Educa.

[13] Thomas, J.W. (2000). A review of research on project-based learning. The Autodesk Foundation 111. McInnis Parkway San Rafael, California (online) preuzeto 1.9.2019. sa http://www.bobpearlman.org/BestPractices/PBL_Research.pdf

[14] Vuković, N. (2003.). Mogućnost vrednovanja postignuća projektne nastave. Napredak 144 (2), 225-234.

[15] Zugaj, B. (2014). Utjecaj projektne nastave na promjenu stava kod učenika prema nastavnim sadržajima iz biologije. Educatio Biologiae, 1: 18-26.

Raspberry PI mini računalo kroz nastavnu cjelinu

Danijel Vrbanc, dipl. ing.
Željeznička tehnička škola – Moravice, Vrbovsko, Hrvatska
danijel.vrbanc@skole.hr

Sažetak - U radu je pokazano računalo Raspberry Pi (RPi) koje svojim odlikama i funkcionalnošću omogućava uporabu za različite namjene: od obrazovnog alata, minijaturne verzije stolnog računala, uporabe u industrijskim postrojenjima gdje se koristi sam ili integriran s PLC uređajima, pa do sve više prisutne IoT (Internet of Things) uporabe. Računalo u ovom radu predstavljeno je kroz nastavnu cjelinu koja se bavi njegovim glavnim odlikama. Predočena je nastavna cjelina kojom se obrađuje RPi, zamišljena na način da se rabe neki Microsoftovi alati sadržani u paketu Office 365. Pokazan je prezentacijski alat Sway za prezentaciju računala. Za bilješke na satu predavanja rabi se OneNote Class Notebook bilježnica (OneNote). Pokazana je njena uporaba kroz različite platforme: stolnih PC računala, hibridnih uređaja, tableta ili varijanti na pametnom telefonu. Za provjeru znanja tj. za provjeru očekivanih ishoda učenja prezentiran je Microsoft Forms, koji je namijenjen za uporabu na bilo kojoj od platformi prije spomenutih, a najspretnija uporaba pokazana je kroz korištenje pametnog telefona učitavanjem QR koda prikazanog na zaslonu projektora. U radu je pokazana bezpapirna (paper less) varijanta obrade nastavne teme, uz nužnu uporabu pomoćnih papirnatih materijala kod onih učenika i studenata koji nemaju uređaj s aktivnom olovkom (stylus).

Ključne riječi: OneNote Class Notebook, IoT, Raspberry PI, Office 365, Sway

I. UVOD

Arhitektura računala RPi se temelji na ARM procesoru koje može raditi u više načina rada. Uobičajeni način rada je Raspbian, službeni operacijski sustav (OS), koji predstavlja u biti Linux sustav. Međutim, osim što može raditi pod navedenim sustavom, on može koristiti i druge sustave kao što su OSMC (Open Source Media Center), OpenELEC (Open Embedded Linux Entertainment Center), RISC OS, Lakka, RaspBSD, RetroPie, Ubuntu Core (verzija Ubunto core distribucije predviđene za rad s IoT), Linutop OS, i što je najzanimljivije Windows IoT Core, što je u biti pojednostavnjena verzija Windows 10 OS-a.

OS je smješten na SD kartici, ali isto tako ih može biti više instalirano, te se kroz izbornik može odabrati onaj koji je potreban za pojedinu aplikaciju. Svrhovitije je rabiti pojedini operacijski sustav na zasebnoj SD kartici te mijenjati samo SD karticu - prema potrebi.

Kroz nastavnu cjelinu o RPi računalu pokazat će se njegove osnovne funkcionalnosti, kao običnog desktop računala kao i primjer uporabe kao IoT (Internet of Things – Internet stvari).

Nastavna cjelina je osmišljena kao bezpapirna varijanta nastavne lekcije, osim u slučaju kada student (učenik) sa sobom nema nikakav uređaj, nije mu u učionici (predavaonici) na raspolaganju, ili ne želi koristiti tehnologiju.

Najpoželjnija varijanta uporabe uređaja je BYOD (Bring Your Own Device). Tako se može rabiti hibridni uređaj s aktivnom olovkom, tablet s aktivnom olovkom, laptop, desktop računalo, ili običan pametni telefon s predinstaliranom OneNote [3] Microsoftovom aplikacijom.

Kao prezentacijski alat koristi se Sway [5] prezentacijski alat kojim se na jedan vrlo efektan način može u kratkom vremenu kreirati prezentacijski sadržaj.

II. UVOD U RPI RAČUNALO

RPi računalo (slika 1) predstavlja jedno potpuno računalo na kojem se uobičajeno koristi Raspbian operacijski sustav [1], makar se drugi OS-ovi također mogu koristiti kao što je u uvodu spomenuto. Raspbian predstavlja inačicu Linuxa i službeni je OS za RPi. Za taj OS napravljeno je najviše aplikacija kako programske, tako i dodatne hardverske podrške, a ujedno koristi najviše raspoloživih komponenti koje su na računalu dostupne.

Postoji više inačica RPi računala. U ovom radu će se koristiti inačica RPi Model 3, dok je najnovija inačica Model 4.

Računalo se koristi na način da se na njega spajaju vanjske ulazno izlazne jedinice: zaslon računala preko HDMI ulaza, tipkovnicu i miš preko USB ulaza, napajanje je standardno preko pretvornika 230V~/5V= (ne manje



Slika 1. RPi računalo Model 3

struje od 2,5A). Pored žičnog mrežnog priključka, na pločici se nalazi ugrađen Bluetooth i WIFI adapter te ga je moguće spojiti preko WIFI-ja na bežičnu mrežu. Pored utora za SD karticu, RPi ima konektor za spajanje vlastite kamere, kao i posebni izlaz za spajanje monitora bilo koje veličine kompatibilnog sa sustavom RPI-ja.

Kako ovakvo malo računalo nije uvijek spretno rabiti spajanjem svih vanjskih ulazno izlaznih jedinica (tipkovnice, miša, zaslona), može se koristiti bežičnim putem koristeći njegovu sposobnost priključka na WIFI. Tako se može koristiti VNC pristup [7] ili WinSCP za pristup [8] (slika 2). Time se omogućava udaljeni pristup računalu i kada nije fizički dostupan bez obzira je li uređaj unutar lokalne mreže ili mu se pristupa na bilo kojoj lokaciji preko usmjernika (routera). Ovaj slučaj je vrlo čest kada se koristi računalo kao IoT.

RPi-om se može udaljeno upravljati kroz naredbeni redak (command prompt), ili kroz grafičko sučelje. Oba pristupa omogućena su kroz to, ili slično programsko internetsko sučelje.

Najzanimljiviji dio uređaja predstavljaju *General Purpose Input Output* (GPIO) portovi: set od 40 iglica (pinova) preko kojih se upravlja ili s računalom, ili računalo upravlja različitim uređajima (slika 3). To je najveća snaga tog računala i njegova mogućnost punog iskorištavanja.

U uvodu ove nastavne lekcije sve prikazane slike o Raspberry računalu nalaze se u OneNote-u „Biblioteci sadržaja“ pod sekcijom i stranicom, koja je dostupna svim polaznicima. Prema vlastitom načinu rada, polaznici (dok predavač „vrti“ svoju prezentaciju, preslikavaju slike iz te biblioteke te ga slažu na svoju stranicu po svom nahođenju, te pišu bilješke predavača i unose ih, bilo stylusom ili tipkovnicom (slika 4). Pregled rada polaznika lako se provjeri uvidom u OneNote-u od strane nastavnika.

III. RPI U IOT APLIKACIJI

Primjer uporabe kao i sama učinkovitost i svrhovitost ovog malog računala pokazat će se kroz primjer realiziranja mini meteorološke postaje. Meteorološka postaja u ovom primjeru mjeri i prikuplja podatke o tlaku zraka i temperaturi.

A. Opis projekta mini meteorološke postaje

Mini meteorološka postaja sastoji se od četiri osnovne grupe elementa: RPi računala, senzora, pokaznika (displeja), te programske podrške.

U sklopu rada ove aplikacije postoje dva senzora: senzor temperature i senzor tlaka zraka. Senzori pomoću mjernog osjetnika detektiraju promjenu temperature i tlaka zraka te ih šalju putem određenih GPIO portova u RPi računalo.

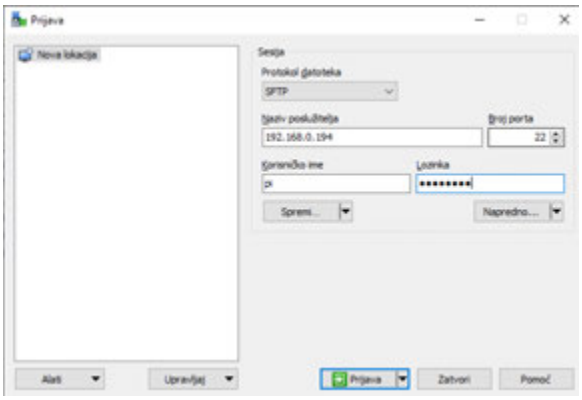
Rezultati mjerenja se očitavaju određenim tempom te se prikazuju na zaslonu koji se nalazi u području nazočnosti osoba koji su potencijalno zainteresirani za informacije o vremenskim prilikama. Prijenos podataka od računala do pokaznika je definiran putem određenih GPIO portova.

Na računalu je pokrenut pozadinski skriptni program koji prikupljene podatke upisuje u datoteku tempom očitavanja mjerenih rezultata. Taj program se pokreće prilikom dizanja operacijskog sustava.

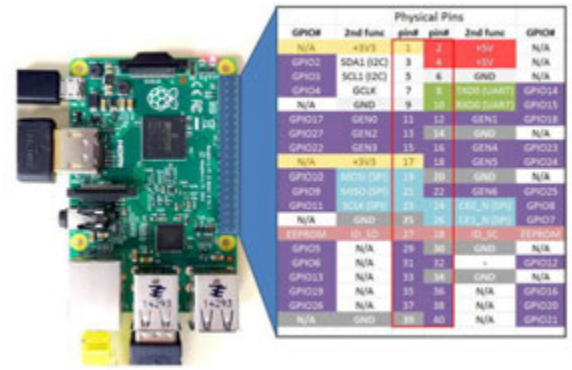
Iz tih podataka u kojima se nalazi i vrijeme očitavanja rezultata, svakih 15 minuta se uzima podatak o temperaturi i tlaku zraka te se sve šalje na poslužitelj u za to pripremljenu bazu podataka.

Izgled samog sučelja i očitani podaci dostupni su na https://malina.vrbanc.eu (slika 5).

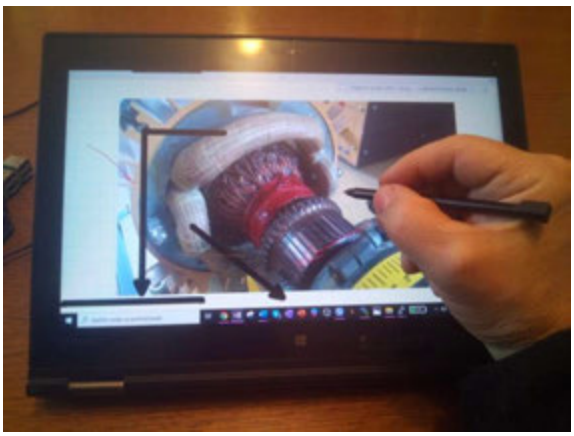
U ovom dijelu nastavne cjeline učenici iz Biblioteke sadržaja OneNote-a kopiraju slike koje je nastavnik stavio na raspolaganje u svoje bilježnice, te ih prema svome



Slika 2. WinSCP sučelje za udaljeni pristup RPi preko udaljenog računala pod Windows operacijskim sustavom



Slika 3. GPIO portovi RPi računala preko kojih se RPi spaja na razne ulazno izlazne jedinice



Slika 4. OneNote na hibridnom računalu u primjeru uporabe aktivne olovke (stylusa)

nahođenju, kao i u uvodu, razmještaju u svoj prostor (svoje osobne stranice bilježnice) te upotpunjavaju stranice s dodatnim sadržajima, tekstom, skicama i sličnim materijalima bitnima za nastavnu lekciju.

B. Idejna realizacija projekta mini meteorološke postaje

Pri izradi projekta učenici pregledavaju raspoloživi materijal za izradu: računalo, senzore, mehanički materijal za izradu meteorološke postaje, pričvrsni materijal te razmatraju mogućnosti izrade.

Učenici se mogu grupirati u parove, trojke ili četvorke u raznim segmentima realizacije projekta. Nastavnik ih upućuje u zasebne cjeline projekta, te se učenici prema svojim afinitetima i sklonostima pridružuju u navedene grupe.

Jedna grupa treba osmisliti poziciju meteorološke postaje obzirom na različite čimbenike kao što su: osunčanost pozicije mjerenja i dostupnost napona napajanja. Druga grupa se posvećuje mehaničkoj realizaciji samog mjernog uređaja postaje. Treća grupa se posvećuje povezivanju senzora s računalom te softverskom podrškom. Četvrta grupa ima zadatak osmišljavanja web sučelja tj. izrade web stranice koja će omogućavati što jednostavniji prikaz mjerenih rezultata. Peta grupa ostvaruje povezivanje mini meteorološke postaje na hosting poslužitelj kako bi se rezultati pohranjivali i pratili kroz određeno vremensko razdoblje. Moguće je osmisliti još podgrupa koje mogu podijeliti navedene grupe. Kreiranje grupa učenika kao njihova veličina varira od razreda do razreda. Ovisi o broju učenika, njihovom afinitetu za izradu pojedine cjeline projekta i o senzibilitetu samog razreda koji se razlikuje od razreda do razreda.

Učenici sve svoje ideje, skice, crteže snimaju putem mobitela ili crtaju putem digitalne olovke (stylusa) i to u suradničkom dijelu OneNote-a gdje svi mogu raditi sve (*read, write, update*). Naravno da se od učenika očekuje odgovornost što je vrlo lako postići čim se radi na projektu koji predstavlja konkretan zadatak.

C. Realizacija projekta mini meteorološke postaje

Nakon što su određene sve cjeline projekta, pristupa se realizaciji pojedinih cjelina. Svaka grupa učenika prezentira cijelom razredu svoj dio odrađenih zadataka te se razvija diskusija oko načina realizacije i eventualnih nekih dvojbenih načina izvedbe. Nastavnik ima ulogu moderatora diskusije kao i upravitelja cijelog projekta

Svi segmenti izrade upisuju se putem bilježaka, slikanja fotografija ili videa te upisuju u OneNote.

Kako se radi o prilično složenom projektu, nastavnik sugerira način uporabe pojedinih senzora preko već ostvarenih drugih projekata koji rabe RPi računalo. Obzirom da je svjetska zajednica koja se bavi projektima vezanim u RPi računalo velika, način realizacije je olakšan primjerima iz prakse u svrhu ostvarivanja vlastitog projekta koji opet ima jedinstvene odlike.

Web sučelje je izrađeno programiranjem u *html*-u i moguće ga je iščitati u bilo kojem pregledniku. Web sučelje je prilagođeno prikazu na bilo kojem uređaju: stolnom računalu, tabletu, pametnom telefonu (slika 5).

Detalj realizacije spajanja temperaturnog osjetnika na računalo prikazana je na slici 6.

Realizacija spajanja pokaznika u holu škole prikazana je na slici 7.

D. Puštanje u pogon i testiranje

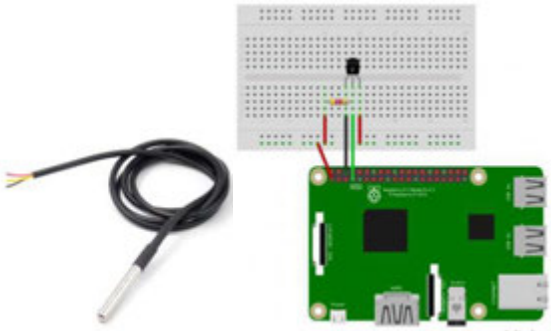
Svaki uređaj treba određeno vrijeme kako bi ga se testiralo te time otklonili njegovi nedostaci. Kroz jedno određeno vrijeme (otprilike dva tjedna), uređaj se testira. Ustanovljeno je kako je dovoljno spremati u bazu podataka na serveru izmjerene temperaturne i vrijednosti tlaka zraka svakih 15 minuta. Ti podaci su vrijedni te se u nekoj budućnosti mogu analizirati podaci o mikroklimi mjerenog područja.

Na pokazniku se prikazuju trenutni podaci o temperaturi, tlaku zraka, te aktualnoj prognozi za tekući dan koji se uzimaju sa službenih stranica DHMZ-a.

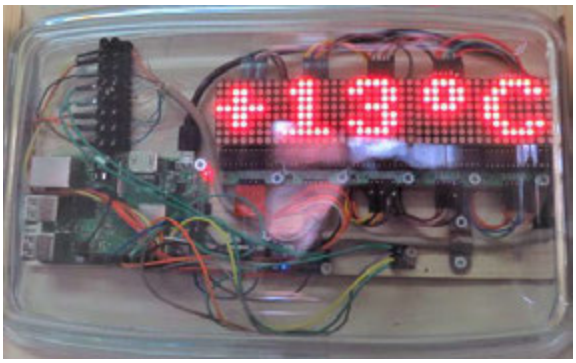
Postaja radi stalno tj. 24/7/365 dana u godini. Pristup postaji omogućen je preko bilo kojeg programa za pristup RPi, npr. preko lokalnog računala ili udaljenog računala preko *remote.it* [7] protokola koji nam daje vanjsku IP



Slika 5. Slika sučelja na web pregledniku meteorološke postaje



Slika 6. Temperaturni osjetnik i shema spajanja na Rpi



Slika 7. Pokaznik sa „srcem“ meteorološke postaje smješten u holu škole (RPi, senzor tlaka zraka i matrični pokaznik)

adresu i port preko kojeg se ulazi preko usmjernika (routera) (slika 8).

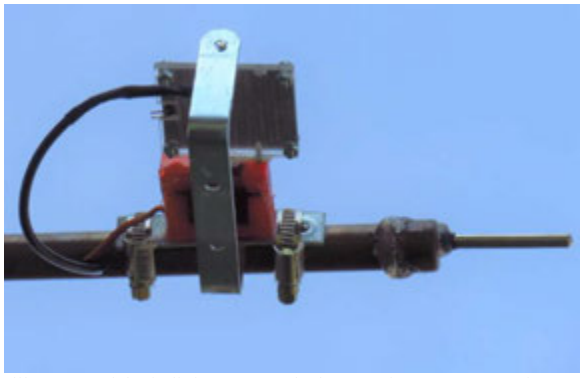
Izgled vanjske jedinice (temperaturnog mjernog senzora) prikazan je na slici 9.

Kao i u ostalim cjelinama izrade projekta, tako i u ovoj učenici sve što su radili, slikaju, snimaju, čine bilješke i sve stavljaju u OneNote.

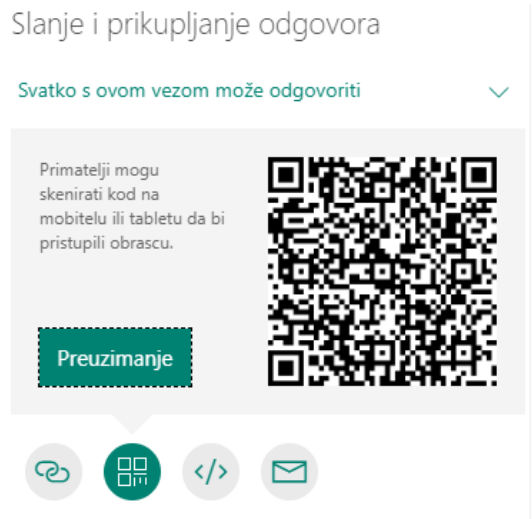
Treba napomenuti kako se svi podaci u OneNote-u, bez obzira koji se uređaj rabi, sinkroniziraju te su svi podaci osvježeni bez obzira koju vrstu uređaja rabimo, te jesu li unutar WIFI mreže, spojeni mrežnim kablom ili je računalo spojeno preko podatkovnog uređaja na GSM mrežu. Ako uređaj trenutno nije spojen na navedene pristupne točke, svi podaci u OneNote-u se spremaju lokalno, te se u trenutku spajanja na Internet sinkroniziraju.



Slika 8. Sučelje remote.it servisa kojim se pristupa udaljenom RPi-ju (plavo je lokalna IP adresa, a crveno vanjska IP adresa)



Slika 9. Vanjska jedinica meteorološke postaje s temperaturnim senzorom



Slika 10. QR kod za pristup provjeri znanja skeniranjem

IV. Prezentacijski alat Sway

Pored sveprisutnog PowerPointa, Office 365 [2] nudi jedan zanimljiv alat za prezentiranje. Radi se o Sway-u [3]. Pomoću vrlo jednostavnih radnji mogu se pokazati dojmljive prezentacije. U nastavnoj lekciji o uporabi RPi-ja može se rabiti taj alat, makar se može koristiti već spomenuti PowerPoint ili naprimjer Prezi [9].

U usporedbi s PowerPointom koji može imati lokalnu ili mrežnu inačicu, Sway nudi samo mrežnu inačicu. U njemu postoje gotovi predlošci u koje možemo umetati slike, ugrađivati poveznice na video uratke, uvesti PDF datoteke, PowerPoint datoteke kao i poveznice na servise: YouTube, Flickr, OneDrive, BingImages, Pickit. Prilagođen je uporabi na bilo kojem zaslonu te izgled prezentacije nije isti na stolnom računalu kao naprimjer na pametnom telefonu.

Alat je vrlo intuitivan, te se prezentacije mogu kreirati u vrlo kratkom vremenu.

Prijelaza sa slike na sliku ima nekoliko vrsta, kao i pomicanje slajdova. Ako se želi mijenjati izgled prezentacije, to se može učiniti i kada je prezentacija gotova, te se mogućnošću promjene teme mijenja kompletan izgled.

Pomicanje slajdova može biti horizontalno i vertikalno.

V. Provjera znanja

Obzirom da je Office 365 dostupan svim nastavnicima i učenicima u školskom sustavu RH, to je pored OneNotea dostupan i MS Forms [5] kojim se mogu utvrditi stečene spoznaje i činjenice koje su se usvojile tijekom izrade projekta.

Nastavnik može po završetku kreiranja provjere znanja kreirati QR kod, pokazati ga na projektorskom platnu kojeg učenici skeniraju svojim pametnim telefonima te pristupaju rješavanju zadanih pitanja.

Time je i u ovom segmentu nastavnog procesa izbjegnuta uporaba papirnatog medija (slika 10).

VI. Kratak pregled izvedenog projekta

Ovaj projekt namijenjen je učenicima trećeg razreda smjerova Tehničar za mehatroniku ili Tehničara za računalstvo i objedinjuje nekoliko predmeta koji se poučavaju u trećem razredu navedenih zanimanja: „Računalstvo“, „Informacije i komunikacije“, „Senzorika“, „Elektronički sklopovi“, „Baze podataka“, „Algoritmi i programiranje“.

Ujedno nije na odmet spomenuti cijenu koštanja materijala za izradbu koja uključuje: RPi računalo, senzor temperature, senzor tlaka, pokaznik, ožičenje, pričvrsni pribor, kanalice za kablove, vanjska prozirna maska ispred displeja: te sve to iznosi otprilike oko 900,00 – 1.000,00 Kn.

Vrijeme izrade projekta s osmišljavanjem sustava je otprilike 40 – 50 čovjek-sati. U razrednom odjeljenju od 10 učenika, projekt se može odraditi za 7 – 8 dana s cjelodnevnom satnicom.

ZAKLJUČAK

Maleni RPi uređaj ima sve odlike klasičnog stolnog ili prijenosnog računala. Posebno je specijalizirano za IoT projekte jer sadrži poseban set od 40 iglica kojima ostvarujemo povezivanje uređaja s vanjskim uređajima kao i upravljanje računala putem tih iglica (pinova) koji u biti predstavljaju ulazno izlazne sabirnice podataka.

Upravljanje RPi računalom se može izvesti lokalno ili udaljeno; rabeći mrežni kabel, WIFI, ili GSM signal.

Svojim mogućnostima RPi je sve više prisutan kako kod hobista tako i u industriji. Postoji široka zajednica koja se bavi njegovom primjenom te je lako moguće ostvariti širok spektar različitih projekata iz već ostvarenih projekata.

U ovom radu prikazan je rad mini meteorološke postaje čijom realizacijom u razredu je pokazan pored ostalog i primjer uporabe tehnologije u nastavi na način da tehnologija ne ugrožava nastavni proces, već da mu je jedna vrlo istaknuta nadogradnja. Svrha tehnologije u obrazovanju je biti prvenstveno nastavno pomagalo. Dakle obrazovni proces se koncentrira na ono što želimo objasniti, pokazati, istražiti, a tehnologija je tu koja nam samo omogućava brži pristup i zornost informacija.

Dakle tehnologija – IKT (informacijsko komunikacijska tehnologija) mora biti „nevidljivi“ pratilac nastavnog procesa. Ako to nije, najbolje je primijeniti „klasične“ načine poučavanja.

LITERATURA

[1] Raspberry Pi Foundation (bez dat.) [Na internetu]. Dostupno: <*https://www.raspberrypi.org/*> [pristupano 11.01.2020.]

[2] Microsoft Corporation (bez dat.) [Na internetu]. Dostupno: <*https://office365.skole.hr/*> [pristupano 11.01.2020.]

[3] Microsoft Corporation (bez dat.) Dostupno: <*https://www.onenote.com/classnotebook*> [pristupano 11.01.2020.]

[4] Microsoft Corporation (bez dat.) [Na internetu]. Dostupno: <*https://sway.office.com/* > [pristupano 11.01.2020.]

[5] Microsoft Corporation (bez dat.) [Na internetu]. Dostupno: <*https://forms.office.com/*> [pristupano 11.01.2020.]

[6] RealVNC Ltd (bez dat.) [Na internetu]. Dostupno: <*https://www.realvnc.com/en/*> [pristupano 11.01.2020.]

[7] remot3.it, Inc, Computer Networking (bez dat.) [Na internetu]. Dostupno: <*https://app.remote.it/*> [pristupano 11.01.2020.]

[8] WinSCP - Information Technology and Services (bez dat.) [Na internetu]. Dostupno: <*https://winscp.net/eng/index.php*> [pristupano 11.01.2020.]

[9] Prezi Inc. [Na internetu]. (bez dat.) Dostupno: <*https://prezi.com/*> [pristupano 11.01.2020.]

Sustav upravljanja zadacima iz područja micro:bit

Tomislav Adamović
Veleučilište u Bjelovaru/Računarstvo, Bjelovar, Hrvatska
tadamovic@vub.hr

***Sažetak* - U ovom radu govori se o karakteristikama sustava za prikupljanje zadataka iz područja micro:bit korištenjem edukacijske platforme Microsoft makecode te njihova analiza pomoću alata za procjenu kompleksnosti koda. Opisani su načini kategorizacije zadataka na temelju upotrijebljenih ključnih riječi u programima iz čega proizlaze ishodi učenja. Zaključno, prikazana je procjena kvalitete zadataka za svaku kategoriju.**

Ključne riječi - micro:bit; kompleksnost koda; ishodi učenja; procjena kvalitete;

I. UVOD

Micro:bit je programabilni elektronički sklop čija je namjena jednostavno učenje programiranja i upoznavanje s mikroračunalima.

Iako postoji veliki broj uređivača programskog koda za micro:bit, Microsoftov internetski uređivač Makecode https://makecode.microbit.org je prvi preporučeni uređivač na službenoj internetskoj stranici micro:bita https://microbit.org. Makecode koristi programske blokove u obliku grafičkih elemenata za izrađivanje programskog koda. Programski blokovi su raspoređeni u kategorije koje sadrže logičke cjeline programskih naredbi poput naredbi za prikaz i grananje, te naredbi za izradu funkcija, polja i sl.
U pozadini programskih blokova nalazi se JavaScript programski jezik.

Pored uređivanja programskog koda, MakeCode omogućuje dijeljenje programskih rješenja u obliku jedinstvenih poveznica te API-a preko kojeg se može pristupiti JavaScript kodu programskog rješenja.

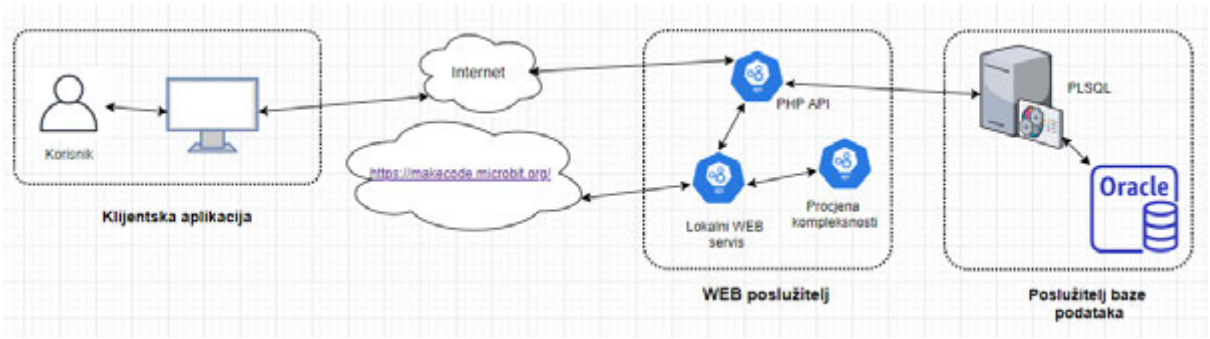
Navedene karakteristike MakeCode uređivača daju osnovu za izgradnju sustava za evidenciju, analizu i upravljanje programskim zadacima i rješenjima iz micro:bita.

II. ARHITEKTURA SUSATAVA ZA UPRAVLJANJE ZADACIMA IZ MICRO:BITA

Sustav za upravljanje zadacima iz micro:bita je razvijen na poslužiteljima Veleučilišta u Bjelovaru i bit će dostupan putem poveznice https://vub.hr/microbit. Izgrađen je po principima troslojne arhitekture prikazane na slici 1 koja na klijentskoj strani koristi HTML5, BootStrap i jQuery, na srednjem sloju PHP, a na poslužiteljskoj strani PL/SQL i Oracle bazu podataka.

Model baze podataka prikazan je na slici 2. Preko klijentske aplikacije dostupni su dijelovi modela baze podataka koji osiguravaju prijavu korisnika na sustav, spremanje programskih rješenja i zadataka iz Makecode uređivača, generiranja zadataka na temelju određenih ishoda te razne vrste pregleda, izvještaja i analiza. Ostali dio baze podatka namijenjen je za pozadinsku obradu dostavljenih programskih zadataka i rješenja.
U sustavu je omogućeno dijeljenje i dodavanje novih zadataka, a prilikom unosa novog zadatka preko poveznice vrši se analiza dostavljenog programa koja će razvrstati programske zadatke u pripadajuće kategorije. Sustav se trenutno sadrži 3 kategorije i 18 potkategorija programskih zadatka koje proizlaze iz kategorizacije u MakeCode uređivaču. Popis svih trenutnih kategorija i potkategorija može se vidjeti u tablici 1.

Analiza programa se temelji na provjeri skupa naredbi koji se nalaze u dostavljenom programskom kodu putem 194 regularna izraza. Trenutno, sustav sadrži 63 kategorizirana zadatka.



Slika 1 Troslojna arhitektura sustava za upravljanje zadacima iz micro:bita



Slika 2 Model baze podataka za sustav upravljanja zadacima iz micro:bita

Osim navedene analize, svaki programski kod zadatka prolazi kroz alat za procjenu kompleksnosti i ta se procjena zapisuje u bazu. Rezultati ove procjene raspoređuju zadatke unutar kategorije.

Korištenjem kategorizacije i procjene kompleksnosti izabiru se ishodi učenja. Sustav je tada spreman za pojedini ishod učenja kvalitetno i individualno generirati zadatke za polaznike edukacije.

III. PROCJENA KOMPLEKSNOSTI PROGRAMSKOG ZADATKA

Za svaki programski jezik moguće je pronaći nekoliko besplatno dostupnih alata za provjeru kompleksnosti programskog koda. Svi ovi alati se temelje na tri načina analize kompleksnosti: McCabeova ciklomatska složenost, Halsteadova metrika kompleksnosti programa i Indeks održivosti.

A. McCabeova ciklomatska složenost

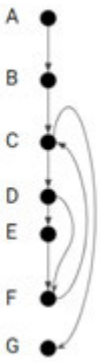
Ovu metriku kompleksnosti razvio je Thomas J. McCabe 1976. godine [1] i temelji se na analizi dijagrama toka programskog koda. Dijagram toka prikazuje program u obliku čvorova i usmjerenih grana. Čvorovi predstavljaju zadatke obrade dok usmjerene grane predstavljaju tok upravljanja između čvorova. Na slici 3. naveden je primjer kako iz programskog koda dobiti dijagram toka. Nakon određivanja čvorova i usmjerenih grana, određuje se broj neovisnih putanja odnosno putanja koje imaju barem jednu granu kojom se prije nije prošlo na bilo kojem drugom putu.
Iz primjera na slici 3 može se zaključiti kako dijagram toka ima 8 neovisnih putanja i 7 čvorova.
Kompleksnost se računa kao

$$K = \text{neovisne putanje} - \text{čvorovi} + 2. \quad (1)$$

što ovom primjeru daje kompleksnost

$$K = 8 - 7 + 2 = 3 \quad (2)$$

```
function isPrime(n) {
    $result = true;                    A
    $int = 2;                          B
    while ($i < $n){                   C
      if ($n % $i == 0){               D
         $result = false;              E
      }
      $i++;                            F
   }
   return $result;                     G
}
```



Slika 3. Primjer analize kompleksnosti programskog koda

B. Halsteadova metrika kompleksnosti programskog koda

Halsteadova metrika [2] svaki programski kod promatra kao zbirku tokena koji se klasificiraju kao operandi ili operatori. Operandi su znakovi koji imaju vrijednost, a obično su to varijable i konstante. Sve ostalo se smatra operatorom. Primjer operatora mogu biti zarezi, zagrade, aritmetički operatori i sl. Ukoliko se operatori pojavljuju kao par tada se računaju kao jedan token npr. lijeva i desna zagrada smatraju se jednim pojavljivanjem zagrada. Na slici 4. je naveden primjer programskog koda za analizu Halsteadove metrike.

```
main(){
  int a, b, c, avg;
  scanf("%d %d %d", &a, &b, &c);
  avg = (a + b + c) / 3;
  printf("avg = %d", avg);
}
```

Slika 4. Primjer programskog koda za analizu Halsteadove metrike

C. *Index održivosti*

Kako se navodi u [4] Oman i Hagemeister sa Sveučilišta Idaho projektirali su indeks održivosti programskog koda [3]. Procjena indeksa održivosti se temelji na nekoliko mjernih podataka kao što su: Halsteadov volumen, ciklomatska složenost, prosječni broj redaka koda i postotak komentara u odnosu na prosječni broj redaka.

Kategorija	Potkategorija	Broj zadataka	Težina	Ciklomatska složenost	Broj logičkih linija koda	Indeks održivosti
Osnovna	1 Basic	7	1 - 5.192	1	1 - 13	62.609 - 100
Osnovna	2 Input	6	1 - 4.125	1	1 - 13	74.483 - 95.096
Osnovna	3 Music	0	-	-	-	-
Osnovna	4 Led	3	3.5 - 7.059	1	10 - 22	65.112 - 65.112
Osnovna	5 Radio	3	1.65 - 4	1	3 - 8	85.799 - 95.096
Osnovna	6. Loops	0	-	-	-	-
Osnovna	7. Logic	1	9.579	5	15	59.177 - 59.177
Osnovna	8. Variables	17	1.773 - 17.778	1 - 12	2 - 31	56.561 - 96.386
Osnovna	9 Math	4	1.5 - 6.957	1 - 6	2 - 19	58.089 - 93.198
Napredna	10 Function	7	3.5 - 13.912	2 - 4	7 - 15	66.028 - 83.182
Napredna	11 Array	0	-	-	-	-
Napredna	12 Text	1	5.714	2	6	82.316 - 82.316
Napredna	13 Game	9	3.269 - 24.939	1 - 12	4 - 56	52.625 - 95.096
Napredna	14 Images	4	7.077 - 18	1 - 17	8 - 26	51.785 - 71.84
Napredna	15 Pins	0	-	-	-	-
Napredna	16 Serial	0	-	-	-	-
Napredna	17 Control	1	19.474	5	33	70.149 - 70.149
Proširenja	18 Bluetooth	0	-	-	-	-

Tablica 1. Procjena kvalitete zadataka za svaku potkategoriju

Navedeni primjer ima 12 jedinstvenih operatora (main, (), {}, int, scanf, &, =, +, /, printf, ",", ;) i 7 jedinstvenih operanada (a, b, c, avg, "%d %d %d", 3, "avg = %d") što ukupno daje 19 tokena. Ova tri parametra su temelj za procjenu sljedećih Halsteadovih metrika: veličina programa, volumen, težina, uloženi rad, vrijeme i broj mogućih pogrešaka.

IV KVALITETA PROGRAMSKIH ZADATAKA

Prilikom unosa poveznice zadatka u sustav za upravljanje zadataka na srednjem sloju se automatski poziva Makecode API koji vraća programski kod zadatka. Ovaj programski kod predstavlja ulazni parametar u „plato“ https://github.com/es-analysis/plato, alat za procjenu kompleksnosti JavaScript programskog koda. Karakteristike i primjena alata „plato“ mogu se vidjeti u

U tablici 1. se mogu vidjeti neki od tih parametara: halsteadova težina, ciklomatska složenost, broj logičkih linija koda i indeks održivosti. Osim parametara kompleksnosti predstavljene su i kategorije i potkategorije zadataka te trenutni ukupan broj zadataka u sustavu za pojedinu potkategoriju.

Zadaci iz tablice 1 su dobiveni po principu da sljedeća potkategorija podrazumijeva znanje i iskustvo iz svih prethodnih potkategorija. Međutim, moguće je odabrati zadatke koji pripadaju samo određenoj potkategoriji bez poveznice na prethodne potkategorije. Ako se u ovaj odabir uključi i odabir razina kompleksnosti programskog koda tada postoji veliki broj kombinacija različitih vrsta ishoda učenja.

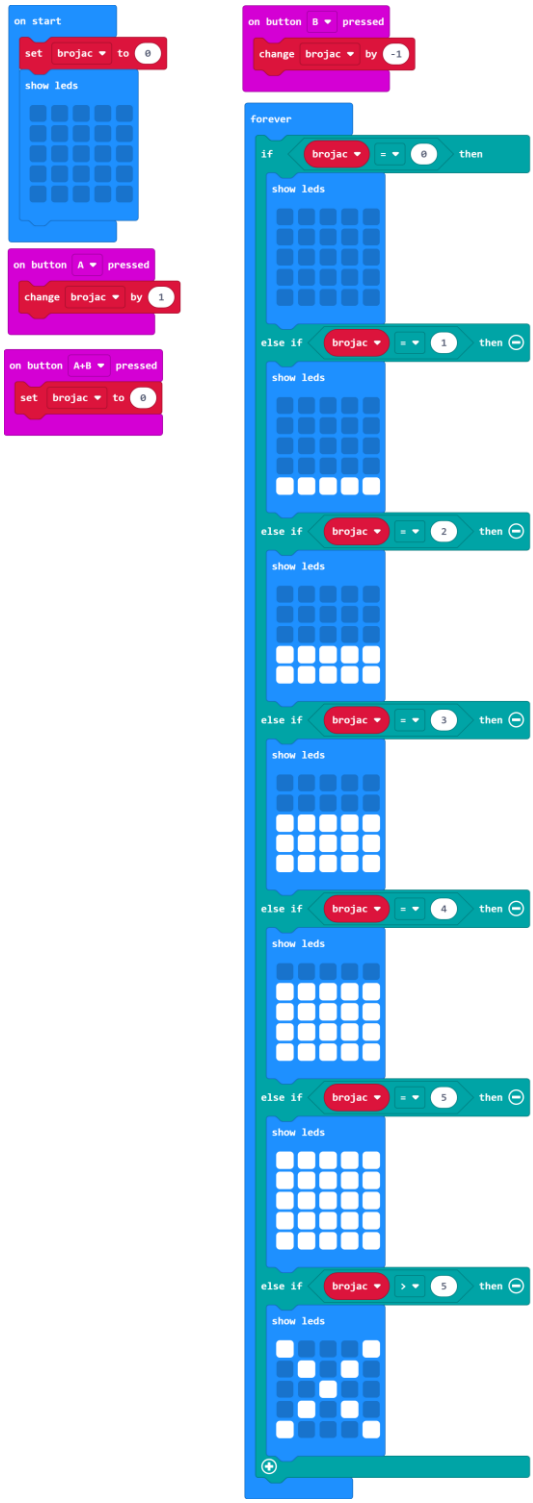
V KORIŠTENJE SUSTAVA

Kvaliteta sustava ovisi o količini i raznovrsnosti zadataka. Kada se prikupi dovoljno velika baza zadataka predavač će moći odabrati kategoriju koju želi predavati i razinu kompleksnosti zadataka. Na temelju ovih podataka, sustav će ponuditi zadatke iz odabrane kategorije i rasporediti ih po razini kompleksnosti unutar odabranih granica kompleksnosti. Na ovaj način sustav se može prilagoditi raznim skupinama polaznika predavanja.

U sljedećem primjeru prikazanom na slici 5 i 6 može se vidjeti kako sustav korištenjem parametara kompleksnosti raspoređuje zadatke. Na slici 5 se vidi zadatak čija je McCabeova ciklomatska složenost jednaka 2, indeks održivosti je 25.104, a Halsteadova težina je 4.125



Slika 5. Primjer zadatka s Halsteadovom težinom od 4.125



Slika 6. Primjer zadatka s Halsteadovom težinom od 14.52

Na slici 6 prikazan je zadatak kojemu je McCabeova ciklomatska složenost jednaka 8, indeks održivosti je 70.832, a Healsteadova težina je 14.52.

Upotrebom ovih parametara sustav će najprije učeniku ponuditi zadatak sa slike 5, a zatim zadatak sa slike 6. Sljedeći zadatak učenik dobiva uzimajući u obzir analizu kompleksnosti zadataka i ponuđenog rješenja i vrijeme rješavanja prethodnog zadatka. Veća

raznovrsnost i veći broj zadataka omogućit će bolju prilagodbu učenicima u njihovom radu.

VI Zaključak

Navedeni sustav može objektivno procijeniti kompleksnost, kategorizirati zadatke i dati procjenu napretka učenika. Za kvalitetnu upotrebu sustava trebat će velik broj raznovrsnih zadataka i podrška zajednice koja se bavi edukacijom u području micro:bita. S vremenom, sustav će korištenjem prikupljenih podataka moći procijeniti koji od parametara kompleksnosti ima najveću težinu u odabiru sljedećeg zadatka, koja kategorija zadataka je najteža te predložiti rješenja za optimalno učenje programiranja.

Literatura

[1] T. J. McCabe, "A Complexity Measure," IEEE Trans. Softw. Eng., vol. SE-2, no. 4, pp. 308–320, 1976.

[2] M. H. Halstead, Elements of software science. Elsevier New York, 1977.

[3] P. Oman and J. Hagemeister, "Construction and testing of polynomials predicting software maintainability," J. Syst. Softw., vol. 24, no. 3, pp. 251–266, 1994.

[4] S. A. Mousavi, "Maintainability Evaluation of Single Page Application Frameworks - Angular2 vs. React" Bachelor Thesis Project, 2016.

Aplikacije za fotografsko prepoznavanje i rješavanje matematičkih zadataka

Hrvoje Kovač*, Ivan Nađ **
* Visoko učilište Algebra, Zagreb, Hrvatska
** Visoko učilište Algebra, Zagreb, Hrvatska; Osnovna škola Eugena Kvaternika, Velika Gorica, Hrvatska

***Sažetak* - Ovaj rad opisuje funkcionalnosti, prednosti i ograničenja aplikacija za fotografsko prepoznavanje i rješavanje matematičkih zadataka. Posebno se analizira metodičnost rješavanja zadataka i mogućnost odabira metoda rješavanja. U radu je veliki naglasak stavljen na primjenjivost opisanih aplikacija u nastavi matematike u osnovnoj i srednjoj školi.**

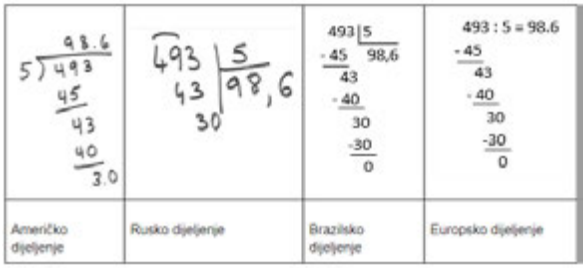
Ključne riječi – računalom podržano obrazovanje; matematika; aplikacije; fotografiranje; skeniranje

I. UVOD

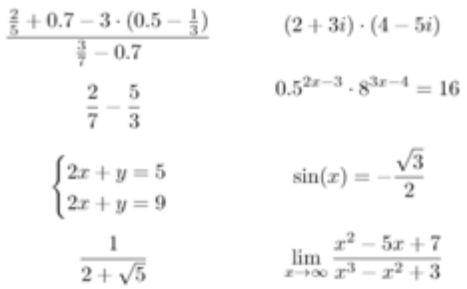
U današnje vrijeme su vještine vezane uz STEM područje, a pogotovo znanje i korištenje matematike, prepoznate u svijetu kao ključna kompetencija za najviše plaćena zanimanja na tržištu.

Taj trend prepoznale su i mnoge *startup* kompanije, pa samo u SAD-u trenutno postoji više od 4200 kompanija koje mijenjaju koncept tradicionalne edukacije nudeći alternativna rješenja u obliku aplikacija za pomoć s domaćom zadaćom, aplikacija za rješavanje zadataka, aplikacija za podučavanje, itd. Pritom svaka aplikacija pokušava privući što veći broj korisnika nudeći različite popratne usluge poput automatskog generiranja problema za vježbu, dodatnih materijala za učenje, vizualno atraktivnih metoda rješavanja zadataka (na primjer tzv. „pisane" tj. *long* računske operacije) i sl. Zbog toga je učenik (a i nastavnik) stavljen pred težak izbor jedne od više desetaka dostupnih aplikacija za učenje nekog školskog predmeta, ne znajući pritom je li aplikacija kvalitetna i dobra za njega.

Cilj ovog rada bio je istražiti postojeće matematičke aplikacije za automatizirano rješavanje matematičkih zadataka. Automatizirano rješavanje matematičkih zadataka u ovom kontekstu podrazumijeva rješenje generirano nekom vrstom programskog alata (*software engine*), za razliku od rješenja napisanih od strane čovjeka.



Slika 1. Različiti algoritmi pisanog dijeljenja (izvor: [7])



Slika 2. Primjeri pitanja korištenih u testu

Aplikacije koje „strojno" rješavaju matematičke zadatke općenito se mogu diferencirati prema tome dozvoljavaju li „fotografiranje" zadatka ili podrazumijevaju samo tekstualni unos sadržaja. Dio aplikacija koje smo istraživali u našem radu koristi neki oblik OCR algoritma (*optical character recognition*) za pretvaranje slike u niz tekstualnih simbola. S obzirom da je matematika izuzetno specifično područje s puno specijalnih znakova, razvijanje OCR algoritma koji će prepoznavati matematički izraz zahtjeva veliku količinu fotografiranih zadataka.

Iako je općeprihvaćeno da je matematika univerzalna, matematička notacija čak i najjednostavnijih operacija nije usklađena. Na Slici 1. su prikazani različiti algoritmi dijeljenja u svijetu.

Tehnologija OCR-a je bez sumnje doprinijela popularnosti matematičkih aplikacija za rješavanje problema zbog jednostavnosti unosa, no opseg i kvaliteta sadržaja takvih aplikacija ostali su neistraženo pitanje. Strojno generiranje rješenja svakako nosi limitacije budući da učenici dobivaju „generičke" metode s tekstovima koji su ponekad nerazumljivi ili loše formulirani. U ovom radu željeli smo se, između ostalog posvetiti i istraživanju metodičkog aspekta ovakvih aplikacija i primjerenosti za korištenje kako za učenike, tako i za nastavnike.

II. METODOLOGIJA

Kako bismo testirali kvalitetu aplikacije, sastavili smo set od 80 pitanja (zadataka) podijeljenih u 4 dijela (prva dva dijela za osnovnu školu te treći i četvrti za tehničke škole i gimnazije). S obzirom na to da većina aplikacija ne podržava iščitavanje i rješavanje geometrije, Popis matematičkih zadataka korištenih za provjeru aplikacija može se pronaći na poveznici http://bit.ly/mathappcomp. Na Slici 2. je nekoliko primjera pitanja korištenih u testu.

Zbog razlika u načinu zapisa i algoritmima rješavanja u svijetu, OCR aplikacije nije uvijek dobro mjerilo

kvalitete (prepoznavanje uvelike ovisi o tome gdje je aplikacija kreirana), ali dobar OCR koji prepoznaje više zapisa predstavlja prednost u odnosu na tekstualni unos sadržaja. U aplikacijama koje ne podržavaju OCR je svaki zadatak unesen tipkovnicom. Također, unos nekih zadataka podrazumijevao je prilagodbu zbog načina zapisa izraza unutar aplikacije. Primjerice, zadatak GCD(24,18) su neke aplikacije znale riješiti (*greatest common divisor* – najveći zajednički djelitelj), no nisu znale iščitati taj zahtjev iz unesenog izraza, nego se onda izraz morao zapisati nekako drugačije kako bi ga aplikacija prepoznala.

Prilikom ispitivanja rješava li određena aplikacija zadatak, promatrali smo dva segmenta:
1) daje li aplikacija rješenje

2) daje li aplikacija matematički postupak "korak po korak"

Naglasak smo stavili na drugi segment jer bolja aplikacija može dati odgovor na većinu unesenih pitanja, imala ona OCR ili ne.

U Hrvatskoj je provedeno istraživanje, za nastavni predmet matematiku, kod učenika nižih razreda osnovne škole, koje proučava upotrebu tehnologije u nastavi matematike. Prema radu T.Pavičić - Zajec, „obzirom na dobivene rezultate, ne može se sa sigurnošću zaključiti da učenici koji koriste tehnologiju u nastavi matematike postižu bolji uspjeh na ispitima znanja od učenika koji rade na klasičan način (bez tehnologije).".

III. ISTRAŽIVANJE

A. Odabir aplikacija

Osnovni kriterij odabira aplikacija bio je strojni način rješavanja pa su zbog toga otpale neke druge popularne aplikacije za rješavanje matematičkih zadataka, a u kojima ljudi rješavaju zadatke, kao što su Chegg, Studydaddy, ili Slader. Drugi kriterij odabira bila je popularnost aplikacije, mjerena ocjenama korisnika. Sve osim jedne aplikacije su se mogle besplatno preuzeti iz trgovine pa su testirane korištenjem mobilnog uređaja s operacijskim sustavom Android. Preuzimanje mobilne aplikacije Wolframalpha se naplaćuje pa je testiranje provedeno na web-stranici, gdje je dostupna besplatna verzija. Program Wolframalpha testiran je zbog svoje izuzetne širine pokrivenosti matematičkog gradiva, i rasprostranjenog korištenja.

B. Rezultati

Iz Tablice 1. je vidljivo da je najveći broj bodova ostvarila aplikacija WolframAlpha. Kao što je već ranije napomenuto, mobilna verzija programa WolframAlpha se plaća i sadrži OCR, dok besplatna verzija za mobilne uređaje ne postoji. Zbog toga nismo uistinu mogli provjeriti ni kvalitetu OCR-a, niti smo mogli vidjeti korake rješavanja.

Prema uspješnosti, više od 85% riješenih zadataka imaju još samo 3 aplikacije: Mathway, Photomath i Symbolab. Iako su razlike u bodovima male, razlike u kvaliteti sadržaja postoje i bit će opisane kasnije.

Aplikacije su u prosjeku najbolje rješavale gradivo 7. i 8. razreda osnovne škole, dok je najslabiji rezultat postignut za zadatke završnih razreda srednje škole.

C. Analiza rezultata

Najbolji rezultati se javljaju u 7. i 8. razredu zato što u nastavno gradivo tada ulaze jednadžbe, nejednadžbe, sustavi, te algebarske operacije s realnim brojevima. Većina ovog gradiva korespondira s gradivom Algebre 1 (osnove algebre) u SAD-u. Algebra 1 obavezan je srednjoškolski predmet u svim školama u Americi i predstavlja najširu skupinu korisnika koji trebaju pomoć u rješavanju matematičkih zadataka. Najslabiji rezultat većina je aplikacija ostvarila na gradivu trećeg i četvrtog razreda srednje škole. Najvjerojatniji uzrok, prema našem mišljenju, je razlika u gradivu u hrvatskim i američkim školama u završnim razredima srednje škole. U hrvatskom školstvu, u trećem razredu srednje škole, izuzetno je zastupljena trigonometrija, od brojevne kružnice do raznih tipova trigonometrijskih jednadžbi i nejednadžbi. Ovo gradivo je u Algebri 2 (naprednoj algebri) u Americi zastupljeno u puno manjoj mjeri. Također, gradivo 4. razreda srednje škole (derivacije i integrali) je u Americi zastupljeno najčešće u kolegiju Calculus koji se većinom drži na fakultetskoj razini.

D. Prednosti i nedostaci

Svaka aplikacija ima svoje prednosti i nedostatke i više je ili manje prilagođena hrvatskom obrazovnom sustavu, a ovdje smo analizirali glavne karakteristike svake od testiranih aplikacija.

Automath ima vrlo loš OCR, okvir za fotografiranje/skeniranje je mutan i u većini slučajeva se ne vidi što se skenira. Teško prepoznaje već i osnovne računske operacije. Opcija ručnog unosa je mnogo kvalitetnija, a glavni joj je nedostatak što se svaki zadatak mora unijeti kao naredba (opis naredbi je većinom nejasan) te što sva rješenja prikazuje samo u decimalnom zapisu. Osim toga, ovo je jedina aplikacija unutar koje su se pojavljivale reklame. Posebnost aplikacije je *Ask a tutor* opcija, a unutar koje možete za 8 kn zatražiti rješenje zadatka, a za 43 kn i korake.

Chegg math solver ima nešto bolji OCR, a problemi se javljaju većinom zbog osvjetljenja. Prikaz koraka rješavanja se naplaćuje, bez mogućnosti besplatnog isprobavanja aplikacije (*free trial*). Nakon skeniranja nudi nekoliko opcija obrade izraza (npr. rješavanje, deriviranje

Aplikacija	OCR	Rješenje	Prikaz koraka	Broj bodova
Automath	DA	DA	Besplatno	36
Chegg Math Solver	DA	DA	Uz plaćanje	56
Cymath	DA	DA	Besplatno	60
Math Solver	DA	DA	Besplatno	37
Mathway	DA	DA	Uz plaćanje	71
Photo Mathematics	DA	DA	Besplatno	48
Photomath	DA	DA	Besplatno	74
Symbolab	DA	DA	Uz plaćanje	72
WolframAlpha	DA	DA	Uz plaćanje	78

Tablica 1. bodovi po zadacima

i sl.) od kojih su mnogi generički i nisu korisni (primjerice, integiranje numeričkog izraza).

Za razliku od prošle aplikacije, Cymath korake rješavanja nudi besplatno. Osnovnoškolsko gradivo je dobro metodički razrađeno po koracima, uz detaljna objašnjenja. OCR ima problema s prepoznavanjem potencija, a složenije srednjoškolske zadatke kao što su, primjerice, eksponencijalno jednadžbe i nejednadžbe, niti ne rješava pa je diskutabilna korisnost ove aplikacije u srednjoškolskoj nastavi uopće.

Microsoft Math Solver, osim opcije unosa i OCR-a, nudi i opciju unosa zadatka pisanjem po zaslonu. Ta je opcija vrlo praktična za korištenje na tabletima i ostalim uređajima s većim ekranima osjetljivima na dodir. Pogotovo ako uzmemo u obzir da svaki učenik petog razreda osnovne škole u Hrvatskoj u školi dobiva tablet na korištenje. Prepreka masovnom korištenju u školama je što aplikacija nema prijevod na hrvatski jezik. Od ostalih funkcionalnosti, valja spomenuti da aplikacija ispod rješenja nudi poveznice na slične sadržaje na internetu. Iako je aplikacija riješila gotovo sve zadatke, za većinu srednjoškolskih zadataka ne nudi korake rješavanja pa je ukupan broj bodova značajno manji, nego bi bio da smo razmatrali samo rješivost zadataka.

Mathway je još jedna od testiranih aplikacija kojoj nismo mogli analizirati metodičnost zbog naplaćivanja prikaza koraka rješavanja. OCR odlično prepoznaje pitanja i aplikacija daje točna rješenja. Za fotografirani izraz često nudi i dodatne opcije obrade, a koje u nekim slučajevima nisu korisne, primjerice integriranje algebarskih izraza.

Photo mathematics je od svih testiranih aplikacija imao najlošiji OCR. Primjerice, često miješa simbole za zbrajanje i množenje, umjesto zapisane varijable očita neku drugu, ne može prepoznati već niti eksponente i slično. Aplikacija je djelomično prevedena na hrvatski jezik, a koraci rješavanja su besplatni. Iako su besplatni, koraci najčešće nemaju nikakva dodatna objašnjenja uz sebe i prijelaz između koraka nije uvijek sasvim jasan. Aplikacija bi mogla biti vrlo zbunjujuća za učenike jer nudi rješenje čak i kada ne zna riješiti zadatak i to na način da samo promijeni redoslijed pribrojnika ili nekih drugih članova izraza.

Photomath je aplikacija koja se pokazala kao najkorisnija, ne samo zato što je postigla odličan rezultat na testu, nego i zato što korake rješavanja zadataka možete vidjeti besplatno. Radi i bez pristupa internetu što je također prednost pri primjeni u nastavi. OCR je odličan i sve što se može ručno unijeti unutar aplikacije i prepoznaje. Aplikacija je u potpunosti prevedena na hrvatski jezik, što je za pohvalu, ali se neki od prijevoda ne koriste u nastavi. Primjerice, u koracima rješavanja jednadžbe sugerira da se na drugu stranu prebace „poznanice“, umjesto brojevi ili konstante. Aplikacija je jedinstvena po tome što su neki koraci rješavanja čak i animirani, primjerice pisano množenje.

Symbolab je još jedna od aplikacija kojoj nismo mogli procijeniti metodičnost zbog naplate prikaza koraka rješavanja. OCR je vrlo loš i najčešće se zadaci moraju ručno unosit. Aplikacija uopće ne prepoznaje oznaku za stupanj pa nijedan zadatak vezan uz kutove nismo mogli testirati, primjerice trigonometriju.

WolframAlpha je aplikacija koja nije striktno matematička nego se upisuju nekakva pitanja ili izrazi i aplikacija daje odgovor na njih. Budući da se već i instalacija aplikacije naplaćuje, testirali smo njezinu web verziju koja je besplatna. Matematičke zadatke rješava bez ikakvog problema, a prikaz koraka rješavanja se naplaćuje pa niti ovdje nismo mogli testirati metodičnost. Za učenike može biti otežan proces unosa zadatka/pitanja jer se oni unose u jednom redu, što za složenije izraze podrazumijeva mnoštvo zagrada. Osim toga, djeluje zbunjujuće i prikaz rješenja jer nije ponuđeno što biste vi željeli napraviti s nekim izrazom, nego sama odmah obradi i ponudi sva rješenja koja zna.

IV. ZAKLJUČAK

Većina matematičkih aplikacija za rješavanje zadataka u današnje vrijeme posjeduje OCR tehnologiju koja pretvara fotografiju u tekst te potom u matematički izraz koji se automatizirano rješava korištenjem neke vrste programskog alata (*software engine*). Poslovni model većine aplikacija podrazumijeva plaćanje za prikaz koraka, a što bi moglo predstavljati problem za prosječnog korisnika koji nije u mogućnosti procijeniti odgovara li aplikacija njegovim potrebama prije nego li ju mora platiti, ali i zbog financijskog aspekta jer aplikacije u velikoj mjeri koriste i učenici.

Od istraženih aplikacija, WolframAlpha je dobio najvišu ocjenu s obzirom na pokrivenost sadržaja, no veliki je nedostatak činjenica da se mobilna aplikacija naplaćuje, i to što nije mogao biti provjeren OCR aplikacije. Uzimajući u obzir sve parametre (broj metoda, dostupnost koraka rješenja (onda i metodičnost) te složenost korištenja), za korištenje u nastavi najprikladnija aplikacija je Photomath.

Istraživanje je bilo usmjereno na gradivo koje se obrađuje u nastavi matematike u hrvatskim školama, koje ni za koju aplikaciju ne predstavljaju primarno tržište, pa je lošiji rezultat većine aplikacija u nekim područjima (npr. trigonometrija) očekivan. Aplikacije i njihovi ostvareni bodovi mogli bi se ubrzo promijeniti s obzirom na to da se na tržištu pojavljuje sve više aplikacija koje nude automatizirano rješavanje matematičkih problema. Korisnost takvih aplikacija za razumijevanje matematičkog gradiva je vrlo značajna. Kurikularna reforma oprema učenike tabletima za korištenje u nastavi te je to idealna prilika za integriranje ovakvih aplikacija u nastavu, pogotovo onih koje detaljno prikazuju korake rješavanja. Nastavnici mogu, primjerice, uputiti učenike da provjeravaju svoje metode (i korake) rješavanja, proučavati svojstva funkcija bez potrebe rješavanja na ploči, generalizirati nepotpunom indukcijom na nizu rješenih primjera, lako prelaziti iz jednog načina zapisa u drugi i slično. No, tu je velika odgovornost na nastavnicima da upoznaju učenike s načinom funkcioniranja aplikacija, ali i sa svim prednostima i manama ovakvih aplikacija. S pojavom ovakvih aplikacija se naravno javlja i potencijalni problem zbog moguće zloupotrebe na ispitima znanja, a što je česta tema diskusije među nastavnicima.

Neovisno o percepciji takvih aplikacija, jasno je da će u budućnosti one predstavljati važan oblik učenja, što povlači nužnu provjeru kvalitete sadržaja takvih aplikacija kako bismo osigurali da učenici u budućnosti usvajaju nastavne sadržaje na primjeren način.

LITERATURA

[1] T.Pavičić-Zajec, "Učinkovitost korištenja informacijsko komunikacijske tehnologije u nastavi matematike nižih razreda osnovne" doktorski rad

[2] https://www.forbes.com/sites/mikemontgomery/2019/09/03/this-math-edtech-says-its-found-a-winning-formula-with-on-demand-tutoring/#692ea751716d (preuzeto 21.1.2020.)

[3] https://narodne-novine.nn.hr/clanci/sluzbeni/2019_01_7_146.htmlI (preuzeto 21.1.2020.)

[4] http://www.corestandards.org/Math/Content/HSA/introduction/ (preuzeto 21.1.2020.)

[5] http://www.corestandards.org/Math/Content/HSN/introduction/ (preuzeto 21.1.2020.)

[6] http://www.corestandards.org/Math/Content/HSF/introduction/ (preuzeto: 21.1.2020.)

[7] https://en.wikipedia.org/wiki/Long_division (preuzeto:21.1.2020.)

Micro:bit-na biometrija

Branko Raičković, IV. osnovna škola Bjelovar, Bjelovar, Hrvatska, Bernarda Musović, IV. osnovna škola Bjelovar, Bjelovar, Hrvatska, brarai@gmail.com

***Sažetak* - Ovaj stručni rad se temelji na korištenju mikroračunala "micro:bit-a", kao mjernog instrumenta za mjerenje visine stabla (objekta).**

Za indirektno mjerenje visine objekta koristi se uređaj klinometar. Za izračun visine potrebno je izmjeriti inklinaciju, a zatim izračunati visinu pomoću udaljenosti od objekta i tangensa izmjerenog kuta.

Micro:bit je malo računalo veličine kutije šibica. U sebi sadrži mnogobrojne senzore koji se mogu iskoristiti na različite načine, serijsku i radio komunikaciju te displej za prikaz rezultata. Kako micro:bit sadrži senzor žiroskopa moguća je izračunati visinu stabla pomoću kuta elevacije. Micro:bit možemo u potpunosti iskoristiti za ovakav način mjerenja jer je moguće rezultate prikazati na displeju. Uz računalo je potrebno napisati softversku aplikaciju te prilagoditi mjerne jedinice.

Elektronički klinometar je izrađen za potrebe već poznatog projekta „Drvokod“ u kojemu ga za terenski rad koriste učenici za određivanje visine drveta. Kako su u projektu sve vrste drveća opisane i određene, potrebno je bilo točno odrediti visinu svakog stabla. Upravo zbog toga je ušao u upotrebu ovaj digitalni klinometar.

Ključne riječi – micro:bit; klinometar; mjerenje visine drveta; biometrija; kutomjer

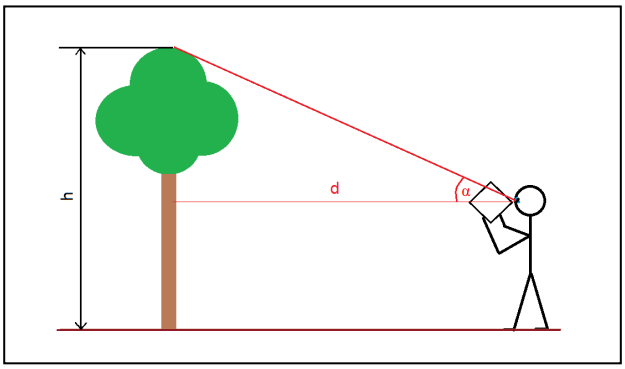
I. UVOD

U sklopu provedbe projekta Drvokod, koji se zasniva na determinaciji i elektroničkom označavanju drveća u našem školskom okolišu, veliki dio nastave se odvija terenski, sa uporabom mobitela, tableta, i ostalih digitalnih uređaja.

Učenici u sklopu terenske nastave u školskom dvorištu mjere temperaturu zraka i tla, određuju dominantne i kodominantne vrste drveća, promatraju promjene boje listova i izrađuju zbirke listova, mjere prsni promjer stabla i visinu stabla.

Budući da su učenici imenovali sve vrste drveća, želja je bila izmjeriti točnu visinu svakog drveta.

Visina stabla se dosada određivala na nekoliko načina. Jedan od načina je bila usporedba dva jednakostranična trokuta. Prvi trokut činila je udaljenost između stabla i sjene, a drugi trokut udaljenost između mjerača i njegove sjene. Kada se znaju omjeri $a_1:a_2=b_1:b_2$ može se izračunati visina stabla.



Slika 1. Određivanje visine stabla mjerenjem kuta elevacije od razine oka do vrha stabla

Za drugi način mjerenja visine može se koristiti klinometrom. (Slika 1.) To je naprava koju je lako konstruirati, a upotrebljava se za mjerenje visine objekata. To je pojednostavljena verzija kvadranta, važnog srednjevjekovnog instrumenta i sekstanta, instrumenta za određivanje položaja brodova. Svaka od ovih naprava ima lukove koji su podijeljeni na stupnjeve za mjerenje kuta elevacije (uzdizanja). Luk klinometra označen je od 0 do 90 stupnjeva. Kada se objekt vidi kroz ciljnik, broj stupnjeva u kutu elevacije može se očitati s luka . Kut elevacije α je kut između zamišljene horizontale povučene od oka mjerača do objekta mjerenja i zamišljene linije koja oko mjerača spaja s vrhom objekta.

Radi se o kutomjeru i malim viskom u vrhu pravog kuta.

Visina drveta = izračunata visina + visina očiju mjerača

Mjerač promatra vrh stabla držeći klinometar u visinu očiju. Konac koji visi, određuje kut od razine oka do visine stabla tj. kut elevacije. Na taj način se očita veličina kuta i odredi se tangens.

Na slici je prikazano načelo uporabe klinometra za određivanje kuta elevacije od razine oka do vrha stabla. Trigonometrijski se određuje onaj dio visine stabla iznad zamišljene horizontale od oka mjerača. (Slika 1.)

Visina = tangens kuta α × udaljenost mjerača od baze stabla

No, da bi se odredila konačna visina stabla, visini moramo pribrojiti visinu mjeračevog oka.

Postojećim načinima mjerenja ponekad možemo naići na poteškoće. Neko drveće može biti preblizu ili predaleko. Može se naići na prepreku između drveća. Tada nije moguće izvršiti točno mjerenje visine.

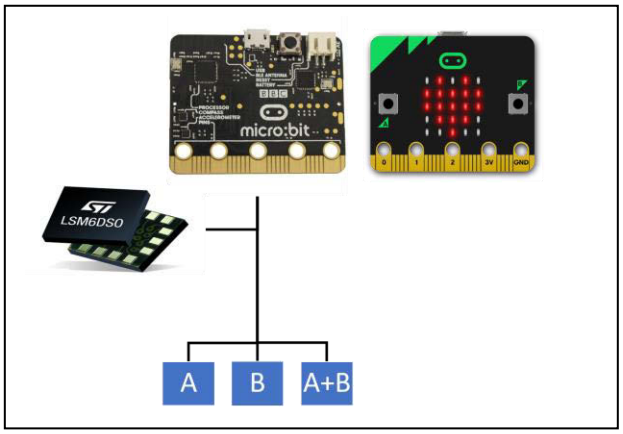
Zbog svih tih poteškoća sinula je ideja o pokušaju korištenja micro:bit-a kao uređaja za mjerenje visine stabla, čime bi olakšati rad na projektu.

II. MICRO:BIT

A. *Hardverski dio*

Hardver se zasniva na korištenju micro:bit-a, malog računala - kontroler sa integriranim senzorima. Micro:bit se također može povezivati sa vanjskim senzorima na različite načine kao što su serijska, ISP, 1-wire, micro-BUS te I^2C te BlueTooth komunikacijom.

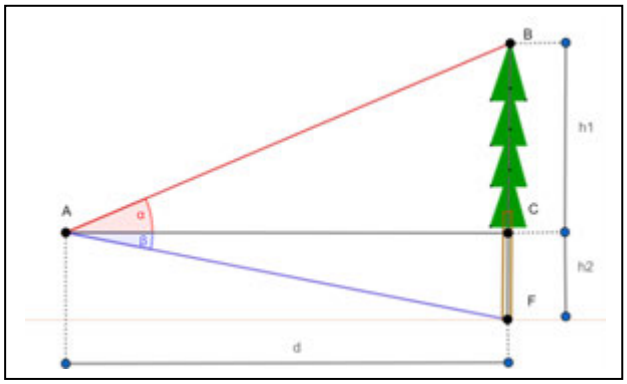
Kako je bila potreba napraviti uređaj koji će očitati kut elevacije između mjerača i vrha drveta, korišten je unutarnji senzor sa žiroskopom namijenjen očitavanju x, y i z-osi.



Slika 2. dijagram spajanja komponenti

B. *Matematički dio*

Korištene su trigonometrijske jednadžbe za matematički izračun. Za izračunavanje duljine nasuprotne katete pravokutnog trokuta (h_1 – visina drveta umanjena za visinu mjerača) pomoću kuta α. Zatim je trebalo izračunati duljinu priležeće katete (d – duljinu od drveta do mjerača) uz pomoć kuta β i duljine nasuprotne katete (h_2 – visina mjerača). Visina drveta je zbroj $h_1 + h_2$.



Slika 3. geometrija mjerenja

Poznate vrijednosti su: h_2 – visina mjerača, kut α – kut između horizontale i vrha stabla i kut β – kut između horizontale i baze stabla. Nepoznanica je h – ukupna visina.

$$h_1 = \tan(\alpha) \times d \quad (1)$$

$$d = \cot(\beta) \times h_2 \quad (2)$$

$$h_1 = \tan(\alpha) \times \cot(\beta) \times h_2 \quad (3)$$

$$h = h_1 + h_2 \quad (4)$$

C. *Softverski dio*

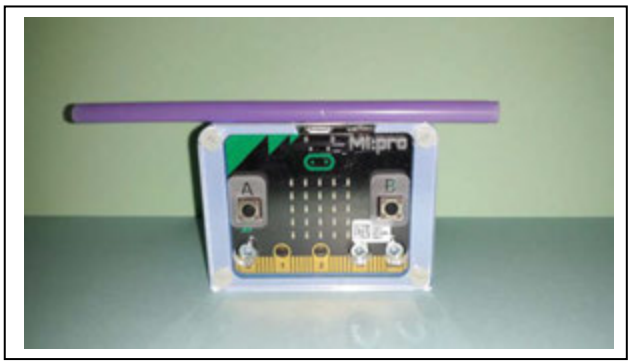
Softverski dio je baziran na povezivanju čipa s micro:bit-om. Prije svega trebalo je predinstalirati upravljački program za povezivanje čipa pomoću I^2C sabirnice te odrediti hexadecimalne adrese za inicijalizaciju i iščitavanje podataka.

Zatim je potrebno pronaći tangens kuta α i kotangens kuta β. Podatci se dalje obrađuju matematički po gore navedenim jednadžbama.

Program je podijeljen na tri dijela. Prvi dio očitava kut elevacije (pritisnuti tipku A) za visinu drveta (1), drugi dio kut za određivanje udaljenosti (2) do drveta (pritisnuti tipku B) i treći za izračun ukupne visine (pritisnuti tipke A i B zajedno (4). Slijedi prikaz podataka na displeju micro:bit-a.

Program je kodiran u JavaScript-u i Micropython-u, sa proširenjem za priloženi senzor.

D. Praktički dio



Slika 4. fotografija uređaja

Cijeli uređaj je smješten u plastično kućište zajedno s litij baterijom.

S gornje strane kućišta pričvršćena je plastična cjevčica koja nam služi za ciljanje mjerne točke.

Na slici 4. vidljiv je displej te tipke A i B.

E. Mjerenje

Klinometar se postavlja vodoravno, u visini očiju te definira se točka C (slika 3.). Gledajući kroz cjevčici nacilja se vrh drveta - točka B (slika 3.) i pritisne tipka A na uređaju, zatim se nacilja baza drveta - točka F (slika 3.) i pritisne tipka B.

Nakon pritiska tipke A i tipke B skupa, na displej uređaja se prikaže ukupna visina mjerenog objekta.

F. Nadogradnja

Kao i svaki uređaj, tako i ovaj već ima mogućnost nadogradnje. Da bi se smanjile pogreške pri mjerenju predviđeno je višestruko mjerenje te izračunavanje srednje vrijednosti. Kako micro:bit podržava radio komunikaciju pomoću BlueTooth-a, može se upariti s drugim uređajem na kojem bi se spremali podatci ili slali direktno na Internet (mobitel, laptop, tablet ili neko drugo mikroračunalo). Za bolje ciljanje moguće je uz cjevčicu koristiti i laserski pokazivač

III. ZAKLJUČAK

Ovakvim načinom rada učenike se potiče na aktivno uključivanje u istraživanje i otkrivanje novih nastavnih sadržaja. Razvijaju se i kompetencije uporabe različitih tehnologija.

Optičkim napravama, alatima i sličnim uređajima, stvara se ozračje za inovativno i kreativno rješavanje problema povezanih s ekološkim temama te popularizacija ekologije kroz primjenu informacijsko–komunikacijske tehnologije.

Izrada svakog novog uređaja povlači za sobom bezbroj nepredvidljivih zadataka i problema. Posebno u nabavci i programiranju nestandardne i nedokumentirane opreme.

Unatoč svemu, ovaj uređaj je našao svoje mjesto u našem projektu, a nadamo se i šire.

LITERATURA

[1] G. Halfacree, "BBC MICRO:BIT službeni priručnik," Zagreb, 2018.

Glazba2Go, mobilna aplikacija za promoviranje glazbene kulture

Ivan Britvić, Filip Meštrović
Visoka škola za informacijske tehnologije, Klaićeva 7, Zagreb, Croatia
ivan.britvic@vsite.hr, fimestro@vsite.hr

***Sažetak* - Glazbeni odgoj u osnovnom obrazovanju često je nedovoljan s obzirom da su učenici suočeni sa poplavom sadržaja upitne kvalitete u raznim medijima. Autori su stoga pokušali razvojem mobilne aplikacije Glazba2Go promovirati kvalitetne glazbene sadržaje. Besplatna mobilna glazbena enciklopedija za Android mobilne sustave izrađena je kako bi ponudila, ne samo učenicima osnovnih škola, već korisnicima svih dobnih skupina mogućnost educiranja o glazbi gdje god se našli.**

Ključne riječi – edukacija; mobilna aplikacija; glazba;

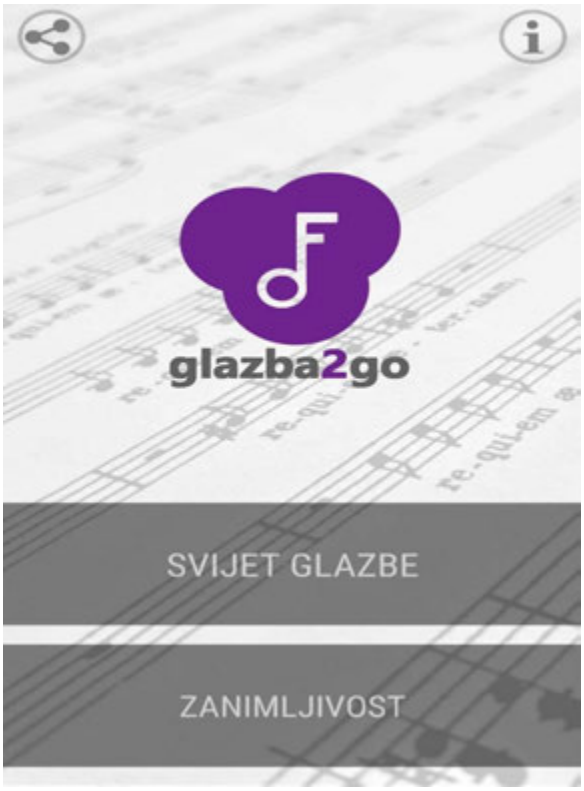
I. UVOD

A. Nastanak ideje

S obzirom na to da je jedan od autora ovog rada, Filip Meštrović, za vrijeme završnih godina na Visokoj školi bio zaposlen kao učitelj Glazbene kulture u Osnovnoj školi Ivana Perkovca u Šenkovcu, tražio je način kako povezati tehnologiju, glazbu i nastavu. Uz to, tendencije Ministarstva znanosti i obrazovanja za uvođenjem i dodatnom integracijom informacijskih i komunikacijskih tehnologija u nastavu, kao i činjenica da se nedovoljno koriste mobilne aplikacije u svrhu edukacije, dodatno su potakle potragu za najboljim načinom uključenja modernih tehnologija u sadržaj nastave što je na kraju rezultiralo javno dostupnom, mobilnom glazbenom enciklopedijom Glazba2Go.

B. Mobilna aplikacija kao medij

Prevladavajući obrasci ponašanja školske djece i mladeži danas uključuju intenzivno korištenje mobilnih uređaja u slobodno vrijeme nerijetko na vrlo štetne i zatupljujuće načine. Autori su u kontaktima s učenicima utvrdili da se jako malo pristupa glazbenim sadržajima iz područja klasične muzike, jazz-a, blues-a, kao i originalnih glazbenih djela razvijenih specifično za potrebe glazbene pratnje filma. S druge strane često pregledavaju sadržaje popularne glazbe kao i sadržaje upitne kvalitete koje nazivaju „narodna muzika“. Ovakav obrazac razmišljanja doveo je do ideje kako bi se korištenjem mobilnih uređaja mogli prezentirati glazbeni uradci koje malo ljudi mlađih generacija pregledava a mogu proširiti znanje o glazbi i razviti glazbeni ukus. Glazba2Go trebala bi svojom interaktivnošću i jednostavnošću doprijeti do mladih umova, ali ne samo mladih već i svih zainteresiranih.



Slika 1 – Glavni izbornik aplikacije

C. *Priprema*

U trenutku osmišljanja ideje za Glazba2Go prof. Meštrović nije bio upoznat s programiranjem za Android mobilne sustave. Unatoč tome, dobri temelji sa Visoke škole za IT i želja za programiranjem mobilnih aplikacija te znatni uloženi trud omogućili su izradu Glazba2Go mobilne glazbene enciklopedije. S obzirom na obrazovanje na Muzičkoj akademiji, dio početne pripreme, što se tiče stručnih materijala u vezi glazbenih žanrova, bio je odrađen. Preostalo je osmisliti vizualni identitet aplikacije, način unosa podataka kao i način prikaza. Nakon toga, projekt Glazba2Go bio je spreman za razvoj.

II. DIZAJN I IZRADA APLIKACIJE

A. *Odabir platforme*

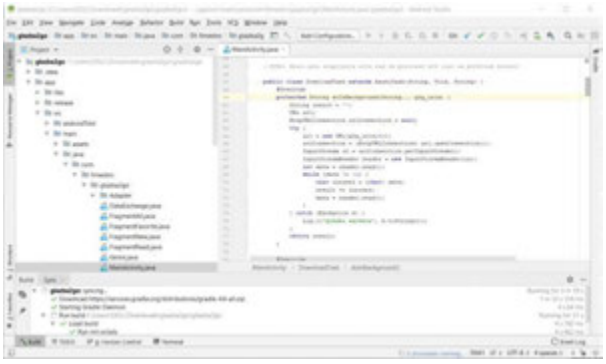
Kako bi se prezentirali i promovirali umjetnički kvalitetni glazbeni sadržaji razmatran je pristup kreiranja specijaliziranog mrežnog mjesta kao i mobilna aplikacija. Oba pristupa imaju vlastite prednosti, mrežno mjesto je šire dostupno na svim platformama ali ga je lako previdjeti u ogromnom mnoštvu sadržaja, optimizacija za tražilice kao i promocija na društvenim mrežama zahtijevaju značajne ljudske i financijske resurse. Mobilna aplikacija zahtijeva korisničku instalaciju na vlastiti uređaj čemu su neki korisnici neskloni ali kada se nadvlada ova barijera, aplikacija omogućuje kvalitetnije korisničko iskustvo na individualnom mobilnom uređaju korisnika.

Dodatna motivacija bili su provedeni razgovori s učenicima osnovne škole u kojima su se jednoglasno opredijelili za odabrani pristup.

Kada je donesena odluka o izradi mobilne aplikacije logični korak bio je odabir Android platforme. Kako i prema konzervativnim procjenama istraživanja tržišta iz 2019. godine udio Android platforme kao operativnog sustava na mobilnim uređajima prelazi 70% korisnika [1] i ima trend povećanja autori su odlučili da razvoj aplikacije na ovoj platformi ima prioritet. Nakon što je aplikacija razvijena i objavljena na platformi za preuzimanje aplikacija *Google Play* trgovini planira se dalje postavljanje *Apple App Store* i *Huawei AppGallery* trgovine.

B. *Osnovne komponente aplikacije*

Mobilna aplikacija Glazba2Go razvijena je u programskom jeziku Java unutar Android studio okruženja. Sama aplikacija namijenjena je prikazu i istraživanju glazbenih žanrova a kako bi isti bili dostupni potrebno je bilo razviti i *backend*, poslužiteljski kod koji omogućuje unos žanrova i preko specijaliziranih servisa priprema ove podatke mobilnoj aplikaciji za korištenje. Poslužiteljski dio sadrži i web sučelje za unos podataka implementirano korištenjem HTML i CSS jezika za definiciju web stranica kao i *Bootstrap* okruženja za responzivni prikaz te PHP poslužiteljskog skripta. Nakon unošenja novog žanra, korištenjem alata *Push notification* iz *Google*-ovog paketa *Firebase*, ista se trenutno prezentira korisniku mobilne aplikacije putem nove obavijesti koja nakon klika otvara aplikaciju. Svi novi i neotvoreni žanrovi su označeni s *Novo!* tako da se korisnik motivira da pregleda što više različitih žanrova kako bi razvijao svoje poznavanje glazbene kulture.



Slika 2 – Android studio zaslon

C. *Dizajn mobilne aplikacije*

Mobilna aplikacija izrađena je koristeći primjere najbolje prakse u Android ekosustavu. Osnovne blokove aplikacije čine tzv. *aktivnosti* (definicije korisničkog sučelja na pojedinačnom zaslonu sa pripadajućom logikom). Aktivnosti u aplikaciji su:

- *SplashActivity* - početni zaslon koji služi kao uvod u aplikaciju, sadrži sliku logotipa s pripadajućom animacijom koja postupno prikazuje logotip (tzv. *fade-in*) i translatira ga po Y osi.
- *MainActivity* – glavni zaslon aplikacije s pripadajućim izbornikom gdje korisnik odabire koju radnju želi pokrenuti: čitanje zanimljivosti, čitanje o glazbenim žanrovima i slušanje glazbenih primjera pojedinog žanra i dijeljenje aplikacije putem drugih aplikacija. Aktivnost asinkrono dohvaća trenutne zanimljivosti iz svijeta glazbe i prezentira ih korisniku.

- *SvijetGlazbeActivity* - prikazivanje različitih načina grupiranja glazbenih žanrova. Žanrove je moguće grupirati prema pročitanima, nepročitanima, novim žanrovima te prikaz svih žanrova. Aktivnost se koristeći asinkroni model pristupa, povezuje s logikom na poslužitelju i preuzima popis žanrova.

Kako bi se korisnicima omogućilo preslušavanje odabranih glazbenih zapisa koji najbolje predstavljaju pojedine žanrove upogonjena je *YouTube Android Player API* biblioteka [2]. Kako bi se ovaj dio odvojio od ostalog koda aktivnosti korištena je Android fragment arhitektura.[3]. S takvom modularnom arhitekturom olakšava se održavanje aplikacije i poboljšava iskoristivost koda. Aplikacija pamti koji žanrovi su otvarani i koji su favoriti korisnika te informacije sprema u vlastitoj bazi podataka koja je radi jednostavnosti implementirana korištenjem *TinyDb* knjižnice koja koristi ugrađeni Android mehanizam *SharedPreferences*.

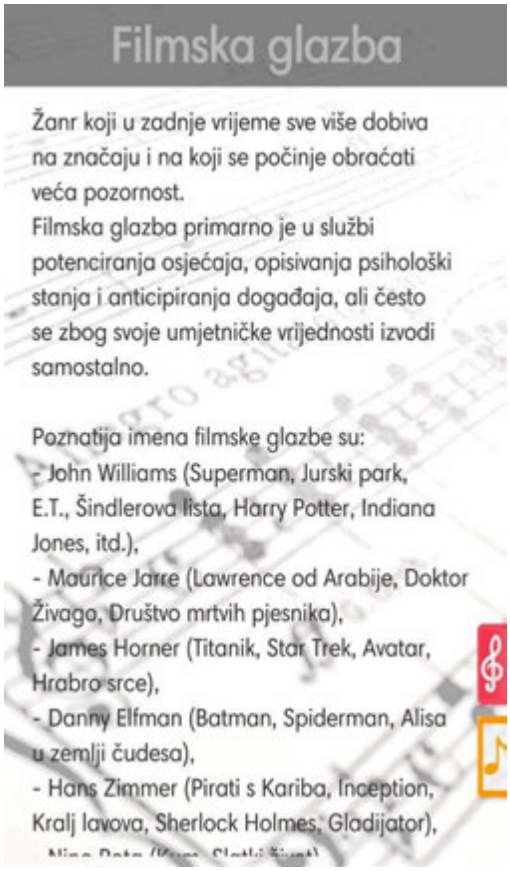
D. *Planirana poboljšanja*

Na popisu budućih funkcionalnih proširenja Glazba2Go mobilne glazbene enciklopedije nalaze se ideje poput mini-igre, bogatije mogućnosti dijeljenja sadržaja i sl. Dakako, idejna okosnica svega je glazbena edukacija, s ciljem da bude prikladno korisnicima svih dobnih skupina. Shodno tome, najvažnije nadolazeće poboljšanje dolazi u videu dodatka o instrumentariju svakog pojedinog žanra. Na taj način korisnici mogu vizualno i auditivno povezati specifične instrumente koji se koriste prilikom izvođenja skladbi svakoga žanra. U vrijeme pisanja ovoga članka, nova verzija aplikacije je u razvoju te će uskoro biti objavljena u *Google Play* trgovini. S obzirom na to da je trenutno ograničeno korištenje na Android operacijske sustave, u tijeku je izrada inačice aplikacije za iOS operacijske sustave. Time bi i korisnici Appleovih proizvoda imali mogućnost korištenja aplikacije.

Prateći tehnološke trendove, ali u isto vrijeme uzimajući u obzir objektivnu potrebe, u razvoju je potpuno nova tehnološka arhitektura Glazba2Go mobilne aplikacije, kao i poslužiteljskog *backend*-a. Nova verzija mobilne aplikacije razvija se koristeći Googleov komplet za razvoj softvera *Flutter* i *Dart* programski jezik, administratorsko sučelje u React.js-u. Poslužiteljska strana sustava prilagođuje se kako bi koristila *Node.js* te *MongoDB* kao bazu podataka.

III. OBJAVA I PRIJEM APLIKACIJE

Aplikacija Glazba2Go objavljena je na *Google Play* trgovini u travnju 2018. godine.[4] S obzirom na minimalnu reklamu i ideju da se jedina prezentacija i promocija korištenja aplikacije odvija u nastavi autora, statistika je vrlo pozitivna i poticajna za daljnji rad. Aplikacija broji otprilike 200 instalacija, od kojih je trenutno aktivno 30-ak. Dodatnim poboljšanjima i sadržajima očekuje se povećanje broja instalacija i veća uporaba u nastavi.



Slika 3 – Pregled glazbenog žanra

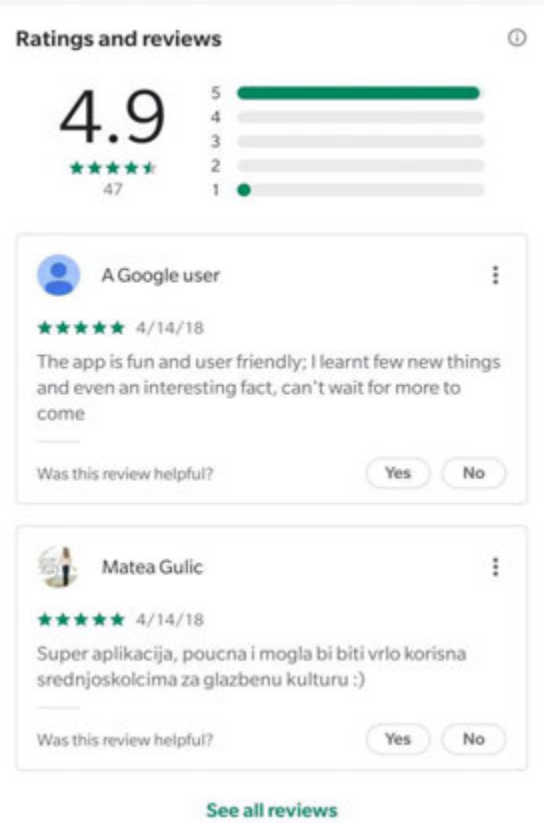
A. *Google Play* trgovina

Postupak objave na *Google Play* trgovini zapravo je vrlo jednostavan. Sučelje je intuitivno, no za samu objavu bilo je potrebno obaviti nekoliko koraka i pripremiti sve materijale potrebne za objavu (instalacijska datoteka, logotipi različitih veličina, tekstovi i sl.). Dodatno, bilo je potrebno ispuniti i minimalne uvjete što se tiče zaštite privatnosti korisnika. To se izvodi u obliku izjave o privatnosti kojom se jamči zaštita osobnih podataka

korisnika koji se koriste u svrhu analize stabilnosti aplikacije i zadovoljstva korisnika korištenjem iste. Nedugo nakon toga Glazba2Go bila je dostupna na *Google Play* trgovini.

B. Recenzije i komentari korisnika

Komentari korisnika vrlo su pozitivni (trenutno aplikacija ima ocjenu 4.9 na ukupno 47 recenzija). Mišljenja korisnika izrazito su pozitivna iako se mogu naći i interesantne primjedbe koje su poslužile za razvijanje ideja za daljnju doradu aplikacije. S obzirom na to da se baza glazbenih žanrova u aplikaciji povećava, neki korisnici čak i predlažu koje žanrovi bi voljeli vidjeti u aplikaciji. Ovo je još jedan primjer koji daje do znanja kako korisnici imaju želju za učenjem, možda ne toliko za onim klasičnim, namijenjenim samo đacima i studentskoj populaciji, već cjeloživotnim.



Slika 4 – Recenzije i komentari korisnika

IV. Zaključak

Interes koji je Glazba2Go dobila pokazao je da edukativne aplikacije ove vrste imaju potencijal na tržištu. Potrebno je prije upuštanja u pothvat izrade edukativnih mobilnih sadržaja istražiti ciljane skupine korisnika te svakako pratiti smjernice stručnjaka za što bolje korisničko iskustvo. Izrada mobilnih aplikacija nije trivijalan proces, no uz uloženi trud i određen napor mnoge korisne ideje mogu biti kvalitetno prezentirane ciljanim skupinama što svojim primjerom dokazuje i prijem Glazba2Go aplikacije.

LITERATURA

[1] https://www.idc.com/promo/smartphone-market-share/os - pristupljeno 28.04.2020

[2] https://developer.android.com/guide/components/fragments- pristupljeno 27.06.2019

[3] https://developers.google.com/youtube/android/player/ - pristupljeno 27.06.2019

[4]https://play.google.com/store/apps/details?id=com.fimestro.glazba2go - pristupljeno 27.06.2019

Mobilne aplikacije u obrazovnom okruženju

Jozica Pažanin*, Marko Rosić **
* Ekonomska i upravna škola, Split, Hrvatska, jozica.pazanin@skole.hr
** Prirodoslovno – matematički fakultet / Informatika, Split, Hrvatska, marko.rosic@pmfst.hr

Sažetak: **Razvoj tehnologije značajno utječe na promjene u obrazovanju koje se sve više odmiče od tradicionalnog pristupa poučavanja. Suvremeno obrazovanje koristi tehnologiju da bi učenika postavilo u središte nastavnog procesa te mu omogućilo personalizirano poučavanje. U ovom radu se razmatra uloga mobilnih tehnologija u suvremenom obrazovanju.**

Cilj rada je prikazati pregled istraživačke literature o integraciji mobilnih tehnologija u obrazovno okruženje te analizirati upotrebu mobilnih aplikacija u obrazovanju. Na temelju istraživačkih studija koje proučavaju rezultate aplikativnog učenja ili ocjenjivanja napravljena je analiza implementiranih aplikacija u nastavnom procesu po dobi sudionika, području učenja, namjeni aplikacije te podržanoj tehnologiji. Također, u radu se ispituje i utjecaj upotrebe aplikacije na učinak učenja, motivaciju i stavove učenika. Integracija mobilnih tehnologija u obrazovno okruženje pruža nove poboljšane mogućnosti učenja kao što su personalizacija, interaktivnost, prilagodljivost. Istraživanja pokazuju da upotreba aplikacija u nastavnom procesu pospješuje ostvarivanje obrazovnih ciljeva te povećava motivaciju i zainteresiranost učenika za učenjem. Nove tehnologije poput proširene stvarnosti predstavljaju izvrsnu platformu za prikazivanje nastavnih sadržaja, čineći sadržaj učenja zanimljivijim a učenike pažljivijim prilikom učenja. Osim za prikazivanje nastavnih sadržaja, aplikacije se koriste za ponavljanje te provjeravanje ostvarenosti ishoda učenja.

Ključne riječi: aplikacije za učenje, mobilne aplikacije, proširena stvarnost u obrazovanju, BYOD, m-učenje, obrazovne tehnologije

I. UVOD

Promjene uzrokovane razvojem tehnologije značajno utječu na obrazovanje. Suvremeno obrazovanje, tzv. obrazovanje 3.0 podrazumijeva osobno orijentirano obrazovanje utemeljeno na web tehnologijama [1]. U obrazovanju 3.0 u kojem je virtualno okruženje uobičajeno obrazovno okruženje, učenici aktivno koriste mogućnosti mobilnog učenja. Brzim rastom mobilnih tehnologija i opsežnom upotrebom mobilnih uređaja neprestano se povećava broj korisnika mobilnih aplikacija u formalnom i neformalnom obrazovnom okruženju [2]. Mobilne tehnologije pružaju nove i poboljšane mogućnosti učenja, kao što su personalizacija i prilagodljivost, interaktivnost, komunikacija te suradnja među učenicima [3], [4]. Mobilni uređaji smatraju se osobnim digitalnim pomoćnicima i poput prijenosnih računala postaju alati za učenje s velikim potencijalom te se BYOD (engl. Bring Your Own Device) pristup sve više koristi u nastavi.

Neovisno o uređaju na kojem se prikazuju (tablet / pametni telefon), posljednjih godina u obrazovanju sve više koristimo aplikacije u nastavi kako za prikazivanje i usvajanje nastavnih sadržaja tako i za provjeru ostvarenosti ishoda učenja. Istraživanja pokazuju da upotreba aplikacija u nastavi povećava zainteresiranost i motivaciju učenika te bolje pamćenje sadržaja. Upotrebom aplikacija u nastavnom procesu postiže se personalizacija kojom se učenika postavlja u centar nastavnog procesa gdje on sam stvara, uređuje i modificira edukacijski proces. Također, upotrebom aplikativno simuliranih situacija, učenici mogu provjeravati vlastite ideje, stvarati pretpostavke na temelju viđenog te elaborirati informacije do razine potpunog razumijevanja onoga što se istražuje. Na taj način ishodi učenja ne obuhvaćaju samo činjenice i definicije već i iskustva potaknuta stjecanjem određenog znanja.

U ovom radu razmatra se uloga tehnologije u suvremenom obrazovanju i način na koje se mobilne tehnologije integriraju u suvremeno obrazovanje te se ispituje utjecaj upotrebe aplikacija u nastavnom procesu na učinak učenja te motivaciju i stavove učenika.

II. TRADICIONALNO VS SUVREMENO OBRAZOVANJE

Već dulje vrijeme javlja se potreba za odmakom od tradicionalnih modela poučavanja. Tradicionalnom nastavom koja se dominantno temelji na predavanjima nastavnika kao primarnom načinu prenošenja znanja nije moguće uspostaviti kontinuiranu interakciju između nastavnika i učenika te nema dovoljno vremena za samostalne aktivnosti učenika koje bi pomogle boljem i kvalitetnijem usvajanju nastavnih sadržaja [5].

Ograničavajući se samo na udžbenike i frontalni način rada ne mogu se razviti učeničke vještine, kritičko razmišljanje, rješavanje problema i pripremanje za cjeloživotno učenje [6]. Univerzalni model poučavanja koji „odgovara svima“, razred shvaća kao homogenu skupinu pojedinaca ne uzima u obzir raznolikost učeničke populacije te kao takav onemogućava postizanje najboljih rezultata učenika. Takav model poučavanja stvara prosječne učenike te ne potiče izvrsnost.

Za razliku od tradicionalnih modela u kojima je nastavnik prenositelj znanja, a učenici pasivni slušači, suvremeni nastavni modeli poučavanja usmjeravaju se na potrebe današnjih učenika kako bi ih se potaklo na samostalnost u učenju, poticalo njihovu kreativnost, učenje putem otkrivanja i problemskog rješavanja [7]. Kako bi to bilo moguće, najprije je potrebno navesti ih da sudjeluju u

nastavi i aktivnostima koje se provode. Čak i kada se određena nastavna metoda, istraživana u kontroliranim uvjetima, pokaže efikasnom u postizanju dobrih rezultata, ona postaje beznačajna ukoliko učenici ne žele aktivno sudjelovati u nastavnom procesu [8]. Rezultati istraživanja pokazuju da motivacija, angažiranost i postignuti rezultati opadaju kada nastavnik naglasak stavljaju na kontrolirajuće oblike rada i kada minimaliziraju mogućnost izbora samim učenicima [9]. Cilj suvremenih modela poučavanja je upravo povećanje angažiranosti učenika u formiranju i modificiranju vlastitog obrazovnog procesa [3]. Angažman je koncept koji uključuje kombinaciju pažnje, interesa i uživanja, obuhvaća samo usmjeravanje, interakciju, emocije i izbor koji su prirodno motivirani poticanjem aktivnosti/akcije [10].

Personalizirano poučavanje, kao neizostavan dio suvremene nastave i moderne škole, pruža brojne mogućnosti svakom učeniku koje su usmjerene na kreativniji i aktivniji odnos prema učenju te na buduće životne izazove [11]. Personalizacija je pojam koji se odnosi na procese i proizvode koji su prilagođeni pojedenoj pojedinoj osobi, bilo od strane drugog ljudskog bića ili tehnologije. Za postizanje personalizacije u obrazovanju, sve više se koriste multimedijske mogućnosti najnovijih tehnologija koje mogu individualizirati informacije u skladu s profilima učenja djece, s čime se povećava aktivno sudjelovanje djece u učenju [12].

Uloga tehnologije u suvremenom obrazovanju

Integracija tehnologije u obrazovanje nastala je kao potreba za odmakom od tradicionalnih modela poučavanja. Različiti suvremeni modeli poučavanja na različite načine koriste tehnologiju kao potporu u realizaciji nastavnog procesa. Na slici 1. vidimo ilustrativno prikazane suvremene metode poučavanja. U diferenciranom modelu nastavnici su ti koji biraju oblik tehnologije kojom će se učenici koristiti pri obradi određenog sadržaja. U individualiziranom modelu, korištenje tehnologije često se veže uz korištenje posebnih softvera koji će smanjiti utjecaj određene teškoće na sudjelovanje u nastavi. U personaliziranom modelu korištenje tehnologije se povezuje sa motivacijom za samostalnim istraživanjem, razmjenom podataka te većim izborom kako će se prikazati gotov rad učenika.



Slika 1. Suvremeni modeli poučavanja (Przemyslaw et al., 2015)

III. INTEGRACIJA MOBILNIH TEHNOLOGIJA U OBRAZOVNO OKRUŽENJE

Istraživanjem literature o integraciji mobilnih tehnologija u obrazovnom okruženju objavljene u recenziranim časopisima koji bi trebali biti među najboljim publikacijama u području obrazovne tehnologije nailazimo na brojne studije koje se ovisno o svrsi istraživanja studije mogu sintetizirati u tri skupine:

- Studije koji razmatraju implementirane aplikacije u različitim razinama obrazovanja te rezultate ostvarenog aplikativnog učenja ili ocjenjivanja
- Studije koji razmatraju upotrebu mobilnih tehnologija (pametnih telefona / tableta) u nastavnom procesu ali ne uključuju upotrebu konkretnih aplikacija već je naglasak na upotrebi uređaja odnosno na stvaranju učinkovitih obrazovnih okruženja potpomognutih mobilnom tehnologijom
- Studije koje razmatraju razvojni dizajn te modeliranje aplikacije za nastavu

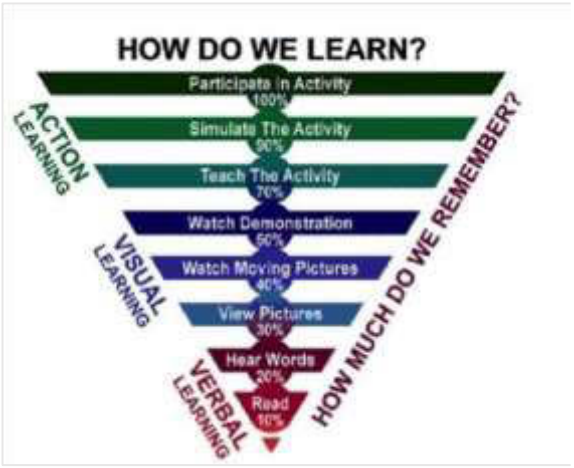
Najveći broj objavljenih studija pripadaj drugoj skupini. Postoje brojni pregledi literature i meta analize o mobilnom učenju koje se usredotočuju na učinkovitost mobilnog učenja i razvoj mobilnih uređaja sustava učenja koji pomažu u učenju učenika [13], [14], [15], [16], [17].

Postoje studije o mobilnom učenju koje se fokusiraju na obrazovanje K-12 [18], [19], visoko obrazovanje [20], specifične domene učenja [21] ili pregledna studija o ocjenjivanju utemeljenom na mobilnim uređajima [22], pregledna studija o učenju istraživanjem putem mobilnih aktivnosti [23], pregled mobilnog učenja: aplikacije i problemi [24]. Većina tih preglednih studija sugerira da je mobilno učenje obećavajuća metoda u poboljšanju učenja učeničkih postignuća, motivacije i stavova učenja. [25] u svojem radu razvijaju dizajn učenja za mobilne uređaje te predstavljaju metodologiju za vizualizaciju i proizvodnju elektroničkih materijala u mobilnom okruženjima. Zbog brzog rasta pouzdanih mobilnih tehnologija, BYOD pristup (donijeti vlastiti uređaj) postao je izvodljiva nastavna strategija za promicanje aktivnog sudjelovanja učenika tijekom procesa učenja [26] te je mobilno učenje postalo izvrstan alat za isporuku obrazovnih resursa.

Također, postoje razna istraživanja koja su usmjerena na stvaranje učinkovitih okruženja za učenje i podučavanje koja će unaprijediti tradicionalni pristup. Jedno od preporučenih koje poboljšava interes i motivaciju djece u učenju je učenje temeljeno na igri (GBL) [27]. Istraživanja pokazuju da se učenjem temeljeno na igri putem mobilnih tehnologija (tablet / pametni telefon) mogu usvojiti određeni obrazovni koncepti na zabavni način te da je mobilna tehnologija ugodna platforma za učenje. Mobilne i tablet platforme slične su na mnogo načina. Obje platforme pružaju interakcija s višestrukim dodirom, pojednostavljuje ulazne naredbe (npr. tipkanje i klikanje) i smanjenje poteškoća i frustracija u učenju uzrokovane netočnim ciljevima djelovanja [30].

Mobilno učenje (m-učenje) može se definirati kao upotreba mobilnih uređaja kao posrednika u procesu učenja

i poučavanja, što znači da se ova uporaba može promatrati kao učenje iz mobilnog i učenje s mobilnim uređajem [25].



Slika 2. Obrazovanje 3.0 i pedagogija mobilnog učenja

Učenje s mobilnim uređajem podrazumijevalo bi upotrebu različitih mrežnih stranica te komunikacijskih platformi za dobivanje obrazovnih resursa, dok bi učenje iz mobilnog uređaja podrazumijevalo upotrebu obrazovnih aplikacija koje su napravljene s ciljem usvajanja određenog obrazovnog sadržaja.

IV. APLIKACIJE U OBRAZOVANJU

Posljednjih godina u obrazovanju sve više koristimo aplikacije u nastavi kako za prikazivanje i usvajanje nastavnih sadržaja tako i za provjeru ostvarenosti ishoda učenja. Upotrebom aplikacija u nastavnom procesu može se postići personalizacija kojom se učenika postavlja u centar nastavnog procesa gdje on sam stvara, uređuje i modificira edukacijski proces. Također, upotrebom aplikativno simuliranih situacija, učenici mogu provjeravati vlastite ideje, stvarati pretpostavke na temelju viđenog te elaborirati informacije do razine potpunog razumijevanja onoga što se istražuje. Na taj način ishodi učenja ne obuhvaćaju samo činjenice i definicije već i iskustva potaknuta stjecanjem određenog znanja. Istraživanja pokazuju da upotreba aplikacija u nastavi povećava zainteresiranost i motivaciju učenika te bolje pamćenje sadržaja. Upotreba aplikacija osobito može biti značajna u obrazovnom kontekstu učenika sa posebnim potrebama [32] gdje se upotrebom aplikacije nastoje umanjiti utjecaji određene teškoće na sudjelovanje u nastavi. Postoji posebno područje istraživanja aplikacija namijenjenih učenicima s teškoćama. Osim toga, razlika u postignućima u uspješnosti učenja je drugo područje koje treba obraditi istraživanjem [33]. Aplikacije osjetljive na dodir mogu biti izuzetno moćan alat za usvajanje obrazovnih koncepata djece predškolske dobi prije ulaska u formalno školovanje. Razvojno razdoblje vrtićkog i predškolskog uzrasta može biti posebno prikladno za ovu vrstu učenja te su napravljena brojna istraživanja o tome [34]. Iako se broj obrazovnih aplikacija svakodnevno povećava, mnoge od tih aplikacija nemaju odgovarajuće pedagoške temelje i ne ispunjavaju obrazovne ciljeve. Nedavni pregled aplikacija o obrazovanju je otkrio da većina razmatranih aplikacija nije prikladna za djecu. Neke aplikacije nisu zadovoljile smjernice okvira, dok neke nisu čak ni usredotočene na cilj učenja [35]. Drugim riječima, te aplikacije nisu bile namijenjene promicanju učenja i razvoja, već se koriste za druge svrhe kao što su zabava ili popunjavanje slobodnog vremena za djecu.

V. PREGLED I ANALIZA ODABRANIH IMPLEMENTIRANIH APLIKACIJA

Istraživanjem literature o upotrebi aplikacija u obrazovanju, odabrani su radovi koji razmatraju rezultate implementiranih aplikacija u nastavi koji prikazuju ostvarene rezultate aplikativnog učenja ili ocjenjivanja. Cilj analize bio je usporediti 15 studija koji razmatraju značajke implementiranih aplikacija u nastavnom procesu po dobi sudionika, području učenja aplikacije, vrsti (namjeni) aplikacije, te podržanoj tehnologiji. Također, cilj je ispitati utjecaj upotrebe aplikacija u nastavi na ostvarivanje rezultata aplikativnog učenja ili ocjenjivanja. Ovom se cilju pristupa rješavanjem tri istraživačka pitanja.

RQ1: Kakav je utjecaj aplikativnog učenja na uspješnost učenja učenika?

RQ2: Koji je učinak upotrebe aplikacije na motivaciju učenika za učenje?

RQ2: Kakvi su stavovi i percepcije učenika o upotrebi aplikacija u nastavnom procesu?

Analiza je napravljena na način da su na temelju definiranih istraživačkih pitanja istraživane sljedeće značajke odabranih radova:

i Naziv aplikacije i autori studije

ii Sudionici – dob sudionika (predškolski uzrast, osnovna škola, srednja škola, fakultet) i broj sudionika

iii Područje za koje je napravljena aplikacija (Programiranje, IT, STEM, Anatomija, ostali predmeti)

iv Namjena (vrsta) aplikacije – aplikacija za usvajanje nastavnih sadržaja, aplikacija za ponavljanje, aplikacija za ocjenjivanje (sumativna, formativna), mobilno anketiranje u razredu, samoprocjena

v Tehnologija (npr. prilagodljiva, kontekstualna, proširena stvarnost, virtualna stvarnost)

vi Utjecaj upotrebe aplikacije na uspješnost usvajanja nastavnog sadržaja – RQ1 (podrazumijeva poboljšanje / povećanje)

vii Utjecaj upotrebe aplikacije na motivaciju učenika prilikom učenja te stavovi / percepcije učenika (RQ2 i RQ3) kodirani su kao: pozitivan (podrazumijeva poboljšanje / povećanje), negativan (implicira slabljenje / smanjenje), neutralan (ne implicirajući ni pozitivan ni negativan utjecaj) ili nije određeno

Tablica 1. prikazuje detaljnu analizu razmatranih aplikacija.

Tablica 1. Analiza razmatranih aplikacija

R. br.	Autor	Aplikacija	Sudionici – dob (broj)			Područje	Namjena aplikacije		Tehnologija	RQ3_1	RQ3_2	RQ3_3	Napomena
			Vrtić	OŠ	SS		Učenje	Ocjenjivanje					
1.	S. Pila, F. Aladé, K. J. Sheehan, A. R. Lauricella, E. A. Wartella , 2019.	DAISY	5 god (28)	-	-	Programiranje	+	-		+	+	Ovisno o privlačnosti aplikacije	Mali uzorak (mala i napredna djeca)
2.		CODABLE		-	-	Programiranje	+	-		+	+		
3.	Salahli M.A., Yildirim E., Gasimzadeh T., Alasgarova F., Guliyev A, 2016.	SCRATCH	-	-	+	Programiranje	+	+		+	Nije poznato	Nije poznato	
4.	Kelly J. Sheehan, Sarah C. Pila, Alexis R. Lauricella, Ellen A. Wartella, 2019.	PBS KIDS ScratchJr	4,5 - 5 (31)	-	-	Programiranje	+	-		+	Nije poznato	Nije poznato	Učenje u interakciji s roditeljima
5.	M. C. Juan; D. Furió, L. Alem; P. Ashworth, 20	ARGreenet	-	+ (38)	-	Recikliranje STEM	Stavovi (utjecaj na ponašanje)		AR	+	+	+	
6.		BasicGreenet	-		-				AR	+	+	+	
7.	Ardito C., Buono P., Costabile M. F., Lanzilotti R. ,Piccinno, A., 2009.	Explore	-	-	+ (42)	Arheologija	+	-	AR	-	+	+	
8.	W. Chen, X. Gu and L.-H. Wong, 2019	CBMS	-	+	-	Pedagogija	Upravljanje razredom			+	Nije poznato	Nije poznato	
9.	Nazatul A. A. M., Hazura M. i Rossilawati S., 2015.	µprocesor	Studenti			IT - Hardware	+	-	AR	Nije poznato	+	+	Nisu svi imali adekvatan uređaj
10.	Z. Bogdanović, D. Barać, B. Jovanić, S. Popović, 2013.	Moodle_LMS	Studenti				+	+		+	+	+	
11.	Pao-Nan Chou, ChiCheng Chang, ChingHsin Lin , 2017	Socrative	-	-	+ (46)	Engleski jezik	-	+		Ovisno o vrsti vrednovanja	+	+	
12.	Fernández-López, Á., RodríGuez-Fórtiz, M. J., Rodríguez-A, M. L., ve Martínez-S, M. J.,2013.	PICAA	-	+ Posebne škole za djecu s poteškoćama (39)	-	Različita područja (jezik, matematika…)	+	+		Nije poznato	+	Nije poznato	Za djecu s posebnim potrebama
13.	W. Tarng, Kuo-Liang O, 2012	Butterfly Ecology	-	+ (60)	-	Ekologija STEM	+	-	AR	+	Nije poznato	Nije poznato	
14.	J. Bacca, S. Baldiris, Fabrega, Kinshuk, Graf, 2015	Paint-CAR	-	-	+ (13)	Struka - automobili	+	-	AR	Nije poznato	+	+	Mali broj ispitanika
15.	S. Jamali, M.F. Shiratuddin, K.W. Wong , C. L. Oskam, 2015	HuMAR	Studenti (30)			Anatomija	+	-	AR	+	+	+	

Rezultati

Analiza nad studijama implementiranih aplikacija razmatra ostvarivanje rezultata aplikativnog učenja ili ocjenjivanja putem eksperimentalnih istraživanja tijekom implementacija aplikacija. Rezultati istraživanja prikazani su prema značajkama definiranih istraživačkih pitanja.

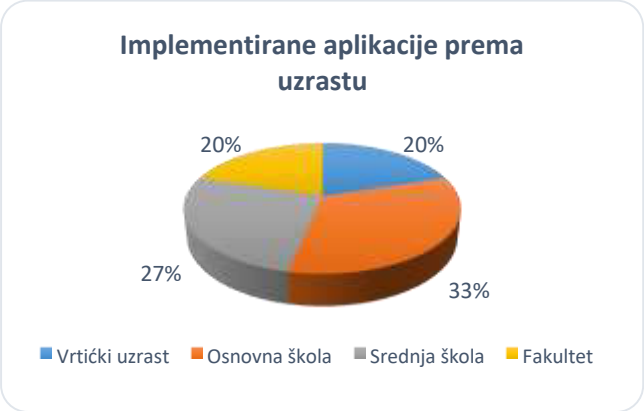
Sudionici

U 15 implementiranih aplikacija podjednako su zastupljene sve dobne skupine, detaljan prikaz po dobi može se vidjeti u Grafikonu 1.

Jedna istraživana skupina osnovno školskog uzrasta su djeca sa poteškoćama u razvoju odnosno djeca s posebnim potrebama.

Zanimljivo je da su sve razmatrane aplikacije predškolskog (vrtićkog) uzrasta vezane za područje učenja programiranje. Učenje programiranja male djece putem aplikacija predstavlja izuzetno atraktivno područje istraživanja.

Grafikon 1. Aplikacije prema dobi sudionika



Područje učenja

Prema području učenja, 5 aplikacija je vezano uz sadržaje nastavnog predmeta Informatika, od čega su 4 aplikacije iz područja programiranja, a jedna aplikacija je iz područja građe računala. Od ostalih aplikacija, tri aplikacije su iz STEM područja, jedna aplikacija je napravljena za strukovno mehaničko obrazovanje, aplikacija za djecu s teškoćama obuhvaća različite predmete učenja, te ostale aplikacije dolaze iz ostalih različitih područja učenja.

Namjena aplikacije

Većina aplikacija je napravljena za usvajanje nastavnih sadržaja (73%), jedna aplikacija je samo za ocjenjivanje, a tri aplikacije su cjelovite te omogućavaju i učenje i ocjenjivanje.

Tehnologija

Od ukupnog broja razmatranih aplikacija njih 46% je napravljeno za tehnologiju proširene stvarnosti (engl. Augmented Reality (AR)). Zanimljivo je da su sve aplikacije iz STEM područja napravljene u AR tehnologiji, koja se i za anatomiju, arheologiju, strukovno obrazovanje (automobili) te područje Hardvera iz IT područja pokazala kao dobar izbor. Učenici koji su koristili aplikacije napravljene u AR-u, iskazuju poboljšano iskustvo učenja te zadovoljstvo upotrjebljenim aplikacijama.

Utjecaj aplikacije na uspješnost usvajanja sadržaja te na motivaciju i stavove učenika

Utjecaj upotrebe aplikacija u nastavi na učinak učenja, motivaciju i stavove prikazan je Grafikonom 2.

Grafikon 2. Utjecaj upotrebe aplikacija u nastavi na učinak učenja, motivaciju i stavove učenika



Gotovo sve aplikacije imaju pozitivan učinak na učenje, dok jedna aplikacija nema ni pozitivnog ni negativnog utjecaja na učenje.

Utjecaj aplikacije na motivaciju poznat je u 73% te je u potpunosti pozitivan, odnosno može se zaključiti da učenici pokazuju povećanu motivaciju za učenjem prilikom upotrebe aplikacije. Također, stavovi učenika prema aplikacijama u nastavi su pozitivni (poznato za 66% aplikacija). Autori [34] ističu da privlačnost određene aplikacije, osobito kod male djece, u velikoj mjeri je povezana sa motivacijom za upotrebu aplikacije odnosno učenjem iz aplikacije. Što je aplikacija privlačnija za upotrebu, veća je motivacija. U razmatranju rezultata dobivenih analizom odabranih aplikacija koja pokazuju da upotreba aplikacija u nastavi pospješuje učenje odnosno da aplikacije imaju izuzetno pozitivan utjecaj na usvajanje obrazovnih sadržaja te povećavaju motivacije učenika za učenjem treba uzeti u obzir da većina razmatranih istraživanja rezultata implementiranih aplikacija u nastavi imaju nedostatke u svojim istraživanjima, bilo u vidu brojnosti ili uzorkovanja populacije istraživanja. Na primjeru aplikacija za programiranje, manjkavost istraživanja je u malom broju djece / učenika koja su sudjelovala u istraživanju a kod aplikacija za malu djecu, kako se radi o skupinama djece koja nisu slučajno odabrana već su njihovi roditelji izabrali da pohađaju malu školu programiranja, uzorak osim što je po brojnosti malen, uzorak je i prigodan odnosno nije reprezentativan. Problem sa prigodnim i malobrojnim uzorkom ima većina studija razmatranih implementiranih aplikacija.

VI. ZAKLJUČAK

Integracija mobilne tehnologije u obrazovno okruženje doprinosi razvoju suvremenog obrazovanja koji se prilagođava učeniku uvažavajući njegove različitosti. Mobilni uređaji poput tableta i pametnih telefona su prepoznati kao alati za učenje s velikim potencijalom u

formalnom i neformalnom okruženju te je veliki broj istraživanja usmjeren upravo na integraciju mobilnih tehnologija u obrazovanje a popularizacijom BYOD pristupa broj publikacija o mobilnom učenju osobito raste.

Upotreba aplikacija u nastavi osim što pospješuje sam proces usvajanja obrazovnih sadržaja i povećava motivaciju učenika za usvajanjem sadržaja omogućava i personalizirano poučavanje u kojem učenik modificira obrazovni proces svojim potrebama. Sveukupno gledano, prema razmatranim analizama odabrane literature, može se zaključiti da su digitalne aplikacije ugodna platforma za učenje putem koje se mogu usvojiti obrazovani sadržaji na zabavan način a suvremene tehnologija poput proširene stvarnosti predstavljaju izvrsnu podloga za prikazivanje nastavnih sadržaja.

Posljednjih godina značajno se povećava broj edukativnih aplikacija na različitim razinama obrazovanja. Iako se često koriste kao dopuna formalnoj nastavi odnosno kao dodatni nastavni alati, mnogobrojne aplikacije ne udovoljavaju temeljnim pedagoškim načelima. Postavljanje metrike kojom će se formirati zahtjevi kojima treba udovoljiti određena obrazovna aplikacija, mjerenje znanja koje se stvarno može provjeriti aplikacijama prema određenoj taksonomiji znanja, istraživanje interakcije s korisnikom s ciljem povećanja privlačnosti aplikacije samo su neka od mogućih područja istraživanja vezanih za aplikacije.

VII. LITERATURA

[1.] Irina Golitsyna (2015), Application of Web Services in Teaching of IT – Discipline, Procedia – Social and Behavioral Sciences

[2.] Adams Becker, S., Cummins, M., Davis, A., Freeman, A., Hall Giesinger, C., & Ananthanarayanan, V. (2017). NMC horizon report: 2017 higher education edition Austin, Texas: The New Media Consortium

[3.] Sung, Y.-T., Chang, K.-E., & Liu, T.-C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. Computers & Education, 94, 252–275.

[4.] West, M., & Vosloo, S. E. (2013). UNESCO policy guidelines for mobile learning.Paris: UNESCO.

[5.] Jensen, E. (2003). Super-nastava: nastavne strategije za kvalitetnu školu i uspješno učenje. Zagreb: Educa

[6.] Bognar, L., Matijević, M. (2002). Didaktika. Zagreb: Školska knjiga

[7.] Matić, T. (2014). Metode aktivnog učenja u razrednoj nastavi. Osijek: Učiteljski fakultet.

[8.] Ferlazzo, L. (2017). Student Engagement: Key to Personalized Learning. Getting Personalization Right, 74 (6), str. 28-33.

[9.] Netcoh, S. (2017). Balancing freedom and limitations: A case study of choice provision in a personalized learning class. Teaching nad Teacher Education, 66 (2017), str. 383-392.

[10.] Noorhidawati, A., Ghalebandi, SG, i Siti Hajar, R. (2015): How Do Young Children Engage with Mobile Apps? Cognitive, Psychomotor, and Affective Perspective, Computers & Education

[11.] Przemyslaw, C. et al. (2015). The Book of Trends in Education 2.0. Gdynia: Young Digital Planet SA a Sanoma Company.

[12.] Natalia Kucirkova (2019): Children's agency by design: Design parameters for personalization in story-making apps, International Journal of Child-Computer Interaction

[13.] Chee, K. N., Yahaya, N., Ibrahim, N. H., & Noor Hassan, M. (2017). Review of mobile learning trends 2010-2015: A meta-analysis. Educational Technology & Society, 20(2)

[14.] Chiang, F.-K., Zhu, G., Wang, Q., Cui, Z., Cai, S., & Yu, S. (2016). Research and trends in mobile learning from 1976 to 2013: A content analysis of patents in selected databases. British Journal of Educational Technology, 47, 1006– 1019.

[15.] Hwang, G.-J., & Wu, P.-H. (2014). Applications, impacts and trends of mobile technologyenhanced learning: A review of 2008–2012 publications in selected SSCI journals. International Journal Mobile Learning and organization

[16.] Sung, Y.-T., Chang, K.-E., & Liu, T.-C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. Computers & Education, 94, 252–275.

[17.] 17. Wu, W.-H., Wu, Y.-C. J., Chen, C.-Y., Kao, H.Y., Lin, C.-H., & Huang, S.-H. (2012). Review of trends from mobile learning studies: A metaanalysis. Computers & Education, 59(2), 817– 827.

[18.] Crompton, H., Burke, D., & Gregory, K. H. (2017). The use of mobile learning in PK-12 education: A systematic review. Computers & Education,

[19.] Liu, M., Scordino, R., Geurtz, R., Navarrete, C., Ko, Y., & Lim, M. (2014). A Look at research on mobile learning in K–12 education from 2007 to the present. Journal of Research on Technology in Education, 46(4), 325–372.

[20.] Pimmer, C., Mateescu, M., & Gröhbiel, U. (2016). Mobile and ubiquitous learning in higher education settings. A systematic review of empirical studies. Computers in Human Behavior, 63, 490–501.

[21.] Helen Crompton, Diane Burke, Kristen H. Gregory (2017): The use of mobile learning in PK-12 education: A systematic review, Computers & Education

[22.] Stavros A. Nikou, Anastasios A. Economides (2018): Mobile-based assessment: A literature review of publications in major referred journals from 2009 to 2018, Computers & Education

[23.] Ángel Suárez, Marcus Specht, Fleur Prinsen, Marco Kalz, Stefaan Ternier (2018): A review of the types of mobile activities in mobile inquirybased learning, Computers & Education

[24.] Liyana Shuib, Shahaboddin Shamshirband, Mohammad Hafiz Ismail (2015): A review of mobile pervasive learning: Applications and issues, Computers in Human Behevior

[25.] Nauris Paulins, Signe Balina, Irina Arhipova (2015): Learning Content Development Methodology for Mobile Devices, Procedia Computer Science

[26.] Pao-Nan Chou, Chi-Cheng Chang, Ching-Hsin Lin (2017): BYOD or not: A comparison of two assessment strategies for student learning, Computers in Human Behavior

[27.] Chen & Law, 2016; Ching-Huei Chen, Victor Law (2016): Scaffolding individual and collaborative game-based learning in learning performance and intrinsic motivation, Computers in Human Behavior

[28.] Salahli M.A., Yildirim E., Gasimzadeh T., Alasgarova F., Guliyev A. (2017) : One mobile application for the development of programming skills of secondary school students, Procedia Computer Science

[29.] Weiyun Chen, Xiaoqing Gu and Lung-Hsiang Wong (2019): To click or not to click: Effectiveness of rating classroom behaviors on academic achievement with tablets, British Journal of Educational Technology

[30.] Ngan Kuen Lai, Tan Fong Ang, Lip Yee Por, Chee Sun Liew (2018): Learning through intuitive interface: A case study on preschool learning, Computers & Education

[31.] Zorica Bogdanović, Dušan Barać, Branislav Jovanić, Snežana Popović (2013): Evaluation of mobile assessment in a learning management system, British Journal of Educational Technology

[32.] Álvaro Fernández-López, María José RodríguezFórtiz, María Luisa Rodríguez-Almendros, María José Martínez-Segura (2013): Mobile learning technology based on iOS devices to support students with special education needs, Computers & Education

[33.] Hung, H.-C., Young, S. S.-C., & Lin, C.-P. (2015). No student left behind: a collaborative and competitive game-based learning environment to reduce the achievement gap of EFL students in Taiwan. *Technology, pedagogy and Education, 24*(1), 35-49.

[34.] Sarah Pila, Fashina Aladé, Kelly J. Sheehan, Alexis R. Lauricella, Ellen A. Wartella (2019): Learning to code via tablet applications: An evaluation of Daisy the Dinosaur and Kodable as learning tools for you, Computers & Education

[35.] Stamatios Papadakis, Michail Kalogiannakis, Nicholas Zaranis (2018): Educational apps from the Android Google Play for Greek, Computers & Education

[36.] Flannery, L., & Bers, M. U. (2013). Let's dance the "robot hokey pokey!": Children's programming approaches and achievement throughout early cognitive development. Journal of Research on Technology in Education

[37.] Portelance, D. J., Strawhacker, A. L., & Bers, M. U. (2015). Constructing the ScratchJr programming language in the early childhood classroom. International Journal of Technology and Design Education,

[38.] Strawhacker, A., Lee, M., Caine, C., & Bers, M. U. (2015). ScratchJr demo: A coding language for kindergarden. Paper presented at the international conference on interaction design and children, boston, MA.

[39.] Common Sense Media. (n.d.). Cool coding apps and websites for kids. Retrieved from

[40.] Huizenga, J., Admiraal, W., Akkerman, S., & Dam, G. T. (2009). Mobile game-based learning in secondary education: engagement, motivation and learning in a mobile city game. *Journal of Computer Assisted Learning, 25*(4), 332-344

[41.] Nazatul Aini Abd Majid, Hazura Mohammed i Rossilawati Sulaiman (2015): Students' perception of mobile augmented reality applications in learning computer organization, Procedia - Social and Behavioral Sciences

[42.] Netcoh, S. (2017). Balancing freedom and limitations: A case study of choice provision in a personalized learning class. Teaching nad Teacher Education, 66 (2017), str. 383-392.

[43.] Danakorn Nincarean, Mohamad Bilal Ali, Noor Dayana Abdul Halim, Mohd Hishamuddin Abdul Rahman (2013): Mobile augmented reality_the potential for education, Procedia - Social and Behavioral Sciences

[44.] Monika Simkova, Hana Tomaskovaa, Zuzana Nemcovaa (2012): Mobile education in tools , Social and Behavioral Sciences

[45.] Yanjie SONG & Cheuk Lun Alvin MA (2013): Bring Your Own Device (BYOD) for Seamless Science Inquiry: A case study in a Primary School

[46.] Stavros A. Nikou and Anastasios A. Economides (2018): Factors that influence behavioral intention to use mobile - based assessment: A STEM teachers' perspective, British Journal of Educational Technology

[47.] Tsung-Chou Chang, Ya-Fen Tsai, and FongGong Wu (2015): Interaction Design of Digital Teaching Improves Teaching and Learning Effectiveness

[48.] Liliya Andreicheva, Rustam Latypov (2015): Design Of E-Learning System: M-Learning Component , Procedia - Social and Behavioral Sciences

[49.] Kelly J. Sheehan, Sarah C. Pila, Alexis R. Lauricella, Ellen A. Wartella (2019): ParentChild Interaction and Children's Learning from a Coding Application, Computers & Education

[50.] Marta Gómez Domingo, Antoni Badia Garganté (2016): Exploring the use of educational technology in primary education: Teachers' perception of mobile technology learning impacts and applications' use in the classroom, Computers in Human Behavior

[51.] Carmelo Ardito & Paolo Buono & Maria Francesca Costabile & Rosa Lanzilotti & Antonio Piccinno (2009): Enabling Interactive Exploration of Cultural Heritage: An Experience of Designing Systems for Mobile Devices, Springer Science + Business Media B.V

[52.] Hung, H.-C., Young, S. S.-C., & Lin, C.-P. (2015). No student left behind: a collaborative and competitive game-based learning environment to reduce the achievement gap of EFL students in Taiwan. Technology, pedagogy and Education, 24(1), 35-49.

[53.] Flannery, L., Kazakoff, E. R., Bonta, P., Silverman, B., Bers, M. U., & Resnick, M. (2013). Designing ScratchJr: Support for early childhood learning through computer programming. Paper presented at the interational conference on instruction design and children, New York, NY.

[54.] Malina A. (2018). Personalizirano poučavanje i modeli suvremenog poučavanja [Diplomski rad]. Zagreb: Edukacijsko-rehabilitacijski fakultet

[55.] M. Carmen Juan; David Furió, Leila Alem; Peta Ashworth, Juan Cano (2009): ARGreenet and BasicGreenet: Two mobile games for learning how to recycle

[56.] Wernhuar Tarng, Kuo-Liang Ou (2012): A Study of Campus Butterfly Ecology Learning System based on Augmented Reality and Mobile Learning

[57.] Tomáš Jeřábek, Vladimír Rambousek, Radka Wildová (2014): Specifics of Visual Perception of The Augmented Reality in The Context of Education, Procedia - Social and Behavioral Sciences

[58.] Siti Salmi Jamali, Mohd Fairuz Shiratuddin, Kok Wai Wong, Charlotte L. Oskam (2015): Utilising Mobile-Augmented Reality for Learning Human Anatomy, Procedia - Social and Behavioral Sciences

[59.] Jorge Bacca, Silvia Baldiris, Ramon Fabregat, Kinshuk, Sabine Graf (2015): Mobile Augmented Reality in Vocational Education and Training, Procedia Computer Science

Intermedijalnost autorskih slikovnica Ivane Guljašević Kuman[1]

Maja Verdonik*, Viktorija Vitas**
*Sveučilište u Rijeci/Učiteljski fakultet, Rijeka, Hrvatska
**DV „Sušak" Rijeka, Hrvatska
mverdonik@ufri.uniri.hr

***Sažetak* - Suvremena hrvatska ilustratorica Ivana Guljašević Kuman autorica je tekstova i ilustracija više slikovnica kao i animiranih filmova objavljenih na multimedijskim DVD izdanjima. U radu se propituje intermedijalna povezanost Autoričinih tiskanih slikovnica s digitalnim interaktivnim sadržajima uključenim u DVD izdanja koja prate ove slikovnice, te mogućnosti njihove primjene radu s djecom rane i predškolske dobi kao i u nastavnom procesu s učenicima mlađe školske dobi. Provedene analize pokazuju očitovanje intermedijalnosti autorskih slikovnica Ivane Guljašević Kuman u primjeni izražajnih sredstava prisutnih u filmskoj kao i u likovnim umjetnostima. Ujedno, s obzirom na mogućnosti multimedijskih DVD izdanja, pridodane tiskanim slikovnicama, kao što su: slušanje teksta i popratne glazbe, pjevanje uz instrumentalno izvođenje glazbe, doživljavanje animiranih filmova kao svojevrsnih oživljenih ilustracija itd. – dijete čitatelj/ slušatelj/ gledatelj/ izvođač, potaknuto intermedijalnošću slikovnica Ivane Guljašević Kuman, ima mogućnost cjelovitog razvoja putem uporabe kvalitetnih medijskih ostvarenja.**

Ključne riječi - intermedijalnost; slikovnica; dječja književnost; medijska kultura; Ivana Guljašević Kuman

I. UVOD

Ivana Guljašević Kuman suvremena je hrvatska likovna umjetnica – ilustratorica, autorica stripova, književnica i redateljica animiranih filmova. Osebujnost njezinoga djelovanja u kontekstu dječje književnosti, između ostalog, objavljivanje je multimedijskih DVD izdanja animiranih filmova, s popratnim sadržajima, čiji su sastavni dio i tiskane slikovnice. Na primjeru crtanog animiranog filma *Čovječuljak Snovuljak* (2009) i istoimene tiskane slikovnice (2012) Ivane Guljašević Kuman, cilj je rada pokazati intermedijalne poveznice ovih ostvarenja te predložiti mogućnosti njihove primjene u radu s djecom rane i predškolske, te mlađe školske dobi. Ujedno, u radu se predstavlja i mjesto i uloga medijske kulture u aktualnim nacionalnim odgojno-obrazovnim kurikulumima za rani i predškolski odgoj i obrazovanje, te za osnovne škole.

II. ŽANROVSKE ODREDNICE DIGITALNE/ELEKTRONIČKE KNJIŽEVNOSTI

Prema riječima Dubravke Težak i Marine Gabelice: „Pojavom digitalnih medija povratno se mijenja i tradicionalna književnost. Ove se promjene očituju u načinima konstrukcije pripovjedi, jeziku i stilu, a svakako i tematici koja korespondira s interesima suvremenih čitatelja. U isto vrijeme, na digitalnim se medijima razvija jedan posve novi književni sustav – riječ je o digitalnoj ili elektroničkoj književnosti" [17]. Autorice predlažu podjelu digitalnih djela namijenjenih djeci i mladima s obzirom na zastupljenost književnoga teksta u odnosu na druge medijske objekte, smatrajući ovaj kriterij ključnim budući da se njime ukazuje na mogućnost njihove implementacije u postojeći obrazovni sustav [17]. Riječ je o sljedećoj podjeli:

„1. Digitalna djela koja književnost koriste paratekstualno (žanr: poezija i proza u Flashu, *casual* igre)
2. Digitalna djela koja književnost koriste kao svoje uporište (žanr: interaktivne priče)
3. Digitalna djela koja proširuju književno djelo (žanr: hipertekstualna djela, vizualni romani)" [17].

U okviru ove podjele autorice Težak i Gabelica spominju i autorske slikovnice i multimedijska DVD izdanja Ivane Guljašević Kuman, u skupini digitalnih djela koja književnost koriste kao svoje uporište (žanr: interaktivne priče), iako je u njima riječ o minimalnom udjelu interakcije [17].

III. O AUTORSKIM SLIKOVNICAMA IVANE GULJAŠEVIĆ KUMAN

Ivana Guljašević Kuman svestrana je hrvatska umjetnica koja samostalno i u suradnji s Brunom Kumanom objavljuje slikovnice za djecu nastojeći spojiti priču i audiovizualnu umjetnost [3]. Tiskana izdanja njezinih slikovnica čine cjelinu s multimedijskim DVD izdanjima koja sadrže animirane filmove prema kojima su nastale tiskane slikovnice,

[1] Rad je financiralo Sveučilište u Rijeci sredstvima dodijeljenim projektu broj: uniri-pr-human-19-9.

digitalne inačice slikovnica, pjesmice, igre i drugo, i to na hrvatskom i na engleskom jeziku [3].

Većina crtanih animiranih filmova Ivane Guljašević Kuman najprije je bila objavljena na CD izdanjima, kao prilozi uz dječji časopis *Smib*. Riječ je o animiranim filmovima umjetničkih autorskih vrijednosti, snimljenima u maniri Zagrebačke škole crtanog filma, koju Autorica njeguje kao likovni i tematski izričaj u svojim projektima [2] [7] [8].

Kao što ističe Božidar Trkulja, Ivana Guljašević Kuman jedna je od rijetkih autorica u potpunosti posvećenih djeci vrtićke i mlađe osnovnoškolske dobi [18]. „Iako ih nimalo ne štedi, a to je vidljivo u svim filmovima, boje u autoričinu radu vrlo su skladne, što je također jedno od obilježja prepoznatljivosti njezinih ilustracija i filmova. Priče su smještene u seoska dvorišta, šume, polarna prostranstva, pod morem, iznad grada, u kuće, dvorce i oblake, a gotovo svaka ima i odgojnu ulogu, nenametljivu, prijemčivu i na koncu društveno korisnu, a to zaista treba cijeniti u poplavi podilazećih komercijalnih animiranih sadržaja. Ivana Guljašević Kuman u svojim filmovima naglašava važnost prijateljstva, zajedništva, očuvanja prirode, učenja i igre (što je) s obzirom na vrijeme u kojem živimo, ali prije svega na dob kojoj se obraća Guljašević Kuman, itekako važno", zaključuje Trkulja [18].

IV. Intermedijalnost autorskih slikovnica Ivane Guljašević Kuman

Intermedijalne poveznice između crtanih animiranih filmova i tiskanih slikovnica Ivane Guljašević Kuman očituju se u komponentama zajedničkim animiranom filmu i slikovnici kao filmskim odnosno književno-likovnim umjetničkim vrstama.

Prema Hrvatskoj enciklopediji Leksikografskog zavoda Miroslav Krleža intermedijalnost: „počinje tek onda kada se u jednome mediju, upravo radi oživljavanja njegove specifičnosti, primjenjuju zakonitosti i pravila drugoga medija (npr. filmska montaža u književnosti ili kazališna režija u filmu)" [6]. U animiranim filmovima nalazimo likovne komponente, primjerice u crtanom filmu: liniju, boju, plohu, volumen, kao i filmske: kadar, plan, kut snimanja, pokret kamere, zvuk, montažu. Stoga pojedini autori govore o animiranom filmu kao o graničnom području – kombinaciji likovne i filmske umjetnosti, ali i kao o posebnoj, osmoj umjetnosti [9].

S obzirom na dominantnu vizualnu komponentu (uz koju je zvuk ravnopravni element izražavanja) – film, posebice animirani film najbliži je slikarstvu [9] što ga povezuje i sa slikovnicom. Filmu i slikarstvu, odnosno ilustraciji, zajednička je komponenta okvir koji ih odjeljuje od zbilje te je, u procesu stvaranja, riječ o izboru onoga što će se naći u tom okviru te onoga što će ostati izvan njega. Pri tom se poglavito misli na kompoziciju filmskog kadra/slike odnosno na organizaciju sadržaja (elemenata) unutar okvira [9].

Imajući na umu da autorske slikovnice Ivane Guljašević Kuman pripadaju skupini pripovjednih slikovnica odnosno da su njihove fabule zasnovane na spoju slike (ilustracije) i pripovjednog teksta, poveznicu s filmom čini i komponenta narativnosti, neovisno o drugim različitostima [9].

A. Crtani animirani film i slikovnica Čovječuljak Snovuljak (2009/2012)

Kratkometražni crtani animirani film u boji, dužine trajanja oko 5 minuta, *Čovječuljak Snovuljak* odabran je za analizu među istovrsnim filmovima Ivane Guljašević Kuman, kao reprezentativni primjer filma koji sadrži karakteristike većine autoričinih multimedijskih DVD izdanja. Fabula filma – priča o čarobnom biću – Čovječuljku koji noću donosi lijepe snove djeci i odraslima, tjerajući od njih one nelagodne, uključuje i intermedijalni citat [14] Andersenove bajke *Sanak Sanković* [1].

Intermedijalne poveznice filma *Čovječuljak Snovuljak* i istoimene tiskane slikovnice prisutne su, uz gotovo istovjetne crteže i odabir boja, u vizualnoj komponenti uporabom odabranih filmskih planova i kutova snimanja unutar filmskog okvira, prenesenih na dvostranice slikovnice [11].

Film započinje planom total koji prelazi u polutotal dolaskom *Čovječuljka Snovuljka* na planet Zemlju. Njegovo putovanje i obilazak spavača u središnjem dijelu filma prikazano je srednjim, blizu, krupnim i detalj planovima da bi u završnici još jednom preko polutotala prešlo u plan total prikazom svemira kojim Čovječuljak nastavlja putovanje i odlazi. Od kutova snimanja najčešće je prisutan horizontalni kut te povremeno donji kut snimanja. Pokreti kamere prate Čovječuljka na njegovom putovanju dok se njegov animirani lik kreće na statičnoj pozadini.

Ujedno, kao zvučna komponenta filma, uz glazbu, prisutan je pripovjedačev glas koji iz tzv. *offa* kazuje rimovane stihove teksta [4].

Usporedi li se vizualni diskurs dvostranica slikovnice *Čovječuljak Snovuljak* sa strukturom filmskih kadrova, u njima se mogu prepoznati planovi: total, polutotal, srednji, krupni i blizu. Pri tom pripovijedanje u vizualnom diskursu započinje već na prednjoj strani korica slikovnice prizorom Čovječuljka koji putuje svemirom, nastavlja se u peritekstu na naslovnici koja uz ime i prezime Autorice i naslov slikovnice sadrži i Autoričinu posvetu ocu te: „I za sve tate na svijetu" [5]. U slikovnici se izmjenjuju srednji, blizu i krupni planovi, a dojam perspektive treće dimenzije Ivana Guljašević Kuman postiže kombiniranjem veličine nacrtanih likova na pojedinim stranicama što se može povezati s filmskim dubinskim kadrovima. Zadnje stranice slikovnice sadrže impresum i podatke o Autorici, dopunjene vizualnim diskursom – prizorom lika Čovječuljka koji nastavlja putovanje na sljedećoj, zadnjoj dvostranici, naslikanoj u formi filmskog plana totala.

Verbalni diskurs [11] slikovnice istovjetan je tekstu kazivanom u filmu, te se sastoji od rimovanih četverostihova tiskanih velikim tiskanim slovima [5].

B. Primjena crtanog animiranog filma i slikovnice Čovječuljak Snovuljak u odgojno-obrazovnom procesu

Višemodalno [17] multimedijsko DVD izdanje, uz crtani animirani film *Čovječuljak Snovuljak* sadrži i sljedeće priloge: Crtani film *Little man Dreamsy*, Pričalicu *Čovječuljak Snovuljak* (čitanje teksta slikovnice uz snimljene ilustracije), Animiranu pjesmicu *Čovječuljak Snovuljak*, Pjevaj s *Čovječuljkom Snovuljkom* i Karaoke *Čovječuljak Snovuljak*. Kao što Autorica navodi na omotnici izdanja: „Ovaj DVD sastavni je dio slikovnice ČOVJEČULJAK SNOVULJAK" [4]. Stražnja strana korica slikovnice sadrži Autoričinu napomenu komplementarnu prethodno navedenoj: „Čovječuljak Snovuljak nastao je kao animirani film 2009. godine, a sada se preselio i u slikovnicu nakon uspješnog prikazivanja diljem cijelog svijeta" [5].

U okviru ključnih kompetencija za cjeloživotno učenje, predviđenih Nacionalnim kurikulumom za rani i predškolski odgoj i obrazovanje (2014), ovako strukturiran multimedijski sklop doprinosi razvoju digitalnih kompetencija djeteta te razvoju kompetencija kulturne svijesti i izražavanja poticanjem stvaralačkog izražavanja ideja, iskustva i emocija djeteta u nizu umjetničkih područja koja uključuju glazbu, ples, kazališnu, književnu i vizualnu umjetnost [10].

U radu s učenicima mlađe školske dobi, u okviru nastavnog predmeta Hrvatski jezik te nastavnih područja Književnost i Medijska kultura, moguće je primijeniti filmsko-književni pristup djelima, kao nastavni proces koji se izvodi radi spoznavanja osobitosti dvaju medija: filma i knjige, filmske i književne umjetnosti, te filmsko-likovni pristup radi usporedbe likovnih i filmskih obilježja crtanog animiranog filma i slikovnice [16]. Uz međupodručnu, moguća je i međupredmetna korelacija nastave Hrvatskoga jezika i Likovne kulture.

V. MEDIJSKA KULTURA U DJEČJEM VRTIĆU I OSNOVNOJ ŠKOLI

Temelj rada svake odgojno-obrazovne ustanove te sama njezina srž i polazište nalazi se u kurikulumu usvojenom od strane određene ustanove. Nikola Pastuović kurikulum definira kao organizirani angažman procesa učenja i sadržaja s obzirom na određene svrhe i ciljeve te strukturirani niz željenih ishoda (znanja, vještina, sposobnosti i ponašanja) [15].

Primjerice, Nacionalni kurikulum za rani i predškolski odgoj i obrazovanje (2014) sastoji se od polazišta, načela, vrijednosti i ciljeva koji čine temelj povezanosti Nacionalnog kurikuluma kao službenog dokumenta u Republici Hrvatskoj s individualnim kurikulumom dječjeg vrtića – odgojno-obrazovne koncepcije koja se zajednički razvija (sukonstruira) i korespondira s kvalitetom uvjeta u tom vrtiću. Kao cilj Nacionalnog kurikuluma za rani i predškolski odgoj i obrazovanje također se spominje imperativna važnost razvoja ključnih kompetencija za cjeloživotno učenje među kojima se nalaze i digitalne kompetencije kao sastavni dio upoznavanja djece s medijskom kulturom. Ovaj Dokument razvoj digitalnih kompetencija određuje kao upoznavanje djeteta s informacijsko-komunikacijskom tehnologijom i mogućnostima njezine uporabe u različitim aktivnostima. Nadalje, ističe se važnost organizacije rada u vrtićima kojom su djeci informatička sredstva poput računala dostupna u svakodnevnim aktivnostima i radu pod nadzorom odgajatelja, ali i u slobodi korištenja prema interesima djece. Odgajateljeva je dužnost pripremiti raznovrstan i dobno primjeren sadržaj za djecu iz kojeg će ona imati slobodu biranja sadržaja prema vlastitom interesu [10].

Prema Nastavnom planu i programu za osnovnu školu (2006) ciljevi i zadaće odgoja i obrazovanja u osnovnoj školi usmjereni su na cjeloviti razvoj učenika, što znači da učenike valja poučiti i osposobiti za življenje prema najsuvremenijim spoznajama iz svih područja znanja i učenja što uključuje i područje medijske kulture. Školska knjižnica ističe se kao informacijsko, medijsko i komunikacijsko središte škole te mjesto promicanja načela prema kojima su sloboda mišljenja i slobodan pristup informacijama polazišta za uspješno i odgovorno sudjelovanje u građanskom demokratskom društvu. Drugim riječima, školska knjižnica epicentar je medijske kulture svake odgojno-obrazovne ustanove [12].

Odluka o donošenju kurikuluma za nastavni predmet Hrvatski jezik za osnovne škole i gimnazije u Republici hrvatskoj (2019) navodi da se nastavni predmet Hrvatski jezik sastoji od predmetnih područja: Hrvatskoga jezika i komunikacije, Književnosti i stvaralaštva te Kulture i medija. Predmetno područje Kultura i mediji obuhvaća kritički odnos prema medijskim porukama, razumijevanje utjecaja medija i njihovih poruka na društvo i pojedinca te stvaranje medijskih poruka i njihovo odgovorno odašiljanje [13].

Mediji su sadašnjost i budućnost odgoja i obrazovanja djece stoga što su nepresušan izvor informacija i doživljaja koji formiraju znanje, iskustva, vrijednosti i stav djeteta o svijetu u kojem živi. Zato je iznimno važno da mediji budu dio svakog kurikuluma svake odgojno-obrazovne ustanove. Iz prethodno navedenog može se zaključiti da se dijete od svoje najranije dobi razvija cjelovito u svim odgojno-obrazovnim ustanovama koje pohađa te da se medijima pri tome pridaje velika važnost.

VI. ZAKLJUČAK

Višemodalna multimedijska izdanja slikovnica intermedijalnim komponentama nadograđuju verbalni i vizualni diskurs tiskanih slikovnica, čineći ih dodatno zanimljivima djeci čitateljima. S obzirom na pridodane mogućnosti slušanja teksta i popratne glazbe, pjevanja

uz instrumentalno izvođenje glazbe, doživljavanja animiranih filmova kao svojevrsnih, uvjetno rečeno oživljenih ilustracija – dijete koje čita, sluša, promatra ili pjeva, potaknuto intermedijalnošću slikovnice, ima mogućnost cjelovitog razvoja putem uporabe kvalitetnih medijskih ostvarenja. Intermedijalnost autorskih slikovnica Ivane Guljašević Kuman očituje se u primjeni izražajnih sredstava prisutnih u filmskoj kao i u likovnim umjetnostima (okvir, plan, kut snimanja, narativnost). Svojim umjetničkim dosezima te primjerenim sadržajem i izrazom Autoričina su multimedijska DVD izdanja prikladna za primjenu u odgojno-obrazovnom procesu rada s djecom rane i predškolske, te mlađe školske dobi, uklapajući se ujedno i u zadatosti aktualnih nacionalnih kurikuluma.

LITERATURA

[1] H. Ch. Andersen, „Sanak Sanković", u Zaboravljene bajke, Zagreb: Znanje, 2005., str. 31-44.

[2] Društvo hrvatskih filmskih redatelja, https://dhfr.hr/redatelj/ivana-guljasevic-kuman/ (pristupljeno 10.12.2019.)

[3] M. Gabelica, „Dječja književnost Y generacije", u Hrvatski iseljenički zbornik 2015. (e-izdanje). Zagreb: Hrvatska matica iseljenika, 2014, str. 61-74.

[4] I. Guljašević Kuman (rež.), Čovječuljak Snovuljak, animirani film, DVD audiovizualna građa. Zagreb: IGUBUKA D.o.o., 2009.

[5] I. Guljašević Kuman, Čovječuljak Snovuljak. Zagreb: IGUBUKA D.o.o., 2012.

[6] Hrvatska enciklopedija Leksikografskog zavoda Miroslav Krleža, http://www.enciklopedija.hr/Natuknica.aspx?ID=27638 (pristupljeno: 10.12.2019.)

[7] Hrvatsko društvo književnika za djecu i mlade, http://www.hdkdm-klubprvihpisaca.hr/guljasevic-ivana/ (pristupljeno 10.12.2019.)

[8] Igubuka, http://www.igubuka.hr/page/read/id/3 (pristupljeno 10.12.2019.)

[9] K. Mikić, Film u nastavi medijske kulture. Zagreb: Educa, 2001.

[10] Nacionalni kurikulum za rani i predškolski odgoj i obrazovanje. Zagreb: Republika Hrvatska, Ministarstvo znanosti, obrazovanja i sporta, 2014.

[11] S. Narančić Kovač, Jedna priča – dva pripovjedača. Zagreb: Artresor naklada, 2015.

[12] Nastavni plan i program za osnovnu školu. Zagreb: Republika Hrvatska, Ministarstvo znanosti, obrazovanja i športa, 2006.

[13] Odluka o donošenju kurikuluma za nastavni predmet Hrvatski jezik za osnovne škole i gimnazije u Republici hrvatskoj, Zagreb: Narodne novine, 10/2019 (29.1.2019.), https://narodne-novine.nn.hr/clanci/sluzbeni/full/2019_01_10_216.html (pristupljeno 3.9.2019.)

[14] D. Oraić Tolić, Citatnost u književnosti, umjetnosti i kulturi. Zagreb: Ljevak, 2019.

[15] N. Pastuović, Edukologija: integrativna znanost o sustavu cjeloživotnog obrazovanja i odgoja. Zagreb: Znamen, 1999.

[16] S. Težak, Metodika nastave filma na općeobrazovnoj razini. Zagreb: Školska knjiga, 2002.

[17] D. Težak i M. Gabelica, „Digitalna dječja književnost", u Modern Approaches to Teaching Coming Generation, International conference. Ljubljana: EDUvision, Stanislav Jurjevčič s.p., 2015., str. 856-866.

[18] B. Trkulja, „Šareni svjetovi dobroćudnih stvorenja, Filmovi Ivane Guljašević", u Zapis, bilten Hrvatskog filmskog saveza, br. 74, 2012., http://www.hfs.hr/nakladnistvo_zapis_detail.aspx?sif_clanci=34140#.Xi19Foh7mM8 (pristupljeno 5.1.2020.)

Predikcija uspjeha u studiranju primjenom Bayesovih mreža

D. Šimić* i J. Gusić*
* Faculty of Organization and Informatics University of Zagreb, Varaždin, Croatia
diana.simic@foi.unizg.hr, jelena.gusic@foi.unizg.hr

***Sažetak* - Analizirani su podaci o uspješnosti studiranja i prolasku ispita na prvoj godini preddiplomskog studija za 8 generacija studenata. Ciljevi istraživanja bili su: napraviti prediktivni model koji će omogućiti identifikaciju studenata za koje postoji velika vjerojatnost da neće ostvariti 30 ECTS bodova tijekom akademske godine i pružiti studentima informaciju o vjerojatnosti prolaza na pojedinom ispitu, odnosno postizanja ciljanog broja ECTS-a na kraju akademske godine. Za modeliranje združene distribucije korištena je Bayesova mreža. Struktura mreže sadržavala je kao čvorove rezultate ispita pojedinih predmeta (nije upisan, nije položen, položen), a ishod akademske godine reprezentiran je kvalitativnom varijablom s vrijednostima „manje od 30 ECTS-a", „od 30 do 54 ECTS-a" i „55 i više ECTS-a". Za učenje strukture i parametara korišten je paket bnlearn u R-u. Provedena je 20-struka unakrsna validacija. Rezultati su pokazali da su mjera F_1 i odziv najbolji za najuspješnije studente (F_1 između 0,81 i 0,91; odziv između 0,91 i 1,0). Specifičnost je bila najbolja za dvije grupe studenata koji su ostvarili manje od 55 ECTS-a (0,81 do 0,98). Točnost predikcije u unakrsnoj validaciji bila je između 69% i 87%. Za najuspješniju grupu studenata točnost predikcije bila je između 0,79 i 0,92.**

Ključne riječi* - *analitika učenja; Bayesove mreže; predikcija uspjeha studiranja

I. UVOD

Analitika učenja novija je znanstvena disciplina koja se bavi primjenom suvremenih metoda podatkovne znanosti, statističkog učenja i umjetne inteligencije na podatke iz obrazovnog procesa u svrhu boljeg razumijevanja i unapređivanja procesa podučavanja i učenja [1]. Područje primjene analitike učenja vrlo je široko, što je vidljivo već i iz samog sadržaja zbornika međunarodne konferencije o analitici učenja i znanju koja je po deveti puta održana prošle godine (LAK'19) [2]. Između ostalog analitika učenja koristi se za razvoj kurikula i silabusa, automatizirano vrednovanje, praćenje rada studenata, analizu čimbenika koji utječu na duljinu ili uspješnost studiranja.

U posljednje vrijeme poboljšanje uspješnosti studiranja profilirano je kao jedan od važnijih ciljeva koje se želi postići primjenom analitike učenja [3]. Razvijaju se modeli za identificiranje čimbenika koji najviše doprinose uspješnosti i za ranu identifikaciju studenata „u riziku" (tj. onih za koje je vrlo vjerojatno da neće položiti neki od predmeta, neće upisati sljedeću akademsku godinu ili će odustati od studija)[3]–[10]. Pravovremena identifikacija „studenata u riziku" jako je bitna jer otvara prostor za brzu reakciju koja može pomoći tim studentima da savladaju svoje poteškoće u učenju i poboljšaju ishode učenja. Njihov napredak će onda doprinijeti i napretku obrazovne institucije čiji su članovi. Sve navedeno je dovelo do toga da se sve više primjenjuje metode rudarenja velikih skupova edukacijskih podataka (eng. *educational data mining*, EDM) i analitika učenja (eng. *learning analytics*) ([7], [11]–[13]).

Abu Saa i sur. [11] prikazuju pregled metoda korištenih za predviđanje uspješnost studenata u visokom obrazovanju. Analizirali su detaljno 36 radova u razdoblju od 2009. do 2018. godine. Kao najčešće značajke koje su korištene za predviđanje uspješnosti studenata u visokom obrazovanju identificirali su: uspjeh koji je student ostvario na kolegijima iz prethodnog semestra, akademske godine i/ili iz osnovne i srednje škole; aktivnosti na sustavima za e-učenje (kao što su broj registracija u sustav, broj riješenih zadaća, kvizova, broj poruka koje su slali jedni drugima kroz sustav i sl.); demografske karakteristike (kao što su spol, dob, nacionalnost i sl.) te informacije iz društvenog života (kao što je broj prijatelja, je li student pušač ili ne i sl.). Što se tiče metoda strojnog učenja koje su korištene, najčešće se spominju stabla odlučivanja (eng. *decision trees*), naivni Bayesovi klasifikatori (eng. *Naïve Bayes classifiers*) te neuronske mreže (eng. *neural networks*). Nema puno radova u kojima se koriste Bayesove mreže.

Jako je bitno već na prvoj godini identificirati studente u riziku, tj. studente koji neće ostvariti uvjete za upis u višu godinu. Lacave i sur. [14] proveli su istraživanje koje značajke su najčešće povezane s odustajanjem od studija. To je jedno od rijetkih istraživanja koje koristi Bayesove mreže, a u predikciji osim akademskih značajki (trenutna godina studija, najčešća ocjena, broj položenih predmeta i sl.) koristili su i demografske značajke (spol, dob kod upisa na studij, obrazovanje roditelja, je li student zaposlen ili ne i sl.). Iz njihovog modela slijedi da je odluka o odustajanju od studija najčešće povezana s godinom u koju je student upisan te brojem položenih predmeta.

U Hrvatskoj su redoviti studenti oslobođeni plaćanja participacije školarine. Da bi zadržali pravo na oslobođenje od plaćanja participacije školarine nakon prve godine studija, studenti moraju ostvariti najmanje 55 ECTS bodova. Studenti koji ostvare između 30 i 54 ECTS boda sudjeluju u plaćanju participacije, dok oni koji ostvare manje od 30 ECTS bodova plaćaju puni iznos participacije školarine. S obzirom na to da povećanje troškova studiranja može biti razlog za odustajanje od studija, važno je što ranije prepoznati studente koji su „u riziku" da ne ostvare dovoljan broj ECTS bodova. Pojačan rad s takvim studentima može smanjiti rizik neuspjeha, odnosno odustajanja od studija. Ciljevi ovog istraživanja su sljedeći:

- Izraditi model za predikciju koji će identificirati studente prve godine „u riziku“ (studenti za koje postoji velika vjerojatnost da neće ostvariti 30 ECTS-a u toku akademske godine) koristeći podatke o uspješnosti polaganja ispita.
- Na temelju izrađenog prediktivnog modela, informirati studente o vjerojatnosti prolaza na pojedinom ispitu kao i o vjerojatnosti postizanja željenog broja ECTS bodova na kraju akademske godine.

Ostatak rada organiziran je na sljedeći način. U poglavlju II opisani su korišteni podaci i metode. Slijedi poglavlje III u kojem su prikazani rezultati učenja i validacije prediktivnog modela. U završnom IV. Poglavlju prikazani su rasprava i zaključak.

II. MATERIJALI I METODE

A. Podaci

Prikupljeni su podaci o stanju položenih ispita na kraju prve godine preddiplomskog studija Informatike u akademskim godinama od 2008/2009 do 2015/2016. Ukupno je bilo 1800 studenata s podacima o prolaznosti na 15 predmeta i ukupnom broju stečenih ECTS bodova. Prolaznost na predmetu kodirana je kao „Položen“ ili „Nije položen“, a za izborne predmete i „Nije upisan“. Varijabla ukupan broj ECTS bodova podijeljena je u tri kategorije: broj ECTS bodova manji od 30, broj ECTS bodova od 30 do 54 i broj ECTS bodova barem 55. Od 15 predmeta 11 je obaveznih i 4 izborna, od kojih se dio održava u zimskom, a dio u ljetnom semestru.

B. Metode

S obzirom da su neki od predmeta prvog semestra preduvjet za upis / polaganje predmeta drugog semestra, nije bilo moguće koristiti neki od naivnih Bayesovih klasifikatora jer je preduvjet za njihovu primjenu nezavisnost značajki koje se koriste za predikciju [15]. Stoga je za modeliranje združene distribucije analiziranih značajki korištena Bayesova mreža.

U analizi podataka korišten je jezik R [16] i integrirano razvojno okruženje RStudio [17]. Za učenja strukture i parametara Bayesove mreže, vizualizaciju te predikciju korišteni su paketi bnlearn [18] i Rgraphviz [19]. Za evaluaciju modela korišten je paket mltest [20].

U procesu učenja strukture moguće je definirati koji bridovi moraju biti prisutni u mreži (tzv. bijela lista), odnosno koji bridovi ne smiju biti prisutni u mreži (tzv. crna lista).

Bijela lista zadana je tako da:

- postoji brid iz svakog od predmeta prema broju ECTS-a te
- postoji brid iz svakog predmeta prvog semestra prema onim predmetima drugog semestra kojima je prvi predmet preduvjet.

Bridovi koji predstavljaju preduvjete na grafičkom prikazu Bayesove mreže prikazani su crvenim crtkanim linijama.

Crna lista zadana je tako da su zabranjeni bridovi iz broja ECTS bodova prema predmetima kao i bridovi iz predmeta ljetnog semestra prema predmetima zimskog semestra. Struktura Bayesove mreže naučena je metodom „grow-shrink“ [21].

Za procjenu točnosti predikcije modela korištena je 20-struka unakrsna validacija s pogreškom klasifikacije (eng. *expected loss*) kao funkcijom cilja. Za svaki testni skup izračunate su uobičajene mjere uspješnosti klasifikacije, kao što su: točnost (eng. *accuracy*), specifičnost (eng. *specificity*), preciznost (eng. *precision*), osjetljivost ili odziv (eng. *sensitivity, recall*) te F_β skor [22]. F_β skor se najčešće računa uz $\beta = 1$ i tada predstavlja harmonijsku sredinu preciznosti i odziva. Tako je računat i u ovom radu. Ako se provodi klasifikacija u više klasa, mjere uspješnosti klasifikacije se mogu računati za svaku klasu posebno.

III. REZULTATI

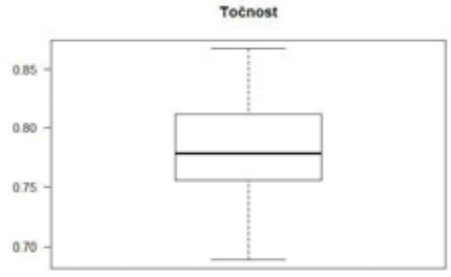
Studenti nisu bili ravnomjerno raspodijeljeni u tri grupe s obzirom na ostvareni broj ECTS-a. Najveći udio činili su studenti koju su ostvarili barem 55 ECTS bodova (44,8%), sljedeći su bili studenti koji su ostvarili od 30 do 54 ECTS boda (28,6%), a najmanji udio činili su studenti koji su ostvarili manje od 30 ECTS bodova (26,6%).

Procijenjena očekivana stopa pogreške klasifikacije u 20-strukoj unakrsnoj validaciji bila je 21,8%. Struktura naučene Bayesove mreže bila je stabilna. Od 20 Bayesovih mreža, samo se u jednom slučaju struktura naučene Bayesove mreže razlikovala i to na način da je sadržavala jedan brid više od ostalih 19 Bayesovih mreža (brid iz Matematike 1 prema Teoriji sustava).

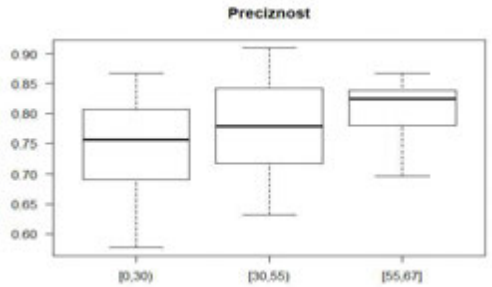
Kao što je vidljivo na Slika 1, točnost klasifikacije se kretala između 69% i 87%, dok je aritmetička sredina točnosti klasifikacije bila 78,2%. S obzirom na to da je važno i kakva je pogreška klasifikacije u podgrupama studenata, ukupna točnost klasifikacije nije prikladna mjera u ovoj situaciji. U nastavku su prikazane distribucije preciznosti, odziva, F_1 i specifičnosti iz unakrsne validacije.

Slika 2 predstavlja distribuciju preciznosti (proporciju dobro klasificiranih objekata u skupu svih objekata klasificiranih u neku klasu). Preciznost je najniža upravo u grupi studenata s najlošijim rezultatom, a najbolja u grupi studenata s najboljim rezultatom.

Slika 3 predstavlja distribuciju odziva (proporcije dobro klasificiranih objekata među onima koji doista pripadaju pojedinoj klasi). Odziv je, slično kao preciznost, najveći u



Slika 1. Grafički prikaz distribucije ukupne točnosti klasifikacije u 20 unakrsnih validacija modela



Slika 2. Grafički prikaz distribucije preciznosti po klasama u 20 unakrsnih validacija modela

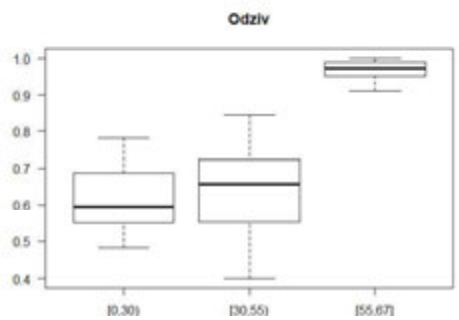
klasi najuspješnijih studenata dok je u preostale dvije klase podjednak.

S obzirom da preciznost i odziv pokazuju sličan obrazac vrijednosti u odnosu na uspješnost studenata, i F_1, kao njihova harmonijska sredina, prikazuje isti trend (Slika 4).

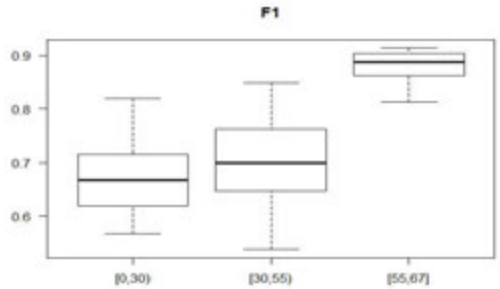
Slika 5 predstavlja distribuciju specifičnosti (proporcija svih objekata koji su dobro klasificirani u skupu svih objekata koji nisu klasificirani u neku klasu – stvarno negativni). Za razliku od F_1 skora i preciznosti, specifičnost je najmanja za najuspješnije studente i povećava se smanjenjem uspješnosti studenata te je najveća za klasu studenata koji su najneuspješniji, tj. koji su u riziku.

Uz jednake postavke i metode za učenje strukture i parametara Bayesove mreže kao kod unakrsne validacije, struktura naučene Bayesove mreže na cijelom skupu podataka prikazana je na Slika 6. Bayesova mreža naučena na cijelom skupu podataka jednaka je kao 19 Bayesovih mreža dobivenih postupkom 20-struke unakrsne validacije. Žutom bojom označeni su izborni predmeti, a crvenim isprekidanim linijama bridovi iz predmeta zimskog semestra prema predmetima ljetnog semestra kojima su preduvjeti. Uz bridove uključene na bijelu listu u strukturi Bayesove mreže pojavljuju se i sljedeći bridovi među predmetima: brid iz Osnova ekonomije prema Organizaciji, iz Poslovnog komuniciranja prema Komuniciranju u organizaciji, iz Komuniciranja u organizaciji i iz Teorije sustava prema Informatici 2 te iz Programiranja 1 prema Obradi teksta i slike.

Bayesova mreža predstavlja reprezentaciju združene distribucije svih analiziranih značajki. Stoga je moguće postavljati upite o uvjetnoj distribuciji pojedine značajke uz zadane vrijednosti nekih ili svih ostalih značajki. Npr. možemo postaviti upit koja je uvjetna distribucija broja



Slika 3. Grafički prikaz distribucije odziva po klasama u 20 unakrsnih validacija modela



Slika 4. Grafički prikaz distribucije mjere F1 po klasama u 20 unakrsnih validacija modela

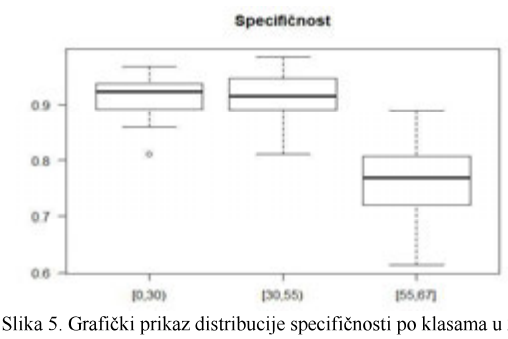
ECTS bodova za studenta za kojeg znamo da je položio neki skup predmeta. Primjer takvog upita u R-u je:

```
> cpquery(la02.gs.fit,
  event=(ects_f=="[30,55)" |
         ects_f=="[55,67]"),
  evidence = list(inf1 = "Passed",
                  math2 = "Not passed"),
  method = "lw")
[1] 0.6769102
```

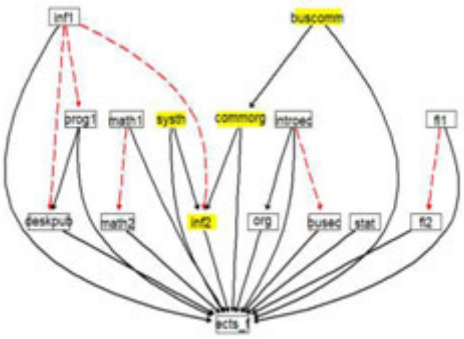
Postavljen je upit kolika je vjerojatnost da student koji je položio Informatiku 1, a do kraja akademske godine nije položio Matematiku 2 stekne više od 30 ECTS bodova. Procijenjena uvjetna vjerojatnost iznosi 0,68. Takve studente ne bi se klasificiralo među studente u riziku.

IV. ZAKLJUČAK

Cilj ovog istraživanja bio je izraditi model za predikciju koji će identificirati „studente u riziku“ (one za koje postoji velika vjerojatnost da neće ostvariti 30 ECTS-a u toku akademske godine). Za modeliranje združene distribucije značajki korištena je Bayesova mreža. Rezultati 20-struke unakrsne validacije pokazali su da su odziv i F_1 bili izrazito visoki za najuspješnije studente. Više od 90% studenata koji su postigli više od 54 ECTS boda su ispravno klasificirani. S druge strane, zato je specifičnost bila bolja upravo za manje uspješne studente, jer specifičnost za jednu klasu odgovara odzivu za drugu klasu. To se može interpretirati na sljedeći način: među onima koje model nije klasificirao u klasu studenata koji su ostvarili manje od 30 ECTS bodova doista ima malo studenata koji su neuspješni i spadaju među studente u riziku. To znači da je vjerojatnost da propustimo identificirati studente u riziku zadovoljavajuće mala, bez obzira na ukupnu točnost



Slika 5. Grafički prikaz distribucije specifičnosti po klasama u 20 unakrsnih validacija modela



Slika 6. Naučena struktura Bayesove mreže. Crvene isprekidane linije predstavljaju bridove između predmeta i njegovog preduvjeta, a žuti čvorovi predstavljaju izborne predmete.

predikcije. Preciznost je, slično odzivu, bila bolja za uspješnije studente. Kad bismo proveli slučajnu alokaciju studenata u klase uz jednaku vjerojatnost klasifikacije u svaku od klasa, preciznost bi za svaku klasu varirala oko proporcije studenata koji pripadaju toj klasi u ukupnom uzorku. Preciznost iz unakrsne validacije za svaku je klasu znatno viša od očekivane temeljem takve slučajne alokacije.

U prethodnim istraživanjima primjenom raznih tehnika i skupova značajki raspon točnosti predikcije kretao se između 66% i čak 97% [3], [6], [23]. Hussain i sur. [23] karakteriziraju modele koji su imali točnost predikcije od 85% kao modele s visokom točnošću. Točnost predikcije našeg modela u unakrsnoj validaciji kretala se između 69% i 87%, a u prosjeku je iznosila 78,2% te ju možemo smatrati dobrom i usporedivom s prethodnim istraživanjima. Slično vrijedi i za ostale mjere prikazane u ovom radu (preciznost, odziv, specifičnost i F_1).

Prednost Bayesovih mreža, u odnosu na ostale prediktivne modele, je da one predstavljaju reprezentaciju združene distribucije svih korištenih značajki. Za svaku značajku je stoga moguće procijeniti uvjetnu distribuciju uz zadavanje svih ili samo nekih vrijednosti drugih značajki. Stoga je Bayesove mreže moguće koristiti i za dopunjavanje podataka koji nedostaju. Iako strukturu Bayesove mreže možemo učiti iz podataka, pri učenju možemo koristiti i ekspertno znanje. U ovom radu koristili smo podatke o vremenskom redoslijedu izvođenja predmeta i preduvjetima pri učenju strukture mreže.

Buduća istraživanja ići će u smjeru proširenja skupa značajki, modeliranja vremenskog slijeda polaganja ispita te primjene drugih prediktivnih modela.

LITERATURA

[1] C. Lang, G. Siemens, A. Wise, and D. Gasevic, Eds., *Handbook of Learning Analytics*, First. Society for Learning Analytics Research (SoLAR), 2017.

[2] D. Azcona and R. Chung, Eds., *LAK 19: proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK'19) : Learning Analytics to Promote Inclusion and Success : March 4-8, 2019, Tempe, Arizona, United States*. 2019.

[3] V. Nida Uzel, S. Sevgi Turgut, and S. Ayse Ozel, "Prediction of Students' Academic Success Using Data Mining Methods," in *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Adana, 2018, pp. 1–5, doi: 10.1109/ASYU.2018.8554006.

[4] A. I. Adekitan and E. Noma-Osaghae, "Data mining approach to predicting the performance of first year student in a university using the admission requirements," *Educ. Inf. Technol.*, vol. 24, no. 2, pp. 1527–1543, Mar. 2019, doi: 10.1007/s10639-018-9839-7.

[5] A. F. Agudo-Peregrina, S. Iglesias-Pradas, M. Angel Conde-Gonzalez, and A. Hernandez-Garcia, "Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning," *Comput. Hum. Behav.*, vol. 31, pp. 542–550, Feb. 2014, doi: 10.1016/j.chb.2013.05.031.

[6] D. Bañeres and M. Serra, "On the Design of a System to Predict Student's Success," in *Complex, Intelligent, and Software Intensive Systems*, vol. 611, L. Barolli and O. Terzo, Eds. Cham: Springer International Publishing, 2018, pp. 274–286.

[7] Y. Cui, F. Chen, A. Shiri, and Y. Fan, "Predictive analytic models of student success in higher education: A review of methodology," *Inf. Learn. Sci.*, vol. 120, no. 3/4, pp. 208–227, Mar. 2019, doi: 10.1108/ILS-10-2018-0104.

[8] B. Dietz-Uhler and J. E. Hurn, "Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective," *J. Interact. Online Learn.*, vol. 12, no. 1, pp. 17–26, SPR 2013.

[9] S. Kotsiantis, N. Tselios, A. Filippidi, and V. Komis, "Using learning analytics to identify successful learners in a blended learning course," *Int. J. Technol. Enhanc. Learn.*, vol. 5, no. 2, pp. 133–150, 2013, doi: 10.1504/IJTEL.2013.059088.

[10] L. Zhuhadar, J. Daday, S. Marklin, B. Kessler, and T. Helbig, "Using survival analysis to discovering pathways to success in mathematics," *Comput. Hum. Behav.*, vol. 92, pp. 487–495, Mar. 2019, doi: 10.1016/j.chb.2017.12.016.

[11] A. Abu Saa, M. Al-Emran, and K. Shaalan, "Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques," *Technol. Knowl. Learn.*, vol. 24, no. 4, pp. 567–598, Dec. 2019, doi: 10.1007/s10758-019-09408-7.

[12] T. Devasia, Vinushree T P, and V. Hegde, "Prediction of students performance using Educational Data Mining," in *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, Ernakulam, India, 2016, pp. 91–95, doi: 10.1109/SAPIENCE.2016.7684167.

[13] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telemat. Inform.*, vol. 37, pp. 13–49, Apr. 2019, doi: 10.1016/j.tele.2019.01.007.

[14] C. Lacave, A. I. Molina, and J. A. Cruz-Lemus, "Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks," *Behav. Inf. Technol.*, vol. 37, no. 10–11, pp. 993–1007, Nov. 2018, doi: 10.1080/0144929X.2018.1485053.

[15] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," *Mach. Learn.*, vol. 29, no. 2/3, pp. 131–163, 1997, doi: 10.1023/A:1007465528199.

[16] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2019.

[17] RStudio Team, *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc., 2018.

[18] M. Scutari, "Learning Bayesian Networks with the bnlearn R Package," *J. Stat. Softw.*, vol. 35, no. 3, pp. 1–22, 2010, doi: 10.18637/jss.v035.i03.

[19] K. D. Hansen *et al.*, *Rgraphviz: Provides plotting capabilities for R graph objects*. 2019.

[20] G. Dudnik, *mltest: Classification Evaluation Metrics*. 2018.

[21] D. Margaritis and S. Thrun, "Bayesian Network Induction via local Neighborhoods," in *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. K. Leen, and K. R. Müller, Eds. MIT Press, 2000, pp. 505–511.

[22] P. Škrabánek and P. Doležel, "ON REPORTING PERFORMANCE OF BINARY CLASSIFIERS," *Sci. Pap. Univ. Pardubice Ser. Fac. Econ. Adm.*, vol. 41, 2017.

[23] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, "Using machine learning to predict student difficulties from learning session data," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 381–407, Jun. 2019, doi: 10.1007/s10462-018-9620-8.

Modifikacija arkadnih igara *Space Invaders* i *Super Mario* u edukativne inačice za učenje matematike i hrvatskog jezika

I. Franković, M. Ivašić-Kos
Odjel za informatiku, Sveučilište u Rijeci, Rijeka, Hrvatska
ifrankovic@inf.uniri.hr, marinai@uniri.hr

***Sažetak* – Novi trendovi obrazovanja nastoje iskoristiti naviku učenika da se redovito igraju na različitim platformama i primijeniti motivacijske elemente digitalnih igara u obrazovanju kako bi se povećala motivacija, zainteresiranost i usredotočenost učenika kod učenja i utvrđivanja nastavnih sadržaja. Računalne igre koje se koriste u obrazovnom kontekstu nazivaju se edukativne igre i dizajnirane su u skladu s ishodima učenja kako bi pomogle učenicima da nauče određeno gradivo, da prošire znanje o određenim konceptima ili da im olakšaju usvajanje.**

Visok stupanj angažiranosti učenika koji se javlja tijekom igranja igara smatra se poželjnim za stvaranje plodnog okruženja za učenje, međutim pokazalo se da edukativne igre učenicima nisu jednako zanimljive niti zabavne kao i komercijalne igre, ne osiguravanju jednaku emociju, adrenalin, uključenost, interakciju i motivaciju za igranjem. Kako bi se prevladala razlika između edukativnih i komercijalnih igara, predlaže se prilagodba poznatih igara, s kojima su se učenici navikli igrati, u njihove edukativne inačice. U ovom radu će se predstaviti prototipovi edukativnih verzija poznatih igara *Super Mario* i *Space Invaders*, koji su namijenjeni učenicima mlađih razreda osnovne škole za vježbanje jednostavnih zadataka iz matematike ili hrvatskog jezika.

Ključne riječi – edukativne igre; učenje uz pomoć igara; edukativne inačice poznatih igara

I. UVOD

Današnje generacije učenika odrasle su u potpuno digitaliziranom i neprestano povezanom svijetu. Konstantno su okruženi novim digitalnim tehnologijama poput računala, računalnih konzola, pametnih telefona te ostalih uređaja digitalnog doba.

To su generacije učenika koje su upoznate s digitalnom tehnologijom od rane dobi pa im je informacijska i komunikacijska tehnologija postala „materinji jezik“, tj. jezik pomoću kojeg komuniciraju, izražavaju se i doživljavaju svijet oko sebe [1]. I u slobodno vrijeme, za razonodu i zabavu provode brojne sate igrajući neku vrstu igre na računalima i mobitelima ili komunicirajući na društvenim mrežama. Za razliku od prethodnih generacija, postali su više vizualni tipovi te je to imalo utjecaja i na način komunikacije. Više komuniciraju vizualnim putem – pomoću videa, slika, emotikona i GIF-ova [2]. Sve se ranije susreću s informacijskom i komunikacijskom tehnologijom i razina njihovog znanja povezanog s digitalnim tehnologijama je sve viša.

Radi velikog utjecaja i upotrebe digitalnih tehnologija, nove generacije učenika odlično prikupljaju i istodobno koriste više izvora informacija integrirajući virtualna i offline iskustva, žele čestu i brzu interakciju sa sadržajem te imaju dobre vizualne sposobnosti [3]. Takvo okruženje i količina interakcije promijenili su način učenja pa je to bio jedan od motiva da se postojeće tradicionalne metode učenja promijene i prilagode.

U suvremenom obrazovanju u središte se stavljaju učenici što zahtjeva prelazak s tradicionalnog prenošenja informacija na formu aktivnog učenja kod koje je naglasak na sudjelovanju [4]. Time se značajno mijenja uloga učitelja te umjesto tradicionalnog, didaktičkog podučavanja, učitelj priprema suvremeno okruženje i zadatke za samostalno učenje kroz koje vodi učenike i daje konstruktivne povratne informacije [5]. Kao posljedica toga javio se novi način učenja uz pomoć računala. Računalo ne može sasvim zamijeniti klasični način učenja u školama, ali istraživanja su pokazala da ga dobro nadopunjuje i proširuje [6].

Obzirom da djeca veliki dio svog vremena provode igrajući se na računalu, poseban interes usmjeren je ka učenje uz pomoć računalnih igara (engl. Game Based Learning – GBL) [3][7] te razvoju računalnih igara koje podržavaju učenje i podučavanje (tzv. edukativne igre). Cilj svake igre je postići zadani cilj, izbjegavajući prepreke i pridržavajući se definiranih pravila pa se nastojalo iskoristiti ta svojstva igre kao i kreativnost, motivaciju i angažman koji pobuđuju kod učenika u obrazovne svrhe.

Edukativne igre moguće je primjenjivati u svim etapama nastavnog procesa te se koriste kao motivacija učenika u uvodnom dijelu sata ili prilikom ponavljanja i utvrđivanja gradiva [24]. Igra u nastavi može biti i simulacija koja uključuje problem rješavanja određenog zadatka koji je izazovan za učenika [8].

U ovome radu opisuje se modifikacije poznatih igara *Space Invaders* i *Super Mario* u edukativne inačice za učenje matematike i hrvatskog jezika. U drugom poglavlju daje se definicija i karakteristike edukativnih igara te se prezentiraju poznate edukativne igre i edukativne inačice poznatih igara. U trećem poglavlju predstavljen je zadatak da se iskoristi „gameplay“ originalnih igara i prenamijeni igre u edukativne inačice dan na kolegiju Objektno orijentirano programiranje studentima Odjela za informatiku Sveučilišta u Rijeci. U četvrtom poglavlju

predstavljen je edukativni prototip igre *Super Mario*, a u petom poglavlju pet edukativnih prototipova igre *Space Invaders*. Rad završava s zaključkom i smjernicama za daljnji razvoj i prilagodbu igara u edukativne svrhe.

II. EDUKATIVNE IGRE I EDUKATIVNE INAČICE POZNATIH IGARA

Edukativne igre se mogu definirati kao interaktivne, kompetitivne lekcije s definiranim ishodima učenja koje učeniku dopuštaju zabavu tijekom stjecanja znanja. Njihov cilj nije isključivo zabava, već sadrže edukativnu komponentu. Moguće ih je razviti imajući u vidu različite ciljeve: – poticanje učenja, razvijanje kognitivnih vještina ili kao simulacije koje omogućuju stjecanje vještina u virtualnom okruženju [9].

Pojam edukativne računalne igre se odnosi na računalne igre koje osim obrazovnih institucija koriste i zdravstvene, vojne, industrijske i poslovne organizacije [10]. Edukativne računalne igre u nastavi (engl. Serious Games) prvenstveno služe za učenje i usvajanje gradiva na zanimljiviji i zabavniji način od klasičnog načina učenja. One motiviraju učenike tako da kroz igru uče, ponavljaju i rješavaju određene zadatke. Pritom je potrebno paziti da su prilikom izrade edukativne računalne igre ispunjeni svi ishodi učenja koji su prethodno postavljeni kako bi gradivo odgovaralo nastavnoj temi/cjelini. Svaka edukativna računalna igra mora imati dobro definirane ciljeve učenja i promicati razvoj važnih strategija i vještina kako bi se povećale kognitivne i intelektualne sposobnosti učenika [11]. Važni elementi koji doprinose obrazovnim vrijednostima igre su stimulativni poticaji, fantazija, izazov i radoznalost [11].

Neke od najpopularniji edukativnih igara su Minecraft: Education Edition, Lightbot, Run Marco!, Codemonkey. Minecraft: Education Edition [12] je inačica popularne igre koja je prilagođena za učenje u virtualnim 3D svjetovima, a omogućuje rješavanje problema i suradnju učenika. Edukativne igre Lightbot [13], RunMarco [14] i CodeMonkey [15] namijenjene su za učenje koncepata programiranja i u njima učenici moraju slagati blokove kôda kako bi se glavni lik kretao po zadanoj putanji. Problem kod ovih igara je što učenici takav tip igre ne doživljaju na isti način kao i igre namijenjene za zabavu te im nisu jednako zanimljive [16].

Istraživanja su pokazala da edukativne inačice poznatih igara mogu biti učinkovite u mnogim domenama. Rad [17] prikazuje edukativne inačice igara Tetris, Zmija, Heksagon i Space Invaders. Sve opisane igre namijenjene su učenju redoslijeda izvođenja aktivnosti. U igrama se prikazuje npr. radnja ispijanja kave koja uključuje aktivnost „A“ (točenje kave) koja se mora dogoditi prije akrivnosti „B“ (pijenje kave), igre moraju primijeniti formalno pravilo da se aktivnosti izvrše točno tim redoslijedom. Rezultati istraživanja su pokazali da su bolje rezultate postigli učenici koji su koristili igre u nastavi u odnosu na učenike koji su gradivo trebali savladati tradicionalnim poučavanjem. Istraživanje u radu [18] predstavlja edukativne inačice križaljke i Duck Hunt (igre pucačine) za učenje uvoda u programiranje. Pokazalo se da je učenicima sadržaj bio zanimljiviji ako se prikazao kroz igre. Još jedna edukativna inačica poznate igre je *Prog&Play* [19]. Namijenjena je za vježbanje programskog jezika C++, a temelji se na igri Kernel Panic koja pripada strateškim igrama u realnom vremenu (engl. Real-time strategic game, RTS).

III. MODIFIKACIJA POZNATIH IGARA U EDUKATIVNE INAČICE

U ovom radu odlučili smo predstaviti mogućnosti transformacije dviju vrlo popularnih i poznatih igara različitih žanrova u edukativne inačice. Kriterij za odabir igara nije bila samo njihova popularnost već i njihova igrivost, mali broj kontrola za upravljanje likom, jednostavna pravila i mogućnost pokretanja bez velikih računalnih ili memorijskih resursa.

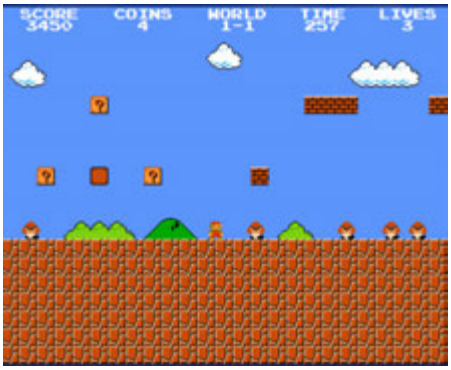
Studenti Odjelu za informatiku, Sveučilišta u Rijeci su dobili zadatak na kolegiju Objektno orijentirano programiranje da iskoriste samo „gameplay“ originalne igre i prenamijene ga u edukativnu igru za savladavanje ili vježbanje nastavnih sadržaja osnovne škole.

Cilj zadataka bio je dvojak. Najprije da studenti na konkretnom projektu primijene principe objektne paradigme kao što su klasa, objekt, ponašanje i svojstva objekta, nasljeđivanje, nadjačavanje i polimorfizam te na taj način uvježbaju i prošire stečene vještine iz programiranja. A drugi da razviju edukativnu igru koja će učenicima nižih razreda osnovne škole savladavanje i uvježbavanje gradiva učiniti zabavnijim i izazovnijim.

Studenti su osmislili priču, razvili prototip igre koristeći Unity game engine [22] i izradili dokumentaciju u skladu s predloškom GDD (Game Design Documet) dokumenta koji ima je zadan. Prilagodili su dizajn i napravili nove likove u igri, po potrebi koristeći Blender [23]. Igre su uglavnom izrađivane u paru.

IV. MODIFIKACIJA PLATFORMSKE IGRE SUPER MARIO U EDUKATIVNE INAČICE

Prva ponuđena igra je Super Mario [20], igra platforme koju je 1985. godine razvio Nintendo. Igre platforme su tip računalnih igara i podtip akcijske igre u kojima igrač kontrolira glavnog lika koji se kreće po platformama kako bi napredovao kroz igru. Pri tom mora skakati i penjati se između razdvojenih platformi, izbjegavajući postavljene prepreke. Cilj igre Super Mario (Slika 1) je da glavni lik Mario prođe kroz kraljevstvo gljiva i spasi princezu. Igrač ima određeni broj života i tijekom igre mora sakupljati novčiće i izbjegavati neprijatelje kako ne bi izgubio život. Igra završava kada Mario spasi princezu ili kada izgubi sve živote.



Slika 1. Originalna igra *Super Mario*

A. *Igra Princ matematike*

Igra Princ matematike je edukativna inačice igre Super Mario. Namijenjena je učenicima 2 razreda osnovne škole za vježbanje jednostavnih zadataka iz matematike – razlikovanje parnih i neparnih brojeva, te višekratnika broja tri. Igra prikazuje mladog princa koji želi pronaći izgubljeno zlato. Kako bi došao do cilja, mora proći kroz tri razine rješavajući zadane matematičke izazove – sakupljanje paketa s određenim brojevima. Paketi se pojavljuju na nasumično odabranim pozicijama, a na svakom paketu napisan je broj koji može biti paran, neparan ili višekratnik broja tri. Slika 2 prikazuje prvu razinu na kojoj je potrebno sakupiti parne brojeve. Igrač mora znati koji je broj paran koji neparan i prema tome sakupiti paket koji je obilježen parnim brojem, a izbjeći paket koji je obilježen s neparnim.



Slika 2. Sakupljanje parnih brojeva

Na drugoj razini igrač mora sakupiti neparne brojeve, a izbjegavati parne (slika 3). Posljednja, treća razina traži od učenika da zna koji je broj djeljiv s tri kako bi mogao sakupiti samo one pakete na kojima je napisan višekratnik broja tri (slika 4). Za svaki točno sakupljeni paket, igrač dobiva 10 bodova, a za svaki pogrešan, oduzima se 5 bodova.



Slika 3. Sakupljanje neparnih brojeva



Slika 4. Sakupljanje višekratnika broja 3

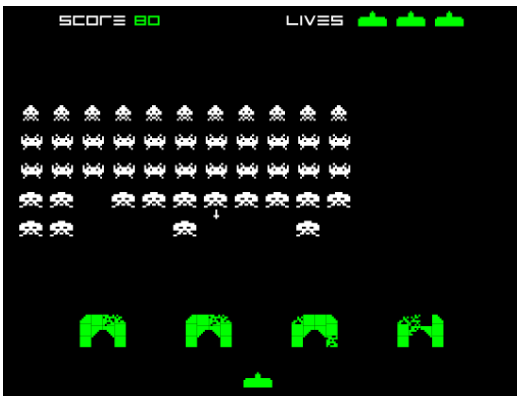
Kako bi igrač prešao na novu razinu potrebno je sakupiti minimalan broj bodova zadan na pojedinoj razini i pronaći portal koji vodi na višu razinu. Ukoliko na određenoj razini igrač ne sakupi dovoljan broj bodova, vraća se na početak igre.

Sa svakom razinom mapa postaje sve složenija, potrebno je sakupiti više bodova te su brojevi koji pišu na paketima u sve većim rasponima. Na prvoj razini se pojavljuju brojevi do 20, na drugoj razini su brojevi do 50 i na trenutno posljednjoj, trećoj razini pojavljuju se brojevi do 100.

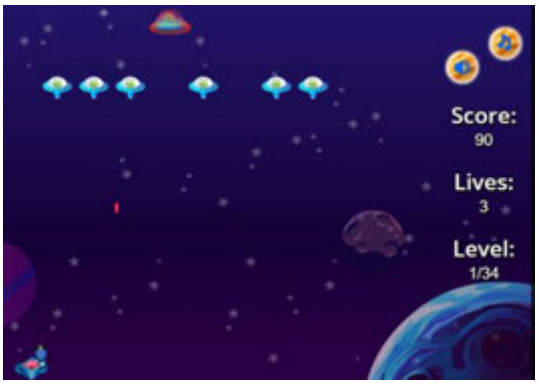
Pojavljivanje brojeva na paketima nije unaprijed definirano, već se pojavljuju nasumično. Igra se može dalje nadograđivati i prilagoditi gradivu koje učenici usvajaju tako da uključuje brojeve djeljive s odabranim brojem, npr. s brojem 5 ili 10 i da se fleksibilno definira raspon brojeva unutar kojih je potrebno pronaći odgovarajuće brojeve.

V. MODIFIKACIJA ARKADNE IGRE SPACE INVADERS U EDUKATIVNE INAČICE

Druga ponuđena igra je *Space Invaders* [21], arkadna igra nastala 1978. godine. Arkadna igra je vrsta računalne igre koja radi na principu da se u nju umetne novčić ili žeton za koje se dobije određeno vrijeme igranja ili određen broj života. *Space Invaders* je dvodimenzionalna igra pucačine u kojoj igrač kontrolira laserski top kako bi porazio gomilu napadača – vanzemaljaca. Cilj je uništiti 5 redova po 11 vanzemaljaca. Za uspješno eliminiranje vanzemaljaca igrač je nagrađivan bodovima. Igrač ima određeni broj života koje može izgubiti ako ga pogodi projektil vanzemaljca. Igra završava kada vanzemaljci dođu do dna ili kada igrač izgubi sve živote. Slika 5 prikazuje originalnu igru, dok se ne slici 6 može vidjeti novija verzija igre.



Slika 5. Original Space Invaders

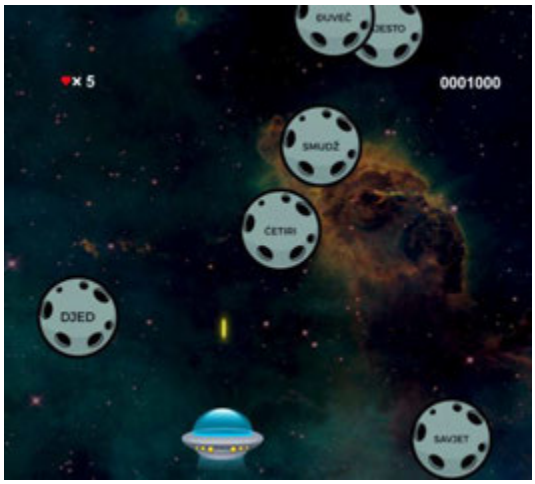


Slika 6. Novija verzija igre Space Invaders

Edukativne inačice igre *Space Invaders* također su razvijene u sklopu kolegija Objektno orijentirano programiranje na Odjelu za informatiku. Zanimljivost ponuđenih igara je da spadaju u žanrove igre koji u svojoj osnovi nemaju nikakve poveznice s edukacijom, a mogu se uspješno transformirati u edukativne inačice.

A. Igra RocketSpelling

Igra *RocketSpelling* je namijenjena učenicima 3. ili 4. razreda osnovne škole. Cilj igre je vježbanje pravilno napisanih riječi hrvatskog jezika. Pravopisna pravila koja se provjeravaju u igri su ije/je, č/ć i dž/đ. Igra prikazuje malog vanzemaljca čiji planet su napali asteroidi obilježeni riječima. Na nekim asteroidima su riječi pogrešno napisane, npr. „ćetiri“, „smudž“ (slika 7). Kako bi spasio svoj planet, vanzemaljac mora uništiti asteroide na kojima su riječi pogrešno napisane, a asteroide s pravilno napisanim riječima treba izbjeći. Za svaki točno uništeni asteroid, dobije se 100 bodova, a za svaki pogrešno uništeni asteroid igrač gubi 100 bodova. Asteroidi se pojavljuju na nasumičnim pozicijama.



Slika 7. Igra RocketSpelling

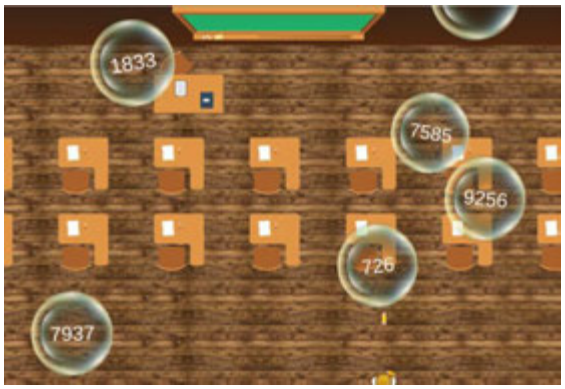
Riječi koje se ispisuju na asteroidima učitavaju se iz datoteke te je moguće dodavanje novih riječi. Kako igra napreduje, asteroidi se pojavljuju sve češće i sve većom brzinom što zahtjeva od učenika ne samo dobro usvojeno gradivo hrvatskog jezika već i dobru koncentraciju, umješnost upravljanja s mišem.

Igrač u igri ima 5 života. Svaki puta kada neuništeni asteroid dotakne igrača gubi jedan život. Igra traje sve dok igrač ima života. Kako bi igra bila što dulje interesantna treba biti predviđeno nasumično generiranje i ispravno i neispravno napisanih riječi te mogućnost učitavanja dodatnog skupa riječi. Moguće je definirati nove razine igre uvodeći dodatna pravopisna pravila, kao što su npr. veliko i malo slovo, sastavljeno i nesastavljeno pisanje, uvećanice i umanjenice i sl.

B. Igra EduBubblePopper

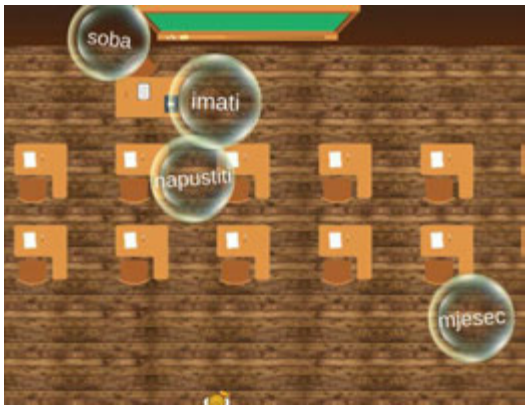
Igra *EduBubblePopper* namijenjena je učenicima mlađih razreda osnovne škole. Glavni lik u igri je učenik u učionici kojeg napadaju balončići s napisanim brojevima ili riječima. Kako bi uništio zadane balone, učenik ispaljuje olovke koje buše balone. Iz matematike se mogu vježbati parni i neparni brojevi, a iz hrvatskog jezika imenice i glagoli. Slika 8 prikazuje varijantu kada igrač igrajući igru može vježbati matematiku i to parne brojeve. U ovom slučaju cilj je uništiti što više balona s parnim brojevima.

Baloni se pojavljuju na nasumičnim pozicijama i s nasumično napisanim brojevima u rasponu od 1 do 10.000. Kako bi igra bila što izazovnija i zanimljivija, s vremenom se baloni počinju pojavljivati češće, poveća im se brzina padanja te veličina balona varira. Kada se pogodi točan balon, dobije se 100 bodova, a pogotkom pogrešnog balona gubi se jedan život. Igrač ima 5 života, i ukoliko izgubi svih 5 života igra završava i vraća se na početak. Kada igrač sakupi 1500 bodova, igrač je pobijedio i igra završava.



Slika 8. Prepoznavanje parnih brojeva

Na slici 9 prikazana je druga varijanta iste igre namijenjena vježbanju prepoznavanja vrsta riječi. Potrebno je sakupiti što više balona na kojima su napisani glagoli, a pri tome izbjegavati balone na kojima pišu imenice.



Slika 9. Prepoznavanje vrsta riječi

Obje inačice igre se mogu dalje razvijati i nadograđivati uvodeći nove razine igre u kojima se vježba dijeljenje brojeva s brojevima do 10, odnosno provjerava poznavanje i ostalih vrsta riječi kao što su npr. pridjevi, zamjenice, veznici itd.

C. Igra PrimeHunter

Igra *PrimeHunter* je namijenjena učenicima petih razreda osnovne škole za vježbanje dijeljenja s prim

brojevima 2, 3, 5, 7. Cilj igre je uništiti sve brodove koji napadaju bazu. Brodovi sadrže brojeve koji predstavljaju njihovu energiju, a mogu se uništiti jedino ako se brod pogodi s djeliteljem njegove trenutne energije. Ako je brod pogođen s djeliteljem, energija broda će se umanjiti i poprima vrijednost količnika. Postupka se ponavlja sve dok energija broda nije 1 kada se brod uništava.

Na Slici 10. prikazano je uništavanje broja 9 s djeliteljem 3. Ispaljivanjem hica s brojem 3 u brod s energijom 9, brod će se djelomično uništiti i energija broda će pasti na 3 (9:3=3) te je brod potrebno još jednom pogoditi s istim djeliteljem da bi se uništio (9:3=1).



Slika 10. Uništavanje broja 9

Djelitelj se može odabrati na početku igre i postavljen je za vrijeme trajanja igre ili se može dinamički mijenjati tijekom igre. Samo ta mala modifikacija igre koja omogućuje dinamički odabir djelitelja tijekom igre uvelike otežava igrivost tako da iako se radi o primjeni znanja iz gradiva mlađih razreda osnovne škole, zahtjeva puno veću koncentraciju i koordinaciju pokreta i upravljanja s mišem te je u tom obliku bolje primijenjena starijem uzrastu učenika.

Na Slici 11. dan je primjer prototipa igre za stariji uzrast učenika osnovne i srednje škole. Djelitelj je na početku postavljen na 1 ali se može mijenjati tijekom igre pritiskom na određeno slovo koje predstavlja jedan od ponuđenih prim brojeva. Postoji i mogućnost da se više puta klikne na isto slovo čime se odabrani djelitelj uvećava te se dobivaju višekratnici odabranog djelitelja. Djelitelje je moguće međusobno kombinirati kako bi se dobili i ostali brojevi.



Slika 11. Uništavanje broja 15

U primjeru na slici 11 prikazano je uništavanje broda s djeliteljem 15. Djelitelj 15 se mogao dobiti kao višekratnik broja 5, dva put klikanjem po oznaci za prim broj 5 ili kombiniranjem prim brojeva 3 i 5.

Ako se broj podijeli s krivim djeliteljem, njegova brzina se poveća. Bazu štite tri štita koji predstavljaju živote. Svaki puta kada se broj ne uništi, on dotakne štit čime se gubi jedan život. Igru je moguće igrati u tri različite težine: *easy, medium* i *hard*. Razine se međusobno razlikuju po brojevima koje je potrebno uništiti – što je teža razina to su veći brojevi te po učestalosti pojavljivanja brojeva – što je veća razina, brojevi se češće pojavljuju.

VI. ZAKLJUČAK

Informatička tehnologija postala je neizostavni dio svakodnevice. Računalne igre dostupne su na različitim platformama, a igranje digitalnih igara nezaobilazni je dio svakodnevnih aktivnosti današnjih učenika.

Cilj igre je da igrač poštujući pravila postigne što bolji rezultat. Motivacijski elementi igre potiču visok stupanj aktivnosti i usredotočenosti igrača, razvijaju njegovu kreativnost, prilagodbu i želju za otkrivanjem novih metoda rješavana problema ali zahtijevaju i upornosti i ponavljanje nekih rutina kako bi njima ovladao

Edukativne računalne igre imaju veliki potencijal za primjenu u nastavi. Međutim i kod edukativnih igara počeo se pojavljivati problem jer učenicima nisu jednako zanimljive niti zabavne kao komercijalne igre. Kako bi se prevladala razlika između poznatih igara i edukativnih igara, poznate igre su se počele prilagođavati u edukativne svrhe.

U ovom radu predstavljeno je pet različitih edukativnih inačica igre *Space Invaders* te jedna edukativna verzija igre *Super Mario.* Igre su namijenjene učenicima mlađih razreda osnovne škole za vježbanje jednostavnih zadataka iz matematike ili hrvatskog jezika. Sve igre su razvili studenti Odjela za informatiku, Sveučilišta u Rijeci u sklopu kolegija Objektno programiranje. Cilj razvijanja ovih igara bio je studentima omogućiti praktičnu primjenu naučenih koncepte iz područja objektno orijentiranog programiranja na projektima iz realnog života za odabranog korisnika.

Razvoj prototipa edukativnih verzija poznatih igara pokazao je da su jednostavne i popularne igre vrlo dobar materijal za razvoj edukativnih igara i to čak i u slučaju žanra koji nije primjeren obrazovnoj ustanovi. U danim primjerima pokazano je da se komercijalne jednostavne igre mogu jednostavno nadograditi i prilagoditi zahtjevima edukativnog procesa u školi, posebno usvajanju gradiva nižih razreda osnovne škole.

U idućim koracima planiraju se razviti predložene nadogradnje igara te provesti testiranje s učenicima kako bi se ispitala igrivost i primjerenost sadržaja. Važno je u budućem radu istražiti i razviti načine kojima bi se pratili, vrednovali i ocjenjivali rezultati koje učenici postižu tijekom igranja edukativnih igara te razviti metode kojima će se edukativne igre što bolje prilagoditi učeniku-igraču da ga što više motiviraju i potaknu na usvajanje novih sadržaja.

REFERENCE

[1] P. Felicia, *Digitalne igre u školama - Priručnik za učitelje*. Brussels, Belgija: European Schoolnet, 2009.

[2] T. Francis and F. Hoefel, "'True Gen': Generation Z and its implications for companies," *McKinsey Co.*, p. 10, 2018.

[3] M. Prensky, "Digital game-based learning," *Comput. Entertain.*, vol. 1, no. 1, p. 21, 2003, doi: 10.1145/950566.950596.

[4] J. Rugelj, "Serious computer games in computer science education," *EAI Endorsed Trans. Game-Based Learn.*, vol. 2, no. 6, p. 150613, 2015, doi: 10.4108/eai.5-11-2015.150613.

[5] O. P. Shefer, L. S. Nosova, and T. N. Lebedeva, "A Modern Methodology for Teaching Programming at a University," *Sci. Tech. Inf. Process.*, vol. 45, no. 2, pp. 81–86, 2018, doi: 10.3103/S0147688218020077.

[6] M. Farah, G. Ireson, and R. Richards, "Developing a Predictive Model for the Enhanced Learning Outcomes by the Use of Technology," *Imp. J. Interdiscip. Res.*, vol. 2, no. 5, pp. 2454–1362, 2016.

[7] J. Alghamdi and C. Holland, "Game-Play: Effects of Online Gamified and Game-Based Learning on Dispositions, Abilities and Behaviours of Primary Learners," vol. 515, pp. 55–63, 2017, doi: 10.1007/978-3-319-74310-3.

[8] M. E. del M. Pérez, A. P. Guzmán Duque, and L. C. F. García, "Game-based learning: Increasing the logical-mathematical, naturalistic, and linguistic learning levels of primary school students," *J. New Approaches Educ. Res.*, vol. 7, no. 1, pp. 31–39, 2018, doi: 10.7821/naer.2018.1.248.

[9] A. Khan, F. H. Ahmad, and M. M. Malik, "Use of digital game based learning and gamification in secondary school science: The effect on student engagement, learning and gender difference," *Educ. Inf. Technol.*, vol. 22, no. 6, pp. 2767–2804, 2017, doi: 10.1007/s10639-017-9622-1.

[10] T. Hainey, T. M. Connolly, E. A. Boyle, A. Wilson, and A. Razak, "A systematic literature review of games-based learning empirical evidence in primary education," *Comput. Educ.*, vol. 102, pp. 202–223, 2016, doi: 10.1016/j.compedu.2016.09.001.

[11] N. Iten and D. Petko, "Learning with serious games: Is fun playing the game a predictor of learning success?," *Br. J. Educ. Technol.*, vol. 47, no. 1, pp. 151–163, 2016, doi: 10.1111/bjet.12226.

[12] "Hompage_Minecraft_Education_Edition," 2020. .

[13] "Lightbot." [Online]. Available: http://lightbot.com/flash.html. [Accessed: 20-Apr-2019].

[14] "Run Marco." [Online]. Available: https://runmarco.allcancode.com/. [Accessed: 20-Apr-2019].

[15] "CodeMonkey." [Online]. Available: https://www.playcodemonkey.com/home. [Accessed: 20-Apr-2019].

[16] B. Kim, H. Park, and Y. Baek, "Not just fun, but serious strategies: Using meta-cognitive strategies in game-based learning," *Comput. Educ.*, vol. 52, no. 4, pp. 800–810, 2009, doi: 10.1016/j.compedu.2008.12.004.

[17] I. A. Zualkernan, "A framework and a methodology for developing authentic constructivist e- Learning environments," vol. 9, pp. 198–212, 2006.

[18] R. Ibrahim, R. Che Mohd Yusoff, H. Mohamaed@Omar, and A. Jaafar, "Students Perceptions of Using Educational Games to Learn Introductory Programming," *Science (80-.).*, vol. 4, no. 1, pp. 205–216, 2011, doi: 10.5539/cis.v4n1p205.

[19] M. Muratet *et al.*, "Experimental Feedback on Prog & Play , a Serious Game for Programming Practice To cite this version : Experimental feedback on Prog & Play , a serious game for programming practice," 2016.

[20] "The official home for Mario - Home," 2019. .

[21] "Invader - Home," 2014. .

Izazovi poučavanja o računalnim mrežama

K. Blažeka, Tehnička škola Čakovec, Tehnička škola Ruđera Boškovića, Zagreb, Hrvatska
kristinka.blazeka@skole.hr

Sažetak - Računalne mreže su bitan i ključni stručni predmet u području obrazovanja za zanimanje Tehničar za računalstvo. Kao i mnogi drugi predmeti struke, zahtjeva savladavanje teoretskih znanja i praktičnih vještina. Pri tom je učenje o mrežnim protokolima i drugim važnim konceptima računalnih mreža učenicima često teško i apstraktno. Tradicionalne metode poučavanju nisu osobito prilagođene pa njihova primjena za posljedicu ima nedostatak razumijevanja kod učenika što u konačnici vodi prema gubitku interesa i motivacije za učenje. Kako bi se to prevladalo i sadržaje učinilo zanimljivijim i razumljivijim, potrebno je pripremiti odgovarajuće materijale i vježbe, te prilagoditi nastavne metode. Poželjno je poraditi na oblikovanju sadržaja, kao i primjenjivati prilagođeni pristup i alternativne metode učenja i poučavanja osnovnih koncepata u području umrežavanja računala, pa su npr. tradicionalne metode učenja primjerene razvoju razumijevanja ključnih mrežnih komponenata i arhitekture, a problemski bazirano učenje, kao preporučen pedagoški alat, omogućuje učenicima razvoj vještina dizajniranja u području računalnih mreža. U radu je ukratko, kroz diskusije sa učenicima i nastavnicima, razmotrena učinkovitost metoda s obzirom na specifičnosti predmeta i komentiran njihov utjecaj na učenje, interes i razumijevanje učenika. Ciljevi su povećanje interesa, poboljšanje rezultata učenja i promocija razvoja kurikuluma.

I. UVOD

Tvrtke suočene s rapidnim razvojem tehnologije očekuju neupitnu i odgovarajuću stručnost školovanih zaposlenika, pa je suvereno ovladavanje tehnološkim vještinama tijekom i nakon završetka obrazovanja u strukovnom području računalstva od iznimne važnosti. Stručne kompetencije se tijekom obrazovanja uglavnom nastoje ostvariti stavljanjem naglaska na stjecanje upotrebljivih vještina sukladno zahtjevima autentičnog okruženja realnog svijeta, pri tom ne zanemarujući niz drugih, u modernom svijetu nužnih vještina kao što su komunikacijske, vještine timskog rada, inicijativa i motivacija, analitičke i druge vještine.

Računalne mreže su predmet obaveznog dijela kurikuluma za zanimanje Tehničar za računalstvo. S obzirom da su interdisciplinaran predmet, podrazumijevaju značajno poznavanje sklopovske računalne tehnologije, elemenata softverskog inženjerstva, elektroničke i komunikacijske tehnologije kao i dijelova drugih srodnih disciplina. Nedovoljno poznavanje bilo kojeg od ovih sastavnih dijelova otežava učenje.

Značaj računalnih mreža porastao je kroz veliku raširenost i utjecaj Interneta, te mobilnih i bežičnih tehnologija za svakodnevne poslove i u svakodnevnom životu, pa učenje o računalnim mrežama stavlja naglasak na razvoj sposobnosti primjene mrežnih tehnologija na situacije iz stvarnog života. Potreba za dubokim tehničkim znanjima određenim je dijelom pomaknuta prema primjeni i upotrebi mreža, tako da se u novijm kurikulima očekuje svojevrstan pomak sa usvajanja tehničkog znanja i razumijevanja na dizajn sustava računalnih mreža koje udovoljavaju zahtjevima korisnika.

Važnost ovog područja računarstva ističe se i kroz primjetnu potražnju za stručno osposobljenim tehničkim osobljem u području umrežavanja koja je u stalnom porastu, kako kvalitativno, tako i kvantitativno. Pri tom su ostvarene kompetencije u izravnoj i proporcionalnoj vezi sa primjenjenim načinima, odnosno metodama učenja i poučavanja.

II. UČENJE I KOMPETENCIJE

Učenje je općenito složen, nelinearan i rekurzivan proces koji je uspješniji u okolinama u kojima se učenici osjećaju sigurno i u kojima ostvaruju osjećaj vlasništva i kontrole nad vlastitim procesom učenja u svrhu stjecanja kompetencija u znanju, vještinama, izvedbi i stručnosti.

Većina metodologija koje promiču uspješno učenje oblikovana je tako da potiče prijenos znanja i informacija. Ako cilj učenja dodatno uključuje promociju vještina i stručnosti pri obavljanju određenih zadataka, u nastavi je važno simulirati njihovo rješavanje tako da učenici uvježbaju izvođenje predviđenim koracima. Osim toga, na tom stupnju učenici bi morali naučiti i kako prilagoditi i primijeniti stečene vještine i u različitim (nekim drugim) kontekstima.

Pri tom bi trebalo respektirati različitost i složenost stilova [1] i procesa učenja pojedinih učenika. Učenici mogu preferirati vizualno (grafovi, modeli, slike), verbalno (tiskani materijali), taktičko ili kinestetičko (veze sa stvarnošću) ili osjetilno (praktično) učenje. Prema shvaćanju, učenike također možemo razlikovati na aktivne (raspravljaju, primjenjuju, objašnjavaju), refleksivne (proučavaju razmišljanjem), globalne (odlični su u sintezi, viđenju cjeline), te sekvencijalne (dobro analiziraju koncepte primjenom linearnog učenja, dobro se prilagođavaju klasičnoj, sekvencijalno organiziranoj nastavi) [2].

No, neovisno o tome da li je različitost ostvarena samo kroz različitost stilova učenja ili se učenici razlikuju i kulturološki, nastavnik bi pristup predavačkoj nastavi trebao pokušati obogatiti primjerima i metaforama s obzirom da pažnja učenika vrlo jednostavno obično nestaje već nakon nekoliko minuta nastave.

Jačanjem participacije učenika moguće je nastavu učiniti interaktivnijom, pa iako je tada količina prenijetih informacija učenicima nešto manja, kvaliteta njihovog

učenja i razumijevanje značajno su bolji. U takve, na učenika usmjerene metode, ubrajaju se diskusije ili laboratorijske vježbe kod kojih se nastavnik suočava sa pitanjima, komentarima i idejama učenika. Ovo nastavniku omogućuje da kod učenika stimulira kritičko razmišljanje, stvaranje, razvoj i jačanje ideja. Suvremeno participirajuće poučavanje koristi razne metode kao što su debate, suprotstavljene diskusije, igre, igranje uloga, diskusije u malim grupama, suradničko učenje i drugo, što su sve metode primjenjive na sekundarnoj razini obrazovanja.

Kako nedostatno razumijevanje sadržaja nastavnih predmeta sa puno teoretskih i tehničkih koncepata često dovodi do nepoželjnog gubitka interesa i motivacije, iskustvo u učionici upućuje na primjenu alternativnih nastavnih metoda. Takav pristup pojačava interes učenika prilikom uvođenja temeljnih koncepata u području određenog nastavnog predmeta.

III. SPECIFIČNOSTI PREDMETA

Nastavni predmet računalne mreže jedan je od ključnih stručnih predmeta, a pokriva područje koje je u rapidnom razvoju. Uočljiva ekspanzija i mogućnosti računalnih mreža traže uspješnije obrazovanje primjenom novih pristupa poučavanju kroz aktivnije i intenzivnije uključivanje učenika u proces učenja.

Uz svoju naglašenu praktičnu orijentaciju ka usvajanju vještina rješavanja praktičnih problema, ovaj predmet je karakteristično i teoretski jer počiva na nizu apstraktnih koncepata. Zato kod učenika treba podjednako razvijati i razumijevanje teorijskih koncepata (pri čemu je važno paziti na odgovarajuću dubinu sadržaja, te povezanost određenih sadržaja sa sadržajima drugih povezanih područja, primjerice teorijom podatkovnih struktura i algoritama) kao i na primjenu.

Predmet složenim čine i mnogi sadržaji koji su međusobno jako povezani i ovisni, npr. slojevi mreža koji se poučavaju: fizički sloj, sloj podatkovne veze, mrežni sloj, transportni sloj i aplikacijski sloj.

Nadalje, mrežni protokoli apstraktni su za učenje te tako uglavnom prilično teški za razumijevanje.

Nezanemariv izazov stvara i velika širina područja računalnih mreža tako da je iznimno važan način organizacije kao i stupanj detaljizacije sadržaja pri čemu dodatno nastavnici trebaju obraćati veliku pažnju podjednako na saznanja o trenutnim istraživanjima i razvoju kao i na kvalitetu poučavanja.

Uz to, mnogi sadržaji su evidentno paralelni jedni drugima (interdisciplinarno ili među različitim disciplinama), što za posljedicu ima mogućnost predviđanja varijabilnog redoslijeda učenja (npr. mrežni protokoli vezani su uz arhitekturu mreže tako da nedovoljno poznavanje koncepta arhitekture može otežavati razumijevanje protokola i tako izravno utjecati na uspješnost procesa učenja i poučavanja).

Može se istaknuti i kako situacije sa velikim grupama učenika (što je najčešće posljedica organizacijskih i/ili prostornih mogućnosti obrazovnih ustanova) ne potiču dijalog između nastavnika i učenika, što umanjuje aktivnost i uključenost polaznika i ne potiče učenje.

IV. POTICAJNO UČENJE

Za vrijeme predavačke nastave, nastavnik samostalno određuje ciljeve, sadržaj i organizaciju prezentacije, dopunjava i objašnjava tekstualne materijale. Na taj način takva je nastava zapravo jednosmjerna komunikacija koja se oslanja na gledište nastavnika bez značajnijeg osvrta na potrebe učenika za interakcijom kroz diskusiju, upite ili praktične vježbe.

Bez verbalne povratne veze nastavnik uglavnom nije svjestan problema, odnosno nerazumijevanja sadržaja koje učenici mogu imati. Dodatno, iskustvo pokazuje da učenici koji imaju pristup kopijama bilješki s predavanja ili priručniku u značajnom postotku uglavnom pokušavaju izbjegavati ovu pasivnu, jednosmjernu metodu prijenosa informacija, prilikom koje nema ili je vrlo malo interakcije s nastavnikom [3]. Ipak, učinkovitost takve, predavačke nastave, može se u određenoj mjeri povećati odgovarajućim treninzima.

Da bi obrada nastavnog sadržaja bila zanimljivija, lekcije je dobro potkrijepiti slikama, shematskim prikazima, grafovima i jednostavnim skicama, a po mogućnosti i kratkim filmovima, te, kad je to odgovarajuće, demonstracijama postupaka, rada instrumenata ili uređaja ili specifičnostima vještina vezanih uz samu profesiju.

Osim prepisivanja, učenicima treba omogućiti diskusije, postavljati im pitanja, zadavati vježbe, provoditi *brainstorming* u grupama i drugo.

Za domaći ili seminarski rad mogu se pripremati i grupni projekti koji potiču međuobnu suradnju učenika, jer kroz međusobne interakcije učenici postaju aktivniji. Poticajna školska okolina i veće iskustvo nastavnika pozitivno utječu na ovaj proces.

Važno je naglasiti i uzeti u obzir da priprema ovakvih metoda često zahtjeva dio vremena predviđen za obradu sadržaja, te da uspješnost ovisi i o pripremljenosti i spremnosti učenika za ovakve pristupe.

Nadalje, ne bi se smjelo zanemariti ni motivacijski učinak svojevrsnog nagrađivanja nastavnika koji implementiraju ovakve pristupe, ne zanemarujući ni potrebu dodatne komunikacije, suradnju i interakciju kroz diskusije izvan učionice prije ili nakon nastave, koristeći mogućnosti koje nudi suvremena tehnologija. Pri tom, ekspresivna, neverbalna komunikacija u stvarnoj učionici, licem u lice, kroz koju nastavnik dobiva povratnu informaciju o razumijevanju učenika, ovim načinima ne može biti nadomještena.

Važno je osigurati i da se u svim situacijama učenici osjećaju udobno i ugodno sa ravnopravnim mogućnostima za savjetovanje i konzultacije, što je važan preduvjet za uspjeh, te da čitavo vrijeme nastavnici spremno i otvoreno upotrebljavaju ideje i različite načine alternativnog poučavanja kako bi zadržali interes učenika, njihovu participaciju i motivirali dijalog.

Vrlo je bitno i odgovarajuće vrednovanje uspješnosti u postizanju predviđenih ishoda procesa učenja kako bi i učenici i nastavnici dobili neposrednu povratnu informaciju o svojoj uspješnosti.

V. MOGUĆNOSTI ZA KVALITETNIJE POUČAVANJE

Kakvo je poučavanje računalnih mreža dobro? Kako ga postići?

Slično drugim predmetima, nastavnik u predmetu računalne mreže u mogućnosti je poboljšati učenje učenika ako je svjestan različitih načina kako učenik razumijeva (percipira) koncepte (npr. protokole) kojima ga poučava [4], mijenjajući prema tome vlastito poučavanje.

Također, radi uspješnijeg rješavanja složenih ili nekih novih problema u području računalnih mreža kod prosuđivanja relevantnosti i kvalitete rješenja, važno je usvajati različite načine razmišljanja o mrežnim konceptima [4].

U tu svrhu, nastavnik može promovirati dobro razumijevanje i poboljšavati učenje nudeći varijacije u načinima na koje učenike poučava mrežne protokole ili druge koncepte sukladno različitostima njihovih shvaćanja. Stvaranjem varijacija u poučavanja promovira se bolje učenje učenika [4].

Načini za učenje i poučavanje računalnih mreža mogu uključivati e-learning, upotrebu simulacijskog softvera (npr. Cisco Packet Tracer), projektno bazirano učenje, problemski bazirano učenje, laboratorijske vježbe i drugo.

U nastavku su dani neki prijedlozi:

A. Laboratorij za računalne mreže

Dobar je pristup (sada već uvriježen u praksi) korištenje laboratorija za računalne mreže, uz pretpostavku posjedovanja odgovarajuće infrastrukture i opreme, te da su učenici uspješno ovladali temeljnim konceptima.

Pri tom upotreba besplatnih softverskih alata (npr. OPNET IT guru, Cisco Packet tracer, NS2 i sl.) pomaže učenicima učenje praktičnih vještina i izvan vremena predviđenog u školi, ponavljanje lekcija i dublje razumijevanje predmeta.

Poznato je kako za najbolje učinke u razumijevanju veza između znanosti i tehnologije, obrazovanje u tehnologiji, te učenje o sadržajima, procesima i vještinama treba biti objedinjeno, odnosno učenje teorije treba biti popraćeno aktivnim vježbama u laboratoriju.

Kvalitetan praktičan rad u laboratoriju od središnjeg je značaja u procesu učenja i poučavanja za razvoj razumijevanja procesa i koncepata, stimulaciju i stvaranje interesa za učenje, te razvija kritičko i logičko razmišljanje.

Upotreba simulacijskog softvera važna je jer daje priliku učenicima da predstavljaju ideje realnog svijeta ili okoline, no još učinkovitije je učenje kroz praktične aktivnosti (npr. instaliranje i konfiguriranje softvera, mrežnih uređaja ili instaliranje i konfiguriranje stvarne računalne mreže).

Praktične aktivnosti jačaju interes učenika, pomažu im razvijati kritičko i logičko razmišljanje i lakše razumijevanje znanstvenih koncepata, pojava i procesa, učenici usvajaju konkretne vještine umrežavanja, razumijevanje je dublje, jača se motivacija učenika za učenjem računalnih mreža, promovira vještine rješavanja problema, promovira se interakcija među učenicima i nastavnicima, te kreativnost učenika i jača njihovo pamćenje [5].

U laboratorijima učenici promišljaju, diskutiraju, simuliraju ili implementiraju sadržaje naučene u učionici i povezuju ih sa primjenama u stvarnom svijetu. Učenici praktičnim radom stječu vještine, usvajaju koncepte, razvijaju svoje kognitivne sposobnosti kritičkog razmišljanja, rješavanja problema, primjene, analize i sinteze, razvijaju stavove vezane uz znatiželju, interes, preuzimanje rizika, objektivnost, odgovornost, suradnju i drugo. Na neki način, te su vježbe prvi istraživački projekti učenika.

B. Projekti u nastavi

Projekte u nastavi može se realizirati uvođenjem odgovarajućih projekata u područja obuhvaćena nastavnim kurikulumom (npr. podatkovna komunikacija, mrežni modeli, osnove LAN-a, topologija i mrežne tehnologije, komunikacijski protokoli, konfiguriranje mrežnih uređaja i mreža i dr.).

Neki primjeri projekata su rješavanje problema uspostave bežične mreže u učionici, izrada *chat* aplikacije, te izrada *web* aplikacije koja omogućava prijenos video sadržaja upotrebom *web*-kamere u realnom vremenu. Pri tom se učenici izlažu stvaranju osjećaja odgovornosti za određivanje ciljeva svog učenja, izlažu se iskustvu planiranja postupaka potrebnih za rješavanje problema, te ohrabrivanju suradničkog učenja među timovima [6].

C. Problemski bazirano učenje

Kako bi učenici bolje razumjeli sadržaje te bili zainteresiraniji za učenje, može se predložiti i metoda problemski utemeljenog učenja rješavanjem problema analogijom. Primjer metode je demonstracija stvaranja paralela između računalne mreže i sustava brze dostave. Učenicima se postavlja problem o kojem razmišljaju i za koji traže nova znanja i vještine da bi ga riješili razvijajući naviku promišljanja i kreativni duh te tako postaju zainteresiraniji za učenje [7].

Može se primijeniti i problemski bazirano učenje (PBL) dizajnirano pomoću odgovarajućih scenarija sa visokim stupnjem realističnosti [8].

D. Upotreba multimedije

Mogućnosti multimedijskog poučavanja su životne, zanimljive i informativne te proširuju mogućnosti praktičnog rada u učionici, no ne bi trebale biti restriktivne u odnosu na improvizaciju nastavnika u određenim situacijama, kao što je to npr. kod izvođenja ili dokazivanja formula kad učenik slijedi razmišljanja nastavnika kako bi shvatio bit problema. [9]

E. Upotreba mrežne platforme

Mrežna platforma kao virtualna okolina za učenje osigurava dobar sekundarni izvor informacija za učenje (za neovisna istraživanja, višestruke interakcije, suradničko učenje, dijeljenje kvalitetnih resursa za učenje i drugo).

Platforma može sadržavati nastavne materijale, plan poučavanja, relevantne informacije i vježbe, pitanja za provjeru, zanimljive priče, pokuse, može omogućiti interakciju, stvaranje baze ispita i online vježbe, testove za samoispitivanje itd. Takva se okolina treba jednostavno koristiti, biti dostupna i mora biti djeljiva. [9]

F. Igre i druge metode

Ostvarenje zanimljive i izazovne okoline učenja ovisno je o kreativnosti nastavnika. Jedan od ciljeva participatornih metoda poučavanja je poticanje aktivnog učenja, interesa i razumijevanja kod učenika. Tako npr. učenici mogu učiti protokole kroz igru jer igra izaziva, budi i maštu i znatiželju. Izazov je u neizvjesnosti cilja, pri čijem se dostizanju učenici osjećaju bolje i samouvjereniji su.

U [3] dan je opis nekih pristupa i procedura koje se mogu koristiti kako bi kod učenika održali interes i olakšali im usvajanje osnovnih teoretskih koncepata u području računalnih mreža (obogaćeno igrama, čarolijama i dramatičnim predstavama).

Znatiželja je uglavnom izravno povezana sa motivacijom za učenje. No, znatiželja će biti stimulirana samo ako je igra optimalne složenosti. Zato, okolina igre ili zadatka nikad ne smije biti ni prejednostavna ni presložena.

Mnoge se jednostavne igre mogu provesti i samo uz pomoć krede i ploče (npr. primjena teorije grafova u komunikacijskim, odnosno računalnim mrežama). Također, učenici mogu učiti i pomoću analogija koje jačaju pažnju, pri čemu je potrebno osvijestiti kako učenici lakše pamte priče nego tehničke koncepte. [3]

VI. Iskustvo u izvedbi

U svrhu uvida i donošenja zaključaka o stvarnom stanju i situacijama u nastavi, provedeno je ispitivanje među učenicima polaznicima i kolegama koji predaju predmete Uvod u računalne mreže i Računalne mreže u zanimanju Tehničar za računalstvo.

Zamišljeno je napraviti određenu procjenu stanja i stvoriti ideje o mogućim putevima za promjene i poboljšanje. Učenici i nastavnici zato su intervjuirani početno u slobodnoj formi, bez unaprijed preciziranih pitanja, tek sa zamolbom da općenito iznesu mišljenje o predmetu s obzirom na sadržaje i način izvedbe (predavanja u učionici, odnosno laboratorijske vježbe). Dobivena mišljenja trebala bi poslužiti uobličavanju stvarnog upitnika za buduća istraživanja prema temama koje su istaknute, odnosno koje su se češće javljale kao sugestije i promišljanja intervjuiranih osoba.

Prikupljena su mišljenja 110 učenika, te su intervjuirana i dva nastavnika. Analiza koja slijedi uključuje neka zapažanja intervjuiranih osoba i neka osobna zapažanja. S obzirom da se radilo o intervjuu otvorenog tipa, nisu iznijeti brojčani podaci. Takvo je istraživanje planirano u slijedećoj fazi, nakon što se prema odgovorima učenika precizno oblikuje sadržaj upitnika.

U nastavku su iznijeta neka zapažanja, grupirano prema razmatranim područjima.

A. Složenost i značaj predmeta

Učenici koji su komentirali težinu i složenost predmeta isticali su kako im se predmet čini umjereno do vrlo težak. Očekivano, tek je manji dio učenika naglasio kako iznijete sadržaje usvoji uglavnom već na nastavi.

Također, neki učenički odgovori navode na zaključak kako ni velika raširenost i važna uloga, ali niti značaj računalnih mreža u suvremenom svijetu nisu dovoljni za njihovu motivaciju i buđenje interesa za ovo područje.

B. Teoretski dio nastave

Uočljivo je da dio učenika (što je i inače prilično čest slučaj u strukovnim školama) teže i s manje motivacije prati teoretsku nastavu u učionici. Takvi učenici često su skloni pokazivati ili isticati probleme vezano uz nedostatak pažnje, nerazumijevanje sadržaja, preveliku količinu sadržaja, pa čak iznositi kako oni zapravo misle u svojoj struci raditi nešto drugo, te im to nije predmet interesa.

Kako se u slučaju predmeta Računalne mreže, kao što je već spomenuto, radi o predmetu sa opširnijom i vrlo važnom teoretskom podlogom, ovo svakako predstavlja određeni problem za kreiranje kurikuluma i stvarnu izvedbu u učionici.

C. Laboratorijske vježbe

Gotovo je pravilo da su laboratorijske vježbe način učenja koji će primjetan broj učenika isticati kao zanimljiviji i privlačniji.

Bolja opremljenost laboratorija opremom (novija i funkcionalna oprema, brža računala, dovoljan broj računala) dodatno potiču zadovoljstvo učenika.

U slučaju predmeta Računalne mreže čini se važnim (što je sukladno prije u tekstu navedenim istraživanjima) naglasiti i potvrditi kako je praktičan rad sa uređajima (ukoliko to opremljenost dozvoljava) u ovom predmetu važniji od simulacijskog načina.

Naime, već na jednostavnoj vježbi konfiguriranja mreže jednakopravnih (*peer to peer* korisnika) izvedenoj na dva načina (simulacijom i stvarnom izvedbom) uočljivo je bilo kako su učenici nailazili na povećan broj vrlo konkretnih problema kad su vježbu izvodili stvarnom izvedbom što je omogućilo i više njihove međusobne diskusije, izraženiju potrebu pristupanja rješavanju stvarnih i konkretnih problema, uz pojačanu potrebu za dijagnostikom i analizom. U konačnici, ove situacije dovode do usvajanja kvalitetnijih vještina, kako u struci tako i vještina važnih za suvremenu svakodnevicu.

Analiza načina izvođenja praktičnog rada (vježbi) s obzirom na broj sudionika ukazuje na potrebu za različitim oblicima rada. Često odluka ovisi o konkretnoj vježbi, pa tako npr. određene vježbe daju najbolje rezultate kod rada u grupi, za neke je najbolji rad u paru (npr. mreža dva jednakopravna korisnika), dok se neke uspješno izvode pojedinačno (rad na programskom simulatoru). Trebalo bi ipak omogućiti da se vježbe koje se izvode primarno preko simulacijske aplikacije realiziraju tako da svaki učenik radi sam na računalu, jer samo tako može steći potrebnu vještinu.

D. Pojmovi i terminologija

Manji broj učenika se prilikom intervjuiranja osvrnuo na terminologiju predmeta. Naime, i ovdje je, kao i u većini drugih slučajeva, na djelu prilagođavanje engleskih termina hrvatskom nazivlju.

Tako se u mrežne uređaje ubraja usmjernik (*router*), preklopnik (ili prespojnik, engl. s*witch*), u opremu ulaze i parice (*twisted pair*) i drugi izrazi i nazivlje.

Kako se u tehnici i tehnologiji teško može obuhvatiti sva širina discipline bez da se makar paralelno ne usvaja i engleske nazive, trebalo bi o tome možda nekako ujednačiti pristup u nastavi, jer, posebno početnike, ovakva situacija može zbunjivati, a može ih i udaljiti od engleske terminologije i tako otežati njihovo buduće djelovanje i uspješnost u struci s obzirom da su priručnici i upute, novosti te razne rasprave na forumima i sl. najčešće dostupni samo na engleskom jeziku.

Ovdje bi bilo dobro spomenuti i to, kako ne bi trebalo zanemarivati ni mogućnosti za primjereniju pripremu učenika za strukovno obrazovanje u području računalstva (u ovom slučaju računalnih mreža) i uspješnije usvajanje odgovarajuće terminologije i procesa, već tijekom obrazovanja u području informatike i programiranja u ranijim godinama školovanja. Tako bi se, primjerice, prilikom obrade odgovarajućih sadržaja mogao stavljati značajniji naglasak na npr. pojam računalne mreže i mrežnih uređaja, važnost komunikacije i dijeljenja resursa, uspostavljanje bežičnih mreža, operacijske sustave i njihovu vezu sa konfiguriranjem mrežnih postavki, zadavanje zadataka u kojima bi se davali i primjeri programiranja za web i sl., a što bi učenicima moglo biti od koristi za bolje i lakše prihvaćanje tema i terminologije vezano uz područje računalnih mreža.

E. Alternativni putovi tumačenja koncepata i usklađenost teoretske nastave sa laboratorijskim vježbama

Ove su teme analizirane sa kolegama nastavnicima.

Naime, prilično je karakteristično u ovom predmetu da se osnovni model koji se provlači prilikom usvajanja koncepata (sedam slojni OSI model) može objašnjavati i tumačiti bilo od fizičkog sloja naviše prema aplikacijskom, bilo obrnuto, od aplikacijskog sloja prema fizičkom. Koliko se čini, postoje pobornici, odnosno činjenice za i protiv obje metode.

Nastava predmeta Računalne mreže, odnosno Uvod u računalne mreže, izvodi se počevši od fizičkog sloja, što zapravo preskače sam početni zahtjev za prijenosom ili razmjenom podataka među računalima koji će potaknuti, odnosno dobro demonstrirati funkcioniranje mrežne komunikacije. Možda bi se moglo napraviti odgovarajuću analizu te usporediti oba načina prihvaćajući bolji način.

Dodatno, određene poteškoće predstavlja usvajanje pojedinih koncepata. Primjerice, usvajanje pojma IP adrese i adresiranja traži se i provlači se kroz vježbe mnogo ranije nego što se objašnjavaju slojevi, odnosno protokoli za koje su te adrese karakteristične.

Općenito, čini se da bi trebalo organizacijom nastave omogućiti bolju i svrsishodniju usklađenost teoretske nastave s praktičnom, jer tjedan za tjednom ne jamči da će se na satovima teorije odraditi sve što je potrebno da bi slijedeći tjedan učenici mogli spremno pristupiti izvođenju vježbe vezane uz neko teoretsko izlaganje.

VII. ZAKLJUČAK

Učenje i poučavanje informacijskih i računalnih tehnologija može biti otežano zbog različitih čimbenika (nedostatak obrazovanog i osposobljenog kadra, zastarjeli kurikulum, slab pedagoški pristup poučavanju, teoretski orijentirana izvedba, nedostatak opreme i drugo).

U slučaju nastavnog predmeta Računalne mreže, kombinacija laboratorijskih vježbi (koje uključuju praktičan rad i simulacije) i predavačke (teorijske) nastave uz uključenje suvremenih nastavnih metoda može učinkovito pomoći učenicima razumijevanje apstraktnih teorija, te jačati njihov interes i svijest o temeljnim konceptima relevantnog područja.

Promišljeni odabir puta i smjera koji će se predlagati kurikulumom, a kojim se će svladati predviđeni sadržaji može pomoći i učenicima i nastavnicima. Pri tom bi bilo dobro usklađivanje osim na razini redoslijeda tumačenja tema, izvesti i na razini organizacije paralelnog teoretskog rada sa praktičnim radom u laboratoriju, pazeći da se koncepti potrebni za izvedbu vježbi prvo objasne u nastavnom dijelu, ne zanemarujući mogućnost da učenici izvode vježbe (simulacije) samostalnim (ali mentoriranim) radom na računalu.

Ovaj je rad ukratko opis odnosno razmatranje prve faze pristupa nekim specifičnostima poučavanja ovih sadržaja, odnosno discipline. U slijedećoj fazi predlaže se stručna rasprava uz pokazatelje proizašle iz istraživanja koje bi dalo konkretne podatke izvedene iz mišljenja učenika i nastavnika o poučavanju predmeta Računalne mreže formalno strukturiranim anketiranjem.

LITERATURA

[1] https://www.learning-styles-online.com/overview (posjećeno veljača 2020.)

[2] http://tess.sbccd.org/Students/Self-Assessments/Learning_Style (posjećeno veljača 2020.)

[3] C. S. Hilas, A. Politis, „Motivating students' participation in a computer networks course by means of magic, drama and games“, SpringerPlus, 2014.

[4] A. Berglund, „What is good teaching of computer networks?“, ASEE/IEEE Frontiers in Education Conference 2003.

[5] C. Sanga, M. M. Shimba, M. J. Mahenge, „Virtual labs versus hands-on labs for teaching and learning computer networking: A comparison study“, International Journal of Research Studies in Educational Technology, 2017.

[6] A. N. Greca, S. E. Butler, J. K. Harris, „A proposed methodology to teach networking using portable network programming projects“, Issues In Information Systems, 2005.

[7] L. Tang, „ Study on Teaching Methods of Computer Network Course Based on Analogy and Problem-based Approach“, International Conference on Education and Development (ICED 2017)

[8] N. Linge, D. J. Parsons, „Problembased learning as an effective tool for teaching computer network design“, IEEE transactions on education, 2006.

[9] X. Zhang, Z. Xiong, Y. Ding, G. Wang, C. Ye, "Teaching content and methods in a *Computer Network* course and its effect on a local university", World Transactions on Engineering and Technology Education, 2014.

Rudarenje edukacijskih podataka: korištenje klasteriranja za predikciju studentskog uspjeha

K. Kačapor*, Z.Lagumdžija**
Univerzitet u Sarajevu, Ekonomski fakultet, Sarajevo, Bosna i Hercegovina,
*kemal.kacapor@efsa.unsa.ba
**zlatko.lagumdzija@efsa.unsa.ba

Sažetak - Ovaj rad istražuje cjelokupni postupak korištenja rudarjenja podataka (K-means klasteriranjem) za ekstrahiranje i pronalaženje korisnog znanja i zanimljivih uvida iz velike količine podataka o studentima prikupljenih tijekom online učenja, zabilježenih u datotekama dnevnika događaja (log files) poslužitelja Moodle LMS-a koji se koristi u programima učenja na daljinu na Ekonomskom fakultetu. Kao nesupervizirana tehnika, algoritam grupiranja K-means algoritam nema „učitelja“, moguće je da ti odvojivi klasteri zajedničke distribucije tih četiriju varijabli ne slijede jasno razdvajanje pri prvom pokušaju studentskog ispita. Stoga su rezultati algoritma grupiranja potvrđeni slijedećim kriterijem - uspješnost na prvom pokušaju studentskog ispita. Dakle, na kraju je provjereno da pronađeni klasteri studenata koji koriste četiri varijable također odgovaraju razlici uspjeha na prvom ispitnom pokušaju, tj. da li i ti klasteri "razdvajaju" studente u svojoj procjeni kao na prvom ispitu. Takvi uvidi predstavljaju korisne informacije nastavnicima i menadžmentu visokih učilišta pri donošenju informiranih odluka o tome kako unaprijediti nastavu ili podržati manje uspješne studente da ispune zahtjeve nastavnog plana i programa.

Ključne riječi - tehnika klasteriranja; rudarenje obrazovnih podataka; edukacijska analitika; blended learning.

I. UVOD

Zahvaljujući razvoju mrežnih i distribuiranih računarskih sistema obrazovni sistem je ušao u novu eru računarski potpomognute edukacije – online edukacije (eng. E-learning, online learning) i učenja na daljinu (eng. Distance learning - DL) pa smo svjedoci ekspanzije novih poslovnih modela koji se oslanjaju samo na online edukaciju, a proizvod su etabliranih i historijski dokazanih visokoškolskih i istraživačkih institucija. Edukacijski portali koji nude masovne otvorene online kurseve (eng. Masive Open Online Courses - MOOC) osnovani su i podržani od strane visoko rangiranih i svjetski priznatih univerziteta. Oni nude besplatne kurseve univerzitetskog nivoa, koje vode univerzitetski predavači, a kompanije kao što su Coursera i Udacity i edX su vodeći u ovom trenutku, prihode generišu kroz certifikaciju. Širi značaj MOOC-ova je da su oni dio trenda “raspakivanja” visokog obrazovanja.

Prelazak sa klasične na online edukaciju postavlja nove zahtjeve pred akademske institucije koje moraju usvojiti nove standarde organizacije i upravljanja nastavnim i administrativnim procesima. Studentski uspjeh nesumnjivo je najvažniji problem s kojim se institucije za visoko obrazovanje danas suočavaju. Izgleda da i pored izvrsne prihvaćenosti i rasta učenja na daljinu, izrazito postaje važno da pedagozi sistematiziraju njegovu upotrebu kako bi se smanjila neefikasnost i povećala efektivnost. Prisutno je povećanje potrebe za vještinama koje bi osigurale organizacijski uspjeh na svim nivoima kako bi se bolje suočili s pitanjima o obrazovanju, zadržavanju veće i raznovrsnije studentske populacije, prijemu studenata, prikupljanju novčanih sredstava i operativnoj efikasnosti [1].

U konstelaciji online komunikacije, jedina dodirna tačka između studenta i osoblja edukacijske institucije postaju LMS (eng. Learning Management Systems - LMS), web i mobilne platforme prepune tehnoloških inovacija, mrežnih i mobilnih komunikacija i serverskih procesa koji pružajući usluge svim učesnicima u procesu, istovremeno evidentiraju svaki klik i svaku transakciju svakog učesnika ponaosob.

Podaci koji proizlaze iz tih tehnologija potencijalno su dostupni za dubinsko istraživanje podataka - rudarenje podataka (eng. Data Mining), analizu (i tumačenje) pa je njihovo prikupljanje zaokupilo pažnju rukovodilaca obrazovnih institucija, akademika, istraživača i vladinih agencija.

II. SISTEMATSKI PREGLED LITERATURE

Uže područje interesovanja ovog rada je upotreba edukacijske analitike (LA) i rudarenja edukacijskih podataka (EDM) prikupljenih u informacionim sistemima univerzitetske ustanove tokom procesa online edukacije, konkretno Ekonomskog fakulteta Univerziteta u Sarajevu. Zadatak je istražiti korištenje rudarenja podataka za ekstrakciju i pronalaženje korisnih i interesantnih saznanja / znanja iz podataka prikupljenih u procesu online edukacije, a koji se nalaze u bazama podataka Moodle LMS-a koji se koristi za potrebe Studija učenja na daljinu na Ekonomskom fakultetu Univerziteta u Sarajevu, a koje mogu sa novim znanjima biti korisne učesnicima u procesu edukacije, u unaprijeđenju samog procesa nastave i efikasnosti online edukacije.

Za analizu literature korištena je metodologija sistematske i mrežna analiza literature (eng. Systematic literature network analysis - SLNA) predložena od strane [5]. Istraživanje sekundarnih podatka iz literature, radova i naučnih članaka objavljenih u okviru tematika data

mininga, edukacijskog data mininga, learning analyticsa te knowledge managementa rađeno je kroz pretrage u online bazama: Web of Science, Francis & Taylor, EconLit, Springer, JSTOR, Ebsco, Emerald, ProQuest, Science Direct i Google scholar.

Rezultati analize literature pokazuju da se rudarenje podataka kao historijski stariji fenomen, pojavljuje prvenstevno u oblasti informacionih tehnologija i računarski podržanih algoritama mašinskog učenja [6]–[8]. Korisnu primjenu istraživača prvenstevno u oblasti informacionih tehnologija i u akademskom sektoru su prepoznali mnogi autori [7], [9]–[14]. Prema [2] "edukacijska analitika se bavi prikupljanjem tragova koje učenici ostave iza sebe i korištenje tih tragova za unapređenje samog učenja", dok [3] definira edukacijsku analitiku kao "proces razvoja akcijskih uvida kroz definiranje problema i primjene statističkih modela i analiza nad postojećim historijskim ili simuliranim budućim podacima". U širem smislu prema [4] kristaliziraju se tri polja u oblasti analitike u edukaciji i to edukacijska analitika, akademska analitika i rudarenje edukacijskih podataka.

Većina problema koji se rešavaju rudarenjem podataka, mogu se svrstati u dva modela i to deskriptivni modeli i prediktivni modeli [14]. Deskriptivni modeli obično koriste nesupervizirano mašinsko učenje gdje nema ciljane referentne vrijednosti i gdje je zadatak računara da identificira uzorke u podacima koji mogu biti značajni. Drugim riječima, nesupervizirano mašinsko učenje koristi se za prepoznavanje odnosa u podacima, strukture i međupovezanosti podataka [15]. Prediktivni modeli najčešće koriste supervizirano mašinsko učenje koje je usmjereno prema cilju, poznatoj vrijednosti, a zadatak je u tom slučaju objasniti vrijednost ciljane varijable i usmjeravamo računar, putem modela da kaže kako ga procijeniti, klasificirati ili predvidjeti u odnosu na poznate vrijednosti nekih drugih varijabli. Supervizirano mašinsko učenje se koristi za objašnjavanje tih odnosa koji su pronađeni ili postoje.

Klasteriranje ili grupiranje je zadatak segmentiranja različitih grupa u slične podgrupe ili klastere [17]. Intuitivno, uzorci unutar valjanog klastera sličniji su međusobno nego uzorcima koji pripadaju različitim klasterima [18]. Kod grupiranja nema predviđenih klasa i nema primjera. Zapisi su grupirani zajedno bazirajući se na vlastitoj sličnosti. Grupe moraju međusobno biti što manje slične. Ovdje ne postoje prethodno definisane klase niti atribute po kojima će se podaci svrstavati u grupe klastera. Metoda se koristi za početno segmentiranje, kao i za praćenje trendova u drugim situacijama. Na analitičaru je da utvrdi koje značenje, ako ga ima, treba dati dobijenom klasteru.

Veliki broj opcija i kombinacija čini eksperiment "pokušati sve" neisplativim za bilo koji praktični problem rudarenja podataka, a u rudarenju edukacijskih podataka a prema [16] najčešće se koriste klasifikacija i klasteriranje koji u zbiru čine najčešći izbor u 69% pregledanih istraživačkih projekata. Za provođenje klaster metode postoje različiti algoritmi od kojih je korišten k-Means klasteriranje. Za klasteriranje je K-prosječni (eng. K-means algorithm) algoritam korišten u 19 slučajeva [16] rudarenja edukacijskih podataka kao i u [19], [20] te konkretno nad podacima Moodle sistema [21].

III. DEFINISANJE HIPOTEZE

U skladu sa rezultatima analize literature mogu se definisati sljedeći ciljevi istraživanja (a) istražiti načine koji omogućavaju ekstrakciju, transformaciju i pripremu podataka iz sistemske evidencije događaja (eng. log files) Moodle LMS okruženja koji će se koristiti u daljem procesu rudarenja podataka korištenjem K-means clustering tehnike, (b) u ekstrahiranim podacima iz LMS-a korištenjem K-means clustering tehnike identificirati slične uzorke korištenja sistema i identificirati grupe studenata sa sličnim navikama i načinom korištenja sistema.

U okviru tako determiniranog problema i predmeta istraživanja postavlja se i temeljna hipoteza:

- H: Korištenjem nesuperviziranog mašinskog učenja moguće je izvršiti prepoznavanja studenata za preraspodjelu po grupama sličnih preferencija i dinamike usvajanja znanja.

Argumenti koji podupiru navvedenu hipotezu su (a) podaci o ponašanju studenata i posjetama korisnika sistema koji su sadržani u bazama podataka LMS-a su preobimni za pregledanje i analizu na osnovu standardnih tehnika izvještavanja i (b) dalje rapidno širenje online edukacije nagovještava sve veće nagomilavanje podataka u sistemima koji će zahtijevati dodatne metode i alate procesiranja kao pomoć nastavnicima i administraciji u donošenju odluka, (c) upotreba rudarenja podataka u edukacijskim ustanovama i nastavnom procesu, može omogućiti otkrivanje skrivenih i interesantnih znanja koja su sadržana u edukacijskim podacima, uključivo i podacima LMS-a, a koja mogu doprinijeti poboljšanju efikasnosti online edukacije.

IV. EMPIRIJSKO ISTRAŽIVANJE

Proces rudarenja podataka u ovom radu vođen je CRISP metodološkim okvirom za rudarenje podataka, koji nije intelektualno vlasništvo nekog od proizvođača softverskih rješenja za rudarenje podataka, već je slobodno dostupan za besplatno korištenje i aplikativno je invarijantan. CRISP-DM je nastao kao rezultat najboljih praksi u različitim industrijama i korišten je kao vodilja u mnogim studijama rudarenja edukacijskih podataka [24], [25], [26], [27], [28], [29], [30], [31].

Za potrebe istraživanja izvršena je selekcija tehnologije i tehnoloških alata. Za ekstrakciju i agregaciju podataka sa izborom relevantnih atributa iz posmatranih skupova podataka korišten je programski jezik Python i SQL upiti nad MySQL bazom Moodle LMS-a. Za modeliranje, rudarenje podataka / mašinsko učenje korišten je RapidMiner i analitički alat R. Uz ove alate korištene su i druge tehnike manipulacije podacima potrebne za dodatnu pripremu trening i testnog skupa podataka iz Moodle LMS-a.

Empirijsko istraživanje je izvedeno na Ekonomskom fakultetu Univerziteta u Sarajevu uz korištenje podataka iz baze podataka (eng. Database management system -

DBMS) Moodle LMS okruženja, nastalim tokom vremenskog perioda od januara 2015 tokom dva semestra, uzimajući u obzir predmete (ispite, kurseve) zajedničkih smjerova Distance learning studija. Svi kursevi na Moodle sistemu DL EFSA, su kombinirani (eng. Blended learning) kursevi jednako vremenski raspoređeni i sastoje se od ukupno 15 sedmica nastave, od kojih je 1., 5., 8., 12., i 15. sedmica in-class, a ostale su online, to znači 30% face to face, a 70% online. Svi kursevi imaju vrlo sličnu temporalnu strukturu (edukacijski dizajn) sedmične organizacije resursa za čitanje, nekog audio-video materijala za slušanje/gledanje te nekih aktivnosti (kvizova) i diskusija. Postoje razlike i odstupanja u strukturi koje su dio akademske slobode nastavnika i njegovih pedagoških principa. Na osnovu analize edukacijskog dizajna kurseva na Moodle sistemu DL studija EFSA, moguće je identificirati module koje nastavnici koriste tokom edukacijskog procesa. Sve ove identificirane module/aktivnosti na sistemu su na neki način zabilježene u evidencijama događaja. Ako ih pogledamo i grupiramo po nekom logičnom pragmatičnom redoslijedu mogu se identificirati slijedeće funkcionalne i logičke kategorije podataka, a samim tim logičke i funkcionalne kategorije varijabli koje će u procesu rudarenja podataka biti kreirane:

- Društvena aktivnost, to su forumi, chat i personalne poruke koje korisnici razmijenjiju (student – student, student – nastavnik)
- Reading resursi, svi nastavni materijali koji se objavljuju kao tekstualne, prezentacijske ili slikovne fajle
- Multimedija, ili materijali koji omogućavaju pregled komvinacije audio-video materijala na našem sistemu najčešće u SCORM formatu
- Aktivnosti, ili moduli koji potiču kognitivno prisustvo kao što su assignmenti ili zadaće
- Testiranje znanja, koje se na LMS-u EFSA izvodi putem sedmičnih kvizova
- Predanost (eng. Commitment), sveukupni kontinuitet posjeta, učestalost i motivisanost za učenje.

U skladu nalazima iz pregleda literature i teoretskog okvira može se zaključiti da je jedan od najznačajnijih koraka u prediktivnoj analitici i modeliranju napraviti pažljiv odabir mjernih varijabli. Usvajajući deduktivni pristup istraživanju, modeliranje je u potpunosti bazirano na teorijskim pretpostavkama [7], [16], [22], ali i u skladu sa preporukama [23] p.81, zdravorazumskim pretpostavkama (iskustvo autora) koje određuju izbor mjernih varijabli. U isto vrijeme moramo misliti i o interpretabilnosti rezultirajućih klastera i misljenja smo da bi klastering na više od 4 varijable bilo teško interpretirati i grafički kompaktno predstaviti pa su kao varijable odabrane:

- prosječni broj sesija po sedmici,
- prosječni broj pregleda bilo kojeg resursa u sedmici,
- prosječno sedmično trajanje sesije u minutama,
- prosječan sedmični broj klikova.

Izvršena je obrada nesuperviziranim algoritmom mašinskog učenja prema tako identificiranim varijablama i konačno, pretpostavljen njihov uticaj na uspjeh studenata.

Ako se uzme u obzir šta su značajniji istraživači [6], [16], [32], [33] usvojim istraživanjima koristili kao mjerne varijable kao i aktuelne postavke korištenja Moodle LMSa na DL studiju EFSA bilo je neophodno utvrditi slijedeće:

- Identificirati vremenski i datumski početak i kraj svake sedimice tokom semestra.
- Identificirati sesiju svakog korisnika [34]. Sesija počinje prvim klikom korisnika na konkretnom kursu (ne na nekom drugom dijelu sistema) do trenutka odjave (bilo sistemske ili da se korisnik sam odjavi sa sistema) ili najviše 30 minuta od posljednje akcije na kursu. Identificirati sve sesije korisnika tokom jedne sedmice, pobrojati sve klikove koje je korisnik imao tokom svake sesije, napraviti sumu svih klikova i izračunati prosječan broj klikova po sesiji, broj sesija u toj sedmici i prosječno trajanje sesije.
- Pobrojati klikove koji se odnose na pregled edukacijskih resursa (readings) kao što su dokumenti, prezentacije, i sl. (resource, book, page, folder, folder download, external url) korisnika, na konkretnom kursu u toj sedmici.

Ovo su ujedno i jedini MOGUĆI podaci koji se mogu agregirati iz postojećih podataka zbog limitiranosti evidentiranih događaja u skladu sa mogućnošću evidencije događaja Moodle logova. Na osnovu eksplorativnog pregleda podataka može se zaključiti da treba uzeti u obzir samo kurseve na kojima je bilo više od 70 studenata, a ostali kursevi su zbog malog uzorka studenata izuzeti iz razmatranja. Ova odluka se donosi na osnovu prirode rada sa podacima od strane algoritama mašinskog učenja, koji kako je ranije navedeno, zahtjevaju skup podataka za trening i skup podataka za testiranje, čime se inicijalni skup podataka znatno umanjuje, tj svodi se na broj uzoraka koji je ispod prihvatljivog nivoa. Osim toga, obuhvaćeni su kursevi prve (zimski i ljetni semestar) i druge godine studija (samo zimski semestar sa zajedničkim predmetima). Kako studenti slušaju iste (obavezne) tj. zajedničke predmete moguće je da se u više kurseva pojavljuje isti student, barem je ovo situacija koja odgovara prvi put upisanom studentu prve godine.

Nadalje, za analizu je planirano ekstrahirati slijedeće semestre u periodu dvije školske godine, čime su obuhvaćeni jedan ljetni i jedan ljetnji semestar a to znači i svi kursevi zajedničke osnove na prvoj i drugoj godini. Podaci moraju biti podijeljeni u vremenske periode po semestrima, sa jasnim oznakama kurseva, te pripadnosti po korisniku.

Za struktuiranu strategiju ispitivanja postojanja klastera unutar naših podataka koristili smo prethodno opisan izbor 4 varijable u modelu i nakon toga na njima primijeniti nesupervizirani klastering algoritam [18]. Na ovaj način ćemo bar u nekoj mjeri struktuirati nesupervizirani klastering algoritam sa stajališta selekcije varijabli koje bi identifikaciju pojedinih klastera činile relevantnim s obzirom na naš primarni ishod prolaznosti na prvom ispitnom pokušaju.

Naime, za razliku od superviziranog učenja, koje rezultira modelom čije je performanse na kraju moguće i provjeriti uspoređujući predviđene ishode s onim postojećim stvarnim ishodima studenata, klastering algoritmi kao nesupervizirane metode učenja, nose određenu dozu proizvoljnosti sa sobom. U obje faze, selekcije algoritma i validacije rezultata i zaključaka nesuperviziranog učenja (što klastering algoritmi gotovo isključivo jesu) neophodno je osloniti se na heuristike. U naučnom kontekstu to često znači teško prihvatljivu činjenicu, da je efektivnost nerijetko pitanje mišljenja. S tim na umu, pokušali smo na prethodno opisani način, makar prvi dio, selekciju varijabli na kojima ćemo vršiti klastering, struktuirati na neproizvoljan način. Vođeni smo idejom da ukoliko postoje jasno odvojive grupe unutar zajedničke distribucije 4 najvažnije varijable iz predikcijskog modela, njih će biti moguće detektovati pomoću klastering algoritma. Kako klastering algoritam nema "učitelja" moguće je da ove odvojive komponente zajedničke distribucije 4 najvažnije varijable ne slijede i po prolaznosti na prvom ispitnom pokušaju sličnu

```
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  493

                  | prolaznaPrvomIspitu
          cluster |         0 |         1 | Row Total |
------------------|-----------|-----------|-----------|
                1 |        39 |        68 |       107 |
                  |    12.761 |    22.783 |           |
                  |     0.364 |     0.636 |     0.217 |
                  |     0.123 |     0.384 |           |
                  |     0.079 |     0.138 |           |
------------------|-----------|-----------|-----------|
                2 |       277 |       109 |       386 |
                  |     3.537 |     6.315 |           |
                  |     0.718 |     0.282 |     0.783 |
                  |     0.877 |     0.616 |           |
                  |     0.562 |     0.221 |           |
------------------|-----------|-----------|-----------|
     Column Total |       316 |       177 |       493 |
                  |     0.641 |     0.359 |           |
------------------|-----------|-----------|-----------|

Statistics for All Table Factors

Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  45.39713     d.f. =  1     p =  1.608703e-11

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  43.87559     d.f. =  1     p =  3.499267e-11
```

Figure 1. Distribucija glavnog ishoda po pripadnosti rezultirajućem klasteru

separaciju. Ishode našeg klastering algoritma validiramo upravo ovim kriterijem. Provjerili smo da li se pronađeni klasteri studenata koristeći 4 najvažnije varijable podudaraju i po razlici uspjeha na prvom ispitnom pokušaju, tj. da li ovi klasteri također "dobro odvajaju" studente po njihovoj ocjeni na ispitu.

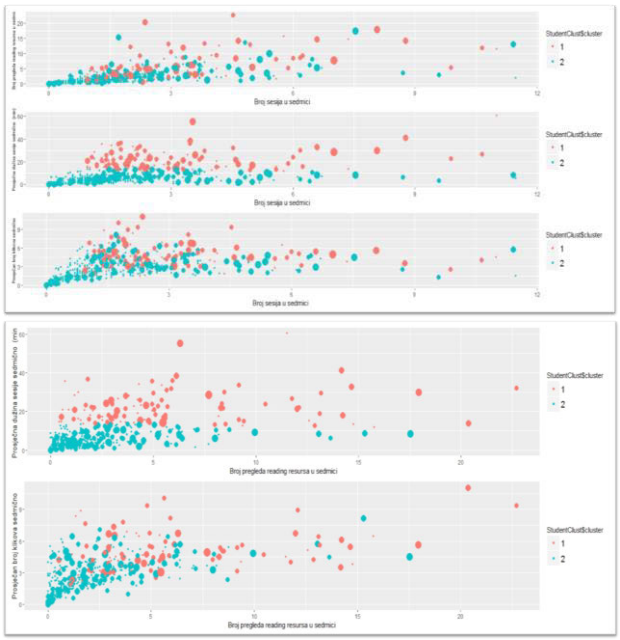
Za klastering koristimo K-means algoritam. Ukratko, algoritam počinje s nekoliko nasumičnih particija uzorka (mi smo izabrali 20) od kojih bira onu particiju koja ima najmanju varijabilnost unutar klastera (ovo je suma kvadratnih odstupanja svake observacije od centra svog klastera sabrana za sve klastere). U drugom koraku za svaku tačku u uzorku nalazimo klaster čiji je trenutni centar (prosječna vrijednost) klastera najbliži toj tački (blizina je definirana kao euklidski norm). Ako su trenutni klaster u kojem tačka jeste i klaster čiji je centar najbliži različiti, tačka mijenja klaster. Nakon što iteriramo sve tačke u uzorku na ovaj način, nanovo izračunamo centre klastera. Ovo se ponavlja dok nijedna tačka ne promijeni klaster i u zadnjoj iteraciji podjela na klastere finalna [18].

K-means algoritam zahtijeva specifikaciju broja klastera koje želimo pronaći. U nasem slučaju isprobali smo 2, 3 i 4 klastera. Odlučili smo se za 2 klastera jer je povećanje klastera od 2 do 3 znakovito smanjilo onaj udio ukupne sume kvadratnih odstupanja koji se nalazi u klasterima, dok je povećanje s 3 na 4 ovaj udio samo neznatno smanjilo.

K-means algoritmu je potreban početni input o broju klastera koje želimo naći. Algoritam smo započeli tražeći 2 klastera. Figure 1 Distribucija glavnog ishoda po pripadnosti rezultirajućem klasteru prikazuje distribuciju glavnog ishoda po pripadnosti rezultirajućem klasteru.

V. ANALIZA REZULTATA

Nakon što je algoritam našao klastere bazirane na 4 najvažnije varijable iz RF algoritma, koristeci Pearson Chi-Square test provjerili smo da li je distribucija prolaznosti na prvom ispitu signifkantno različita u zavisnosti od pripadnosti nesupervizirano pronađenih klastera.



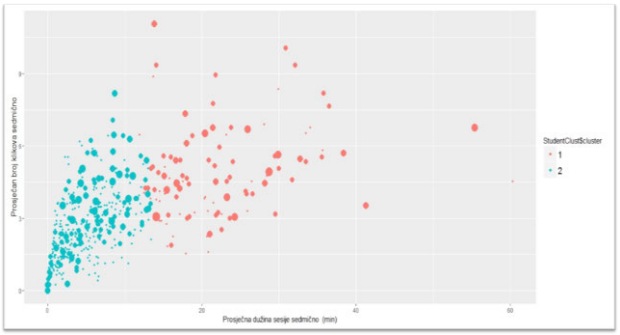


Figure 2. Distribucija glavnog ishoda po pripadnosti rezultirajućem klasteru

Iz Chi-square p-vrijednosti koja je $p<0.001$ možemo vidjeti da je distribucija glavnog ishoda statistički signifikantno nejednaka u dva nesupervizirano pronađena klastera. Prvi ili akademski bolji klaster ima prolaznost od oko 64% dok je u drugom ili akademski lošijem klasteru prolaznost na prvom ispitnom pokušaju oko 28%.

Ovo je ohrabrujuće saznanje jer iako RF algoritam ne nudi tačan funkcionalni oblik kompozicije prediktora, no iz nesuperviziranog klastering algoritma se čini da je podudarnost zatvorenih klastera ovih važnih varijabli i akademskog uspjeha studenta na prvom ispitnom pokušaju dovoljno podudarna da bi mogli pričati o postojanju makar djelimičnih funkcionalnih oblika koje ove četiri varijable kvalificira kao dobre prediktore akademskog uspjeha.

Slika 2 prikazuje bivarijatne grafove svih 6 mogućih parova 4 varijable na kojima smo implementirali klastering algoritam. Iz grafova u kojima sudjeluje varijabla „Prosječna dužina sesije sedmično (min)“ čini se da postoji gotovo linearna separacija u pronađenim klasterima po ovoj varijabli.

U grafovima sa slike 2 veličina tačke je proporcionalna ocjeni iz prvog ispita. U skladu s zaključkom koji smo donijeli pomoću Pearsonovog Chi-Square testa o raspodjeli akademske uspješnosti na prvom ispitu među klasterima u klasteru broj 1 veličine tačaka su preovladavajući veće nego u klasteru 2.

VI. ZAKLJUČAK

Ovaj rad na izvjestan sistematičan način pruža uvid u trentno stanje znanja o edukacijskom rudarenju podataka, kao oblasti kojoj je primarni cilj bolje razumijevanje učenja i obezbjeđivanja informacija o procesu učenja kroz korištenje velikih edukacijskih setova podataka.

Ono što je evidentno je da pojedinačna naučna disciplina nije dovoljna sama po sebi da bi se napravilo istraživanje edukacijskog rudarenja podataka. Kompjuterska nauka obezbjeđuje ekspertizu o radu sa velikim setovima podataka, u kontekstu mašinskog učenja i tehnika rudarenje podataka koje se mogu fantastično primijeniti na setovima podataka koji imaju milione zapisa, ali jednako dobro i za rješavanje problema iz stvarnog svijeta „češljanjem“ izvornih podataka kako bi se eliminisale greške u podacima na samom početku kojim se sprječava propagiranje tih grešaka do krajnjih rezultata.

Korištenjem klasteriranja kao metode nesupervizirane klasifikacije, pokazano je da i nesupevizirani algoritmi, u pretpostavljenoj situaciji nepoznate target varijable, prihvatljivo dobro vrše odvajanje studenata u različite klastere. Validacijom ovog algoritma, korištenjem poznatih target varijabli iz seta podataka ishoda na ispitima pokazano je da je tačnost ovog algoritma na prihvatljivom nivou, čime je dokazana druga pomoćna hipoteza. Ovdje je zaključeno da se preciznijim zahtjevom na isti način mogu kreirati varijable koje će studente izdvojiti u grupe sličnih karakteristika po tim varijablama.

LITERATURA

[1] A. Van Barneveld, K. Arnold, and J. Campbell, “Analytics in higher education: Establishing a common language,” Educ. Learn. Initiat., no. January, pp. 1–11, 2012.

[2] E. Duval, “Learning Analytics and Educational Data Mining,” 2012. [Online]. Available: https://erikduval.wordpress.com/2012/01/30/learning-analytics-and-educational-data-mining/. [Accessed: 01-Jan-2016].

[3] A. Cooper, “What is Analytics? Definition and Essential Characteristics,” CETIS Anal. Ser., vol. 1, no. 5, pp. 1–10, 2012.

[4] G. Siemens and P. Long, “Penetrating the Fog: Analytics in Learning and Education,” Educ. Rev., vol. 46, pp. 30–32, 2011.

[5] C. Colicchia and F. Strozzi, “Supply chain risk management: a new methodology for a systematic literature review,” Supply Chain Manag., vol. 17, no. 4, pp. 403–418, 2012.

[6] C. Romero, S. Ventura, and E. García, “Data mining in course management systems: Moodle case study and tutorial,” Comput. Educ., vol. 51, no. 1, pp. 368–384, 2008.

[7] C. Romero and S. Ventura, “Educational data mining: A survey from 1995 to 2005,” Expert Syst. Appl., vol. 33, no. 1, pp. 135–146, 2007.

[8] J. Alcalá-Fdez et al., “KEEL: a software tool to assess evolutionary algorithms for data mining problems,” Soft Comput., vol. 13, no. 3, pp. 307–318, 2008.

[9] M. Blagojević and Ž. Micić, “A web-based intelligent report e-learning system using data mining techniques,” Comput. Electr. Eng., vol. 39, no. 2, pp. 465–474, 2013.

[10] Y. H. Wang and H. C. Liao, “Data mining for adaptive learning in a TESL-based e-learning system,” Expert Syst. Appl., vol. 38, no. 6, pp. 6480–6485, 2011.

[11] Y. Aydogdu and Z. Tanrikulu, “Corporate E-Learning Success Model Development by Using Data Mining Methodologies,” Egit. Ve Bilim. Sci., vol. 38, no. 170, pp. 95–111, 2013.

[12] M. Jovanovic, M. Vukicevic, M. Milovanovic, and M. Minovic, “Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study,” Int. J. Comput. Intell. Syst., vol. 5, no. December 2012, pp. 597–610, 2012.

[13] J. Srivastava and D. A. K. Srivastava, “Data Mining in Education Sector: A Review,” Spec. Conf. Issue Natl. Conf. Cloud Comput. Big Data, pp. 184–190, 2013.

[14] P. Thakar, A. Mehta, and Manisha, “Performance analysis and prediction in educational data mining: A research travelogue,” Int. J. Comput. Appl., vol. 110, no. 15, pp. 60–68, 2015.

[15] T. Ben-Zvi, Data analysis: A Roadmap to better decision-making methods, techniques and applications. Germany: LAP LAMBERT Academic Publishing, 2012.

[16] A. Peña-Ayala, “Educational data mining: A survey and a data mining-based analysis of recent works,” Expert Syst. Appl., vol. 41, no. 4 PART 1, pp. 1432–1462, 2014.

[17] G. K. Bapu, M. B. Ashok, S. P. Shamrao, and S. G. Tanaji, “Clustering Moodle Data As a Tool For Profiling Students,” Int. J. Adv. Res. Comput. Eng. Technol., vol. 4, no. 3, pp. 121–126, 2015.

[18] T. Hastie, R. Tibshirani, and J. Friedman, “The Elements of Statistical Learning,” Elements, vol. 1, pp. 337–387, 2009.

[19] V. Kumar and N. Rathi, “Knowledge discovery from database Using an integration of clustering and classification,” Int. J. Adv. Comput. Sci. Appl., vol. 2, no. 3, pp. 29–33, 2011.

[20] J. Sancho, “Learning Opportunities for Mass Collaboration Projects Through Learning Analytics: A Case Study,” Rev. Iberoam. Tecnol. del Aprendiz., vol. 11, no. 3, pp. 148–158, 2016.

[21] S. Sisovic, M. Matetic, and M. B. Bakaric, “Clustering of imbalanced moodle data for early alert of student failure,” SAMI 2016 - IEEE 14th Int. Symp. Appl. Mach. Intell. Informatics - Proc., pp. 165–170, 2016.

[22] J. W. You, “Identifying significant indicators using LMS data to predict course achievement in online learning,” Internet High. Educ., vol. 29, 2016.

[23] L. Cohen, L. Manion, and K. Morrison, Research Methods in Education, Sixth Edit. 2007.

[24] D. Delen, “A comparative analysis of machine learning techniques for student retention management,” Decis. Support Syst., vol. 49, no. 4, pp. 498–506, 2010.

[25] C. Vialardi et al., “A data mining approach to guide students through the enrollment process based on academic performance,” User Model. User-adapt. Interact., vol. 21, no. 1–2, pp. 217–248, 2011.

[26] S. Agarwal, G. N. Pandey, and M. D. Tiwari, “Data Mining in Education: Data Classification and Decision Tree Approach,” Int. J. e-Education, e-Management e-Learning, vol. 2, no. 2, pp. 140–145, 2012.

[27] R. A. Huebner, “A survey of educational data mining Research,” Nature, vol. 194, pp. 1006–1006, 2013.

[28] M. Chalaris, S. Gritzalis, M. Maragoudakis, C. Sgouropoulou, and A. Tsolakidis, “Improving Quality of Educational Processes Providing New Knowledge using Data Mining Techniques,” Procedia - Soc. Behav. Sci., vol. 147, pp. 390–397, 2014.

[29] M. A. Yehuala, “Application Of Data Mining Techniques For Student Success And Failure Prediction (The Case Of Debre_Markos University),” Int. J. Sci. Technol. Res., vol. 4, no. 04, 2015.

[30] C. Y. Weston, “ON THE K-NEAREST NEIGHBOR APPROACH TO THE GENERATION OF FUZZY RULES FOR COLLEGE STUDENT PERFORMANCE PREDICTION,” 2015.

[31] J. Sykurska, “Data Mining in Learning Analytics,” 2016.

[32] C. Romero and S. Ventura, “Data mining in education,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 3, no. 1, 2013.

[33] J. W. You, “Identifying significant indicators using LMS data to predict course achievement in online learning,” Internet High. Educ., vol. 29, pp. 23–30, 2016.

[34] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, “Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS,” IEEE Trans. Learn. Technol., vol. 1382, no. c, pp. 1–1, 2016.

Robotika u funkciji razvoja vještina 21. stoljeća

Ivana Ružić
I. osnovna škola Čakovec, Čakovec, Hrvatska
ivana.ruzic@skole.hr

***Sažetak* - Robotika u osnovnom obrazovanju predstavlja transformacijski alat ključan za učenje STEM-a, razvoj računalnog razmišljanja i vještina programiranja. Iako je robotika u osnovnom obrazovanju prisutna od kasnih 1990.-ih, ona posljednjih godina postaje sve popularnija među učenicima i učiteljima. Često robotika nije dobro integrirana kao tehnološki alat za učenje u redovitim školskim uvjetima. Cilj rada je ukazati na važnost uvođenja robotike u osnovno obrazovanje kao naprednog tehnološkog alata za učenje, u redovne kurikulume i objasniti na koji način robotika pomaže učenicima da se pripreme za budućnost.**

Ključne riječi – robotika, osnovno obrazovanje, STEM, računalno razmišljanje, programiranje

I. UVOD

Tehnologija je sveprisutna i integrirana u svaki aspekt našeg života. Ubrzani tehnološki napredak jača međusobnu svepovezanost koju donosi moć interneta i društvenih medija. Novi tehnološki alati uvode se u naš život brže nego ikad prije. Naslovi vijesti s različitim robotskim inovacijama snažan su pokazatelj koliko je popularna tehnologija robotike u industriji, obrazovanju, kućanstvu i obavljanju svakodnevnih uobičajenih poslova.

U bližoj prošlosti dogodili su se veliki zaokreti u svakodnevnoj primjeni robotike. Primjerice, 2013. godine kada je Google započeo intenzivnu suradnju s osam kompanija renomiranih u području robotike. Jedan od ključnih datuma je i 5. lipanj 2014. kada je japanska tvrtka Softbank Mobile, u suradnji s francuskom tvrtkom Aldebaran Robotics, predstavila je Pepper, na svijetu prvi osobni humanoidni robot sposoban pomoći ljudima čitanjem i odgovaranjem na njihove ljudske emocije. U veljači 2015. Amazon je predstavio svoj sustav isporuke pomoću dronova. 2014. godine NAO Aldebaran Robotics predstavio je autonomni i humanoid robot kojeg korisnici mogu samostalno programirati, koristi se obrazovnim ustanovama za razvoj algoritama ali i za istraživanje djece s autizmom.

Naši učenici digitalni su urođenici koji odrastaju koristeći pametne uređaje. Iako su njihovi životi ispunjeni tehnologijom, oni rijetko razmišljaju i razumiju kako njihovi uređaji rade. Kada njihov pametni uređaj prestaje raditi, oni traže novi uređaj ne razmišljajući da bi se mogao otkloniti kvar i poteškoća u radu postojećeg, kao da su pametni uređaji jednokratni. Naše učenike nismo naučili da propitkuju i promišljaju o tehnologiji koju koriste. Oni su uglavnom pasivni korisnici pametnih uređaja i njihovih alata.

U posljednje vrijeme interes za robotiku porastao je zadivljujućom brzinom. Tehnologija robotike uvedena je u razna područja poput medicine, njege starijih i nemoćnih osoba, rehabilitacije, obrazovanja, kućanskih aparata, potrage i spašavanja, automobilske industrije itd. Svijet i globalna ekonomija mijenjaju se takvom brzinom da je nemoguće predvidjeti kako će izgledati završetak sljedećeg tjedna [1]. Svijet se ubrzano mijenja ali javno obrazovanje održava gotovo isti sustav od svog organiziranog uvođenja. Iako se ulažu ogromni napori u obrazovnim reformama širom svijeta, problem uglavnom ostaje isti: većina škola pokušava pripremiti učenike za budućnost temeljeno na onome što je učinjeno u prošlosti [2].

Budući da moderna globalna ekonomija snažno utječe na tehnološki povezane industrije, stjecanje znanja i razvoj vještina programiranja i računalnog razmišljanja presudno je za uspjeh budućih generacija učenika.

Trenutno, 44% europskih građana nema razvijene osnovne digitalne vještine. 37% ljudi na tržištu rada - poljoprivrednici, zaposlenici u bankama i industriji- također nemaju dovoljno razvijenih digitalnih vještina, usprkos sve većoj potrebi za takvim vještinama na svim poslovima [3].

Europi također nedostaje kvalificiranih stručnjaka iz područja ICT kako bi se popunio sve veći broj slobodnih radnih mjesta u svim sektorima gospodarstva. Ključno pitanje na kojem se temelji je potreba za modernizacijom naših sustava obrazovanja i osposobljavanja koji trenutno ne pripremaju mlade ljude dovoljno za digitalnu ekonomiju i društvo, te prelazak na cjeloživotno učenje kako bi ljudi mogli prilagoditi svoje skupove vještina njihova životna vremena po potrebi [4].

U cilju ranog razvoja računalnog razmišljanja i programiranja te općenito STEM područja u obrazovanju, posljednjih godina se sve više razvijaju i jačaju globalne inicijative i pokreti: Hour of Code, EU Code Week, MakeX, First Lego League, World Robot Olympiad, i mnogi drugi.

Korištenje robotike u obrazovanju učenicima omogućuje da razvijaju i stvaraju vlastite inovativne tehnološke kreacije, omogućuju poticanje njihovog aktivnog sudjelovanja u samom činu izrade, što ih potiče da postanu aktivni stvaratelji, a ne potrošači tehnoloških proizvoda u budućnost [5].

II. ROBOTIKA U OBRAZOVANJU KAO MEĐUPREDMETNO PODRUČJE

STEM u obrazovanju je uobičajeno shvaćen kao obrazovni pristup koji integrira znanost, tehnologiju, inženjerstvo i matematiku. STEM se uvodi u obrazovanje sa ciljem razvoja većeg interesa i motivacije učenika za to područje kako bi učenici razvijali vlastita znanja, vještine

i stavove za buduće obrazovanje i buduća zanimanja u STEM području. STEM se uvodi u obrazovanje kako bi se zadovoljile potrebe na tržištu rada. Na tržištu rada, svakodnevno se povećava nedostatak obrazovanih STEM stručnjaka.

STEM u obrazovanju zahtijeva cjeloviti i međupredmetni pristup u integriranju znanja i vještina, traženje i predlaganje rješenja za probleme iz stvarnog svijeta. U kontekstu STEM obrazovanja, potrebno je povezivanje i suradnja između obrazovnih ustanova, zajednice, lokalnih tvrtki i globalnih kompanija.

Robotika je učinkovit alat za učenje temeljeno na projektima gdje je nužno povezivanje znanja i vještina iz područja STEM-a, programiranja, računalnog razmišljanja i inženjerstva. Uvođenjem robotike u obrazovanje učenicima pružamo mogućnosti istraživanja i promišljanja o načinima funkcioniranja tehnologije u stvarnom životu. Prilikom dizajniranja, konstruiranja, programiranja i dokumentiranja rada s autonomnim robotima, učenici ne samo da uče kako tehnologija funkcionira, već također primjenjuju vještine i sadržajno znanje naučeno u školi na smislen i uzbudljiv način. Robotika u obrazovanju je bogata mogućnostima za integraciju ne samo u području STEM-a, nego i mnogih drugih područja, uključujući društvene predmete, ples, glazbu, sport i umjetnost. Istovremeno, uvođenje robotike u obrazovanje učenicima se pruža prilika za pronalaženje novih načina za zajednički, timski rad na poticanju razvoja sposobnosti međusobne suradnje, rada u timu, komunikacije i izražavanja pomoću tehnološki naprednih alata, razvoja vještina rješavanja problema, kritičkog razmišljanja, kreativnosti i inovacija.

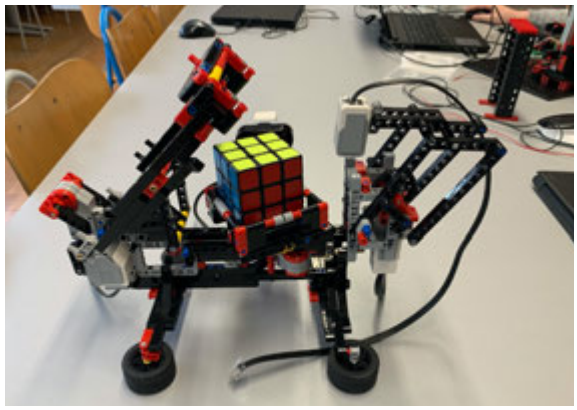
Robotika u obrazovanju je alat za učenje koji poboljšava učenikovo iskustvo putem praktičnog učenja. Ona pruža zabavno i uzbudljivo okruženje za učenje i stvaranje. Takvo okruženje učenike kontinuirano motivira za učenje, promišljanje i istraživanje, kako bi razvijali vlastite vještine i znanja koja su im potrebna za ostvarivanje vlastitih ciljeva na putu prema uspješnom rješavanju stvarnih problema u području STEM-a.

Robotika u obrazovanju nailazi i na poteškoće i probleme. Prije svega poteškoće se očituju u nedostatku obrazovnih materijala. Razvoj, stvaranje i dijeljenje obrazovnih materijala ne prati razvoj tehnologije. Značajan problem je nabava najnovije opreme. Na tržištima malih država poput Hrvatske, često nije moguće nabaviti opremu koja prati najnovije trendove u razvoju tehnologije. Online kupovina još uvijek nije svakodnevnica u našim obrazovnim ustanovama i često se u strahu od eventualnih problema u hrvatskim školama odustaje od ovakvog načina kupovine. Značajan problem je motivacija učitelja i nastavnika za dodatnim usavršavanjem i praćenjem razvoja tehnologije. Oni često o vlastitom trošku nabavljaju opremu i stručno se usavršavaju u svoje slobodno vrijeme.

Nastavak rada prikazuje primjere integracija STEM-a, računalnog razmišljanja, programiranja i inženjerskih vještina u obrazovanju kako bi učenici upoznali, naučili i razumjeli na koji način funkcionira tehnologija uz pomoć projektnih zadataka iz robotike.

A. World Robotic Olympiad

World Robotic Olympiad je globalna inicijativa koja učenicima daje priliku i mogućnost da razvijaju vlastitu kreativnost i vještine rješavanja problema koristeći omiljene Lego kockice. Učenici u timu samostalno sastavljaju i grade odgovarajućeg robota, programiraju ga koristeći programsko sučelje Lego Mindstorms Education ili programski jezik MicroPython u programskom sučelju Visual Code Studio, kako bi robot u što kraćem vremenu uspješno ostvario unaprijed zadanu misiju: spašavanje ozlijeđenih, pametni prijevoz u urbanim sredinama, prepoznavanje i pravilno zbrinjavanje otpada i dr.



Slika 1. Lego robot koji slaže Rubikovu kocku

Na taj način učenicima se daje prilika da prošire svoje vidike istraživanjem robota i robotskih sustava, učenike se potiče da budu naši budući znanstvenici, inženjeri, tvorci i izumitelji, pomaže se razvoj i stjecanje vještina 21. stoljeća poput kreativnog razmišljanja, suradnje i komunikacije.



Slika 2. Programiranje Lego robota u Lego Mindstorms Education

B. Code Kids Croatia 2019.

Code Cids Croatia 2019. nacionalna je inicijativa za uvođenje robotike u obrazovanje koju provodi Hrvatski robotički savez kao sastavni dio EU Code Week-a. Kroz navedenu inicijativu učenici u osnovnim školama dobili su priliku samostalno izgraditi svog robota po imenu Marty. Marty-a je moguće programirati u programskom jeziku Scratch koji je namijenjen učenicima mlađeg uzrasta. U jednostavnom programskom sučelju učenici mogu osmisliti i izraditi različite programe za robota koji proizvodi zvuk, kreće se, vježba, pleše,…

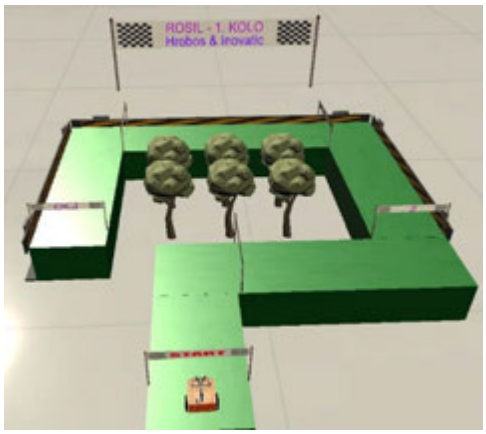


Slika 3. Robot Marty i programski blokovi u Scratchu

C. *Miranda*

Miranda je online simulacijsko sučelje koje omogućuje stvaranje i programiranje različitih vrsta robota pomoću programskih jezika Scratch i Python. Miranda učenicima omogućuje da stvore simulacije postojećih ili vlastitih imaginarnih robota, podijele svoje kreacije, rješavaju postojeće i stvaraju vlastite izazove. Simulacije nisu zamjene stvarnih STEM robota, već obogaćivanje aktivnosti s mogućnošću simulacije više robota istovremeno. Učenicima je omogućeno da sami vježbaju, suoče se s nizom neuspjeha i uspjeha u programiranju robota. Učenici samostalno biraju simulacije unaprijed definiranih robota i bespilotnih dronova koji se nude u Mirandi (mBot, Thymio, Ozobot, Lego, Dash, Codey Rocky, Edison) ili samostalno osmišljavaju i grade vlastite virtualne robote uz pomoć kojih rješavaju unaprijed zadane zadatke i izazove za razvoj vještina računalnog razmišljanja i programiranja.

Simulacija omogućava obogaćivanje aktivnosti na stvarnoj opremi, učenje i rad kada stvarni resursi nisu dostupni. Stvaranje vlastitih simulacija također je učenje stvaranja obrazovnih sadržaja za druge sa željom da ih unaprijede i inteligentno odvrate.



Slika 4. Programiranje robota Edison u okružju Miranda

Stvaranje vlastitih simuliranih robota STEM također znači eksperimentiranje s mehaničkim rješenjima i njihovo korištenje u simulacijama za analizu njihovog ponašanja.

III. ZAKLJUČAK

Predstavljeni primjeri samo su neki od mnogih uspješnih primjera primjene robotike u obrazovanju, mnogih uspješnih programa i projekata koji povezuju STEM, vještine programiranja, računalnog razmišljanja i inženjerske vještine. Robotika učinkovito uključuje učenike u učenje STEM koncepata, kodiranje, računsko razmišljanje i inženjerske vještine, sva potrebna znanja i vještine kako bi studenti u budućnosti postali uspješni članovi radne snage. Obrazovna robotika sveobuhvatno je tehnološko sredstvo za učenje koje promovira budući uspjeh naših učenika i treba ga sve više integrirati u školu Nastavni plan i program..

LITERATURA

[1] Robinson, K., Changing education paradigms, 2010, TED Talk.

[2] Robinson, K., Changing education paradigms, 2010, TED Talk.

[3] European Commision: The Digital Skills and Jobs Coalition, https://ec.europa.eu/digital-single-market/en/digital-skills-jobs-coalition, preuzeto: svibanj 2020.

[4] European Commision: The Digital Skills and Jobs Coalition, https://ec.europa.eu/digital-single-market/en/digital-skills-jobs-coalition, preuzeto: svibanj 2020.

[5] Amy Eguchi, "Robotics as a Learning Tool for Educational Transformation," Proceedings of 4th International Workshop Teaching Robotics, Teaching with Robotics & 5th International Conference Robotics in Education, Padova, Italy, July 2014.

Informacijska i komunikacijska tehnologija (ICT) – ključna karika između različitih oblika obrazovanja

Mario Božurić[1], Romana Bogut [2] i Mario Tretinjak[3]
[1,2,3] Elektrotehnička škola, Zagreb, Hrvatska
mario.bozuric@gmail.com, romanabogut@gmail.com, mario.tretinjak@gmail.com

***Sažetak* - Na 39. međunarodnom skupu MIPRO 2016 održanom u Opatiji, autori ovog rada objavili su članak pod nazivom „E-obrazovanjem do fleksibilnog modela učenja“, u kojem su prikazali konkretne učinke uvođenja e-obrazovanja u obrazovni sustav kako bi se održao i razvio fleksibilni model učenja i poučavanja u srednjim strukovnim školama. Nakon provedenog istraživanja nad različitim dobnim skupinama nastavnika, te nastavnicima struke i nastavnicima općeobrazovnih predmeta autori su došli do zaključka da će škole s iskustvom u implementaciji ICT tehnologije u provedbi, praćenju i vrednovanju obrazovnih postignuća puno lakše implementirati nove oblike nastave u budućnosti. Suočeni s izazovima provedbe nastave na daljinu kao posljedicu pandemije COVID-19 autori u ovom radu prikazuju korelaciju između različitih oblika izvođenja klasične nastave u učionici i prelaska na potpuni model nastave na daljinu uz pomoć ICT tehnologije. Autori su uvjereni da iskustva opisana u ovom radu mogu pomoći prilikom planiranja i provedbe različitih metoda i oblika izvođenja nastave, kako bi se nastavnicima i učenicima uz korištenje ICT tehnologije omogućio što lakši i fleksibilniji prelazak iz jednog obrazovnog modela u drugi.**

Ključne riječi - nastava; nastava na daljinu; online; e-obrazovanje; m-obrazovanje; ICT; tehnologija; obrazovanje; informacijsko-komunikacijska tehnologija

I. UVOD

Učionica u prostorima škole donedavno je bila središnje mjesto izvođenja školskog kurikuluma u osnovnim i srednjim školama. Malo tko je mogao pretpostaviti da će se cjelokupna redovna nastava u samo par dana premjestiti iz klasične nastave u online nastavu. Prostor škole je u pravilu bilo mjesto koje je osiguravalo sve potrebne uvjete i opremu za izvođenje redovne nastave, kako za nastavnike tako i za njihove učenike. Obrazovanje kao ključni faktor i motor modernog društva mijenjalo se s razvojem tehnologija kako bi moglo pozitivno utjecati i odgovoriti na sve zahtjeve i izazove koje su donosile društvene i gospodarske promjene. Kako bi se osigurala fleksibilnost obrazovnog procesa, te njegova horizontalna i vertikalna prohodnost mijenjali su se uvjeti i oblici izvođenja nastave, ali nikad tako naglo kao tog 19. ožujka 2020. godine kada je Vlada Republike Hrvatske donijela Odluku o izmjeni Odluke o obustavi izvođenja nastave u visokim učilištima, srednjim i osnovnim školama te redovnog rada ustanova predškolskog odgoja i obrazovanja i uspostavi nastave na daljinu[1]. Ravnatelji škola su umjesto prostora škole za privremeno mjesto rada trebali odrediti neko drugo mjesto prema okolnostima slučaja, ali uz uvjet da se nastava na daljinu kontinuirano i nesmetano održava.

Za nastavnike i njihove učenike dogodila se ključna promjena u obliku i načinu izvođenja nastavnih sadržaja jer je školska učionica prestala biti mjesto odvijanja nastavnog procesa. Za vrijeme izvođenja klasične nastave (nastava u učionicama) ovisno o školskom kurikulumu i osobnom pristupu nastavnici su digitalnu tehnologiju koristili za poboljšanje klasične nastave ili kao pomoć u kombinaciji klasične nastave u učionici i raznih oblika e-učenja i m-učenja. Nastava koja se odvijala u školskim prostorima zamijenjena je nastavom na daljinu, odnosno potpuno novim oblikom odvijanja nastavnog procesa gdje učenici uče samostalno prema uputama svojih nastavnika, te uz pomoć nastavnih sadržaja dostupnih u online okruženju. Prilikom odvijanja i izvođenja nastave na daljinu računalo i internet, odnosno informacijsko - komunikacijska tehnologija (ICT) postaju ključna veza između nastavnika i njihovih učenika.

Članak prikazuje konkretna iskustva srednjoškolskih nastavnika i učenika prilikom prelaska iz nastavnog procesa koji se odvijao u prostorima škole ili sa školom povezanim prostorima u potpuno novi način odvijanja nastave koji se uz pomoć ICT tehnologije odvija u virtualnom okruženju kao nastava na daljinu. U članku je prikazano kako su nastavnici u srednjim školama prihvatili i implementirali ICT tehnologiju koja je postala ključna karika u odvijanju nastave na daljinu, koliki je utjecaj digitalnih tehnologija i alata na tijek provedbe i izvođenja nastavnog procesa, kao i korelacija između uporabe različitih oblika ICT tehnologija i kvalitete izvođenja nastave na daljinu, prilikom čega je dat poseban osvrt na stavove i iskustva učenika tijekom usvajanja i slušanja nastavnih sadržaja iz različitih predmeta uz pomoć računala i raznih digitalnih alata, odnosno ICT tehnologije. Osim analize odvijanja i provođenja nastave na daljinu navode se stavovi i prijedlozi nastavnika i učenika kao glavnih čimbenika provedbe online nastave koji se odnose na planiranje školskog kurikuluma i uporabe ICT tehnologije, a temelje se na njihovom dosadašnjem iskustvu slušanja, odnosno provođenja nastave na daljinu.

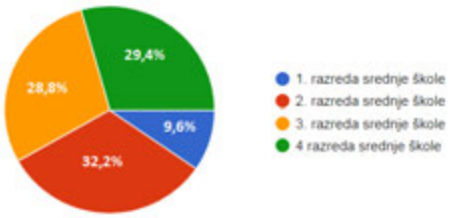
II. OPIS PROVEDENIH AKTIVNOSTI

Nakon četrdeset dana izvođenja nastave na daljinu zbog proglašene pandemije uzrokovane pojavom COVID-19 autori ovog rada proveli su anketu o izazovima provedbe i izvođenja nastave na daljinu u kojoj je sudjelovalo 70

srednjoškolskih nastavnika iz više obrazovnih grupa predmeta (slika 1.) i 177 srednjoškolskih učenika od prvog do četvrtog razreda (slika 2.). Zbog preporučenih i propisanih mjera Stožera civilne zaštite Republike Hrvatske za sprečavanje širenja zaraze koronavirusom anketa je provedena u online okruženju, a za kreiranje i dijeljenje ankete korištena je aplikacija Google Forms. Nastavnici i učenici koristeći dostupne ICT tehnologije mogli su pristupiti poveznici za popunjavanje iste. Anketna pitanja, kao i ponuđene odgovore zatvorenog i otvorenog tipa osmislili su autori članka. Anketa namijenjena nastavnicima sadrži 21 pitanje od čega je 19 pitanja s ponuđenim odgovorima zatvorenog tipa i dva pitanja s odgovorima otvorenog tipa u kojima su nastavnici mogli dati duži tekstualni odgovor. Anketa za učenike sadrži 22 pitanja od kojih je 19 pitanja s ponuđenim odgovorima zatvorenog tipa i tri pitanja s odgovorima otvorenog tipa u kojima su učenici mogli iznositi osobna mišljenja i stavove vezane za tijek i provedbu nastave na daljinu. Za učenike i nastavnike anketa je bila potpuno anonimna.



Slika 1. Anketa nastavnici – prema grupi predmeta u kojoj predaju



Slika 2. Anketa učenici – prema razredu koji pohađaju

III. EVALUACIJA ISTRAŽIVANJA

A. Evaluacija - nastavnici

Od 19 anketnih pitanja zatvorenog tipa u evaluaciji su prikazani odgovori na 10 pitanja iz kojih se može uočiti izazov prihvaćanja i implementacije ICT tehnologije kao ključne karike između svih dionika obrazovanja, kao i utjecaj digitalnih alata na tijek izvođenja nastave na daljinu. Evaluacija obuhvaća i dva pitanja otvorenog tipa na koja su nastavnici dali ukupno 95 odgovora. Odgovori na pitanja otvorenog tipa ukazuju na potrebnu pomoć nastavnicima tijekom izvođenja nastave na daljinu i njihove prijedloge za poboljšanje nastave na daljinu i što lakši prijelaz iz klasične nastave u online nastavu i obrnuto.

Tijekom izvođenja klasične nastave (nastava u učionici) svi nastavnici koji su ispunili anketu koristili su neki oblik ICT tehnologije prilikom pripreme i izvođenja nastave. Najveći broj nastavnika, njih 50,9% ICT tehnologiju koristili su za poboljšanje klasične nastave, 29,5% za pripremu nastave, 5,4% nastavnika provodilo je kombinaciju klasične nastave i nastave na daljinu (e-učenje, m-učenje), dok je 5,4% ispitanika imalo iskustva s provedbom „čistog“ obrazovanja na daljinu (Slika 3.). Više od trećine nastavnika, njih 36,7% uglavnom se ne slaže ili se uopće ne slaže s izjavom da su bili pripremljeni za izvođenje nastave na daljinu (Slika 4.). Da su koristili ICT tehnologiju potrebnu za održavanje nastave na daljinu prije pojave pandemije, 39,6% nastavnika uglavnom se ne slaže ili se uopće ne slaže (Slika 5.) Čak 92,9% nastavnika kaže da im za pripremanje online nastave treba više vremena, a preostalih 7,1% kaže da im treba jednako vremena iz čega proizlazi da je priprema i održavanje nastave u ovakvim uvjetima za nastavnike izrazito složena, te zahtjeva puno više rada i vremena u odnosu na klasičnu nastavu (Slika 6.). Rezultati pokazuju da se veliki dio nastavnika oslanja na gotove materijale za učenje (Slika 7.) te na komunikaciju elektroničkom poštom ili putem obavijesti na portalima za učenje (Slika 8.). Digitalne alate (Zoom, Microsoft Teams i sl.) za izvođenje nastave uživo redovito koristi 29,6% nastavnika, rijetko njih 16,9%, dok 39,4% nastavnika uopće ne izvodi nastavu uživo (Slika 9.). Da je komunikacija s učenicima vrlo uspješna u potpunosti se slaže ili se uglavnom slaže 69,1% nastavnika (Slika 10.), a pritom većina učenika sve zadatke obavlja na vrijeme. Veliki broj nastavnika, njih 77,5% se u potpunosti slaže ili se uglavnom slaže da bi im održavanje jednog dijela nastave u online okruženju za vrijeme trajanja klasične nastave (nastava u učionici) pomoglo prilikom prelaska na "čistu" online nastavu (Slika 11.), dok 40,8% nastavnika smatra da bi u budućnosti jedan dio nastavnih sadržaja, odnosno nekoliko sati nastave trebalo planirati tako da se izvodi u online okruženju kao "čista" nastava na daljinu (Slika 12.).

Odgovori na dva pitanjima otvorenog tipa ukazuju na potrebnu edukaciju i pomoć nastavnicima prilikom korištenja ICT tehnologije, digitalnih alata i obrazovnih platformi za provođenje i izvođenje nastave na daljinu.

Odabrana (značajna) anketna pitanja na koja su odgovarali nastavnici prilikom ispunjavanja ankete:

- Za vrijeme održavanja klasične nastave (nastava u učionici) ICT tehnologije koristio sam za (mogućnost više odgovora):

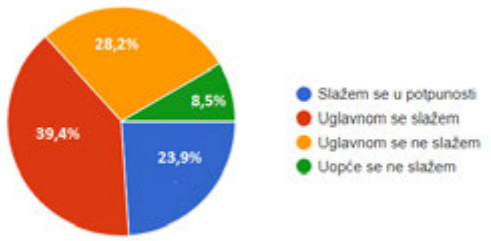
Odgovori:



Slika 3. Anketa nastavnici - korištenje ICT tijekom klasične nastave

- Bio (-la) sam dovoljno pripremljen (-na) za provedbu i održavanje nastave na daljinu (uzrokovanu pojavom pandemije COVID-19) jer sam prije održavanja nastave na daljinu bio (-la) educiran (-a) i upoznat (-a) s većinom ICT tehnologije potrebne za održavanje online nastave.

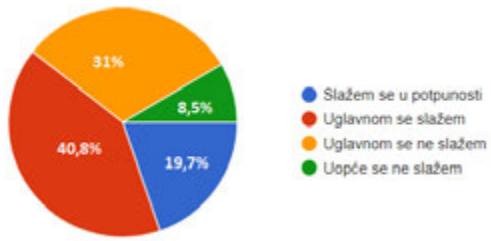
Odgovori:



Slika 4. Nastavnici su bili dovoljno pripremljeni za održavanje nastave na daljinu

- Većinu ICT tehnologije potrebne za održavanje nastave na daljinu koristio (-la) sam prije pojave pandemije uzrokovane COVID-19.

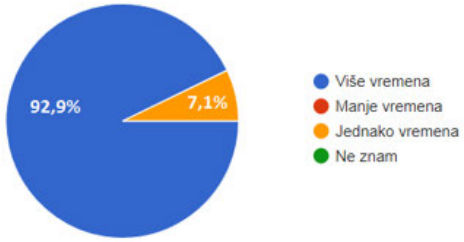
Odgovori:



Slika 5. Nastavnici su većinu ICT koristili i prije pandemije

- Pripremanje nastave na daljinu u odnosu na klasičnu nastavu zahtjeva:

Odgovori:



Slika 6. Nastavnici odgovaraju koliko im vremena treba za pripremu nastave na daljinu

- Za pripremu i održavanje nastave na daljinu koristim sljedeće nastavne materijale (na ovo pitanje nastavnici su imali mogućnost više odgovora):

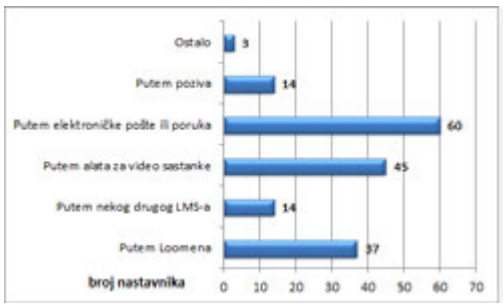
Odgovori:



Slika 7. Materijali koje nastavnici koriste u nastavi na daljinu

- Za provedbu nastave na daljinu komunikaciju s učenicima ostvarujem (na ovo pitanje nastavnici su imali mogućnost više odgovora):

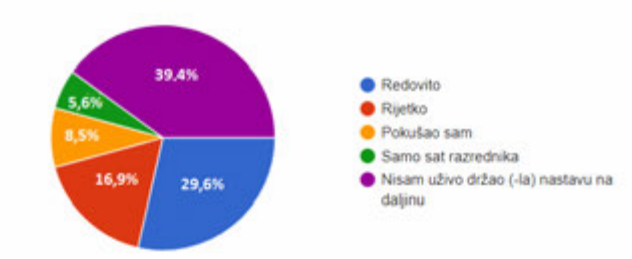
Odgovori:



Slika 8. Kako nastavnici ostvaruju komunikaciju s učenicima prilikom provedbe nastave na daljinu

- Korištenjem digitalnih alata (Zoom, Microsoft Teams i sl.) s učenicima održavam nastavu uživo:

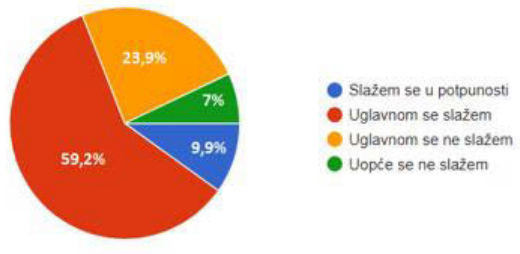
Odgovori:



Slika 9. Korištenjem digitalih alata nastavnici održavaju nastavu uživo

- Za vrijeme održavanja nastave na daljinu komunikacija s većinom učenika je vrlo uspješna.

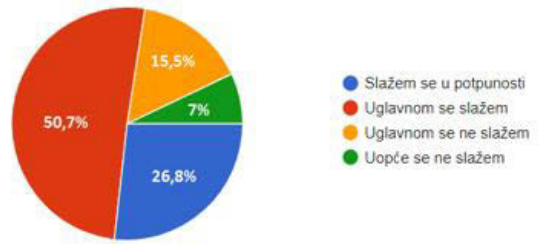
Odgovori:



Slika 10. Komunikacija s većinom učenika je vrlo uspješna

- Za vrijeme trajanja klasične nastave (nastava u učionici) održavanje jednog dijela nastave u online okruženju pomoglo bi mi prilikom prelaska na "čistu" online nastavu.

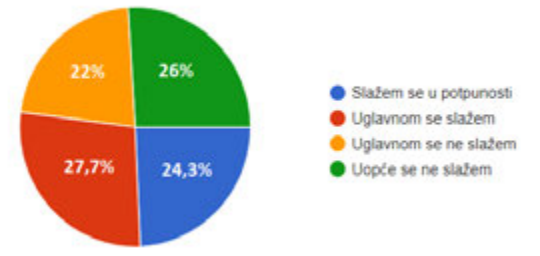
Odgovori:



Slika 11. Održavanje jednog dijela klasične nastave u online okruženju pomoglo bi prilikom prelaska na "čistu" online nastavu

- U budućnosti bi jedan dio nastavnih sadržaja, odnosno nekoliko sati nastave trebalo planirati tako da se izvode u online okruženju kao "čista" nastava na daljinu.

Odgovori:



Slika 12. U budućnosti bi jedan dio nastave trebalo izvoditi kao nastava na daljinu

- Što bi vam pomoglo prilikom provedbe i izvođenja nastave na daljinu (pitanje otvorenog tipa)?

Značajni odgovori (od ukupno 50 odgovora):

- „Bolja pripremljenost - da sam otprije poznavala aplikacije koje koristim"
- „Edukacija"
- „Radionica (obuka) – Zoom i sl."
- „Sustav za online učenje manje opterećen od Loomena"
- „Jedna platforma s više mogućnosti i većom brzinom (u odnosu na Loomen), više unificiranog sadržaja, a ne da svatko radi po svom, što dodatno zbunjuje učenike."
- „Više predznanja i iskustva u korištenju različitih ICT alata"
- „Podrška nakladnika izradom online materijala."
- „Dodatni mjeseci usavršavanja i pripreme te zakonski okvir (realni)!"

- Napišite sve vaše prijedloge vezane uz lakšu prilagodbu nastavnika i učenika prilikom prelaska iz jednog obrazovnog modela u drugi, odnosno iz klasične nastave u online nastavu i obrnuto (pitanje otvorenog tipa).

Značajni odgovori (od ukupno 45 odgovora):

- „Bolja priprema - u ovom slučaju prijelaz je bio iznenadan i bilo je potrebno vrijeme da da nastavnici pronađu optimalan način rada te da se i nastavnici i učenici prilagode situaciji."
- „Kombinacija "klasične" i on line nastave bila bi dobitna kombinacija!"
- „Kvalitetnija tehnička podrška. Detaljnije i jasnije upute oko vođenja izostanaka učenika i ocjenjivanja."
- „Izrada kvalitetnih materijala pojedinih programskih sadržaja prilagođenih za nastavu na daljinu."
- „Kurikulum bi trebalo podijeliti na klasičnu nastavu i e-nastavu (cca. 50 – 50%)."

B. Evaluacija učenici

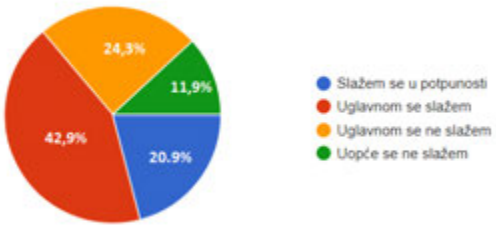
Od 19 anketnih pitanja zatvorenog tipa na koja su odgovarali učenici u evaluaciji su prikazani odgovori na 5 pitanja. Iz odgovora na pitanja može se uočiti izazov uporabe i korištenja ICT tehnologije i digitalnih nastavnih sadržaja tijekom odvijanja i praćenja nastave na daljinu. Evaluacija obuhvaća i tri pitanja otvorenog tipa na koja su učenici dali više od 300 odgovora. Odgovori na pitanja otvorenog tipa ukazuju na poteškoće koje učenici imaju, te na potrebnu pomoć i podršku prilikom praćenja nastave na daljinu. Učenici su u odgovorima davali svoje prijedloge vezane uz poboljšanje odvijanja nastave na daljinu i što lakši prijelaz iz klasične nastave u online nastavu i obrnuto.

Uglavnom se slaže 42,9% ili se u potpunosti slaže 20,9% učenika da su i prije izvođenja nastave na daljinu koja je posljedica pandemije koristili neku vrstu ICT tehnologije prilikom samostalnog učenja (Slika 13.). Ranije iskustvo korištenja digitalnih alata omogućilo je učenicima lakši prijelaz na online nastavu. Svaki peti učenik nije se uspio prilagoditi na udaljenu nastavu, a otprilike toliki postotak učenika ne izvršava redovno svoje školske obaveze. Premda 96,6% učenika u online nastavi koristi računalo, postoji 3,6% učenika koji koriste isključivo mobitel ili tablet. Za uređaje koje koriste u online nastavi 68,4% učenika kažu da su na raspolaganju samo njima, 21,5% učenika dijeli uređaje i opremu s još jednom osobom, 4,5% učenika s još dvije osobe, a preostalih 5,6% dijeli uređaje s tri ili više osoba (Slika 14.). Većina učenika kaže da im je za praćenje nastave na daljinu i usvajanje nastavnih sadržaja u odnosu na klasičnu nastavu treba više vremena. Najveći broj učenika za preuzimanje nastavnih sadržaja i zadatka, te slanje zadaća koristi Loomen (oko 81,4%). Zbog čestog opterećenja na Loomenu učenici i nastavnici u paraleli koriste i druge portale za međusobnu komunikaciju, dok jedan dio učenika i nastavnika uopće ne koristi Loomen, za koji kažu da je često opterećen i spor. Gotovo 60% učenika komunicira s nastavnicima putem elektroničke pošte ili putem aplikacija na mobilnim telefonima kako bi ostvarili bržu interakciju s predmetnim nastavnicima (Slika 15.). Više od polovine učenika smatra da je komunikacija s većinom nastavnika vrlo uspješna. Prema navodima učenika samo iz pojedinih predmeta 52,5% nastavnika izvodi nastavu uživo, dok u 29,9% predmeta nastavnici uopće ne koriste digitalne alate (Zoom, Microsoft Teams i sl.) kako bi izvodili nastavu uživo (Slika 16.). Učenici navode da su im snimljeni materijali u obliku video prezentacija ili snimke predavanja njihova nastavnika dostupni samo kod polovine predmeta, dok ostale sadržaje čine materijali dostupni na internetu. Vezano uz dosadašnja iskustva praćenja nastave 52% učenika uglavnom se slaže ili se u potpunosti slaže da bi u budućnosti dio nastavnih sati trebalo održavati kao nastavu na daljinu jer bi im to pomoglo prilikom prilagodbe na neku buduću online nastavu (Slika 17.).

Odabrana (značajna) anketna pitanja na koja su odgovarali učenici prilikom ispunjavanja ankete:

- Uz pomoć ICT tehnologije dio nastavnog gradiva učio (-la) sam samostalno i online (e-učenje, m-učenje) i prije prelaska na udaljenu nastavu uzrokovanu pandemijom COVID-19.

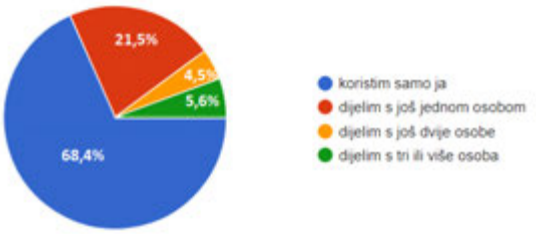
Odgovori:



Slika 13. Učenici su uz pomoć ICT tehnologije dio nastavnog gradiva učili samostalno i online (e-učenje, m-učenje) i prije prelaska na udaljenu nastavu uzrokovanu pandemijom COVID-19.

- Uređaje koje koristim u online nastavi:

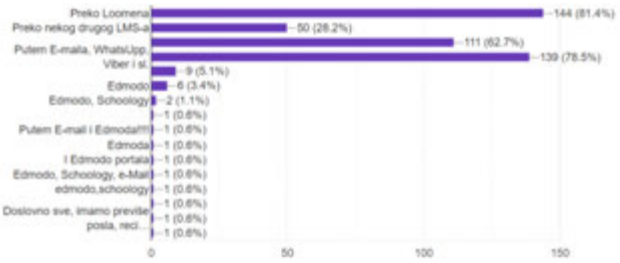
Odgovori:



Slika 14. Računalo i ostalu tehničku opremu učenik dijeli

- Za praćenje nastave na daljinu komunikaciju s nastavnicima ostvarujem (uz mogućnost više odgovora):

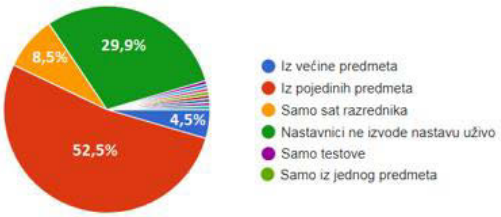
Odgovori:



Slika 15. Alati za komunikaciju s učenicima u provedbi nastave na daljinu

- Korištenjem digitalnih alata (Zoom, Microsoft Teams i sl.) nastavu pratim uživo:

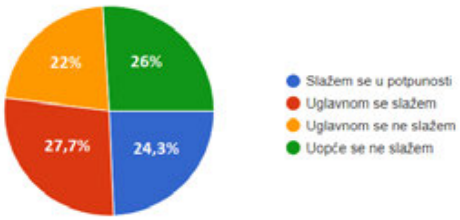
Odgovori:



Slika 16. Korištenjem digitalnih alata učenici prate nastavu uživo

- U budućnosti bi jedan dio nastave trebalo planirati tako da se izvode u online okruženju kao "čista" nastava na daljinu.

- Odgovori:



Slika 17. U budućnosti bi jedan dio nastave trebalo izvoditi kao nastava na daljinu

- Što bi ti pomoglo prilikom provedbe i praćenja nastave na daljinu?

Značajni odgovori (od ukupno 115 odgovora):

- „Da svi profesori koriste jednu platformu za online nastavu, korištenje Zoom-a za nastavu uživo."
- „Organiziranost profesora, pridržavanje termina postavljanja materijala."

- Napiši svoje prijedloge vezane uz lakšu prilagodbu učenika prilikom prelaska iz klasične nastave u online nastavu i obrnuto.

Značajni odgovori (od ukupno 100 odgovora):

- „Imati video predavanja s profesorima svaki dan kada imamo tog profesora/icu po rasporedu, manje pojedinačnog rada."
- „Organiziranost nastave, pridržavanje rasporeda i termina postavljanja materijala."
- Određena gradiva bi trebali usvajati online, iako smo u klasičnoj nastavi.

- Napiši što bi poručio (-la) nastavnicima vezano uz nastavu na daljinu:

Značajni odgovori (od ukupno 116 odgovora):

- „Nastavnici bi se trebali bolje organizirati, neki od nastavnika šalju predmetni sadržaj kada se sjete, ovo je ONLINE NASTAVA, vjerujem da vrijedi isti raspored kao i u klasičnoj nastavi…"
- „I mi smo ljudi, dajte nam malo odmora i razmišljajte da nemamo samo vaš predmet već još 14 ostalih."

IV. RASPRAVA

Iz rezultata provedene ankete i dobivenih odgovora može se uočiti da je većina nastavnika koristila neku vrstu digitalnih alata, odnosno ICT tehnologija prilikom pripreme i održavanja klasične nastave. Nastavnici ističu da su im ta znanja i vještine bili od pomoći prilikom održavanja nastave na daljinu premda ne u tolikoj mjeri da bi nastavu na daljinu održavali jednako kvalitetno kao i klasičnu nastavu. Kako bi ostvarili željenu kvalitetu nastave potrebno im je puno više vremena i truda zbog različitih vanjskih faktora i okolnosti koje utječu na nesmetano odvijanje obrazovnog procesa. Kao otežavajuće faktore navode nestabilne platforme za učenje na daljinu, nedovoljno poznavanje ICT tehnologije potrebne za nesmetano odvijanje online nastave, poteškoće s brzinom interneta, te mali broj gotovih i standardiziranih edukativnih sadržaja kao što su video prezentacije za nastavne jedinice u pojedinim strukovnim predmetima. Nastavnici ističu da čak i gotove nastavne materijale koje su ranije koristili za izvođenje nastave moraju prilagođavati novim uvjetima izvođenja nastave. Iz odgovora na ponuđena anketna pitanja vidljivo je da određeni broj nastavnika ne koristi ili u dovoljnoj mjeri ne poznaje ICT tehnologije potrebne za održavanje online nastave, a posebno je to vidljivo kod izvođenja nastave uživo. Nastavnici koriste različite alate i sredstva za komunikaciju kako bi ostvarili željeni kontakt, odnosno interakciju s učenicima. Upravo postojanje različitih i nestandardiziranih alata za komunikaciju učenici navode kao veliki problem jer im na različitim digitalnim sredstvima za komunikaciju svakodnevno dolaze obavijesti o materijalima, zadacima i zadaćama od više nastavnika. Ovaj problem je još veći kod učenika koji dijele opremu i računalo s nekim od članova svoje obitelji. Učenici gotovo u pravilu ističu kako bi im rad u online okruženju bio znatno lakši kad bi svi nastavnici koristili iste digitalne alate i platforme za komunikaciju, prijenos i obradu nastavnih sadržaja. Navikli su na predavanja i način rada svojih nastavnika pa stoga predlažu veći broj snimljenih predavanja ili izvođenja nastave uživo koja bi se održavala u definiranim terminima jer bi tako mogli lakše učiti i organizirali vrijeme za rad, što je posebno važno učenicima koji dijele računalo i opremu s članovima obitelji. Više od polovine nastavnika i učenika slaže se da bi u budućnosti jedan dio nastavnih sadržaja, odnosno nekoliko sati nastave trebalo planirati tako da se izvodi u online okruženju kao "čista" nastava na daljinu jer bi im to znatno olakšalo prijelaz iz klasične nastave na online nastavu u izvanrednim i neplaniranim okolnostima. Nastavnici i učenici uglavnom podržavaju održavanje online nastave ali uz jasna pravila izvođenja i standardizirane obrazovne platforme i alate.

V. ZAKLJUČAK

Iako je vidljivo iz evaluacije provedene ankete da su gotovo svi učenici i nastavnici koristili neki oblik e-obrazovanja za vrijeme odvijanja klasične nastave, i kod jednih i drugih može se uočiti značajna razlika prilikom korištenja ICT tehnologije za vrijeme održavana nastave na daljinu. Različiti oblici obrazovanja, od klasične nastave do online nastave, koji se provode uz pomoć računalne tehnologije i interneta mogu se definirati kao e-obrazovanje. S obzirom da svaka razina e-učenja zahtijeva različitu vrstu ICT tehnologije, poznavanje i korištenje ICT tehnologije na jednoj razini nije garancija za uspješnu provedbu e-učenja na višim razinama, odnosno u „čistoj" online nastavi gdje učenici uče potpuno samostalno, a tehnologija je jedina veza u komunikaciji između nastavnika i učenika. Iako je provedba nastave na daljinu nezamisliva bez uporabe tehnologije, ne smije se zanemariti važnost kvalitetnog učenja i poučavanja, kao i važnost socijalne i odgojne komponente. Upotreba ICT tehnologije kao alata za održavanje online nastave može ukloniti fizičku barijeru između učenika i nastavnika, ali uz uvjet da učenik i nastavnik jednako dobro poznaju digitalne alate za izvođenje i praćenje nastave uživo. Kvalitetno izvođenje i praćenje nastave uživo uz poznavanje digitalnih alata zahtjeva određene vještine i iskustvo korištenja istih prilikom izvođenja i praćenja nastave na daljinu. Planiranje jednog dijela redovne nastave u online okruženju koja bi se uz pomoć ICT tehnologije u budućnosti izvodila uživo, može znatno olakšati učenicima i nastavnicima prelazak iz klasičnog oblika nastave u online nastavu i obrnuto.

VI. LITERATURA

[1] Vlada Republike Hrvatske, MZO: Odluka o izmjeni Odluke o obustavi izvođenja nastave u visokim učilištima, srednjim i osnovnim školama te redovnog rada ustanova predškolskog odgoja i obrazovanja i uspostavi nastave na daljinu, NN 32/220, ožujak 2020.

Laboratorijski 3-osni model realiziran pomoću servo sustava

Mario Lučan* Ivica Vlašić* i Goran Malčić*
* Tehničko veleučilište u Zagrebu, Zagreb, Hrvatska
mlucan@tvz.hr ivica.vlasic@tvz.hr goran.malcic@tvz.hr

Sažetak **U okviru ovog rada opisana je implementacija servo sustava kao dijela sustava industrijske automatizacije na laboratorijskoj maketi sa tri osi. Na osnovu mehaničkih zahtjeva konstrukcije makete izvedeno je tehničko rješenje koje obuhvaća odabir sklopne opreme, kontrolera, senzora i praktičnu implementaciju servo pogona na primjeru troosnog manipulatora. Odabrani primjer identičan je stroju za manipulaciju paketa gdje se isti slažu na paletu za transport. Prema danom tehničkom rješenju izrađena je mehanička konstrukcija i električna shema po kojoj je napravljena upravljačka ploča sa sklopnom opremom i kontrolerom. Prema zahtjevima tehničkog rješenja izrađena je upravljačko-nadzorna programska podrška. Opisana je logika upravljanja i posebnosti koje nam pruža odabrana opreme prilikom izrade programa.**

I. ZAHTJEVI RADA

U sklopu rada potrebno je bilo osmisliti laboratorijsku maketu sa tri servo sustava koja će imati smislenu funkcionalnost sličnu nekoj iz industrije. Izraditi električnu shemu i po njoj napraviti upravljačku ploću sa sklopnom opremom i kontrolerom te prema zahtjevima tehničkog rješenja izraditi upravljačko-nadzornu podršku.

II. TEHNIČKO RJEŠENJE

A. *Mehanička konstrukcija*

Prvi problem je bio izraditi mehaničku konstrukciju na koju će biti dovoljno čvrsta i stabilna kako bi se mogli montirani na nju motri i drugi elementi automatizacije potrebni za realizaciju funkcionalne makete. Kako su profesionalni elementi za takve sustave i projekte pre skupi bilo je potrebno osmisliti jeftinije rješenje. Za baznu konstrukciju je odabran stari 3D printer dostupan na veleučilištu. Veoma je stabilne i čvrste konstrukcije te s time idealan za ovaj rad.

Uz svoje odlične statičke karakteristike 3D printer je konstruiran sa tri osi tako da su već postojali mehanički elementi pomoću kojih je izvedeno kretanje uz neke manje izmjene. No položaj osi nije bio dobar pa je bilo potrebno rastaviti cijelu konstrukciju, nadograditi i prenamijeniti pojedine osi tako da maketa može obavljati željene pokrete.

Maketa je preuređena tako da je hvataljka s kojom bi se objekti transportirali ima omogućena horizontalna i vertikalna a postolje gdje se slažu objekti ima mogućnost gibanja naprijed i nazad. Takvim preuređenjem pokrila su se sva tri gibanja koja su bila potrebna kako bi se ostvarila mogućnost slaganja objekata u tri dimenzije.

Na svaku os je montiran jedan servo motor koji služi za pokretanje te osi. Prijenos pokreta motora na pomične dijelovi osi je izveden preko zupčastog remena. Na jednom kraju osi se nalazi servo motor sa pričvršćenim zupčanikom a na drugom kraju ležaj, između njih se polože i steže zupčasti remen. Pomični dijelovi osi su na jednom dijelu pričvršćeni na napeti remen čime se ostvaruje prijenos gibanja motora na os.



Slika 1 Vertikalna os

U većini postrojenja strojevi za pakiranje uzimaju objekte sa neke vrste transportnog pogona koji transportira objekte do njih. U ovom primjeru zbog ograničenih resursa i vremena nije napravljen sličan primjer tome nego je izrađena konstrukcija u kojoj se mogu pohraniti objekti te su na raspolaganju za transport.



Slika 2 Konstrukcija za skladišenje objekata

B. *Servo sutavi*

Servo pretvarači korišteni u ovom radu su SINAMICS S210 model 6SL3210-5HB10-4UF0 sa motorima modela 1FK2103-4AG00-1MA0 od proizvođača SIEMENS. Pretvarač se napaja iz jednofaznog izvora 230V , 5A te pretvara i napaja motor sa trofaznim naponom 110V. Pošto su napravljeni da rade jedan sa drugim dizajnirani su tako da se mogu spojiti sa jednim kabelom u kojemu su vodovi za napajanje, enkoder i za kočnicu implementirani zajedno u tom jednom kabelu.

Odabrani servo sustav nam pruža jednostavno upravljnje pozicijom motora kako nije potrebna izrada programa koja će pratit njegovu poziciju preko enkodera već je taj dio obrađen unutar samog pretvarača. Uz jednostavno upravljanje pruža nam niz drugih parametara koji se mogu pratiti i implementirati u izradi upravljačkog programa.

C. *Programibilni logički kontroler – PLC*

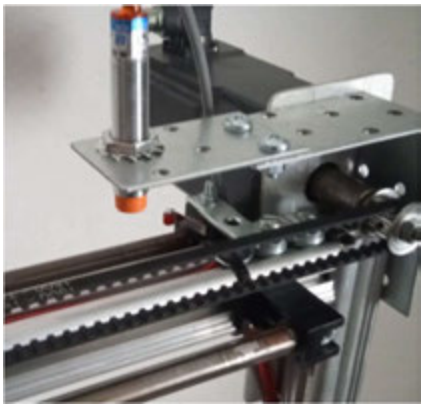
U radu se koristi kontroler *SIMATIC S7 1500* model *CPU 1516-3 PN/DP* sa modularnom jedinicom za proširenje digitalnih ulaza *DI-32x24VDC-HF* sa modularnom jedinicom za proširenje digitalnih izlaza *DQ-32x24VDC/0.5A* sa modularnom jedinicom za proširenje analognih izlaza *AI 8xU/I/RTD/TC ST* sa modularnom jedinicom za proširenje analognih ulaza *AQ 4xU/I ST*. Ovaj tip PLC-a je pogodan za ovaj rad jer podržava posebnu vrstu blokova iz kategorije *Motion Control* koja pruža veoma korisne mogućnosti upravljanja sa servo sustavima i olakšava izradu upravljačkog programa.

D. *Senzori*

Radi osiguravanja sigurnog rada i implementiranja svih mogućnosti navedenog servo sustava na maketu su ugrađena dva tipa senzora mikro prekidači i bez kontaktni induktivni senzori *LJ12A3-4-Z/BY* . Mikro prekidači su implementirani kao granični senzori te služe za zaustavljanje kretanja ako su okinuti. Induktivni senzori se koriste kao davači početne pozicije i koriste se kod inicijalizacije pozicija osi. Osim za inicijalizaciju ugrađeni su induktivni senzori na hvataljku i na postolje gdje se pohranjuju objekti. Svrha senzora na hvataljci je signaliziranje nestanka objekta prilikom transporta a senzori na postolju daju podatak o količini i prisutnosti dostupnih objekta. Svi senzori su spojeni na digitalne ulaze PLC-a te se koriste u programu za navedene funkcije.

E. *Signalizacija i upravljanje*

U postrojenjima sustavi mogu biti u potpunosti automatizirani ili polu automatizirani sa potrebnom prisutnošću operatera koji će pokrenuti sustav i nadopuniti potrebne materijale za nastavak procesa. Kako nije bilo moguće izvesti maketu koja će od početka do kraja sama odraditi posao implementirana su tipkala za upravljanje i lampice za signalizaciju kako bi se olakšalo upravljanje i nadzor makete. Ugrađena su dva tipkala START i STOP pomoću kojih se pokreće posao i pauzira odnosno zakoči sustav. Uz ta dva tipkala ugrađena je i preklopka u slučaju kvara s kojom se prekida rad motora i ne dopušta ponovno kretanje u rad dok se ne deaktivira preklopka. Za



Slika 3 Granični senzor i senzor početnog položaja

signalizaciju stanja makete. Za signalizaciju stanja makete montirana su i tri indikatora zelene žute i crvene boje. Zelenim indikatorom se signalizira rad i čekanje u slučaju potrebe za nadopunom objekta za transport. Kontinuiranim svijetljenjem se signalizira rad a čekanje sa treptanjem frekvencije 2Hz. Žutim indikatorom se signalizira spremnost za početak rada i stanje pauze kada se pritisne tipkalo STOP. Kao kod zelenog indikatora stanje spremnosti se signalizira kontinuiranim svijetljenjem a stanje pauze treptanjem od 2Hz. Crveni indikator signalizira kvar u sustavu.

F. *Korisničko sučelje*

Za nadzor sustava se koristi se HMI (engl. *Human Machine Interface*) uređaj *SIMATIC TP700 Comfort.* Na njegovom zaslonu se prikazuje stanje sustava i u slučaju kvara mogu odabrati neke posebne opcije. Izgled ekrana će biti opisanu kasnijem dijelu rada.



Slika 4 Laboratorijski 3-osni model raliziran pomoću servo sustava

G. *Upravljačka ploća*

Prema tehničkom rješenju napravljena j upravljačka ploča sa sklopnom opremom, kontrolerom i servo pretvaračima. Kako je za napajanje jednog pretvarača potrebna samo jedna faza odabrano je trofazno napajanje kako bi se ravnomjerno rasporedilo opterećenje na tri faze. Takvim rasporedom opterećenja mogla se odabrati sklopna oprema nižih kategorija te kombinirana u jedan

element pošto se štite tri ista pretvarača. Mana takvog pristupa je da ukoliko jedan pretvarač okine zaštitu ostali će pretvarači biti također isključeni. Za zaštitu od preopterećenja ugrađeni su rastalni osigurači 10A u kombinaciji za motornom zaštitnom skloplom 4,5-6A.

Za napajanje kontrolera i ostalih djelova koristi se napajanje 24VDC 6EP1333-4BA0 nazivne snage 190W. Spojeno je na jednu od faza i štićeno automatskim prekidačem C6A-2 polni.



Slika 5 Upravljačka ploća makete

III. KONFIGURACIJA KOMPONENATA

A. *TIA- portal*

Za postavljanje komunikacije, konfiguriranje uređaja i izradu programa koristi se programsko okruženje TIA portal V15.1(engl. *Totally Integrated Automation Portal*). U tom okruženju mogu se instalirati programski alati potrebni za izradu aplikacije. Alati instalirani potrebi za ovaj rad su bili SIMATIC Startdrive V15.1 za konfiguraciju servo pretvarača, SIMATIC Step7 V15.1 za izradu upravljačkog programa i SIMATIC WinCC V15.1 za izradu korisničkog sučelja.

B. *Konfiguracija servo sustava*

Kako bi se moglo upravljati pojedinim servo pretvaračima bilo je potrebno dodati svakoga u TIA-portal projekt te im je bilo potrebno dodati IP- adrese u istom u istom sub netu kao što je kontroler. U TIA-portalu taj je problem jednostavno riješen kanon dodavanja svih potrebnih uređaja odlaskom na *Networked view* otvara se prozor sa vizualnom mogućnošću povezivanja uređaja. Odabirom međusobnog povezivanja svakom uređaju nakon kontrolera dodjeljuje se IP adresa koja odgovara ispravnom mrežnom području za komunikaciju. Uređaji međusobno komuniciraju preko PRROFINET-a industrijskog komunikacijskog standarda.

Komunikacija između kontrolera i pretvarača se vrši pomoću telegrama u ovom slučaju sa čak dva. Jedan telegram je *SIMENS telegtam 105* sastavljen od 10 kontrolnih riječi i 10 upravljačkih riječi. Pomoću tog telegrama se vrši glavno upravljanje i nadzor nad pretvaračem tipa trenutna pozicija motora, brzina vrtnje, status pretvarača itd. Drugi telegram je dodatni manji telegram *SIMENS telegtam 750* sastavljen od jedne kontrolne riječi i 3 statusne. Pomoću ovog telegrama se prati i upravlja moment motora. Specifično u ovom radu se taj telegram koristi za pokrivanje ne predvidivih greška kolizija sa objektima.

Nakon konfiguracije komunikacije potrebno je konfigurirati priključenu periferiju na pretvarač odnosno motor. Odabirom na model motora dodaju se svi njegovi podaci i ostvaruje se veza između enkodera i pretvarača.

Sva konfiguracija do sad navedena je konfiguracija koja se izvodi unutar alata *SIMATIC Startdrive.* Kako bi kontroler mogao upravljati pretvaračem potrebno ga je definirati unutar alata *SIMATIC Step 7*. Pretvarač se može definirati kao poseban *Tehnology object* iz SIMENS-ove serije *Motion control* blokova.

Unutar te konfiguracije pretvarača pruža nam se dodavanje graničnih senzora i senzora početne pozicije na koje se dodaju stanja digitalnih ulaza konrolera na koje su priključeni mikro prekidači i induktivni senzor.

Jedna od zgodnih opcija koja nam pruža ovaj servo sustav je automatsko postavljanje nulte ili početne pozicije. Nije potrebno pisati kod s kojim bi se izvršila ta radnja nego već postoji poseban blok koji pruža tu mogućnost. Unutar konfiguracije automatskog postavljanja početnog položaja potrebno je definirati na koji način će se odrediti nulta pozicija pomoću senzora.

Prvi slučaj je kada senzor dođe u visoko stanje da se kretanje zaustavi i sa niskom brzinom kreće u nazad sve dok se senzor ne vrati u nisko stanje i tada je ta pozicija definirana kao početna.

Drugi slučaj je da kada senzor dođe u visoko stanje nastavlja se gibanje u istom smjeru ali nižom brzinom dok senzor se ne vrati u nisko stanje i tada je ta pozicija definirana kao početna.

U slučaju da se os kreće u krivom smjeru postoji mogućnost odabira promjene smjera kada se okine jedan od graničnih kontakata. Brzina kretanja prilikom postavljanja početne pozicije se može također namjestiti po želji. U ovom radu odabrana je prva opcija kako bi se maksimalno moga iskoristiti duljina osi.



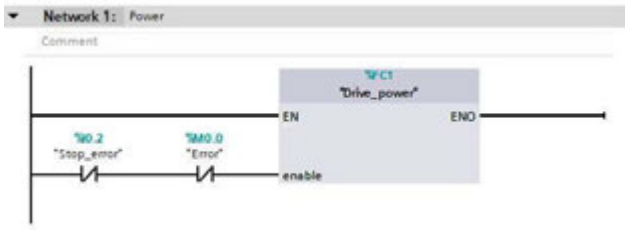
Slika 6 Postavke za automatsko postavljanje početne pozicije

IV. UPRAVLJAČKI PROGRAM

U sklopu upravljačkog programa važno je istaknuti 5 važnih programskih bloka pomoću kojih se pojednostavljuje i olakšava upravljanje servo sustavima.

A. *MC POWER*

Pomoću ovog bloka se aktivira mogućnost upravljanjem servo sustavom . Za svaki od pretvarača je potrebno definirati po jedan takav blok. Ukoliko se to ne učini nije moguće koristiti druge blokove niti upravljati pomoću njih servo sustav. MC POWER blok je aktivan sve dok na njgovom ulazu je visoki nivo. Zbog toga na ulaze svih MC POWER blokova je spojeno u logički ILI invertirano stanje preklopke u slučaju kvara„Stop_error“ i memorijski bit „Error“ koji se postavlja u visoko stanje ukoliko dođe do bilo neke vrste kvara.



Slika 8 Funkcija sa implementiranim MC POWER blokovioma

B. *MC HOME*

Ovim blokom se pokreće već prije spomenuto automatsko postavljanje početne pozicije. Blok se aktivira rastućim bridom na njegovom ulazu. Nakon završetka postavljanja početne pozicije na izlazu bloka „Done“ javlja se impulsni signal a kao svaki o servo sustava ima svoj blok za postavljanje u početnu poziciju blokovi su postavljeni tako da su na ulaz svakog prijašnjeg bloka spojeni izlazi prijašnjeg. Tako je osigurano da se inicijalizacija pozicije izvede sekvencijalno bez potrebe ručne aktivacije pojedinog bloka. Na istom principu su spojeni i ostali tipovi blokova osim bloka MC POWER:

C. *MC RESET i MC STOP*

Blokom MC RESET se u slučaju kvara sekvencijalno resetiraju kvarovi na svim pretvaračima kada na ulaz prvog bloka dođe rastući brid. Pomoću bloka MC STOP se zaustavlja kretanje svih motora. Važno je napomenuti da se ne isključuju nego zaustavljaju na trenutnoj poziciji što je iskorišteno kao funkcija za pauzu sustava.

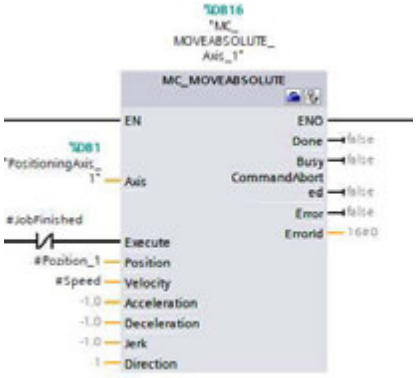
D. *MC MOVEABSOLUTE*

Ovim blokom se upravlja pozicioniranje pojedinog servo sustava. Uz poziciju može se zadati željena brzina kretanja, akceleracija, moment i usporavanje. MC MOVEABSOLUTE blokovi su također povezani jedan za drugoa tako da se sva kretanja izvršavaju sekvencijalno.

E. *Upravljački algoritam*

Upravljački algoritam je sastavljen od kombinacije svih ranije navedenih blokova. Najvažniji blok je svakako MC MOVEABSOLUTE pomoću kojeg se zadaje pozicija na koju se treba os pomaknuti. Kako bi se izvelo trebalo je izmisliti algoritam pomoću kojeg će kontroler sam moći odrediti potrebna kretanja kako bi se objekt sa skladišta uzeo i prenio na postolje tako da se slože kao paketi na paleti.

U prvoj izvedbi odabrano je rješenje sa poljem potrebnih pozicija za svaku os. Takvim rješenjem odradio se zamišljeni zadatak ali rezultat je bio sustav koji uvijek radi iste pokrete neovisno o broju objekta na skladištu te je svako polje sa pozicijama osi imalo po 86 vrijednosti u sebi. Kako bi se unaprijedilo rješenje osmišljen je



Slika 7 Blok za upravljane pokretima servo motora

algoritam koji dinamično sastavlja polja sa pozicijama ovisno o broju objekata na skladištu i postavljenih na postolje pazeći pri tome da ne dođe do kolizije prilikom transporta između hvataljke i već postavljenih objekata.

Kako bi se mogao napraviti taj algoritam analizom svih potrebnih pokreta napravljena su polja sa već predodređenim pozicijama. Radi veće modularnosti polja su podijeljena na dva dijela na „dolazna“ polja i na „transportna“ polja. Podjelom na te dvije kategorije smanjena je potreba za velikim brojem polja za svaku mogućnost dohvata i transporata na određenu poziciju.

Pri prvom pokretanju ovisno o broju objekta na skladištu i objekta postavljenih na postolje algoritam sastavlja polje za prijenos samo jednog objekta iz ponuđenih polja za dohvat i transport. Kada je taj jedan objekt prenesen sastavlja se opet novo polje uzimajući u obzir stanje na skladištu, broj postavljenih objekata i trenutnu poziciju na kojoj se nalazi hvataljaka. Tako se ponavlja algoritam dok više nema objekata na skladištu ili su svih 12 objekata složeni u dimenzije 4 x 3. Kada složi sve objekte ili kada više nema objekata za prijenos zadnje koordinate se postavljaju u nultu poziciju te se sustav prebacuje u stanje čekanja dok se ne nadopuni stanje



Slika 9 Prikaz sliženih objekata pomoću makete

skladištenih objekata ili resetira broj postavljenih objekata kako bi ponovno mogao početi sa slaganjem.

F. *Ekrani korisničkog sučelje*

Za nadzor i praćenje izrađeno je korisničko sučelje na kojem se može pratiti trenutna pozicija i stanje sustava.

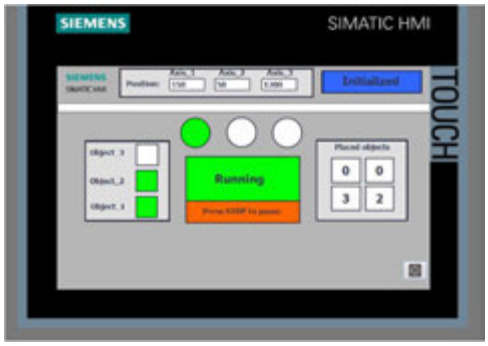
Na ekranu se dinamički pojavljuju različite poruke trenutnog stanja sa uputama što treba napraviti kako bi se nastavilo sa radom.

Kada je barem jedan objekt dostupan za transport i nema nikakvih drugih kvarova na ekranu se prikazuje poruka da je sustav spreman za pokretanje te kako ga pokrenuti. Dok nema nikakvih kretanja moguće je kliknuti na tipku „Reset objects“ kako bi se postavio broj postavljenih objekata na 0.



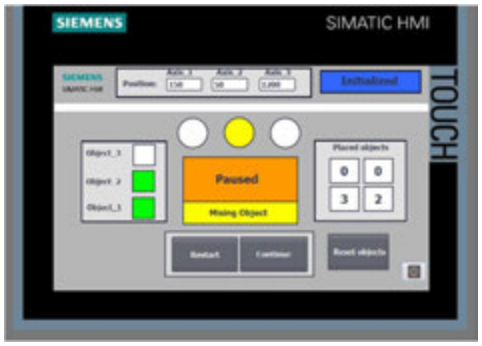
Slika 10 Ekran korisničkog sučelja kada je sustav spreman za rad

Prilikom kretanja poruka se mijena te se onemogućuje resetiranje broja složenih objekta kako bi to izazvalo grešku ukoliko ima već postavljenih objekata a sustav misli da ih nema.

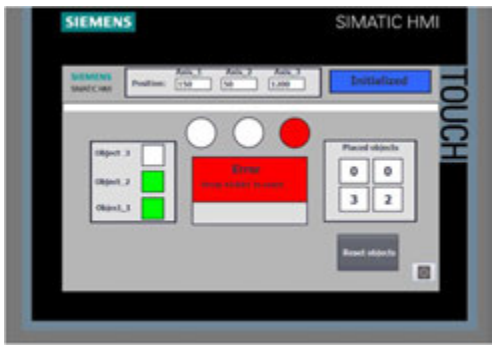


Slika 11 Ekran korisničkog sučelja kada se sustav u pokretu

Ukoliko se tijekom kretanja pritisne tipkalo STOP ili nestane objekt sa hvataljke sustav se prebacuje u stanje pauze te zaustavlja sva kretanja. Na ekranu je naznačeno to stanje i daje se mogućnost između nastavka sa kretanjem ako se vrati objekt ili resetiranja i kretanja iz početka. Ukoliko dođe do neke greške prikazuje se poruka o grešci i kako ju ukloniti. Kada je složeno svih 12 objekata osi se vračaju u početnu poziciju a na ekranu se ispisuje poruka da je posao gotov. Nije moguće ponovno pokrenuti slaganje dok se ne resetira broj složenih objekata.



Slika 12 Ekran korisničkog sučelja za vrijeme stanke



Slika 10 Ekran korisničkog sučelja u slučaju greške

V. ZAKLJUČAK

Svi zahtjevi koji su bili potrebni su realizirani ali što ne znači da ne postoje mane i nedostaci u radu te prostor za unaprjeđenje. Jedan veliki nedostatak ovog sustava je ne mogućnost fizičkog praćenja postavljenih objekata. Broj postavljenih objekta je mijenja samo programski po uspješnom prijenosu objekta. Ne postoji nikakvo drugo praćenje stanja kao što je napravljeno kod detekcije količine objekta na skladištu. Moguće poboljšanje bi bilo ugraditi industrijske kamere koje bi prepoznavale oblike i davale stvarnu povratnu informaciju o količini složenih objekata. Te velika nadogradnja bi bila dodavanje pokretne trake kako bi maketa kontinuirano mogla raditi do završetka slaganja jednog seta palete.

LITERATURA

[1] M. Lučan - Laboratorijski 3-osni model realiziran pomoću servo sustava SINAMICS S210 i kontrolera SIMATIC S7-150 Tehničko veleučilište u Zagrebu, Elektrotehnički odjel, Zagreb.

[2] G. Malčić, D. Maršić: Programirljivi logički kontroleri, interna skripta, Tehničko veleučilište u Zagrebu, Elektrotehnički odjel, Zagreb.

[3] SINAMICS S210 URL: https://cache.industry.siemens.com/dl/files/297/109763297/att_971661/v1/s210_1FK2_op_instr_01_2019_en-US.pdf (18.05.2020.)

[4] SIMATIC S7-1500 S7-1500T Motion Control V4.0 in TIA Portal V15 URL: https://support.industry.siemens.com/cs/document/109749263/simatic-s7-1500-s7-1500t-motion-control-v4-0-in-tia-portal-v15?dti=0&lc=en-GR (18.05.2020.)

Laboratorijska maketa redundantnog sustava upravljanja

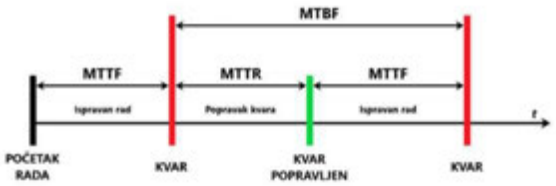
Antun Omrčen*, Mario Lučan* i Goran Malčić*
* Tehničko veleučilište u Zagrebu, Zagreb, Hrvatska
antun.omrcen@tvz.hr mlucan@tvz.hr goran.malcic@tvz.hr

***Sažetak* - Sadržaj rada opisuje opći pristup sustavima redundancije u industrijskim postrojenjima i problematiku koja je usko povezana sa ispravnim radom sustava te tehničko rješenje prikazano na stvarnom modelu laboratorijske makete. Dano tehničko rješenje je primjer pozicioniranja realizirano pomoću industrijske opreme pritom koristeći industrijsku računalnu mrežu kao komunikacijski kanal između upravljačkih jedinica sustava. Za praktičnu izvedbu tehničkog rješenja također je bilo nužno izraditi pasivne elemente u vidu nosača pojedinih elemenata te element za vizualnu prezentaciju trenutne pozicije. Radi se o jedinstvenom riješenu te je elemente bilo potrebno dizajnirati i izraditi u skladu potrebama tehničkog rješenja. Prilikom izrade tehničkog rješenja korišteni su razni programski alati te u konačnici i 3D printer za izradu elemenata.**

I. OPĆI KONCEPT REDUNDANCIJE

Zajednička karakteristika svih industrijskih sustava nevezana sa njihovom kompleksnosti i dizajniranim funkcijama imaju kvarove ili ispade pojedinih komponenti te ih je potrebno pratiti kako bi barem otprilike predvidjeli taj događaj. Kroz godine rada pokazalo se ključnim za uspjeh očuvanja sustava u radu praćenje pouzdanosti sustava. Mjerne vrijednosti kojima opisujemo događanje kvarova su nazvane mjerne vrijednosti neuspjeha. Preko tri parametra postiže se zadovoljavajuća informacija pomoću koje se organiziraju i izvode ključne aktivnosti u cilju prevencije neplaniranih zaustavljanja sustava. Ključni parametri su: MTTR (engl. *Mean Time To Repair*, hrv. srednje vrijeme za popravak), MTBF (engl. *Mean Time Between Failures*, hrv. srednje vrijeme između kvarova) te MTTF (eng. Mean Time To Failure, hrv. srednje vrijeme kvara).

Svaka spomenuta vrijednost ima svoju namjenu, razlog i način kako pridonosi ključnim informacijama za praćenje pouzdanosti. U cilju shvaćanja potrebama sustava za redundantnim komponentama i elementima detaljnije je potrebno razumjeti parametar MTTF. Spomenuti parametar je od velike važnosti jer se koristi za procjenu životnog vijeka elementa ili određenog sustava. Moguće ga je primijeniti na sve oblike komponenti, elemenata i sustava. Osobito je važan prilikom izračuna pouzdanosti pojedinih komponenti koje se dalje ugrađuju u veće elemente ili sustava, koje su ugrađene u procese koji su izuzetno osjetljivi na zastoje i kvar opreme. Jednako tako se primjenjuje u slučajevima kod kojih nije moguć trenutni popravak ili uopće nije moguć popravak te se ovisno o elementima može implementirat neki od oblika redundancije o kojima će nešto kasnije biti spomenuti.



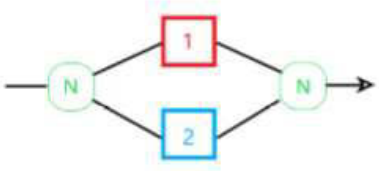
Slika 1 Odnos parametara MTTR, MTTF i MTBF

II. MODELI REDUNDANCIJE PREMA AKTIVNOSTI U SUSTAVU

A. Redundancija u stanju pripravnosti

Karakteristika za redundanciju u stanju pripravnosti je ta da je sekundarna jedinica isključuje te se na taj način čuva njena pouzdanost. Kroz isključivanje sekundarne jedinice smanjujemo njen radni vijek te samim time do određene mjere čuvamo uređaj do onog trenutka kada je stvarno potreban njegov radu unutar samog sustava. Nedostatak ovog dizajna je što je vrijeme preuzimanja primarne funkcije duže nego u aktivnom stanju.

Na slici 2 prikazan je model, element 1 je u aktivnom stanju te tok informacija ide preko njega dok je element 2



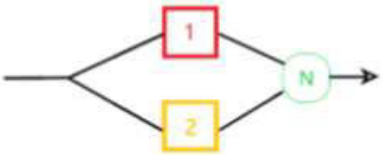
Slika 2 Model redundancije u stanju pripravnosti

u pasivnom stanju te čeka on nadzorne aplikacije ili sustava naredbu za paljenje. Nadzorna aplikacija ili sustav na modelu je označen sa N, te kroz njega ulaze i izlaze informacije kako bi adekvatno mogao izvršiti sinkronizaciju ukoliko do dođe do toga da se element 2 aktivira te preuzme ulogu elementa 1.

B. Pasivna redundancija

Primjena pasivne redundancije je prikladna za sustave i procese u kojima su važni vremenski odzivi ali trenutni ili kratki prekid je i dalje prihvatljiv.

U ovakvom primjeru redundantni elementi češće su opterećeni sa manje zadataka ili manjim opterećenjem od glavnog elementa sve dok glavni element ne zakaže ili prestane ispravno obavljati dizajniranu funkciju. Tako se na sličan način kao i kod redundancije u stanju pripravnosti produžuje životni vijek elemenata ali ne u tolikoj mjeri kao što je to kod pasivnog stanja. Prihvatljivi vremenski period ispada elementa iz dizajnirane funkcije može potrajati sekundu, minutu ili već koliko je dopušteno kroz tehnološki proces ali općenito postupak preuzimanje glavne uloge odrađuje se što je moguće brže sa implementiranim elementima.

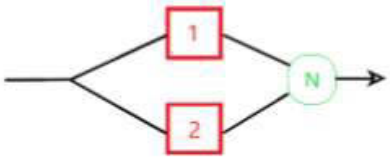


Slika 3 Model pasivne redundancije

C. Aktivna redundancija

Karakteristično za aktivnu redundanciju je da su primarna i sekundarna jedinica upaljene i aktivno sinkronizirane. Ukoliko sekundarna jedinica postavljena tako da može odlučiti kada napraviti prebacivanje tada nema potrebe za dodatnim programom koji bi obavljao tu funkciju. Ovaj dizajn ne čuva uređaj i njegovu pouzdanost kao što je to slučaj kod pasivnog stanja pripravnosti. Velika prednost ovog dizajna je što skraćuje zastoje u radu dizajniranog sustava te u konačnici zauzvrat povećava raspoloživost sustava.

Glavna razlika između pasivnog sustava se odnosi na stupanj sinkroniziranosti između primarne i sekundarne jedinice. Aktivna sinkronizacija postavlja komunikacijskim kanalima zahtijevan zadatak koji dovodi do revolucionarnih rješenja i vrlo velikih brzina komunikacije između postavljenih jedinica. Također izvedbe upotrijebljenih jedinica moraju biti izvedene na način da mogu preuzeti primarnu ulogu u sustavu bez najmanjih vremenskih odgoda, kroz takvu brzu promjenu ostatak sustava nesmetano nastavlja sa dizajniranim radom kao da se kvar nije niti dogodio.



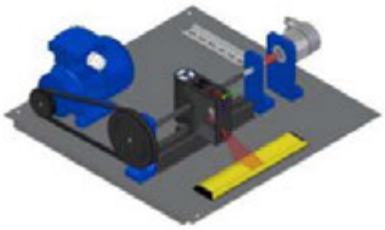
Slika 4 Model aktivne redundancije

III. OPIS TEHNIČKOG PROBLEMA

Projektni zadatak je zadan u svrhu dokazivanja jednog od ranije navedenih modela redundantnih sustava. Sa opremom o kojoj će biti riječi nešto kasnije, izveden je model dvostruke pasivne redundancije. Idejno rješenje je bilo da ukoliko se dogodi do ispada jednog od upravljačkih jedinica ostali uređaji i aktuatori unutar sustava nastave obavljati svoju dizajniranu funkciju bez ikakvih spoznaja da je došli do bilo kakvog prekida ili ispada pojedinog PLC-a. Dizajnirana funkcija sustava je automatsko pozicioniranje elementa sa promjenjivim parametrima. Kako bi uređaji mogli međusobno komunicirati također je potrebno izvesti neki tip industrijske računalne mreže. Također uz izvedenu maketu potrebno izraditi elektro shemu na kojoj će biti opisano kako je koji od pojedinih uređaja spojen.

IV. OPIS TEHNIČKOG RJEŠENJA

Prilikom izrade tehničkog rješenja glavni cilj bio je dokazati funkcionalnost redundancije u vidu PLC-a i upravljanjem sustavom ukoliko dođe do ispada jednog od upravljačkih jedinica. Za adekvatnu prezentaciju tog cilja potrebno je bilo osmisliti neki tip industrijskog procesa kako bi se ispravno adekvatno mogao dokazati ostvareno rješenje. Samim time što se radi o simulaciji industrijskog rješenja korišteni su elementi predviđeni za tu namjenu. Za upravljanje motorom koristi se standardni pretvarač frekvencija koji upravlja asinkronim kaveznim motorom, to je primjer je se vrlo često pojavljuje u industrijskim postrojenjima kao kvalitetno rješenje ukoliko postoji potreba za pokretima u procesu što je vrlo čest slučaj. Prijenos okretnog momenta sa motora na osovinu koja pokreće element za kontrolu pozicije izveo sam pomoću zupčastog remena i dvije remenice. Također prilikom prijenosa remenice su izvedene u omjeru 0.5 što smanjuje brzinu elementa za 50 % u odnosu na elektro motor. Maketa na kojoj se izvodi prezentacija tehničkog rješenja pruža opciju vizualne kontrole pozicije elementa koji laserskim snopom pokazuje trenutnu poziciju prikazanu na mjernoj letvi lociranoj na lijevoj strani makete kao što je vidljivo na slici 5. Kao što je spomenuto rješenje simulira industrijske uvjete rada te je za točnu preciznost pozicije zadužen apsolutni pretvornik pokreta.

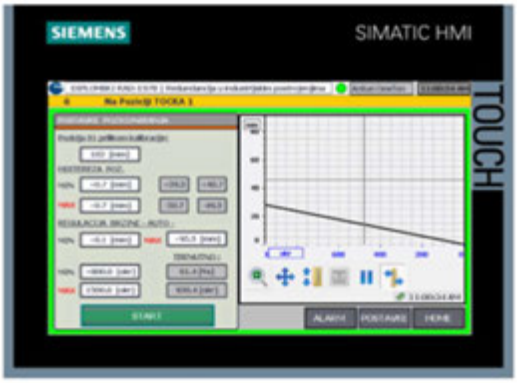


Slika 5 Inicijalni model makete

A. HMI sučelje

Zamišljena funkcija aplikacije je prilikom inicijalnog paljenja upravljačkih elemenata potrebna je kalibracija sustava. Preko HMI (engl. *Human Machine Interface*) aplikacije moguće je postaviti određene parametre prema kojima bi se sustav trebao korigirati kao što su dozvoljena odstupanja prilikom pozicioniranja i brzine prilikom automatskog rada sustava.

Preko početnog ekrana moguće je odabrati način rada sustava između ručnog rada sustava ili automatskog rada sustava.



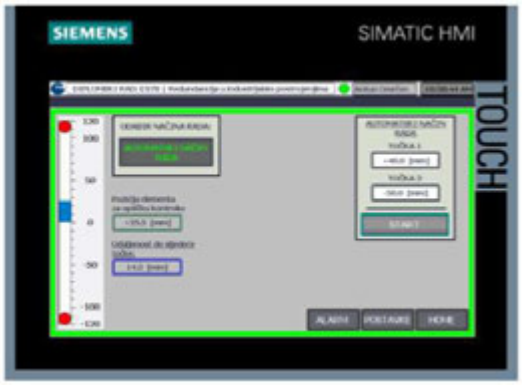
Slika 6 Ekran za podešavanje parametara

Ukoliko je odabran ručni način rada, korisnim može unijeti željenu vrijednost brzine skalirane u rasponu od 0 – 100 % dok je ona na daljnjim elementima dodatno reducirana. Nakon što je odabran odgovarajući broj, korisnik ima mogućnost upravljat elementom za optičku kontrolu pozicije. Taj element se nalazi na osovini sa navojem te ovisno o pritisnutoj tipici element se kreće gore ili dolje u limitiranim okvirima. U sustav su također ugrađeni senzori koji se nalaze na rubnim točkama hoda elementa za optičku kontrolu. Ukoliko element u svome radu dođe do primjerice donjeg senzora, sustav operateru ne dozvoljava daljnje kretanje elementa prema dolje. Ta situacija vrijedi i za gornju poziciju sustava, ukoliko element dođe do gornje točke te aktivira senzor, sustav neće dozvoliti operateru daljnje kretanje elementa prema gore.



Slika 7 Ekran za ručni način upravljanja

Automatski dio aplikacije se odnosi na dio aplikacije gdje je operateru ponuđeno određeno područje u kojem će se element kretati počevši prvo prema pozitivnoj ili prvoj zadanoj točci te nakon što dođe do nje krenut će prema negativnoj odnosno drugoj zadanoj točci. Prije nego element počne raditi u automatskom načinu rada potrebnoj napraviti automatsku kalibraciju elementa. Ukoliko nije zadovoljena na HMI-u će se pojaviti tipka koja pokreće automatsku kalibraciju. Automatska kalibracija unaprijed postavljenom brzinom pomiče element prema gornjem senzoru. Kada element dođe do krajnje točke sustav je spreman za automatski rad. Također u parametrima programa moguće je jednostavno unijeti korekciju gornje točke ukoliko je to potrebno. Trajanje automatskog rada nije ograničeno te sustav radi sve dok ima zadovoljene uvijete za normalan rad.



Slika 8 Ekran za automatso upravljanje

B. Mehanička izvedba tehničkog rješenja

Radi fizičkih ograničenja u koje je upravljačka oprema morala stati te sama testna maketa, idejno rješenje makete prošireno je na dvije zasebne montažne ploče. Na prvoj ili upravljačkoj ploči ugrađena je upravljačka oprema potrebna za obavljanje logičkog dijela projektnog zadatka dok je na drugoj ploči izvedena maketa koja se sastoji od izvršnog dijela aplikacije kao što je motor, zupčasti prijenosnici, element preko kojeg dokazujemo optičku poziciju elementa te također pretvornik pomaka preko kojeg pratimo stvarnu poziciju predmeta. Elementi za osmišljen projektni zadatak nisu standardni na tržištu već se ih je bilo potrebno projektirati prema projektnom zadatku. Zadatak je specifičan po tome što inače u industrijskim postrojenjima se raspolaže sa nešto više prostora za ugradnju opreme te je bilo potrebno realizirati zadatak u relativno malo fizičkog prostora.



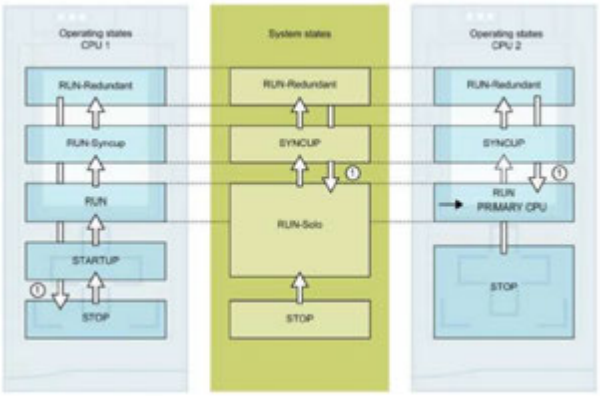
Slika 9 Maketa redundantnog sustava upravljanja

C. PLC u redundantnom sustavu

Tehničko rješenje sam izveo pomoću dva ista kontrolera *S7-1513R-1 PN*. Radi se o osnovnom modelu PLC-a koji podržava funkcionalnosti potrebne za izvedbu redundantnog rješenja kao što su redundancija CPU-a (engl. *Central Processing Unit*, hrv. centrala procesna jedinica) i protokole vezane za *PROFINET* redundanciju koje smo ranije spominjali. Spomenuti kontroler nema mogućnost korištenja dodatnih ulazno/izlaznih jedinica direktno povezanih u internu sabirnicu kao što je to slučaj koji je vidljiv prilikom izrade rješenja koristeći kontrolere iz serije *S7-1500* iz razloga što je njegova namjena

isključivo upravljanje procesom te kao takav ne podržava taj tip funkcija. Za očitanje i kontrolu ulazno/izlaznih komponenti zadužene su ostali uređaji spojeni u istu mrežu. Iz tog razloga na uređaju su izvedena samo tri konektora, prvi za osiguravanje napona potrebnog za rad uređaja te ostala dva za izvedbu industrije računalne mreže pomoću koje uređaji upravljaju procesnim jedinicama. Također bitno je napomenuti iako uređaj posjeduje dva spojna mjesta riječ je o jednom *PROFINET* sučelju. Za izvedbu ovog tehničkog rješenja to je bilo sasvim dovoljno ali ukoliko se žele implementirati napredana rješenja napredi modeli podržavaju više razlučih sučelja. Svaki uređaj također sadrži memorijsku karticu na kojoj su spremljeni upisani parametri te iz koje se iščitava izrađeni program.

Realizirano tehničko rješenje radi na način da prilikom ispada zamjenskog PLC-a sustav izbaci grešku kako je došlo do problema sa redundantnim sustava ali ostaje u normalnom radu. Ukoliko dođe do ispada glavnog PLC-a stanica postavlja zamjenski PLC u glavnog te se program nastavlja sa odvijanjem kao da se niša nije dogodili. Prilikom prebacivanja stanica upućuje uređaje povezane u mreži kako svoje telegrame sada šalju na drugu adresu. Prilikom toga dolazi do ponovne inicijalizacije ukoliko uređaj ne šalje apsolutu vrijednosti ili ne podržava S2 oblik komunikacije. Ukoliko je uzrok ispada PLC-a otkriven ili je uređaj popravljen prilikom vračanja uređaja na mrežu ne dolazi do zastoja nego uz nesmetano odvijanje procesa stanica pokreče sinkronizaciju zamjenskog PLC-a sa realnim informacijama koje su u glavnom PLC-u. Na slici 10 prikazana je slika sinkronizacije dva PLC-a dok je aplikacija u radu, sustav javlja grešku koja označava da sustav nije redundantan ali i dalje je u radu. Nakon što je sinkronizacija izvršena u cijelosti tada se sustav ponovo smatra redundantnim. Također bitno je spomenuti kako se programski blokovi nalaze samo na jednom PLC-u. Prilikom spuštanja moguće je program spustiti samo u glavni PLC, te nakon što je program uspješno spušten stanica prebacuje promjene sa glavnog PLC-a na zamjenski PLC. Preko te funkcionalnosti olakšava se izvedba programskog rješenja te se osigurana je vjerodostojnost podataka na oba PLC-a.



Slika 10 Grafički prikaz prebacivanja upravljanja između dva PLC-a

D. 3D printani elementi implementirani u tehničko rješenje

Za izvedbu tehničkog riješe potrebno je bilo realizirati određene mehaničke komponente kako bi cjelokupna testna maketa postigla određenu formu potrebnu za smislenu simulaciju industrijskog okruženja. Riječ je o jedinstvenim prototipnim elementima te za njihovu izvedbu idealno odgovara tehnologija 3D ispisa. Elementi su izrađeni od različitih polimera koji omogućavaju dovoljnu čvrstoću i žilavost materijala kako bi meketa moga ispravno obavljati dizajniranu funkciju.

Filamentni korišteni u realizaciji tehničkog rješenja se uvelike razlikuju jedan od drugoga. Razlikuju se u čvrstoći, žilavosti, otpornosti na habanje i po mnogim drugim kriterijima. Za realizaciju cjelokupnog rješenja izrađeno je 15 elemenata te ukupni rad 3D printera prelazi 62 radna sata. Korišteni filamentni su *PETG* (engl. *Polyethylene Terephthalate (PET) with Glycol*), *TPE* (engl. *Thermoplastic elastomer*), *NGEN* (*sastavljen od Eastman Amphora™ AM3300 3D polymera*) te *PLA* (engl. *Polylactic Acid*).

V. PROGRAMI KORIŠTENI ZA POSTIZANJE TEHNIČKOG RJEŠENJA

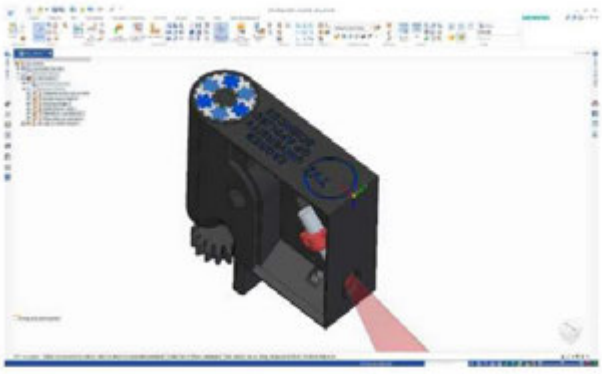
A. TIA portal v16

TIA portal (engl. Totally Integrated Automation) je programska platforma razvijena u cilju realizacije automatizacijskih zadataka, pritom pružajući neograničeni pristup cjelovitom nizu digitalnih automatizacijskih usluga. Funkcionalnost, performanse i učinkovitost u složenim industrijskim procesima su traženih zahtjevi te je kompanija Siemens odgovorila na te zahtjeve sa *TIA portalom*. Pomoću ovog alata izrađen je program koji upravlja ranije spomenutom aplikacijom. Prema standardnoj normi *IEC 61131-3* u alatu su integrirane standardne programske metode: ljestvičasti dijagram, funkcijski blok dijagram, instrukcijske liste, strukturirani tekst te sekvencijalni funkcijski dijagram. Prilikom izrade spomenute aplikacije korišten sam ljestvičasti dijagram za realizaciju programskog koda. Kroz alat se postavlja industrijsku računalna mreža *PROFINET* te svim uređajima u mreži definira adresa i ostali parametri potrebni za ispravnu komunikaciju.

B. Solid Edge 2020

Alat za izradu modela *Solid Edge* je zajednička platforma softverskih alata koji se koriste i koji se bave svim aspektima procesa razvoja proizvoda odnosno 3D dizajnom, simulacijom, proizvodnjom i još mnogo toga. *Solid Edge* omogućuje brzu i jednostavnu izradu 3D modela uz fleksibilnost i kontrolu parametrijskog dizajna. Postoje različiti načini za izradu modela ali u ovom radu se koristi postupak modeliranja koji započinje osnovnom značajkom koju kontrolira 2D skica, koja je linearna, okretna, podignuta ili istisnuta ekstruzija. Također značajka je to da je svaki sljedeći dodatak na crtež izgrađen na prethodnoj značajki. Kako bi bilo vidljiva međusobna korelacija i funkcija elemenata Solid Edge podržava izradu zajedničkog modela u kojem se nalaze ranije modelirani

modeli. Organizacija izrađenih dijelova postavljena je u karakterističnu slijednu strukturu te je jednostavan pregled dijelova, pod dijelova te samih radnji izvršenih na pojedinim elementima. Kroz taj dio alata moguće je simulirati određene pokrete, te uvidjeti ukoliko nešto ne funkcionira na zamišljeni način. Također moguće je izraditi točan plan i poziciju elemenata, na taj način se olakšava priprema elementa za montažu.



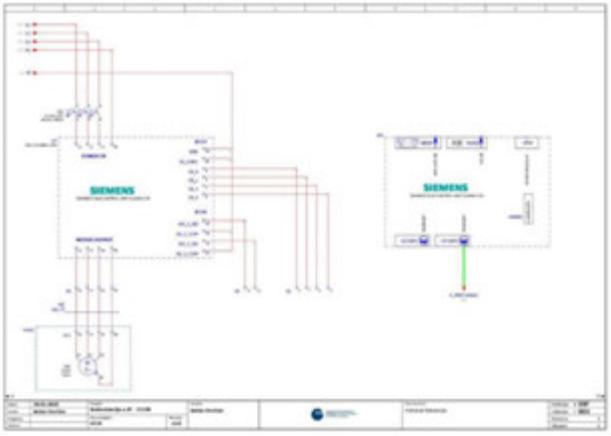
Slika 11 Sučelje alata Solid Edge sa elementom za optičku kontrolu pozicije

C. *PrusaSlicer*

Nakon što je model preveden u neki od formata primjerice STL (engl. *Standard Tessellation Language*) ili 3MF (engl. *3D Manufacturing Format*) potrebno je dizajnirani model pretvoriti u kod koji 3D printer razumije a riječ je o *G-kodu*. Također specifična tehnologija 3D ispisa zahtijeva nešto drugačije parametre od klasičnih alata koje koriste G-kod ali princip čitanja koda je isti. Prednosti alata je ta što je proizveden od istog dobavljača kao što je i 3D printer što je izuzetno bitno radi nivoa integracije alata i parametara na printeru. Nakon što je model uvezen u slicer podesio sam niz parametara kako bi uvezeni element bio uspješno isprintan. Osnovne postavke kao temperatura ekstruzivne glave, temperatura podloge, brzine kretanja i akceleracije glave, količina ispune elementa samo su među prvima u nizu parametara potrebnih za uspješan print. Također svaki je filament različit te se postavke odstupaju za svaki pojedini materijal. Nakon izrađenog programa alat generira kod te prikazuje element u nivoima kako bi se mogla izvršiti pojedina prepravka ukoliko je nužna.

D. *EPLAN P8 v2.8*

Programski inženjerski alat čije su usluga široko primjenjuju na području elektrotehnike, automatizacije i mehatronike. Dosljedna digitalna platforma osnova je za učinkovitu izradu električne dokumentacije neophodne za izradu projekta u industrijskim postrojenjima. Programska platforma podržava niz dodatnih mogućnosti i nadogradnji u cilju efikasnije izvedbe tražene dokumentacije. Danas u širokoj primjeni su izrada 3D modela elektro ormara radi jednostavnije prezentacije i vizualizacije komponenti koje trebaju biti ugrađene na pojedinu lokaciju te usluge podatkovnog portala na kojem se nalaze aktualne komponente svih vodećih proizvođača komponenata. Alat pruža automatsku integraciju spomenutih komponenti u dokumentaciju. Pomoću njega sam izradio elektro dokumentaciju koja realiziranom tehničkom rješenju pridodaje kompletnost rješenja. Izrada elektro dokumentacije obavezna praksa prilikom izrade tehničkih rješenja koja u sebi sadrže komponente koje su povezane raznim spojevima jedne sa drugima. Pomoću elektro dokumentacije jednostavnije je vidjeti kako i sa kojim signalima je oprema povezana.



Slika 12 Shema spajanja motra i pretvaraća frekvencije

VI. ZAKLJUČAK

Prilikom izrade tehničkog rješenja uočeni su pojedini nedostatci prilikom sinkronizacije u vidu brzine izmjene podataka između glavnog i zamjenskog PLC-a te problemi sam komunikacijom između sustava i HMI-a. Problem nije utjecao na ishod rješenja ali ukoliko se radi o brzim očitanjima svakako je potrebno izvesti napredniji redundantni sustav. Izvedba naprednijeg sustava potreban je dodatni hardver, ali ispravnom implementacijom moguće je postići veliki napredak u stabilnosti sustava. Primjer implementacije funkcionalnijeg hardvera omogućava dodatno proširenje *PROFINET* sučelja te za sinkronizacijski kanal ne koristi standardno *PROFINET* sučelje već je implementiran dodatni modul koji sinkronizaciju glavnog i zamjenskog PLC-a izvršava putem optičkih vlakana i korištenjem naprednijih protokola. Dodatna proširenja u *PROFINET* sučelju omogućavaju danas tražene funkcionalnosti praćenja sustava te funkcionalnost da je sustav poveziv na vanjske uređaje poput dodatnih servera preko kojih je dalje omogućena dubinska analiza i detaljnije nadgledanje cjelokupnog proizvodnog procesa.

LITERATURA

[1] A. Omrčen -Redundantni sustavi upravljanja u industrijskim postrojenjima , Tehničko veleučilište u Zagrebu, Elektrotehnički odjel, Zagreb.

[2] A. Birolini, „Reliability Engineering“ šesto izdanje, Springer Berlin Heidelberg 2010.

[3] D.Cindori - Redudantni sustavi u automatiziranim postrojenjima 2012., diplomski rad, diplomski, Elektrotehnički odjel Tehničkog veleučilišta u Zagrebu, Zagreb

[4] National Instruments, Redundant System Basic Concept URL: https://www.ni.com/en-rs/innovations/white-papers/08/redundant-system-basic-concepts.html (12.02.2020.)

[5] SIEMENS; SIMATIC S7-1500 S7-1500R/H redundant system URL: https://support.industry.siemens.com/cs/document/109754833/simatic-s7-1500-s7-1500r-h-redundant-system?dti=0&lc=en-WW (16.05.2020.)

Samoregulacija učenika u virtualnoj učionici nastave matematike

Ružica Soldo*, Jelena Domac** i Ivana Olujić***
* Strojarska tehnička škola Fausta Vrančića, Zagreb, Hrvatska
** Osnovna škola Stjepana Antolovića, Privlaka, Hrvatska
***Ministarstvo znanosti i obrazovanja, Zagreb, Hrvatska
ruzica.soldo@skole.hr, jelena.domac@skole.hr, iolujic81@gmail.com

Sažetak - **Ovim istraživanjem pokazalo se kako rad u virtualnoj učionici uz inovativnu pedagošku metodu WTL ("Write To Learn", odnosno "Pisanjem učiti") utječe na samoregulaciju učenja u nastavi matematike. U istraživanju je sudjelovalo 156 učenika iz dvije srednje strukovne škola u Gradu Zagrebu, primjenio se eksperimentalni nacrt istraživanja na dvije skupine ispitanika. Kontrolnu skupinu čini 73 učenika, koji koriste IKT u virtualnoj učionici bez digitalno socijalne interakcije i formativne procjene kroz rad na virtualnoj ploči. Eksperimentalnu skupinu čine 83 učenika koji koriste IKT uz formativno vrednovanje i društvene interakcije ističući interakciju učenik-učenik i vršnjačko vrednovanje kroz rad na virtualnoj ploči. Rezultati istraživanja pokazuju kako korištenje IKT-a u kombinaciji s pedagoškom metodom koja uključuje formativno vrednovanje kao digitalnu socijalnu interakciju među vršnjacima kroz korištenje virtualne ploče povećava samoregulaciju učenika u virtualnoj učionici nastave matematike.**

Ključne riječi - formativna procjena; samoregulacija; virtualna ploča;WTL (metoda pisanjem učiti)

I. UVOD

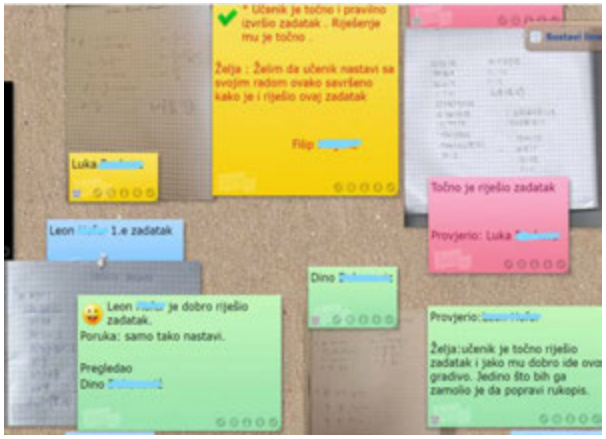
Koncept samoregulacije učenja može se definirati kao aktivni proces u kojem učenici sustavno usmjeravaju svoje misli, osjećaje i postupke kako bi postigli prethodno postavljene ciljeve [10]. Akademska samoregulacija označava sposobnost regulacije vlastitog ponašanja, specifično vezanog za učenje u školi. Samoregulacija kao mjera motivacije najbolje se odražava u intrinzičnoj i ekstrizičnoj regulaciji motivacije [11]. Intrizično motivirani učenici sudjeluju u aktivnosti kako bi zadovoljili svoju znatiželju i zanimanje za nastavno gradivo te kako bi razvili kompetenciju ili sposobnost u odnosu na zahtjeve koji se pred njih stavljaju, dok za ekstrinzičnu motivaciju ljudi za svoja ponašanja dobiju poticaje izvana [12]. Razlikujemo četiri tipa ekstrizične regulacije: eksternalnu, introjeciranu, identificiranu i integriranu regulaciju [11]. Eksternalna regulacija uključuje ona ponašanja koja pojedinac izvodi u radu da dobije neku vrstu nagrade ili da ugodi drugima. Introjeciranu regulaciju prepoznajemo kroz aktivnosti koje pojedinac izvodi kako bi izbjegao krivnju ili ojačao ego dobivanjem društvene potpore, dok u integriranoj regulaciji pojedinac pokazuje visok stupanj autonomije i obavlja aktivnosti jer on to stvarno želi. Identificirana regulacija postiže se kada se neko ponašanje osobe priznaje kao vrijedno, dakle osoba nešto ne čini jer ima osjećaj da se to mora učiniti, nego zbog toga što se to smatra važnim [13]. Pregledom istraživanja koja su proučavala samoregulirano učenje i njegovu školsku primjenu zaključuje se kako poučavanje vršnjaka može doprinijeti povećanom korištenju učinkovitih strategija učenja kod učenika [7]. Nastavnici mogu učenicima pružati podršku i kroz suradnju s vršnjacima, koji međusobno dijele ideje i strategije rješavanja problema. Učenicima na taj način daju priliku međusobnog pomaganje čime jačaju njihovu samoregulaciju i motivaciju za rad. Boekaerts i Cascallar [5] ističu da je interakcija koju učenici ostvaruju s vršnjacima i nastavnicima u razredu ključna za razvoj samoregulacijskih vještina. U novije vrijeme se interakcija spominje u kontekstu poučavanja na daljinu. Govori se o tri tipa interakcije: učenik-sadržaj, učenik-učenik i učenik-nastavnik [3]. Unapređivanjem interakcije učenik-učenik kroz aktivnosti učenja, povećaju se ne samo postignuća učenika, već i njihova samoefikasnost učenja [9].

Inovativna pedagoška metoda WTL ("Pisanjem učiti") uključuje IKT, kombinirajući ga sa socijalnom interakcijom učenika. Ova metoda je razvijena 2011. godine pod nazivom "Write To Learn" i cilj joj je bio strukturirati ključne aspekte koji pokazuju kako integrirati obrazovanje s informatičkom tehnologijom, kako bi što bolje iskoristili mogućnosti koje nudi tehnologija, uz cilj smanjivanja različitosti među rodnim razlikama učenika[1]. WTL propisuje društvenu uporabu IKT-a koja uključuje komunikaciju i formativne povratne informacije i procjene među učenicima, kao i između učenika i nastavnika. Učenici jedni drugima daju pisanu formativnu povratnu informaciju kroz zajedničke digitalne dokumente interaktivno. Studija koja je uspoređivala učenička postignuća u švedskim nacionalnim standardnim testovima pismenosti i matematike došla je do zaključka kako je upotreba IKT-a u školama bez strukturirane integracije s pedagogijom neučinkovita, čak i lošija od tradicionalne metode poučavanja bez IKT-a [2].

Black i Wiliam ističu u svom istraživačkom radu da se značajni pozitivni učinci na uspješnost učenika u nastavi matematike mogu postići formativnim vrednovanjem stavljajući naglasak na vršnjačko vrednovanje [4]. Formativne povratne informacije i procjene koriste se sustavno , a IKT je važan medij za povećanje učinaka formativne povratne informacije. Primjenom zajedničkih digitalnih dokumenata Google diska i virtualnih ploča: Padlet, Linoit, Whiteboard u matematici omogućeno je učenicima stvaranje i rješavanje matematičkih problema (Slika1.) pismenom suradnjom i međusobnim komuniciranjem (Slika 2.). Riješene probleme objavljuju i formativno međusobno vrednuju. Obzirom da je dobivanje povratnih informacija složen proces te zahtjeva određene vještine, učenike se kontinuirano podučavalo pomoću uputa i prakse kako bi razvijali vještine preispitivanja tuđeg rada na formativan način. Učitelj u cijelom procesu ima bitnu ulogu i donosi zaključnu formativnu procjenu nakon što učenici naprave međusobne formativne procjene.



Slika 1. Prikaz problemskog zadatka na virtualnoj ploči Linoit



Slika 2. Prikaz formativne procjene na virtualnoj ploči Linoit

II. METODOLOGIJA

Cilj ovog istraživanja je pokazati kako korištenje IKT-a u kombinaciji s pedagoškom metodom koja uključuje formativno vrednovanje kao digitalnu socijalnu interakciju među vršnjacima utječe na samoregulaciju učenika u virtualnoj učionici nastave matematike.

A. Hipoteza

H1: Korištenje IKT-a u kombinaciji s pedagoškom metodom koja uključuje formativno vrednovanje kao digitalnu socijalnu interakciju među vršnjacima povećava samoregulaciju učenika u virtualnoj učionici nastave matematike.

B. Uzorak istraživanja

Prigodni uzorak čini 156 učenika i učenica dvije strukovne škola u Gradu Zagrebu. U jednoj školi tri su razredna odjeljenja koristila virtualnu ploču i formativno vršnjačko vrednovanje, a u drugoj školi tri razredna odjeljenja su radila virtualnu nastavu uz IKT bez korištenja virtualne ploče i formativnog vršnjačkog vrednovanja kao digitalno socijalne interakcije među vršnjacima.
Istraživanje se provelo koristeći eksperimentalni nacrt na dvije skupine ispitanika. Kontrolnu skupinu čine 73 učenika koji koriste IKT kroz rad u virtualnoj učionici, a eksperimentalnu skupinu čini 83 učenika koji koristi IKT uz formativno vrednovanje i društvene interakcije ističući interakciju učenik-učenik i vršnjačko vrednovanje.

C. Instrument istraživanja

U istraživanju je korišten upitnik Ryan i Connell [8] "Upitnik za akademsku samoregulaciju" verzije SRQ-A. Connel i Ryan su valjanost skale dokazali na 4 različita uzorka, koje je činilo oko 2000 učenika. [8]
Upitnik je za istraživanje u Republici Hrvatskoj koristio i preveo Goldin [6] na uzorku od 132 učenika srednje škole i pokazao kako koeficijenti pouzdanosti za subskale akademske samoregulacije ukazuju na dobru pouzdanost pa se kao takav koristi i u ovom istraživanju.
Postavljaju se četiri različita pitanja učenicima, a svakom pitanju slijedi osam odgovora koji su prikazani na Likertovoj skali vrijednostima od 1 do 4, dakle imamo ukupno 32 čestice. Vrijednosti su numerirane slijedećim redoslijedom ; 1= ne slažem se, 2= donekle se ne slažem,3= donekle se slažem, 4= slažem se. Ispitanik mora odabrati tvrdnju koja odgovara njegovom stavu. Tvrdnje predstavljaju jednu od 4 stupnja regulacije motivacije: eksternalnu regulaciju, introjeciranu regulaciju, identificiranu regulaciju i intrinzičnu regulaciju. Eksternalna i introjecirana regulacija čine

kontroliranu motivaciju, jer kontrola dolazi izvan osobe, dok se identificirana i intrizična motivacija nazivaju autonomnom motivacijom, jer njom upravlja sama osoba [14]. Intrgrirana regulacija nije korištena jer se rijetko pronalazi među školskom djecom.

D. Rezultati istraživanja

Na temelju teorijskih indikacija, konstruirali smo dva aditivna indeksa za dva generalna tipa regulacije, na način da smo zbrojili referentne čestice svakog teorijskog generalnog tipa regulacija te dobivenu sumu podijelili s brojem čestica, kako bi ujednačili skalni raspon oba generalna indeksa regulacije. Stvorena su dva indeksa regulacije, kontrolirana i autonomna regulacija. Kontroliranu regulaciju čine eksternalna i introjecirana regulacija dok autonomna regulacija sadrži identificiranu regulaciju i intrinzičnu regulaciju. Sudeći prema prosječnim vrijednostima, ispitani učenici su u većoj mjeri regulirani kontrolnom regulacijom, dok su u manjoj mjeri regulirani autonomnom regulacijom (Tablica I).

TABLICA I. DESKRIPTIVNI STATISTIČKI POKAZATELJI ZA KONTROLIRANU I AUTONOMNU REGULACIJU

Deskriptivni statistički pokazatelji	**Vrsta regulacije**	
	Kontrolirana	***Autonomna***
Valjane frekvencije	156	156
Nedostajuće frekvencije	0	0
Prosječna vrijednost	2,6026	2,4886
Medijalna vrijednost	2,6667	2,5000
Modalna vrijednost	2,67a	2,14a
Standardna devijacija	,62581	,68022
Varijanca	,392	,463
Asimetrija (Skewness)	-,240	-,110
Standardna pogreška asimetrije	,194	,194
Spljoštenost (Kurtosis)	,013	-,482
Standardna pogreška spljoštenosti	,386	,386
Minimum	1,00	1,00
Maksimum	4,00	4,00
Pouzdanost (Cronbach Alpha koeficijent)	0,885	0,908
Testiranje normalnosti distribucije		
Kolmogorov-Smirnov Z	,830	,724
Dvosmjerna statistička značajnost	,497	,670

Korelaciju smo računali primjenom Spearmanovog koeficijenta korelacije za ordinalne varijable. Na razini rizika od 5% korelacije su statistički značajne i indeksi autonomne i kontrolne regulacije iznosi 0,448 , što znači povećanjem kontrolirane regulacije povećava se i autonomna regulacija (Tablica II.).

TABLICA II. POKAZATELJI KORELACIJE ZA KONTROLIRANE I AUTONOMNE REGULACIJE

Korelacija				
			kontrolirana	***autonomna***
Spearmanov koeficijent korelacije	***kontrolirana***	Koeficijent korelacije	1,000	,448**
		Dvosmjerna signifikantnost	.	,000
		N	156	156
	autonomna	Koeficijent korelacije	,448**	1,000
		Dvosmjerna signifikantnost	,000	.
		N	156	156

** *Korelacija je statistički značajna uz razinu rizika od 1% (dvosmjerno testiranje).*

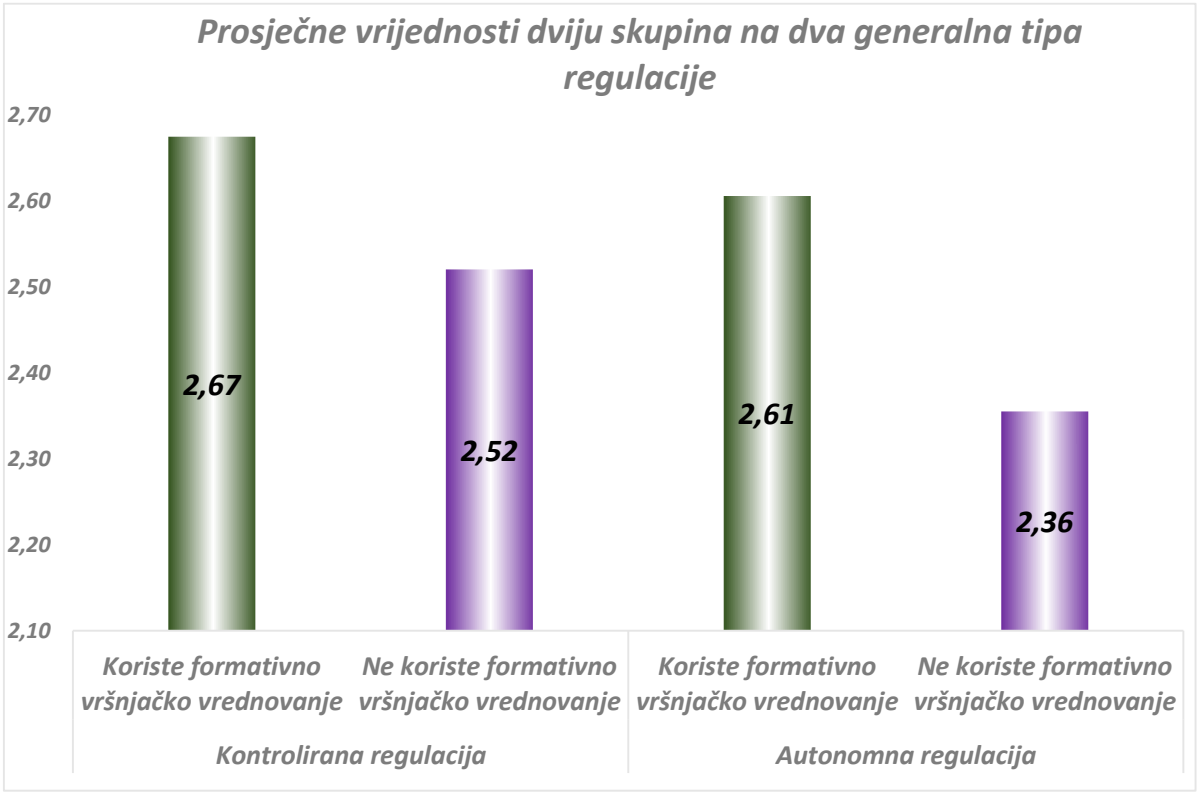
Testiranjem normalnosti distribucija uz razinu rizika od 5% utvrđeno kako su obje skale, skala kontrolirane regulacije (Kolmogorov-Smirnov Z = 0,830; p > 0,05) i skala autonomne regulacije (Kolmogorov-Smirnov Z = 0,724; p > 0,05) normalno distribuirane, odnosno, ne odstupaju statistički značajno od normalne distribucije. Testiranje statističke značajnosti razlika između učenika koji u radu nastave matematike koriste formativno vršnjačko vrednovanje te onih koji u radu nastave matematike ne koriste formativno vršnjačko vrednovanje provedeno je parametrijskim testom (T-testom za dva nezavisna uzorka). Rezultati testiranja prikazani su u tablici III. Na skali kontrolirane regulacije nije utvrđena statistički značajna razlika prosječnih rezultata učenika koji u radu u nastavi matematike koriste formativno vršnjačko vrednovanje te onih koji u radu u nastavi matematike ne koriste formativno vršnjačko vrednovanje, ali na skali autonomne regulacije, uz razinu rizika od 5%, prosječna vrijednost učenika koji u radu u nastavi matematike koriste formativno vršnjačko vrednovanje statistički se značajno razlikuje od prosječne vrijednosti učenika koji u radu u nastavi matematike ne koriste formativno vršnjačko vrednovanje (homogene varijance, T omjer = 2,329; p < 0,05). Pri tome, učenici koji u radu u nastavi matematike koriste formativno vršnjačko vrednovanje, međusobno vrednuju rješenja zadataka na virtualnoj ploči, u većoj su mjeri autonomno regulirani, dok su učenici koji u radu u nastavi matematike ne koriste formativno vršnjačko vrednovanje u manjoj mjeri autonomno regulirani i postižu nižu prosječnu vrijednost na indeksu autonomne regulacije (Slika 3., Tablica IV.).

TABLICA III. TESTRANJE STATISTIČKE ZNAČAJNOSTI RAZLIKE PROSJEČNIH REZULTATA UČENIKA KOJI U NASTAVI MATEMATIKE KORISTE FORMATIVNO VREDNOVANJE UZ RAD NA VIRTUALNOJ PLOČI

		Leveneov test homogenosti varijanci		**T-test statističke značajnosti razlike prosječnih vrijednosti**						
		F omjer	Signifikantnost	T omjer	Stupnjevi slobode	Signifikantnost (dvosmjerna)	Razlika prosječnih vrijednosti	Standardna pogreška razlike	95% interval pouzdanosti razlike prosječnih vrijednosti	
									Donja granica	Gornja granica
Kontrolirana regulacija	Homogene varijance	2,696	,103	1,542	154	,125	,15415	,09997	-,04335	,35165
	Nehomogene varijance			1,561	153,433	,121	,15415	,09877	-,04097	,34927
Autonomna regulacija	Homogene varijance	,514	,475	2,329	154	,021	,25067	,10762	,03806	,46327
	Nehomogene varijance			2,326	150,592	,021	,25067	,10777	,03773	,46360

TABLICA IV. PROSJEČNE VRIJEDNOSTI DVIJU SKUPINA ISPITANIKA NA DVA GENERALNA TIPA REGULACIJE

Grupirajuća varijabla: Formativno vršnjačko vrednovanje – „Međusobno vrednujemo rješenja zadataka na virtualnoj ploči“					
Dva generalna tipa regulacije	Međusobno vrednujemo rješenja zadataka na virtualnoj ploči	N	Prosječna vrijednost	Standardna devijacija	Standardna pogreška aritmetičke sredine
Kontrolirana regulacija	Da	83	2,6747	,67503	,07409
	Ne	73	2,5205	,55798	,06531
Autonomna regulacija	Da	83	2,6059	,66403	,07289
	Ne	73	2,3552	,67826	,07938



Slika 3. Prikaz prosječne vrijednosti dviju skupina na dva generalna tipa regulacije

III. ZAKLJUČAK

Pokazalo se ovim istraživanjem kako se IKT može učinkovito koristiti sa pedagoškim metodama koje uključuju vršnjačko vrednovanje te da kao takvo povećava samoregulaciju učenika. Učenici koji u radu u nastavi matematike koriste formativno vršnjačko vrednovanje, odnosno međusobno vrednuju rješenja zadataka na virtualnoj ploči u većoj su mjeri autonomno regulirani, imaju veću intrizičnu i identificiranu regulaciju u odnosu na učenike koji ne koriste virtualnu ploču s metodom vršnjačkog vrednovanja. Preporuka za buduća istraživanja je ponovit ovo istraživanje na većem uzorku ispitanika te u istraživanje uključiti profesore s približno istim radnim iskustvom, sposobnostima i otvorenosti novim metodama poučavanja.

LITERATURA

[1] A. Agélii Genlott and Å. Grönlund, "Improving literacy skills through learning reading by writing: the iWTR method presented and tested," Computers &Education,vol.67, pp.8-10, 2013.

[2] A. Agélii Genlott and Å. Grönlund, "Closing the gaps : Improving literacy and mathematics by ict-enhanced collaboration", Computers and Education,vol.99, pp. 68–80, 2016.

[3] T. Anderson,"Getting the mix right again: An updated and theoretical rationale for interaction,"International Review of Research on Open and Distance Learning, vol. 4, no. 2, pp.1-14, 2003.

[4] P. Black, D. Wiliam, "Assessment and classroom learning," Assessment in education, vol.5, no.1, pp.7-74, 1998.

[5] M. Boekaerts and E. Cascallar, "How far have we moved toward the integration of theory and practice in self-regulation?" Educational Psychology Review, vol. 18, no. 3, pp. 199-210, 2006.

[6] T. Goldin "Povezanost akademske samoregulacije, učeničke percepcije roditelja i školskog uspjeha." Sveučilište u Zagrebu, Filozofski fakultet, Odsjek za psihologiju, Diplomski rad, Zagreb, 2007.

[7] S. G. Paris and A.H. Paris, "Classroom applications of research on self-regulated learning," Educational psychologist, vol.36, no.2, pp. 89–101, 2001.

[8] R. M. Ryan and J.P. Connell," Perceived locus of causality and internalization: Examining reasons for acting in two domains." Journal of Personality and Social Psychology, 57, pp. 749-761, 1989.

[9] R. Slavin, "Cooperative learning theory, research, and practice," Boston: Allyn and Bacon, 1995.

[10] B. J. Zimmerman, "Attaining self-regulation: a social cognitive perspective," In: Boekaerts, M.; Pintrich, P.; Zeidner, M. (Orgs.), Handbook of self-regulation, San Diego: Academic Press, pp.13-39, 2000.

[11] R. M. Ryan and E. L. Deci, " Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being," American Psychologist*, 55*, pp.68-78, 2000.

[12] V. Vidović, V. Vlahović-Štetić, M. Rijavec, D. Miljković, "Psihologija obrazovanja," Zagreb: IEP-VERN', 2003.

[13] M. Palekčić i F. Müller, "Uvjeti i efekti interesa za studij i motivacije za učenje (motivi za izbor studija i motivacijski regulacijski stilovi) kod hrvatskih i njemačkih studenata", Pedagogijska istraživanja, vol.1, br. 2, pp. 159-192, 2004.

[14] T.Veliki, "The correlation considering the degree of autonomus motivation, academic achievement and mental health." Croatian Journal of Education, vol.13, no.3, pp. 56-87, 2011.

Primjena alata Xmpl za brzo pokretanje naredbi u podučavanju Linuxa

I. Krpan, I. Capan i M. Davidović
Visoka škola za informacijske tehnologije, Zagreb, Hrvatska
ivan.krpan@vsite.hr, ivan.capan@vsite.hr, milan.davidovic@vsite.hr

***Sažetak* - Podučavanje alata otvorenog koda i Linux operacijskog sustava prate razni problemi. Rad u Linux ljusci bazira se na unošenju naredbi s više parametara i opcija. Njihova sintaksa je stroga i ne tolerira pogreške, a od korisnika se očekuje da te opcije znaju napamet. Iako postoji dokumentacija za svaku naredbu, to nisu kratke upute već treba potrošiti neko vrijeme za stjecanje početnih znanja. Upravo iz tih razloga pojavila se potreba da se razvije alat za nastavu. Njime bi se po upitu pronašla potrebna naredba i kratko objašnjenje s osnovnim opcijama i nužnim parametrima za pomoć u sistemskom programiranju. Tako je nastao slobodni edukacijski alat nazvan Xmpl koji kroz primjere naredbi i njihovo pronalaženje u bazi provjerenih primjera korisnicima rješava probleme različitog stupnja složenosti.**

U radu su opisane funkcionalnosti tog alata i način kako se instalira i koristi, te o kojim alatima i platformi ovisi. Web izvedba omogućuje interaktivnu nastavu i cjeloživotno učenje.

Ključne riječi - Linux, nastava, Linux ljuska, naredbe

I. UVOD

Linux i tehnologije otvorenog koda sve su prisutnije u poslovnom i privatnom svijetu. Tvrtke takve sustave sve više koriste zbog njihove pouzdanosti, stabilnosti, niske cijene i sve većih mogućnosti. Linux je sve popularniji i kod privatnih korisnika jer se povećava baza aplikacija koje se mogu koristiti i u tipičnoj kućnoj primjeni (na primjer, multimedija i igre). Poslodavci slijedom toga sve više traže znanje Linuxa, pa se studenti i osobe koje traže posao prilagođavaju tome i žele dobiti ta specifična znanja i vještine. Organiziraju se tečajevi, predaje se kao poseban predmet u studijskim programima i koriste se drugi načini prijenosa znanja Linuxa.

Za operacijski sustav Linux mnogima je prva asocijacija "težak za korištenje". Budući da je nastao iz operacijskih sustava UNIX, koji su se pojavili ranih 70-tih godina kad nisu postojala grafička sučelja, računalni miš i slična pomagala, Linux je preuzeo njihovu praksu. Osnovna postavka UNIX-a i njemu pripadajućih alata je: “radi jednu stvar i radi ju dobro”. Tako dobivamo desetine i stotine alata koji pojedinačno rade samo jednu stvar. Primjerice, za kopiranje datoteka na sustavu koristi se naredba *cp*, a za njihovo premještanje naredba *mv*. U slučajevima kad se pokuša naredba reciklirati, dodaju se nove opcije. Za brisanje datoteka koristi se naredba *rm*, ali kad se briše direktorij, mora se dodati opcija r što znači rekurzivno, inače javlja nedovoljno intuitivnu grešku. Čak i danas, iako je tehnologija jako napredovala, Linux se i dalje primarno koristi bez grafičkog sučelja jer je daleko češći na poslužiteljima kojima se pristupa udaljeno. Osnovni način komunikacije korisnika i Linuxa je redak za upisivanje tekstualnih naredbi koji se naziva ljuska (engl. *shell*). Najčešća se ljuska, odnosno softver koji ju pokreće, naziva Bash. Korisnik, kad se spoji na sustav, mora utipkavati razne naredbe da bi postigao čak i neku osnovnu funkcionalnost.

Krivulja učenja Linuxa je vrlo strma. Svaki novi zadatak je novi izazov upoznavanja nove naredbe sa svim njenim opcijama. Prethodno naučena naredba odnosno stečeno znanje ne pomaže kod idućih naredbi. Taj način korištenja zahtijeva strukturirani pristup stjecanju novih znanja, što je poseban izazov za predavače.

Pregledom raznih komercijalnih tečajeva na internetu može se ustvrditi da se uglavnom Linux podučava na način da se jednostavno daje popis naredbi koje treba memorirati. Interaktivnost je svedena na minimum, a nakon odrađenog primjera ide se dalje. Takva forma podučavanja nije optimalna za specifične situacije s brojnim naredbama u Linux sustavu. Iako za svaku naredbu postoje upute na samom sustavu (tzv. *man* stranice) [1] to je zapravo dokumentacija o pojedinoj naredbi, a ne brza pomoć za izbor najbolje opcije. Te upute, dakle, ne rješavaju problem popisa naredbi koje i dalje treba učiti napamet. Popis bitnih naredbi s kratkim opisom i samo s bitnim opcijama koji je stalno dostupan na sustavu rješenje je prethodno opisanih problema. Taj popis može biti u formi lokalne tekstualne datoteke, ali to ne daje nikakve napredne mogućnosti ili prednosti pred podsjetnikom sa istim sadržajem koji korisnik drži sa strane.

Alat Xmpl, čije ime je izvedeno iz engleske riječi *example*, nastao je i dalje se razvija kroz proces nastave kao program koji početnicima pruža pristup primjerima naredbi i olakšava njihovo izravno korištenje. Iskusnijim korisnicima pak omogućuje razvoj i testiranje prototipova programa i kreiranje vlastitih primjera naredbi kojih nema u bazi provjerenih primjera. Alat je detaljno opisan u završnom radu njegovog autora [2]. Kao podrška daljnjem razvoju alata i njegovom kontinuiranom poboljšavanju, nastala je i zajednica istog naziva [3]. Xmpl je još jedan primjer korištenja GitHuba kao platforme za razvoj kreativnosti i suradnje u podučavanju [4].

Xmpl se može i planira koristiti u podučavanju Linuxa, a njegova Web inačica i za nastavu na daljinu i

cjeloživotno učenje kroz preuzimanje najbolje prakse i proaktivnu suradnju u rješavanju i najsloženijih problema.

II. XMPL ALAT

Xmpl alat je izveden kao Bash skripta koja se može koristiti bez potrebe za instalacijom ili prevođenjem (engl. *compiling*). Koristi GitHub API za dohvaćanje primjera Linux programa iz repozitorija na GitHubu.

Primarne funkcionalnosti Xmpl alata su pretraživanje svih dostupnih primjera u određenom repozitoriju i izvršavanje dohvaćenog programa na sustavu kroz Xmpl.

Za pretraživanje GitHub repozitorija koristi se *code search* API [5] i određeni filteri za dohvaćanja primjera iz repozitorija (na primjer, ekstenzija datoteka, lokacija unutar repozitorija ili naziv datoteke).

Xmpl repozitorij je baza primjera korištenja Unix alata za čuvanje i pretraživanje u GitHub repozitoriju. Strukturu baze čine ispravno organizirane *.repo , *.desc i *.xmpl datoteke. Svaka od njih opisuje i sadrži neke entitete baze, poput imenovanja paketa u bazi primjera, njihovih opisa ili pojmova za pretragu.

III. KORIŠTENJE XMPL ALATA

A. *Pretraživanje Xmpl repozitorija*

Za pretragu po repozitoriju koriste se pojmovi kojima se s Xmpl alatom dohvaća popis primjera koji odgovaraju proslijeđenom upitu, kao na slici 1.

```
user@system:~/xmpl-script$ ./xmpl.sh download file
Online mode!
1 curl Download a file from URL and save it to local destination
2 curl Download a file using its original filename, follow all redirects, and
previously failed
3 wget Download a file from a webserver and save to hard drive
4 wget Download file into specific directory
Please select example number:
4
wget: Download file into specific directory
wget -P {:destination directory:} {:remote file:}
user@system:~/xmpl-script$
```

Slika 1. Pretraživanje Xmpl repozitorija

B. *Izvršavanje primjera*

Nakon pretrage i dohvaćanja, taj se primjer može i izvršiti direktno iz alata. U primjeru na slici 2 se pretražuje postoji li naredba koja sadrži ključne riječi show my ip koja se potom i izvršava. Naredba se, kao što je vidljivo, zove ip addr show.

C. *Priprema s korisničkim unosom*

Naredba za prikaz adrese iz prethodnog primjera nije zahtijevala nikakve parametre. U slučaju da naredba traži

```
user@system:~/xmpl-script$ ./xmpl.sh show my ip -x
Online mode!
ip: Display all network interfaces and ip addresses
ip addr show
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group d
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: enp5s0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc pfifo_fast st
    link/ether 90:2b:34:5a:d0:91 brd ff:ff:ff:ff:ff:ff
3: wlp4s0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq state UP grou
    link/ether 00:13:46:93:9e:9b brd ff:ff:ff:ff:ff:ff
    inet 192.168.2.25/24 brd 192.168.2.255 scope global dynamic wlp4s0
       valid_lft 259083sec preferred_lft 259083sec
    inet6 fe80::1c0b:c35e:62f2:9b2d/64 scope link
       valid_lft forever preferred_lft forever
EXECUTED
user@system:~/xmpl-script$
```

Slika 2. Izvršavanje primjera

```
user@system:~/xmpl-script$ . ./xmpl.sh ping -i
Online mode!
1 nmap Ping scans the network, listing machines that respon
2 ping Flood ping
3 ping Test the reachability of a host on network
Please select example number:
3
ping: Test the reachability of a host on network
ping {:remote host:} -c 4
remote host:
www.github.com
ping www.github.com -c 4
user@system:~/xmpl-script$
```

Slika 3.Izvršavanje cjelovite naredbe

neki parametar, možemo i njega navesti interaktivno u sučelju. Konačni je rezultat pokretanje cjelovite naredbe prikazan na slici 3. U primjeru se može uočiti i višestruki rezultat pretraživanja te odabir rednog broja naredbe koja nam treba.

U sljedećem primjeru (slika 4) koristi se prethodno pripremljena naredba. Posljednje unesene parametre moguće je dohvatiti koristeći se strelicama gore/dolje. Parametri su uneseni unutar Xmpl alata i naredba se izravno izvodi bez dodatnih promjena.

```
user@system:~/xmpl-script$ . ./xmpl.sh -X
ping: Test the reachability of a host on network
ping {:remote host:} -c 4
remote host:
www.github.com
ping www.github.com -c 4
PING github.com (192.30.253.113) 56(84) bytes of data.
64 bytes from lb-192-30-253-113-iad.github.com (192.30.253.113):
64 bytes from lb-192-30-253-113-iad.github.com (192.30.253.113):
64 bytes from lb-192-30-253-113-iad.github.com (192.30.253.113):
64 bytes from lb-192-30-253-113-iad.github.com (192.30.253.113):

--- github.com ping statistics ---
4 packets transmitted, 4 received, 0% packet loss, time 3004ms
rtt min/avg/max/mdev = 122.222/124.711/129.554/2.984 ms
EXECUTED
user@system:~/xmpl-script$
```

Slika 4. Izvršavanje pripremljene naredbe

U posljednjem promjeru (slika 5) alat pozivamo tako da zadamo zadnje korištenu naredbu, ali joj samo mijenjamo prvi parametar. Njega zadajemo kod pozivanja alata Xmpl i tako se primjer odmah izvršava, a korisnika se naknadno ne pita unos varijable. Cjelovita dokumentacija može se naći na [6].

```
user@system:~/xmpl-script$ . ./xmpl.sh -l -x 8.8.8.8
ping: Test the reachability of a host on network
ping {:remote host:} -c 4
remote host:
8.8.8.8
ping 8.8.8.8 -c 4
PING 8.8.8.8 (8.8.8.8) 56(84) bytes of data.
64 bytes from 8.8.8.8: icmp_seq=1 ttl=123 time=22.6 ms
64 bytes from 8.8.8.8: icmp_seq=2 ttl=123 time=22.6 ms
64 bytes from 8.8.8.8: icmp_seq=3 ttl=123 time=23.3 ms
64 bytes from 8.8.8.8: icmp_seq=4 ttl=123 time=22.3 ms

--- 8.8.8.8 ping statistics ---
4 packets transmitted, 4 received, 0% packet loss, time 3005ms
rtt min/avg/max/mdev = 22.309/22.755/23.343/0.415 ms
EXECUTED
user@system:~/xmpl-script$
```

Slika 5. Prosljeđivanje parametara prilikom prosljeđivanja

IV. INSTALACIJA XMPL ALATA

Alat Xmpl instalira se povlačenjem sa službenog repozitorija na GitHubu korištenjem jasno dokumentiranih



Slika 6. Instaliranje Xmpl alata i dodavanje repozitorija

procedura (slika 6). S alatom se instalira skripta i njeno zamjensko ime (alias) s kojim se skripta poziva. Tad se postavljaju pravila za prosljeđivanje određenih Xmpl varijabli kroz *sudo* naredbu, kako bi se povijest unosa mogla koristiti kroz *sudo* alat. Osim osnovne instalacije, alat nudi i mogućnost instalacije lokalnih repozitorija na računalu koji omogućuju način rada bez internet povezivosti, te privatnih repozitorija na GitHubu preko kojih je moguće uređivati i dodavati nove primjere u vlastiti ili službeni GitHub Xmpl repozitorij. Preporučljivo je, ali ne i nužno, imati GitHub korisnički račun kako bi se mogli dodavati novi primjeri ili poboljšavati postojeći.

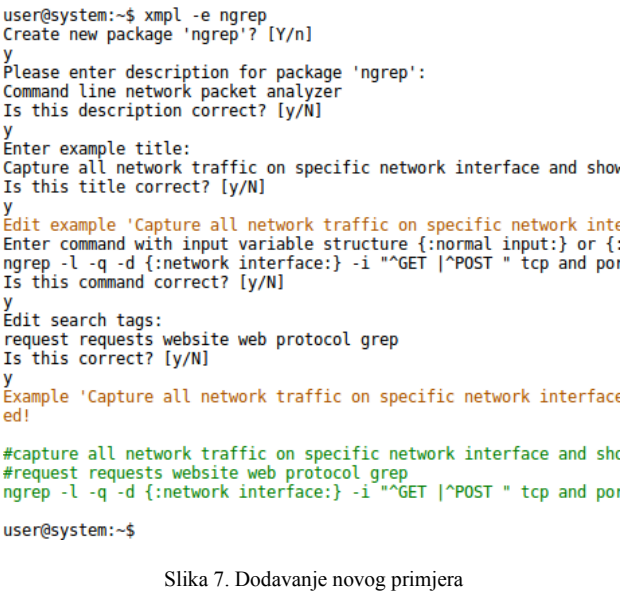
Nakon završene instalacije, Xmpl alat je potpuno konfiguriran za korištenje i dostupan s bilo koje lokacije na sustavu pod nazivom xmpl. S instaliranom verzijom Xmpl alata mogu se koristiti dodatne funkcionalnosti poput manipulacije lokalnim repozitorijima i uređivanje ili dodavanje novih primjera.

Popis svih lokalno instaliranih repozitorija prikazuje se sa parametrom -R. Tako se može vidjeti da postoje *main* i *dev* repozitoriji. S istom naredbom moguće je promijeniti trenutni repozitorij iz *main* u *dev*. Cijela naredba glasi: *xmpl -R dev*. Po ovoj naredbi Xmpl alat koristi razvojni repozitorij koji je moguće i uređivati korištenjem Xmpl-a.

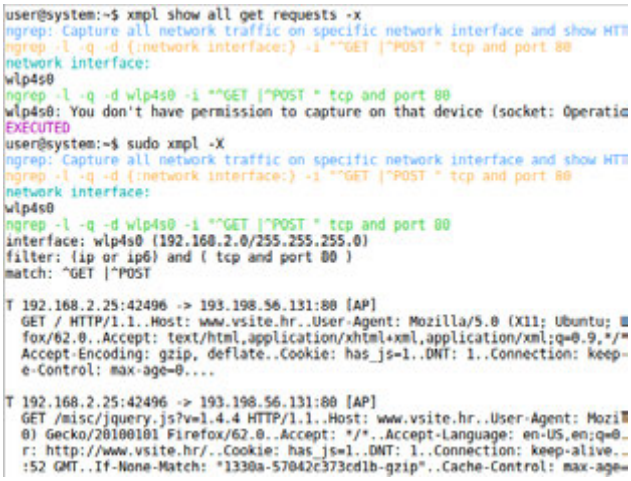
V. UREĐIVANJE PRIMJERA

Za demonstraciju unosa primjera u repozitorij unosi se novi paket. Dodat ćemo primjer izvođenja programa *ngrep* (nadzor mrežnog prometa). To je softver za prikaz svih POST i GET zahtjeva na HTTP protokolu. S naredbom *xmpl -e ngrep* započinje postupak dodavanja primjera (Slika 7). Nakon toga korisnika interaktivno sučelje traži dodatna pitanja: opis, ključne riječi za lakše pretraživanje, varijable koje treba ubaciti i slično.

Na primjeru korištenja (Slika 8) može se vidjeti kako se pokreće pretraživanje s općenitim ključnim riječima i testiranje novo dodanog primjera.



Slika 7. Dodavanje novog primjera



Slika 8. Testiranje novog primjera

Kako bi naš primjer mogli podijeliti sa zajednicom, moramo ga poslati na centralni repozitorij. Tako ćemo Xmpl alatom pokrenuti sinkronizaciju privatnog lokalnog repozitorija s vlastitim na GitHubu koji je kreiran kod instalacije Xmpl alata. Naredba za to je *xmpl -S*. Na udaljenom privatnom repozitoriju će se pojaviti nove .desc i .xmpl datoteke, kao što je već opisano ranije. Sljedeći je korak slanje zahtjeva za uključivanjem tog primjera u službeni repozitorij na GitHubu.

Glavni, odnosno službeni repozitorij dostupan je svim korisnicima kod instalacije i ažuriranja Xmpl alata. Bitno je da se tamo nalaze samo provjereni i relevantni primjeri. Alat je zamišljen kao kolaboracijski projekt i postoje autori sustava i zainteresirani administratori i nastavnici koji nadziru taj repozitorij. Netko će zahtjev za dodavanjem novog primjera zaprimiti, testirati, verificirati i propustiti u glavni repozitorij. Time se garantira da novi primjer nema nepotrebnih opcija, da je opis dobar i ključne riječi precizne te da korisnici ne ubacuju gluposti na popis. Nakon svih tih provjera će i taj primjer biti dostupan u Xmpl alatu svih ostalih korisnika, odnosno njegovom lokalnom repozitoriju nakon sinkronizacije. Sve ovo je nužno zbog kontrole kvalitete izvornog programa i

njegove relevantnosti. GitHub također posjeduje i web sučelje te administrator može provjeravati novi primjer, bez Xmpl alata, pa čak i s mobilnog uređaja.

Dodavanje novog primjera od strane studenta bitno je u procesu učenja jer tako studenti prolaze kroz proces tzv. „igrifikacije“ [7] (engl. *gamification*). To je proces u kojem se neka radnja, zadatak ili rješenje problema nagrađuju ne materijalnim već nekim virtualnim bodovanjem, podizanjem statusa ili sličnim elementima preuzetim iz računalnih igara. U našem slučaju, dodavanje nekog primjera može biti popraćeno javnom pohvalom nakon što se primjer prihvati, objavom neke top-liste najboljih predlagatelja primjera ili davanjem virtualnih „znački“ (engl. *badges*). Ti studenti imat će na svom GitHub računu i pismeni trag suradnje s Xmpl alatom što bi moglo biti važno kod budućeg traženja posla. U raznim se oglasima za posao sve češće traži i adresa GitHub računa jer se tamo mogu vidjeti kandidatovi interesi, razina znanja i realizirani projekti. Naposlijetku, istraživanje [8] tvrdi da su studenti koji su sudjelovali u igrificiranim sustavima za učenje imali 14% bolje rezultate na testovima vještina, postigli 11% bolji rezultat na teorijskim testovima te imali 9% bolje usvajanje znanja, odnosno dugoročno pamćenje.

VI. Planovi za razvoj

Iako su već razvijene osnovne funkcionalnosti alata Xmpl, moguća su daljnja poboljšanja i razvoj novih funkcionalnosti. Za korištenje u nastavi treba uvesti mogućnost ocjenjivanja te provjere znanja i ishoda učenja. Studenti bi izravno preko sučelja dobivali pitanja iz fonda postojećih primjera, a odgovarali bi s imenom naredbe, točnom sintaksom ili pak odabirom jednog od točnih odgovora. Nastavnik bi imao uvid u sve odgovore u stvarnom vremenu, preko GitHub sučelja ili nekog sličnog mehanizma za računanje točnih odgovora odnosno bodova. Osim praćenja ishoda učenja, potrebno je uključiti i istraživanja zadovoljstva studenata i nastavnika alatom Xmpl. Na temelju njihovih ocjena i prijedloga poboljšanja analiziralo bi se i ocijenilo u kojoj mjeri su studenti brže i bolje svladali Linux korištenjem tog alata.

Sadašnju verziju Xmpl-a je manji broj studenata ocijenio zanimljivim i korisnim za rad s Linuxom. Na besplatnim otvorenim radionicama polaznici će isprobati Xmpl i aktivno se uključiti u testiranje i razvoj alata te njegovog repozitorija primjera. Korisnost Xmpl-a u podučavanju Linuxa ocijenila bi se anketiranjem studenata i nastavnika nakon održanih radionica.

VII. Zaključak

U ovom radu opisan je softver Xmpl i njegova primjena u podučavanju studenata i raznih korisnika koji žele koristiti Linux i tehnologije otvorenog koda. Naglašeni su izazovi koje korisnici susreću pri učenju takvog gradiva i kako ih mogu premostiti uz ovaj alat. Cjeloživotno učenje je još jedna od primjena alata kojim se trebaju poslužiti i korisnici koji nisu dugo koristili znanje stečeno u akademskoj ustanovi.

Xmpl je alat koji je jednostavno postaviti i koristiti, a pokazano je da se i jednostavno može proširiti baza znanja i studenta i alata. Korištenje igrifikacije rezultira bržim stjecanjem vještina u sistemskom programiranju, boljim pamćenjem Linux naredbi i njihovih opcija te podupire rast posebno talentiranih studenata.

Cijeli ekosustav ovog alata se u radu oslanja na Github portal kao dokazano koristan kolaboracijski alat.

Za afirmaciju alata Xmpl u podučavanju Linuxa neophodno je provesti anketiranje nastavnika Linuxa i studenata i na temelju toga ocijeniti koliko primjena Xmpl-a pomaže u nastavi Linuxa.

Literatura

[1] Man (engl. *manual*) stranice, https://linux.die.net/man/, (pristupljeno 28. Svibnja 2020.)

[2] Ivan Krpan: „Projekt otvorenog koda Xmpl“, završni rad, Visoka škola za informacijske tehnologije, 2019.

[3] Xmpl project - https://xmpl-tool.github.io/ (pristupljeno 20. svibnja 2020.)

[4] K. Blaženka, Razvoj kreativnosti i suradnje upotrebom GitHub platforme u srednjoškolskom strukovnom obrazovanju, 42nd International Convention MIPRO 2019, Mipro Proceedings, Computers in Education, Opatija, 2019

[5] GitHub API documentation, https://developer.github.com/ /v3/search/#search-code (pristupljeno 20. svibnja 2020.)

[6] Xmpl tool documentation, https://xmpl-tool.github.io/usage.html (pristupljeno 20. svibnja 2020.)

[7] S.Lovrenčić et.al., Igrifikacija: prema sistematizaciji termina na hrvatskom jeziku, Računalne igre 2018., Zbornik radova, Varaždin, 2018.

[8] V.Bušelić, K.Župan, Metode gamifikacije u online sustavima učenja programiranja – osobno iskustvo, Polytechnic and Design, Vol. 6, No. 4, 2018

Održivi razvoj ruralnog područja – Croduino setovi i dronovi

Jelena Domac. mag.prim.educ., OŠ Stjepana Antolovića, Privlaka
jelena.domac@hotmail.com

Sažetak- **Održivi razvoj ruralnog područja – Croduino setovi i dronovi – stručni rad – koji je nastao kao rezultat uključivanja u projekt LOKO LAG[1].**

Razvoj ruralnih područja danas je od velike važnosti pa su tako i projekti jedan od segmenata njihova razvoja. Održivi razvoj ruralnog područja je projekt u kojem je obuhvaćeno 14 škola i 140 učenika s područja LAG – a Bosutski niz (mjesta kojima protječe rijeka Bosut). Ovim projektom želio se potaknuti razvoj tehnološkog dijela informatike kod djece u osnovnim školama te buđenje interesa za tehnološki napredak ruralnih područja. U projektu je osigurano 14 dronova (po jedan za svaku školu) i 70 Croduino (Arduino) setova (po 5 za svaku školu) te je po planu projekta predviđena edukacija 140 djece (po 10 iz svake škole) za rad s dronovima i Croduino (Arduino) setovima.

Cilj projekta bio je upoznati i educirati učenike ruralnih područja s najnovijim tehnologijama te time potaknuti njihov razvoj i potencijal te usmjeriti ih na programiranje, razvoj njihovih digitalnih kompetencija u daljnjem školovanju i povećanju konkurentnosti pri odabiru zanimanja i zapošljavanju. Po završetku projekta pokazalo se kako su djeca zainteresirana za daljnji rad s Croduino (Arduino) setovima i rukovanje dronovima te kako vide da im je tehnologija saveznik u njihovom daljnjem razvoju i razvitku.

Ovaj projekt LAG-a sufinanciran je iz Europskih strukturnih i investicijskih fondova za učinkovitost i ljudski potencijal, zajedno s 14 općina koje pripadaju Bosutskom nizu i Hrvatskim zavodom za zapošljavanje.

Ključne riječi: održivi razvoj, ruralna područja, Croduinu (Arduino) setovi, dronovi, digitalne kompetencije, tehnološki napredak

I. UVOD

Kako su na području Republike Hrvatske u kategoriji nerazvijenih područja uglavnom ruralni dijelovi kojima pripada 52% stanovništva i 73% površine [1], važno je krenuti u njihovo pravovremeno razvijanje uz pomoć različitih strategija u kojima bi se zalagalo za iskorištavanje onoga čime je to područje bogato, sačuvati njegovu tradiciju i njegovalo kraj ali i nasuprot tome potrebno ga je dovesti na razinu na kojoj može konkurirati pomodnim zahtjevima današnjice te ga modernizirati, ali ne zanemariti ono što taj kraj čini posebnim. Vodeći se različitim strategijama razvoja nastala je i inicijativa o osnivanju LAG-a, koja je došla na poticaj Regionalne razvojne agencije Vukovarsko-srijemske županije (HRAST), u stvari je rezultat interesa, okupljanja i suradnje šireg kruga dionika, predstavnika javnog, privatnog i civilnog sektora iz jedinica lokalne samouprave s područja LAG-a Bosutski niz [2]. LAG u svom sastavu ima 11 jedinica lokalne samouprave, a to su Općine: Andrijaševci, Babina Greda, Cerna, Gradište, Jarmina, Ivankovo, Nijemci, Privlaka, Stari Mikanovci, Vođinci i grad Otok.

Tako je LAG žudeći za modernim strategijama u razvoju ruralnih područja Istočne Slavonije konkurirao za projekt LOKO LAG koji je sufinancirala Europska unija iz Europskog socijalnog fonda s 85% sredstava i Republika Hrvatska - Ministarstvo rada i mirovinskog sustava s 15% sredstava.

II. PROJEKT

Održivi razvoj ruralnog područja (LOKO LAG) nastao je kako bi se učenici u ruralnim područjima Istočne Slavonije, koja nisu zahvaćena različitim robotičkim klubovima i tečajevima programiranja, naučili osnovama rada sa dobivenim setovima, rukovanju s dronovima, upoznali se s njihovom širokom primjenom u različitim sferama društva te da bi se osvijestili kako je programiranje jedno među traženijim zanimanjima na tržištu rada.
U projekt je uključeno 14 osnovnih škola (Jarmina, Nijemci, Lipovac, Ivankovo, Retkovci, Babina Greda, Stari Mikanovci, Gradište, Cerna, Rokovci, Vođinci, Privlaka, Otok i Komletinci), njihovih 140 učenika i 14 učitelja. Projekt se provodio kao

[1]LOKO – kratica za lokalno, misli se na lokalne škole koje su okupljene u akcijskoj grupi
LAG – Lokalna akcijska grupa Bosutski niz

izvannastavna/izvanškolska aktivnost za učenike od 5. do 8. razreda tijekom školske godine 2019./20. u kojem su sudionici (škole) opremljeni s 5 Croduino (Arduino) setova i 1 dron.

A. Cilj projekta

U skladu sa zahtjevima života u 21. stoljeću važno je i neophodno na individualnoj razini razvijati informatičke vještine i informatičku pismenost te su vještine usko povezane uz sustav širih kompetencija (komunikacijske kompetencije, kompetencije za timski rad i suradnju, kreativno i kritičko mišljenje, rješavanje problema, mrežno povezivanje, kompetencije za profesionalno cjeloživotno učenje i sl.) [3]. Tako se kod učenika potaklo razvijanje informatičkih vještina i znanja te različitih digitalnih kompetencija, među njima se stavio naglasak na programiranje u C programskom jeziku te pravilno rukovanje s dronom.

B. Oprema

Za ovaj projekt je uzeta prva hrvatska inačica Arduina. Croduino set za početnike (CSP) je sustav za učenje programiranja i osnovnih elektroničkih principa. Croduino je verzija Arduina. Kako je Arduino open-source, derivacije, izmjene i drukčije verzije Arduina su dozvoljene –tako je kreiran Croduino; prva hrvatska Arduino kompatibilna pločica [4]. Mogućnosti seta je spajanje različitih elektroničkih komponenti i programiranje u C programskom jeziku. Treba napomenuti, da za rukovanje i uporabu ovim setom učenicima nije bilo potrebno nikakvo iskustvo u programiranju ili elektronici kako bi počeli s radom. Dron marke JJRC x7 koji dolazi s osnovnim uputama, upravljačem, utorom za SSD karticu te ima mogućnost povezivanja s pametnim telefonom preko vlastite aplikacije.

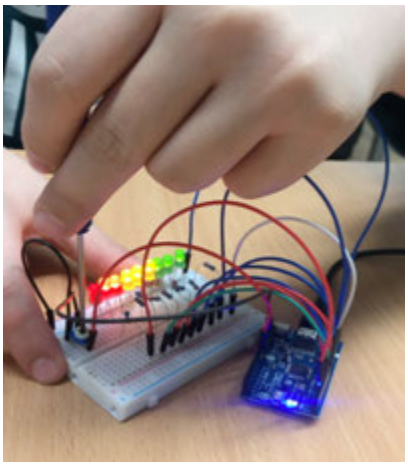
C. Edukacija

Učenici i njihovi učitelji mogli su se kroz 20 sati edukacije, koje im je držao vanjski suradnik, upoznati s osnovama rada Croduino seta te s osnovama programiranja C programskog jezika. Kako su naši učenici imali set za početnike on je sadržavao dovoljno sadržaja za uvođenje u svijet programiranja. Za pravilno rukovanje trebali su proučiti i upoznati se prvo sa svim elementima i dijelovima koje ovaj set nudi. S različitim pinovima koji se nalaze na pločici, mikrokontrolerom, USB konverterom, konektorom, reset tipkalom, različitim LED diodama, senzorima otpornicima… Nakon upoznavanja s opremom, učenici su na svoja računala trebali instalirati potreban softver za programiranje. Kroz cijelo trajanje projekta učenici su bili vođeni, ali su sve zadatke izvršavali sami jer se stavio naglasak na iskustveno učenje u kojem je težište na procesu i uključenosti učenika u sam proces (Slika 1.).



Slika 1. Rad učenika s Croduino setovima

Budući da se učenici do sada nisu sreli s programskim jezikom C i nekim „ozbiljnijim“ programiranjem trebalo ih je upoznati i s osnovnim pravilima pisanja koda. Važno je napomenuti kako su učenici također upoznati s najvažnijim strategijama programiranja i postupkom izrade programa za računalom u kojem prednjači metoda pokušaja i pogreški. U svom radu s Croduino setovima imali su različite projekte kao automatsko noćno svjetlo, semafor, serijska komunikacija, putujuće svijetlo (Slika 2.), senzori za temperaturu i vlagu.



Slika 2. Prikaz putujućeg svijetla – uz pomoć odvijača se regulira svjetlost

Što se tiče rada s dronovima, tu je učiteljima bilo prepušteno da se sami upoznaju s njegovim pravilnim rukovanjem. To je bio izazov za sve učitelje koji su sudjelovali u ovom projektu, no uz detaljno čitanje i praćenje uputa to nije predstavljalo problem. Učitelji su svoja znanja u rukovanju dronom prenijeli na svoje učenike i upoznali ih s osnovama korištenja, instaliranja aplikacije na pametni telefon, korištenjem upravljača, njegovim mogućnostima i kalibriranjem. Osnovno rukovanje dronom učenici su mogli naučiti kroz 10 školskih sati. Važno je napomenuti kako dron ima GPS pozicioniranje u stvarnom vremenu pa ga se tako može pratiti na krati gdje god se on nalazio. Isto tako jedna od važnijih značajki koja se učenicima posebno svidjela je *„Follow-up me“* mogućnost koja omogućava dronu praćenje nekog objekta. Učenici su također istaknuli kako je im bila jako važna opcija *„home“* koja vraća dron na početni položaj uzlijetanja. Potrebno je napomenuti kako su ovoga puta učenici iskustveno učili iako ih je rukovanje dronom u početku predstavljalo veći izazov nego programiranje. Učenici su cijelo vrijeme radili u parovima te su tako dodatno stekli različite kompetencije timskog rada i suradnje, razvijali su kritičko mišljenje te važne kompetencije za cjeloživotno učenje. Pri samom kraju ovog projekta učenicima je održala predavanje zaposlenica Hrvatskog zavoda za zapošljavanje te je intervjuirala učenike. U svom predavanju istaknula je mogućnost tehnološkog razvoja ruralnog područja upravo kroz različite edukacije i projekte u kojima se povećava konkurentnosti pri odabiru zanimanja i zapošljavanju.

D. Projekt unutar projekta

U svim školama, zadnju srijedu u veljači, obilježava se Međunarodni dan borbe protiv vršnjačkog nasilja - dan ružičastih majica. Ove je godine logo bio ispružena ruka koja pokazuje stop. Tako su učenici koji su sudjelovali u ovom projektu osmislili projekt za cijelu osnovnu školu Stjepana Antolovića, Privlaka u kojem su željeli drugima pokazati svoje znanje rukovanjem dronom. Kako za dan ružičastih majica svi učenici i učitelji obuku ružičastu majicu u znak borbe protiv vršnjačkog nasilja tako su to učinili i ove godine. Na školskom dvorištu nacrtali smo veliku raširenu ruku, koja pokazuje stop. Svi učenici i njihovi učitelji simbolično su stali u raširenu ruku i time pokazali jedno zajedništvo. Učenici koji su naučili rukovati dronom snimili su fotografiju te time ovjekovječili ovaj trenutak (Slika 3.). Također su učenici koji su sudjelovali u ovom projektu već počeli razmišljati o različitim upotrebama Croduino setova kroz različite projekte koji bi nastavili razvijati i time svoja stečena znanja produbili i dalje razvijali.



Slika 3. Snimka dronom – Dan ružičastih majica

III. ZAKLJUČAK

Pri završetku projekta pokazalo se kako su djeca zainteresirana za daljnji rad s Croduino (Arduino) setovima i rukovanje dronovima. Planira se i nastavak ovog projekta te proširenja s različitim mogućnostima koje današnja tehnologija sa sobom nosi. Učenici vide kako im je tehnologija saveznik u njihovom daljnjem razvoju i razvitku. Nadasve, njihove digitalne kompetencije, informatička znanja i vještine stečene ovim projektom mogu se kroz sljedeće projekte implementirati (što su učenici već i sami pokazali). Krajnji rezultat rukovanja tehnologijom ne treba biti program, kod ili nešto drugo. Nego tehnologija treba služiti čovjeku i pomoći mu u njegovom daljnjem razvitku. Važno je da učenici steknu tu osobinu i da oni budu ti koji će na kreativne i nove, nikada do sad viđene načine pronalazi rješenja i upravljati tehnologijom za bolji razvitak svoga područja.

LITERATURA

[1] Čavrak, V., „Održivi razvoj ruralnih područja Hrvatske“, Zbornik Ekonomskog fakulteta u Zagrebu, Godina 1, broj 1, Ekonomski fakultet Zagreb, Zagreb, str. 61-77, prosinac 2003.

[2] Službena internetska stranica LAG Bosutski niz https://lag-bosutskiniz.hr/ (pristupljeno 29.5.2020.)

[3] Vrkić Dimić, J. „Kompetencije učenika i nastavnika za 21. stoljeće“, Acta ladertina, 10 , str. 49-60, 2013.

[4] Službena internetska stranica Croduina https://e-radionica.com/hr/ (pristupljeno 29.5.2020.)

Stavovi studenata o računalnom načinu razmišljanja i vještine rješavanja testa računalnog načina razmišljanja

N. Tatković *, P. Radulović**, S. Tatković*
*Sveučilište Jurja Dobrile u Puli, Pula, Republika Hrvatska
**Ekonomska škola Pula, Pula, Republika Hrvatska
ntatkovic85@gmail.com, petko.radulovic@skole.hr, statkovic@gmail.com

***Sažetak* – Računalni način razmišljanja jedna je od temeljnih računalnih vještina, ali i važan aspekt pismenosti 21. stoljeća, potreban u svim područjima ljudskog života i rada. Cilj istraživanja ovog rada bio je utvrditi postoje li statistički značajne razlike u stavovima o računalnom načinu razmišljanja između studenata informatike i predškolskog odgoja. U istraživanju je korištena deskriptivna i kauzalno-neeksperimentalna metoda. Rezultati su pokazali da studenti informatike imaju statistički značajno pozitivniji stav o računalnom načinu razmišljanja od studenata predškolskog odgoja. S aspekta računalnih vještina, zanimala nas je uspješnost rješavanja testa računalnog razmišljanja. Dobiveni rezultati računalnih vještina nisu zadovoljavajući, a studenti predškolskog odgoja postigli nešto bolje rezultate. Istraživanje upućuje na važnost i potrebu razvoja računalnog načina razmišljanja tijekom cijelog procesa obrazovanja i potrebu daljnjeg istraživanja ovog problema.**

Ključne riječi: algoritmi, apstrakcija, dekompozicija, prepoznavanje uzoraka, programiranje

I. UVOD

Uz dotadašnje neosporne pismenosti: čitanja, računanja i pisanja, informatička i informacijska pismenosti predstavlja novu, izuzetno važnu pismenost 21. stoljeća koja zahtjeva dobro razvijene vještine računalnog načina razmišljanja. Računalni način razmišljanja predstavlja skup stavova i vještina potrebnih svakome, a ne samo informatičarima, za uspješno korištenje informatičko komunikacijske tehnologije, učenje i poučavanje. U svom članaku o računalnom načinu razmišljanja (Computational Thinking) autorica Wing daje osnovni okvir za razumijevanje toga pojma, ističući njegovu važnost kao jednog od temeljnih ciljeva učenja i poučavanja [16]. Danas računalni način razmišljanja nalazi svoje mjesto u obrazovnim kurikulumima država širom svijeta [17]. Uočeno je kako računala mogu unaprijediti obrazovni proces ako se promijeni način poučavanja i razvije računalni način razmišljanja [8]. Prepoznata je važnost i nužnost primjene toga načina razmišljanja u svim područjima ljudskog života i rada, od STEM područja, do društvenog, humanističkog i umjetničkog područja [17]. Zaostajanje u procesu učenja i poučavanja računalnog načina razmišljanja vodi države i društva u gospodarsku stagnaciju i „nepismenost“ 21. stoljeća. Primjena računalnog načina razmišljanja u tijesnoj je vezi i s modernim teorijama učenja, osobito konstruktivističkim učenjem [8,9]. Razmišljanja o uvođenju nastave programiranja u osnovne škole, kao i o razvoju algoritamskog razmišljanja kod svih učenika, postoje od ranije, ali nisu značajnije zaživjela u kurikulumima i nisu isto što i računalni način razmišljanja [18].

II. ŠTO JEST, A ŠTO NIJE RAČUNALNI NAČIN RAZMIŠLJANJA?

Računalno razmišljanje nije razmišljanje na način kako računalo „razmišlja“. Računalo radi odnosno „razmišlja“ na jedan izrazito jednostavan način gdje se koristi binarni brojevni sustav, a operacije se izvode korištenjem velikog broja jednostavnih logičkih sklopova (i, ili, ne). Računalo ne zna što računa, već samo izvodi veliki broj najjednostavnijih operacija, velikom brzinom. Na taj način radi i najmodernije računalo te ne postoji nitko tko može u potpunosti razumjeti složene operacije (kao skup najjednostavnijih) koje izvodi računalo. Očigledno je da to nije razmišljanje i da pojam računalnog razmišljanja ne podrazumijeva razumijevanje rada računala [13].

Računalno razmišljanje je razmišljanje poput računalnog znanstvenika, odnosno to su misaoni procesi koji omogućavaju razumijevanje i rješavanje problema na način na koji može djelovati i čovjek, ali i računalo [6]. Naglasak nije samo na rješavanju problema, nego i na njegovom uočavanju i formuliranju. Ljudi mogu koristiti računalni način razmišljanja i neovisno o stroju- računalu. Međutim, računalni način razmišljanja omogućit će, ukoliko želimo, pisanje programskog koda i rješavanje zadatka uz pomoć računala.

III. TEMELJNE TEHNIKE RAČUNALNOG NAČINA RAZMIŠLJANJA

Temeljne tehnike računalnog razmišljanja jesu: dekompozicija, prepoznavanje uzoraka, apstrakcija i algoritmi [2,4].

Dekompozicija je razbijanje kompleksnog problema na niz manjih. Manji problemi su lakše razumljivi i rješivi, a kasnije se mogu integrirati i dovesti do rješenja složenog problema.

Prepoznavanje uzorka slijedi nakon dekompozicije složenog problema. Naime, traže se sličnosti između problema i u tim sličnostima prepoznaju se zakonitosti. Te se zakonitosti mogu koristiti u drugim, sličnim zadacima [4].

Apstrakcija omogućava usredotočenost na bitne informacije, a odbacivanje nevažnih detalja. Prema Wing, apstrakcija je najznačajnija za računalno razmišljanje. Informatika je automatizacija apstrakcije, najviša razina misaonog procesa u računalnom razmišljanju te se i algoritam opisuje kao apstrakcija procesa ulaza, obrade i izlaza rezultata [17].

Algoritmi podrazumijevaju način rješavanja problema korak po korak, a na osnovi apstrakcije. Algoritam je rješenje koje vrijedi za puno sličnih problema koji se ponavljaju uz varijacije varijabli. Vještina razvijanja algoritama najbolje se uvježbava učenjem programiranja [6].

Programiranje razvija računalni način razmišljanja, posebno zato što razvija logičko razmišljanje i vještinu osmišljavanja algoritma kao koraka za rješavanje određenog problema na način koji „razumije“ računalo. Kako je računalo stroj koji izvodi naredbe, a te su naredbe napisane u nekom programskom jeziku, jasno je koliko je znanje programiranja značajno za punu i kreativnu uporabu IKT-a. Računalni način razmišljanja može se razvijati i bez znanja programiranja, ali je kroz učenje programiranja, usvajanje računalnog načina razmišljanja brže i jednostavnije. Stoga se, kao preteča računalnom načinu razmišljanja već u osnovne škole uvodi učenje programiranja, kroz programski jezik LOGO, a u novije vrijeme i Python. Računalni način razmišljanja treba početi razvijati od najranijeg razdoblja školovanja, prije sposobnosti za učenje programiranja. S razvojem računalnog načina razmišljanja treba krenuti što je moguće ranije, u ranom djetinjstvu, kako kasnije, učenici ne bi bili „bez sluha“ za računalni način razmišljanja i programiranje [10].

Dakle, računalno razmišljanje nije programiranje i kodiranje, iako može rezultirati pisanjem programa. Računalo razmišljanje je puno složeniji proces od programiranja i pisanja algoritama i može se primijeniti u svim područjima ljudskog života i rada te mora imati značajno mjesto u učenju i poučavanju [7].

Sastavnice računalnog načina razmišljanja su apstrakcija, dekompozicija, uočavanje uzoraka i algoritam. Kroz jedan primjer dat je prikaz koraka računalnog načina razmišljanja i rješavanja problema.

Primjer zadatka:

Borna je od novca koji je dobio od roditelja uložio 7/10 toga novca i kupio računalnu igricu. Nakon toga dao je 1/3 ostatka novca i kupio strip. Ostalo mu je 100 kn.

Koliko je novaca Borna dobio od roditelja? Koliko je dao za igricu, a koliko za strip?

Apstrakcija omogućava odvajanje bitnog od nebitnog. Ovdje je bitno uočiti i shvatiti odnose između veličina i omjera koji se odnose na trošenje novca za kupnju računalne igrice i stripa. Dakle, Borna je dao sedam desetina ukupnog novca koji je dobio od roditelja za kupnju računalne igrice. Zatim je jednu trećinu novca koji je ostao nakon kupnje igrice dao za strip. Kada je kupio i igricu i strip ostalo mu je 100 kn.

Dekompozicija omogućava razbijanje cjelokupnog problema na manje, rješive dijelove. U ovom slučaju, treba cjelovit problem razbiti na tri dijela. Prvi dio je iznos koji je dat za računalnu igricu. Drugi dio je novac koji je utrošen za kupnju stripa. Treće dio je ostatak koji je ostao nakon kupnje.

Sva tri dijela zajedno, zbrojena daju ukupan iznos, prikazan Slikom 2 (lokomotiva koja vuče kompoziciju).

X- ukupan novac koji je Borna dobio od roditelja

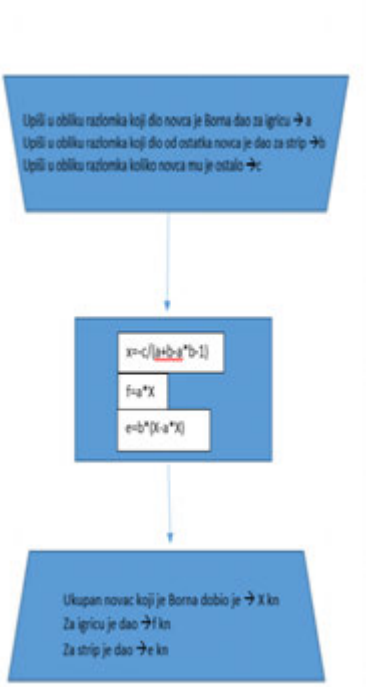
a*X- novac koji je dao za igricu (a je dio od X)

b*(x-a*x)- novac koji je dao za strip

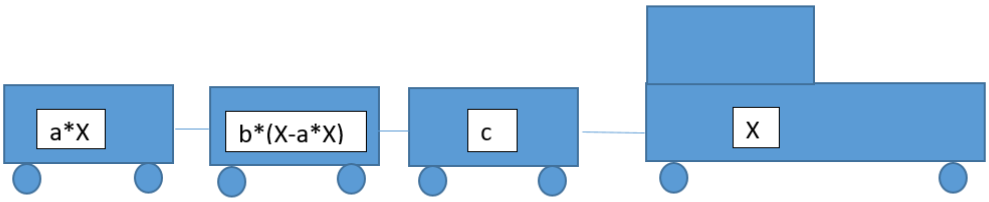
c- novac koji mu je ostao nakon što je kupio igricu i strip

Nakon izvršene dekompozicije treba uočiti mogućnost primjene uzorka za slične zadatke. Stoga je napravljen algoritam koji omogućava, prema uzorku, rješavanje i drugih zadataka gdje se mijenjaju omjeri novca koji se daje za kupnju igrice i/ili stripa. Također može se mijenjati i iznos ostatka novca koji ostaje nakon kupnje.

Nakon toga napravljen je algoritam za rješavanje ovog (i sličnih) zadataka i prikazan Slikom 1.



Slika 1 Algoritam



Slika 2. Dekompozicija

Računalni način razmišljanja pogodan je za rješavanje problema i često prethodi izradi računalnog programa. Zapravo veći dio računalnog programa ovime je i napravljen pa je sama izrada programa, samo stvar poznavanja sintakse nekog programskog jezika. Nebitno je u kojemu se programskom jeziku radi, jer računalo sve instrukcije bilo kojeg višeg programskog jezika pretvara u strojni jezik.

Program je napravljen u razvojnom okruženju Borland Delphi. a Slika 3 prikazuje Glavni prozor programa.

Programski kod:

```
procedure TForm1.Button1Click(Sender: TObject);
vara,b,c,x,e,f:real;
begin
a:=strtofloat(edit1.Text)/strtofloat(edit2.Text);
b:=strtofloat(edit3.Text)/strtofloat(edit4.Text);
c:=strtofloat(edit8.Text);
x:=-c/(a+b-a*b-1);
edit7.text:=floattostrf(x,fffixed,10,2);
f:=a*x;
edit5.Text:=floattostrf(f,fffixed,10,2);
e:=d-strtofloat(edit5.Text)-c;
edit6.text:=floattostrf(e,fffixed,10,2);
end;
```



Slika 3 Glavni prozor programa

Sam program (Project1.exe) dostupan je preko poveznice:

https://drive.google.com/file/d/1dNeqKVJCaUIpu3f_0hS_hJx5RWExhASL/view?usp=sharing

IV. RAČUNALNI NAČIN RAZMIŠLJANJA I OBRAZOVNI SUSTAV REPUBLIKE HRVATSKE

U Republici Hrvatskoj pojam se uvodi u Kurikulum informatike od 2018.g. pa se navodi i razvoj računalnog razmišljanja kao jedan od temeljnih ciljeva učenja i poučavanja nastavnog predmeta Informatike za osnovne škole i gimnazije. Tako uz ciljeve kao što su informatička pismenost, svrhovito i odgovorno korištenje digitalne tehnologije, osposobljenost za sadašnji, ali i budući život u digitalnom svijetu, razvijanje kritičkog mišljenja, kreativnosti i inovativnosti, u kurikulum ulaze i ciljevi razvoja računalnog razmišljanja, sposobnosti rješavanja problema i stjecanja vještine programiranja [12].

Također u Kurikulumu informatike navode se četiri domene kojima će se realizirati ciljevi predmeta Informatika pa je uz domene e-Društvo, Digitalna pismenost i komunikacija, Informacije i digitalna tehnologija navedena i domena Računalno razmišljanje i programiranje [12].

O važnosti računalnog razmišljanja u Kurikulumu nastavnog predmeta informatike govori i činjenica da se pojam računalno razmišljanje nalazi čak 151 put u Kurikulumu.

U Kurikulumu se, o računalnom načinu razmišljanja govori kao o temeljnoj informatičkoj vještini koja razvija pristup rješavanju problema koji je primjenjiv na računalu. Od učenika se očekuje da postanu i stvaratelji različitih računalnih alata, a ne samo korisnici. Za to je potreban razvoj vještina logičkog zaključivanja, modeliranja, apstrahiranja te rješavanja problema. Naglasak je na činjenici da se računalno razmišljanje može primijeniti u svim segmentima života [10]. Temeljni koncept računalnog načina razmišljanja je apstrakcija, gdje uz uporabu metakognitivnih vještina dolazi do rješavanja složenih problema, razdvajanjem u više jednostavnih problema. Računalno razmišljanje nije isto što i programiranje, odnosno kodiranje, ali je neophodno za programiranje [11].

S druge strane, znanje programiranja nije nužno potrebno za razvoj računalnog razmišljanja i njegovu primjenu u svakodnevnom životu.

Preteča uvrštavanja računalnog načina razmišljanja u obrazovni sustav Republike Hrvatske bilo je uvrštavanje

domene Rješavanje problema pomoću računala, u okviru Tehničkog i informatičkog područja, Nacionalnog okvirnog kurikuluma iz 2011.g. [1]. Nakon toga, u prijedlogu cjelovite kurikularne reforme iz 2016.g. objavljen je i Nacionalni kurikulum nastavnog predmeta informatike gdje je predviđena i domena Računalno razmišljanje i programiranje te su specificirani obrazovni ishodi i kompetencije koje bi učenici trebali steći [11].

Prema JRC Science For Policy Report (Developing Computational Thinking in Compulsory Education) iz 2016.g. [5], obrazovni sustav Republike Hrvatske je na vrijeme je reagirao, prepoznao i uvrstio u Kurikulum informatike (najprije u prijedlog kurikuluma) nastavne sadržaje iz računalnog načina razmišljanja [8]. Međutim, istraživačkih radova na temu računalnog načina razmišljanja u Republici Hrvatskoj gotova da i nema. Najviše je radova, prema podacima iz 2019.g., u SAD-u, oko 415 [14].

V. CILJEVI ISTRAŽIVANJA

Cilj istraživanja je ispitati stavove studenata o računalnom načinu razmišljanja te utvrditi postoje li statistički značajne razlike između studenata informatike i studenata predškolskog odgoja.

S aspekta računalnog načina razmišljanja, cilj istraživanja je ispitati vještine navedenih grupa studenata iskazane rješavanjem zadataka koji traže računalni način razmišljanja.

VI. TESTIRANJE VJEŠTINA RAČUNALNOG RAZMIŠLJANJA

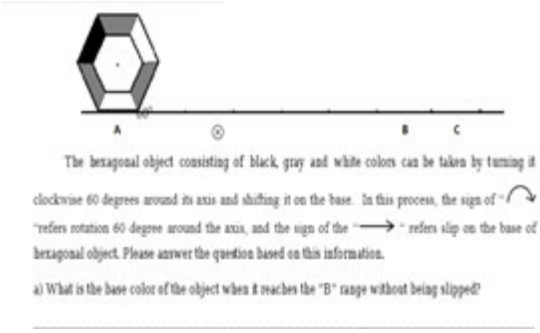
Studentima je ponuđen test računalnog načina razmišljanja. Test je imao 10 zadataka, a svaki zadatak nosio je 10 bodova. Dakle, student je mogao najviše dobiti 100 bodova. Test je, za potrebe testiranja računalnog načina razmišljanja razvio profesor Bati Kaan (Hacettepe University, Department of Science Education, Ankara, Turska) [3]. Slika 4 prikazuje zadatak s testa.

Njime je testirao 104 učenika srednje škole u Ankari. Prosječan rezultat testiranih učenika u Turskoj bio je 33 boda (od ukupno 100 bodova).

Republika Hrvatska i Republika Turska rano su prepoznale značaj računalnog načina razmišljanja i integrirale ga u Kurikulum. Također obije države su se uključile u natjecanja Dabar (Berbas) koja potiču računalni način razmišljanja [15].

Testiranje studenata sa Sveučilišta Juraj Dobrila tim testom je pokazalo kako studenti predškolskog odgoja ne zaostaju za studentima informatike u sposobnostima rješavanja zadataka u kojima je potrebno računalno razmišljanje. Rezultati su bili dosta loši pa je na tom testu prosječan student informatike imao 27,7 bodova, a prosječan student predškolskog odgoja 29,1 bod (od ukupno 100 bodova).

Ovakvi, iznenađujući rezultati gdje studenti predškolskog odgoja imaju bolje rezultate od studenata informatike, bili su povod za daljnje istraživanje.



Slika 4 Zadatak s testa

U tom istraživanju (koje slijedi) istraživani su stavovi o računalnom načinu razmišljanja, tih istih grupa studenata.

VII. METODOLOGIJA ISTRAŽIVANJA

Nakon testiranja istraženi su stavovi tih grupa studenata o računalnom načinu razmišljanja. U istraživanju je korištena deskriptivna i kauzalno-neeksperimentalna metoda. Istraživanje je izvršeno Likertovom skalom sa sedam čestica. Istraživao se pozitivan stav studenata prema računalnom načinu razmišljanja. Istraživanje je bilo anonimno i zatražena je suglasnost Povjerenstva za etičnost istraživanja Sveučilišta Juraj Dobrila u Puli. Provedeno je u veljači 2020.

A. Uzorak istraživanja

Istraživanje je obavljeno na dva nezavisna uzorka. Prvi uzorak su sačinjavala 22 studenta treće godine preddiplomskog studija informatike. Drugi uzorak se sastojao od 33 studenta treće godine predškolskog odgoja.

Istraživanjem treba utvrditi stavove studenta treće godine preddiplomskog studija predškolskog odgoja i treće godine preddiplomskog studija informatike o računalnom načinu razmišljanja kao novoj pismenosti 21. stoljeća.

B. Hipoteza

Postavljena je hipoteza kako nema statistički signifikantne razlike između pozitivnoga stava o računalnom načinu razmišljanja između studenata treće godine predškolskog odgoja i studenata treće godine preddiplomskog studija informatike na Sveučilištu Jurja Dobrile u Puli.

H0: Ne postoji statistički signifikantna razlika između pozitivnoga stava o računalnom načinu razmišljanja između studenata treće godine predškolskog odgoja i studenata treće godine preddiplomskog studija informatike na Sveučilištu Jurja Dobrile u Puli

C. Tijek istraživanja

Istraživanje je provedeno nad dvije nezavisne skupine studenata. Jedna skupina je imala 22 studenta sa studija Informatike, a druga 33 studenta sa studija predškolskog odgoja, oba studija su pri Sveučilištu Jurja Dobrile u Puli.

Za istraživanje je korištena Likertova skala pozitivnih stavova o računalnom načinu razmišljanja s pet stupnjeva (od 1- potpuno se ne slažem do 5- potpuno se slažem).

Najprije je izvršena deskriptivna analiza frekvencija, koja je pokazala kako nema značajnih odstupanja te kako su podaci pogodni za daljnju analizu.

Zatim je provjerena pouzdanost skale. Cronbachov koeficijent Alpha je 0,862 što znači da je skala pouzdana, odnosno da postoji unutarnja suglasnost skale.

Potom je ustanovljena pogodnost skale za faktorsku analizu, odnosno proveden je KMO test. KMO pokazatelj je bio 0,790, a Bartlett-ov test sferičnosti je bio 0,000, odnosno $p<0,05$ pa su podaci prikladni za faktorsku analizu. Correlation Matrix pokazuje da je većina koeficijenata korelacije veća od 0,3. Iz tablice Total Variance Explained pod zaglavljem Initial Eigenvalue vidljivo je da samo jedna komponenta ima vrijednost iznad 1 te da ona objašnjava 55,564 postotka varijance.

Faktorskom analizom utvrđeno je da se svih sedam čestica skale (1-Smatrate li da često koristite računalni način razmišljanja u svakodnevnom životu?, 2-Smatrate li da je potrebno koristiti računalni način razmišljanja u svakodnevnom životu, 3-Smatrate li da su računalni način razmišljanja i programiranje međusobno povezani?, 4-Smatrate li da je za programiranje važan računalni način razmišljanja?, 5-Smatrate li da je za računalni način razmišljanja važno znanje programiranja?, 6-Smatrate li da računalni način razmišljanja ima veliku ulogu u STEM području?, 7-Smatrate li da računalni način razmišljanja ima veliku ulogu u društvenom, humanističkom i umjetničkom području?) može povezati sa samo jednim faktorom, kako pokazuje i SCREE PLOT, Slika 5.

Utvrđena je normalnost distribucije (Kolmogorov-Smirnov test, $p=0,820$) te je zaključeno kako se može provesti t-test za nezavisne uzorke. Izračunata je aritmetička sredina svih čestica skale te je proveden t-test za nezavisne uzorke.

Proveden je t-test za nezavisne skupine kako bi se utvrdilo je li stav studenata treće godine preddiplomskog studija informatike različit od stava studenata treće godine preddiplomskog studija predškolskog odgoja.

D. Rezultati istraživanja

Istraživanje je pokazalo da studenti informatike ($M=3,8384$ $SD=0,71986$) imaju statistički signifikantno različit stav od studenata treće godine ($M=3,0976$ $SD=0,57451$), kada je u pitanju pozitivan stav prema računalnom načinu razmišljanja ($t=4,231$, $p=0,000<0,05$, $df=53$), odnosno studenti informatike imaju pozitivniji stav o računalnom načinu razmišljanja od studenata predškolskog odgoja.

Obrada je vršena programom SPSS for Windows 20.0.0.

E. Diskusija

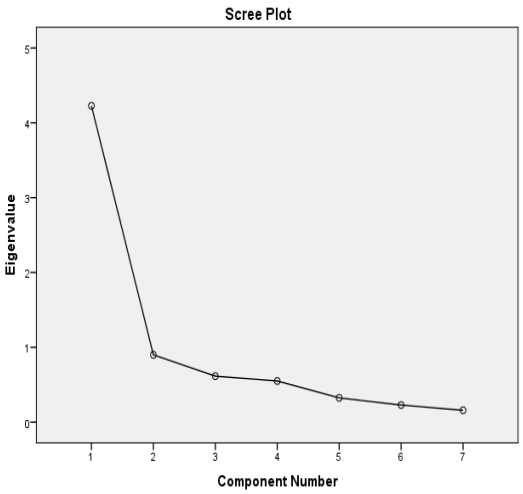
Usporede li se rezultati istraživanja vještina računalnog načina razmišljanja autora B. Kaana provedenog u Turskoj na uzorku srednjoškolaca i ovi provedeni na Sveučilištu u Puli, vidljivo je da su rezultati naših studenata puno lošiji. Kod 104 testiranih učenika u Turskoj (Ankara) prosječan rezultat bio je 33 boda (od ukupno 100 bodova), a na Sveučilištu u Puli prosječan student informatike je imao 27,7 bodova, dok je prosječan student predškolskog odgoja postigao 29,1 bod (od ukupno 100 bodova). Iako je pulski uzorak istraživanja puno manji, naši rezultati nisu zadovoljavajući. Istraživanje bi trebalo provesti na većem, reprezentativnom uzorku.

Ovakvi rezultati otvaraju mnoga pitanja. Osnovno je pitanje kako poboljšati nezadovoljavajuće sposobnosti računalnog načina razmišljanja studenata. Jedan od mogućih načina bio bi, možda, uvođenje zasebnog predmeta/kolegija Računalni način razmišljanja.

Jedan od mogućih razloga dobivenih rezultata o stavovima studenata o računalnom načinu razmišljanja i vještinama njegova korištenja na testovima, mogao bi ležati i u činjenici da naši testirani studenti nisu imali priliku kroz osnovnu i srednju školu koristiti ovaj pojam te učiti o računalnom načinu razmišljanja jer je računalni način razmišljanja uvršten u Kurikulum nastave informatike tek 2018.godine.

Na studijima nastavnog smjera informatike i predškolskog odgoja sada je odgovornost da ukažu studentima na važnost ovog način razmišljanja pri učenju i poučavanju, posebno u kontekstu njihove uloge u budućem odgoju i obrazovanju djece i učenika. Mišljenja smo da u Hrvatskoj nema dovoljno istraživanja i testova za mjerenje sposobnosti računalnog načina razmišljanja. Stoga smo uz odobrenje autora, koristili test profesora Kaana.

Drugi testovi, osim ovog, nisu nam bili dostupni, a ni poznati. Trebalo bi provesti daljnja istraživanja, na većem broju studenata kako bi se ovakvi rezultati potvrdili ili opovrgli. Također, trebalo bi uključiti i studente s drugih sveučilišta u Republici Hrvatskoj.



Slika 5 SCREEPLOT

VIII. ZAKLJUČAK

Računalni način razmišljanja važan je aspekt pismenosti 21.stoljeća. U obrazovni sustav Republike Hrvatske na vrijeme je uvršteno učenje računalnog načina razmišljanja u obrazovne dokumente i kurikulume. Republika Hrvatska time je postala jedna od zemalja Europske unije koja je u potpunosti i na vrijeme prepoznala važnost računalnog načina razmišljanja. Međutim, rezultati koji su dobiveni ovim istraživanjem ukazuju kako je problem u realizaciji obrazovnih kurikuluma. Tako se događa da jedan broj studenata koji će uskoro završiti preddiplomski studij nije do dana testiranja uopće čuo za pojam računalni način razmišljanja. Unatoč tome, studenti informatike pokazuju signifikantno pozitivniji stav od studenata predškolskog odgoja kada je u pitanju razumijevanje uloge računalnog načina razmišljanja. Testirani studenti Sveučilišta Jurja Dobrile u Puli ostvarili su prilično loš rezultat, posebno studenti informatike. Takav rezultat iznenađuje, osobito kada se uzme u obzir da je obrazovni sustav Republike Hrvatske rano prepoznao važnost računalnog načina razmišljanja kako u STEM području tako i u svim drugim područjima života i rada.

LITERATURA

[1] Agencija za odgoj i obrazovanje, Nacionalni okvirni kurikulum, Zagreb, 2011.

[2] M. Anđelić, Didaktičke igre za razvoj računalnog razmišljanja, Sveučilište u Rijeci- Odjel za informatiku, Rijeka, 2018.

[3] K. Bati, Computational Thinking Test (CTT) for Middle School Students, Akdeniz Egitim Arastirmalari Dergisi, vol. 12, Issue 2018

[4] BBC Official Hme Page of BBC, Introduction to computational thinking, preuzeto s https://www.bbc.co.uk/bitesize/guides/zp92mp3/revision/1 (Retrieved on 1 March 2020).

[5] European Commission, Science for Policy Report 2016, Developing Computational Thinking in Compulsory Education

[6] N. Bubica, M. Mladenović,I. Boljat, Programiranje kao alat za razvoj apstraktnog mišljenja, preuzeto s https://bib.irb.hr/datoteka/702093.Programiranje_kao_alat_za_razvoj_apstraktnog_miljenja-CUC-zbornik.pdf (Retrieved on 2 March 2020).

[7] L. Budin., Računalno razmišljanje i programiranje u višim razredima osnovne škole u sklopu novog kurikuluma predmeta informatika, preuzeto s https://mzo.gov.hr/userdocsimages//dokumenti/obrazovanje/nacionalnikurikulum/prezentacijewebinara//prezentacija_webinara_racunalno_razmisljanje_i_programiranje_u_visim_razredima_osnovne_skole.pdf (Retrieved 23 February 2020.)

[8] L. Budin, S. Perić, P. Brođanac, D. Škvorc, Z. Markučić, M. Babić, Računalno razmišljanje i programiranje u Pytonu, Zagreb, Element d.o.o., 2017.

[9] L. Budin., Nastava informatike treba razvijati računalni način razmišljanja, Zagreb, Školske novine, 2017.

[10] Ž. Ivanković, Obrazovanje čini razliku između kreatora i potrošača: informatikom do sposobnosti razmišljanja, preuzeto s https://www.jutarnji.hr/komentari/obrazovanje-cini-razliku-izmedu-kreatora-i-potrosaca-informatikom-do-sposobnosti-razmisljanja/4119849/ (Retrieved 5 March 2020).

[11] Ministarstvo znanosti i obrazovanja, Prijedlog nacionalnog kurikuluma nastavnog predmeta informatike, preuzeto s http://www.kurikulum.hr/wp-content/uploads/2016/03/Informatika.pdf (Retrieved 18 February 2020).

[12] Ministarstvo znanosti i obrazovanja, Kurikulum za nastavni predmet informatike za osnovne škole i gimnazije u RH, preuzeto https://www.azoo.hr/images/strucni_ispiti_2018/Odluka_o_donosenju_kurikuluma_za_nastavni_predmet_Informatike_za_osnovne_skole_i_gimnazije_u_Republici_Hrvatskoj.pdf (Retrieved 2 March 2020).

[13] D. Muraja, Kako radi računalo, Procon, Zagreb, 1996.

[14] M. Pollak, M. Ebner, 2019., The Missing Link to Computational Thinking, Graz, Austria, preuzeto s https://www.mdpi.com/1999-5903/11/12/263 (Retrieved 4 March 2020).

[15] Suradnici u učenju, Dabar (Bebras)- Međunarodno natjecanje u informatici i logičkom razmišljanju, preuzeto s http://www.ucitelji.hr/Naslovnica/Dabar.aspx (Retrieved 3 February 2020).

[16] J. M. Wing, Computational Thinking, Communications of the ACM, viewpoint, vol. 49, 2006., preuzeto s https://www.cs.cmu.edu/~15110-s13/Wing06-ct.pdf (Retrieved 1 March 2020).

[17] J. M. Wing, Computational Thinking, Ten Years Later, Microsoft Research Blog, 2016, preuzeto s https://www.microsoft.com/en-us/research/blog/computational-thinking-10-years-later/ (Retrieved 1 February 2020).

[18] M. Syslo, From Algorithmic to Computationa Thinking, ACM Conference, 2015., preuzeto s https://www.researchgate.net/publication/300481428_From_Algorithmic_to_Computational_Thinking (Retrived 2 March 2020).

Percipirana kvaliteta sustava za upravljanje verzijama programskog koda od strane studenata informatike

T. Orehovački **, S. Babić* i D. Etinger**
** Sveučilište Jurja Dobrile u Puli, Fakultet informatike u Puli, Pula, Croatia
{tihomir.orehovacki, darko.etinger}@unipu.hr
*Veleučilište u Rijeci, Poslovni odjel, Rijeka, Croatia
snjezana.babic@veleri.hr

***Sažetak* - Programiranje je značajno područje informatike u kojem se uvijek iznova traže načini kojima se mogu ubrzati i olakšati procesi usvajanja novih znanja i vještina. Jedan od načina je i primjena sustava za upravljanje verzijama programskog koda koji omogućuje korisniku podršku u svim fazama razvoja softvera kroz suradnju sa drugima iz cijelog svijeta uz pristup skladištu izvornog koda. Prihvaćanje i veća primjena ovog sustava je jednim dijelom ovisno o percipiranoj kvaliteti u korištenju ovog sustava od strane krajnjih korisnika, posebno budućih programera(ki). U ovom radu prikazani su rezultati empirijskog istraživanja percepcije kvalitete u korištenju ovog sustava od strane studenata informatike na dva hrvatska fakulteta, nakon upotrebe sustava GitHub prema zadanim koracima. Rezultati mogu biti od pomoći nastavnicima pri korištenju sustava za upravljanje verzijama programskog koda u nastavnim procesima, kao i svima onima koji imaju za cilj razvoj i korištenje sustava za upravljanje verzijama programskog koda te za daljnje istraživanje u navedenom području.**

Ključne riječi - kvaliteta, sustav za upravljanje verzijama programskog koda, GitHub, programiranje, prihvaćanje, studenti

I. UVOD

Posljednjih godina, sve se više primjenjuje agilni pristup razvoju softvera u kojem je pružanje pravovremenih procjena i povratnih informacija projektnim timovima od iznimne važnosti [1]. Za potrebe razvoja projekata softverskog inženjeringa danas se uvelike koriste sustavi za upravljanje verzijama koda (eng. Code Version Control System), čije funkcionalnosti programerima omogućuju suradničko pisanje kodova, na svjetskoj razini.

U softverskoj industriji i programerskim zajednicama, najpopularniji je GitHub (www.github.com), web osnovan hosting servis za distribuiranu kontrolu verzije Git. Github je servis za suradničko kodiranje koji integrira mnoge funkcionalnosti, uključujući wiki, praćenje izdavanja i pregled kodova [2]. Osim za pisanje kodova, sustavi za upravljanje verzijama mogu se upotrijebiti za suradničko stvaranje i praćenje drugih datoteka [3]. Nadalje, GitHub sadrži mnogo projekata koji se razlikuju obzirom na vrstu jezika, veličinu projekta, dob i broj programera i drugo, a svaki projektni repozitorij sadrži povijesne zapise o projektu, uključujući povijesne zapise doprinosa pojedinaca, autorstvo, ispravljanje kodova i drugo [3].

Procjena doprinosa GitHuba u razvoju softvera može se promatrati kroz utjecaj tehničkih i društvenih činitelja [4]. Kao primjer, prema Tsay i sur. [4] moguće je primjenom GihHuba otkriti potrebe korisnika, kao i ostvariti zapošljavanje programera. Rad programera je transparentan i vidljiv drugima.

Vasilescu i sur. [5] pronalaze da kontinuirana integracija promjena u glavno spremište doprinosa članova poboljšava produktivnost projektnih timova bez uočljivog smanjenja kvalitete koda.

Sustavi za upravljanje verzijama programskog koda pomažu profesionalnim programerima u razvoju softvera, ne samo kao mjesto za praćenje projekata, već i kao mjesto održavanja sastanaka zainteresiranih strana koje podržavaju iste [6].

Premda GitHub sustav temeljen na Gitu ima jednostavno sučelje, za upotrebu istih potrebna su određena znanja, vještine i kompetencije od strane korisnika [7].

Potencijal primjene GitHuba uočen je i od strane nastavnika u okviru mogućnosti transformiranja iskustva učenja programiranja kroz suradnju na projektnim zadacima te povezivanje mnogih dionika u tom procesu [6].

Korisnička percepcija kvalitete softvera i njihovo zadovoljstvo u radu s istima, vodi ka kontinuiranoj primjeni softvera u poslovanju i obrazovanju. Obzirom da je upotreba softvera kao podrška u nastavi ovisna o obrazovnom okruženju, pitanje percepcije studenata je uvijek iznova otvoreno. Glavni cilj u ovom radu bio je utvrditi percepciju kvalitete sustava za upravljanje verzijama programskog koda od strane studenata informatike.

U nastavku rada je dana kratka analiza prethodnih istraživanja, zatim je opisan postupak empirijskog istraživanja i diskusija rezultata istoga. Na kraju rada su dani kratki zaključci, ograničenja i smjernice za budući rad.

II. PREGLED LITERATURE

Općenito, korisnicima prikupljene informacije u servisu GitHub mogu koristiti za razvijanje strategije za koordinaciju njihova rada, unapređivanje tehničkih vještina i upravljanje njihovim ugledom u programerskim i drugim okruženjima [8].

Na postignute rezultate u GitHubu, prema Vasilescu i sur. [5], može utjecati veći broj činitelja, kao što su: veličina koda, dob, veličina tima i interes korisnika.

Borges i sur. [9] naglašavaju da GitHub podržava mnoge tipične značajke društvenih mreža, pri čemu navode mogućnost označavanja spremišta sa zvjezdicom od strane korisnika kao „posrednika" za popularnost projekta, što nije uvijek slučaj.

Primjena GitHuba u obrazovnom procesu, posebno za transparentni i timski razvoj programskih kodova, otvara nove mogućnosti učenja i poučavanja programiranja na svim razinama. Studenti mogu stvarati, uređivati i dijeliti programske kodove, pri čemu stječu važnost njihova doprinosa u istome [10].

Zagalsky i sur. [6] ističu da se Git i GitHub mogu upotrijebiti kao sustavi za e-učenje obzirom na neke slične značajke (npr. izvršavanje zadataka i učitavanje), međutim nedostaju pojedine značajke, kao što su alati za upravljanje procjenom znanja. Primjenom GitHuba u nastavi, nastavnici mogu kontinuirano pratiti rezultate svakog pojedinog studenta tijekom procesa rješavanja programerskih zadataka.

Rezultati Kalliamvakou i sur. [7] pokazuju da svi nastavnici ne koriste na isti način GitHub u njihovoj nastavi, obzirom da je upotreba istoga ovisna o mnogo činitelja, kao primjer, broj studenata i vrsta nastavnog predmeta. Na osnovu rezultata istraživanja kojeg su proveli, Kalliamvakou i sur. [7] navode sljedeće prednosti GitHuba u nastavi: transparentnost aktivnosti, poticanje sudjelovanja, relevantnost industrije, jednostavnost korištenja, besplatne akademske licence, verzija zajedničkog prostora i tečaja. Nadalje, kao izazove u primjeni GitHuba u nastavi navode: zajedničku bazu znanja o predloženim praksama, prepreke za ulazak, podrška za dodatne formate, vanjska ograničenja i upravljanje velikim opsegom. Autori ističu da poteškoće s Gitom obično imaju programeri početnici [7].

Prema nalazima Valdivia [10], korištenje Gita u nastavi može pomoći nastavnicima kao podrška u poučavanju programiranja, posebno za studente s nižom razinom programerskih vještina. Kao neke od prednosti, autori navode mogućnosti komunikacije i suradnje pri pisanju kodova, razvijanje vještine kritičkog razmišljanja, timskog rada te mogućnosti ispravljanja kodova.

Kontinuirani sustav pružanja povratnih informacija upotrebom servisa GitHub je vrlo učinkovit u obrazovanju kod primjene agilnih projekata u nastavi [1].

Rezultati istraživanja koje su proveli Zakiah i Fauzan [11] pokazuju da primjenom njihovog modela zajedničkog učenja softverskog inženjeringa pomoću Githuba za studente informatike, olakšava se proces učenja kroz suradnju studenata u postizanju definiranih ishoda učenja.

III. METODOLOGIJA EMPIRIJSKOG ISTRAŽIVANJA

Glavni cilj u ovom istraživanju je bio utvrditi percepciju kvalitete u korištenju za upravljanje verzijama programskog koda od strane studenata informatike na dva hrvatska fakulteta, nakon upotrebe sustava GitHub prema zadanim koracima.

A. *Postupak*

Provedba istraživanja se sastojala od dva temeljna dijela. Sudionici u istraživanju su trebali najprije dovršiti 27 reprezentativnih koraka scenarija interakcije sa sustavom za upravljanje izvornim kodom GitHub, a potom popuniti upitnik. Isti je bio sačinjen od 5 čestica vezanih uz demografiju ispitanika te 140 čestica namijenjenih vrednovanju prihvaćanja i kvalitete sustava za upravljanje izvornim kodom GitHub od strane studenata. Upitnik je izrađen pomoću aplikacije Google Forms te je ista upotrijebljena i u svrhu prikupljanja podataka od ispitanika. Razinu svojeg slaganja sa česticama iz upitnika sudionici su izražavali preko Likertove skale od pet stupnjeva (1 – u potpunosti se slažem, 5 – u potpunosti se ne slažem).

B. *Instrument istraživanja*

Za izradu upitnika korišteni su konstrukti iz modela za vrednovanje kvalitete u korištenju [12] koji je udovoljio svim psihometrijskim obilježjima u kontekstu vrednovanja društvenih Web aplikacija [13], njihove kvalitete kada se iste primjenjuju u mobilnom okruženju [14] i obrazovnom kontekstu [15], između ostalog. Rezultati Cronbach alpha koeficijenata, vidljivi u Tablici I., potvrđuju da sve skale u korištenom mjernom instrumentu imaju dobru unutarnju konzistentnost (od 0.79 do 0.94).

C. *Ispitanici*

U istraživanju je sudjelovalo ukupno 172 studenata od kojih je 72.1% bilo muškog, a 27.9% ženskog spola. U trenutku provedbe istraživanja, većina (89%) sudionika su bili redovni studenti. Veći dio uzorka ispitanika (64%) studira na Sveučilištu Jurja Dobrile u Puli, dok ih preostalih 36% pohađa jedan od studija na Veleučilišu u Rijeci. Uzorak ispitanika je u većini (53.5%) bio sačinjen od studenata prve godine preddiplomskog studija, 26.2% ih je bilo sa treće godine preddiplomskog studija, 11.6% je u trenutku provedbe istraživanja bilo na prvoj godini diplomskog studija, 5.2% ih je studiralo na drugoj godini diplomskog studija dok ih je 3.5% pohađalo drugu godinu preddiplomskog studija.

Kada je riječ o razini poznavanja korištenja sustavom za upravljanje izvornim kodom GitHub, većina studenata (48.3%) je istu procijenila kao dobru, pri čemu samo 2.3% studenata smatra da se zna izvrsno služiti GitHubom, dok ih 7% vjeruje da je njihovo poznavanje primjene GitHuba jako loše. Obzirom na učestalost korištenja GitHub sustava, 38.9% studenata utroši barem sat vremena tjedno na interakciju sa GitHubom pri čemu ga 39.5% upotrebljava najmanje jednom tjedno, dok ih samo 3.5% upotrebljava GitHub barem jednom dnevno.

TABLICA I. BROJ ČESTICA I UNUTARNJA KONZISTENCIJA SKALA KORIŠTENIH U ISTRAŽIVANJU (N=173)

Skale	**Broj čestica**	**Cronbach alpha koeficijent**
Stav	5	**.89**
Utjecaj društva	3	**.90**
Olakšavajuće okolnosti	5	**.82**
Mogućnost kontrole	6	**.82**
Važnost za posao	5	**.82**
Lakoća korištenja	6	**.91**
Lakoća učenja korištenja	3	**.82**
Pamtljivost	3	**.86**
Razigranost	7	**.86**
Estetika	5	**.91**
Korisnost	5	**.90**
Djelotvornost	5	**.85**
Učinkovitost	5	**.86**
Skalabilnost	4	**.83**
Kvaliteta rezultata	3	**.86**
Pouzdanost	5	**.85**
Sigurnost	7	**.89**
Upravljanje promjenama	4	**.79**
Suradnja	3	**.84**
Obilježja tehnologije	7	**.91**
Usklađenost zadataka i tehnologije	3	**.83**
Obilježja zadataka	5	**.90**
Funkcionalnost	10	**.91**
Potvrda očekivanja	5	**.89**
Užitak	6	**.94**
Zadovoljstvo	4	**.90**
Navika	6	**.90**
Ponašajna namjera	5	**.91**

D. Rezultati

Analiza prikupljenih podataka pokazala je da većina studenata ima pozitivno mišljenje o primjeni GitHuba. Naime, većina studenata smatra da je dobro (84.3%) i mudro (82.5%) upotrebljavati GitHub, njih 81.9% ima pozitivan stav prema primjeni sustava za upravljanje izvornim kodom te ih jednako toliko zagovara ideju o upotrebi GitHuba u tom kontekstu, dok ih se 68% veseli zadacima i projektima zbog kojih će trebati upotrebljavati GitHub.

Rezultati istraživanja impliciraju da ljudi koji okružuju sudionike istraživanja imaju utjecaj na njihovo ponašanje vezano uz upotrebu sustava za upravljanje izvornim kodom. Konkretnije, 35.5% studenata vjeruje da ljudi koji su njima važni misle da oni trebaju koristiti GitHub, 30.9% ispitanika je mišljenja da ljudi koji utječu na njihovo ponašanje smatraju kako oni trebaju upotrebljavati GitHub dok ih 54% misli da ljudi čije mišljenje cijene preferiraju da oni upotrebljavaju GitHub.

Većina sudionika u istraživanju također smatra da ne postoje nikakve prepreke njihovoj interakciji sa sustavom za upravljanje izvornim kodom. Naime, 67.5% ispitanika je mišljenja da raspolažu sa svim relevantnim resursima potrebnim za korištenje GitHubom, 41.8% studenata vjeruje da imaju svo potrebno znanje potrebno za primjenu GitHuba, 48.3% sudionika smatra da raspolažu vještinama potrebnim ua upotrebu sustava za upravljanje izvornim kodom, 68.6% studenata vjeruje da je GitHub kompatibilan sa tehnologijama namijenjenim razvoju programskih proizvoda, dok je 77.9% studenata uvjereno da će im, u slučaju poteškoća u radu sa GitHubom, njihovi kolege moći pomoći riješiti spomenutu poteškoću.

Sudionici u istraživanju također smatraju da upravljaju cijelim procesom interakcije sa sustavom za upravljanje izvornim kodom. Konkretnije, 62.8% studenata vjeruje da ima kontrolu nad upotrebom GitHuba, 62.3% ispitanika smatra da je jednostavno postići da GitHub radi ono što oni žele, 64.6% sudionika misli da redoslijed upotrebe funkcionalnosti sa kojima raspolaže GitHub u potpunosti ovisi o samom korisniku, 77.3% studenata smatra da imaju potpunu slobodu u načinu interakcije sa GitHubom dok 79.1% i 79.7% ispitanika vjeruje da im GitHub omogućuje da zadatke vezane uz razvoj programskih proizvoda izvršavaju tempom i redoslijedom koji im odgovara, respektivno.

Rezultati istraživanja su također otkrili kako studenti smatraju da je od velike važnosti upotrebljavati sustav za upravljanje izvornim kodom tijekom razvoja programskog proizvoda. Naime, 59.3% i 48.8% ispitanika misli kako je tijekom razvoja programskog proizvoda nužno i neophodno upotrebljavati GitHub, respektivno, dok ih 80.8%, 82.5% i 71.5% vjeruje kako je GitHub prikladan, primjeren i relevantan za upotrebu tijekom razvoja programskog proizvoda, respektivno.

Analiza prikupljenih podataka pokazala je da studenti u interakciju sa sustavom za upravljanje izvornim kodom ne trebaju uložiti previše truda. Detaljnije, 58.7% sudionika misli kako je interakcija sa GitHubom jasna i razumljiva, 51.2% ispitanika smatra da je GitHub lagan za korištenje, 51.7% studenata vjeruje kako je primjenom GitHuba lagano izvršavati zadatke vezane uz razvoj programskih proizvoda, 54.7% studenta percipira GitHub jednostavnim za upotrebu, 33.1% sudionika misli kako tijekom upotrebe GitHuba nije potrebno konzultirati sadržaje pomoći korisnicima dok 53.5% ispitanika vjeruje kako im je, kada se u obzir uzmu potrebni resursi, znanje i vještine, lagano koristiti se GitHubom.

Većina studenata također smatra kako nije teško naučiti upotrebljavati sustav za upravljanje izvornim kodom. Naime, 61% studenata misli da je jednostavno postati vješt/a u upotrebi GitHuba, 62.2% ispitanika vjeruje da je u GitHubu lagano naučiti izvršavati zadatke vezane uz razvoj programskih proizvode dok 69.2% sudionika smatra kako je lagano naučiti koristiti se funkcionalnostima sučelja GitHuba.

Analiza podataka vezanih uz pamtljivost sustava za upravljanje izvornim kodom otkrila je da 73.2% studenata misli da je jednostavno zapamtiti kako se upotrebljava GitHub, 75.6% ispitanika smatra da je lagano zapamtiti gdje se nalaze funkcionalnosti sučelja GitHuba dok 61.1% sudionika vjeruje da će biti jednostavno prisjetiti se gdje se nalaze i kako se upotrebljavaju određene funkcionalnosti GitHuba, čak i kada se aplikacije ne bude koristila neko vrijeme.

Kada je bila riječ o ispitivanju razine percipirane razigranosti kroz interakciju sa sustavom za upravljanje izvornim kodom, studenti su odgovorili da se, kada upotrebljavaju GitHub, osjećaju spontano (33.1%), kreativno (51.2%) i originalno (40.1%). Pored navedenog, 45.9% studenata smatra da interakcija sa GitHubom uspješno zadržava njihovu pažnju dulje vrijeme, 63.4% ispitanika vjeruje da je tijekom upotrebe GitHuba u potpunosti koncentrirano na provedbu zadataka namijenjenih razvoju programskih proizvoda, 29.1% sudionika misli da tijekom primjene GitHuba ne primjećuje što se oko njih događa dok 30.2% studenata smatra da ih interakcija sa GitHubom u potpunosti zaokuplja.

Analiza prikupljenih podataka je također otkrila kako se studenti sviđa dizajn sučelja sustava za upravljanje izvornim kodom. Konkretnije, 58.7% ispitanika smatra da je dizajn sučelja GitHuba vizualno privlačan, 72.1% i 72.6% studenata misli da su boje koje prevladavaju na sučelju GitHuba prikladno usklađene i ugodne za oči, respektivno, 52.9% sudionika vjeruje kako GitHub ima atraktivno sučelje dok 66.9% studenata smatra kako GitHub u cjelini lijepo izgleda.

Rezultati istraživanja otkrili su da većina (84.3%) studenata smatra da je tijekom razvoja programskog proizvoda korisno upotrebljavati sustav za upravljanje izvornim kodom. Navedeno proizlazi iz činjenice da ispitanici vjeruju kako GitHub olakšava razvoj programskih proizvoda (83.1%), povećava produktivnost u razvoju programskih proizvoda (69.7%), pospješuje kvalitetu procesa razvoja programskih proizvoda (71.6%) te povećava učinak u razvoju programskih proizvoda (70.9%).

Analizom prikupljenih podataka otkriveno je kako 75% sudionika u istraživanju smatra da primjena sustava za upravljanje izvornim kodom povećava djelotvornost u razvoju programskih proizvoda prvenstveno zbog toga što 64.5% studenata vjeruje kako je primjenom GitHuba moguće u potpunosti razviti programski proizvod, 59.9% i 68% ispitanika smatra da GitHub sadrži sve funkcionalnosti potrebne za razvoj programskih proizvoda te da su iste raznovrsne i relevantne, respektivno, dok 59.3% sudionika misli kako GitHub podržava sve faze razvoja programskih proizvoda.

Podaci prikupljeni od sudionika u istraživanju impliciraju da upotreba sustava za upravljanje izvornim kodom doprinosi povećanju učinkovitosti (72.7%) u razvoju programskih proizvoda. Naime, studenti su nakon interakcije sa GitHubom percipirali kako je njegovom upotrebom zadatke vezane uz razvoj programskih proizvoda moguće izvršiti brzo (73.3%) uz poduzimanje malog broja koraka (54.1%) što rezultira uštedom vremena (67.4%) te ubrzavanjem cjelokupnog procesa razvoja programskih proizvoda (70.3%).

Rezultati istraživanja pokazali su kako je skalabilnost sustava za upravljanje izvornim kodom na prihvatljivoj razini iz razloga što GitHub omogućava pohranu velike količine projekata (77.3%), podržava istovremeni rad velikog broja programera na projektu (89%), omogućava istovremeni rad na većem broju zadataka vezanih uz određeni projekt (87.2%) te omogućava izvršavanje kompleksnih zadataka vezanog uz razvoj programskih proizvoda (80.8%).

Kada je u fokus vrednovanja stavljena na rezultate proizašlih iz upotrebe sustava za upravljanje izvornim kodom, 74.4% odnosno 69.1% studenata smatra da je kvaliteta istih visoka odnosno izvrsna, respektivno, te je stoga 72.7% ispitanika zadovoljno rezultatima proizašlim iz upotrebe GitHuba.

Nadalje, 81.9% studenata misli da je sustav za upravljanje izvornim kodom pouzdan. Razlog tome je što 84.3% i 65.7% ispitanika vjeruje da je GitHub stabilan i bez bugova, respektivno, te što je 78.5% i 72.7% sudionika percipiralo da tijekom upotrebe GitHuba nije dolazilo do prekida u radu te se pogreške nisu javljale, respektivno.

Rezultati istraživanja pokazali su da je sigurnost sustava za upravljanje izvornim kodom na dostatnoj razini. Naime, 75% studenata smatra da GitHub ima implementirane sve potrebne sigurnosne mehanizme, 76.2% ispitanika vjeruje da GitHub štiti njihove projekte od neovlaštenog pristupa, 69.2% sudionika misli da GitHub provjerava identitet korisnika prije nego što omogući pristup pohranjenim projektima, 76.8% sudionika misli da su njihovi projekti pohranjeni na GitHubu sigurni, 69.2% studenata vjeruje da GitHub stvara sigurnosne kopije njihovih projekata pohranjenih na istom, 70.3% ispitanika vjeruje da GitHub štiti njihove projekte i osobne podatke dok 66.9% sudionika misli da GitHub podržava sigurnu pohranu njihovih osobnih podataka i projekata.

Iz analize prikupljenih podataka je evidentno da Github olakšava upravljanje promjenama nad izvornim kodom. Konkretnije, studenti smatraju da primjena GitHuba olakšava rješavanje problema nekompatibilnosti među dijelovima koda (66.9%), nastanak ljudske pogreške tijekom kodiranja svodi na minimum (58.2%), sprječava konflikte do kojih dolazi tijekom istovremenog razvoja programskih proizvoda (64%) te poništava nepoželjne promjene nad kodom (75.6%).

Rezultati istraživanja također pokazuju da 85.5% sudionika smatra kako sustav za upravljanje izvornim kodom omogućava interakciju među programerima koji rada na istom projektu. Navedeno se prvenstveno odnosi na kolaboraciju među različitim timovima koji rade na istom projektu u što vjeruje 89.6% ispitanika te na preraspodjelu zadataka vezanih uz rad na istom projektu što misli 87.2% studenata.

Kada su se vrednovala temeljna obilježja sustava za upravljanje izvornim kodom, rezultati analize prikupljenih podataka pokazali su da 84.9% ispitanika smatra da GitHub omogućava sinkroni razvoj programskih

proizvoda, 86.7% studenata misli da GitHub podržava paralelni razvoj programskih proizvoda, 79% sudionika vjeruje da GitHub omogućava neovisni razvoj programskih proizvoda, 79.1% ispitanika misli da GitHub podržava istovremeni razvoj programskih proizvoda, 86.6% studenata je uvjereno da GitHub omogućava razvoj programskih proizvoda bilo kada i bilo gdje, 81.4% sudionika smatra da GitHub podržava razvoj programskih proizvoda u realnom vremenu dok 88.9% ispitanika vjeruje da GitHub omogućava udaljeni rad na projektu.

Istraživanje je također otkrilo kako ispitanici smatraju kako je sustav za upravljanje izvornim kodom primjeren za upotrebu u kontekstu razvoja programskih proizvoda. Konkretnije, funkcionalnosti sa kojima raspolaže sustav za upravljanje izvornim kodom za 83.2% studenata predstavljaju prikladnu podršku razvoju programskih proizvoda, za 74.4% sudionika predstavljaju dostatnu podršku razvoju programskih proizvoda dok za 71.5% studenta u potpunosti udovoljavaju njihovim potrebama kao programera.

Kada je riječ o specifičnostima zadataka vezanih uz razvoj programskih proizvoda, istraživanje je pokazalo da GitHub omogućava udaljeni pristup izvornom kodu što smatra 86% ispitanika, udaljenu implementaciju izvornog koda što misli 83.2% ispitanika, udaljeno pregledavanje izvornog koda u što vjeruje 85.5% studenata, udaljeno testiranje izvornog koda što smatra 72.1% sudionika te udaljeno izvršavanje izvornog koda što je percipiralo 75% ispitanika.

Rezultati istraživanja su također pokazali kako je većina ispitanika zadovoljno sa funkcionalnostima sa kojima raspolaže sustav za upravljanje izvornim kodom. Naime, 51.7% studenata smatra kako GitHub omogućava automatsko testiranje projekata, 84.9% sudionika vjeruje da GitHub olakšava praćenje promjena nad projektom, 71.5% studenata misli da GitHub omogućava vraćanje prethodne verzije projekta, 80.8% ispitnika smatra kako GitHub omogućava upravljanje kodom većeg broja projekata, 87.2% sudionika je uvjereno kako GitHub olakšava usporedbu verzija projekta, 77.9% studenata misli kako GitHub olakšava analizu i uklanjanje uzroka problema u projektu, 77.3% ispitanika vjeruje kako GitHub olakšava reviziju projekta, 85.5% sudionika misli da GitHub omogućava kreiranje, brisanje i ažuriranje projekta, 83.7% studenata je uvjereno da GitHub omogućava kreiranje, kloniranje i arhiviranje repozitorija, dok 83.7% ispitanika smatra da GitHub omogućava pregledavanje povijesti promjena nad projektom.

Analizom prikupljenih podataka utvrđeno je da su učekivanja sudionika vezana uz interakciju sa sustavom za upravljanje izvornim kodom potvrđena. Naime, 59.9% studenata se izjasnilo kako je njihovo iskustvo upotrebe GitHuba bilo bolje od očekivanog, 69.7% ispitanika je izjavilo da je većina njihovih očekivanja vezanih uz upotrebu GitHuba potvrđeno, 64% studenata je odgovorilo kako je kvaliteta funkcionalnosti sa kojima raspolaže GitHub veća od očekivane dok je 72.6% sudionika potvrdilo da je GitHub ispunio njihova očekivanja.

Kada je riječ o percipiranom užitku tijekom interakcije sa sustavom za upravljanje izvornim kodom, 54.1% studenata smatra da je GitHub zabavan dok ih 58.2% vjeruje kako je isti zanimljiv za korištenje. Pored navedenog, 66.8% ispitanika je potvrdilo kako im je bilo ugodno upotrebljavati GitHub, 67.5% njih se sviđa način izvršavanja zadataka vezanih uz razvoj programskih proizvoda pomoću funkcionalnosti sučelja GitHuba, 49.5% ispitanika se izjasnilo da ih je interakcija sa GitHubom oduševila dok je 40.7% studenata potvrdilo kako ih interakcija sa GitHubom veseli.

Većina studenata (75.6%) je zadovoljno upotrebom sustava za upravljanje izvornim kodom pri čemu se to uglavnom (64%) odnosi na način izvršavanja zadataka vezanih uz razvoj programskih proizvoda pomoću elemenata sučelja GitHuba. Osim toga, GitHub je na 68% ispitanika ostavio dobar dojam, dok je 73.3% studenata potvrdilo kako GitHub udovoljava njihovim potrebama vezanim uz razvoj programskih proizvoda.

Premda je većina studenata upotrebljavala sustav za upravljanje izvornim kodom neko kraće vrijeme, 31.4% ih je odgovorilo da im je upotreba GitHuba u svrhu razvoja programskog proizvoda postala navika, samo 19.2% ispitanika je potvrdilo kako je postalo ovisno o korištenju GitHubom tijekom razvoja programskog proizvoda, 66.8% sudionika je izjavilo kako želi upotrebljavati GitHub tijekom razvoja programskog proizvoda, 44.2% ih je odgovorilo da imaju osjećaj da moraju upotrebljavati GitHub u tom kontekstu, 25% studenata je potvrdilo kako im je upotreba GitHuba postala svakodnevnica dok je 29.1% ispitanika izjavilo kako im je postalo normalno da tijekom razvoja programskog proizvoda upotrebljavaju GitHub.

Kada je riječ o ponašajnoj namjeri vezanoj uz upotrebu sustava za upravljanje izvornim kodom, 43.1% ispitanika je odgovorilo da će koristiti GitHub svaki put kada će razvijati programski proizvod, 59.9% ih namjerava nastaviti upotrebljavati GitHub, 70.9% studenata bi svakome preporučilo GitHub, 52.9% sudionika je izjavilo da namjerava često upotrebljavati GitHub dok je 71.5% ispitanika potvrdilo kako namjerava GitHub upotrebljavati i ubuduće.

IV. ZAKLJUČAK

GitHub se sve više primjenjuje u obrazovnim procesima za učenje i poučavanje programiranja, pri čemu postoji potreba za razumijevanjem činitelja usvajanja istoga od strane svih dionika.

Cilj ovog rada bio je utvrditi percepciju kvalitete sustava za upravljanje verzijama programskog koda od strane studenata informatike, na primjeru korištenja GitHuba.

Rezultati empirijskog istraživanja pokazali su da većina studenata ima pozitivno mišljenje o primjeni GitHuba za razvoj programskih kodova te da ljudi koji okružuju sudionike istraživanja imaju utjecaj na njihovo ponašanje vezano uz upotrebu sustava za upravljanje izvornim kodom.

Važno je naglasiti kako većina sudionika u istraživanju smatra da ne postoje nikakve prepreke njihovoj interakciji sa sustavom za upravljanje izvornim kodom te da je od velike važnosti upotrebljavati sustav za upravljanje izvornim kodom tijekom razvoja programskog proizvoda.

Rezultati analize prikupljenih podataka pokazali su da se studentima sviđa dizajn sučelja, te su mišljenja da sustav doprinosi povećanju njihove učinkovitosti, kao i da sustav ima sve potrebne funkcionalnosti za razvoj programskih proizvoda.

Općenito, većina studenata (75.6%) je zadovoljno upotrebom sustava za upravljanje izvornim kodom. Međutim, samo mali broj sudionika, njih 31.4%, smatra da im je upotreba GitHuba u svrhu razvoja programskog proizvoda postala navika. Navedeno se može povezati s činjenicom da nisu svi studenti u ovome istraživanju imali prethodna iskustva u radu s GitHubom.

Velika većina studenata (71.5%) koji su sudjelovali u istraživanju je potvrdilo kako namjerava GitHub upotrebljavati i ubuduće što je pokazatelj njihove percepcije kvalitete sustava te zadovoljstva rada s GitHubom.

Ograničenje ove studije se ogleda u homogenosti sudionika u studiji te obrazovnom okruženju i načinu primjene GitHuba. Budući rad će se temeljiti na ispitivanju povezanosti činitelja uključenih u istraživački model spram razine usvajanja sustava kontrole verzije razvoja programskog koda. Rezultati ovog rada mogu biti od pomoći svima onima koji žele primjenjivati GitHub u obrazovnom procesu.

LITERATURA

[1] X. Bai, M. Li, D. Pei, S. Li, & D. Ye, "Continuous delivery of personalized assessment and feedback in agile software engineering projects", In Proceedings of the 40th International Conference on Software Engineering: Software Engineering Education and Training, pp. 58-67, 2018.

[2] F. Thung, T. F., Bissyande, D. Lo, & L. Jiang, "Network structure of social coding in github", In 2013 17th European conference on software maintenance and reengineering, pp. 323-326, IEEE, 2013.

[3] J. D. Blischak, E. R. Davenport, & G. Wilson, "A quick introduction to version control with Git and GitHub", PLoS computational biology, 12(1), pp. 1-18, 2016.

[4] J. Tsay, L. Dabbish, & J. Herbsleb, J., "Influence of social and technical factors for evaluating contribution in GitHub", In Proceedings of the 36th international conference on Software engineering, pp. 356-366, 2014.

[5] B. Vasilescu, Y. Yu, H. Wang, P. Devanbu, P., & V. Filkov, "Quality and productivity outcomes relating to continuous integration in GitHub", In Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, pp. 805-816, 2015.

[6] A. Zagalsky, J. Feliciano, M. A. Storey, Y. Zhao, Y., & W. Wang, "The emergence of github as a collaborative platform for education", In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 1906-1917, 2015.

[7] E. Kalliamvakou, G. Gousios, K. Blincoe, L. Singer, D. M. German, & D. Damian, "The promises and perils of mining GitHub", In Proceedings of the 11th working conference on mining software repositories, pp. 92-101, 2014.

[8] I. Dabbish, C. Stuart, J. Tsay, & J. Herbsleb, "Social coding in GitHub: transparency and collaboration in an open software repository", In Proceedings of the ACM 2012 conference on computer supported cooperative work , pp. 1277-1286, 2012.

[9] H. Borges, A. Hora, & M. T. Valente, "Understanding the factors that impact the popularity of GitHub repositories", In 2016 IEEE International Conference on Software Maintenance and Evolution (ICSME), pp. 334-344, IEEE 2016.

[10] R. G. B. Valdivia, "Collaborative Learning Using Git with GitLab in Students of the Engineering Programming Course", pp. Proceedings of the International Congress on Educational and Technology in Sciences 2019, Arequipa, Perú, pp. 92-101, 2019.

[11] A. Zakiah, & M. N. Fauzan, "Collaborative learning model of software engineering using github for informatics student", In 2016 4th International Conference on Cyber and IT Service Management, pp. 1-5, IEEE, 2016.

[12] T. Orehovački, "Metodologija vrjednovanja kvalitete u korištenju aplikacijama Web 2.0", doktorski rad, Fakultet organizacije i informatike u Varaždinu, Sveučilište u Zagrebu, 2013.

[13] T. Orehovački, A. Granić, & D. Kermek, "Evaluating the Perceived and Estimated Quality in Use of Web 2.0 Applications", Journal of Systems and Software, 86(12), pp. 3039-3059, 2013.

[14] T. Orehovački, & S. Babić, "Identifying the Relevance of Quality Dimensions Contributing to Universal Access of Social Web Applications for Collaborative Writing on Mobile Devices: An Empirical Study", Universal Access in the Information Society, 17(3), pp. 453-473, 2018.

[15] T. Orehovački, D. Etinger, & S. Babić, "Modelling an interplay of adoption determinants with respect to social Web applications used in massive online open courses", Universal Access in the Information Society, 18(3), pp. 469-487, 2019.

Usporedba nastave na daljinu i učioničke nastave iz predmeta Informatika i Računalstvo

Saida Deljac*, Mladenka Berović**
* V. gimnazija, Zagreb, Hrvatska
** Prometno-tehnička škola Šibenik, Šibenik, Hrvatska
saida.deljac@skole.hr, mladenka.berovic@skole.hr

***Sažetak* - Ovaj rad daje usporedbu učioničke nastave i nastave na daljinu iz nastavnog predmeta Informatika u V. gimnaziji Zagreb i predmeta Računalstvo u Prometno-tehničkoj školi Šibeniku. Rad opisuje organizaciju rada, nastavne metode, načine vrednovanja i programske alate koji su se koristili u učionici u vrijeme redovne nastave i u modelu nastave na daljinu koja je uslijedila u vrijeme pandemije virusa COVID 19.**

U okviru rada prikazani su rezultati evaluacije učenika nakon provođenja nastave na daljinu. Rezultati evaluacije ukazuju na pozitivne i negativne aspekte nastave na daljinu te će se u slučaju ponovne potrebe takve nastave koristiti za daljnja unaprjeđenja i poboljšanja.

Ključne riječi – učionička nastava, nastava na daljinu; Moodle; Loomen; MS Teams

I. UVOD

V. gimnazija Zagreb je prirodoslovno - matematička gimnazija koja upisuje učenike u tri različita programa:

- program A – 2 strana jezika
- program B – 1 strani jezik i povećani broj sati matematike (za 2) kroz sve četiri godine
- program C – 1 strani jezik, povećani broj sati matematike (za 1) i povećani broj sati informatike (za 1) kroz sve četiri godine

Svi programi provode se kroz 4 godine učenja informatike po 2 sata tjedno, odnosno 3 za program s pojačanom informatikom. Provođenje nastave ostvareno je na način da se jedan sat izvodi u klasičnoj učionici koja je opremljena nastavničkim računalom i projektorom i jedan sat, odnosno dva za program s pojačanom informatikom, u specijaliziranoj informatičkoj učionici opremljenoj s 15 učeničkih računala prema standardu jedan učenik za jednim računalom.

Prometno-tehnička škola Šibenik je srednja strukovna škola koja obrazuje učenike u prometnim zanimanjima. Zanimanja su se s godinama mijenjala, po nazivu, kurikulumu i broju sati Računalstva, čak je i sam predmet mijenjao naziv iz Informatike u Računalstvo. Također je iz skupine općeobrazovnih prešao u skupinu strukovnih predmeta, a mijenjao se i tjedni broj sati i broj godina učenja. Sve se to odrazilo na znanja i vještine koje učenici stječu kroz obrazovanje i na njihovo snalaženje u novonastaloj situaciji nastave na daljinu u kojoj su se našli zbog pandemije virusa COVID 19.

U ovom istraživanju sudjelovali su učenici četiri četverogodišnja zanimanja: tehničar cestovnog prometa, tehničar za poštanske i financijske usluge, tehničar za logistiku i špediciju i pomorski nautičar.

Tehničar cestovnog prometa i tehničar za poštanske i financijske usluge u svom programu nastave imaju Računalstvo samo jednu godinu (kao strukovni sadržaj) dok tehničar za logistiku i špediciju i pomorski nautičar imaju dvije godine kao opće obrazovni sadržaj. Cjelokupna nastava Računalstva održava se u specijaliziranoj informatičkoj učionici opremljenoj potrebnom prezentacijskom opremom i radnim mjestima za učenike, prema pedagoškom standardu, jedan učenik za jednim računalom.

II. UČIONIČKA NASTAVA

A. Informatika

Učionička nastava u V. gimnaziji Zagreb ostvaruje se primjenom različitih programskih alata i sadržaja. Za ostvarivanje nastave, nastavnik sam odabire alate i nastavne metode koje smatra najpovoljnijim za ostvarivanje obrazovnih ishoda koji su zadani kurikulumom. Uz ostale alate koji se primjenjuju u nastavi, autorica koristi i Moodle – online sustav za udaljeno učenje koji CARNet nudi školama pod nazivom Loomen.

Iako je programski alat Moodle (Loomen) prvenstveno namijenjen za realizaciju nastavu na daljinu on se u nastavi informatike u V. gimnaziji Zagreb koristi prvenstveno kao alat za podršku učioničkoj (klasičnoj) nastavi, isprva samo u drugim razredima, a nakon toga i u prvim, trećim i četvrtim razredima. Uvođenje Loomena kao tehnološke potpore za učenje, posljedica je potrebe jedinstvenog virtualnog mjesta u kojem će se nalaziti svi digitalni sadržaji za učenje namijenjeni učenicima, a kojima će oni moći pristupati neovisno o fizičkom mjestu na kojem se trenutno nalaze i u vrijeme kada njima odgovara. Drugi razlog je nužnost podučavanja učenika novim alatima i načinima učenja i samoučenja koje će oni primjenjivati u budućim etapama svoga života (na fakultetu i u poslovnom okruženju).

U Loomen-u se nalaze sve vježbe i zadaci za samostalni rad učenika koje se provode na redovnim blok-

satovima nastave informatike u informatičkoj učionici. Također, u Loomen-u se nalaze i drugi obrazovni sadržaji koji se provode na zajedničkom satu kao što su prezentacije, video zapisi i poveznice na vanjske sadržaje. Izrađene vježbe na nastavi u školi, kao i domaće zadaće, učenici predaju u Loomen-u u obliku zadaće. Za sve napravljene radove, učenici dobivaju povratne informacije, također u Loomen-u. U Loomen-u se provode i različita vrednovanja kao što su testovi, kvizovi i vršnjačka vrednovanja koja se bave različitim informatičkim zadacima i aktivnostima za učenje kao što su programiranje, prezentacije radova, tematska istraživanja i drugo.

B. Računalstvo

Učenici 1. razreda Prometno-tehničke škole Šibenik upisuju se u srednju školu s različitim razinama predznanja. Neki od njih nikad nisu bili uključeni u formalno učenje korištenja računala. Stoga nastavnica pažljivo odabire metode rada kako bi svi učenici mogli ravnopravno napredovati. Nastava iz Računalstva je organizirana u blok satovima koji se provode u specijaliziranoj informatičkoj učionici. U uvodnom dijelu nastavnog blok sata prezentira se novo gradivo koje se uvježbava kroz zajednički primjer nakon kojeg se provodi samostalan rad učenika ili rad u grupama. Većina učeničkih aktivnosti obavlja se u školi, dok se kod kuće dodatno oblikuje nedovršeno. Kako bi učenicima njihovi radovi bili dostupni u školi i kod kuće spremaju ih na Google Drive koji se uz ostale Google alate pokazao kao najbolji izvor radne platforme jer tu platformu učenici najčešće koriste i izvan nastave u školi.

Kroz reformu *Škole za život* nastavnici su osposobljeni za korištenje sustava za online učenje (Loomen) pa je i taj oblik edukacije prezentiran učenicima. Međutim, nije se pokazao kao najbolji izbor. Među učenicima su se pojavile velike razlike u ostvarenosti ishoda. Manjina učenika bi obavila sve zadatke kvalitetno i u rokovima. Pojedini učenici koji nisu uspijevali samostalno riješiti probleme odustajali bi bez traženja dodatne pomoći. Dio učenika nije aktivno sudjelovao u radu pravdajući se nedostatkom adekvatnih uvjeta za rad. Takvo iskustvo pokazalo je da online nastava kroz sustav za učenje Loomen nije dovoljno učinkovit u učioničkoj nastavi Prometno-tehničke škole Šibenik te da učenici ove škole najbolje rezultate u učenju postižu u klasičnoj učionici uz vodstvo i podršku nastavnice.

III. NASTAVA NA DALJINU

A. Informatika

16. ožujka 2020. godine kao posljedica pandemije virusa COVID 19 nastava je u Republici Hrvatskoj prebačena u model nastave na daljinu.

Kao središnje mjesto za realizaciju nastave na daljinu i timski rad s učenicima odabran je Microsoft Teams. Osim u tom alatu nastavnici su i dalje imali mogućnost koristiti druge alate za realizaciju nastave na daljinu pa je tako za nastavu informatike nastavljeno korištenje LMS alata Moodle (Loomen). Za potrebe online nastave za svaki razred je u Loomen-u napravljen zaseban odjeljak u kojem su se objavljivali sadržaji za učenje, vježbanje i ponavljanje gradiva. Sadržaji i aktivnosti za učenike su se odvijali prema tjednom rasporedu sati koji je sastavljen posebno za potrebe realizacije nastave na daljinu. Za vrednovanje nastave na daljinu izrađen je poseban plan rada koji je sadržavao popis aktivnosti, način bodovanja aktivnosti i način izvođenja ocjene za pojedine aktivnosti. Na kraju nastavnih aktivnosti provedena je evaluacija nastave na daljinu.

B. Računalstvo

Kao i u ostalim školama Republike Hrvatske 16. ožujka 2020. godine je zbog pandemije virusa COVID-19 i u Prometno-tehničkoj školi Šibenik napravljen prelazak na nastavu na daljinu. Svi učenici i nastavnici su se, bez prethodnih priprema, našli u sustavu Microsoft Teams. Kroz dva dana svi su se morali prilagoditi novim uvjetima rada. U početku je bilo nedoumica, i učenici i nastavnici su se pitali da li su spremni za takav oblik nastave i kako će se u njemu snaći. Prvi tjedan bilo je i tehničkih problema, zbog preopterećenosti u vremenu od 8 do 18 sati bilo je otežano prenijeti datoteku u sustav ili je preuzeti. Radilo se 24 sata na dan tražeći period kad će sustav pravovaljano funkcionirati. Ubrzo su poteškoće otklonjene i svi su se prilagodilo novonastaloj situaciji.

Sustav Microsoft Teams je omogućio formiranje razreda i predmetnih kanala. U predmetnom kanalu nastavnik ostavlja učenicima radne materijale i komunicira s cijelim razredom, a u odjeljku *Zadaci* se objavljuju učeničke obaveze koje se moraju napraviti u zadanom vremenskom roku. Za svaku predanu zadaću učenik dobiva od nastavnika povratnu informaciju. Komunikacija s učenicima se odvijala kroz privatni chat ili kroz obavijesti u razrednom kanalu. Obavijesti uz navođenje posebnih oznaka (npr: @) mogu biti navedene kao individualne ili razredne što stvara dodatnu pogodnost u komunikaciji s učenicima. Prednost alata Microsoft Team je mogućnost povezivanja s drugim office365 alatima kao što su SharePoint (za dijeljenje materijala i/ili zajednički rad na dokumentu), video poziv i/ili online sastanak (za komunikaciju s učenicima uz mogućnost dijeljenja sadržaja zaslona nastavnika i/ili učenika) te alat Forms (za izradu online obrazaca za samoprocjenu i kao procjena drugih učenika).Online obrasci su se na kraju učenja koristili i za vrednovanje naučenog.

U sustavu nastave na daljini svi učenici se nisu ponašali kao u realnoj učionici. Pojedinci, koji su u realnoj učionici redovito izvršavali svoje obaveze, povremeno bi nestali u online okruženju i trebalo ih je poticati i voditi, a bilo je i onih koji su bili uspješniji nego u realnoj učionici. Nedostatak ovakvog oblika edukacije u predmetu Računalstvo u Prometno-tehničkoj školi Šibenik je bio što nije unaprijed definirano vrijeme kad je nastavnica dostupna učenicima pojedinog razreda, te je svaki pojedinac očekivao odgovor/povratnu informaciju neposredno nakon izvršene aktivnosti ili postavljenog upita. To je stvaralo dodatno opterećenje nastavnici. Uz mnogo uloženog truda učenika i nastavnice, stalne podrške i neograničenog strpljenja nastavnice, završili smo nastavnu godinu s nastavom na daljinu i ostvarili

gotovo iste rezultate kao i prethodnih godina u modelu učioničke nastave.

IV. EVALUACIJA NASTAVE NA DALJINU

A. Informatika

U evaluaciji nastave na daljinu iz informatike sudjelovala su 3 prva razreda, 2 druga i jedan 4 razred, sveukupno 107 učenika. Pitanja na koja su odgovarali bila su razvrstana u dvije grupe: opća pitanja i pitanja vezana za online nastavu. Na pitanje u kojoj mjeri preferiraju on line pristup nastavi u odnosu na klasični pristup na skali od 1 do 5 8% učenika je navelo ocjenu 1, 12% učenika ocjenu 2, 29% učenika ocjenu 3, 24% učenika ocjenu 4 i 23% učenika ocjenu 5 što je rezultiralo prosječnom ocjenom 3.43.

Na pitanje koji način predstavljanja gradiva im je najviše odgovarao u vrijeme nastave na daljinu, učenici su odgovorili da im najviše odgovarao samostalni rad na materijalima, slijede snimljena predavanja, a zatim predavanja u realnom vremenu.

TABLICA I. PREDSTAVLJANJE GRADIVA

	Postotak odgovora
Samostalni rad na materijalima	46%
Snimljena predavanja	30%
On line predavanja /susreti u stvarnom vremenu	18%
Ostalo	6%

Na pitanje što misle o količini gradiva koja se obrađivala u vrijeme nastave na daljinu i količini zadataka i zadaća koje su dobivali, učenici su u velikom postotku odgovorili da im je količina gradiva i zadataka bila optimalna.

TABLICA II. KOLIČINA GRADIVA KOJA SE OBRAĐIVALA U VRIJEME NASTAVE NA DALJINU

	Postotak odgovora
Količina gradiva bila je optimalna	73%
Obrađivalo se malo više gradiva nego inače	15%
Obrađivalo se malo manje gradiva nego inače	8%
Obrađivalo se previše gradiva	4%
Obrađivalo se premalo gradiva	1%

TABLICA III. KOLIČINA ZADATAKA KOJE SU UČENICI DOBIVALI U VRIJEME NASTAV E NA DALJINU

	Postotak odgovora5
Dobivali su premalo zadataka	0%
Mogli su dobivati malo više zadataka	6.5%
Dobivali su optimalnu količinu zadataka	71,02%
Mogli su dobivati malo manje zadataka	16,8%
Dobivali su previše zadataka	5,6%

Za pitanja vezana za jasnoću gradiva, prikladnost zadataka i njihovu pravovremenu dostavu na skali od 1 do 5 dobivene su sljedeće prosječne ocjene:

TABLICA IV. JASNOĆA GRADIVA, PRIKLADNOST ZADATAKA I ISPORUKA SADRŽAJA U VRIJEME NASTAVE NA DALJINU

	Prosječna ocjena na skali od 1 do 5
Gradivo je prezentirano na jasan i razumljiv način	3,73
Materijali koje su dobivali bili su korisni i upotpunili su gradivo koje su učili	4,00
Zadaci koje su dobivali su bili prikladni i doprinijeli su boljem razumijevanju gradiva	4,03
Zadaci su bili pravovremeno dostavljeni na dogovorenoj platformi	4,67

Za dobivanje povratne informacije za tražena objašnjenja od učenika je dobivena prosječna ocjena 4.12, a za kvizove kao pomoć pri učenju u modelu nastave na daljinu prosječna ocjena 3.9.

Na pitanja o postavljenim rokovima za izvršenje zadaća i o primjeni Loomna kao alata za udaljeno učenje učenici su odgovorili:

TABLICA V. ROKOVI ZA IZVRŠENJE ZADAĆA

	Postotak odgovora
Rokovi su bili prekratki	2%
Rokovi su mogli biti duži	8%
Rokovi su bili optimalni	89%
Rokovi su mogli biti kraći	1%
Rokovi su bili predugi	0%

Na pitanja o korištenju Loomena dobiveni su sljedeći odgovori:

TABLICA VI. PITANJA O KORIŠTENJU LOOMEN-A

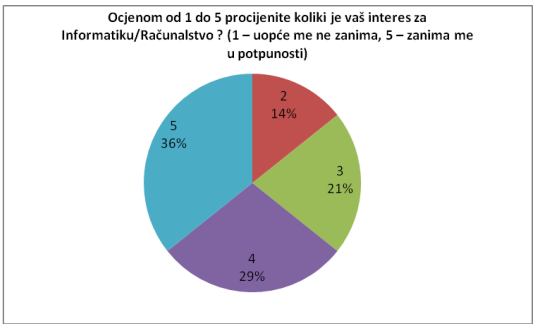
	u potpunosti se slažem	djelomično se slažem	ne slažem se
Znanje stečeno o Loomenu u vrijeme nastave pomoglo mi je u vrijeme online nastave	67,3%	29,9%	2,8%
Digitalni sadržaji u Loomenu mi pomažu u učenju	56,1%	39,3%	4,7%
Stalan pristup materijalima općenito smatram prednošću za moje učenje	82,2%	16,8%	0,9%
Kada dođem na nastavu u školu želio bih nastaviti koristiti Loomen	66,4%	29%	4,7%
Smatram da ću budućnosti više primjenjivati online učenje	30,2%	50%	19,8%

Na pitanja o vrsti materijala koji su im najviše koristili u Loomen-u učenici su u najvećem broju odabrali prezentacije i vježbe, zatim kvizove i video materijale i na kraju zadaće.

B. Računalstvo

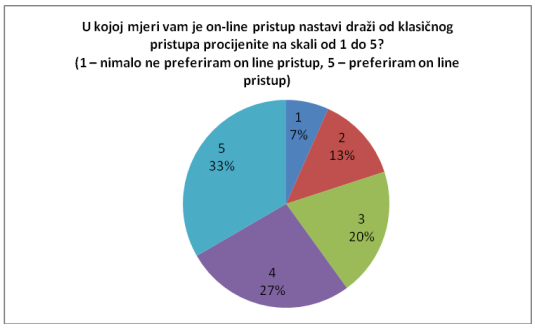
Ocjena nije jedini pokazatelj uspješnosti, zato je u Prometno-tehničkoj škole Šibenik proveden upitnik o zadovoljstvu učenika u online učionici.

Upitnik je popunilo 82 učenika 1. i 2. razreda koji slušaju predmet Računalstvo. Učenici su svoj interes za nastavni predmet računalstvo ocijenili s prosječnom ocjenom 3.62 ili u postotcima kao na Slici 1.



Slika 1.

Preferencije prema online nastavi ocijenili su s prosječnom ocjenom 3.15, a Slika 2 prikazuje raspodjelu ocjena u postotcima



Slika 2.

Kako su učenici Prometno-tehničke škole odgovorili na pitanja vezana za predstavljanje gradiva vidljivo je u sljedećoj tablici.

TABLICA VII. PREDSTAVLJANJE GRADIVA

	Postotak odgovora
Samostalni rad na materijalima	53%
Snimljena predavanja	18%
On line predavanja /susreti u stvarnom vremenu	11%
Ostalo	18%

Količina gradiva i zadataka koje je bilo optimalno isporučiti učenicima u vrijeme nastave na daljinu predstavljala je poseban izazov. Što učenici misle i kako su se izjasnili o količini gradiva i zadataka koje su dobivali u vrijeme nastave na daljinu vidljivo je tablicama VIII i IX..

TABLICA VIII. KOLIČINA GRADIVA KOJA SE OBRAĐIVALA U VRIJEME NASTAVE NA DALJINU

	Postotak odgovora
Količina gradiva bila je optimalna	42%
Obrađivalo se malo više gradiva nego inače	29%
Obrađivalo se malo manje gradiva nego inače	5%
Obrađivalo se previše gradiva	26%
Obrađivalo se premalo gradiva	0%

TABLICA IX. KOLIČINA ZADATAKA U VRIJEME NASTAVE NA DALJINU

	Postotak odgovora
Dobivali su premalo zadataka	0%
Mogli su dobivati malo više zadataka	0%
Dobivali su optimalnu količinu zadataka	41%
Mogli su dobivati malo manje zadataka	37%
Dobivali su previše zadataka	22%

Jasnoća gradiva, prikladnost zadataka i pravovremenost dostave na dogovorenu platformu učenici su ocijenili na skali od 1 do 5 te su dobivene prosječne ocjene koje su prikazane u sljedećoj tablici

TABLICA X. JASNOĆA GRADIVA, PRIKLADNOST ZADATAKA I ISPORUKA SADRŽAJA U VRIJEME NASTAVE NA DALJINU

	Prosječna ocjena na skali od 1 do 5
Gradivo je prezentirano na jasan i razumljiv način	3,43
Materijali koje su dobivali bili su korisni i upotpunili su gradivo koje su učili	3,48
Zadaci koje su dobivali su bili prikladni i doprinijeli su boljem razumijevanju gradiva	3,94
Zadaci su bili pravovremeno dostavljeni na dogovorenoj platformi	4,00

Na pitanje o vremenskim rokovima koje su imali na raspolaganju za izvršavanje zadaća učenici su se izjasnili u postotcima prikazanim u sljedećoj tablici.

TABLICA XI. ROKOVI ZA IZVRŠENJE ZADAĆA

	Postotak odgovora
Rokovi su bili prekratki	6%
Rokovi su mogli biti duži	35%
Rokovi su bili optimalni	58%
Rokovi su mogli biti kraći	0%
Rokovi su bili predugi	1%

Mišljenje učenika o alatu Microsoft Teams razvidno je iz tablice koja slijedi:

TABLICA XII. PITANJA O KORIŠTENJU MS TEAMS

	u potpunosti se slažem	djelomično se slažem	ne slažem se
Za vrijeme nastave u realnoj učionici upoznao sam se s načinom korištenja sustava MS Teams za udaljno učenje	42,7%	41,5%	15,8%
Digitalni sadržaji u MS Teams-ima mi pomažu u učenju	35,3%	48,8%	15,9%
Stalan pristup materijalima općenito smatram prednošću za moje učenje	43,8%	37,6%	18,6%
Kada se vratim na nastavu u školu u realnu učionicu želio bih nastaviti koristiti MS Teams	26,8%	39%	31,7%
Smatram da ću budućnosti više primjenjivati online učenje	31,7%	47,6%	20,7%

Među aktivnostima u Microsoft Teams-u učenici ističu kao najkorisnije prezentacije, zatim zajednički chat nastavnice i učenika cijelog razreda, te individualni chat. Najmanje korisnim su procijenili vršnjačko vrednovanje (ocjenjivanje tuđih radova).

Među onim što bi pohvalili ističe se dostupnost nastavnice i pravovremeno davanje povratnih informacija, a kao nedostatak da ne preferiraju nastavu na daljinu i da bi ipak radije bili u klasičnoj učionici.

V. ZAKLJUČAK

Nastava na daljinu koje se ostvarivala u vrijeme pandemije virusa COVID-19 pokazala je sve prednosti, ali i moguće nedostatke u odnosu na učioničku nastavu.

Prednosti koje se ističu su: fleksibilnost u učenju, dostupnost materijala za učenje, organizacija učenja prema vlastitom rasporedu, razvijanje samostalnosti u učenju.

Kao mogući nedostatci koji mogu negativno utjecati na tijek i kvalitetu nastave na daljinu su: neadekvatna količina gradiva (premalo ili previše), nerazumni rokovi za izvršenje zadataka, nedostatak objašnjenja, nedostatak rasporeda (plana) rada te istovremeni rad na više platformi.

Uspoređujući rezultate upitnika provedenog u ove dvije različite škole (gimnazija i strukovna škola), nameće se zaključak da su se svi učenici aktivno sudjelovali u nastavi na daljinu, ali da im je ipak draže i prirodnije okruženje učioničke nastave, gdje imaju nastavničku podršku u učenju.

LITERATURA

[1] T. Beats, "Upravljanje tehnološkim promjenama: strategije za voditelje visokih učilišta,", CARNet, 2004; preveli Srećko Jelušić, Sanjica Faletar.

[2] I. Blažević, "Suvremeni kurikul i kompetencijski profil učitelja.", Zbornik radova Filozofskog fakulteta u Splitu, 2016. (str. 189-210). Dostupno na https://hrcak.srce.hr/154590, [pristupljeno 11.7.2020.]

[3] M. Bulić, V. Kostović Vranješ, "Utjecaj e-učenja na samoodgovornost učenika pri izvršavanju domaćih zadaća", Školski vjesnik 2019.,(str.112-126). Dostupno na https://hrcak.srce.hr/230624, [pristupljeno 11.7.2020.]

[4] Institut za društvena istraživanja u Zagrebu: Pilot istraživanje učeničkih potreba i suočavanja s izazovima online nastave u ožujku 2020. godine. Dostupno na https://www.idi.hr/ciro-idiz-za-online-skolu/ , [pristupljeno 11.7.2020.]

[5] Grafički prikaz rezultata ankete za V. gimnaziju Zagreb. Dostupno na https://bit.ly/3gNSkfX, [pristupljeno 11.7.2020.]

[6] Grafički prikaz rezultata ankete za Prometno-tehničku školu Šibenik. Dostupno na https://bit.ly/3elXzBX, [pristupljeno 14.7.2020.]

[7] J. Pregrad: Stres, trauma, oporavak: udžbenik programa "Osnove psihosocijalne traume i oporavka" / urednik Jasenka Pregrad; [ilustracije Biserka Tomljenović Belicza] -Zagreb: Društvo za psihološku pomoć, 1996.

CTS

International Conference on

COMPUTERS IN TECHNICAL SYSTEMS

Steering Committee

Chairs: Vlado Sruk, University of Zagreb, Croatia
Dejan Škvorc, University of Zagreb, Croatia
Jadranko F. Novak, Nomen, Rijeka, Croatia

Program Committee

Members: Željko Hocenski, University of Osijek, Croatia
Leonardo Jelenković, University of Zagreb, Croatia
Hrvoje Mlinarić, University of Zagreb, Croatia
Jadranko F. Novak, Nomen, Rijeka, Croatia
Vlado Sruk, University of Zagreb, Croatia
Dejan Škvorc, University of Zagreb, Croatia

PAPERS

ROBOTICS AND AUTOMATION

Application of PandaPower Tool in Evaluating the Potential of Using PV Distributed Generation for Voltage Regulation in Electrical Power Networks

N. Bogunović*, S. Vlahinić**, D. Franković** and V. Komen*
* HEP Distribution System Operator, Elektroprimorje Rijeka, Rijeka, Croatia
** University of Rijeka, Faculty of Engineering, Rijeka, Croatia
nikola.bogunovic@hep.hr, sasa.vlahinic@riteh.hr, dubravko.frankovic@riteh.hr, vitomir.komen@hep.hr

***Abstract* - Traditional electrical power systems were designed with an assumption of unidirectional power flows through electrical networks. For many years power plants had been the only sources of electrical energy, and the energy had been delivered to the consumers by the means of transmission and distribution networks exclusively.**

With the emergence of distributed generation units, power flows have become bidirectional, which has made it impossible to estimate the voltages and currents in substation feeder nodes by measuring only the values at the beginning of the feeder. As the aforementioned estimation process is the basis of traditional voltage regulation techniques, it is obvious that under the new circumstances existing methods have to be upgraded.

The article demonstrates how the open source tool PandaPower along with the Python programming language was utilized to analyze and evaluate several different scenarios of using photovoltaic distributed sources inverters for voltage regulation in low voltage electrical power networks.

Keywords – PandaPower; voltage regulation; voltage stability; photovoltaic; distributed generation

I. INTRODUCTION

The emergence of distributed generation units has transformed the traditional electrical power distribution networks from passive to active. Power flows through the networks have become bidirectional, which has added more complexity to the process of bus voltage estimation. Since the aforementioned estimation process is the basis for traditional voltage regulation techniques, it is obvious that, under the new circumstances, these techniques have to be upgraded.

As the inverters used in photovoltaic (PV) distributed generation (DG) units are designed to operate with different power factors, they can be utilized to support voltage regulation in the networks [1].

The aim of this article is to analyze the influence of reactive power injected by DG units on voltage stability in a typical Croatian low voltage (LV) electrical distribution network.

The analysis was performed with the PandaPower package for the Python programming language, since it is a scriptable electrical power networks calculation tool.

II. VOLTAGE REGULATION IN ELECTRICAL POWER NETWORKS

A. Voltage Drop on LV Distribution Lines

Typical low voltage electrical distribution line can be approximated with a series connection of resistance R and reactance X, as it is shown in Figure 1.

The voltage difference $\Delta\bar{V}$ along the line can be expressed as

$$\Delta\bar{V} = \bar{V}_1 - \bar{V}_2 = \bar{I}R + j\bar{I}X, \tag{1}$$

where $\bar{I}$ is the current flowing through the line and $\bar{V}_1$ and $\bar{V}_2$ are the voltages at the beginning and the end of the line respectively.

Equation (1) can be transformed to take the active power P_1 and reactive power Q_1 transferred by the line into account:

$$\Delta\bar{V} = \frac{P_1R + Q_1X}{V_1} + j\frac{P_1X - Q_1R}{V_1}. \tag{2}$$

Since the R/X ratio in LV distribution networks is relatively high (usually above 10) [1], the imaginary part of (2) is usually neglected. Therefore, the approximate equation for voltage drop along the LV line can be stated as:

$$\Delta V \approx \frac{P_1R + Q_1X}{V_1}. \tag{3}$$

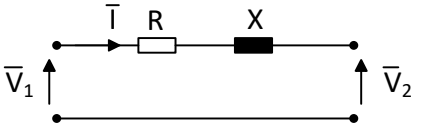


Figure 1. Low voltage distribution line model

As the higher voltage drops are correlated to the higher power losses in distribution networks, it is desirable to keep the voltage drops as low as possible, and that is where the voltage regulation comes in place.

From (3), it is easily observed that the voltage drop can be reduced in three ways: by using conductors with lower impedance levels, by raising the voltage at the beginning of the line, or by reducing the reactive power flow. Active power flows are usually dictated by consumers' demand and they are in general not easy to control.

The aim of voltage regulation is to keep the voltages in all nodes of a network as close as possible to the nominal voltage of a feeder, thus reducing the overall power losses of the network.

B. *Traditional Voltage Regulation Techniques*

Traditional voltage regulation techniques were designed with an assumption of unidirectional power flows (i.e. from power plants to consumers) in electrical networks. In such networks it is relatively easy to estimate the voltages and currents in all nodes on a feeder by measuring the values at its beginning. As the loads in distribution networks are mainly inductive, the voltage drops along a feeder and reaches the lowest value at its end.

In medium voltage (MV) distribution networks, voltage is commonly regulated by installing transformers with automatic tap changers in HV/MV and MV/MV substations. The tap changer keeps the voltage on the secondary coil of a transformer high enough to compensate for voltage drops in a feeder, i.e. to keep all the voltages along the feeder within allowable tolerances.

Similar technique is used in LV distribution networks, with the difference that the tap changers are usually manual, so they will have to be switched by the operator once the conditions on a feeder change.

Along with regulation transformers, shunt capacitor banks for reactive power compensation are usually installed in substations and in industrial plants.

C. *Distributed Generation as a Voltage Stability Support*

With the emergence of DG units, power flows have become bidirectional, and under that circumstances traditional voltage regulation techniques have to be upgraded.

The most common DG type connected to LV networks in Croatia is photovoltaic (PV) generation, which makes almost 86% of total DG installed power [2]. As the PV inverters can easily operate with different power factors, they can be used to support voltage stability by either injecting (if the voltage is too low) or absorbing (if the voltage is too high) the reactive power from a connection point in LV network [3].

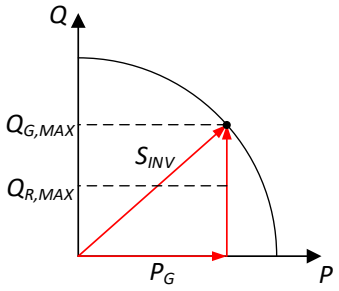


Figure 2. PV inverter reactive power generation limits

Maximum amount of reactive power $Q_{G,MAX}(t)$ that the PV inverter can produce at certain moment t is limited by its rated apparent power S_{INV} and the active power $P_G(t)$ generated at the same moment [4]:

$$Q_{G,MAX}(t) = \sqrt{S_{INV}^2 - P_G^2(t)}. \quad (4)$$

Furthermore, the produced reactive power level is usually limited by country specific regulations, i.e. the Distribution Network Grid Code. In Croatia, all grid-connected generation units must support the operation with a power factor from 0.9 to unity [5]. That requirement leads to the following maximal reactive power injection amount:

$$Q_{R,MAX}(t) \approx 0.484 \cdot P_G(t). \quad (5)$$

Therefore, the maximum theoretical amount of reactive power that can be produced by PV generation unit inverter is given by [6]:

$$Q_{MAX}(t) = \min\left[Q_{G,MAX}(t), Q_{R,MAX}(t)\right]. \quad (6)$$

The limiting values defined by (4) and (5) are shown in Figure 2 [6].

D. *PV Inverter Reactive Power Control Approaches*

There are different approaches to controlling the amount of reactive power produced by PV inverters. Some of the most common are [7,4]:

1. Reactive Power Management (RPM)
2. Power Loss Minimization (PLM)
3. Voltage Rise Minimization (VRM)
4. Power Curtailment Strategy (PCS)

In this article the RPM strategy is analyzed, since it does not require any communication infrastructure for transferring measurement data from network nodes to a centralized controller.

Basically, the RPM strategy is based upon a predefined relation between PV inverter power factor and LV bus voltage, which is stored in the PV inverter controller. The



Figure 3. Typical reactive power control characteristic [8]

inverter controller sets the required power factor according to the voltage measurements at the DG point of connection, and the inverter produces the reactive power which supports the voltage stability [7,4].

A typical RPM $Q(V)$ relation shape used for small PV distributed generation units is shown in Figure 3.

E. Test Cases Comparison Methodology

It is usually not possible to minimize voltage drops and power losses at the same time [7]. Therefore, a modified maximum voltage deviation and loss reduction index introduced in [9] was used to compare different test cases and scenarios.

Maximum voltage deviation of a network with N_n nodes can be calculated from the equation:

$$\Delta V_D = \max\left[\frac{|V_k - V_1|}{V_1}\right], \quad k = 1 \dots N_n. \tag{7}$$

The total power loss of a network with N_b branches is given by:

$$P_L = \sum_{i=1}^{N_b} P_{L,i}, \tag{8}$$

where $P_{L,i}$ is the power loss of branch i.

If P_{L0} denotes the total power loss prior to making changes in a network (e.g. by adding DG units or injecting reactive power), and P_{LC} denotes the total power loss after the specific change in the network is applied, then the loss reduction index can be calculated as follows:

$$\Delta P_{LR} = \frac{P_{LC} - P_{L0}}{P_{L0}}. \tag{9}$$

The total index for test case evaluation is calculated as a summation of (7) and (9):

$$F = \Delta V_D + \Delta P_{LR}. \tag{10}$$

The lower value of F indicates the lower voltage deviation and the lower total power loss, with a theoretical minimal value of -1.

III. Description of the Model Considered

A. Characteristic LV Network without DG Units

The analysis has been performed on a characteristic Croatian LV distribution network model, created from the available data [1,10]. The network model diagram is shown in Figure 4.

Characteristic LV distribution network is connected to the MV network over a 400 kVA 20/0.4 kV transformer. The substation has 4 feeders which supply 86 households with the simultaneous peak power of 1.25 kW and 9 business consumers with the simultaneous peak power of 9.9 kW. Power factor for all consumers in the network is assumed to be 0.9 inductive.

The first feeder is built with underground cables, the second one is a combination of underground cables and overhead lines, while the third and the fourth ones are

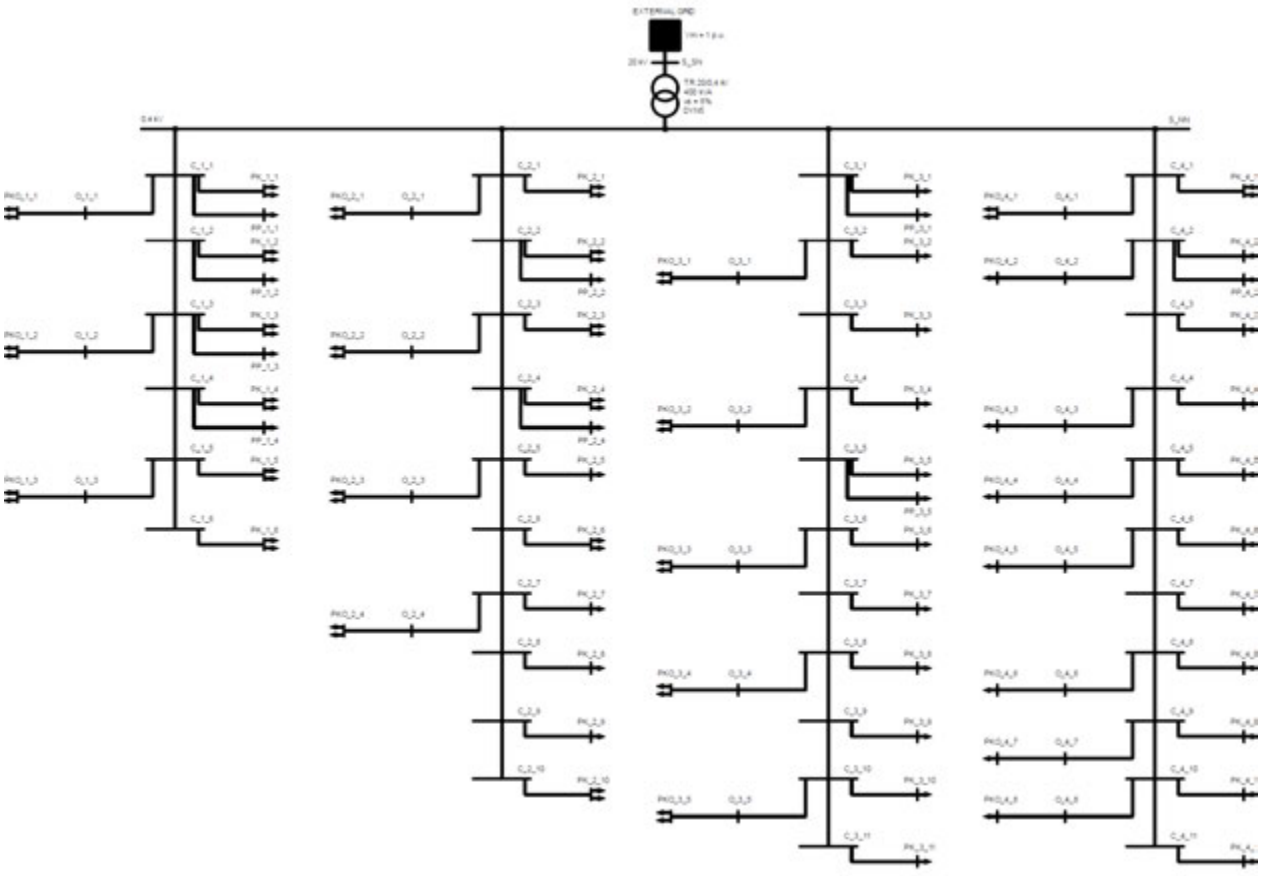


Figure 4. Characteristic LV distribution network in Croatia

entirely built with overhead lines. Total length of underground cables in the network is 1,187 m and total length of overhead lines is 2,690 m.

Without connected DG units the LV network draws 216.68 kW and 110.95 kVAr from MV network. Active and reactive power losses are 7.28 kW and 9.57 kVAr respectively. Maximum voltage drop equals 6.25 p.u.

B. *Distributed Generation Units Penetration Levels*

According to available data [2], installed capacity of PV generation in LV networks in Croatia equals 52,080 kW. As the number of MV/LV transformers is 26,973 with the total installed capacity of 8,954 MVA, it can be shown that the average installed capacity of PV equals 0.58% of installed capacity of a substation transformer.

Similarly, from the electrical energy production and consumption data [10], it can be derived that PV distributed generation averagely covers around 4.14% of typical LV electrical network energy requirements.

Therefore, the analysis was performed for five different PV distributed generation penetration levels: 2%, 5%, 10%, 15% and 20%.

IV. USING PANDAPOWER TOOL FOR LV NETWORK POWER FLOW AND VOLTAGE DROPS CALCULATION

A. *PandaPower Tool*

PandaPower is an open source tool written in Python which has the built-in capabilities for modelling, analysis and optimization of electrical power systems [11].

It is built as a Python module, and it has several advantages and a few disadvantages when compared to other electrical power systems analysis tools. The main disadvantage is that it is generally harder do enter a network model manually because the tool does not have any graphical user interface, so every single network element has to be entered by using the Python programming language.

The scripting capabilities of the tool make its greatest advantage, because it is easily linked with existing geographical and technical information systems, so it is theoretically possible to avoid entering the network data completely.

PandaPower supports the following electrical power systems analysis modules [11]:

- Power flow
- Optimal power flow
- State estimation
- Short circuit calculation

Calculation results can be analyzed and processed with Python and then exported to several spreadsheet applications, where they can be furtherly formatted and processed.

B. *Basic Alghoritm for Data Analysis and Processing*

The main loop of data analysis and processing algorithm is shown in Figure 5. Generally, it creates the test network and then performs the analysis on all test cases. The results are evaluated and stored for later comparison.

As it is depicted in Figure 5, PV generation units are sequentially added to the test network, and therefore the test results are easily comparable. Every PV generation unit added to the test network is defined by two parameters – its location and its installed capacity.

To simulate the inability of a distribution system operator to predict the location requirements of their customers' distributed generation units, the location of every new unit is chosen as a random consumer node in the test network.

The installed capacity of a DG unit is calculated from a normal distribution with the mean value of 6 kW and the standard deviation of 3 kW.

Figure 6 shows the sub-process of adding DG units to the test network. When the total DG installed capacity exceeds the maximum capacity for a test case (defined by

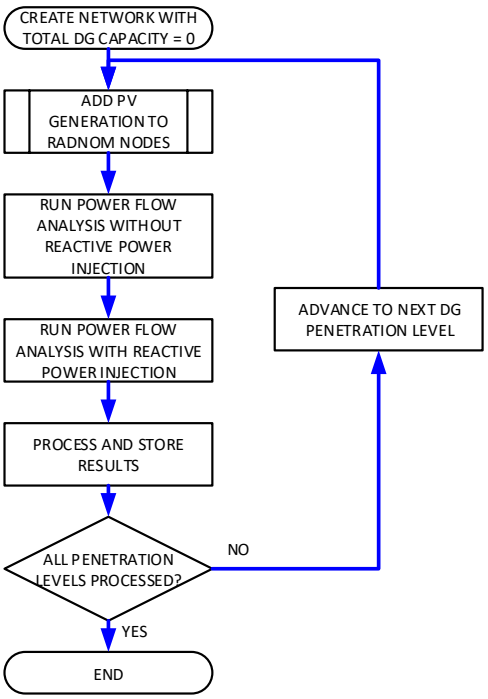


Figure 5. The main loop of data analysis and processing algorithm

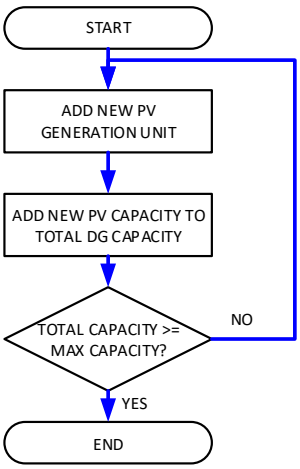


Figure 6. The sub-process for DG unit creation

DG penetration level), the sub-process ends and the results are analyzed and evaluated by the main loop.

V. Simulation Results

Table I displays the installed DG units' capacities and their points of connection in the test network for different DG penetration levels in a single run of the algorithm shown in Figure 5. Each of the higher level test cases keeps all of the installed PV units from lower penetration levels and adds one or more new units, as shown in the table.

As the area of interest in this research has been to analyze benefits the PV distributed generation units have on voltage stability, the calculations were performed with an assumption of high irradiance levels (i.e. during a typical sunny day).

The results of the calculations are summarized in Table II. For each DG penetration level, two different sets of results are displayed – one with DG units producing only active power, and the other with reactive power injection, in accordance with RPM characteristic and limitations described in sections III.C and III.D. The test cases are compared by their respective *F*-index values, which are calculated by the means of (10).

Figures 7 and 8 show the voltage profiles for the lowest (case No. 1) and the highest (case No. 5) DG penetration levels compared to the base LV network without connected DG units (case No. 0).

The results indicate that PV inverters have the potential to reduce voltage drops and power losses in electrical distribution networks, provided that the DG penetration level is high enough.

Table I. PV units installed capacities and locations

DG penetration level	Node number	DG unit capacity [kW]
0	-	0.00
1	30	8.44
2	104	10.44
3	33	5.69
	64	7.13
4	12	4.64
	94	2.48
	120	2.54
5	7	9.30
	92	6.35

Table II. Calculation results summary

DG penetration level	Total DG installed capacity [kW]	Total DG reactive power injection [kVAr]	ΔV_D [%]	P_{loss} [kW]	Q_{loss} [kVAr]	*F*-Index
0	0.00	0.00	6.25	7.28	9.57	0.063
1	8.44	0.00	6.22	7.02	9.01	0.026
		3.52	6.16	6.96	8.88	0.017
2	18.88	0.00	5.23	6.41	8.27	-0.068
		7.87	5.12	6.22	8.00	-0.094
3	31.70	0.00	4.79	5.85	7.48	-0.149
		13.21	4.60	5.57	7.07	-0.190
4	41.37	0.00	4.75	5.51	6.92	-0.196
		17.23	4.51	5.17	6.41	-0.245
5	57.02	0.00	4.69	5.10	6.11	-0.253
		23.75	4.37	4.68	5.46	-0.313

Since the installed DG capacities in Croatian LV networks are still relatively low, PV inverters usually may inject the maximum possible amounts of reactive power without causing overvoltage in the network. Except for the highest irradiance levels when the inverter's rated power is the limiting factor, these amounts are mainly dictated by Croatian Grid Codes (Equation 5).

In such circumstances, the best results could be achieved if the RPM approach is combined with one or more traditional voltage regulation techniques.

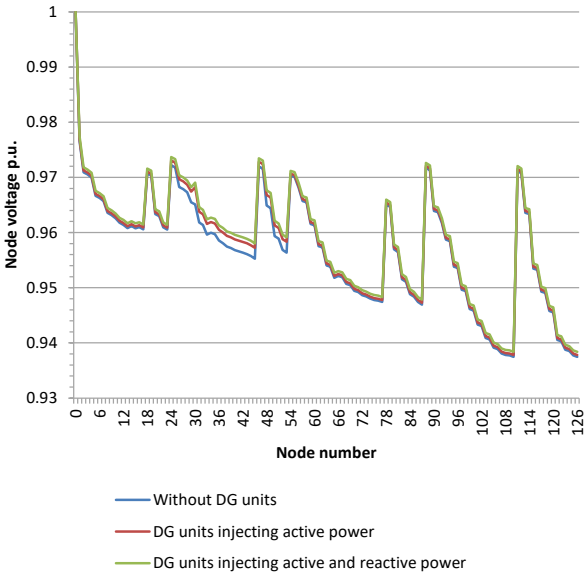


Figure 7. Voltage profiles for low DG penetration level

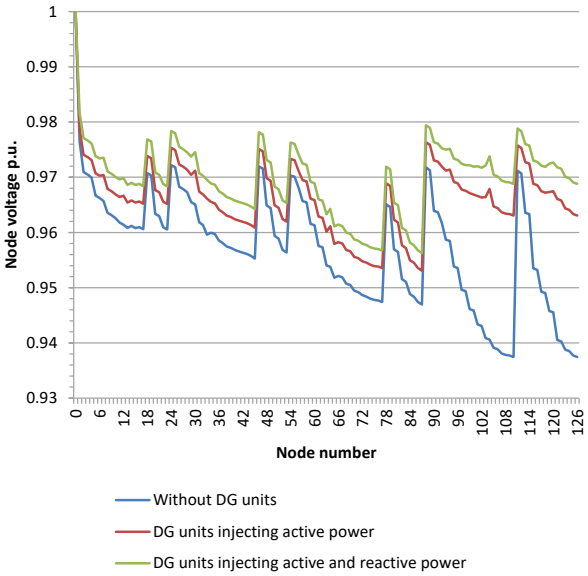


Figure 8. Voltage profiles for high DG penetration level

VI. Conclusion

The emergence of distributed generation units in LV electrical networks allow for some new approaches to voltage regulation and power loss reduction. Since the PV inverters are able to operate with different power factors,

they make a good addition to traditional voltage regulation techniques.

When the DG penetration levels are relatively low, inverter controllers may be set to inject the maximum possible amounts of reactive power to support voltage stability in the network without causing overvoltage in the network buses.

However, for higher penetration levels some kind of controller regulation should be established to keep the voltages in allowable limits.

The article demonstrates an example of using the PandaPower tool to simulate and analyze the influence of reactive power injection from PV distributed generation controllers on voltage stability in a typical LV distribution network in Croatia. Since the PandaPower tool could be easily scripted through the Python programming language, it makes a perfect tool for analyzing and comparing many different test case scenarios.

The algorithm presented in the article is relatively simple, but it could be effortlessly upgraded to simulate and optimize the reactive power injection to achieve the minimal voltage deviations and minimal power losses in the network.

REFERENCES

[1] S. Vlahinić, D. Franković, V. Komen, A. Antonić, "Reactive power compensation with PV inverters for system loss reduction", Energies, vol. 12, no. 21, pp. 4062, 2019.

[2] HEP ODS d.o.o., "Godišnje izvješće 2018", HEP Operator distribucijskog sustava d.o.o., Zagreb, 2019.

[3] O. Gandhi, C. D. Rodríguez-Gallegos, W. Zhang, D. Srinivasan, T. Reindl, "Economic and technical analysis of reactive power provision from distributed energy resources in microgrids", Applied energy, vol. 210, pp. 827-841, 2018.

[4] T. Stetz, F. Marten, M. Braun, "Improved low voltage grid-integration of photovoltaic systems in Germany", IEEE Transactions on sustainable energy, vol. 4, no. 2, pp. 534-542, 2012.

[5] HEP ODS d.o.o., "Mrežna pravila distribucijskog sustava", NN 74/2018

[6] M. Begović, A. Pregelj, A. Rohatgi, D. Novosel, "Impact of renewable distributed generation on power systems", Proceedings of the 34th annual Hawaii International conference on system sciences, IEEE, pp. 654-663, 2001.

[7] R. Kabiri, D. G. Holmes, B. P. McGrath, L. Gunaruwan, "LV grid voltage regulation using transformer electronic tap changing, with PV inverter reactive power injection", IEEE Journal of Emerging and Selected Topics in Power Electronics, vol. 3, no. 4, pp. 1182-1192, 2015.

[8] CENELEC, "Requirements for micro-generating plants to be connected in parallel with public low-voltage distribution networks", EN 50438:2013

[9] A. M. Imran, M. Kowsalya, "A new power system reconfiguration scheme for power loss minimization and voltage profile enhancement using fireworks algorithm", International Journal of Electrical Power & Energy Systems, vol. 62, pp. 312-322, 2014.

[10] B. Vuk et al., "Energy in Croatia – annual energy report", Republic of Croatia, Ministry of environment and energy, 2017.

[11] L. Thurner, A. Scheidler, F. Schäfer, J. H. Menke, J. Dollichon, F. Meier, S. Meinecke, M. Braun, "PandaPower – an open-source Python tool for convenient modelling, analysis, and optimization of electric power systems", IEEE Transactions on Power Systems, vol. 33, no. 6, pp. 6510-6521, 2018.

Resilience to cascading failures: a complex network approach for analysing the Croatian power grid

Zlatan Sičanica[1] and Ivan Vujaklija[2]

[1]Končar – Power Plant and Electric Traction Engineering Inc. Zagreb, Croatia (zlatan.sicanica@koncar-ket.hr)

[2]Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia (ivan.vujaklija@fer.hr)

***Abstract* – A transmission power grid system is arguably the single most important utility of any nation since it provides a basic infrastructure for a wide variety of interconnected and interdependent systems of vital importance. Therefore, its resilience to intentional attacks or random failures is of critical importance. This is especially true with respect to cascading failures which can be initiated by seemingly minor events but can lead to economic losses worth billions of dollars, as recent examples show. In order to gain insight into the resilience of the Croatian transmission power grid we investigated its resilience to three different types of failures, including both intentional attacks and random failures. Our analyses are based on the complex network approach. Extensive numerical simulations were carried out based on three different discrete-time cascading failure models. The first two models were described previously, whereas we substantially modified the third model.**

Keywords - cascading failure, network efficiency, transmission power grid resilience, complex network analysis

I. INTRODUCTION

A transmission power grid comprises a system of high-voltage transmission lines which connect power generating stations and distribution substations. Since electrical energy provides a basic infrastructure for a wide variety of interconnected and interdependent systems of vital importance (such as transportation and communication infrastructure, heating, gas and water supply to name just a few) power grids are arguably the most important infrastructural resource of any country. Power grid systems of developed countries are highly resilient and most power outages are of local character and short duration. However, although rare, large scale blackouts can cause tremendous economic damage and as such it is of vital importance to understand their dynamics.

One of the early examples of a major blackout is the New York, July 1977 blackout which lasted for hours and caused economic damages estimated to be well over $1 billion in today's money. Some of the more recent wide scale blackouts include the August 2003 blackout which affected large parts of Northeastern and Midwestern United States as well as Ontario (with an estimated cost of over $10 billion), the September 2003 Italian blackout (over $1 billion), the July 2012 Indian blackout ($6 billion) and the March 2015 Turkish blackout ($700 million) [1]. The key factor in such major blackouts are cascading failures which are characterized by initial failure which causes redistribution of power flow which triggers further overloads and failures which spread through the entire network or large parts of the network [1]. The causes of initial (trigger) failures vary from component failure (e.g. sagging power lines, Italy 2003 blackout), human error (e.g. US & Canada 2003 blackout), lightning strikes (e.g. New York 1977 blackout), to system failures (e.g. Indian 2012 blackout). Whatever the cause, it is clear that such large scale blackouts have tremendous economic impact.

During the past two decades numerous studies have analysed power grid systems from the complex network perspective ([2], [3], [8]). Among others, national power grid systems of US, Italy, Spain, France, UK, Germany, South Korea, Iran, India and China have been analysed using the complex network approach ([9], [2], [7], [10], [11], [3], [12], [13], [14]). Although modelling a highly complex power grid system as a simple undirected/directed, weighted/unweighted network, without taking into account their electrical and physical properties, is obviously a major simplification and as such has major limitations ([4], [5]), the simplicity of the network approach on the other hand makes the study of their complex dynamics (e.g. cascading failures) more tractable. Indeed, many highly complex systems have been successfully modelled as networks. Examples include metabolic networks, biochemical networks, protein interaction networks, ecological networks, transportation networks, internet, social networks, spread of epidemics and genetic regulatory networks ([5], [3]).

Two modelling approaches based on complex networks have been investigated in the literature: static analysis and dynamic analysis [11]. Static analysis is a simple approach where failure of a node has no influence on neighbouring nodes. Namely, nodes are removed one by one (either according to random failure or intentional attack model) and after each removal, the network efficiency and size of the *giant component* [5] are computed. A major limitation of the static models however is their inherent inability to model cascading failures. In dynamic analysis on the other hand only the initial node (or edge) is removed. After the initial removal, the system is left to behave according to its own dynamics. The goal of dynamic models is to capture the more realistic behaviour where the initial failure of a single node or edge causes a dynamic redistribution of loads throughout the network. This is in line with the more realistic cascading type of failures [7].

In this study we analysed the resilience of the Croatian Power Grid (CPG for short) to cascading failures based on the complex network analysis approach. Both random failures and targeted attacks on particular nodes were investigated using numerical simulations based on three different discrete-time dynamical models, two of which have been described previously ([7], [15]). The third model is based on [16] and [17], however we introduce an important modification.

II. METHODS AND RESULTS

In the complex network analysis approach the power grid is modelled as a graph $G = (N, E)$. Mathematically, a graph consist of a finite set of nodes N (usually called vertices in the mathematical literature) and a finite set of edges E, where each edge connects two nodes. In the power grid context, generators and distribution substations are represented by nodes and transmission lines by edges ([1, 5, 7]). Edges can be weighted or unweighted and directed or undirected depending on the particular model. We follow the same approach as in [7] and distinguish two types of nodes: generator nodes and distribution nodes. Generator nodes are sources of electrical energy whereas distribution nodes distribute electrical energy to end consumers and are thus "sinks of energy". According to this model there are 75 generator nodes and 190 distribution nodes in the CPG, which are shown by red and blue circles respectively (Fig.1).

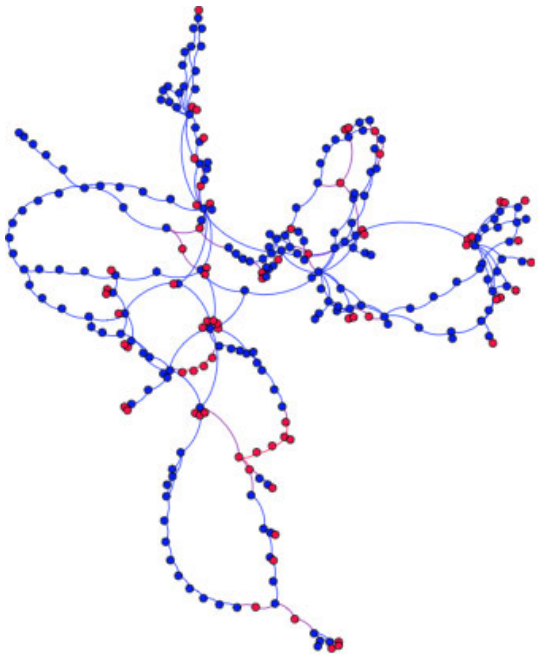


Fig. 1. The model of the Croatian Power Grid (CPG). Generators are represented by red circles and distribution stations by blue circles

Importantly, we assume that electricity is transmitted from all generators to all distribution nodes along the most efficient paths. The efficiency of the path between generator node i and distribution node j, ϵ_{ij} is defined simply as the inverse of their (shortest) distance, i.e. $\epsilon_{ij} = 1/d_{ij}$. Note that efficiency ϵ_{ij} and hence the distance d_{ij} will depend on the particular model being used. We measure the resilience of a network $\mathcal{N}$ by the *network efficiency* $E(\mathcal{N})$, defined as in [7]:

$$E(\mathcal{N}) = \frac{1}{2n_G n_D} \sum_{i \in N_G, j \in N_D} \frac{1}{d_{ij}} = \frac{1}{n_G \cdot n_D} \sum_{i \in N_G, j \in N_D} \epsilon_{ij} \quad (1)$$

Here n_G is the number of generator nodes, n_D is the number of distribution nodes, N_G is the set of all generator nodes, N_D is the set of all distribution nodes, d_{ij} is the distance (i.e. the most efficient path) between a generator node i and a distribution node j and ϵ_{ij} is the efficiency of this path. Note that only paths between generator nodes and distribution nodes are taken into account by the formula (1). We also scale (1) by $1/2$ since we double-count each edge (i.e. both ϵ_{ij} and ϵ_{ji}, are counted although they are equal and represent the same edge).

Generally speaking there are two possible node removal strategies, namely, *random removal*: whereby each node is equally likely to be removed and *targeted removal* where nodes are removed based on their "importance". Random removal simulates random failures whereas targeted removal simulates intentional attack scenarios [9]. Although several targeted removal strategies for both node removal and edge removal (e.g. [14]) have been investigated we follow two standardly used targeted attack approaches based on a node's *degree* and a node's *betweenness centrality* [16]. Node degree (also called degree centrality) is simply a number of edges connected to it. Betweenness centrality of the node is defined as the number of shortest paths (i.e. most efficient paths) between any pair of nodes which pass through it, including the path between the node itself and other nodes [5].

A. Crucitti-Latora-Marchiori model

First we analysed the behaviour of the CPG using the Crucitti-Latora-Marchiori (CLM for short) model and assuming either intentional attack or random failure. The CLM model originally described in [6] is generic. Later it has been applied to analyse the resilience of the Italian [7] and Korean [9] power grid. The model comprises two types of nodes, generators and distribution substations with transmission lines (i.e. edges) between them. All edges are assigned the so-called *edge efficiency*, with values between zero and one where one represents optimal working condition and zero indicates line failure. As mentioned, it is assumed that a given generator transmits power to all distribution substations in the network along most efficient paths. Given path P, consisting of edges $i = 1, \ldots, K$ with their corresponding efficiencies e_i, the path efficiency $E(P)$ is defined as:

$$E(P) = \left[\sum_{i=1}^{K} \left(\frac{1}{e_i}\right)\right]^{-1} \quad (2)$$

Furthermore, each node i is assigned a maximum capacity C_i proportional to its *initial load* $L_i(0)$; i.e. $C_i = \alpha \cdot L_i(0)$; where $\alpha > 1$ is the *tolerance coefficient* (essentially a safety margin). This is quite intuitive since all nodes should have the capacity to sustain disturbances or increased demand (hence $\alpha > 1$). The initial load of node i, $L_i(0)$ corresponds

to standard operating conditions and is defined as its betweenness centrality, assuming all edges having optimal efficiency, $\epsilon_{ij} = 1$, at $t = 0$.

At $t = 1$ a single node k and its connecting edges are removed from the network. In the intentional attack or worst case scenario, k is taken to be either the node with the highest degree or node with the highest betweenness centrality, since both measures are commonly taken to represent the node's relative importance ([9], [7]). In the standard malfunction scenario a random node is removed. Either way, shortest paths which traverse removed edges will have to be reassigned to new edges. Note that the length of the reassigned path can be only greater or equal to the old length. Consequently network efficiency will decrease and is updated according to (1), while still assuming all edges have their initial efficiency of one. Next, the flow of electricity is redistributed along the new most efficient paths. This in turn can cause an overloading of some nodes and a consequent efficiency decrease of their connecting edges. Namely[1],

$$e_{ij}(t+1) = \begin{cases} 1 \text{ if } L_i(t) \le C_i \text{ and } L_j(t) \le C_j \\ e_{ij}(0) \cdot \dfrac{C_i}{L_i(t)} \text{ if } L_i(t) > C_i \text{ and } L_j(t) \le C_j \\ e_{ij}(0) \cdot \dfrac{\min(C_i,\ C_j)}{L_i(t) + L_j(t)} \text{ if } L_i(t) > C_i \text{ and } L_j(t) > C_j \end{cases} \quad (3)$$

Note that the case where both nodes i and j are overloaded was not described in the original work [7], so we extended the original definition to also account for this case. In order to simplify the notation we assume it is clear that $L_j(t) > C_j$ and $L_i(t) \le C_i$ case is defined analogously to the $L_i(t) > C_i$ and $L_j(t) \le C_j$ case with the simple change of subscripts. After computing new edge efficiencies (3) for all edges, new most efficient paths are found for all pairs of generator-substation nodes based on the updated edge efficiencies, and a new network efficiency (1) is computed. Next, the flow of electricity between generator-substation pairs is redistributed along the new most efficient paths. This redistribution of flow can change loads of nodes again, which in turn can cause edge efficiencies to change again, thus decreasing network efficiency further and the whole process is repeated. To summarize, the process is described by:

Remove an initial node and its edges at $t = 1$

Repeat:

- find new most efficient paths for all pairs of nodes
- compute network efficiency (1) based on new paths
- redistribute the flow along the new paths
- update node loads
- update edge efficiencies according to (3)

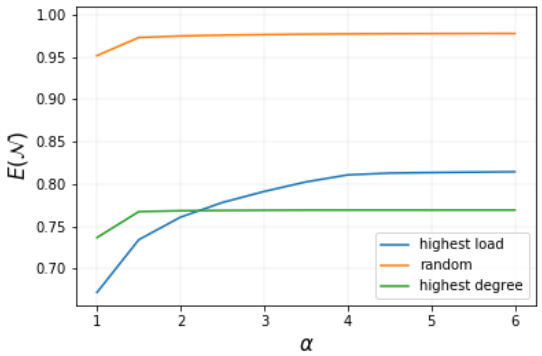


Fig. 2. CLM model. Simulation results for CPG ($\Delta\alpha = 0.5$)

We applied the CLM model to CPG and results are shown in Figure 2. In case of random failure simulation we randomly selected 50 different nodes to be removed during each simulation round (i.e. for each value of α, starting from one and with step $\Delta\alpha = 0.5$). The same sample was used during each simulation round in order to avoid fluctuations of network efficiency due to resampling. This was done since we are interested in the relationship between the network efficiency $E(\mathcal{N})$ and α and resampling would just add unwanted noise. As can be seen in the figure, targeted attack scenarios (blue and green curves) cause greater loss of network efficiency, as expected. Interestingly, there are clear curve elbows[2] for all three curves although their values vary with failure type. In cases of highest degree node failure and random node failure the threshold is $\alpha = 1.5$, whereas in the highest load node failure case there are two thresholds, at $\alpha = 1.5$ and at $\alpha = 4$.

B. *Modelling catastrophic failures*

The second model we used to analyse the CPG resilience was proposed in [15] and was later used to investigate the behaviour of the Iranian power grid under the cascading failure type of events [11]. The node loads in this model are defined in the same way as in the CLM model (i.e. as node betweenness centralities). Node capacity is again defined by $C_i = \alpha \cdot L_i(0)$, where α is the tolerance parameter as before. The major difference in comparison to the CLM model is that according to this model nodes are removed from the network if the load exceeds their capacity. This also removes all edges connected (directly) to them. Note that in the CLM model only a single node with its connected edges was initially (at $t = 1$) removed from the network. Subsequent redistribution of electricity in the CLM model caused deterioration of edge efficiencies and consequently network efficiency but no further complete failures of nodes and their connecting edges from the network. Removing nodes from the network in the case of overload is more in line with "catastrophic failure" where an overload causes complete breakdown of the node and its connecting edges as opposed to only diminished functionality. The algorithm itself proposed in [15] is very simple. Namely, removal of the initial node causes betweenness centralities of some nodes to change. If the updated betweenness centrality exceeds capacity of any node in the network, that node is removed from the network, new shortest distances are computed,

[1] assuming $t = 1$ initially

[2] Sometimes also called the knee of a curve

power flow is redistributed along the new shortest paths and the whole process is repeated either until new equilibrium is reached or until a pre-specified number of simulation steps are exceeded. Note that edges in this model have no efficiencies, i.e. they are either present or have been removed from the network due to failure of one of the two nodes connected to them.

We investigated the CPG network efficiency using the above described model. The same three failure models were used. Namely, two intentional attack scenarios (initial removal of the highest degree node or the node with highest betweenness centrality) and random node removal (as described in the previous subsection), with step $\Delta\alpha = 5$. The results of the simulations are shown in the figure below.

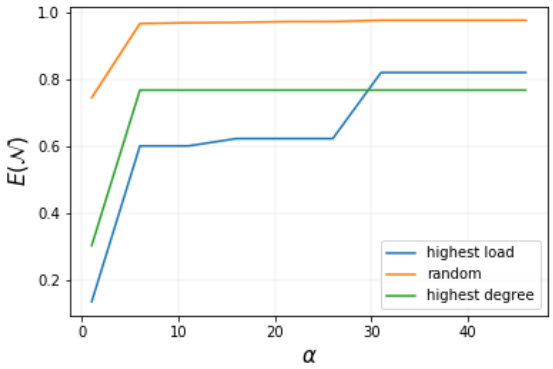


Fig. 3. "Catastrophic failures" model. Simulation results for CPG ($\Delta\alpha = 5$)

As can be seen, the same general pattern can be observed. Namely, targeted attack scenarios (blue and green curves) cause greater loss of network efficiency than random failures. However, a couple of notable differences can be also observed. Firstly, curve elbow values are much higher here. Secondly, in contrast to the CLM model, initial failure of the highest load (i.e. highest betweenness centrality node) causes greater loss of network efficiency than the failure of the highest degree node. Finally, network efficiencies are much lower than in the CLM model for the same range of α values. This is understandable since as already mentioned node overload in this model causes full blown failure of the node and its connecting edges as opposed to simply diminishing functionality like in the CLM case.

C. *Local load redistribution model*

The local load redistribution model (LLR for short) used in our analysis is partially based on [16] and [17] although we introduce some important modifications. Unlike the previous two models which define the initial load of the node to be equal to its betweenness centrality, thus taking into account topology of the whole network, the LLR model is local in the sense that the initial load of each node depends only on its local topology ([16], [17]). In [16] the initial load of node i at $t = 0$ is defined by $L_i(0) = \mu \cdot k_i^{\theta}$, where k_i is the degree of node i and μ and θ are model parameters. In order to give it an intuitive interpretation we set $\mu = \theta = 1$. The capacity of node i equals $C_i = \alpha \cdot L_i(0) = \alpha \cdot k_i$ where $\alpha > 1$ is the tolerance *coefficient*, as before ([16], [17]). Initial conditions are the same as before. Namely, at $t = 1$ a single node fails and is removed from the network. Again, we simulate the same three failure scenarios.

An important difference between the LLR model and the previous two models is that in the LLR extra loads generated by the initial node failure spread "outward" from the failed node i and at the next time step affect only those nodes which are in direct contact with i. Namely, assuming that node i which is connected to node j fails at time t, the proportion of its load redistributed to j, Δ_{ij} is in the original LLR model ([16], [17]) defined by (where Γ_i is the "neighbourhood" of i):

$$\Delta_{ij}(t) = L_i(0) \cdot \frac{L_j(0)}{\sum_{m \in \Gamma_i} L_m(0)} \tag{4}$$

The new load of j at $t + 1$ is given by

$$L_j(t+1) = L_j(t) + \Delta_{ij}(t) \tag{5}$$

Instead of using (4), we define the extra load to be

$$\Delta_{ij}(t) = L_i(t) \cdot \frac{C_j - L_j(t)}{\sum_{m \in \Gamma_i}(C_m - L_m(t))} \tag{6}$$

The reason for using (6) instead of (4) is twofold. Firstly, (6) assumes that the *present* load of i, $L_i(t)$ is redistributed to its neighbours (due to its failure) as opposed to its initial load $L_i(0)$. Namely, taking the sum over $j \in \Gamma_i$ in (4), yields $\sum_{j \in \Gamma_i} \Delta_{ij} = L_i(0)$, which is the load of i at $t = 0$. Taking the same sum in (6), yields $\sum_{j \in \Gamma_i} \Delta_{ij} = L_i(t)$ which is the load of i at time t. We argue the latter is a more realistic assumption since the present load accounts for the present state of the network. Namely, it accounts for the initial node failure and all subsequent load redistributions which took place. Secondly, extra load redistributed to j in (6) is not proportional to its normalized initial load $L_j(0)/\sum_{m \in \Gamma_i} L_m(0)$, as in (4), but rather to its currently available normalized capacity $C_j - L_j(t)/\sum_{m \in \Gamma_i}(C_m - L_m(t))$. This is in effect a mitigation strategy which models human intervention. Intuitively, it makes sense to redirect the extra load to neighbouring nodes in proportion to their currently available capacity as opposed to their initial load. As in the original model we also use (5) to update the weights. Note that if more neighbours of i fail at the same time, the contribution of all of them has to be taken into account. This is done simply by taking the sum $\sum_k \Delta_{kj}(t)$ in (5) where k denotes all nodes which fail at time t.

We simulated the behaviour of CPG based on the above described model (6) and the same three failure scenarios as before (with step $\Delta\alpha = 0.025$). Results are shown in Figure 4 below. Noteworthy, it is inevitable that neighbouring nodes will fail for very small values of α, since in such cases all nodes are running at the very limit of their capacity. Hence, even a very small extra load a node receives from its failed neighbour will inevitably exceed its capacity and will cause it to fail. Indeed, such behaviour can be clearly seen in Figure 4 where network efficiencies for $\alpha = 1$ equal zero, thus indicating total network collapse. An obvious solution to this problem is to simply increase the tolerance coefficient α, which is obviously the case in practice. Interestingly, even a

small increase of tolerance coefficient of about 50% substantially increases the network resilience, as shown in Figure 4.

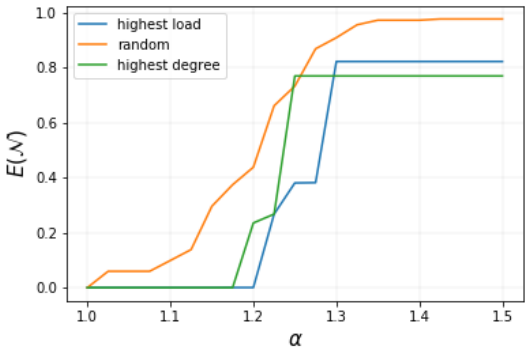


Fig. 4. LLR model. Simulation results for CPG ($\Delta\alpha = 0.025$)

III. CONCLUSION

In this study we analysed the resilience of the Croatian power grid based on the complex network analysis approach. Three different models were used. Major limitation of all three models is that they are discrete-time models. Also the assumption that a generator node transmits power to all distribution nodes in the network is unrealistic. Moreover generators and distribution stations as well as transmission lines are treated as simple nodes and edges respectively, without modelling their physical and electrical properties. That being said however, the expectation of developing both analytically tractable and accurate models of the highly complex cascading failure processes in power grids is at present unrealistic [18]. Therefore, despite all of the above mentioned limitations of simple models described here and other similar models described elsewhere in the literature, they should be viewed as an important stepping stone towards more realistic models.

[1] H. Guo, C. Zheng, H. H.C. Iu and T. Fernando, "A critical review of cascading failure analysis and modelling of power system", Renewable and Sustainable Energy Reviews, Vol. 80 (2017), Pages 9-22, DOI: 10.1016/j.rser.2017.05.206

[2] A. Shahpari, M. Khansari, A. Moeini, "Vulnerability analysis of power grid with the network science approach based on actual grid characteristics: A case study in Iran," Physica A: Statistical Mechanics and its Applications, Vol. 513 (2019), Pages 14-21. DOI:10.1016/j.physa.2018.08.059

[3] G.A. Pagani, and M. Aiello, "The Power Grid as a complex network: A survey", Physica A: Statistical Mechanics and its Applications, Vol. 392 (2013), Pages: 2688-2700. DOI: 10.1016/j.physa.2013.01.023

[4] P. Hines, E. Cotilla-Sanchez and S. Blumsack, "Do topological models provide good information about vulnerability in electric power networks?", Chaos, Vol. 20 (2010), Pages: 033122(1-5). DOI: 10.1063/1.3489887

[5] M. Newman, "Networks", Second edition (2018). Oxford University Press, Oxford, UK. ISBN: 978-0-19-880509-0

[6] V. Latora and M. Marchiori, "Efficient behavior of small-world networks." Phys Rev Lett. 2001; Vol. 87 (2001) DOI:10.1103/PhysRevLett.87.198701

[7] P. Crucitti, V. Latora and M. Marchiori, "A topological analysis of the Italian electric power grid", Physica A: Statistical Mechanics and its Applications, Vol. 338 (2004), Pages: 92-97, DOI: 10.1016/j.physa.2004.02.029

[8] M. Saleh, Y. Esa, and A. Mohamed, "Applications of Complex Network Analysis in Electric Power Systems", *Energies* (2018), DOI: 10.3390/en11061381

[9] D. H. Kim, D. A. Eisenberg Y. H. Chun and J. Park, "Network topology and resilience analysis of South Korean power grid", Physica A: Statistical Mechanics and its Applications, Vol. 465 (2017), Pages: 13-24. DOI: 10.1016/j.physa.2016.08.002

[10] R. Sole, M. Rosas-Casals, B. Corominas-Murtra & S. Valverde, „Robustness of the European power grids under intentional attack", Physical review. E, Statistical, nonlinear, and soft matter physics (2008). DOI: 10.1103/PhysRevE.77.026102

[11] M. A. S. Monfared, M. Jalili and Z. Alipour, "Topology and vulnerability of the Iranian power grid", ", Physica A: Statistical Mechanics and its Applications, Vol. 406 (2014), Pages: 24-33. DOI: 10.1016/j.physa.2014.03.031

[12] H. Das, G.S. Panda, B. Muduli and P.K. Rath, "The Complex Network Analysis of Power Grid: A Case Study of the West Bengal Power Network", Intelligent Computing, Networking, and Informatics. Advances in Intelligent Systems and Computing, Vol 243 (2014), Pages: 17-29, DOI: 10.1007/978-81-322-1665-0_3.

[13] D.P. Chassin and C. Posse, "Evaluating North American Electric Grid Reliability Using the Barabasi-Albert Network Model", Physica A: Statistical Mechanics and its Applications, Vol. 355 (2005), Pages: 667-677. DOI: 10.1016/j.physa.2005.02.051

[14] J.W. Wang and L.L. Rong, "Robustness of the western United States power grid under edge attack strategies due to cascading failures", Safety Science, Vol. 49 (2011), Pages: 807-812, DOI: 10.1016/j.ssci.2010.10.003

[15] A.E. Motter and YC. Lai. "Cascade-based attacks on complex networks", Physical review. E, Statistical, nonlinear, and soft matter physics (2002). DOI:10.1103/PhysRevE.66.065102

[16] J. Wang, L. Rong, L. Zhang, and Z. Zhang, "Attack vulnerability of scale-free networks due to cascading failures". Physica A: Statistical Mechanics and its Applications Vol. 387 (2008), Pages: 6671–6678, DOI:10.1016/j.physa.2008.08.037.

[17] J. Zhang, X. Xu, L. Hong, S. Wang and & Q, Fei",Attack vulnerability of self-organizing networks", Safety Science, Vol. 50 (2012), Pages: 443-447, DOI:10.1016/j.ssci.2011.10.005.

[18] B. Schäfer, D. Witthaut, M. Timme and V. Latora *et al.* "Dynamically induced cascading failures in power grids". *Nat Commun* Vol. 9 (2018). DOI:10.1038/s41467-018-04287-5

Analysis of justification for using capacitor banks in distribution networks with low power demand

D. Ćetković*, S. Vlahinić**, D.Franković***, V.Komen****
* HEP—Distribution system operator, Elektroprimorje, 51000 Rijeka, Croatia; deni.cetkovic@hep.hr
** Faculty of Engineering, University of Rijeka, 51000 Rijeka, Croatia; sasa.vlahinic@riteh.hr
*** Faculty of Engineering, University of Rijeka, 51000 Rijeka, Croatia; dubravko.frankovic@riteh.hr
**** HEP—Distribution system operator, Elektroprimorje, 51000 Rijeka, Croatia; vitomir.komen@hep.hr

***Abstract* - In distribution networks with normal and high load capacitor banks are used in MV/LV substations for reducing power losses. However, underground MV cable lines and capacitor banks have a similar impact on power loss reduction in the distribution system so in specific conditions such as long underground MV cable lines and low electric energy consumption the question arises about the justification for the use of capacitor banks. This paper presents an approach to the analysis of the justification for the use of capacitor banks in indicated circumstances. The approach is based on monitoring statistics of average measured power consumption during the representative period. The operation of the monitored distribution network is simulated for each short interval of the selected period. For each short interval, simulations are made both with and without the use of capacitor banks and energy savings and losses are calculated. On the basis of calculated data, the cumulative sum of saved or lost energy during the whole selected period is calculated and a conclusion is reached on the justification for the use of capacitor banks in the monitored network.**

Keywords – distribution system; power loss reduction; capacitor banks; distribution network simulations;

I. INTRODUCTION

Medium voltage distribution networks are primarily used for the distribution of electricity from the transmission network to consumer hubs or lower level transformer stations. The flow of electric energy through the medium voltage distribution networks generates losses that can generally be divided into losses caused by the active power flow and losses caused by the reactive power flow. One of the standard measures for reducing the losses caused by reactive power flow is the installation of capacitor banks on LV side in MV/LV substations [1]. As shown in Figure 1., capacitor banks generate reactive power that flows in the opposite direction and reduce inductive power flow trough the network.

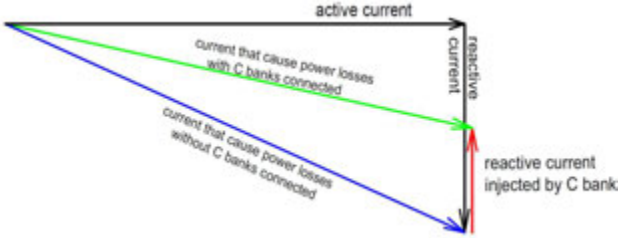


Figure 1. Reactive power compensation with capacitor banks

With the proper allocation of capacitor banks, the losses in the distribution power networks can significantly decrease [2,3]. However, underground cable lines also generate capacitive reactive power, with the total amount of reactive power produced increases with the increase in cable line length. Therefore, in the case of MV network with long underground cable sections and small power consumption, the question arises about the justification for the usage of capacitor banks. In extreme conditions such as very long underground cable network and very low power consumption capacitor banks could increase energy losses instead of reducing them, as shown in Figure 2.

Studies dealing with proper allocation and sizing of capacitor banks [2,3] usually are based only on losses that occur in the MV lines of the distribution networks. In this paper, the distribution system model includes MV/LV transformers with their parameters that can contribute to overall system losses.

This paper focuses on the analysis of the justification for the installation of capacitor banks in MV/LV substations powered from MV networks with long underground cable lines. Simulations are being made using model of an average MV network that can be found in rural regions of Croatia (Gorski Kotar County), whose characteristics are very low loading for long periods and very long underground cable sections.

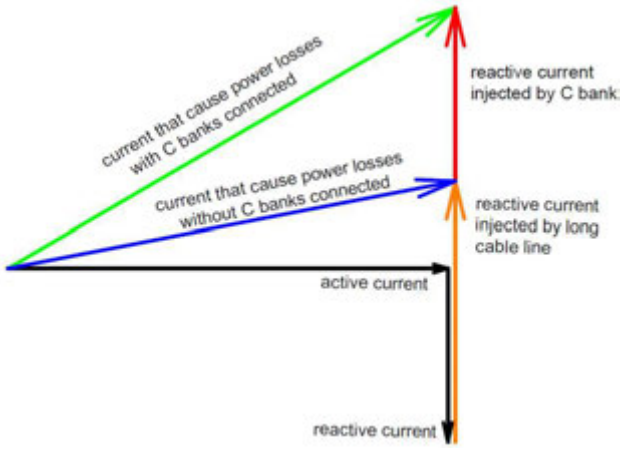


Figure 2. Reactive power compensation in MV networks with very long underground cable line

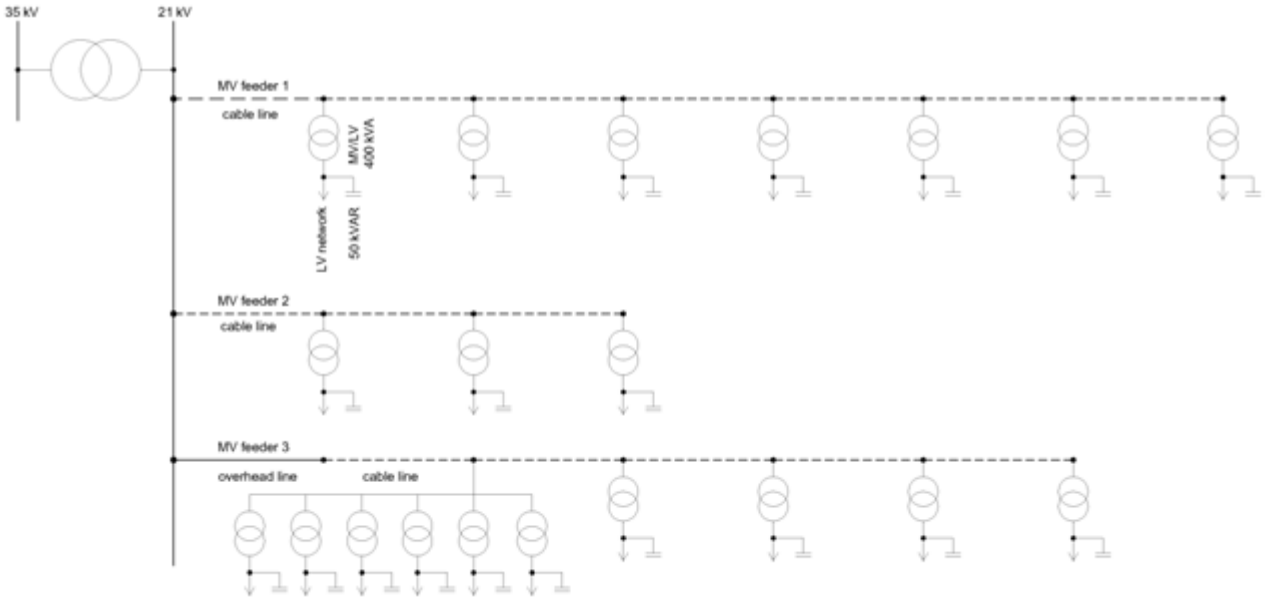


Figure 3. MV distribution network simulation model

II. DISTRIBUTION NETWORK CALCULATION

A. Distribution network model design

In order to calculate the values of currents that generate power losses and to analyze the values obtained in different situations, it is necessary to create a distribution system model and make load flow calculations. Load flow calculations were made in Matlab with „backward/forward sweep“ algorithm [4-6].

The distribution network shown in Figure 3. has 32 nodes. It was modeled in Matlab as a 32x32 matrix (Figure 4.) where the rows of the matrix represent the network lines and transformers, and the columns of the matrix represent the network nodes. The starting point of each line or transformer is valued '-1' and the ending point '1'. Other matrix elements are valued '0'.

```
IM=[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    -1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 -1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 -1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 -1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 -1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 -1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 0 -1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 0 0 -1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    -1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0 0 0 0 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0 0 0 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0 0 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1 0;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 1;
    0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1];
```

Figure 4. Matrix of distribution network topology

Electric parameters of lines and transformers are entered as 1x32 vectors.

$$R_l=[0.325\ 0.86\ 0.206\ ...\ 12.67875]; \quad (1)$$

Maximum power consumption in every node is also entered as a 1x32 vector.

$$P_{l0max}=[0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 400\ 400\ \ ...\ 400]; \quad (2)$$

Actual power consumption in every node depends on power consumption level.

$$P_{lmax} = pcl \cdot P_{l0max} \quad (3)$$

The parameters of consumer power factor and capacitor banks power are also defined. This data can be later easily modified in order to perform analyses for various conditions.

B. Power consumption level data

In order to reach a general conclusion on the justification for the use of capacitor banks in the distribution network, it is necessary to take into account the variability of consumption level.

For the network on which the simulation model of MV network is based, it was assumed that daily power consumption level is the same as obtained data on the average daily power supplied by the main 35/21 kV transformer to the MV grid. For the network on which the simulation model of MV network is based, data were collected on the average daily power supplied by the 35/21 kV transformer to the MV grid. Data were collected for a

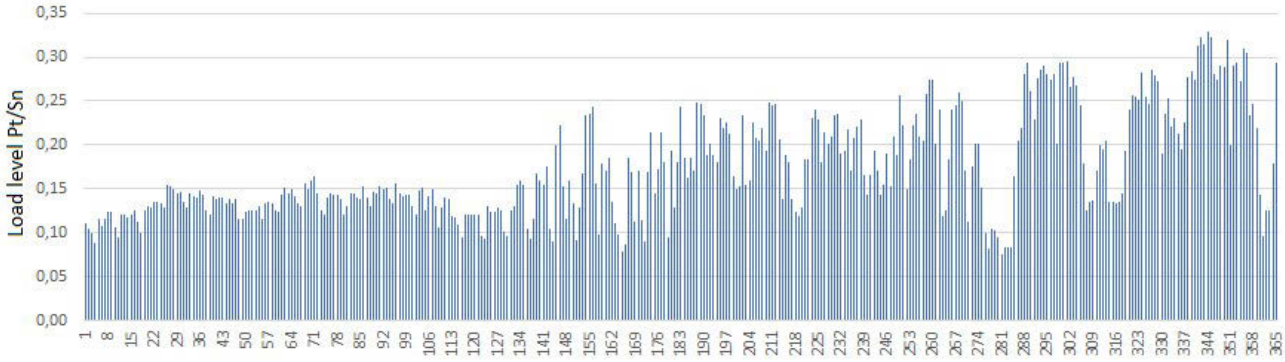


Figure 5. Daily load level of 35/21 kV power transformer during period from 8.11.2018.(Day1) to 7.11.2019.(Day 365)

period of one year, in the period from 8.11.2018. to 7.11.2019. The graphical representation of data is shown in Figure 5.

Power consumption level is modeled as 1 x n vector where *n* represents a number of intervals for which the data is obtained. For the period of one year and daily measured average power consumption data is entered as a 1x365 vector.

$$pcl=[0.111\ 0.104\ 0.100\ 0.088\ ...\ 0.294]; \quad (3)$$

Since the data was measured on a daily basis, a value of 24 hours was entered for the time interval.

III. DETERMINING THE JUSTIFICATION FOR USE OF CAPACITOR BANKS

The proposed method for capacitor banks justification analysis is shown in Figure 6. First, the data defining the network topology and the collected consumption level data are entered. Then the combination of the variable parameters for which the analysis is to be performed is defined.

In order to determine power savings and losses due to the use of capacitor banks simulations are made for each interval that is entered in the consumption level matrix. For each interval, simulations are firstly made without capacitor banks and then repeated with capacitor banks connected. For each interval, the amount of power saved is calculated by subtracting the value of the losses obtained with the capacitor banks connected from the value of the losses with capacitor banks disconnected, as given in equation (3).

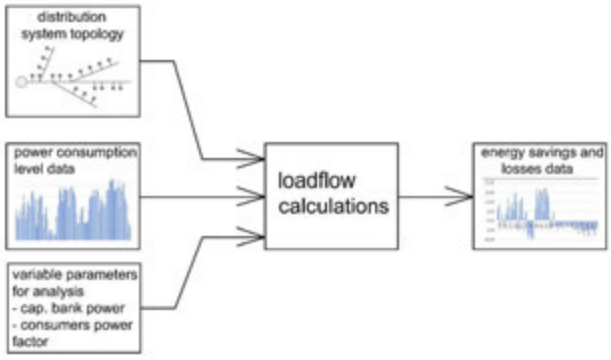


Figure 6. Proposed method for justification analysis

$$P_{savings} = P_{loss_withCbanks} - P_{loss_withoutCbanks} \quad (3)$$

By multiplying saved power with a duration of time interval information about total saved or lost energy is obtained, as given in equation (4).

$$W_{savings} = P_{savings} \cdot time_interval \quad (4)$$

Results of simulations for all intervals are represented as 1 x n vector where *n* represents a number of intervals for which the data about power consumption level was obtained. For example, for the period of one year and daily measured average power consumption level saved power is calculated and given as a 1x365 vector where every column gives the value of total saved or lost energy during one day as a result of the use of capacitor banks

$$W_{savings}=[10.22\ 18.31\ -12.33\ ...\ 2.93]; \quad (5)$$

If the value is positive it means that capacitor banks have made a positive impact on power losses during that

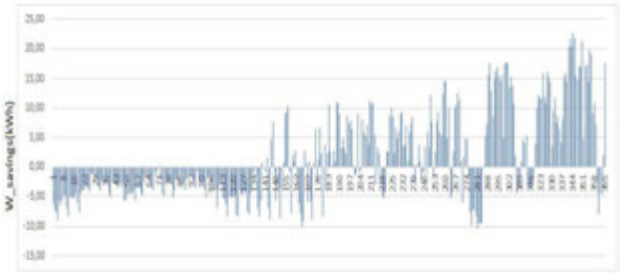


Figure 7. Daily energy savings and losses during one year due to usage of 12.5 kVar capacitor banks (best results)

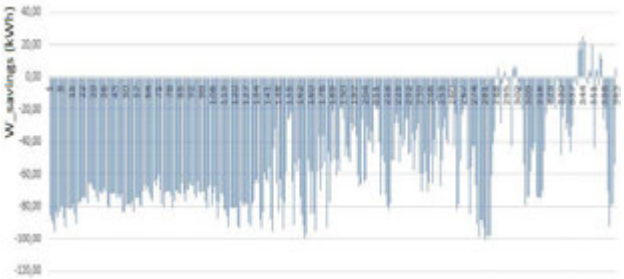


Figure 8. Daily energy savings and losses during one year due to usage of 50 kVar capacitor banks (worst results)

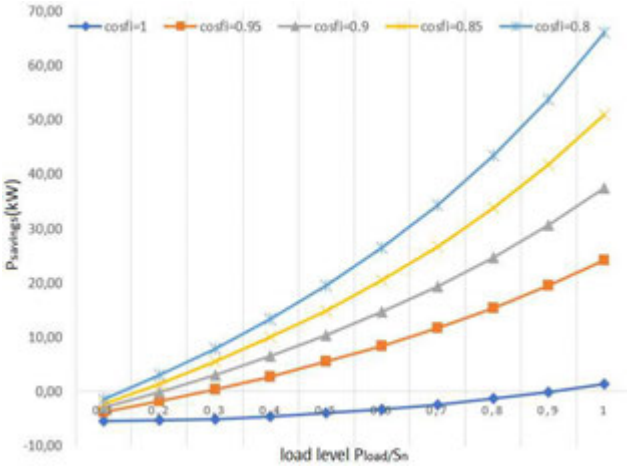


Figure 9. Calculated savings (losses) due to use of 50 kVAr capacitor banks for various consumption levels and power factors

day. If the value is negative it means that capacitor banks have actually increased distribution network losses.

The compensation is considered with 12.5, 25 and 50 kVAr capacitor banks connected on the LV side in MV/LV substations because these values represent the usual installation power that distribution system operator in Croatia use for compensation.

Examples of results are given in Figures 7. and 8. This figures show daily energy savings and losses. Figure 7. show best results for energy savings which are achieved by installation of 12.5 kVAr capacitor banks. Installation of capacitor banks of this power would in general increase energy savings because energy savings in the period of high power demand (days 323.-358. for example) outweigh the increase in losses in periods of low power demand (days 1.-140.).

Figure 8. show worst results for energy savings which are achieved by installation of 50 kVAr capacitor banks. The diagrams clearly indicate that 50 kVAr capacitor banks would produce minor energy savings in the smaller part of the year and large losses in most of the year.

Finally, the total saved or lost energy during the whole observed period is calculated as a cumulative sum of saved or lost energy in each interval, as shown in equation (6). If the given results are positive it means that the use of capacitor banks of that power would be justified in terms of reducing energy losses. If total energy saved has a negative value it means that capacitor banks of that power would only increase power losses so it is not recommended to use capacitor banks of that power.

$$W_{SAVED_TOTAL} = \sum_{1}^{n} W_{savings}[1,n] \qquad (6)$$

Based on the quality of collected data about power consumption level this algorithm could be easily applied for any time interval and amount of collected data. Capacitor banks installation power could be easily changed to analyze network losses behaviour for various conditions.

Based on the simulations performed it is calculated that compensation with 50 kVAr capacitor banks would lead to an energy loss of 20183 kWh during a year period.

Results of simulations for best scenario with 12.5 kVAr capacitor banks installed show that the annual energy savings would be 573,3 kWh.

It is also possible to consider the justification of using a particular capacitor bank under different load conditions. Justification for the usage of capacitor banks could be easily researched in various conditions such as different power factors and different consumption, as shown in Figure 9. Figure show the general power savings and losses due to usage of 50 kVAr capacitor banks for different load levels and power factors. It can be seen that in the case of high loading conditions significant savings can be achieved for all power factors. However, at low loading conditions usage of 50 kVAr capacitor banks can cause power losses.

IV. CONCLUSION

Capacitor banks are a common solution for reducing energy losses in the power distribution systems. However, the analyses of capacitor banks optimization is relatively complex issue and depends on many factors such as power consumption level, configuration and length of cable lines and costs of installing capacitor banks. Proper determination of capacitor banks power can significantly reduce power losses.

This paper presents analyses of the capacitor banks sizing based on monitoring statistics of average measured power consumption during selected period. The operation of the monitored distribution network is simulated for each short interval of the selected period. For each short interval, simulations are made both with and without the use of capacitor banks and energy savings and losses are calculated. On the basis of calculated data, the cumulative sum of saved or lost energy during a whole selected period is calculated and a conclusion is reached on the justification for the use of each capacitor bank in the monitored network. Distribution system model used in this study includes both MV lines and MV/LV transformers models with their parameters that can contribute to overall system losses.

Simulations were made with 12.5, 25, and 50 kVAr capacitor banks because these values represent the usual installation power that distribution system operator in Croatia use for compensation and are currently a common solution in practice.

Simulations showed that usage of 12.5 kVAr capacitor banks, for considered load levels, could be justified because it would create energy savings on an annual basis. Usage of 25 and 50 kVAr capacitor banks would lead only to total energy loss increases. Therefore, it can be concluded that in the case of the presented distribution network and power demand profile the use of 12.5 kVar capacitor banks can be justified in terms of reducing technical losses. The use of 25 and 50 kVAr capacitor banks would not be justified.

The described method may be applied for any MV distribution network for any time interval.

REFERENCES

[1] Miller, T.J.E. Reactive Power Control in Electric Systems; Wiley: New York, NY, USA, 1982.

[2] Ivanov, O.; Neagu, B.C.; Grigoras, G. ; Gavrilas, M. Optimal Capacitor Bank Allocation in Electricity Distribution Networks Using Metaheuristic Algorithms, ENERGIES Vol:12, No: 22, November 2019., Article Number: 4239

[3] Masoum, M.; Ladjevardi, M. ; Jafarian, A.; Fuchs, E. Optimal placement, replacement and sizing of capacitor banks in distorted distribution networks by genetic algorithms, IEEE transactions on power delivery Vol. 19 No: 4 October 2004, Pages: 1794-1801

[4] Haque, M. H. A General Load Flow Method for Distribution System. // Electric Power System Research. 54, 1(2000), pp. 47-54.

[5] Hamouda, S. A.; Zehar, K. Efficient Load Flow Method for Radial Distribution Feeders. // Journal of Applied Sciences. 6, 13(2006), pp. 2471-2478.

[6] Prenc, R., Škrlec, D., Komen, V., "A novel load flow algorithm for radial distribution networks with dispersed generation", Tehnički vjesnik – Technical Gazette, Vol. 20, No., 6, December 2013., str. 969-977.

[7] Vlahinić, S.; Franković, D.; Komen, V.; Antonić, A. Reactive Power Compensation with PV Inverters for System Loss Reduction. Energies 2019, 12, 4062.

AVR and PSS Coordination Strategy by Using Multi-objective Ant Lion Optimizer

T. Špoljarić*, I. Pavić**
* University of Applied Sciences, Zagreb, Croatia
** Faculty of Electrical Engineering and Computing, Zagreb, Croatia
tomislav.spoljaric@tvz.hr, ivica.pavic@fer.hr

***Abstract* - In this paper a novel optimization method called Multi-Objective Ant Lion Optimizer (MOALO) is proposed for tuning synchronous generator excitation controls in multi machine power system. Devices used in excitation control are automatic voltage regulator (AVR) and power system stabilizer (PSS). Two area four machine model (TAFM) is used for observing power system dynamics through several various operating states. In a performance analysis of a proposed algorithm two objective functions are used. First objective function uses integral of time weighted absolute error of rotor speed, voltage and tie line active power data, while second objective function uses mean value of time domain transitional process quality indicators such as overshoot, undershoot and settling time. A proposed algorithm is tested and its performance is compared with performances of two other multi objective swarm intelligence algorithms: Multi-Objective Particle Swarm Optimization (MOPSO) and Multi-Objective Salp Swarm Algorithm (MOSSA). Results are compared and presented as sets of solutions composed in Pareto fronts.**

Keywords - Generator Excitation Controls, Multi-objective optimization, Ant Lion Optimizer, Power System Dynamics

I. INTRODUCTION

Due to a nonlinear dynamic performance of a power system, a stability of such system has proven to be a challenge for observation and control. In recent decades the problems regarding power system stability have been discussed and surveyed in several articles and books, most notably [1], [2], [4] and [5]. Studies in a field of power systems' stability led to formalizing a power system stability classification that has been made by experts in this field [3].

These classified stability categories include small disturbance angle stability and transient stability as two subcategories of rotor angle stability [3]. Rotor angle stability is examined through short term analysis. Primarily, changes in generators' characteristic values are observed while considering small (short term load variations and outages) and large disturbances (short circuits). These disturbances lead to appearance of several oscillatory modes with low frequency of which local (1.5 – 2.5 Hz) and inter-area oscillations (0.2 – 1 Hz) are most significant in observing a rotor angle stability of a studied power system.

In addressing this problem, various control methods have been proposed and implemented. Elements that can be tuned include automatic voltage regulators, power system stabilizers on the excitation systems, and FACTS systems on transmission network side. Lately, several of these tuning methods proposed implementation of coordinated strategies for parameter tuning driven by different AI techniques such as fuzzy regulation [8 – 10], neural networks [8] and swarm intelligence methods [8, 10 – 16]. With the latest blooming of swarm intelligence algorithms [7, 8] this problem has become more interesting in state-of-the-art control techniques.

This paper describes coordinated control design via multi-objective Ant Lion Optimizer (MOALO) of two excitation regulators on synchronous generators. Power system controlled with MOALO technique is tested and compared with two different multi-objective swarm intelligence methods and finally proposed as a solid solution for such problems in power system dynamics.

II. POWER SYSTEM TESTING MODEL

A. Model description

Generator controls' models include turbine governing control and excitation system control. On the excitation side of each generator DC exciters are equipped with AVR and PSS controllers. These controllers are widely used for damping of various local and inter-area modes of frequency. AVR controller uses terminal voltage and PSS uses speed deviations as a feedback information for control.

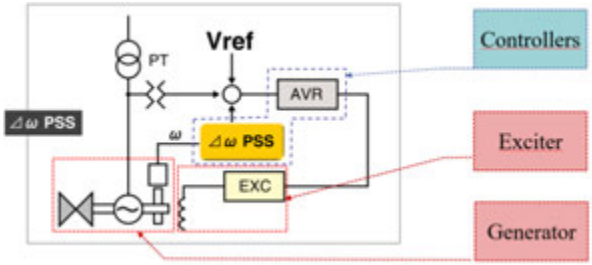


Figure 1. Block schematic of a geneartor and its exciter

For simulation and comparison, a standard TAFM (Two Area Four Machine) model has been selected and adapted for testing. This model, introduced in [2], has been set as a certain standard for power system stability analysis in various articles (with [12], [13] and [16] being examples of swarm intelligence methods applied in power system control).

Described model has two areas of production and consumption. Each area has two generating units. These two areas are interconnected via 220kV transmission tie-line. Observed TAFM model has been implemented previously for performance analysis of different objective functions in multi-objective optimization [19], and is shown on fig. 2.

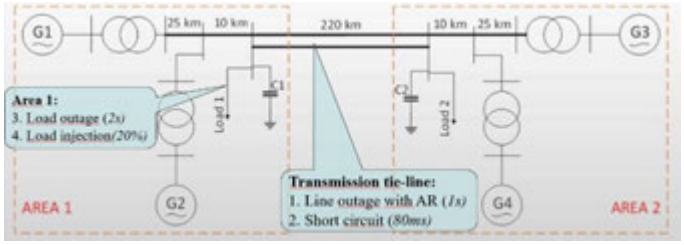


Figure 2. Observed TAFM model [19]

TABLE I. TECHNICAL DATA OF TESTED MODEL [19]

Element	*Area*	*Technical characteristics*
Generator 1	Area 1	$20kV,\ 900MVA,\ 50Hz$ $X_d = 1.8pu,\ X'_d = 0.3pu, X''_d = 0.25pu,\ R_a = 0.0025pu$ $X_q = 1.7pu,\ X'_q = 0.55pu, X''_q = 0.25pu, X_l = 0.2pu$ $T'_{d0} = 8s, T''_{d0} = 0.03s, T'_{q0} = 0.4s, T''_{q0} = 0.05s$ $H_{G1} = H_{G2} = 6.5,\ H_{G3} = H_{G4} = 6.175,\ K_D = 0$
Generator 2	Area 1	
Generator 3	Area 2	
Generator 4	Area 2	
Load 1	Area 1	$P_{load1} = 967MW, Q_{load1} = 100MVAr$
Load 2	Area 2	$P_{load2} = 1767MW, Q_{load2} = 100MVAr$
Transmission lines 1, 2	Area 1	$l_{line1} = l_{line3} = 10km, l_{line2} = l_{line4} = 25km$ $l_{line5} = 220km, V_{nom} = 230kV$ $R_l = \frac{0.001pu}{km},\ L_l = 0.001pu,$ $b_l = 0.00175pu/km$
Transmission lines 3, 4	Area 2	
Transmission line 5	Area 1 – Area 2	
Transformers 1, 2	Area 1	$900MVA, {}^{20kV}/_{230kV},\ 50Hz,\ D1Yg$ $R_1 = R_2 = 10^{-6}\ pu, L_1 = 0pu, L_2 = 0.15pu$
Transformers 3, 4	Area 2	
Capacitor 1	Area 1	$Q_{C1} = 300MVAr$
Capacitor 2	Area 2	$Q_{C2} = 350MVAr$

B. Controller model design

Controller models in coordinated tuning strategy include: PT1 regulator with gain K_a and time constant T_a which is used as AVR. PSS1 type with gain K_{PSS}, washout time constant T_w, and two lead-lag blocks' time constants $T_{1n}, T_{1d}, T_{2n}, T_{2d}$ is used as PSS regulator model. Block schematic of PSS is shown in fig. 3.

C. Scenarios used for algorithm performance

Four scenarios are simulated for evaluation of three different algorithms' performance [19]:

- 1st scenario – outage and automatic reclosure after first second of one of two paralell interconnection lines,
- 2nd scenario – short circuit that occurs on 170th kilometre of a transmission line from Area 1, and lasts for 80 ms (self-clearing fault),
- 3rd scenario – temporary outage of a load in Area 1 which lasts for 2 seconds,
- 4th scenario – Additional 20% load injection on a load bus in Area 1.

Observed system is modelled in MATLAB Simulink®, with duration of simulation that lasts for 25 seconds. All scenarios include contingency that occurs in 5th second of simulation.

III. MULTI-OBJECTIVE ANT LION OPTIMIZER ALGORITHM

A. A priori analysis and scalarization of objective functions

Multi-objective optimization (MOO) provides set of alternate solutions, as opposed to single-objective optimization, where a single best solution is provided (fig. 4). These sets are called Pareto-optimal solutions, since there no dominant solution exists in majority of the multi-optimization problems. MOO is also subjected to equality and inequality constraints that are examined during the optimization process [18]. Also, a priori analysis through scalarization of two independent objective functions is used is in this paper. Therefore, a third objective function, as a sum of two weighted objective functions, is formed:

$$\begin{gathered} F(t) = w_1 \cdot 500 \cdot f_1(t) + w_2 \cdot f_2(t) \begin{cases} w_1 = [0,1] \\ w_2 = 1 - w_1 \end{cases} \\ f_1(t) = \int_{t=t_1}^{25\,s} t \cdot (|\Delta\omega(t)| + 0.005 \cdot |\Delta V_t(t)| + |\Delta P_{tie}(t)|) \\ f_2(t) = 4000 \cdot \overline{os}^2 + 1000 \cdot \overline{us}^2 + \overline{ts}^2 \end{gathered} \quad (1)$$

where deviations in rotor speeds, terminal voltages of generators and tie-line power transfer are considered in first objective function $f_1(t)$ and first overshoot, first undershoot and settling time of rotors speeds' deviations are considered in second objective function $f_2(t)$. Through scalarization with weight factors w_1 and w_2 a third objective function $F(t)$ is introduced (1). Deviations of rotor speeds, terminal voltages and tie line power transfer are defined as:

$$|\Delta\omega(t)| = |\omega_1(t) - \omega_2(t)| + |\omega_1(t) - \omega_3(t)| + |\omega_1(t) - \omega_4(t)| + |\omega_3(t) - \omega_4(t)|$$
$$|\Delta V_t(t)| = 0.005 \cdot (|V_{t1}(t) - V_{t10}| + |V_{t2}(t) - V_{t20}| + |V_{t3}(t) - V_{t30}| + |V_{t4}(t) - V_{t40}|)$$
$$|\Delta P_{tie}(t)| = \frac{|P_{tie}(t) - P_{tie0}|}{10^5}$$

with initial states of terminal voltages and power transfer considered.

Inequality constraints that were used in these multi-objective optimization tests are defined as limits of AVR and PSS parameters which define search space of controller parameters:

$$2 \le K_{a,i} \le 250$$
$$0 \le T_{a,i} \le 0.1$$

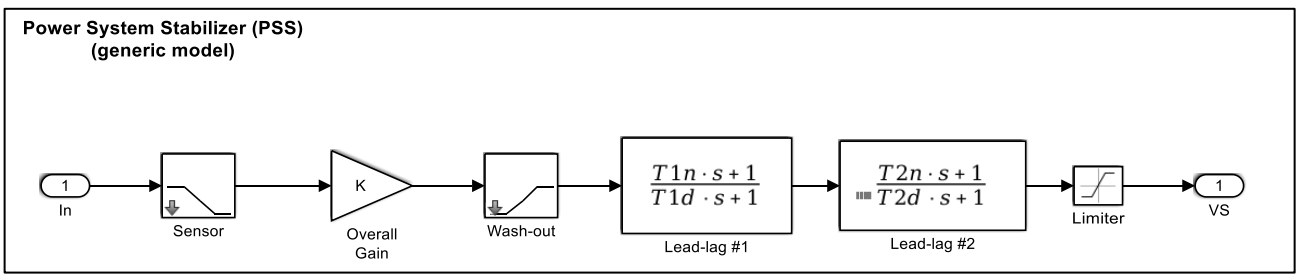


Figure 3. PSS1 model used in observed TAFM system

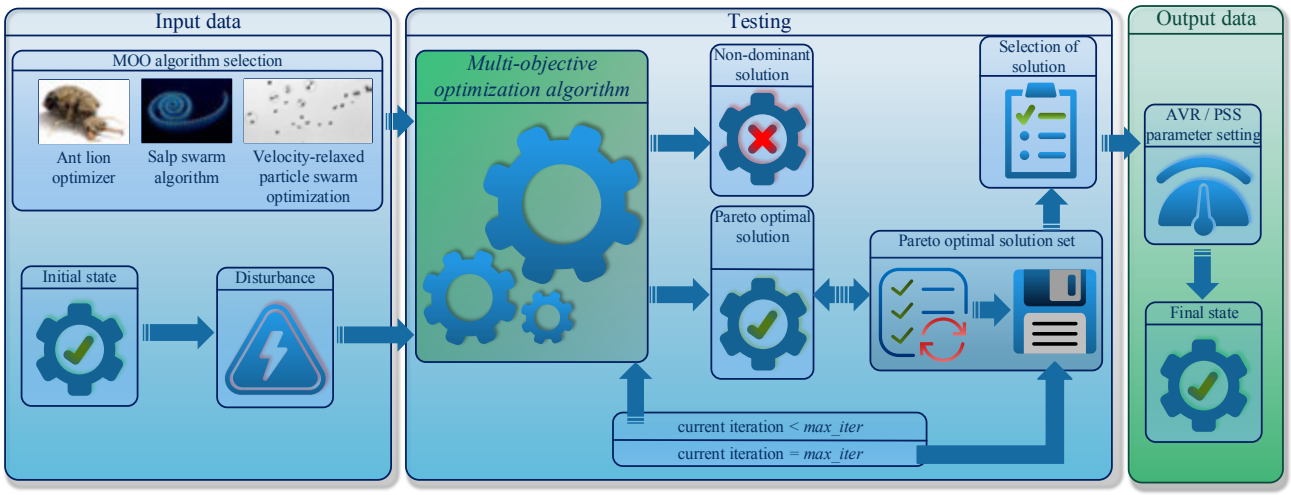


Figure 4. Flow chart of MOO testing algorithms

$$1 \le K_{pss,i} \le 250$$
$$0.005 \le T_{1n,i} \le 5$$
$$0.005 \le T_{1d,i} \le 5$$
$$0.05 \le T_{2n,i} \le 15$$
$$0.05 \le T_{2d,i} \le 25$$

B. MOALO as an optimization tool in AVR/PSS coordination strategy

MOALO is an expansion of an Ant Lion Optimizer algorithm, a nature inspired algorithm for swarm intelligence optimization [6, 18]. In one iteration six phases are implemented which include random walk of prey, trapping in pits, building a trap, sliding of prey towards antlion, catching the prey/rebuilding the pit and defining a best antlion through elitism rule. Initial positions of first ant lions and its corresponding ants are stochastically positioned through defined search space, as is common in all swarm intelligence methods. Detailed description of these stages can be found in [6] and [19].

1) Random walks of ants

Random walks of ants are described as:

$$X(t) = [0, cumsum(2r(t_1)) - 1, \dots, cumsum(2r(t_n)) - 1] \quad (2)$$

where:

$$r(t) = \begin{cases} 1, & rand(0,1) > 0.5 \\ 0, & rand(0,1) \le 0.5 \end{cases}$$

and *cumsum* is a calculation of cumulative sum. A normalization of position in search space is required as a next step and is defined as:

$$X_i^t = \frac{(X_i^t - a_i)\cdot(d_i^t - c_i^t)}{b_i - a_i} + c_i^t \quad (3)$$

where X_i^t is a position of i-th variable in iteration t, a_i and b_i are minimum and maximum of random walks in i-th variable, and c_i^t and d_i^t are minimum and maximum of i-th variable at iteration t.

2) Trapping in pits

Mathematical definition of ant lion's entrapment can be defined as:

$$c_i^t = antlion_j^t + c^t, \quad d_i^t = antlion_j^t + d^t \quad (4)$$

where c^t and d^t are minimum and maximum of all variables at iteration t, and $antlion_j^t$ is a position of j-th selected antlion at iteration t.

3) Building a trap

ALO algorithm presumes that every ant is to be trapped by one hunter. Fitness of an ant therefore corresponds to a fitness of an antlion which is employed by a roulette wheel selection function. This function gives higher probability to fitter ant lions in catching better prey.

4) Sliding of prey in a trap

Ant lions are building sliding traps whose quality corresponds to their fitness. Sliding in a trap is defined as adaptive decrease in a coned pit's limits:

$$c^t = {c^t}/{I}, \quad d^t = {d^t}/{I} \quad (5)$$

where I is a factor depending on a ratio of a current iteration and a total number of iterations and it provides a decrease in search space for achieving a more convergent movement towards optimum.

5) Catching a prey and rebuilding a trap

After sliding of prey in a trap catching takes place and is defined as:

$$f(ant_i^t) < f(antlion_j^t) \rightarrow antlion_j^t = ant_i^t \quad (6)$$

6) Rule of Elitism

Fittest ant lion in every iteration is defined as an elite ant lion. It affects also the next random walks of prey in search space and this is defined as:

$$ant_i^t = \frac{R_A^t + R_E^t}{2} \quad (7)$$

where R_A^t is a random walk of prey around antlion selected through roulette wheel selection function, and R_E^t is a random walk of the same prey around elite ant lion at iteration t.

7) Formation of Pareto front

Pareto optimal vector set calculation starts with comparison of two different vectors of k and l solutions:

$$\overrightarrow{antlion_a} = (antlion_{a,1}, \dots, antlion_{a,k})$$
$$\overrightarrow{antlion_b} = (antlion_{b,1}, \dots, antlion_{b,l})$$

Vector $\overrightarrow{antlion_a}$ dominates $\overrightarrow{antlion_b}$ if a following condition is fulfilled:

$$\forall i = \{1,2,\dots,k\}: f_i(\overrightarrow{antlion_a}) \leq f_i(\overrightarrow{antlion_b})$$
$$\wedge$$
$$\exists i = \{1,2,\dots,k\}: f_i(\overrightarrow{antlion_a}) < f_i(\overrightarrow{antlion_b})$$

Solution set $\overrightarrow{antlion_a}$ is defined as Pareto optimal if:

$$\overrightarrow{antlion_a} \in X \overset{P.O.}{\longleftarrow} \{\nexists \overrightarrow{antlion_b} \in X | \overrightarrow{antlion_b} \prec \overrightarrow{antlion_a}\}$$

Pareto set of solutions and its corresponding front of objective functions f_1 and f_2 are defined as:

$$P_s := \{\overrightarrow{antlion_a}, \overrightarrow{antlion_b} \in X | \nexists \overrightarrow{antlion_b} \prec \overrightarrow{antlion_a}\}$$
$$P_f := \{f_1(\overrightarrow{antlion_a}), f_2(\overrightarrow{antlion_a}) | \overrightarrow{antlion_a} \in P_s\} \quad (8)$$

IV. Initializaton and Simulation Results of TAFM Model

For the testing purpose initial conditions include: 2 trials for each of three weight factors ($w_1 = \{0.25, 0.5, 0.75\}$, with $w_2 = 1 - w_1$), with each trial of 50 iterations, and within each iteration a population of 25 agents (particles, salps, ants/ant lions) is tested with two objective functions (1). Dimension of each agent is equal to number of tuned parameters and is 28 (four PSSs with five parameters each, and four AVRs with two parameters each). These conditions are applied to each of three algorithms described in the next section.

A. Algorithms used in testing and comparison

Model testing is done with four different algorithms: MOALO, velocity relaxed multi-objective particle swarm optimization (MOVRPSO) and a multi-objective salp swarm algorithm (MOSSA), and afterwards results are being compared within these three algorithms.

Initial settings for VRPSO constants include velocity factors according to personal and global best solution and weight factor for velocity of previous iteration $c_1 = 1.2$, $c_2 = 0.9$, $w = 0.9$. VRPSO also uses adaptive speed for weight factor with velocity constraint: $w = w_{max} - a \cdot i$, where $a = 0.05$, and i is current iteration. Also, an algorithm called multi-objective salp swarm algorithm (MOSSA) [17] is introduced for testing and comparison of results.

In MOALO, search space decrement is done according to current iteration (5) and is as follows:

$$I = \begin{cases} 1, & i \leq \frac{max_iter}{3} \\ 1 + 10^{\frac{i}{max_iter}}, & i > \frac{max_iter}{3} \\ 1 + 10^{\frac{2i}{max_iter}}, & i > \frac{max_iter}{2} \\ 1 + 10^{\frac{3i}{max_iter}}, & i > 0.75 \cdot max_iter \\ 1 + 10^{\frac{4i}{max_iter}}, & i > 0.9 \cdot max_iter \\ 1 + 10^{\frac{5i}{max_iter}}, & i > 0.95 \cdot max_iter \end{cases}$$

where i is current iteration and max_iter is total number of iterations.

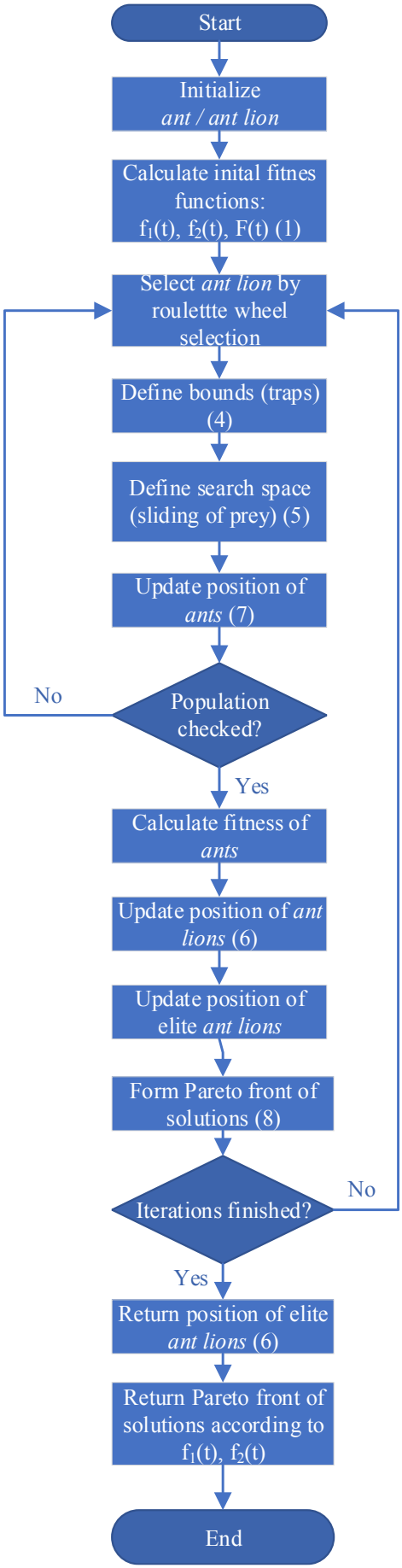


Figure 5. Flow chart of MOALO algorithm

B. Simulation results

In this section, performance of different algorithms is done by comparing graphical and analytical results of simulation. Graphical review mainly includes responses of rotor speed deviations between generators locally inside a certain area and inter-areal, between generators in different areas. Also, terminal voltages are being observed along with a tie line power flow P_{tie}. Graphical responses are presented as a single solution according to best fit by third objective function $F(t)$. Graphical responses that are shown on fig. 6-9 include speed deviations between generators G1 and G2 (local deviation) and between G1 and G3 (inter-area deviation), terminal voltage on a swing type generator G2 and tie-line power transfer between areas

Since there is also a vector of solutions/parameters given to the user by MOO algorithms, a graphical representation of Pareto fronts of each algorithm is given on fig. 10-13.

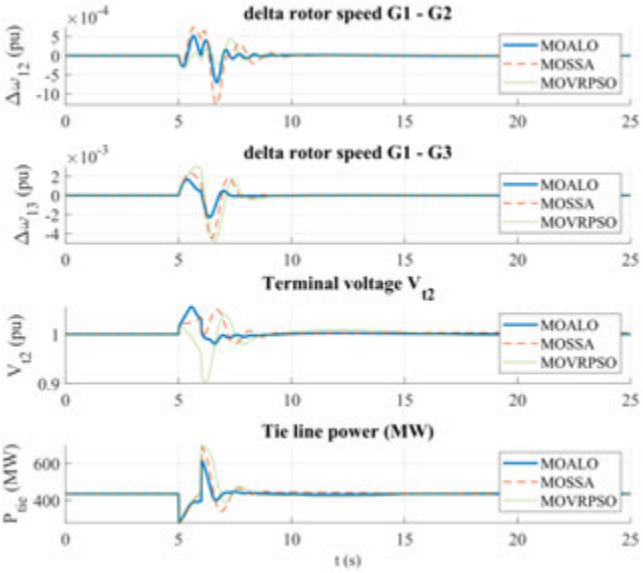


Figure 6. Testing results for 1 s tie-line outage with automatic reclosure, best fit by objective $F(t)$

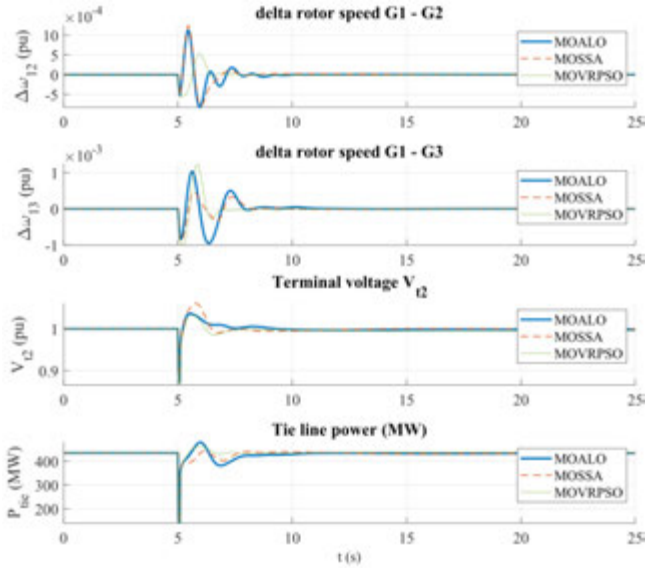


Figure 7. Testing results for 80 ms self clearing short circuit fault, best fit by objective $F(t)$

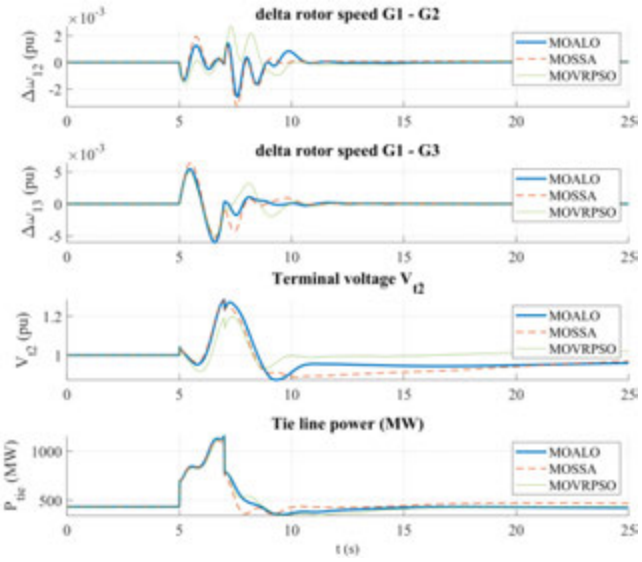


Figure 8. Testing results for 2s Area 1 load outage, best fit by objective $F(t)$

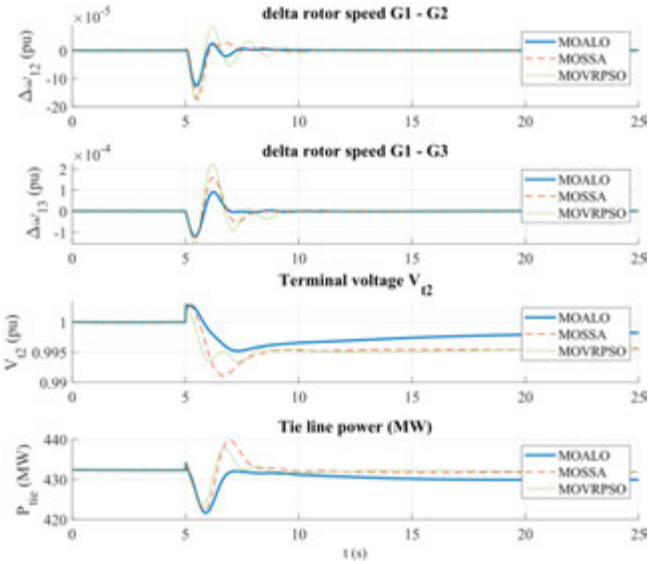


Figure 9. Testing results for additional 20% load injection in Area 1, best fit by objective $F(t)$

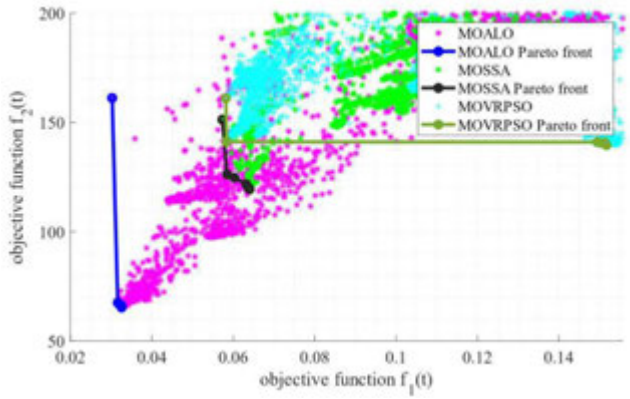


Figure 10. Pareto fronts for 1 s tie-line outage with automatic reclosure

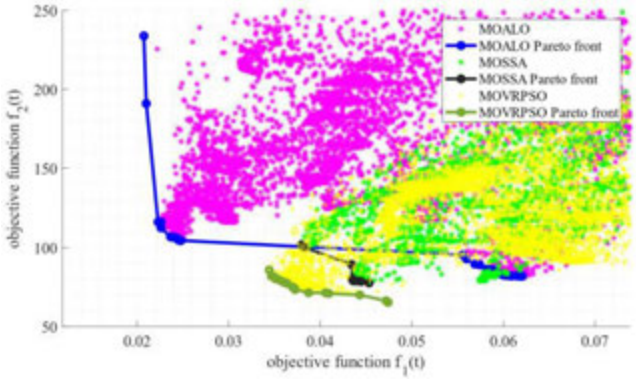


Figure 11. Pareto fronts for 80 ms self clearing short circuit fault

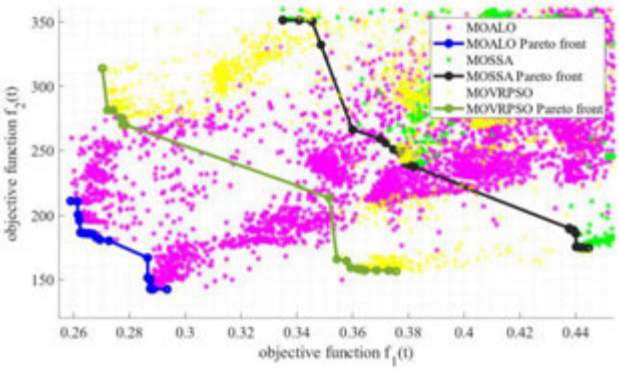


Figure 12. Pareto fronts for 2 s load outage in Area 1

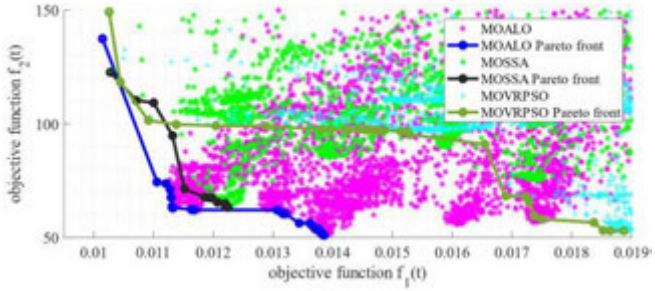


Figure 13. Pareto fronts for 20% additional load injection in Area 1

TABLE II. SIMULATION RESULTS FOR TESTED TAFM MODEL

Scenario	*Algorithm*	*Range of* $f_1(t)$	*Range of* $f_2(t)$	*Number of solutions*
Line outage	**MOALO**	0.0302 ÷ 0.0326	65.32 ÷ 161.16	9
	MOSSA	0.0573 ÷ 0.0640	75.68 ÷ 89.94	7
	MOVRPSO	0.0582 ÷ 0.1517	139.57 ÷ 161.04	10
Short circuit	**MOALO**	0.0207 ÷ 0.0621	61.57 ÷ 233.96	49
	MOSSA	0.0380 ÷ 0.0453	36.14 ÷ 81.28	44
	MOVRPSO	0.0344 ÷ 0.0474	32.34 ÷ 51.52	25
Load outage	**MOALO**	0.2589 ÷ 0.2935	143.23 ÷ 150.72	53
	MOSSA	0.3350 ÷ 0.4447	174.65 ÷ 292.85	61
	MOVRPSO	0.2703 ÷ 0.3756	169.45 ÷ 194.77	20
Load injection	**MOALO**	0.0101 ÷ 0.0104	51.11 ÷ 137.43	100
	MOSSA	0.0103 ÷ 0.0122	63.25 ÷ 122.70	13
	MOVRPSO	0.0103 ÷ 0.0189	52.95 ÷ 149.22	29

V. CONCLUSION

The results of testing simulations show that multi-objective optimization is very effective in AVR/PSS coordinated parameter tuning. In most cases of study multi-objective ant lion optimizer algorithm has proven to be the most solid solution to different problems regarding transient and small signal stability.

When considering a research conducted in [19] every tested MOO algorithm has shown progress with regards to its single-objective variant. Future work in research regarding this problem may use normalization of objective functions and static or dynamic penalty functions which may achieve even better results in power system stability.

REFERENCES

[1] P. Kundur, "Power System Stability and Control", McGraw Hill, USA, 1994

[2] Klein M., Roger G. J., Kundur, P., *"A Fundametal Study of Inter-Area Oscillations in Power Systems"*, IEEE Transactions on Power Systems, Vol. 6, No. 3, August 1991, pp. 914- 921

[3] Kundur P., Paserba J., Ajjarapu V., Andersson G., Bose A., Canizares C., Hatziargyriou N., Hill D., Stankovic A., Taylor C., Van Cutsem T., Vittal V., "*Definition and Classification of Power System Stability"*, IEEE CIGRE Joint Task Force on Stability Terms and Definitions, IEEE Transactions on Power Systems, Vol. 19, No. 2, May 2004, pp. 1387–1401

[4] M.J. Gibbard, P. Pourbeik, D.J. Vowles, *"Small Signal Stability, Control and Dynamic Performance of Power System"*, University of Adelaide, 2015

[5] M. Eremia, M. Shahidehpour, *"Handbook of Electrical Power system Dynamics: Modeling, Stability and Control"*, John Wiley & Sons, USA, 2013

[6] Mirjalili S., *"The Ant Lion Optimizer"*, Advances in Engineering Software, Vol. 83, May 2015, pp. 80-98

[7] B. Ribeiro, R. F. Albrecht, A. Dobnikar, D. W. Pearson, N. C. Steele (editors), *"Adaptive and Natural Computing Algorithms:* Proceedings of the International Conference in Coimbra, Portugal", Springer, Austria, 2005

[8] Hassan L. H., Moghavvemi M., Mohamed H. A. F., *"Power System Stabilization Based on Artificial Intelligent Techniques: A review"*, 2009 International Conference for Technical Postgraduates (TECHPOS), Malaysia, 2009

[9] Murali D., Rajaram M., *"Comparison of Damping Performance of Conventional and Neuro–Fuzzy Based Power System Stabilizers Applied in Multi–Machine Power Systems"*, Journal of Electrical Engineering, Vol. 64, No. 6, 2013, pp. 366–370

[10] Mukherjee V., Ghoshal S. P., *"Velocity Relaxed Swarm Intelligent Tuning of Fuzzy Based Power System Stabilizer"*, IEEE Power India Conference, India, 2006, pp. 108-115

[11] Shayeghi H., Safari A.,Shayanfar H., *"Multimachine Power System Stabilizers Design Using PSO Algorithm"*, International Journal of Electrical Power and Energy Systems Engineering, Vol. 1, No. 4, 2008, pp. 226-233

[12] Shayeghi H., Ghasemi A., Shayanfar H., *"PID Stabillizer Design for Multi Machine Power System Using IPSO Procedure"*, Computer Science and Engineering, vol. 1, 2011, pp. 36-42

[13] Shayeghi H., Shayanfar H. A., Jalilzadeh S., Safari A., "*TCSC Robust Damping Controller Design Based on Particle Swarm Optimization for a Multi-Machine Power System"*, Energy Conversion and Management, vol. 51, 2010, pp. 1873–1882

[14] Mahmoud M. S., Soliman H. M., *"Design of Robust Power System Stabilizer Based on Particle Swarm Optimization"*, Circuits and Systems, vol. 3., 2012, pp. 82-89

[15] Shewtahul H. W., Ah King R. T. W., Rughooputh H. C. S., *"Optimization of AVR Parameters of a Multi-machine Power System Using Particle Swarm Optimization"*, University of Mauritius Research Journal, vol. 16, 2007, pp. 14-28

[16] Gholipour E., Nosratabadi S. M. *"A new coordination strategy of SSSC and PSS controllers in power system using SOA algorithm based on Pareto method"*, Electrical Power & Energy Systems, v. 67, May 2015, pp. 462-471

[17] Mirjalili S., Gandomi A. H., Mirjalili S. Z., Saremi S., Faris H., Mirjalili S. M., *"Salp Swarm Algorithm: A Bio-Inspired Optimizer for Engineering Design Problems"*, Advances in Engineering Software, Vol. 114, December 2015, pp. 163-191

[18] Mirjalili S., Jangir P., Saremi S., *"Multi-objective Ant Lion Optimizer: A Multi-objective Optimization Algorithm for Solving Engineering Problems"*, Applied Intelligence, Vol. 46, July 2016, pp. 79-95

[19] Špoljarić T., Pavić I., *"Performance Analysis of an Ant Lion Optimizer in Tuning Generators' Excitation Controls in Multi Machine Power System"*, Proceedings of 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, May, 2018., pp. 1040-1045

Integrated Production Optimization at INA d.d. Upstream supported by Osisoft PI

B. Žeželj *, Ž. Frkin**, Florentin Beznea***, M. Gojić****
*Plavi Tim d.o.o., Vukovarska 18, 10000 Zagreb
boris.zezelj@plavitim.hr
** Plavi Tim d.o.o., Vukovarska 18, 10000 Zagreb
zeljko.frkin@plavitim.hr
***FIA Integra, 010717 Bucharest Romania
florentin.beznea@fiaintegra.ro
**** IoT Net Adria d.o.o.
marko.gojic@iotnet.hr

Abstract - Following the success of MOL Group PI system implementation logical choice was to choose the same approach on both INA Downstream and Upstream. INA Upstream recognizes the importance and potential of implementing PI system to enable digital transformation as a necessary precondition for the Digital oil field concept. Implementation of the Try Buy PI system took a total of twelve working days, with the benefit of continuing to be used existed PI ecosystem after acquiring permanent licenses and expanding it to an additional four oilfields in 2020. Aim of PI system in Integrated Production Optimization 4.0 (IPO 4.0) project is to ensure continuous and secure data flow from oilfields to business oriented expert systems for modelling, optimization and diagnostics engine of the production system. IPO 4.0 goal is to decrease time of the reaction to underperforming production, enabler for advanced well surveillance and collaborative engineering environment increasing production between 3-5%.

I. INTRODUCTION

Exploration and Production Business at INA d.d. is engaged in exploration, development and production of oil and natural gas in Croatia and abroad. More intensive oil and gas exploration and production in Croatia began after World War II. Former INA Naftaplin, which is synonym for numerous oil and gas fields discovery was founded more than 60 years ago, in 1952. Since it was founded it has been involved in many oilfield operations in Croatia and 20 foreign countries, currently in Angola and Egypt.

During more than 60 years of existence 45 oil and 30 gas fields were equipped and put into operation, around 4,500 exploration and development wells with total depth of about seven million meters, about 1,200 producing oil wells with more than 4,000 km of various pipelines and more than 200 production gas wells and oil facilities and installations were built. Until these days, Naftaplin's production was and still is between 50 and 60 percent of total domestic primary energy production. INA is strongly dedicated to its exploration and production activities in Croatia and is at the moment the only company with the required equipment, experience, knowledge and projects that can speed up exploration activities in Croatia.

Since 2010. INA have drilled 34 wells, discovered three land oil fields and one smaller offshore gas field and confirmed a gas condensate field on the border between Croatia and Hungary. In 2016 company put into production two gas fields in Međimurje and plan on putting a third one in 2017. In the last two years, with its activities and investments in new technologies INA has managed to increase production of Croatian exploitation fields, switching the trend of natural production decline and sure this will continue in the coming years because on June 10, 2016 the Company signed the onshore exploration and production sharing agreement (Block Drava-2) with the Croatian Government.

INA has been present in the northern Adriatic since the 1990's and the production has begun in 1998. Current exploration activities in the northern Adriatic and Aiza Laura Contract areas are performed in cooperation with Italian oil company Eni through joint company INAgip. More than one billion dollars have been invested since the beginning of the project. Also, in the northern Adriatic, in cooperation with the Italian company Edison through joint company Edina, gas from gas field Izabela is exploited.

In the summer of 2016, INA launched test gas production at two gas production fields Vuckovec and Zebanec. The Vukanovec production field will follow early in 2017. The commercially recoverable gas volumes are estimated at roughly 1 billion m³, to be produced by 2024. This increase in INA's total domestic gas production is significant both for the company and for Croatia.

The foundations for new production were laid following the successful implementation of the Medjimurje project, in which works on building the necessary mining and accompanying facilities were completed in two years. Five production wells were drilled and equipped, and 100km of new pipelines were built, including 73km of the evacuation gas pipeline used to transport the gas produced in Medjimurje to the Central Gas Station Molve for further processing and purifying. The successful implementation of the Medjimurje project, in which INA invested HRK 450 million, also means the

return of hydrocarbon production to Medjimurje after 25 years.

INA's gas production in Medjimurje is in synergy with the company's oil production in Moslavina. The carbon dioxide separated from the Medjimurje gas will no longer be released to atmosphere, but instead be shipped via the pipeline to the Zutica and Ivanic oil fields, where it will be re-injected in oil wells to help produce additional oil volumes as part of the EOR project. The closed production system used in the Medjimurje facilities is important for environmental protection, which is in line with INA's commitment to sustainable development to the benefit of all the stakeholders.

Purpose of this paper is to show how it is carried out process of digital transformation as a necessary precondition for the digital oilfield concept implementation. As the digital revolution swept through many industries, it has disrupted established business models and created new winners and losers as fast movers have leveraged the digital age to gain competitive advantage. Just as the Internet has transformed the information and media industries, smartphones have revolutionized communications and social media, and the Industrial Internet has turned assets like jet engines and locomotives into data churning "smart" machines that "talk" to each other and self-diagnose problems, the Digital Oilfield is poised to transform the oil and gas industry.

The Digital Oilfield represents a business-wide reimagining of the oilfield. It requires business leaders to step back from the traditional view of technology advancement siloed within individual departments and adopt a cross-organization approach.

The Digital Oilfield is not merely about computer chips, processors and software. It is about the melding of operations technology with information technology and the Internet of Things. It involves a powerful combination of distributed network sensors, ubiquitous mobile connectivity, cloud computing, advanced big data analytics and artificial intelligence. It could "learn" from what works in the best producing wells and apply those learnings to entire fields. It will predict equipment breakdown before it happens and bring about "condition-based" maintenance rather than "schedule-based" methods. It will track workers in the field, feed them the data they need via various platforms, "coach" their work in real-time and remove them from hazardous situations. Ultimately, it will produce more oil and gas for less cost.

The Digital Oilfield is not a technology play but the combination of several technologies in innovative ways to utilize technology to drive productivity. In short, it requires taking a different company-wide approach to today's challenges. Challenge of digital transformation is only weakly related to product selection. The most important step is to deliver fundamental change in a company's world view. This means doing things differently in order to move to a modern working environment, we must embrace new ways of working and new ways of thinking. Even after all this time, improving collaboration within and across teams, and across companies is still one of the fundamentals.

Supporting this effort three technology pillars of transformation were recognized:

1. Machine learning
2. Industrial Internet of Things
3. Cloud

II. DIGITAL TRANSFORMATION

Following the successful implementation of the Osisoft PI system in the MOL Group refineries, the logical choice was to implement PI system in Rijeka oil refinery, which began in 2011 and successfully completed in 2012.

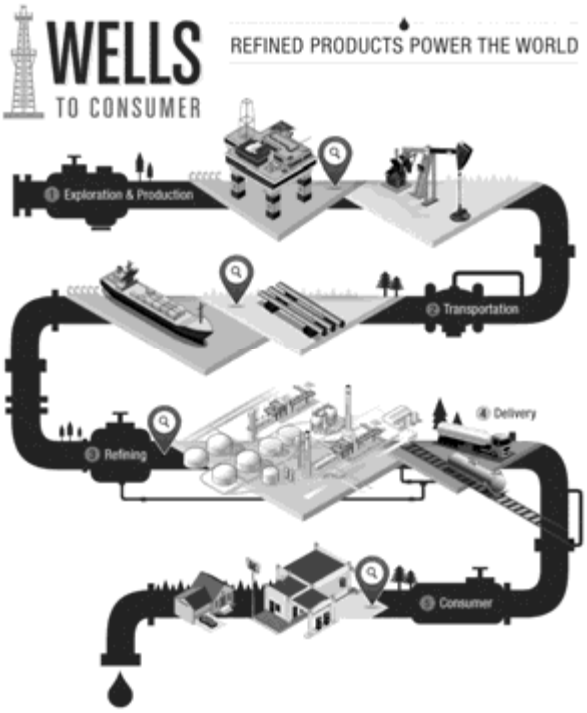


Figure 1. Oil & Gas Process Lifecycle - by Design Assistance Corporation (DAC)

Following the same example based on the established standard it happened that INA Upstream also recognizes the importance and potential of implementing PI system to enable digital transformation as a necessary precondition for the application of Digital oilfield technologies. After some discussions it happened that Osisoft company recognized the importance of the moment and in collaboration with INA made it possible to implement a pilot project at INA Upstream on the Medjimurje gas field based on a 10,000-tags PI server, a PI Vision installation for 10 client computers, a PI Data link installation for 10 client computers, one OPC DA interface connection, one RDBMS interface connection, and one PI integrator license for business analytics. Implementation of the PI ecosystem with OPC DA connectivity to the ABB DCS system took a total of twelve working days, with the greatest benefit for INA Upstream to continue to use the

implemented PI ecosystem after expanding it to an additional four selected oil fields which implementation is in plan for Q2/2020.

Pilot project of PI system implementation at Medjimurje cluster field is selected due to low production system complexity (5 gas producing wells, 1 water disposal injection well) with already installed well surveillance system connected to local SCADA (wellhead pressures, temperatures and production rates) with associated Integrated Production Modelling which is regularly updated twice per year.

Additionally at the same time the proof of concept pilot implementation of the IoT devices on the sucker rod pumps was used and so far, non-remotely monitored pumps have the ability to display the status of the pressure switch (detecting pump failure due to the triggering of installed pressure switches caused by abnormally high pressure in the connecting pipeline), pump general condition and bimetal switch (detecting pump failure due to the triggering of installed bimetal switches caused by the pump pulling abnormally high current) on the mobile device screen provided by the PI vision server. Complete implementation was done using Sigfox cloud, Azure IoT HUB and finally sending json format file to PI UFL connector.

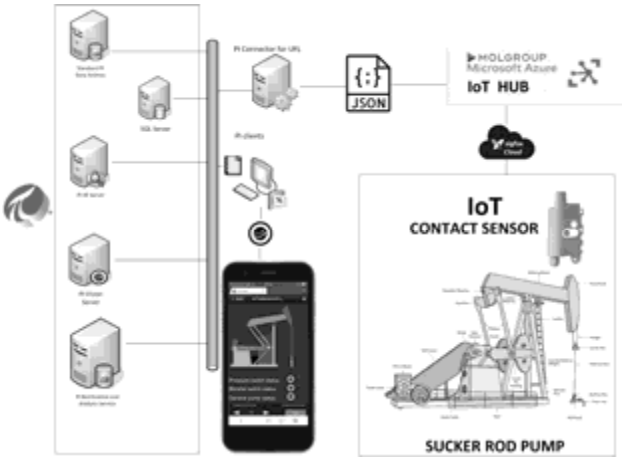


Figure 2. IoT detectors of sucker rod pump activity integration with mobile platform

Along with the PI system implementation in Q2/2020 on four selected INA Upstream oil fields there is ongoing preparation for Integrated Production Optimization 4.0 project which in first phase includes Medjimurje gas field cluster. Aim of PI system in Integrated Production Optimization 4.0 project is to ensure continuous and secure data flow from oil fields to business oriented expert systems for modelling, optimization and diagnostics engine of the production system. It will allow achieving higher engineering efficiency along with improved analytics in considerably shorter time.

III. PI SYSTEM IMPLEMENTATION

The aim was to implement Osisoft PI system that will help collect information from various data islands into one unique Real Time data base via standard interfaces which provide the users data history, findings, analysis, delivering and visualization. Project goal was to ensure continuous and secure data flow from oil fields to business oriented expert systems for modelling, optimization and diagnostics engine of the production system thus achieving higher engineering efficiency along with improved analytics in considerably shorter time.

Whole PI software system was built on Vmware "green IT" infrastructure located on company data center (four times server machines in total each with 8-core CPU, 2,6 GHZ, RAM 8 GB) using standardized company database SQL cluster farm and single interface machine for connection to ABB System 800xA distributed control system using PI OPC DA (data access) interface.

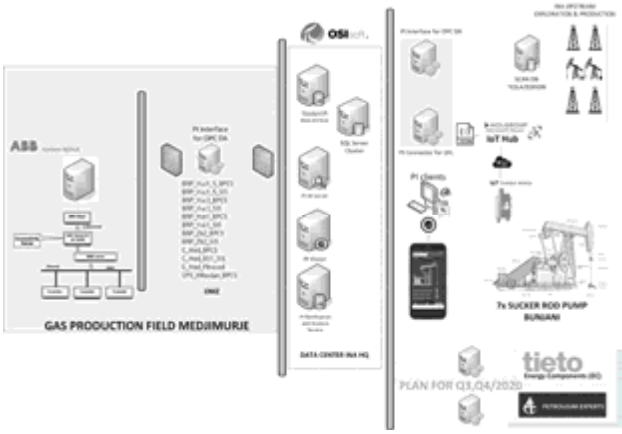


Figure 3. Medjimurje gas field PI system architecture

Implemented system can cover in future different areas of Upstream Oil and Gas business such as:

1) Production and Operations
   - Production Optimization
   - Operational Intelligence
2) Maintenance and Integrity
   - Condition Based & Predictive Maintenance
   - Real-time Equipment Monitoring
   - Inspections and Integrity Management
3) Drilling & Wells
   - Real Time Drilling Data
   - Down Hole Measurements
   - Well Integrity & Well Completions

IV. INTEGRATED PRODUCTION OPTIMIZATION

Expected impact of Integrated Production Optimization 4.0 project implementation is to decrease time of the reaction to underperforming production, enabler for advanced well surveillance and collaborative engineering environment. Financial impact is based on benchmark from implementation currently used in the world and it shows that implementation can increase production between 3 and 5 % with reducing costs in well testing and chemical consumption. Integrated operation project 4.0 is Digital Oil Field technology implementation for engaging engineering workflows that can be categorized in 5 groups by calculation frequency.

High frequency workflow (virtual flow metering) is calculated estimation of production rate in real time that is enabler for better well surveillance and detecting underperforming production in real time.

Medium frequency workflows (multi well allocation) is recalculation of well production based on measured total production allocated to the wells based on pressure and mass calculation from IPM. Virtual flow metering and multi well production allocation is in conjunction as quality check for production data base Energy Component TIETO.

Low frequency workflows (production forecasting) is based on IPM (numerical simulation or material balance) or based on decline curve analysis (DCA). Event driven workflows are calculations of underperforming events and optimizations. On demand workflows are used for field development cases scenarios and investigation of their impact with economic feasibility calculations.

PI system on DOF system diagram by Petroleum Experts figure will act important role between Energy components TIETO software and DOF based systems.

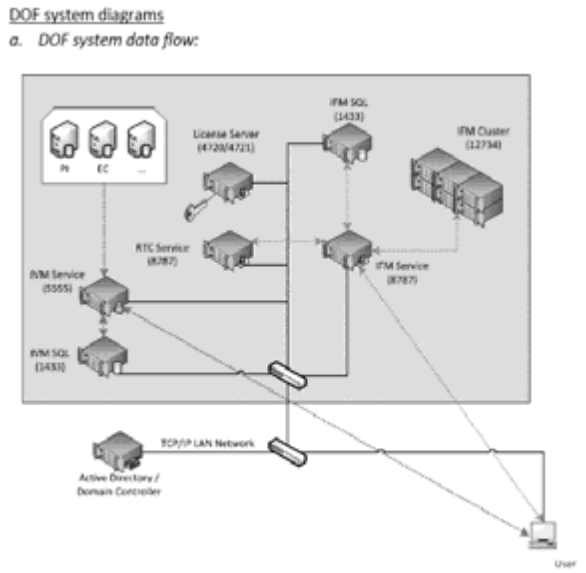


Figure 4. DOF system diagram by Petroleum Experts

V. CONCLUSION

The described paper showed most recent digital transformation example from Oil & Gas industry in INA d.d. company. As it was mentioned Digital Oilfield is combination of several technologies in innovative ways to utilize technology to drive productivity. INA d.d. recognize potential to utilize three technology pillars of transformation which are: machine learning, industrial internet of things and cloud technologies. Two of them are represented in ongoing INA Data innovation project initiative while third one is used as platform where the previous two can meet.

Seems that simplest part of whole concept is to develop central store database called Osisoft PI system which opens up many other possibilities for manipulating data such as machine learning in which area especially for Upstream Oil and Gas business use cases many scientific papers are written and some of them were developed and used in daily production by the leading Oil & Gas industry companies. Basic function and purpose of introducing PI system is to make data from the field available for Digital oil field. The overall objectives of the Digital oil field system are to have a company-wide view of producing assets, from the reservoir to the export point, which would allow the following objectives to be achieved:

1) Field Surveillance to understand at any given time the production from any asset at all levels (Reservoir, Well, Surface Network, Process).
2) Operational Efficiency achieved through surveillance of the field, which will then be used to streamline the operational scheduling of field-based testing (Well, Surface and Process) and capture uptime/downtime accurately.
3) Production Management reconciling and acting on the opportunities that exist at a company level between the field/assets and producing fields in a timely manner (Production guidelines).
4) Optimization to increase production in a measurable and sustainable way, whilst detecting and mitigating any limiting factors (Bottlenecks) that impede this objective from being achieved.
5) Field Development Strategies to maximize long term production and highlight long term opportunities that may exist (Planning).

REFERENCES

[1] INA-Industrija nafte, d.d., Istraživanje i proizvodnja nafte i plina (https://www.ina.hr/o-kompaniji/temeljne-djelatnosti/istrazivanje-i-proizvodnja-nafte-i-plina/28)

[2] Digital Oilfield Outlook Report, Opportunities and challenges for Digital Oilfield transformation, Research study conducted by JWN and commissioned by GE Canada and Accenture

[3] DOF IT Architecture Quick Guide - Petroleum Experts ©Petroleum Experts 2018 Version 1.1

Predictive lane-keeping system for an autonomous vehicle

P. Makarun, Š. Ileš, J. Matuško
University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
Email: {petar.makarun, sandor.iles, jadranko.matusko}@fer.hr

***Abstract*—In this paper, a model predictive control lane-keeping system of an autonomous vehicle is proposed. The control algorithm was tested experimentally on a laboratory scaled model of the vehicle. The desired trajectory during the experiment was determined by computer vision. Using image processing the middle of the driving lane and the horizontal profile of the driving lane was recognized. The control algorithm on the vehicle model was implemented in Python and solved using OSQP mathematical programming solver.**

***Index Terms*—Model predictive control, Lane-keeping, Autonomous driving**

I. Introduction

The idea of the autonomous car is over 100 years old. Leading manufacturers in the automotive industry are already investing in research and development of autonomous vehicles. Even though there are fully autonomous car prototypes, this is still an active research area both in academia and industry [1]. The main advantage of autonomous vehicles over manually driven vehicles is that the human impact on the traffic is reduced. It is expected that increasing traffic autonomy and eliminating human impact would reduce the number of accidents and make traffic safer [2]. There are six levels of autonomy according to the J3016 international standard for measuring the degree of automation in cars issued by the *Society of Automotive Engineers (SAE)* [3]. A vehicle with levels from 0 to 2 contains only advanced driver assistance systems (ADAS), while a vehicle with levels 3 to 4 is a vehicle that can drive autonomously in limited conditions, while the highest level represents complete autonomy in all conditions [3].

Lane Keeping System (LKS) is a central feature of autonomous cars. It belongs to level 1 autonomy, and combined with adaptive cruise control makes level 2 autonomy. For LKS, it is essential to get information about vehicle position in the lane and to identify the lane borders. It usually employs a color camera and computer vision techniques such as Hough transform, Edge detection, etc. or even neural network-based techniques to get this information. The next component is a path planning algorithm, where the goal is to provide a reference for the control algorithm [4], [5].

The control algorithms for lane-keeping vary from simple linear PID controllers to advanced control algorithms such as sliding mode control (SMC) or model predictive control (MPC) [6]–[9].

MPC is a control method that uses a model of the system to predict its future behavior and it solves an optimization problem over the finite horizon at each sampling instant to obtain the optimal finite-horizon control sequence that ensures the optimal finite-horizon trajectory. The first element in the sequence is applied and the process is repeated. One of the main advantages of such type of control is its ability to satisfy the constraints [10]. Autonomous vehicles have been proven to be a suitable area for the application of MPC due to their high performance and safety requirements [11], [12].

Since at each sampling instant an optimization problem has to be solved, the speed of solving is one of the main practical aspects of the MPC. The algorithm should be tested on an experimental setup to investigate for possible losses of feasibility due to the discrepancy of the model and the experimental setup. Furthermore, the speed of solving the optimization problem has to be investigated on the experimental platform. For experimental testing of various control algorithms in the literature various experimental setups have been used. They vary from scaled car models [13], [14] to full-sized car [8].

In this paper, model predictive control is applied to maintain a car within the traffic lane. The paper uses a linear mathematical model and a quadratic objective function. OSQP solver for quadratic programming [15] is used to solve the optimization problem in the simulation and in the experiments. The correctness of the mathematical model and the operation of the algorithm were tested in simulation, while the speed of solving the corresponding optimization problem and the performance of the algorithm was tested in real-time on a custom build scaled vehicle model. Such a custom build scaled vehicle model is suitable for students to design and test the control algorithms.

The paper is organized as follows: In Section II, the mathematical model of the car is given. In Section III. image processing and path generation is described. In Section IV, the model predictive control algorithm is described. Section V. presents the results and Section VI. concludes the paper.

II. Control oriented mathematical model of a vehicle

A. Bicycle model of a vehicle

In this paper, a simplified Bicycle model of the vehicle shown in Fig 1 is used. The model has two degrees of freedom. These are the lateral position of the vehicle y and the rotation of the vehicle around the vertical axis ψ. The lateral position of the vehicle is measured from the imaginary point O, which is also the center of rotation of the vehicle. The front steering

angle is denoted by δ and the desired yaw is denoted by ψ_{des}. The longitudinal speed of the vehicle is denoted by v_x. The l_f and l_r denote the distance of the front and rear wheel from the center of gravity respectively. The mass of the vehicle is denoted by m and the moment of inertia around the z-axis is denoted by I_z.

Assuming small slip angles, the lateral friction forces acting on the front and rear wheel F_{yf} i F_{yr} are assumed to be described by the following expressions:

$$F_{yf} = 2C_{\alpha f}(\delta - \theta_{vf}), \tag{1}$$

$$F_{yr} = 2C_{\alpha r}(-\theta_{vr}), \tag{2}$$

where $C_{\alpha f}$ and $C_{\alpha r}$ denote cornering stiffness of the front and rear tire respectively and θ_{vf} and θ_{vr} represents the angle between the the velocity vector and the longitudinal axis of the vehicle. The state space model is written in terms of control errors:

$$\dot{e} = A(v_x)e + B\delta + B_{ref}(v_x)\dot{\psi}_{des}, \tag{3}$$

where $e = [e_1\, \dot{e}_1\, e_2\, \dot{e}_2]^T$ and $\ddot{e}_1 = \ddot{y} + v_x(\dot{\psi} - \dot{\psi}_{des})$, $e_2 = \psi - \psi_{des}$ and matrices $A(v_x)$, B and B_{ref} are given as follows:

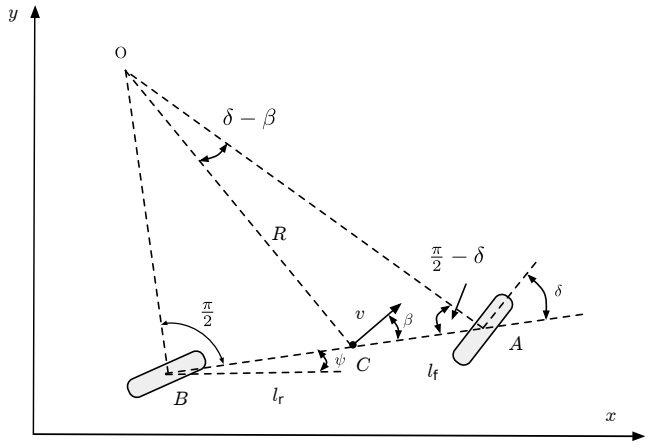


Fig. 1. Bicycle model of a vehicle

$$A(v_x) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -\frac{2C_{\alpha f}+2C_{\alpha r}}{mv_x} & \frac{2C_{\alpha f}+2C_{\alpha r}}{m} & -\frac{2C_{\alpha f}l_f - 2C_{\alpha r}l_r}{mv_x} \\ 0 & 0 & 0 & 1 \\ 0 & -\frac{2C_{\alpha f}l_f - 2C_{\alpha r}l_r}{J_z v_x} & \frac{2C_{\alpha f}l_f - 2C_{\alpha r}l_r}{J_z} & -\frac{2C_{\alpha f}l_f^2 + 2C_{\alpha r}l_r^2}{J_z v_x} \end{bmatrix} \tag{4}$$

$$B = \begin{bmatrix} 0 \\ \frac{2C_{\alpha f}}{m} \\ 0 \\ \frac{2C_{\alpha f}l_f}{J_z} \end{bmatrix}, B_{ref} = \begin{bmatrix} 0 \\ -\frac{2C_{\alpha f}l_f - 2C_{\alpha r}l_r}{mv_x} - v_x \\ 0 \\ -\frac{2C_{\alpha f}l_f^2 + 2C_{\alpha r}l_r^2}{J_z v_x} \end{bmatrix} \tag{5}$$

For more details about the used model, the reader is referred to [16]

III. Model predictive control

In this paper, we treat longitudinal and lateral vehicle control separately. For keeping the vehicle speed as close as possible at the desired reference, a PI controller is used, where the controller gains are selected experimentally. For such a model, its closed-loop response is recorded and used for predicting future values of the vehicle velocity.

The PI controller is responsible for providing the prediction of the longitudinal velocity for the MPC controller. In that way matrices in system model (5) become constant matrices and the model is discretized at each step. The proposed control scheme is shown in Fig 2.

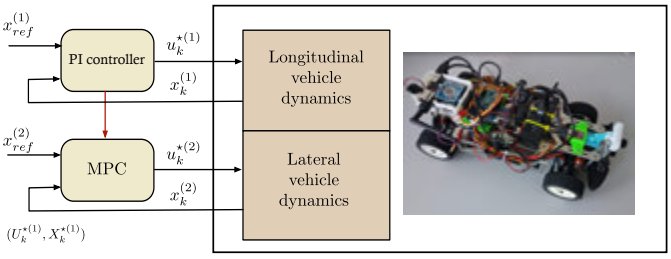


Fig. 2. Proposed control scheme

For the lateral vehicle dynamics control the following objective function is chosen

$$J = e_{k+N}^T P e_{k+N} + \sum_{j=0}^{N-1} \left(e_{k+j}^T Q e_{k+j} + \delta_{k+j}^T R \delta_{k+j} \right) \tag{6}$$

The matrix Q is a diagonal matrix with positive coefficients on its diagonal:

$$Q = diag(\begin{bmatrix} q_1 & q_2 & q_3 & q_4 \end{bmatrix}) \tag{7}$$

where the diagonal terms penalize e_1, e_2 and its derivatives $\dot{e}_1$, $\dot{e}_2$ respectively. The matrix R is positive matrix which penalize the steering angle while matrix P is a positive definite matrix representing the terminal cost. The only constraint used for lateral vehicle dynamics is the control input constraint:

$$\delta \in [-17.5°, 17.5°]. \tag{8}$$

IV. Line detection algorithm

To test the algorithm experimentally a track is made of green electrical insulating tape. The first step in the line detection algorithm is to undistort the raw image. The next step is to isolate track on the image using color segmentation in the HSV color space. The image was then transferred to a bird's eye view, and the sliding window method is used to identify points of interest and eliminate the background noise. A 3rd order spline is identified trough the identified points, and the curvature of the line was calculated. Both the distance from the centerline and the orientation error is computed.

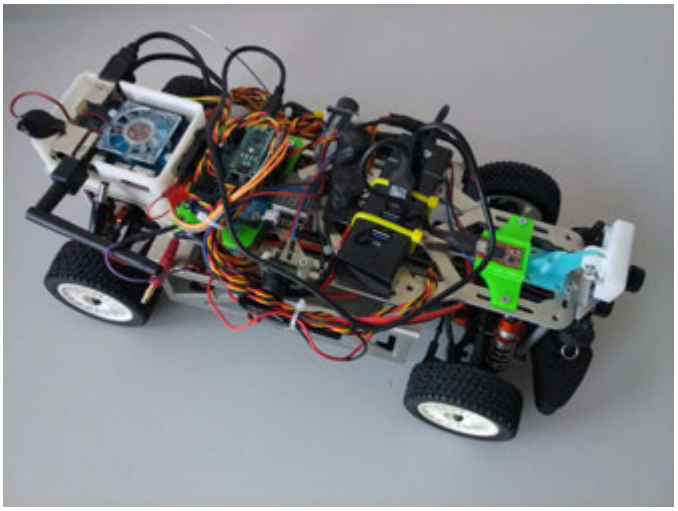


Fig. 3. The scaled model of the car

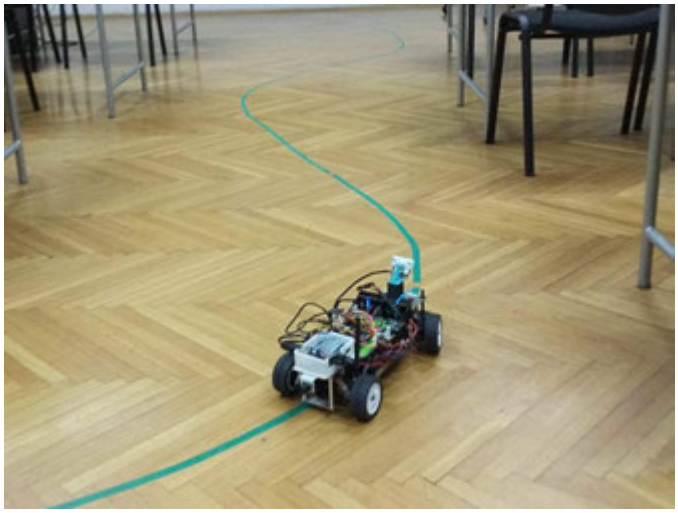


Fig. 4. Experimental driving on the track in the laboratory

TABLE I
PARAMETERS OF THE RC CAR

Parameter	Value	Parameter	Value
$m[kg]$	1.95	$I_z[kgm^2]$	0.24
$C_{\alpha f}[N/rad]$	1.76	$C_{\alpha r}[N/rad]$	1.76
$L_f[m]$	0.125	$L_r[m]$	0.125

V. EXPERIMENTAL SETUP

For the experimental results, a custom-built scaled car shown in Fig. 3 is used. The car was inspired by a similar car [17]. For such car the parameters are given in [18] and are shown in Table I.

The main mechanical components of the car are the chassis, suspension, transmission, and wheels. For producing the driving torque a brushless DC electric motor is used together with an electronic speed controller. For steering, a servo motor is used.

The control part of the electronics consists of an ODROID computer and an Arduino microcontroller. The Arduino microcontroller retrieves data from sensors and sends control values to actuators. The ODROID computer processes data from sensors, evaluates system status, and determines control signals. The state of the system is estimated using sensor data. There is an inertial sensor, a camera, and a hall sensor for sensing the speed of the vehicle.

VI. RESULTS

Fig 4 shows the experimental setup and the corresponding track made of a green electrical insulating tape. The radius of the curves is close to the minimal kinematic radius of the vehicle. Fig 5 shows the process of image processing of the data seen by the camera mounted on the scaled RC car.

For the experiments the following values of the weights are selected for the objective function: $q_1 = 10$, $q_2 = 1$, $q_3 = 1$, $q_4 = 1$, and $r_1 = 1, r_2 = 1$. In that way a higher importance is put on tracking the lateral error. The reference signal was handled as a input with an equality constraint.

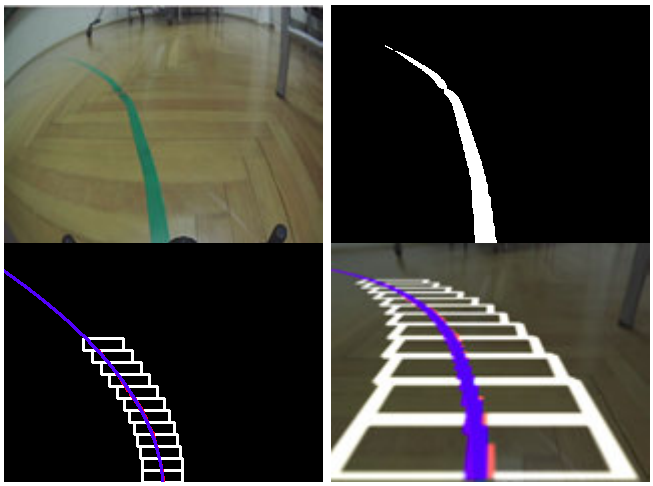


Fig. 5. Line detection experiments

For the cases when the track is outside of the field of view of the camera, the last know spline is used to provide a reference signal.

Fig. 6-10 show experimental results for control of lateral vehicle dynamics with sampling time $T = 0.1$ s, and with prediction horizon length $N = 10$. In Fig. 6 a constant lateral error can be observed during cornering. Fig. 7 shows the lateral error rate. The peak around 37th second occurs when the car is turned manually in the opposite direction to travel back over the same path. The steering angle is shown in Fig. 10. It hits the upper and lower limit since the radius of the curves is close to the minimum kinematic radius of the vehicle. Fig 8 and 9 shows the yaw and yaw rate error respectively. It can be observed that there is a relatively small yaw error during the whole experiment. However, during sharp corners, the track was not in the field of view of the camera. This can be observed around 57th second in Fig 8. Fig. 11 shows

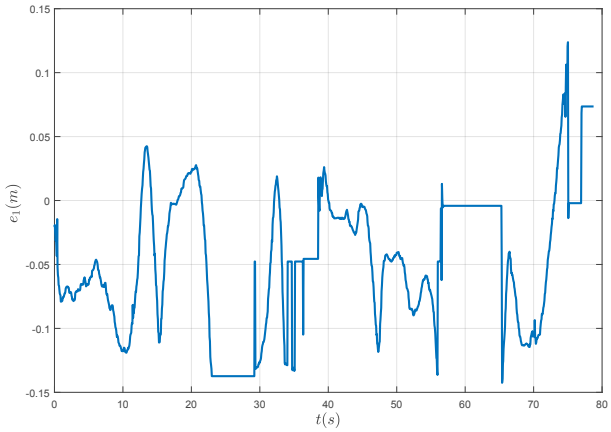


Fig. 6. Experimental results: Lateral error

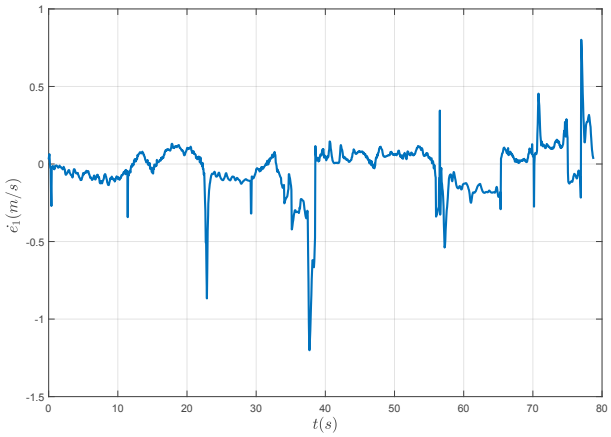


Fig. 7. Experimental results: Lateral error rate

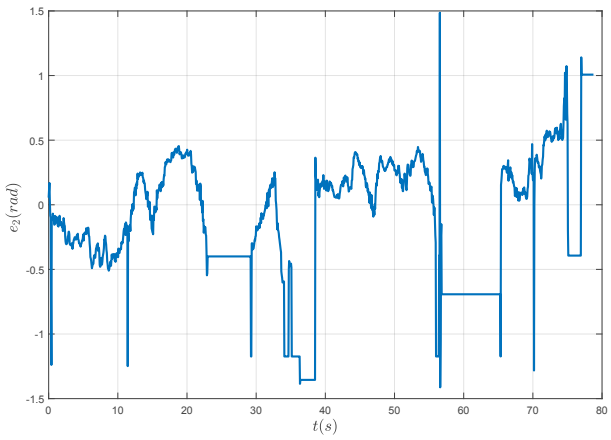


Fig. 8. Experimental results: Yaw error

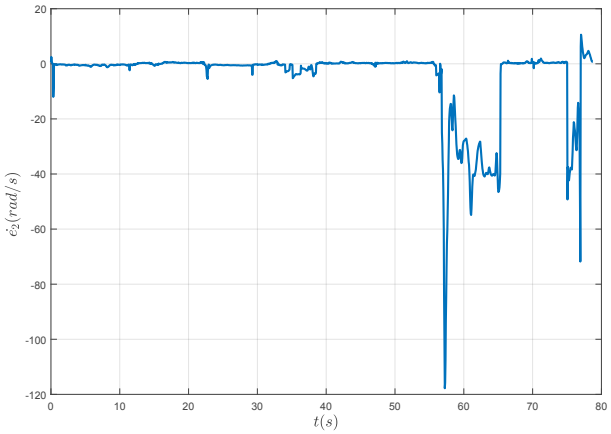


Fig. 9. Experimental results: Yaw error rate

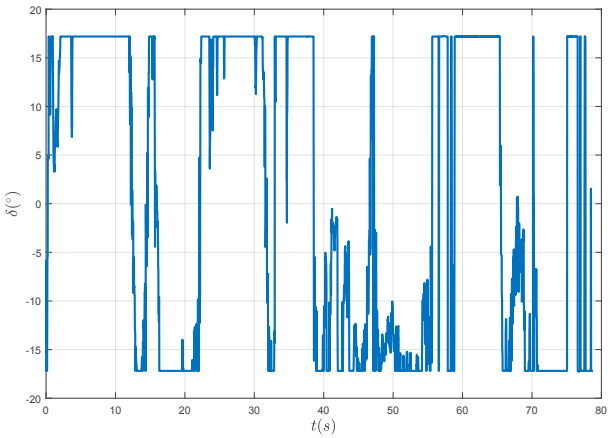


Fig. 10. Experimental results: Steering angle

longitudinal vehicle dynamics control. The speed setpoint is chosen to be 0.25 m/s. Such low speed is selected since the experiments are conducted in a small interior space with obstacles as shown in Fig 4. There are outliers in the measurement of the vehicle speed which can be observed around 23rd and 71st second. Such a low speed is hard to maintain, due to noise introduced by hall sensors and the fact that there are dominant friction forces around zero velocity. The average time required for solving the optimization problem is 0.003 s which is fast enough for real-time experiments.

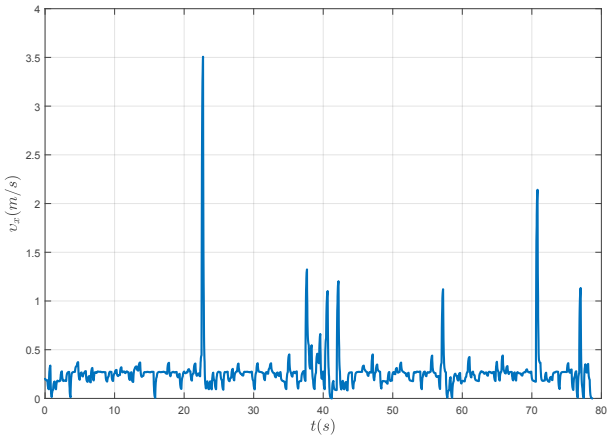


Fig. 11. Experimental results: Vehicle speed

VII. Conclusion

In this paper, a predictive lane-keeping system is presented. A simple controller is used for longitudinal dynamics control and the prediction of the vehicle speed is used to form the system matrices for lateral vehicle dynamics control. The controller was tested experimentally on a scaled RC car. A centerline detection algorithm was developed for a scaled vehicle model and the image was transferred to a bird's eye view where the curvature of the line was calculated. The control algorithm for the scaled vehicle model was tested and the time required to solve the optimization problem was observed. During the experiment, the optimization problem was solved using OSQP mathematical programming solver. To test the algorithm a track is made of green electrical insulating tape and the algorithm is tested in the laboratory. It was shown by simulation that the algorithm keeps the vehicle within the traffic lane. The experimental results suggest that the algorithm can be solved quickly enough and it can track the identified centerline. However, the paper shows only preliminary results. In the future work, a model predictive control of longitudinal vehicle dynamics is predicted to obtain the prediction of the vehicle velocity over the prediction horizon.

Acknowledgement

This work has been supported in part by Croatian Science Foundation under the project UIP-2019-04-6487.

References

[1] M. Walch, T. Sieber, P. Hock, M. Baumann, and M. Weber, "Towards cooperative driving: Involving the driver in an autonomous vehicle's decision making," in *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2016, pp. 261–268.

[2] R. Hussain and S. Zeadally, "Autonomous cars: Research results, issues, and future challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1275–1313, 2018.

[3] S. International, "Automated driving: levels of driving automation are defined in new sae international standard j3016," 2014.

[4] Z. Chen and X. Huang, "End-to-end learning for lane keeping of self-driving cars," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 1856–1860.

[5] M. Samuel, M. Mohamad, S. M. Saad, and M. Hussein, "Development of edge-based lane detection algorithm using image processing," *JOIV: International Journal on Informatics Visualization*, vol. 2, no. 1, pp. 19–22, 2018.

[6] R. Marino, S. Scalzi, and M. Netto, "Nested pid steering control for lane keeping in autonomous vehicles," *Control Engineering Practice*, vol. 19, no. 12, pp. 1459–1467, 2011.

[7] H. Du, Z. Man, J. Zheng, A. Cricenti, H. Wang, and Y. Zhao, "A novel sliding mode control for lane keeping in road vehicles," in *2016 International Conference on Advanced Mechatronic Systems (ICAMechS)*. IEEE, 2016, pp. 289–294.

[8] V. Turri, A. Carvalho, H. E. Tseng, K. H. Johansson, and F. Borrelli, "Linear model predictive control for lane keeping and obstacle avoidance on low curvature roads," in *16th international IEEE conference on intelligent transportation systems (ITSC 2013)*. IEEE, 2013, pp. 378–383.

[9] B.-C. Chen, B.-C. Luan, and K. Lee, "Design of lane keeping system using adaptive model predictive control," in *2014 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE, 2014, pp. 922–926.

[10] D. Q. Mayne, "Model predictive control: Recent developments and future promise," *Automatica*, vol. 50, no. 12, pp. 2967–2986, 2014.

[11] R. Verschueren, S. De Bruyne, M. Zanon, J. V. Frasch, and M. Diehl, "Towards time-optimal race car driving using nonlinear mpc in real-time," in *53rd IEEE conference on decision and control*. IEEE, 2014, pp. 2505–2510.

[12] F. Borrelli, P. Falcone, T. Keviczky, J. Asgari, and D. Hrovat, "Mpc-based approach to active steering for autonomous vehicle systems," *International Journal of Vehicle Autonomous Systems*, vol. 3, no. 2, pp. 265–291, 2005.

[13] D. I. Katzourakis, I. Papaefstathiou, and M. G. Lagoudakis, "An open-source scaled automobile platform for fault-tolerant electronic stability control," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 9, pp. 2303–2314, 2010.

[14] C. E. Beal and J. C. Gerdes, "Model predictive control for vehicle stabilization at the limits of handling," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 4, pp. 1258–1269, 2012.

[15] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd, "Osqp: An operator splitting solver for quadratic programs," in *2018 UKACC 12th international conference on control (CONTROL)*. IEEE, 2018, pp. 339–339.

[16] R. Rajamani, *Vehicle dynamics and control*. Springer Science & Business Media, 2011.

[17] J. Gonzales, F. Zhang, K. Li, and F. Borrelli, "Autonomous drifting with onboard sensors," in *Advanced Vehicle Control: Proceedings of the 13th International Symposium on Advanced Vehicle Control (AVEC16)*, 2016, p. 133.

[18] F. Zhang, J. Gonzales, K. Li, and F. Borrelli, "Autonomous drift cornering with mixed open-loop and closed-loop control," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 1916–1922, 2017.

Sliding Mode Control of Custom Built Rotary Inverted Pendulum

M. Švec, Š. Ileš, J. Matuško
University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
Email: {marko.svec, sandor.iles, jadranko.matusko}@fer.hr

***Abstract*—Inverted pendulums have been used for a long time as a typical laboratory setup for learning and testing of various control system techniques. In this paper, we present a new laboratory rotary inverted pendulum setup. The setup is equipped with a microcontroller containing a code for communicating directly with MATLAB/Simulink using Simulink Desktop Real Time. In this configuration students have the opportunity to implement a controller in MATLAB/Simulink or to implement it directly on the microcontroller. Its capabilities are tested using sliding mode control, first using simulation and then verified on the experimental setup.**

***Keywords*—inverted pendulum, sliding mode control, Furuta pendulum**

I. Introduction

The rotary inverted pendulum is a commonly used benchmark in control systems laboratories for learning and testing of various control system techniques. It consists of an arm driven by a motor which is able to rotate in the horizontal plane and a pendulum suspended from the arm which is able to rotate in the vertical plane. It is also called Furuta pendulum. It has one input and two degrees of freedom and belongs to a class of nonlinear unstable underactuated mechanical systems [1], [2].

For Furuta pendulum a large number of papers are written presenting different linear and nonlinear controllers for swing-up and stabilizing the pendulum in the upright unstable position.

In [3], [4] Linear Quadratic Regulator (LQR) is used for stabilizing the pendulum in the upright position, while additional swing-up controller is used for swinging up the pendulum. In presence of nonlinearities, such as dead-zone caused by static friction, the performance of the controller degrades and a limit cycle can occur. For such case, the limit cycle can be avoided by designing a linear controller in the frequency domain taking into account nonlinearity described by describing function [5]. A design based on differential flatness and root locus is used for selecting state feedback gains in [6].

Besides linear control techniques, various nonlinear techniques are also reported in the literature [7]. Model predictive control is used for swinging up and/or stabilizing the pendulum in upright position [8]–[10]. Proportional retarted controller is reported successful in stabilizing the pendulum in case of time delays [11].

Among nonlinear control techniques the sliding-mode control is a popular nonlinear control technique due to robustness against system uncertainties and external disturbances. However, standard sliding-mode control is not directly applicable for underactuated systems without selecting an appropriate sliding surface [12]. Sliding mode controllers are reported for both swing-up and stabilization of Furuta pendulum in [12]–[14], proving semiglobal asymptotic stability. Furthermore sliding mode controllers are often used for comparison with other controllers such as active disturbance rejection control combined with differential flatness [15].

The simulation model can be written in terms of differential equations representing dynamics of the system, which can be obtained by using Euler-Lagrange or Newton-Euler methods [16]–[18], while some authors resort to making a 3D model in CAD software such as and exporting the simulation model directly into SimMechanics [18]–[20].

Several papers presented building their own inverted pendulum systems (both rotary and pendulum on a cart) [19], [21], [22], concentrating on the mechanical design, design of the simulation model and experiments. Papers use either a commercial rapid prototyping system, such as dSPACE DS1104, [19] or an inexpensive microcontroller, such as Arduino, to implement the controller [22].

In this paper an experimental model of Furuta pendulum built for experimental verification of various control algorithms taught in undergraduate, graduate and postgraduate control courses, such as linear state feedback, model predictive control or sliding-mode control is presented. The built experimental setup is equipped with a microcontroller with prepared code for communicating directly with MATLAB/Simulink using Simulink Desktop Real Time. The microcontroller controls two H-bridges with current sense for driving up to two DC motors and it is able to provide either motor voltage or control the motor current (torque) directly. In the first phase, students use the designed interface to run the control algorithms directly from Simulink using Simulink Desktop Real Time, while in a later phase they can exploit added JTAG connector on the PCB to implement the control algorithm directly on the microcontroller. Sliding mode control is selected to show the possibility for testing a high bandwidth nonlinear control algorithm directly from MATLAB/Simulink.

The rest of the paper is organized as follows. In Section II the mathematical model of the pendulum is presented, Section III presents the experimental setup, Section IV presents the designed control algorithm and Section V presents simulation and experimental results. Section VI concludes the paper.

II. PENDULUM MODEL

The Furuta pendulum, shown in Fig 1, consists of an arm which is able to rotate in the horizontal plane and a pendulum suspended from the arm which is able to rotate in the vertical plane. The pendulum angle is denoted by α and arm angle is denoted by θ, respectively, while their meaning is shown in Fig 1. The arm is driven directly by an electric motor.

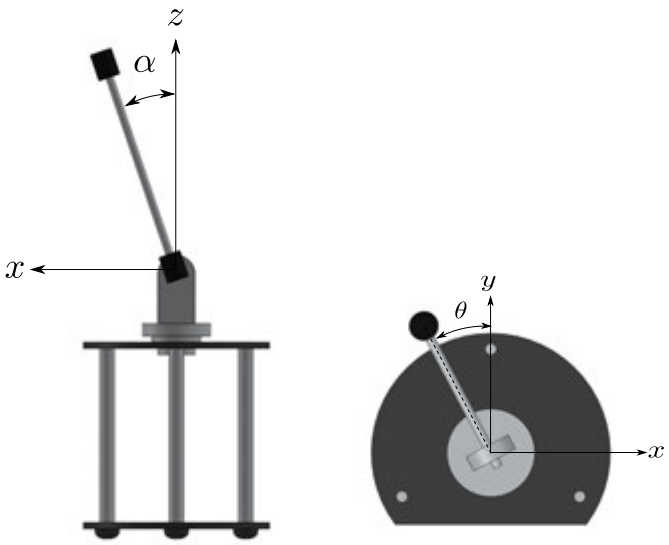


Fig. 1. Furuta pendulum model

Dynamic equations of the system can be obtained using Euler-Lagrange or Newton-Euler methods. The full nonlinear system dynamics for the Furuta pendulum with a complete inertia tensor is reported in [16]. This model can be simplified for simulation purposes by considering only the total moment of inertia of the pendulum arm and the pendulum about their pivot points: $\hat{J}_r = J_r + m_p l_r^2$ and $\hat{J}_p = J_p + 0.25 m_p l_p^2$, respectively. In those expressions, J_r represents the moment of inertia of the arm around z axis, J_p moment of inertia of the pendulum around its center of mass, m_p pednulum mass, l_r length of the arm and l_p length of the pendulum.

The equations are [16]:

$$\begin{aligned}\left(\hat{J}_r + \hat{J}_p \sin^2\alpha\right)\ddot{\theta} - \frac{1}{2} m_p l_p l_r \cos\alpha\ddot{\alpha} + \\ \hat{J}_p \sin(2\alpha)\dot{\theta}\dot{\alpha} + \frac{1}{2} m_p l_p l_r \sin\alpha\dot{\alpha}^2 = \tau - B_r\dot{\theta},\end{aligned} \tag{1}$$

$$\begin{aligned}-\frac{1}{2} m_p l_p l_r \cos\alpha\ddot{\theta} + \hat{J}_p\ddot{\alpha} - \\ \frac{1}{2}\hat{J}_p \sin(2\alpha)\dot{\theta}^2 - \frac{1}{2} m_p l_p g \sin\alpha = -B_p\dot{\alpha}.\end{aligned} \tag{2}$$

B_r is DC motor and B_p pendulum friction coefficient. The torque τ is generated by the DC motor and can be expressed as

$$\tau = \frac{n c_m (u - n c_e \dot{\theta})}{R_a}. \tag{3}$$

Parameter	Value	Unit
m_r	$5.3 \cdot 10^{-3}$	kg
l_r	0.09	m
J_r	$4.6 \cdot 10^{-3}$	$kg \cdot m^2$
m_p	0.107	kg
l_p	0.3395	m
J_p	$3.5 \cdot 10^{-6}$	$kg \cdot m^2$
B_r	0.056	$kg \cdot m^2/s$
B_p	$2.5 \cdot 10^{-4}$	$kg \cdot m^2$
g	9.81	m/s^2
c_m	0.0134	Nm/A
c_e	0.0134	Vs/rad
R_a	3.33	Ω
n	54.7945	

TABLE I
PARAMETERS OF THE ROTARY PENDULUM SIMULATION MODEL

In the upper equation, n is the gearbox ratio, R_a armature resistance, c_e and c_m are back-EMF and torque motor constants.

III. EXPERIMENTAL SETUP

A custom experimental setup shown in Fig 2 has been designed. It consist of a DC motor with gearbox equipped with a magnetic encoder for driving the arm and measuring its position and 1024 ppr incremental encoder for measuring the position of the pendulum.

A custom made PCB is designed, which consist of a microcontroller, dual H-bridge with current sense, USB port, connectors for connecting 3 incremental encoders, and 2 motors. A custom code is written to communicate with MATLAB S-funcition and a Simulink block diagram has been made. However, the PCB is equipped with a JTAG connector to enable students downloading their own custom code. Furthermore, H-bridge with current sense enables for two different modes of operation. The control input can be either the DC motor voltage or armature current (motor torque). The latter, enables students to use equations of motion directly obtained from Euler-Lagrange equations.

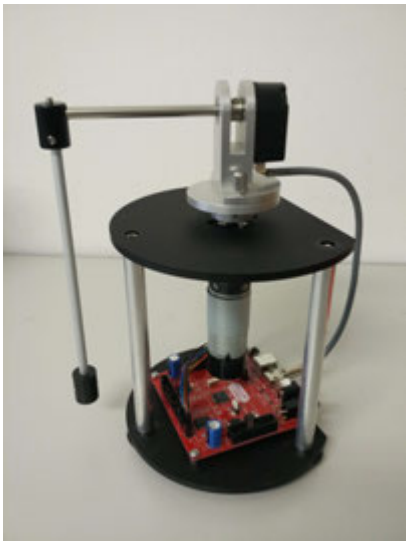


Fig. 2. The experimental setup

IV. Controller Design

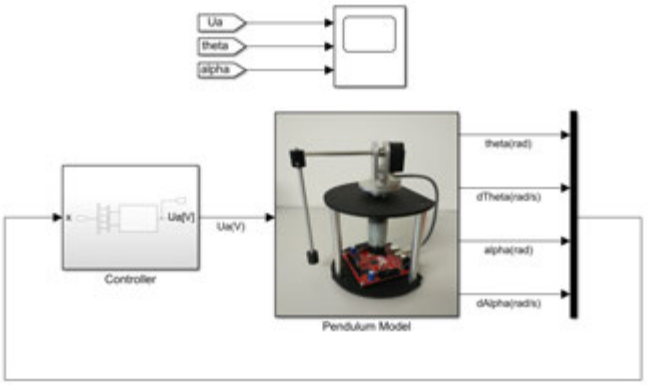


Fig. 3. Simulink model

For controller synthesis, the following simplified model is used:

$$\begin{aligned} \ddot{\theta} &= -a_m\dot{\theta} + c_e u, \\ \ddot{\alpha} &= -\frac{C}{\hat{J}_p}\dot{\alpha} + \frac{m_p g l_p}{2\hat{J}_p}\sin\alpha + \frac{K}{\hat{J}_p}\ddot{\theta}, \end{aligned} \tag{4}$$

where $\hat{J}_p = J_p + 0.25 m_p l_p^2$. The upper equation represents simplified model of the permanent magnet DC motor which is driving the rotating base, where the constant a_m includes back-EMF and friction effects and c_e effect of the input voltage on the angular acceleration of the base. The bottom equation describes the dynamics of the pendulum. C is the pendulum friction, m_p and J_p pendulum mass and moment of inertia and l_p pendulum length. K is a proportionality constant which is $K > 0$ for inverted and $K < 0$ for non-inverted position. The only control input is the DC motor voltage u [23].

The system (4) can be rewritten as

$$\begin{aligned} \ddot{\theta} &= -a_m\dot{\theta} + c_e u, \\ \ddot{\alpha} &= -\frac{C}{\hat{J}_p}\dot{\alpha} + \frac{m_p g l_p}{2\hat{J}_p}\sin\alpha + \frac{K}{\hat{J}_p}a_m\dot{\theta} + \frac{K}{\hat{J}_p}c_e u. \end{aligned} \tag{5}$$

The next step is to transform the system (5) into the new form using [23]

$$\begin{aligned} y &= \theta - \frac{\hat{J}_p}{K}\alpha, \\ x &= \dot{y} - \frac{C}{K}\alpha, \end{aligned} \tag{6}$$

which results in the following system formulation:

$$\begin{aligned} \dot{x} &= -\frac{m_p g l_p}{2K}\sin\alpha, \\ \dot{y} &= x + \frac{C}{K}\alpha. \end{aligned} \tag{7}$$

Parameter	Value	Unit
a_m	20	$1/s$
c_e	20.142	$rad/s^2/V$
m_p	0.107	kg
l_p	0.3395	m
J_p	$3.5 \cdot 10^{-6}$	$kg \cdot m^2$
C	$2 \cdot 10^{-3}$	$kg \cdot m^2/s$
K	$1.9 \cdot 10^{-3}\,\mathrm{sign}(\cos\alpha)$	$kg \cdot m^2$
g	9.81	m/s^2

TABLE II
Parameters of the rotary pendulum control model

A. Sliding Mode Controller

Lets take the state component θ_1 in the upper equation of the system (7) as control input. If the last term of this equation satisfies

$$\sin\alpha = -\lambda_1(x + y) \tag{8}$$

and

$$\lambda_1 = -\frac{2K}{m_p g l_p}, \quad \frac{m_p g l_p}{2C} > \frac{\pi}{2}, \quad K > 0, \tag{9}$$

then the system (7) is asymptotically stable. Consequently, the system (5) is asymptotically stable. For more information regarding system stability, refer to [23].

Equation (8) holds if the function

$$s_1 = \sin\alpha + \lambda_1(x + y) = 0. \tag{10}$$

The condition $s_1 = 0$ will be satisfied, if the function s_1 satisfies the linear first order differential equation

$$\dot{s}_1 = -\lambda s_1, \quad \lambda > 0. \tag{11}$$

The expression (11) will be satisfied if the sliding mode is enforced in the switching surface

$$s = \dot{s}_1 + \lambda s_1 = \cos\alpha\dot{\alpha} + \lambda_1(\dot{x} + \dot{y}) + \lambda s_1 = 0. \tag{12}$$

The time derivative of s equals

$$\dot{s} = \frac{K c_e}{\hat{J}_p}\cos\alpha u + \Psi(x, y, \alpha, \dot{\alpha}), \tag{13}$$

where Ψ is a nonlinear function of the system states. For the controller to achieve a stable behaviour, the functions s and $\dot{s}$ need to have opposite signs. This is achieved if

$$\dot{s} = -\eta\,\mathrm{sign}(s), \quad \eta > 0, \tag{14}$$

in which case the control input equals

$$u = \frac{-\hat{J}_p}{K c_e \cos\alpha}(\Psi + \eta\,\mathrm{sign}(s)). \tag{15}$$

However, such formulation can cause unwanted chattering. In order to reduce this effect, (14) was replaced by

$$\dot{s} = -\eta\tanh(s/2), \quad \eta > 0. \tag{16}$$

In the following chapters, control law realized with (14) will be referred to as "Controller 1" and the one realized with (16) as "Controller 2".

B. Swing-Up Controller

To bring the pendulum in the upright position, we use energy control strategy described in [24]. The energy of the pendulum is

$$E = \frac{1}{2}\hat{J}_p\dot{\alpha}^2 + \frac{1}{2}m_p g l_p(1+\cos\alpha), \tag{17}$$

while the reference energy equals maximum potential energy

$$E_0 = m_p g l_p. \tag{18}$$

There are various control laws which can be used to achieve reference energy (18). The one we use in this work keeps the amplitude of control signal constant, while changing only its sign, and is given by

$$u = -\mu\,\mathrm{sign}((E-E_0)\dot{\alpha}\cos\alpha). \tag{19}$$

Swing-up controller is active until the condition $|\alpha| < 30°$ is satisfied, when it switches to sliding mode control law (15).

V. Simulation and Experiment Results

Verification of described algorithms has been done using both simulation and experimental tests. Controller parameters are $\lambda = 10$, $\eta = 40$ and swing-up gain $\mu = 12$. At the beginning of the experiment, pendulum is in its stable stationary position with all the system states $[\theta\ \dot{\theta}\ \alpha\ \dot{\alpha}]^T = [0\ 0\ 0\ 0]^T$.

Fig. 4 shows simulation results for both controller versions. In the upper graph, it can be seen that Controller 1 yields faster θ angle dynamics, but is unable to hold the base of the pendulum in the zero position, whereas Controller 2 causes oscilations and results in zero steady state tracking error. Pendulum angle α dynamics is shown in the lower graph and no difference in dynamics between the two controllers can be noticed. Both stabilize pendulum very fast and with no steady state error.

Input signals from the simulation can be seen in the Fig. 5. Swing-up phase lasts for cca. 1 second, after which the sliding mode is turned on. There, one can notice already mentioned chattering effect, which causes input signal to have extremely fast switching behaviour and causes high jerk when applied to an electromechanical system. Controller 2 doesn't cause chattering, since around $s = 0$ it behaves similarly to a proportional controller and no switching is enforced.

Fig. 6 shows measured angles θ and α. It can be seen that, in reality, neither of controllers is able to hold zero steady state θ error. Controller 1 makes the system oscillate slightly around $\theta \approx 100°$, while Controller 2 results in a better behaviour and shifts those oscillations around $\theta = 0°$. This behaviour is consistent with observations made based on simulation, but with higher steady state error because of modelling errors and mechanical disturbances (e.g. friction, gearbox backlash etc.). On the other hand, both controllers show almost identical

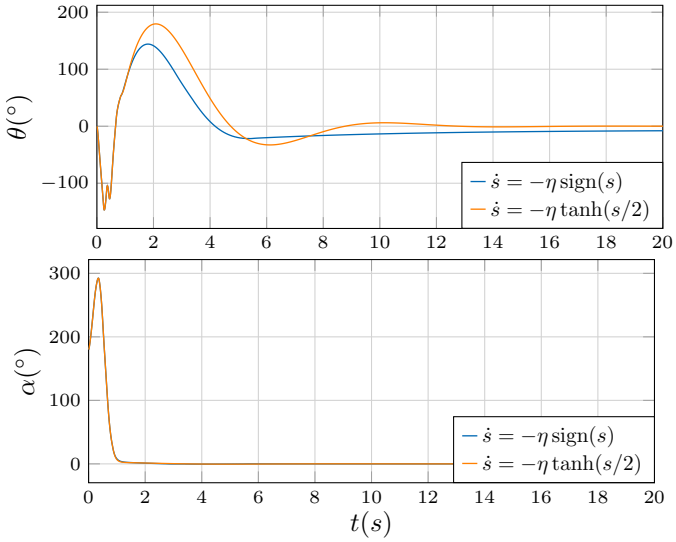


Fig. 4. Simulation results for both controller versions (angles).

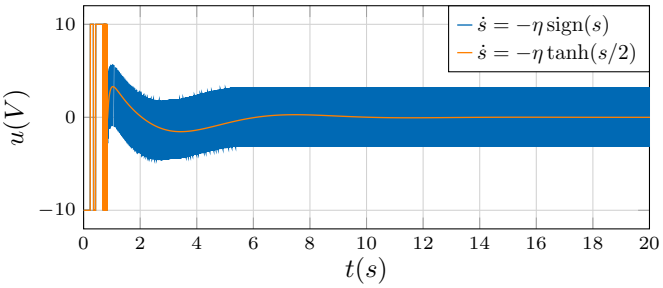


Fig. 5. Simulation results for both controller versions (input).

behaviour when it comes to α angle control. Behaviour is very similar to the one obtained from the simulation.

Input signals in Fig. 7 show that swing-up time in experiment lasted similarly long as in the simulation. After the sliding mode control has been activated, Controller 1 behaves very similar to the simulated one, causing a lot of chattering. However, in reality Controller 2 also causes chattering, although smaller than Controller 1.

VI. Conclusion

In this paper a new rotary inverted pendulum experimental setup has been proposed. The proposed experimental setup has a custom built PCB board equipped with a microcontroller with a code for communicating directly with MATLAB/Simulink using Simulink Desktop Real Time. In this configuration students have opportunity to implement controller in MATLAB/Simulink or implement the controller directly on the microcontroller. To show the capability of implementing high bandwith control algorithms, a sliding mode controller has been synthesised and implemented in MATLAB/Simulink. The controller has been tested in simulations and experimentally for two different versions of the controller. The first version uses sign function in the computation of the control signal while the second version uses tanh for reducing the chattering

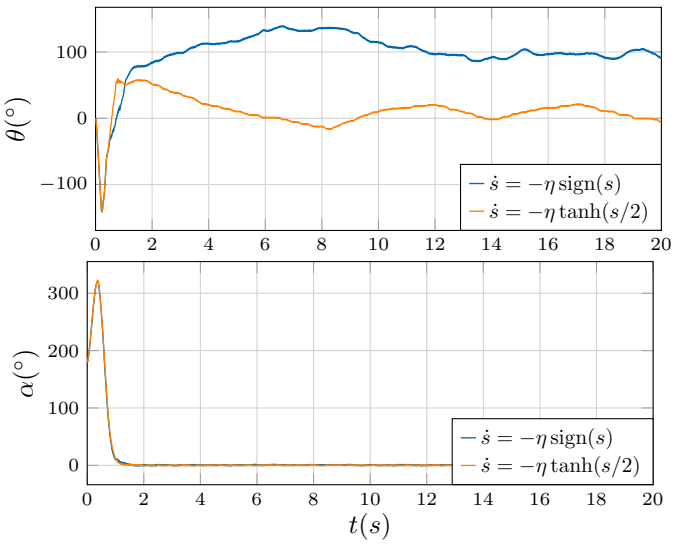


Fig. 6. Experimental results for both controller versions (angles).

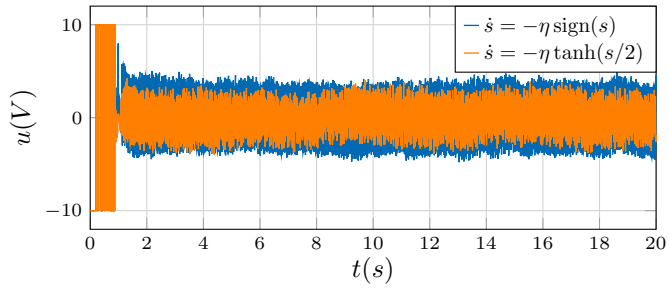


Fig. 7. Experimental results for both controller versions (input).

problem. In both simulation and experimental results the chattering has been reduced by using tanh function. The results on the proposed experimental setup closely resemble simulation results, which shows that the experimental setup can be used for testing various advanced nonlinear control algorithms.

ACKNOWLEDGMENT

This research has been supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS).

REFERENCES

[1] K. Furuta, M. Yamakita, S. Kobayashi, and M. Nishimura, "A new inverted pendulum apparatus for education," in *Advances in Control Education 1991*. Elsevier, 1992, pp. 133–138.

[2] A. T. Azar and F. E. Serrano, "Adaptive sliding mode control of the furuta pendulum," in *Advances and applications in sliding mode control systems*. Springer, 2015, pp. 1–42.

[3] M. Park, Y.-J. Kim, and J.-J. Lee, "Swing-up and lqr stabilization of a rotary inverted pendulum," *Artificial Life and Robotics*, vol. 16, no. 1, pp. 94–97, 2011.

[4] J. Acosta, J. Aracil, and F. Gordillo, "Comparative study of nonlinear control strategies for the furuta pendulum," in *Proceedings of the International Conference Measurement and Control*, 2001, pp. 18–23.

[5] M. Antonio-Cruz, V. M. Hernández-Guzmán, and R. Silva-Ortigoza, "Limit cycle elimination in inverted pendulums: Furuta pendulum and pendubot," *IEEE Access*, vol. 6, pp. 30 317–30 332, 2018.

[6] V. M. Hernández-Guzmán, M. Antonio-Cruz, and R. Silva-Ortigoza, "Linear state feedback regulation of a furuta pendulum: design based on differential flatness and root locus," *IEEE Access*, vol. 4, pp. 8721–8736, 2016.

[7] J. Aracil, J. Acosta, and F. Gordillo, "A nonlinear hybrid controller for swinging-up and stabilizing the furuta pendulum," *Control Engineering Practice*, vol. 21, no. 8, pp. 989–993, 2013.

[8] L. Magni, R. Scattolini, and K. J. Åström, "Global stabilization of the inverted pendulum using model predictive control," *IFAC Proceedings Volumes*, vol. 35, no. 1, pp. 141–146, 2002.

[9] P. Seman, M. Juh, M. Salaj *et al.*, "Swinging up the furuta pendulum and its stabilization via model predictive control," *Journal of Electrical Engineering*, vol. 64, no. 3, pp. 152–158, 2013.

[10] M. Furka, M. Klaučo, and M. Kvasnica, "Stabilization of furuta pendulum using nonlinear mpc," *Research Papers Faculty of Materials Science and Technology Slovak University of Technology*, vol. 27, no. 45, pp. 42–48, 2019.

[11] T. Ortega-Montiel, R. Villafuerte-Segura, C. Vázquez-Aguilera, and L. Freidovich, "Proportional retarded controller to stabilize underactuated systems with measurement delays: Furuta pendulum case study," *Mathematical Problems in Engineering*, vol. 2017, 2017.

[12] M.-S. Park and D. Chwa, "Swing-up and stabilization control of inverted-pendulum systems via coupled sliding-mode control method," *IEEE transactions on industrial electronics*, vol. 56, no. 9, pp. 3541–3555, 2009.

[13] M. Izutsu, Y. Pan, and K. Furuta, "Swing-up of furuta pendulum by nonlinear sliding mode control," *SICE Journal of Control, Measurement, and System Integration*, vol. 1, no. 1, pp. 12–17, 2008.

[14] D. R. Porto, "Angular position control of furuta pendulum with an intelligent sliding modes approach," B.S. thesis, Universidade Federal do Rio Grande do Norte, 2019.

[15] M. Ramírez-Neria, H. Sira-Ramírez, R. Garrido-Moctezuma, and A. Luviano-Juarez, "Linear active disturbance rejection control of underactuated systems: The case of the furuta pendulum," *ISA transactions*, vol. 53, no. 4, pp. 920–928, 2014.

[16] B. S. Cazzolato and Z. Prime, "On the dynamics of the furuta pendulum," *Journal of Control Science and Engineering*, vol. 2011, 2011.

[17] J. Acosta, "Furuta's pendulum: A conservative nonlinear model for theory and practise," *Mathematical Problems in Engineering*, vol. 2010, 2010.

[18] J. Duart, B. Montero, P. Ospina, and E. González, "Dynamic modeling and simulation of a rotational inverted pendulum," in *Journal of Physics: Conference Series*, vol. 792, no. 1. IOP Publishing, 2017, p. 012081.

[19] M. Antonio-Cruz, R. Silva-Ortigoza, C. Merlo-Zapata, M. G. Villarreal-Cervantes, D. Muñoz-Carrillo, and V. Hernández-Guzmán, "Modeling and construction of a furuta pendulum prototype," in *2014 International Conference on Mechatronics, Electronics and Automotive Engineering*. IEEE, 2014, pp. 98–103.

[20] A. Kathpal and A. Singla, "Simmechanics™ based modeling, simulation and real-time control of rotary inverted pendulum," in *2017 11th International Conference on Intelligent Systems and Control (ISCO)*. IEEE, 2017, pp. 166–172.

[21] B. Allotta, L. Pugi, and F. Bartolini, "Reinforcement neural network for the stabilization of a furuta pendulum," in *Proceedings of EUCOMES 08*. Springer, 2009, pp. 287–294.

[22] T. R. Cooley, "A low cost and flexible open source inverted pendulum for feedback control laboratory courses," *age*, vol. 26, p. 1.

[23] V. Utkin, J. Guldner, and J. Shi, *Sliding mode control in electro-mechanical systems*. CRC press, 2009.

[24] K. J. Åström and K. Furuta, "Swinging up a pendulum by energy control," *Automatica*, vol. 36, no. 2, pp. 287–295, 2000.

A Simulator for Training Human Operators of a Remote Controlled Anti-Terrorism Ground Vehicle

Juraj Fulir, Željka Mihajlović, Marija Seder
University of Zagreb Faculty of Electrical Engineering and Computing
Zagreb, Republika Hrvatska
{juraj.fulir, zeljka.mihajlovic, marija.seder}@fer.hr

***Abstract*—Remote controlled robotic vehicles are expensive and require high maintenance. In an effort to reduce costs of training human operators for an anti-terrorism ground vehicle we introduce a simulator implemented in Unreal Engine 4 as an affordable solution. We describe several features that need to be considered when simulating a remotely controlled vehicle guided solely by its integrated cameras. Additional steps are shown for achieving realism and user immersion while maintaining an interactive rendering speed. The cost of rendering multiple cameras is analysed and steps for remedying it are shown.**

***Index Terms*—simulator training; remote controlled ground vehicle; anti-terrorism**

I. Introduction

Remote controlled vehicles are used in dangerous situations and often assist humans through close interaction. This requires fast and precise actions from a safe distance by a well trained operator using a very limited field of view. To acquire such skills, a candidate must practice controlling the vehicle in safe environments. This is costly since the candidate must be brought to the training course, additional resources must be spent to run the vehicle and there is a risk of damaging the vehicle or endangering the training environment (pedestrians, animals, training course objects, etc.). Training in a simulator costs virtually nothing and can be especially useful in the early stages of training, offering an approximate experience of operating the vehicle. Additionally, simulators can offer approximate experience of situations that would otherwise be too dangerous or costly to perform in real-life training, e.g. around explosion and radiation hazards [1]–[3]. Experiences gained in a highly realistic simulator could be transferred to real-life skills adding great value its usage.

Simulators have several components that make them realistic in different aspects. The physical simulation of vehicle's movement and its interaction with the world is a crucial component. It makes the experiences gained in the simulator transferable to real-life skills. A replica of the user interface and the ability to connect to an official controller allow the candidate to adjust to the controls and learn efficient usage of limited information available. Scenarios for gaining targeted experiences are deigned by experts and prepare the candidate for real life situations. Finally, increasing the rendering quality helps immersing the candidate into the virtual environment and treating the scenarios presented as real. To maintain this illusion the rendering speed must be high at all times which poses constraints on the amount of detail and dynamics of all of the above mentioned components [4], [5]. Realistic simulators are additionally useful for training machine learning models used in computer vision and robotics [6]–[8].

Game engines are generally designed to offer various libraries, tools and optimised features. These allow fast development and adjustments for leveraging between visual quality and performance according to need [9]. Generally, they support multiple standards for assets and offer their own which enables fast prototyping with high quality results, making them very handy for building simulators. In this work we describe features, encountered problems and proposed solutions for a simulator built using Unreal Engine 4 (UE4) [10] for operating a remote controlled anti-terrorism ground vehicle. The simulated vehicle is a prototype based on the DOK-ING MV-3 model [11].

In section II we review several popular simulators similar to ours built as standalone tools or projects within game engines with usages in training human or AI operators. In section III we discuss implementation details of various physical properties of the simulator and in section IV how the vehicle interacts with its environment. In section V we discuss performance issues that arise from simultaneous rendering of multiple camera views and in section VI methods for achieving a realistic and immersive environment at small rendering cost. In section VII we review results of proposed solutions and finally propose several directions for further research.

II. Related work

As remarked in [9], game engines contain many readily available elements necessary for building a simulator. While some simulators are built as a standalone tool, most modern simulators just offer a layer of abstraction over the physics engine. Usually they offer a set of implemented components [12] or use specialised physics engines with an API as a plugin for game engines [13], [14]. This way the development of the graphical and physical end of the simulator is separated allowing development teams to specialise in their respective areas. With the ever growing interest in game development,

the game engines continually increase both the usability and performance allowing the development of realistic simulators.

Gazebo [12] is a standalone simulator popular in robotics. It offers various implementations of joints, actuators and sensors by fostering the power of an external physics engine. Programs developed for a simulated model are transferable to the real model. The simulator however offers modest set of tools for realistic graphics which limits its usage for training human operators. Several solutions were developed for using a game engine to improve visual quality while supporting the development of transferable programs [6], [15].

Simulators can implement various levels of physical complexity. Some were developed to achieve high precision simulation of vehicles and their surroundings. These include modelling the deformation of ground under construction vehicles [16] and precise mechanical response for various vehicle performance assessments [13], [14]. These simulators require high-end or even specialised hardware for the complex physical simulation which raises the cost of usage. Other simulators focus on the concept of serious gaming, focusing on immersion and rapid decision making while achieving a decent amount physical reality. These simulators are often used by the military for strategy training of teams. Due to their lower simulation costs, these simulators often offer state of the art real-time graphics quality and can be run on consumer class hardware. Simulators that require both high fidelity graphics and precise physics are often used for training and validating machine learning algorithms [8], however they require specialised hardware clusters. Lower complexity solutions also exist for the domain of machine learning [6], [7].

Multi-view rendering is common in multiplayer games where multiple users remotely interact with the same virtual scene. These are usually centralised systems with a server entity which processes some or all of the physics or rendering related tasks of multiple clients. The server can be one or more specialized machines with various connection strategies. [17] These approaches differ from ours as we focus on single user interaction and require the entire simulator to run on a single consumer class machine. This gives us the possibility of connecting an external remote controller and controlling the simulated vehicle for training purposes. In virtual reality applications two cameras close together are used to create a stereo image of the scene which is a special case of multi-view rendering. Due to proximity of cameras scene elements visible from one camera are similarly visible from the other, so some resources or even pixels can be reused when rendering the other camera. In our case vehicle mounted cameras are distant from each other and their fields of view are mostly exclusive which reduces the effectiveness of reusing information.

III. Physical fidelity and movement

The PhysX Vehicles plugin for UE4 comes packaged with the engine and is a highly configurable plugin, containing implementations of several powertrain models for cars. It uses the PhysX vehicle template for simulating vehicle movement. However, it currently supports up to 4 wheels and doesn't support the application of torque to the left and right wheels independently, referred to as differential steering. This makes the plugin unusable for simulating tracked vehicle movement. The PhysX Vehicle SDK also contains an implementation of tank movement which supports differential steering, however we find it unusable for our application. The wheel collision with the ground is approximated using a ray-cast in the $-\vec{z}$ direction of local space. Based on the length of the ray a suspension spring is simulated and used to suspend the vehicle body. This approximation works well on relatively flat terrain and when vehicles aren't expected to climb on steep surfaces. However, we expect the vehicle will have no suspension and at low speeds on rough terrain ray-casts without suspension perform unnatural hard oscillations. Since we expect the intended environments will contain rubble and sharp fragments we replace the single ray-cast collision with full body collision.

The tracks are approximated with a finite number of wheels and a cuboid shape to fill the gaps between wheels, as shown in Fig. 1. The wheels are represented by spherical collision bodies and are positioned at locations of the wheels supporting the track in the real vehicle, shifted lower so the track's outer perimeter touches the perimeters of wheels. This way the weight distribution necessary for drive force calculation is well approximated. Wheels are connected to the body using a movement restricted bone, allowed only to freely rotate in the direction of movement, around the vehicle Y axis, where the X and Z axes are respectively the forward and up direction. Between the wheels there are gaps present, which can cause the vehicle to get stuck in sufficiently thin and tall obstacles. The real vehicle has hard track guides which prevent this effect. To approximate this we use an additional collision body in the shape of a cuboid. The cuboid is placed slightly above the track's outer perimeter to allow the wheels to contact the ground, but block sharp obstacles from getting between the

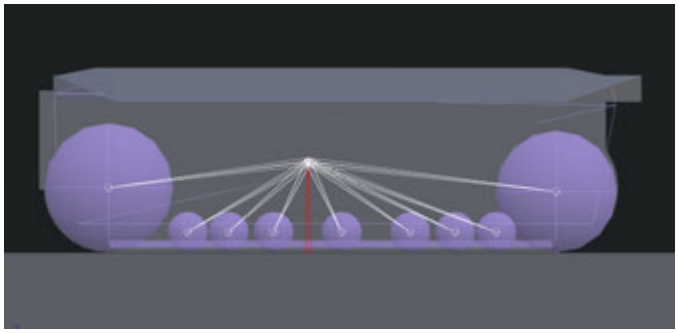


Fig. 1. A view of the vehicle with highlighted collision bodies and bones. The purple objects represent collision bodies, connected to the red root bone by individual white bones. The cuboid fills the space between the wheels and prevents obstacles from jamming them, as expected in the real vehicle. Note that the collision between the overlapping bodies is turned off.

wheels and jamming them. To reduce the energy loss resulting from using only the wheels as force applicators, we remove all friction of the cuboid.

The vehicle can be controlled via discrete and continuous input signals. Discrete signals are used for discrete actions such as changing a gear or setting the handbrake. Continuous signals are used for throttle and steering. However, in absence of an analog input device, discrete input devices such as a keyboard can be used by gradually incrementing the input value through time to simulate analog inputs. This allows us to use both input types in case the controller input specification changes. To simulate throttle we set the constant to positive if throttle is non-zero and to negative when throttle is near zero. This produces a linear change which can be easily adjusted to vehicle specification. The vehicle engine supports two gears: one for normal movement and one used for gaining momentum when punching through objects. The sound of the engine is used to give additional information about the engine to the user so we modulate it based on the throttle value, which corresponds to the change in engine RPM.

IV. Interaction with the world

The terrorists can hide behind a solid object or in a building to achieve a strategic advantage. To overcome this the vehicle has a front mountable spike designed to penetrate through certain vertical barriers, such as glass or thin wooden walls. Destructible objects can be simulated and created almost effortlessly using the PhysX APEX plugin which converts solid meshes into broken meshes when hit with sufficient force. Tuning various parameters allows emulating resistance of different materials. Adding sound cues and smoke particle effects to the broken items gives a more believable look to the action of punching through a wall. To keep the rendering speed in a scene overflowed with fragments, each fragment can be given a random lifetime after which it is removed from the scene. The vehicle can be mounted with a shield used to protect the people and items it carries from enemy fire. Over some terrain types dust can be generated. Dust particles are generated randomly with probability proportional to the vehicle velocity. The landscape information is obtained by shooting a length limited ray-cast in the $-\vec{z}$ direction of the local space.

Due to its high power the vehicle can be used for towing objects in the scene with a chain or a rope. To tow an object the vehicle must first get close enough to the object. To test whether the object is in range for towing an invisible collision volume is placed around the vehicle. The collision volume is set only to query the collision with the world and can be set to filter unwanted collision. If the object is in range and is tagged using a specified string tag, it can be connected by pressing an action button. The object gets connected to the vehicle via a physical constraint in a form of maximum distance of its anchors. To cue the user that a connection was established we add a simulated rope between the two anchors, defined on both objects as sockets.

On the top of the vehicle various items can be carried. The main expected usage is driving wounded people away from the danger zone and bringing heavy equipment where it is needed. To carry an object we define a virtual socket on the vehicle. We reuse the same collision body used for towing to check if an object is in range for carrying. On the press of an action button the object gets attached to the specified socket and can be carried. Prior to attachment we must disable collision of the body and the simulation of physics, otherwise the attachment constraint becomes unstable and breaks. When detaching the item, the collision and physics properties are restored and the object is placed to the right of the vehicle.

V. User interface and display

The vehicle will have multiple cameras mounted on various locations to allow the user a view of the vehicles' surroundings. The video feed must be sent to a distant operator via a low bandwidth communications channel which restricts the view to a single camera at a time. Future work includes optimisations that will enable simultaneous feed from multiple cameras so we focus on that scenario. Additionally, we render an operator's view of the scene which simulates the expected usage. Fig. 2 shows the user interface and the humanoid used for the operator's view. The third-person view towards the vehicle is in the bottom left corner.

To display images from multiple cameras first we render them to textures and then draw them on the head-up display. Each rendered camera requires a separate render query which increases the rendering cost. Since camera views are will be small in the user interface, we reduce the resolution of target textures. The image quality reduction isn't noticeable, but the rendering cost is decreased.

To allow the user to focus on a single camera with more detail, action keys are added to cycle through different



Fig. 2. The user interface displays all camera sources in reduced resolution to preserve interactive rendering speeds. The active camera is indicated by a transparent rectangle, in this case the third person view. The vehicle cameras shown on top are in order: rear left, front left, front, front right and rear right.

cameras. When a camera is selected it is displayed across the screen in its full resolution and its window is marked with a semi-transparent window. The small window of the current camera is not shown because it requires an additional render query which unnecessarily degrades performance, as discussed in section VII-B.

VI. Visual fidelity and performance

Higher visual fidelity allows the user immerse into the environment and approach the simulator training more seriously. High quality models, realistic textures and adding as many varieties as possible are essential parts. Colour varieties provide good variation and can be simulated efficiently using Perlin noise mapping based on the world xy coordinates. Geometry can be shared between instances to save memory at the expense of similarity between them. A different approach would be to build trees procedurally with unique and randomised shapes using algorithms described in [18]. However, the cost of building and rendering lots of trees in this way would be overwhelming considering the number of trees and the amount of work needed for other tasks such as physics simulation and world interaction. On the other hand, simulating a natural environment requires more than just realistic visuals. If a natural scene is static it looks very artificial. Adding just a small amount of branch movement simulates wind and increases fidelity of the natural environment. The wind can be simulated within UE4 using the 'SpeedTree' plugin and compatible tree models.

It makes sense to put certain vegetation on landscape types that in nature support it. This is achieved by limiting on which landscape layer foliage can be put. Instead of putting instances one by one, they can be massively placed using brushes or generated procedurally. Procedurally generated instances can be placed using procedural volumes or using the landscape material. Procedural volumes can be placed in the scene and define the bounds within which vegetation can be spawned. Landscape material can be used to procedurally spawn multiple different instances on specific landscape layers. This is a very simple method that is very configurable and yields great results. However, it still doesn't support collision on spawned meshes which makes it useful only for certain types of vegetation. Brushes are used to paint on the terrain texture where instances can be spawned and the instances are statically placed. This method allows for a finer control of instance distribution and supports meshes with collision, which makes it useful for placing trees, rocks and other objects.

In practice, cameras of remote controlled vehicles can be occluded by clouds of dust from vehicle movement. To simulate dust generation we cast a ray from each track. Based on the type of hit material, we generate an adequately tinted dust sprite depending on probability proportional to track speed. Dust generation should not be too frequent to avoid generating non-transparent clouds and randomisation gives a more natural look to the simulated effect.

VII. Results

All experiments were executed on a single Lenovo Legion Y520 laptop with an Intel i7-7700HQ chip, 8 GB of RAM and a NVIDIA GTX 1050 graphics card with 4 GB of VRAM. These specifications are representative of the expected systems to run the simulator in practice. The Unreal Engine version is 4.22.1 with the OpenGL version 4.6.0 backend and PhysX 3.4 as the physics engine.

A. Track collision using the wheel collision body

Fig. 3 depicts traces of the two approaches in implementing track collision without suspension. The ray cast collision shows a sharp spike when it encounters the edge of an obstacle causing the entire vehicle to jump unnaturally. Using the full collision body for wheel collision the trace is much smoother and provides a more natural transition. In high speed vehicles with suspension the effect of sharp collision is softened using the simulated suspension springs and the jumping effect is unnoticeable. As we expect this tracked vehicle to operate at low velocities without suspension, the full wheel collision is a better approach to implementation.

Fig. 3 also points out a problem with the approach of approximating a track using multiple wheels. In the upper image obstacle clearly penetrates the track and occasionally causes the vehicle to get caught in it. The naive implementation using only wheel colliders suffers from the same problem. The addition of a cuboid to the track collision keeps the vehicle suspended over the obstacles and greatly reduces

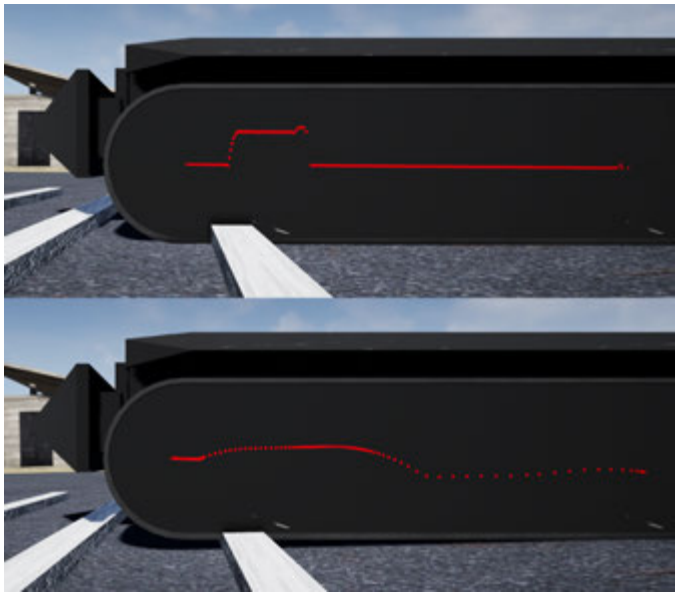


Fig. 3. The trace represents the movement of a front wheel through multiple consecutive frames. The top image shows a sharp wheel jump when the ray-cast collision detects the difference in surface elevation. The bottom image shows a smoother trace using the full collision body of the wheel. The traces in the figure are only indicative and do not match the exact horizontal coordinate of the obstacle due to perspective projection of the camera.

the chance of jamming the wheels. The vehicle still changes elevation as it traverses over individual obstacles due to the height difference between wheels and the added cuboid, but in practice the effect is barely noticeable. When traversing a regular pattern of obstacles, the vehicle still might get stuck if none of the wheels have contact with the ground. However, this scenario is very improbable due to multitude of conditions needed for it to happen that are not common in the expected environments. These include: pattern of obstacles must match the pattern of passive space between the wheels, vehicle must be perpendicular to the obstacle and vehicle must be stationary due to no friction constraint on the cuboid.

B. *Measuring the cost of multiple cameras*

The user interface requires simultaneous display of the feed from multiple vehicle mounted cameras. Rendering multiple cameras causes additional queries for scene rendering which adds a significant penalty to the total time cost per frame. The cost of displaying a high resolution interface is proportional to the expected cost of streaming a video from a remote vehicle through limited bandwidth wireless communication. We test the effect of interface resolution, number of simulated cameras and the resolution of camera views on the drop of the overall frame rate.

We use two resolutions expected to be found in the device for remote control: 1920×1080 and 1280×720. For each device resolution we test several scenarios, depending on the number of cameras: a single camera (front), 3 cameras (front), 5 cameras (3 front and 2 back) and the same 5 cameras with an additional third-person view. Vehicle mounted cameras were rotated from the front facing one by $\pm 45°$ for front ones and $\pm 135°$ for rear ones. The front facing camera was always used as the active camera displayed in the background. For each of these setups we test several resolutions of camera views. Since our interface divides the horizontal screen space into 5 parts, we ensure both tests include the camera resolution that is exactly one fifth of the screen resolution, which will maintain the original pixel density. The camera view resolutions sorted by the number of pixels are: 128×72, 256×144, 384×216, 512×288. Larger resolutions were not tested as they show no visible improvements in image quality in direct display and are not to be expected in practice. For each experiment we record 1000 frames after a first 5 second delay. The delay allows the resources to load and the frame rate to stabilise for an accurate reading. The physics engine in conjunction with the game engine produces stochastic vehicle trajectory for the same starting conditions. To reduce the possibility of unfair estimation while maintaining tractability, each of the above experiments was repeated 5 times. Finally, the temporal cost of each frame over all repetitions were averaged to produce the final results displayed in Fig. 4.

The vehicle was placed into a relatively open part of the map and forced to drive forward with occasional turning right according to a simple delay based flip-flop automata. The turning was stronger then usual to allow for the vehicle to turn quickly enough so that all cameras catch a part of the dust particles before they disappear. The dust particles cause significant drops in frame rate due to being semi-transparent and numerous. By allowing all cameras to capture it, we guarantee that some setups aren't falsely better. Dust particles are important for our testing since we expect dense generation during the training procedures to increase candidate robustness. The surroundings include large and small objects, both of static and dynamic nature. In the beginning of simulation three destructible objects were dropped from the sky and broken to increase the number of dynamic objects and scene complexity.

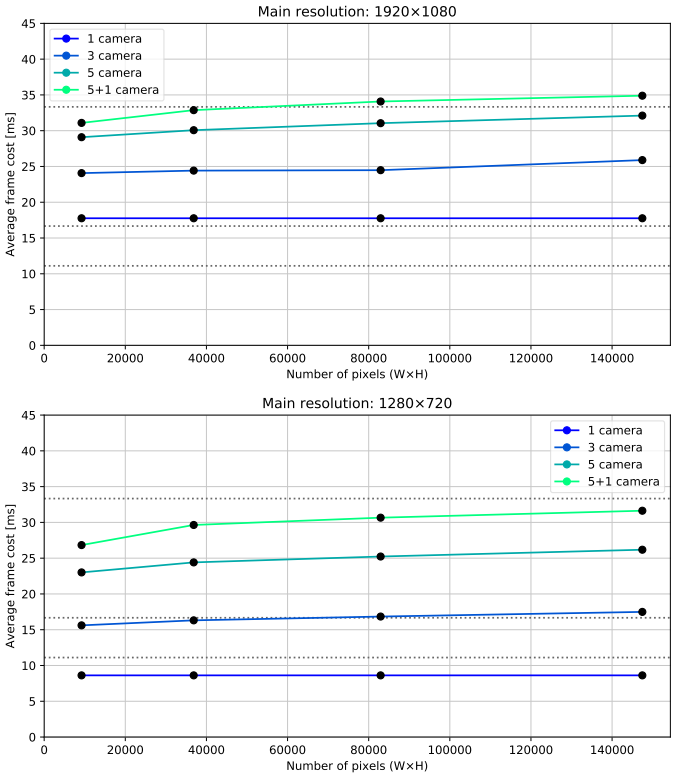


Fig. 4. Graphs represents the average time to render a frame over a number of pixels for each of the camera view resolutions. Each line represents a setup depending on the number of cameras. Each graph presents measurements at a given resolution of the main viewport. The dotted lines represent the rendering costs at 30, 60 and 90 FPS. The values in the X-axis represent the number of pixels in a given camera view resolution.

Fig. 4 shows results of the experiments. The larger resolution of the main viewport shows generally larger costs of rendering a frame. We can see that even the lightest configuration on high resolution viewport doesn't achieve the 60 FPS threshold. This suggests a possibility of using various methods of upsampling to utilise the speed of rendering a low resolution screen by sacrificing some visual details. By adding additional cameras to the user interface the costs rise significantly. Each camera approximately adds additional 2.5 ms, rapidly escaping the 60 FPS threshold for the lower resolution viewport. This indicates a large advantage of hiding as much of camera sources as possible in the user interface.

This is the reason we hide the current view as described in section V. Note that this makes the cost for the one camera configuration same for any camera view resolution since it is hidden from the user interface. The cost of adding additional cameras is more or less the same for all scenarios.

Increasing the camera resolution, within reasonable bounds, has little effect across different number of cameras compared to the cost of drawing the main viewport and requiring several rendering queries. Therefore it is not necessary to reduce the resolution of views at the cost of aliasing effects and detail loss in an already small image. In our case, fifth of the screen resolution maintains pixel density of the views and is enough to display the scenes. However, we observe slightly better image quality on a resolution larger then the optimal one.

VIII. Future work

The current implementation is sufficient to give initial experience in manoeuvring the vehicle over rough terrain and in tight places, using the limited information available through the vehicle mounted cameras. A designed user study of candidates would reveal the true effectiveness of the simulator in training. Due to time and space constraints we believe that the effectiveness study is outside of the intended scope of this paper and is left as future work.

Dynamic scenarios and specific goals should be specified to immerse the candidate into the simulated environment. Both need to be challenging but representative of real life situations the candidate might encounter. Future work might include scenarios designed in collaboration with experienced pilots to increase the value of experiences gained through the simulator.

When constructed, the original vehicle controller will communicate with the simulator. Although rendering cameras directly to the framebuffer could lower the render time, our goal is to use render targets which will later allow us to simulate the image streaming from a vehicle to the original controller. To compensate the cost of image transfer from the vehicle to the controller, smarter approaches to information presentation and transfer will be investigated.

IX. Conclusion

A simulator of a remote controlled anti-terrorism ground vehicle for usage in human operator training is presented. Several features required by the vehicle and their implementation are described. The implementation of tracks is compared to the existing solution for vehicles in UE4. The costs of displaying multiple cameras for remote control are compared and analysed to point out main problems. The largest cost comes from displaying multiple cameras in the user interface so their count should be specified wisely. Their resolution in the user interface has little cost when within reasonable bounds and can be safely increased to match the needs of the interface.

Acknowledgement

This work has been supported by the European Regional Development Fund under the grant KK.01.2.1.01.0138 - Development of a multi-functional anti-terrorism system (MAS). We gratefully acknowledge the NVIDIA Corporation for their donation of the Titan V GPU used in this research.

References

[1] M. Đakulović and I. Petrović, "Motion planning of mobile robots for humanitarian demining," in *10th International IARP Workshop HUDEM'2012*, 2012.

[2] A. Šelek, D. Jurić, A. Čirjak, F. Marić, M. Seder, I. Marković, and I. Petrović, "Control architecture of a remotely controlled vehicle in extreme CBRNE conditions," in *2019 International Conference on Electrical Drives Power Electronics (EDPE)*, pp. 273–278, Sep. 2019.

[3] J. Peterson, W. Li, B. Cesar-Tondreau, J. Bird, K. Kochersberger, W. Czaja, and M. McLean, "Experiments in unmanned aerial vehicle/unmanned ground vehicle radiation search," *Journal of Field Robotics*, vol. 36, no. 4, pp. 818–845, 2019.

[4] M. Meehan, S. Razzaque, M. C. Whitton, and F. P. Brooks, "Effect of latency on presence in stressful virtual environments," in *IEEE Virtual Reality, 2003. Proceedings.*, pp. 141–148, March 2003.

[5] S. P. Sudarsan, C. J. Cohen, L. Q. Du, P. N. Cobb, E. S. Yager, and C. J. Jacobus, "Influence of video characteristics of simulator images on remote driving performance," in *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, vol. 2, pp. 1062–1066 vol.2, Oct 1997.

[6] Microsoft, "AirSim." https://microsoft.github.io/AirSim. Accessed: 2020-02-06.

[7] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16, 2017.

[8] Nvidia, "Nvidia Drive." https://developer.nvidia.com/drive. Accessed: 2020-02-06.

[9] J. Craighead, R. Murphy, J. Burke, and B. Goldiez, "A survey of commercial open source unmanned vehicle simulators," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pp. 852–857, April 2007.

[10] Epic Games, "Unreal Engine." https://www.unrealengine.com. Accessed: 2019-12-18.

[11] "DOK-ING." https://www.dok-ing.hr. Accessed: 2020-02-06.

[12] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*, vol. 3, pp. 2149–2154 vol.3, Sep. 2004.

[13] CM Labs, "Vortex Studio." https://www.cm-labs.com/vortex-studio. Accessed: 2020-01-20.

[14] Mechanical Simulation, "CarSim." https://www.carsim.com. Accessed: 2020-01-20.

[15] S. Carpin, M. Lewis, J. Wang, S. Balakirsky, and C. Scrapper, "Usarsim: a robot simulator for research and education," in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pp. 1400–1405, April 2007.

[16] D. Holz, A. Azimi, and M. Teichmann, "Advances in physically-based modeling of deformable soil for real-time operator training simulators," in *2015 International Conference on Virtual Reality and Visualization (ICVRV)*, pp. 166–172, Oct 2015.

[17] W. Cai, R. Shea, C. Huang, K. Chen, J. Liu, V. C. M. Leung, and C. Hsu, "A survey on cloud gaming: Future of computer games," *IEEE Access*, vol. 4, pp. 7605–7620, 2016.

[18] H. Nuić and Ž. Mihajlović, "Algorithms for procedural generation and display of trees," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 230–235, May 2019.

INTERNET BASED SYSTEMS, APPLICATIONS AND TECHNOLOGIES

Tools for Analytics and Cognition Framework for a Car-Sharing Use Case

A. Karadimce, D. Capeska Bogatinoska, M. Sefidansoki, N. Paunkoska Dimoska and N. Marina
University of Information Science and Technology "St. Paul the Apostle"
Ohrid, R.N. Macedonia
aleksandar.karadimce@uist.edu.mk, dijana.c.bogatinoska@uist.edu.mk,
mirsat.sefidanoski@cse.uist.edu.mk, natasa.paunkoska@uist.edu.mk, Rector@uist.edu.mk

***Abstract*—The development of tools that can improve efficiency and inject intelligent insights into operational and mission-critical social media businesses through guided analytics is crucial for consumers, prosumers, and business markets. These tools will provide contextualised socially aware and spatial-temporal data aggregation, knowledge extraction, cognitive learning about users' behaviour, and risk quantification for the car-sharing use case. The proposed Tools for Analytics and Cognition (TAC) framework will provide a tool-set of guided analytics software for smart aggregation, cognition and interactive visualisation with a monitoring dashboard for the car-sharing use cases. The proposed TAC framework uses the dashboard to visually analyse the behaviour and engagement of the social media actors, diagnose performance risks and provide guided analytics to consumer prosumers and application providers to improve collaboration and revenues, using the established car-sharing qualitative mapping model. This framework has supplied a seamless coupling with distributed blockchain-based services for early alert, real-time tracking and updated data triggers for reach and engagement analysis of car-sharing events. Moreover, the TAC framework will allow car-sharing providers to analyse, control and track their investment to enhance monetary inclusion in the collaborative social media ecosystem.**

***Index Terms*—Guided Analytics, Data Aggregation, car-sharing, Augmented Cognitive, Microservices, Social Media**

I. Introduction

The ARTICONF 'smART socIal media eCOsystem in a blockchaiN Federated environment' [1] is an EU project that addresses the urgent industrial need for a novel set of trustworthy, resilient, and globally sustainable decentralised social media services. These popular services consider issues of trust, time-criticality and democratisation for a new generation of federated infrastructure. By, fulfilling the privacy, robustness, and autonomy related promises that proprietary social media platforms have failed to deliver so far. The social media platforms have the potential to shape and mobilise patterns of communication, practices of exchange and business creation and learning and knowledge acquisition. Typically, social media are centralised platforms with a single proprietary organisation controlling the network. This poses critical issues of trust and governance over created and propagated content. This is becoming very problematic when there occur data breaches that are happening regularly at the hands of centralised intermediaries.

The ARTICONF project is found by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825134

The ARTICONF project has been envisioned to overcome the common barriers for the existing social platforms, such as (i) non-transparent and centralised infrastructure control when defining the social media network (ii) inefficient methodologies for identification of malicious actors (iii) lack of a democratic paradigm in existing social media networks, (iv) an inefficient resource provisioning for dynamic and time-critical social media applications such as the one based on crowds and (v) lack of mechanisms to support social cohesion. The proposed ARTICONF platform will provide a universal backbone for an open and federated platform that accommodates trust, democratisation, and time-critical based QoS requirements with incentivization opportunities for a large and diverse user base, including consumers, prosumers and business markets. The main goal of this project is to introduce a new type of social media network with democratic decision making, dynamic infrastructure provisioning and addressing trust and privacy issues. The project aims to create a decentralized and federated social media ecosystem, supported by an underlying blockchain technology that will ensure portable intra-platform and cross-platform social media data, interpretable in a range of different contexts for preservation, analysis and visualization via four frameworks (i) Trust and Integration Controller (TIC), (ii) Co-located and Orchestrated Network Fabric (CONF), (iii) Semantic Model with self-adaptive and Autonomous Relevant Technology (SMART) and (iv) Tools for Analytics and Cognition (TAC).

The TAC is a crucial tool-set of a decentralized social media ecosystem which enhances business productivity by tracking updated data triggers from diverse social media events. The TAC is seamlessly coupled with SMART and provisions socially-contextual geospatial and temporal data aggregation microservices to gain intelligent insights and prediction. The SMART tool initiates the TAC configuration providing aggregation, monitoring, cognitive reasoning and learning modules that analyse the behaviour and engagement of the social media actors. The TAC tool interacts with the CONF component in continuity to avail intelligently provisioned tool based on social media application requirements, operational conditions at the infrastructure level, and time-critical event triggering. TAC

has a microservices-based architecture organised in modules displayed in Fig. 1.

The SMART framework initiates the TAC configuration providing aggregation, monitoring, cognitive reasoning and learning modules that analyse the behaviour and engagement of the car-sharing social media actors. The outcome is to diagnose performance risks, and provide guided analytics to consumer, prosumers, and application providers to improve collaboration and revenues for a car-sharing use case. The TAC framework will interact with the CONF component in continuity to avail intelligently provisioned frameworks based on social media car-sharing application requirements, operational conditions at the infrastructure level, and time-critical event triggering. The TAC is a fundamental part of the ARTICONF platform which is used to collect and aggregate data for the car-sharing pilot use case scenario, run data aggregation analysis based on the TIC, SMART and CONF frameworks (see Figure 1). The ARTICONF data flow architecture model handles the pre-processing microservices for geospatial and temporal data to be visualized in a guided analytics dashboard, part of the TAC framework.



Fig. 1. The ARTICONF data flow architecture model

The remainder of the paper is organised as follows. Section II gives an overview of the car-sharing use-case in the context of TAC ARTICONF framework. Section III explains the TAC microservices, providing details for the geospatial, temporal and visualization components. Section IV describes TAC implementation for a car-sharing scenario. Finally, section V concludes the paper and provides future directions.

II. Car-sharing use-case in context of TAC ARTICONF framework

This section will provide a short overview of the car-sharing platform and give insight into how the car-sharing use-case can be deployed in the context of the TAC framework.

The car-sharing platform is a new collaborative model providing an alternative solution to private car ownership. This model allows customers to temporarily use and share a vehicle (on-demand) at a variable fee, charged depending on the distance travelled or time used. It is based on a sharing economy concept, which is an umbrella term for new business models where the most important thing is saving money and reducing pollution.

The car-sharing use case is composed of four microservices: (i) social network, (ii) blockchain, (iii) storage, and (iv) prescriptive microservice.

The social network is used by customers to interact, plan (where and when a vehicle is available), hire a service, or share content like photos and short videos. The availability of a social network will improve user communication and will be intended to create a community in each city. Using this network, a user can report any issue detected by posting in this social network. This includes providing information about the service, asking other users for a route, or reporting problems with the vehicle (battery, engine, etc). In order to facilitate the use of the social network and to reduce risk about the user data, this social network would be used anonymously.

Companies or private users of this system carry out payments while maintaining anonymity through permissionless blockchain networks. They can deploy smart contracts in the blockchain which will be verified and resolved automatically. These contracts will have coded penalties which will be applied automatically in case of contract breaches.

The data coming from the social media network, service utilization and geolocation system (user geolocation data) are stored in the data storage system, implemented using the CouchDB database in Fabric.

Several algorithms will run over these data providing information for management and maintenance of the system. The prescriptive microservice aims to provide suggestions to the system administrators or users to increase revenues from the system.

In conclusion, using these microservices, the use-case allows the user to rent and share a car at any place and time through smart contracts based on blockchain. Thanks to this technology, their money will be safe throughout the process and the contracts will be fulfilled automatically, including penalties. Through social networks, users can interact with others and report issues directly to the company, vehicle owner or other users: the platform will gather all the information generated by the anonymous user (geolocation, social network interaction, external events...) and this data will be used by ARTICONF, which will help to classify users. Finally, the AI (Artificial Intelligence) system will analyse this data and will use it to boost user experience and the economic benefits for users and vehicle providers involved in the model.

The proof of concept of the proposed TAC framework depicted in Fig. 1 is done using a car-sharing use-case.

The proposed TAC framework will develop and design a qualitative mapping model with context knowledge and context-sensitive interpretation. It takes car-sharing use-case data as input, structured according to the dimension's data model, and computes output data, which are also structured according to the metrics' output data modules (geospatial, temporal and cognitive). Default dimensions for the modules come with default generic functions that remain at the partners' disposal for creating custom use-case functions. For

example, the default function scenario rating to the default dimensions calculates the average score per scenario, which is the filtering input data end-user rating. A typical example of an analytical platform for integrating and computing spatio-temporal metrics, which supports intensive cluster-based data processing is given in [2]. Here the use of a microservices is supported. The stream-processing method is part of the data ingestion microservice intended for storing collected data, this can be implemented and realised as a distributed, micro-service based cloud architecture.

The TAC framework is developed using multiple individual microservices internally interconnected among them, as is depicted in Fig. 1. For the purpose of developing and designing a qualitative mapping model with context knowledge and context-sensitive interpretation, three separate microservices are provided: geospatial, temporal and visualisation. Their aim is explained in the following section.

III. TAC Microservices

The main purpose of the TAC service, as the part of the general ARTICONF platform, is supporting the analytic system, injecting additional information to improve operational tasks, planning and management of the fleet. The TAC framework is developed as a concept of interconnected microservices. The aim is producing meaningful insights for providers to improve their businesses and profits; and for users to enhance their experience and earn some extra revenues by using the ARTICONF platform. This paper put focus on creating microserices for the concrete car-sharing use case. The work can be extended for various use cases that willing to be added in the ARTICONF platform.

Thus, three separate TAC microserveses are developed: geospatial, temporal and visualisation. In the sections below a comprehensive state-of-the-art for each component and descriptions of the developed car-sharing TAC microservices are given.

A. Geospatial microservice

1) Geospatial Component - overview: Geospatial analysis can essentially be described as the gathering, display, and manipulation of imagery, GPS, satellite photography, and historical data. This information is usually represented in terms of geographic coordinates or implicitly, in terms of a street address, postal code, or forest stand identifier as they are applied to geographic models. There exist a variety of use cases where the geospatial analysis is applicable, such as crisis management, weather monitoring, climate change modelling, sales analysis, human population forecasting, and the list goes on. Usually, the geospatial data is represent as Geographic Information Systems (GIS) data. Other geospatial data can originate from GPS data, satellite imagery, and geotagging. There are several software and tools available today which can assist with the utilization of geographically referenced data in population research. Some of them are FalconView - an open source GIS software that displays various types of maps and geographically referenced overlays, GeoDa's GIS software for introducing new users into spatial data analysis with main functionality of geostatistics, GRASS GIS designed as tool for environment planning and land management, MapWindow software performs map viewer, identify features, processing tools, and print layout and other given in [3]. Geospatial data usage and analysis are crucial to be applied in all ARTICONF use-cases: crowd journalism, car-sharing, smart energy and co-created video for successful sustainability and appropriate functioning of the same. Regarding the crowd journalism scenario, the geospatial data are significant for on-time crisis informing, like the Nepal earthquake by providing a medium for citizens to communicate with one another and with those seeking to help victims. The benefits of affected peoples are immediate information sharing and visualization of dire and urgent events [4]. Therefore, crowdsourcing geospatial data has a key role in such situations. Traditional mapping is nearly exclusively coordinated and often also carried out by large organisations, crowdsourcing geospatial data refers to generating a map using informal social networks and web 2.0 technology [5]. Geospatial data can also contribute to the car-sharing use-case, helping in finding effective ways to relieve the problems of traffic jams, parking difficulties, and air pollution. Using the GIS data are targeted potential users, potential travel demand, potential travel purposes, hence, distance from existing stations are selected as the decision criteria [6]. The new model of urban transport, gives users access for short periods of rental, thus providing the benefits of using private vehicles while avoiding the inherent property charges of a vehicle. The adopted methodology of demand analysis identifies the spatial patterns of the intervening variables of socioeconomic information, transportation and land use, to understand the current panorama of the demand for transport in the wanted location [7]. The integration of energy system models and Geographic Information Systems (GIS) is still in its infancy. This integration can help in boosting the essential trends concerning energy infrastructure planning, energy generation and storage: from planned production towards fluctuating production based on renewable energy sources, from centralized generation towards decentralized generation and expensive energy carriers towards cost-free energy carriers [8].

2) TAC geospatial microservice: The main purpose of this microservice is to handles the gathering, display, and manipulation of all data consisting of longitude and latitude as information. Considering the needs of the specific car-sharing scenario the data are reused for creating meaningful information available for the car companies and the vehicle owners for boosting the experience and improving the economic benefits.

The insights that can be offered based on the geospatial information are coordinates and address of the place where the car is parked; all possible offered places where the trip can end (the company can follow whether the car is left on a right position, otherwise the user can be penalised); the reward that can be gain if the trip ends on one of the suggested destinations; list of users who started the trip and users who checked for that trip but still have not joint (this goes with the exact dates and coordinates); and the location from where

some user sends message to the platform.

B. Temporal microservice

1) Temporal Component - overview: Visualization is a powerful tool for analysing complex social networks to provide users with actionable insights even against a large amount of data. Numerous social network analysis (SNA) methods have been coupled with visualization to uncover influential actors, find helpful bridging people, and identify destructive spammers. There are many research tools and a growing number of commercial software packages, some designed for innovative large scale data analysis, while others deal with common business intelligence needs. However, few approaches or tools sufficiently address the problem of how to analyse the social dynamics of change over time. Communities are not static. Like living organisms, they evolve because of cultural, environmental, economic, or political trends, external interventions, or unexpected events. Technological developments have also strong impacts on social changes, a phenomenon that has become influential with the arrival of mobile communications devices and social networking services [9].

Temporal data usage and analysis are crucial to be applied in all ARTICONF use-cases: crowd journalism, car-sharing, smart energy and co-created video for successful sustainability and revenue growth. Regarding the crowd journalism scenario the geospatial temporal are significant for anomaly detection, correlation of events, prediction over time series data and event stream [13], allowing the tracking of verified active accounts, and allowing the system to push notifications to proven contributors and track, and increase the engagement of users. Both active posters and lurkers are almost twice as likely to return to the site when they have received an alert [14]. Temporal data can also contribute to the car-sharing use-case, helping in finding an effective way to relieve the problems of traffic jams, and seasonality is a key feature in achieving high accuracy traffic prediction [15]. Also, it makes it possible to statistically model the users' demand in terms of drop-off and pickup rates, and the parking times of vehicles to understand the current panorama of the demand for transport in the wanted location [16]. The original Bitcoin community made much out of the "trustless" nature of the technology—the fact that it does not rely on trusted central intermediaries—but newer groups are expanding the vision into one of the trust-enabling decentralized cooperatives, or "distributed collaborative organizations" [17], improving trust and security in open marketplaces and shared economies.

2) TAC temporal microservice: This microservice is designed such that to support complex analyses of the car-sharing social network. Providers and users will benefit from actionable insights, although it is about a large amount of data produced over a short time.

Considering the performed analysis coupled with visualisation users can follow the amount of their savings; when there is an available car for renting or when some certain car will become unavailable; the status of some travel whether is booked, started, finished, checked or cancel. The provider can identify the rating of the user, which reveal their behaviour during the trip; the date when the trip should start or end; or if the trip has not started or finished in a range of five minutes then to penalised the user by losing their deposit. Hence, the company can follow kilometres travelled by a concrete car; the rating of the travel given by the passengers; list of users who already started the trip together with the dates and the coordinates; plus the date of publications made by the users.

C. Visualisation microservice

1) Visualisation Component-overview: Modern applications produce large amounts of data in the form of logs and events to facilitate quick failure diagnosis and mitigation. Special Big Data database systems are needed to store and manage these logs. Through such a system, the data is used to query different events and trace down issues that appear in the application. There are several software and tools available today which can assist with the utilization of time series data in population research. A popular one is a Jupyter, an open-source project that displays various types of maps and geographically referenced overlays, the enabling Big Data analysis, visualization and real-time collaboration on software development across more than a dozen of programming languages [10].

Other commonly used monitoring tools in this field can be found presented in papers [11]. Some of them are Splunk used for storing and analyzing high amounts of log data, which has support for numerous related applications and plugins; Sumo Logic able to process logs and generate real-time insights from them, which makes it useful in quick analysis and troubleshooting of infrastructure issues; Loggly a cloud-based log analyzer. This tool compared to others let the administrators know of what issues have occurred in their system, Loggly focuses more on finding the source of the problem; Graphite a scalable tool for storing data, creating graphs based on it and monitoring performance parameters; and Cacti designed as a front-end application and it is able to store system data and display graphs constructed from it.

2) TAC visualisation microservice: The process of guiding the social network users through the ARTOCONF workflow will analyse only the parameters of its interest through a car-sharing use case recommendation. In this way, the analysis moves beyond reporting shallow summary data to acquire strong and actionable insights from users. The aggregated and summarized information will be shown on the ARTICONF dashboards to keep the car-sharing use case providers aware of what happens in the platform with the users and to summarize what happened in the past period with the scenario. As a background technology for the realisation of the above mentioned is chosen the Elastic Stack (formally known as ELK Stack). The structure of this technology is explained in [12].

The TAC visualisation microservice is responsible for aggregating and exploiting the car-sharing content diffusion supply chain across different providers, communities, groups and users. It provides information that supports user engagement in collaborative economies with monetary inclusion and increase

the provider's awareness of the users' activity on the platform and helps them track the rating and functioning of their application.

The TAC car-sharing visualisation dashboard enables the following insights depicted in Fig. 2

- A timeline series that shows the average, minimum, maximum and middle of the price for time rent.
- A gauge that gives information about the average satisfaction rating score for a concrete travel.
- Which brands of cars are the most used, hence, the types of car the most used from that same brand are depicted with a pie chart.
- A heat map showing what are the most common places where the travels start.
- A group of words, sized according to their importance, that indicates users who are active in the use-case platform.
- A horizontal bar chart that gives view of car brands and types that produce the most revenue.

IV. TAC IMPLEMENTATION FOR CAR-SHARING

This section covers all needed technical setups for development of the TAC microservices, elaborated in section III. Hence, gives results in a final form of a dashboard for the concrete car-sharing scenario.

A. Background technology and configuration setup

For achieving the results presented in this paper the Elastic Stack (ELK) technology is chosen. The ELK represent an open-source tool that provides end-to-end data analysis solutions, and helps with deep searching, analyzing and visualization of data provided across different machines. The Elastic Stack is structured of three main parts [12]:

- **Logstash** which is specialized in data collection and transformation.
- **Elasticsearch** an open-source distributed database system capable of real-time full-text search and analytics. It is built in Java based on Apache Lucene library [18].
- **Kibana** a GUI tool developed by elastic team that connects to an Elasticsearch server to navigate and display stored data [19]. Kibana offers various visualisation type outputs. The most frequently used visualisations include: line, area and bar charts; pie chart; data table; metric, goal and gauge; tag cloud, etc. Most of them are used for displaying our results.

One advantage of the ELK is that can be shared amongst the team to enable internal collaboration and boost the teamwork on the data analysis. That can be achieved by use of Beats, which is collection of lightweight data shippers.

The configuration setup used for development of the TAC framework is following: first, the Elastic Stack version 7.4.0 was installed on a Docker container on a local machine with basic license and with the following specification: CPU with 4 cores, 4 threads and 3.3GHz clock speed, 8GB RAM, 1TB HDD, internet connection with 30Mbps download speed and 40Mbps upload speed, Ubuntu 18.04. Additionally, for assuring certain security in the work-environment a basic authentication was set up. For starting the Elasticsearch node port 9200, for the Logstash port 9600, and Kibana interface port 5601 were used. In order the server to be accessible not only internally, but also externally an IP tunnel was created.

In the process of the configuration the server was set up in such manner to be capable to put, delete and get documents stored on the Elasticsearch node. Further, the Logstash was also connected with the Elasticsearch node, and now this piece of software can ingest and send data to the Elasticsearch node. By default Kibana is configured to cooperate with the Logstash that makes it ready for visualization activities.

B. Results and discussion

The TAC microservices developed for the specific car-sharing use-case, explained in section III, were integrated in the ELK. The most meaningful insights for this provider were created and presented in a form of a visualisation dashboard depicted in Fig. 2.

- Fig. 2.1 is created by use of timeline series aggregation. In this insight can be notice how the average, maximum, minimum and median price for time rent was changing in respect to time. Base on this information the car provider can plan some future concept of working and charging. The type of visualisation used here is a line chart.
- Fig. 2.2 uses metric aggregation for depicting the final result. The aim of this insight is to be find the average score of all present ARTICONF users authenticated on the car-sharing network. The result here is shown by help of a gauge visualisation type.
- Fig. 2.3 was done using the bucket aggregation. The result is visualised by a pie chart, and represents which brand car is most used by the users, and also what type of car of a certain brand was the favourite one amongst the users.
- Fig. 2.4 was created using the metric aggregation, adding now plus the count aggregation that considers the geographical points on a map. This visualisation of type heat map enables to be find what are the most common locations on the map for a starting places of all travels.
- Fig. 2.5 uses again the metric aggregation, but now for visualisation type is taken the tag cloud. Using this aggregation and visualisation, can be revealed the user who most have been checked for travel.
- Fig. 2.6 for its creation uses bucket aggregation. Bar chart type of visualisation is used, which illustrates the car that made the most revenue, and also, from the top five cars brands the producing the top level revenue, what are the most productive car types.

V. CONCLUSION AND FUTURE WORK

The ARTICONF ecosystem has provided a robust tool for analytics and cognition (TAC) monitoring for a car-sharing use case scenario. The main contribution of the proposed TAC framework was to provide social media consumers with analytics dashboard data from the collaborative participation of the car-sharing scenario. In this way, the qualitative mapping

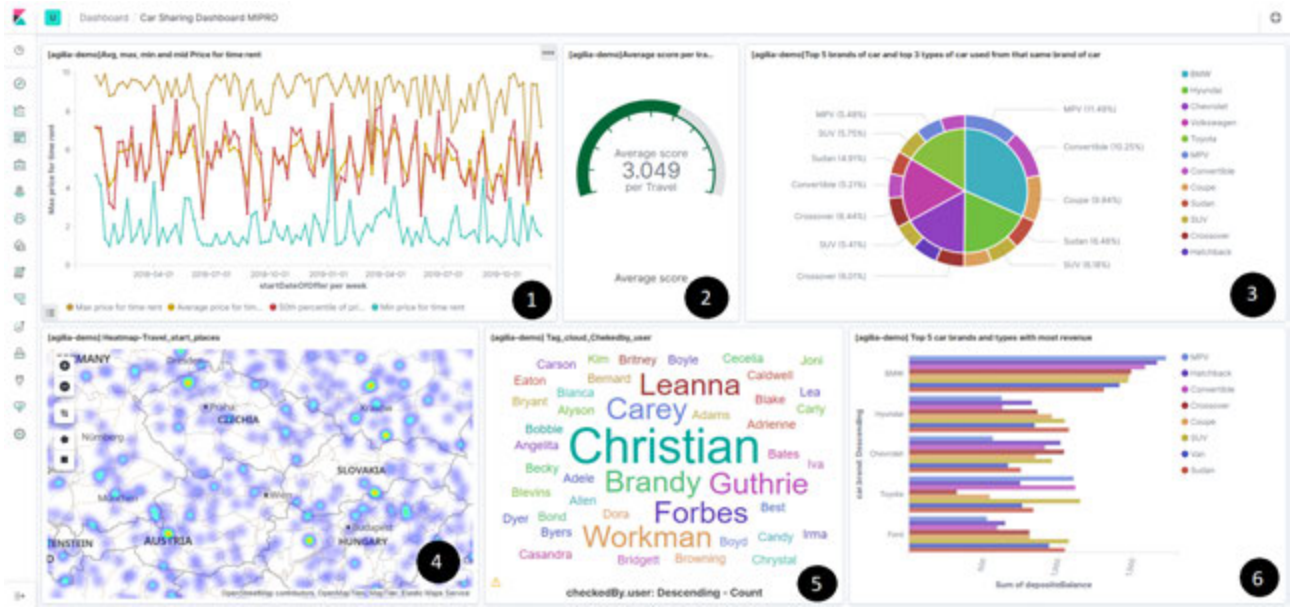


Fig. 2. TAC dashboard car-sharing implementation

model with context knowledge and context-sensitive interpretation that will improve collaboration amongst intelligently defined car-sharing communities.

The TAC framework has used the geospatial, temporal and visualization microservices to deliver a strategy for car-sharing data aggregation and group recommendation generation that is most appropriate when dealing with a specific group of intended users. The proposed tools for analytics and cognition has successfully been demonstrated in a car-sharing use case. In future, the ARTICONF platform will be piloting use cases for a crowd journalism platform; a smart energy sharing platform and video commenting technologies scenarios.

ACKNOWLEDGMENT

This research is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825134, the ARTICONF Project (https://articonf.eu/).

REFERENCES

[1] R. Prodan, N. Saurabh, Z. Zhao, K. O. Johnson, A. Chakravorty, A. Karadimce and A. Ulisses, "ARTICONF: Towards a Smart Social Media Ecosystem in a Blockchain Federated Environment", in the 7th Workshop on Large Scale Distributed Virtual Environments, Euro-Par 2019, Göttingen, Germmany, 26-30 August 2019.

[2] L. Rodríguez-Pupo, C. Granell, and S. Casteleyn, "An Analytics Platform for Integrating and Computing Spatio-Temporal Metrics," ISPRS Int. J. Geo-Inf., vol. 8, no. 2, p. 54, Jan. 2019.

[3] A. Chowdhury, "Top 10 best Geospatial tools for GIS, mapping and data visualization, Analytics India Magazine."

[4] A. Brandusescu, R. E. Sieber, and S. Jochems, "Confronting the hype: The use of crisis mapping for community development," Converg. Int. J. Res. New Media Technol., vol. 22, no. 6, pp. 616–632, Dec. 2016.

[5] C. Heipke, "Crowdsourcing geospatial data," ISPRS J. Photogramm. Remote Sens., vol. 65, no. 6, pp. 550–557, Nov. 2010.

[6] W. Li, Y. Li, J. Fan, and H. Deng, "Siting of Carsharing Stations Based on Spatial Multi-Criteria Evaluation: A Case Study of Shanghai EVCARD," Sustainability, vol. 9, no. 1, p. 152, Jan. 2017.

[7] M. O. Lage, C. A. S. Machado, F. Berssaneti, and J. A. Quintanilha, "A METHOD TO DEFINE THE SPATIAL STATIONS LOCATION IN A CARSHARING SYSTEM IN SÃO PAULO – BRAZIL," ISPRS - Int. Arch. Photo. Rem. Sens. Spat. Inf. Sci., pp. 27–32, Sep. 2018.

[8] B. Resch et al., "GIS-Based Planning and Modeling for Renewable Energy: Challenges and Future Research Avenues," ISPRS Int. J. Geo-Inf., vol. 3, no. 2, pp. 662–692, May 2014.

[9] J. Ahn, M. Taieb-Maimon, A. Sopan, C. Plaisant, and B. Shneiderman, "Temporal Visualization of Social Network Dynamics: Prototypes for Nation of Neighbors," in Social Comp., Beha.-Cult. Model. and Predic., 2011, pp. 309–316.

[10] V. Fedak, "Top 4 Popular Big Data Visualization Tools," Towards Data Science, 09-Jan-2018. [Online]. Available: https://towardsdatascience.com/top-4-popular-big-data-visualization-tools-4ee945fe207d. [Accessed: 11-Apr-2019].

[11] H. Akshaya, J. Vidya, and K. Veena, "A basic introduction to devops tools," Inter. Jour. of Comp. Sci. & Infor. Tech., vol. 6, no. 3, pp. 05–06, 2015

[12] V.-A. Zamfir, M. Carabas, C. Carabas, and N. Tapus, (2019). Systems Monitoring and Big Data Analysis Using the Elasticsearch System. 22nd Inter. Conf. on Control Sys. and Comp. Science. 2019

[13] A. Zaslavsky, P. P. Jayaraman, and S. Krishnaswamy, "ShareLikesCrowd: Mobile analytics for participatory sensing and crowd-sourcing applications," in IEEE 29th Inter. Conf. on Data Engi. Work., pp. 128–135, 2013.

[14] D. R. Millen and J. F. Patterson, "Stimulating Social Engagement in a Community Network," in Pro. of the 2002 ACM Conf. on Comp. Supp. Coop. Work, New York, NY, USA, 2002, pp. 306–313.

[15] M. Lippi, M. Bertini, and P. Frasconi, "Short-Term Traffic Flow Forecasting: An Experimental Comparison of Time-Series Analysis and Supervised Learning," IEEE Trans. Intell. Transp. Syst., vol. 14, no. 2, pp. 871–882, Jun. 2013.

[16] C. Boldrini, R. Bruno, and M. Conti, "Characterising demand and usage patterns in a large station-based car sharing system," in IEEE Conf. on Comp. Comm. Workshops (INFOCOM WKSHPS), 2016, pp. 572–577.

[17] B. Scott, "How can cryptocurrency and blockchain technology play a role in building social and solidarity finance?," UNRISD Working Paper, 2016.

[18] A. Kumar, A. Bandyopadhyay, H. Bhoomika, I. Singhania, and K. Shah, "Analysis of network traffic and security through log aggregation," Inter. Jour. of Comp. Sci. and Infor. Sec. (IJCSIS), vol. 16, no. 6, 2018.

[19] B. Akdal, Z.G.C. Keskin, E.E. Ekinci, and G. Kardast, "Model-driven query generation for elasticsearch," in (FedCSIS). IEEE, 2018, pp. 853–862.

Experimenting with Means to Store and Monitor IoT based Measurement Results for Energy Saving

M. Saari *, J. Grönman*, J. Soini *, P. Rantanen, and T. Mäkinen*
* Tampere University/Faculty of Information Technology and Communication Sciences, Pori, Finland
Mika.saari@tuni.fi

***Abstract* - Nowadays, energy consumption and especially energy saving are important issues. The news of global warming has increased the need to save energy in many areas of our living community. Therefore, several research projects on various aspects have been established. This study is part of our ongoing project with the purpose of developing a means of reducing energy consumption in houses and apartments. In particular, the project aims to find ways to increase energy savings without compromising comfortable living. This requires the constant measuring of living conditions, which produces a huge amount of data that has to be monitored continuously and stored for later analysis. In this paper, we report our experiments to select appropriate tools for storing, monitoring, and visualizing data on living conditions. In our test arrangements, the data is gathered by IoT sensors in real locations. Different types of database systems and dashboard software have been used in the trials. Each of these is discussed from the IoT data processing perspective and is presented with examples of gathered data. Finally, we present the toolset that we considered a suitable choice for our context.**

Keywords - Internet of Things; IoT; database; visualization; sensors

I. INTRODUCTION

The news of global warming has increased people's awareness of energy consumption and energy saving. These important issues should influence our everyday decisions. The Internet of Things (IoT) has introduced a lot of tools to measure the environment and its state. In the academic world, several research projects from different aspects of energy consumption and energy saving have been established. We are focusing on this issue in our ongoing KIEMI ("Vähemmällä Enemmän – Kohti Kiinteistöjen Energiaminimiä", or "Less is More: Towards Energy Minimum of Properties" in English) project. The aim of the project is to develop proof-of-concept demonstrations and prototype applications that illustrate how cost-effective, open, and modular solutions could be utilized to improve the energy efficiency of existing, older buildings. [1]

The developed prototypes were placed in indoor spaces with the purpose of collecting environmental data, such as temperature and humidity. At the beginning of the project it became clear that the project partners were also interested in knowing the quality of the indoor air, for example the CO_2 and VOC (volatile organic compound) content. This kind of continuous measuring of living conditions produces a huge amount of data. The data must be stored for later analysis. In this research we report on our experiments to select appropriate tools for storing, monitoring, and visualizing the data.

The structure of this paper is as follows: In Section II, we review the related research. In Section III, we introduce three use cases for storing IoT data. In Section IV, we include a discussion and suggestions for future research on the topic and finally, Section VI summarizes the study.

II. RELATED RESEARCH

This section introduces the related studies in this research area.

In Finland, one of the major sources of energy consumption is housing. In fact, the heating of residential buildings accounts for up to 68% of housing energy consumption. Obviously, the main reason for this is our northern geographical location. This background of energy saving in the area of real estate and housing was handled in our earlier study [1]. In addition, the introduced prototype systems [2] were developed during the KIEMI project.

The prototype systems introduced in this study were developed for testing purposes using cost-effective, open, and modular solutions. Therefore, the prototype developing process could be called rapid prototyping. This kind of prototyping and the prototypes developed were handled in more detail in an earlier research paper [1].

In the IoT context dealt with in this paper, the collected stream of data is potentially useful for a large number of applications. The data coming from each connected device can be seen as collections of time series.

A time series can be specified as a sequential set of data points, typically measured over successive time periods. The mathematical definition is a set of vectors x(t),t = 0,1, 2,... where t represents the time elapsed. The variable x(t) is treated as a random variable. The measurements taken during an event in a time series are arranged in chronological order. [3] [4]

There are several options for storing time series. Namiot [5] provides a survey of several data persistence options for time series data. He discusses a couple of traditional

relational databases as well as many NoSQL solutions. Bader et al. [6] performed a survey and comparison of open source time series databases (TSDB). Their systematic search yielded a total of 83 TSDBs, which were compared using 27 criteria.

Di Martino et al. [7] compare the performance of three different types of database systems in processing time series data: the document database MongoDB, the column family database Cassandra, and the time series database InfluxDB. Musa et al. [8] conducted a performance comparison of InfluxDB and an object-relational database, PostgreSQL. Ramesh et al. [9] present data modeling schemas for Cassandra and MongoDB to store time series data. Rinaldi et al. [4] investigate how the data model used affects the performance of database queries, using InfluxDB in their evaluation.

III. TESTED DATABASES AND VISUALIZATION

This section presents three different use cases: RuuviTAG [10] with a time series database, a data gathering system with a cloud based document database, and an embedded relational database prototype system. The proposed ways to structure data are presented at the end of the section.

A. Time series database

A time series database is commonly used in IoT related systems. This section goes through a time series data table schema, the open source time series database InfluxDB, and finally data visualization with time series databases.

1) Time series databases

Time series data is the order of the values of a variable (e.g., temperature) at regular or set intervals (e.g., hourly). Thus, it is a sequence of discrete time data. For example, time-stamped information such as log files and IoT sensor measurements can be regarded as time series data. The measurements that make up the time series are organized on a timeline that reveals information about the underlying patterns. Organized data is important because there is a relationship between time and measurements and changing the order can change the meaning of the data. An example of a time series would be hourly temperature measurements at a specific weather station etc. [4]

During the rise of IoT technologies, a new trend in data storage has also emerged in the form of time series databases. A time series database (TSDB) is a database used specifically to manage and store values of variables that change over time. Time series data can be process measurements, triggers and alarms, system and application metrics, or revenue flows. They occur in all aspects of life, so the need to monitor, follow, process, and connect with them is enough. However, because the values of the variables are constantly changing, the database must meet specific requirements in order to process such a large amount of data. The write speed must be considerably high in order to track the large amount of data that is sent to the server every second or even milliseconds. In addition, the TSDB should have high capacity and potential scalability, as the amount of data increases rapidly over time. In addition, it provides quick access to real-time analytics, which is mostly used in trending machine learning applications. [4]

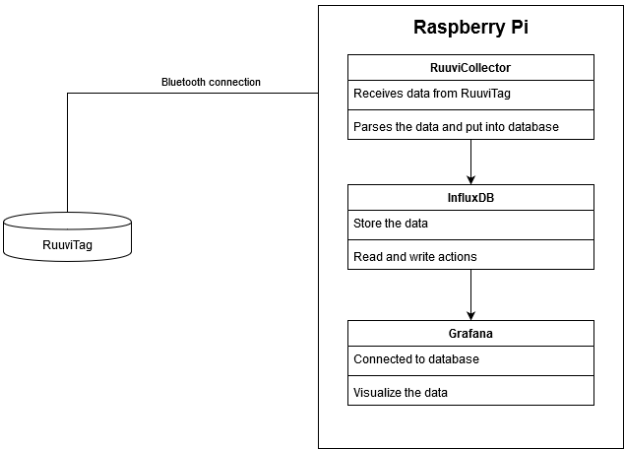


Figure 1. Open-Source sensing solution.

A TSDB can be designed using a standard SQL-based relational database, such as MySQL or PostgreSQL. Recently, NoSQL solutions have had success in designing TSDBs instead of relational databases. NoSQL databases have some advantages over relational databases: they are schema-free, with a simple DB design (no relational model is required); simple horizontal DB scaling in server clusters; easy support for availability. NoSQL databases use different data structures. Such an approach can benefit from the flexibility to represent the data and quick access to the data stored in the database, although this last point depends on the application and the structure of the data. [4]

The most effective TSDB at this time is InfluxDB [11].

2) InfluxDB

InfluxDB is an open source TSDB developed by Influx Data. It is written in the Go programming language developed by Google and is optimized to handle large write and read transactions. InfluxDB is intended for use in any application involving large amounts of time-stamped data. Examples of similar uses include DevOps monitoring, IoT sensor data, and real-time analysis. A time series database is, as its name implies, a database for storing time-stamped data, in which the data is stored and indexed according to the timestamp. In InfluxDB, data is stored in databases that contain measurements. It is possible to define retention policies for these databases. Retention instructions tell the database how long data will be retained and how many copies of that data will be retained in the cluster. The measurements contained in the databases can be compared to the tables in SQL-based databases, because at the conceptual level they are very similar to each other. [12]

Database data is processed using the Influx query language. The syntax of the language is expediently very similar to the SQL query language so that users that have experience with other SQL query languages can easily adopt its use. Database searches are performed using the SELECT statement, and it is possible to use other refinement criteria in addition to it, such as GROUP BY, INTO, or WHERE. The language also includes other useful functions for retrieving information. One database consists of at least two mandatory fields: a timestamp and a field. The timestamp is stored in the time column for each data point in the format RFC3339 and there can be only one for each data point. The field consists of a key value pair,



Figure 2. Example of a Grafana dashboard layout with charts.

where the column name acts as a key (Field Key) and the corresponding data points as values (Field Value). The keys are stored as strings, while values can be stored as either strings, truth value variables, or as floats or integers. The database can also contain tags. According to the fields, the tags consist of a key-value pair, where the column name acts as a key (Tag Key) and the corresponding data points as values (Tag Value). All tags are stored as strings. Although tags are not mandatory, their use is desirable mainly because, like fields, the columns are not indexed. As a result, database searches for tags are faster and are ideal for storing frequently retrieved data. All of the above together form a point. The point is thus in practice one line in the same series with the same timestamp. [12]

3) Data visualization

Data visualization is the presentation of data and data in the form of graphical components such as graphs, charts, histograms, gauges, geographic maps, etc. Data visualization helps users understand the meaning, structures, and correlation of data by displaying raw data more comprehensibly.

Grafana is a general-purpose open source visualization tool developed by Grafana Labs. It makes it possible to create different dashboards that allow time-stamped data to be visualized. Grafana runs as a web application and officially supports eight different data sources including Elasticsearch and InfluxDB, among others. Since version 3.0, Grafana has been able to install other data sources as plugins, but they are not officially supported. Dashboards play a key role in Grafana. A single dashboard is in practice an organized collection of different panels that together form a visual representation of the data collected. There can be one or more dashboards and it is possible to import or export them from the system as JSON documents. Panels are individual components that visualize data from a desired source, for example in the form of a graph. Each panel includes a Query Editor with slightly different functionality depending on the data source used. The Query Editor makes it easier to retrieve data from the data source used, and any changes you make to it in the panel Query are immediately reflected. Like dashboards, panels can be imported or exported from the system as JSON documents. [13]

Grönman et al. [14] present a low-cost and flexible solution for environmental sensing. The implementation is based on open-source components. The solution utilizes an ROS (Robot Operating System), RuuviTag sensor, and Raspberry Pi 3 computer to achieve a low-cost solution. By implementing a pilot project, they concretized the potential of their approach to environmental sensing. Potential application areas of this solution include the microclimate control of greenhouses and warehouses; for example, robots can also be used in environments where it is unsafe for humans to enter. Compact robots may also reach places that are inaccessible to humans. The robotics solution can utilize the information transmitted by IoT devices and sensors. RuuviTag is connected via Bluetooth to the Raspberry Pi and the collected data is transferred to the TSDB database. The stored data can be visualized by the visualization software.

In this case, they decided to choose three main software programs to take care of the functionality of the sensing solution. On the software level, the solution has three important programs: RuuviCollector, InfluxDB, and Grafana. First, the RuuviTag is switched to “RAW” mode, which means data comes straight from the sensor and no

modification is made on the way. RuuviCollector uses a Raspberry Pi 3 computer with built-in Bluetooth to listen to the data coming from RuuviTag. RuuviCollector needs to parse the data into the correct format and then puts the data in the InfluxDB database. When the sensor data is stored in InfluxDB, then the connection between Grafana and the database needs to be made. Then Grafana simply reads the database and the user can create dashboards to visualize the data. The software and hardware connections are shown in Fig. 1.

Fig. 2. shows the Grafana user interface in the browser window. The user interface is designed to work in a web browser and can also be accessed from public or private networks. Information content can be displayed on the user interface. Fig. 2. also shows the RuuviTag measurements of temperature, humidity, pressure, acceleration x, acceleration y, acceleration z, battery voltage, RSSI, and counts of measurements per hour.

B. Document database

In this section, one developed prototype measurement system is presented. A document database test environment was installed and tested in a real environment, a detached house. The focus was to test the data flow and usability from the sensors to a cloud database service. In addition, during the test, a mobile application was developed to inform the user.

The document database was tested in a detached house where the main heating system is a fossil fuel (wood) fired water heater. The heating is managed by the resident and the wood burning is not automated. The house has central heating with water circulation, which is automated. The water circulation power depends on the outside temperature. The measurement system is illustrated in Fig. 3.

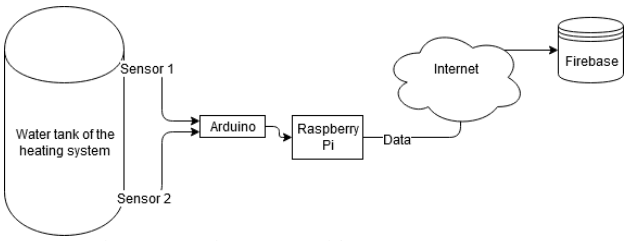


Figure 3. Heating system with measurement system.

The measurement system consists of sensors. The data from the temperature sensors is collected with an Arduino, a single board microcontroller that can handle several types of sensors. The Arduino is connected to the Raspberry Pi, which is small single-board computer (SBC) with a Linux Operating system (OS). The SBC can handle several Arduinos with wired or wireless connections. In the presented case we used a wired connection, but if a wireless connection is used the Arduino is able to work for long periods powered by battery.

The Raspberry Pi is commonly used in rapid prototype test cases due to its ease of use. The Raspberry Pi and the peripherals are off-the-shelf devices, which are low-cost and easily available. In this scenario, the Raspberry Pi was connected to the Internet with a WiFi connection. This connection is needed because the collected data is sent to the Google Firebase Realtime Database. The data includes a timestamp with date and time, and two temperature sensor values.

In the usage test, the measurement system collected 4.3 MB of data during about 10 months. The data was collected once every ten minutes and the database included over 40 thousand datapoints.

The data in the database could be handled in several ways. The Firebase Realtime database provides an interface for reading, modifying, and saving data. In this example case, the data was used with the TempApp mobile application. The TempApp is an Android application and is coded for testing purposes. The main idea was to test the data usage possibilities. The basic functionality of the TempApp application is to retrieve data on a mobile device and show it to the user.

The software framework of the measurement prototype consists of several software components. Fig 4. lists the software used with the devices.

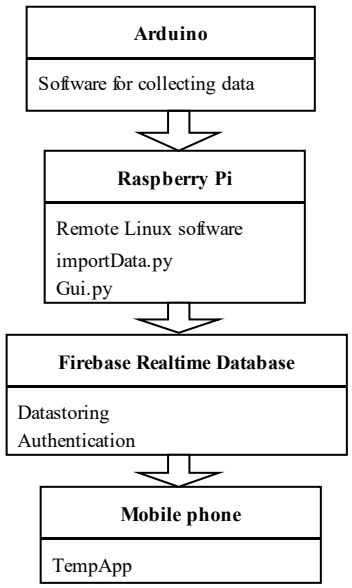


Figure 4. Hardware with software components.

The software in Raspberry Pi has several tasks: importData.py gets data provided by Arduino and sends the data to the database. Gui.py was developed for testing purposes to show that the data collected is acceptable. This was necessary during application development. Raspberry Pi runs with Linux OS and therefore the remote management capabilities provided by Linux were utilized. A remote desktop is one useful tool for the testing phase. Raspberry Pi also stores the data in a file, which can be retrieved easily. Fig 5. shows a bar graph of about one week of data.

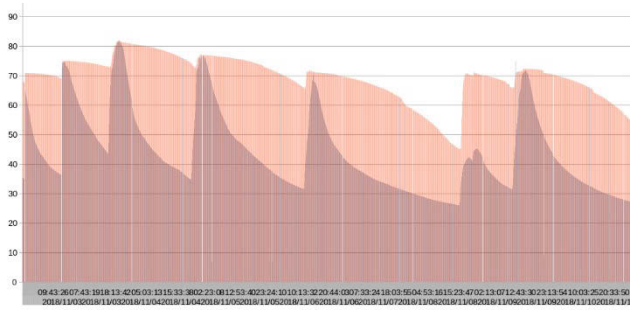


Figure 5. Collected data from measurement prototype.

The database handles data storing and offers the necessary interfaces for using the data. Furthermore, Firebase provides the authentication services that were utilized in the Firebase connections with Raspberry Pi and the TempApp mobile application. The mobile phone runs the TempApp software, the main purpose of which was to provide the user with information.

C. Relational database

In one of our experiments [2], we used the embedded relational database, SQLite, for persistent time-series data. The aim of the study was to test cost-effective, off-the-shelf components for measuring indoor air quality. In one of the cases, a sensor system was installed in an apartment building used for eldercare. In this section, we present a simplified schema of the database of the experiment and discuss corresponding schemas for document and time series databases, especially for MongoDB and InfluxDB.

Table 1 shows an excerpt of the data collected by the sensor system. The structure of the data follows the following relational schema - tables and their columns, primary keys (PK), alternate keys (AK), and foreign keys (FK):

sensor_nodes: {node_id^{PK}, description}
sensor_data: {data_id^{PK}, timestamp^{AK}, eco2, humidity, pressure, temperature, node_id^{FK}}

Table 1. Excerpt of data captured by the sensors in the experiment.

ECO_2	Humidity	Pressure	Temp. (°C)	Timestamp
541	57.96	1007.44	23.54	2019-09-12 T12:00:37.000+0000
600	57.88	1007.50	23.55	2019-09-12 T12:03:57.000+0000
646	57.82	1007.82	23.57	2019-09-12 T12:07:17.000+0000

The conceptual representation of the data is shown in Figure 6. There are two entity types, *Sensor Node* and *Sensor Data*. *Sensor Node* has two attributes, *ID* and *Description*, of which the second one is key. The key attribute of *Sensor Data* is *Timestamp*. The other four attributes represent the information collected at the timestamp, *ECO2* (CO_2 level), *Humidity*, *Pressure*, and *Temperature*. In the conceptual representation, the foreign key column of the relational schema has been replaced with the relationship type, *collects*. The primary key column of the *sensor_data* table is a surrogate key and therefore it has been omitted.

A record in a traditional relational database is a row, whose counterpart in *MongoDB* is a *document*. The structure of the document is composed of field and value pairs, being similar to JSON objects. The values of fields may include other documents, arrays, and arrays of documents.

From the conceptual schema in Fig. 6, we can derive different kind of schemas for the MongoDB document database. One of them is similar to the relational schema shown earlier:

```
sensor_node: {node_id, description}
sensor_data: {timestamp, eco2, humidity, pressure, temperature, node_id}
```

Like the rows in relational databases, the documents following the structure above are flat: the values of fields are simple – no documents or arrays. The associations between the documents are implemented by storing the key field of the *sensor_node* document in the *sensor_data* document, which corresponds to the principle in relational data structures.

Another way to structure data is to use only one type of document:

```
sensor_node: {
   node_id, description,
   sensor_data: [{
        timestamp, eco2, humidity, pressure, temperature}]
}
```

The document stores information on both the sensor nodes and the data they collect. The *sensor_data* is now a field of the *sensor_node* document. The field is an array of

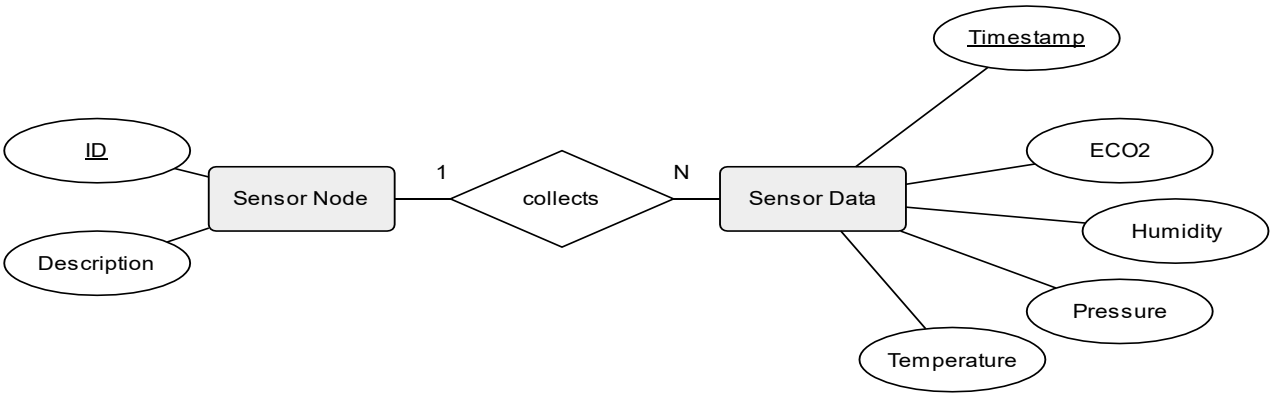


Figure 6. Conceptual schema of the database using prototype system.

documents based on the data collected by a sensor node. Alternatively, the information about the sensor node could be included in the *sensor_data* document as follows:

```
sensor_data: {
   timestamp, eco2, humidity, pressure, temperature,
   sensor_node: {node_id, description}
}
```

The InfluxDB *time series database* stores *measurements*, which are containers of *time*, a *field set,* and an optional *tag set*. *Time* contains timestamps, and the *field sets* and *tag sets* are collections of key-value pairs. The value of a *field* can be of string, float, integer, or Boolean type, while a *tag* value can only be string type. Unlike fields, tags are indexed. [8]

```
sensor_data: {
  time,
  field_set: {eco2, humidity, pressure, temperature},
  tag_set: {sensor_node}
}
```

Converting the relational structure of our experiment for a time series database is relatively straightforward, as shown above. As in the corresponding relational schema, the data structure contains *time*. The data collected by the system forms the *field* set. The data is classified based on the sensor nodes, which represent *tags*.

IV. DISCUSSION AND CONCLUSION

The "best choice" depends on several factors: data size, frequency of data saving, and usage of collected data. Therefore, each use case needs a different kind of data storage system. The previous section presented suitable ways to store data in our use cases.

The study points out several possibilities for future research: How to define a general data structure for storing data? This issue is touched on in the conclusion of use cases. Furthermore, the study raises a question about a technical issue: How to manage databases–cloud based or in one's own managed servers?

The study is part of our ongoing project, the purpose of which is to develop a means of reducing energy consumption in houses and apartments. The presented practical cases focus on the continuous measuring of living conditions, which produces a large amount of data. The data has to be monitored continuously and stored for later analysis.

The study has introduced three different use cases of storing and handling IoT data. Appropriate tools for storing, monitoring, and visualizing the data of living conditions have been presented. In our test arrangements, the data was gathered by IoT sensors in real locations.

Different types of database systems and dashboard software were used in the trials. As the conclusion of our study, we found the *time series database* with the general-purpose visualization tool for a suitable toolset for this kind of context.

ACKNOWLEDGMENT

This work is part of the KIEMI project and has been funded by the European Regional Development Fund and the Regional Council of Satakunta.

REFERENCES

[1] M. Saari, P. Sillberg, J. Grönman, M. Kuusisto, P. Rantanen, H. Jaakkola and J. Henno, "Reducing energy consumption with IoT prototyping", Acta Polytechnica Hungarica, Volume 16, Issue Number 9, 2019.

[2] P. Rantanen and M. Saari, "Towards the utilization of cost-effective off-the-shelf devices for achieving energy savings in existing buildings", Proceedings of 2020 IEEE 10th International Conference on Intelligent Systems (IS20), Varna, Bulgaria, 26-28 June 2020. submitted.

[3] R. Adhikari and R. K. Agrawal, "An Introductory Study on Time Series Modeling and Forecasting," arXiv Prepr. arXiv1302.6613, vol. 1302.6613, pp. 1–68, Feb. 2013.

[4] S. Rinaldi, F. Bonafini, P. Ferrari, A. Flammini, E. Sisinni, and D. Bianchini, "Impact of Data Model on Performance of Time Series Database for Internet of Things Applications," in 2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), 2019, pp. 1–6.

[5] D. Namiot, "Time series databases," Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL2015), vol. 1536, pp. 132–137, 2015.

[6] A. Bader, O. Kopp, and M. Falkenthal, "Survey and comparison of open source time series databases," Lecture Notes Informatics (LNI), Gesellschaft fur Informatik, vol. 266, pp. 249–268, 2017.

[7] S. Di Martino, L. Fiadone, A. Peron, V. N. Vitale, and A. Riccabone, "Industrial Internet of Things: Persistence for Time Series with NoSQL Databases," Proceedings - 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE 2019, pp. 340–345, 2019.

[8] E. Musa, G. Delač, M. Šilić, and K. Vladimir, "Comparison of Relational and Time-Series Databases for Real-Time Massive Datasets," in 2019 42th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2019, pp. 1065–1070.

[9] D. Ramesh, A. Sinha, and S. Singh, "Data modelling for discrete time series data using Cassandra and MongoDB," in 2016 3rd International Conference on Recent Advances in Information Technology (RAIT), 2016, pp. 598–601.

[10] Ruuvitag, retrieved May 12, 2020, from https://ruuvi.com/files/ruuvitag-tech-spec-2019-7.pdf

[11] DB-Engines Ranking of Time Series DBMS, retrieved May 12, 2020, from https://db-engines.com/en/ranking/time+series+dbms

[12] InfluxDB 1.8 documentation, retrieved May 12, 2020, from https://docs.influxdata.com/influxdb/v1.8/

[13] Grafana, retrieved May 12, 2020, from https://grafana.com/grafana/.

[14] J. Grönman, M. Saari, J. Vihervaara and J. Viljanen, "An Open-Source Solution for Mobile Robot Based Environmental Sensing" in 2020 43th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2020, submitted.

An Open-Source Solution for Mobile Robot Based Environmental Sensing

J. Grönman*, J. Viljanen*, J. Vihervaara*, and M. Saari*
* Tampere University/Pori Campus, Pori, Finland
jere.gronman@tuni.fi

Abstract - **Internet of Things-based devices are equipped with sensors and network connections to collect data and store this data to cloud services. This collected data allows operational decision-making processes to be based on a more accurate picture of the current state of physical environment than before. Existing data collection methods often rely on sensors in fixed locations to obtain environmental measurements. Such a solution is not very scalable in terms of the density of the measuring points. In a dynamic environment, the cost of maintaining such fixed-based solutions may also be high. A more flexible solution will be achieved by equipping an autonomous mobile robot with sensors. Mobile robots are increasingly being used for data-gathering in a wide range of environmental applications. However, a robotic solution can be expensive. In this paper, we present a low-cost solution based on open-source components. The solution utilizes a robot operating system, sensor and IoT-board based on open-source implementation. By implementing a pilot project, we concretize the potential of our approach to environmental monitoring. Potential application areas of this solution include the microclimate control of greenhouses and warehouses, for example.**

Keywords – environmental sensing; mobile robot; open-source

I. INTRODUCTION

Environmental sensing includes the processes and activities to characterize and monitor the quality of the environment. Recent advances in environmental sensing technologies have introduced a wide range of new environmental sensors with reduced cost and size. This opens new possibilities to put into service a wide variety of smart applications which improve the efficiency of production processes and human well-being.

The Internet of Things [1][2] connects environmental sensors to the Internet allowing interaction between the physical and digital world. The computer-based digital world includes cloud services, data analytics and other software-based intelligent services. This interaction enables real-time data collection from the physical environment and to store the collected data to cloud services. Therefore, environmental sensing with IoT approach enables a continuous view of the current state of the environment. The collected real-time data allows decision-making processes to be based on a more accurate picture of the current state of physical environment than before.

Existing environmental sensing methods often rely on sensors in fixed locations to obtain measurements. For example, Wireless Sensor Network have been utilized to implement such measurement solutions [3][4]. However, such solutions are not very scalable in terms of the density of the measuring points. In a dynamic environment, the cost of maintaining such fixed-based solutions may also be high. A more flexible solution will be achieved by equipping an autonomous mobile robot with sensors. Mobile robots are increasingly being used for data-gathering in a wide range of environmental applications, because robotics is seen as a promising tool to enhance environmental data collection [5].

Unfortunately, a fine robot-based implementation can sometimes be too expensive. Large-scale utilization of open-source components can offer a solution to this problem. Open-source software and components are often cost-competitive. In addition, open-source is a development model that has shown many other significant benefits in many application areas [6][7][8]. The functions contained in an open-source implementation are manageable. It can be ensured that the code contains only the functionalities required for implementation. Open-source applications can be developed with an iterative approach. This allows implementation to be modified during development to meet new requirements. The small-scale development process remains flexible enough to accept case-by-case changes. Open-source implementations are also often reliable because of the motivated team of experts involved in development and testing. Obviously, the open-source approach also has its drawbacks, as disclosed in the papers [6][7].

This paper presents a low-cost and flexible solution for environmental sensing. The implementation is based on open-source components. The solution utilizes a ROS robot operating system, RuuviTag sensor and Raspberry Pi IoT-board to achieve a low-cost solution. By implementing a pilot project, we concretize the potential of our approach to environmental sensing. Potential application areas of this solution include the microclimate control of greenhouses and warehouses, for example. Robots can also be used in environments where it is unsafe for humans to enter. Compact robots may also reach places that are inaccessible to humans.

The rest of the paper is organized as follows. Section 2 introduces the backgrounds techniques of our pilot project. It also briefly presents some related works.

Section 3 presents the pilot project. Section 4 presents the main conclusions of the project.

II. BACKGROUNDS

A. Components of the pilot project

The mobile robot used in the pilot project is Turtlebot 2. The robot moves by adjusting the speed of the drive wheels. The robot's movement and positioning utilize odometry, which analyzes the distance traveled by the drive wheels [9]. In addition, the robot has sensors to detect obstacles. It is powered by the battery. The battery is recharged at a docking station where the robot can dock independently when the battery charge level is too low. Turtlebot2 is provided with different kinds of sensors, including Microsoft Kinect 360 3D depth camera.

ROS (Robot Operating System) [10] is an open-source operating system developed for robot control and management. ROS is a meta operating system that runs on top of the Linux operating system. It has a modular structure and is supported by a large number of special software packages. Its role is to manage software packages and control communication between different processes. Software packages can be utilized in autonomous navigation and in the design of motion paths, for example. Its programming is not tied to a specific programming language.

Raspberry Pi is an open-source IoT board for providing a complete Linux computer at a very low cost [11]. It is a card-sized minicomputer that can operate also with the power of a battery. Raspberry Pi provides an interface for sensors and actuators through the general purpose I/O pins. Sensors can also be connected to it via wireless connections such as Bluetooth. For control purposes, it can be connected wirelessly via a WLAN port by using a secure shell (SSH) session, for example.

RuuviTag is an IoT-sensor including a circuit board that comes with a waterproof and breathable plastic housing. The circuit board contains a sensor for measuring ambient temperature, relative humidity and barometric pressure. In addition, the circuit board includes its own sensor for measuring acceleration. It is powered by a battery that is promised to function for a few years. For communication purposes, the circuit board includes a Bluetooth chip that enables communication with other devices, such as smartphones or Raspberry Pi. RuuviTag does not have its own IP address. If its data must be read via the Internet, it must be connected via Bluetooth to a Raspberry Pi-based Internet server, for example, RuuviTag is able to communicate with Raspberry Pi with the help of an open-source Python library. Above all, RuuviTag has an open-source firmware. This way it can be customized to suit the needs of the application area, as is done in the paper [12].

InfluxDB [13] is an open-source time series database that provides storage of automatic measurements of Ruuvitag parameters, including the results of calculations as well as assigning a timestamp to each measurement.

Grafana is a program that can be used to easily create graphical control screens [14]. Grafana is an open-source and free program. Grafana is designed for use with time series databases and it supports many different time series databases, such as InfluxDB. Grafana was originally made for monitoring data centers, but its use has spread to other industries.

B. Related studies

There are several studies on mobile robots [15][16]. Several studies also exist to deploy mobile robots for environmental sensing both indoor and outdoor. Here are some of them briefly outlined. However, a few studies focus on the research area from an open-source perspective.

The study [17] demonstrated the mobile platform in a laboratory experiment of measuring air-change effectiveness. By comparing the measurements from the mobile platform and those from a standard dense sensor network, they showed that the automated mobile sensing approach was able to determine the air-change effectiveness with high spatial granularity and accuracy. The paper [18] presets a study where three prototype robots were used to monitor pollution. The robots were capable to autonomously navigate in real urban environments. Different kinds of sensors were integrated into these robots for environmental sensing. Because the solution used sophisticated technology, it could not be considered a low-cost solution.

The study [19] presents a robotics platform for performing environmental monitoring in data centers. They present a platform based on the Robot Operating System, in which a mobile robot is able to autonomously navigate in a data center room for executing measurements at different locations. They observed that the robot avoided fixed and dynamic obstacles successfully. The paper [20] proposes a robot for human rescue operation in an environment which is unsuitable for any human intervention. The robot is capable of autonomous operation by being guided using sensors. The robot is mounted with sensors that help it navigate while operating autonomously. The Sensors also enables the robot to detect a human in need of rescue. The robot was designed for rescue operations by allowing a minimal threat to human life.

III. IMPLEMENTATION

Initially, the robotics solution was configured for a healthcare pilot. The pilot focused on an autonomous mobile robot that could assist nurses in their work. In this study, a modified version was made outlining a solution to various practical cases where a robot can operate when a human is not capable of those conditions, for example oxygen-free warehouses, hazard or hot/cold areas of the industry.

The goal was to create a robotics solution with an autonomously operating mobile robot capable of utilizing IoT devices and sensors. The solution supports browser-based implementation and the robot is remotely controllable. It is possible to connect different sensors to the unit, including temperature, humidity, motion or light sensors. The implementation follows open-source principles and the robotics solution utilizes software found

on the official ROS website. The user interface is implemented according to the current web design trends by using HTML5, CSS and JavaScript.

The robot platform used is the Turtlebot 2 mobile robot with Kinect Xbox 360 3D sensor. The mobile robot is controlled by a Raspberry Pi 3 Model B+ computer that has an open-source ROS operating system developed for robot control. The mobile robot operates in a wireless network, moves autonomously within a defined area, and can be tracked and controlled remotely. The robotics solution is also capable of utilizing the information transmitted by IoT devices and sensors. The user interface is designed to work in a web browser and can be accessed also from a touch screen mounted on the robot. Information content can be displayed in the user interface. RuuviTag is connected via Bluetooth to the Raspberry Pi and the collected data is transferred to the Influx database. The stored data can be visualized by the Grafana software. The Grafana software runs on a website, which is accessible to all devices of the same wireless LAN on a browser basis.

When configurating the robotic solution, a base map must first be created for the area. The map is created by using the RViz (ROS Visualization) software that uses the robot's sensor data to visualize the area. With the RViz software the user can monitor the robot's movements on the map in real time. By combining this with the live image produced by the camera, the new area can be mapped entirely from a remote device via a wireless network. Because the RViz software is constantly updating the map image, one option was to use the program on a VNC (Virtual Network Computing) connection.

ROS consists of processes represented by nodes that are programs that communicate with each other. The program can send or receive messages belonging to a specific topic. ROS is designed to be modular and the robot control system consists of several nodes, each with its own function. For example, one node controls the electric motor and another node controls the route.

There are many different example programs available for Turtlebot2 that can be used as explained above. The example programs can be seen in Mark Silliman's Turtlebot2 github-page [21. The example programs are not always ready for specified needs. Therefore, the code must be modified. This robotics solution uses modified version of the follow the_route.py program. In our solution, each defined point on the map must be capable of performing a specific function. Therefore, the arrival of the robot at the spesific point of the map must be detectable by the program code. For this purpose, we created a program that sends a unique message with its own topic when the robot reaches the map point. To listen to these messages, a separate Python program was created. This program executes the actions defined by the code when the program detects the message associated with the map point. Figure 1 shows the files and programs associated with the customized follow_the_route.py program.

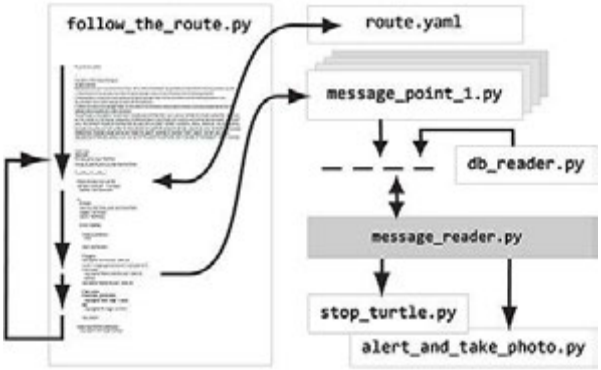


Figure 1. The program files of the implementation

Follow_the_route.py controls the robot in the desired area delimited by predefined map points. The route.yaml file contains the x and y coordinates of predefined points.

In the original follow_the_route.py program, the robot took a picture at each point, but it was modified for this implementation to send a unique message. This will keep track of the points that have been written in the log file and, at the same time, use the message_reader.py program to launch new programs at each point as needed.

At the same time, the RuuviTag sensor sends temperature and other data to the database that is read actively by db reader.py. If a predetermined value of the RuuviTag sensor is exceeded in the database, the db_reader.py program will send a message such as "Alert, temperature 44."

Message_reader.py also reads messages continuously. If the message contains "Alert," the program starts the stop_turtle.py program that stops the robot. In addition, the alert_and_take_photo.py program is launched. The program sends an alert to the user interface and takes a picture of the place.

Because almost every mobile device has a web browser, it was decided to implement a web user interface. The user interface and its development do not require installation of separate development environments. Because the ROS operating system does not support the web browser based communication model, the browser is not capable of communicating with the ROS operating system. A rosbridge interface has been developed to enable communication between the ROS operating system and web applications. Figure 2 shows a typical web application that uses rosbridge.

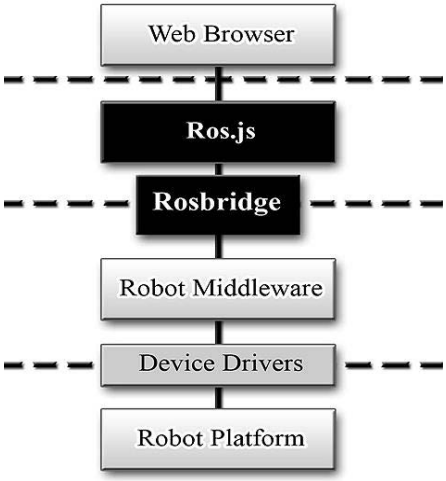


Figure 2. The rosbridge interface between the ROS operating system and web applications

There is a 3rd party Javascript library known as ros.js that has been developed to help creating web applications. The library is designed to integrate seamlessly with the ROS web application. Although it is versatile and extensive, ros.js can be used with a very simple code.

Figure 3 shows the complete layout of the user interface in the browser window. This view acts as the front page of the user interface. It displays temperature, humidity, barometric pressure, and acceleration data as measured by the RuuviTag sensor. With browser-based implementation, the same action logic can be utilized both locally on the touch screen of the robot and in the interfaces of remote devices.



Figure 3. The layout of the user interface

IV. Results and Discussion

The robot created a sufficiently accurate map of the area, but the mapping could not detect all the obstacles in the area to be mapped. When creating the map, it was found that the robot detects thin obstacles poorly. Such objects should either be marked on the map manually or removed from the environment. However, the objects that are manually marked on the map can cause problems because the robot may not be able to locate itself if it cannot find the object shown on the map. In addition, the sensor could not detect very low obstacles. These results are consistent with the findings of other studies and our previous research findings.

When environmental measurements were taken, the robot was able to locate and find the desired measurement points with sufficient accuracy. This was facilitated by the limited size of the area used for piloting, which was about 20X20 meters. A good enough map from a small space can be created with a reasonable amount of work compared to a large space. In addition, errors due to odometry do not become a big problem in a map made of a small space. The area contained a sufficient number of detectable obstacles to ensure reliable testing.

The used mobile robot Turtlebot 2 is inexpensive and well suited for testing. Because it is intended only for indoor use and flat floor surfaces, demanding conditions require a more robust and expensive physical robot implementation than the one used in this study. However, the pilot project showed that the low-cost robot can also provide advanced functions with the help of ROS and sophisticated software components.

It was also found that the processing power of the Raspberry Pi 3 computer is very limited. Its graphics processing unit (GPU) will not be suitable if there is a lot of graphics content to take care of. An example if there are a lot of obstacles that need to be identified from the video or image. Much better results can be produced using a variety of low-cost computers such as the NVIDIA Jetson Nano or Jetson Xavier NX. These computers include the more efficient GPU needed for video and image processing. More processing power will enable the use of neural networks in the future studies.

The sensor used in the pilot project was cheap and easy to deploy. However, the sensor was able to measure the environmental parameters with sufficient accuracy. In addition, the sensor can be customized to better fit the needs of the application area because its implementation is based on the open-source approach. The Bluetooth data connection offered by the sensor met the requirements of the application, because the mapped area was compact.

V. Conclusions

The main objective of this study was to provide a low cost and flexible solution for environmental sensing. The study introduced the low-cost environmental sensing solution based on a mobile robot and open-source components. We implemented a pilot project to ensure that the proposed solution met our design target. The pilot project showed that it is possible to create mobile solutions for environmental sensing using open source

components, but the development project requires a reasonably large amount of work.

REFERENCES

[1] A. Whitmore, A. Agarwal, and L. Da Xu, The Internet of Things-A survey of topics and trends, Information Systems Frontiers, 2014, pp. 1-14.

[2] H. Saha, A. Mandal, and A. Sinha, Recent trends in the Internet of Things, IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), 2017, 4 pages.

[3] T. Alhmiedat, A survey on environmental monitoring systems using wireless sensor networks, Journal of Networks, 2015 10(11), pp. 606-615.

[4] E. Ganesh, IOT based environment monitoring using wireless sensor network, International Journal of Advanced Research, 2015, vol. 5, no. 2, pp. 964-970.

[5] M. Dunbabin and L. Marques, Robots for environmental monitoring: Significant advancements and applications, IEEE Robotics & Automation Magazine, 2012, 19(1), pp. 24-39.

[6] J. Đurković, V. Vuković, and L. Raković, Open source approach in software development - advantages and disadvantages, Management Information Systems, 2008, Vol. 3 No. 2, pp.29-33.

[7] M. Heron, V. Hanson, and I. Ricketts, Open source and accessibility: advantages and limitations, Journal of interaction Science, SpringerOpen, 2013.

[8] G. Signorini, Open Source and Sustainability: the Role of Universities, 2019, https://arxiv.org/abs/1910.06073, 20 pages.

[9] S. Mohamed, M. Haghbayan, T. Westerlund, J. Heikkonen, H. Tenhunen, and J. Plosila, A survey on odometry for autonomous navigation systems. IEEE Access, 7, 2019, pp. 97466-97486.

[10] M. Quigley, B. Gerkey, and W. D. Smart, Programming robots with ROS: a practical introduction to the Robot Operating System, O'Reilly Media, 2015.

[11] A. Kyaw, H. Truong, and J. Joseph, Low-cost computing using Raspberry Pi 2 Model B, Journal of Computers, 2018, 13(3), pp. 287-299.

[12] F. Álvarez, L. Almon, H. Radtki, and M. Hollick, Bluemergency: Mediating post-disaster communication systems using the Internet of Things and Bluetooth mesh, arXiv preprint arXiv:1909.08094, 2019.

[13] S. Naqvi, S. Yfantidou, and E. Zimányi, Time series databases and influxdb, Studienarbeit, Université Libre de Bruxelles, 2017, 44 pages.

[14] E. Betke and J. Kunkel, Real-time I/O-monitoring of HPC applications with SIOX, elasticsearch, Grafana and FUSE, In International Conference on High Performance Computing, 2017 , pp. 174-186.

[15] G. Nagymate, Android based autonomous mobile robot, Recent Innovations in Mechatronics, 2015, 4 pages.

[16] J. Martins, MRSLAM - Multi-Robot Simultaneous Localization and Mapping, 2013, https://ap.isr.uc.pt/archive/MRSLAM_dissertacao_Joao_Martins-vfinal-040913_235208.pdf, 51 pages.

[17] M. Jin, S. Liu, S. Schiavon, and C. Spanos, Automated mobile sensing: Towards high-granularity agile indoor environmental quality monitoring, Building and Environment, 2018, 127, pp. 268-276.

[18] M. Reggente, A. Mondini, G. Ferri, B. Mazzolai, A. Manzi, M. Gabelletti, P. Dario, and A. Lilienthal, The dustbot system: using mobile robots to monitor pollution in pedestrian area, 3rd Biannual International Conference on Environmental Odour Monitoring and Control, 2010, pp. 273-278.

[19] L. Terrissa, S. Ayad, and N. Zerhouni, Robotics based Solution for data center e-monitoring, In International Conference on Advanced Systems and Emergent Technologies, 2019, pp. 201-208.

[20] A. Maurya, M. Sonkusare, A. Raut, D. Tamhane, and D. Palase, Surveillance robot with human detection. In Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 107-110.

[21] M. Silliman, Mark Silliman's Turtlebot github-page for developers interested in robotics, https://github.com/markwsilliman/turtlebot.

Providing Facilities in Health Care via Brain-Computer Interface and Internet of Things

Mehar Ullah*, Amin Hekmatmanesh*, Dimitri Savchenko*, Renan Moioli**, Pedro Nardelli*,
Heikki Handroos* and Huapeng Wu*
* LUT University, Lappeenranta, Finland
** Federal University of Rio Grande do Norte, Natal, Brazil
Contact: mehar.ullah@lut.fi

***Abstract*—Internet of Things (IoT) technologies are steadily growing. Millions of devices are now connected to each other and the internet; they communicate and share data. Another technology that has been taking traction is Brain-Computer Interface (BCI), which get brain signals and convert them into commands that are employed for control of different applications. There is then a clear link between IoT and BCI: the IoT device can receive the commands generated by the BCI and perform the desired action. This combination is useful in health care for supporting disabled patients and old people to have more autonomy. This research aims to assess the combination of the BCI and IoT, specially focusing on health care to facilitate everyday life activities by developing simple home appliances such as door opening and turning on the lights. In the present study, three voluntary candidates participated the experiment of controlling a light based on BCI and IoT systems. The average accuracy results obtained 87.33% and 73.33% with ANOVA $P < 0.05$ for offline and real-time processing modes. Our results indicates that the speed of IoT is fast enough for a BCI real-time system for the case under study.**

***Index Terms*—Brain-Computer Interface, EEG Signal Processing, IoT**

I. Introduction

The Internet of Things (IoT) is a network, in which Things are connected to each other and the internet and communicate with each other and share data. The IoT is not the network of computer only but it is also the network of devices of different types and sizes such as home appliances, toys, cameras, vehicles, medical instruments, and industrial system [1]. The key characteristics of the IoT are interconnectivity, things related services, heterogeneity, dynamic change and safety [2]. It is expected that the pervasiveness of the IoT devices will lead to several positive impacts in healthcare. People will be able to interact and control a large number of objects through different interactions including, for example, software running on their smartphones, wearable devices and voice recognition devices [3].

To efficiently utilize the IoT in the field of healthcare, Brain-Computer Interface (BCI) has emerged and is working very effectively for communication between IoT and individuals. The BCI in the IoT is the communication link between the brain and IoT devices, it converts the brain signals to commands and sends it to the IoT devices. Then, the concerned IoT devices recognize the command sent by the BCI and performed the required action [1], [4]. The BCI is a technology that can help disabled patients by a wide variety of devastating neuromuscular disorders and improve functions in healthy individuals [5]. Recently, there has been a great research interest in how BCI and IoT can effectively communicates. BCI allows to create maps between brain activities and actions. This can be the basis for a translation of human thinking capabilities into physical actions by the help of IoT-enabled machines such as wheelchairs, bionic hand and etc. The cyber-physical system built upon BCI-IoT has great benefits. For example, the BCI communication is in most cases secure as the brain activities are invisible and cannot be copied or hacked [6]. Besides, the BCI can convert the thinking idea about a deliberate action related to a IoT-controlled device into a command in a real-time interaction [7].

In our previous BCI experiences, a bionic hand and a remote vehicle are controlled through thinking. In order to have better accuracy results, different algorithms have been implemented such as wavelet, Common Spatial Pattern (CSP), optimization algorithms and chaotic approaches [8]–[13]. In the present research, we extend the previous studies by combining the benefits of the BCI and IoT to switch on and off a light bulb as an example of an home appliance.

The rest of the paper is presented as follows: Section II contains related studies; Section III explains the components of the BCI; Section IV contains the components of the IoT; Section V explains the proposed research process and discussion; Section VI contains results; Section VII contains discussion, and section VIII contains conclusions and future work.

II. Related Studies

According to the literature, BCI is a system that analyses brain neuron's activities in the central nervous system (CNS) that can also change them into an artificial output. The CNS is responsible of responding to the events in the environment or inside the body and then produces an output [5]. According to the research, the first BCI presentation occurred in 1960 to control a slide projector using electroencephalogram (EEG) signals. Later in 1970, another experiment has been conducted by scientists to control the movement of the computer cursor using the human eyes [14]. In 1988, the BCI paradigm known

as "P300-speller" was developed, which was used for the spelling letters based on Event-Related Potentials (ERPs), which are the EEG deflections in the response of an activity or event [15].

The BCI field is growing at fast speed over the past few years and especially the European Commission is providing outstanding funds for the improvements in the field of the BCI and EEG sensors. In USA, Nicolelis' group performs research where rats and monkeys can control robotic arms using the neural signals recorded from their motor cortex neurons while the electrodes are implanted in their brains [16]. In 1999, the BCI research groups were trying to organize themselves and the first international BCI meeting was conducted in the USA. About 50 participants from 22 groups participated in that meeting [17]. The number of participants and the ground numbers increased very rapidly and in the 6th international BCI meeting, there were about 400 participants and 188 research groups that were participating in the conference [18]. The EEG signal processing algorithms are improved and the scope of the BCI also expands [19] and many new BCI applications with free software [20] are explored for stroke rehabilitation, gaming, mobile BCI [21], [22] applications.

One primitive achieved successful developed algorithms were the CSP spatial filtering algorithm [23] and machine learning methods such as Support Vector Machine (SVM) and neural classifiers [24], [25]. In the previous years, the developed algorithms and discovered patterns in the EEG lead researches into a deeper understanding of brain functionality such as visual and auditory evoked potentials patterns based on the BCI was discovered [26], Steady-State Visually Evoked Potential (SSVEP) [27], Event Related Desynchronization (ERD) and Event Related Synchronization (ERS) [9]. The ERD patterns appears in the EEG, when human has an intention of movement and the ERS appears, when the intention of movement turns to action. In order to improve the detection results, hybrid BCI technique is also developed, which is a combination of different biometrics techniques and different type of biosignals such as using functional magnetic resonance imaging, near-infrared spectroscopy, SSVEP and ERD in one algorithm [28].

Our aim in the BCI research area is rehabilitation for patients with movement disabilities. The BCI can be used with the combination of other systems, in particular IoT-enabled applications. Recently, the BCI is used with the IoT so that the brain signals are converted into commands using the BCI techniques and then received by the IoT application (through mobile device etc.). Finally, the IoT application send the commands to the aim smart objects for the required actions [29]–[32].

III. Components of BCI

The main purpose of the BCI based on the EEG is to collect the brain signals that contain human intentions and converts signals to the hardware commands that can fulfill the human intent. To convert the input signals into the output, the BCI algorithms consist of following steps [5], [17]:

1) **Signal Acquisition and Preprocessing:** In the signal acquisition, the EEG signals are amplified, transferred and recorded using the Enobio32 amplifier that the topography of the channel locations are presented in Figure 1. Then, in the preprocessing step following steps are applied: normalizing data between zero and one, segmentation and filtering are applied to remove the noise and reach to the frequency range of 8-16 Hz. The ERD/ERS patterns are observable in the frequency range of 8-16 Hz. Computing power of the ERD, gives information about the location of neuron's activation during the task as shown in Figure 2. The preprocessed signals are then ready for post processing.
2) **Feature Extraction and Selection:** In the feature extraction process, the segmented and filtered signals are analyzed to separate pertinent signal characteristics from extraneous content and represent them in a suitable way for translation into output commands. These features should have a strong correlation with the user intent because most of the brain activities are either transient or oscillatory. Sometimes the features are computed based on a signal, which is highly affected by noise. Feature selection helps to remove the useless features from the feature space. In this study, a Filter Bank CSP using a DSLVQ (FBCSP-DSLVQ) weighting algorithm is employed to compute features. In this research the Kernel Linear Discriminant Analysis (KLDA) is utilized as the feature selection approach [11].
3) **Feature Translation:** The computed features from the feature extraction process are passed through the feature translation algorithm, which converts those features into suitable commands for the output devices. The feature translation algorithm changes the independent variables (features) to dependent variables (output device commands). If the feature translation algorithm is dynamic, it accommodates spontaneous changes in the signal features and provides better results. In this study, a combination of soft-margin SVM classifier [33] with the generalized radial basis function (SSVM-GRBF) [34], [35] in is employed for categorizing the features [9].

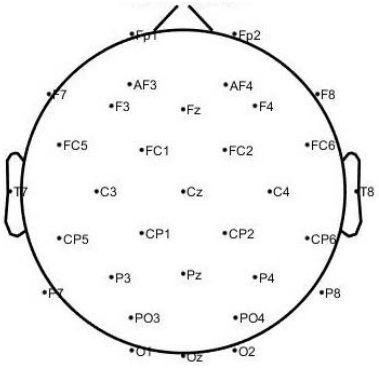


Fig. 1. The location of the EEG sensors on the scalp map for the Enobio32 device.

4) **Device Output:** The device output is the commands from the feature translation to the concerned external device (in this research the external device is the IoT application). The external device provides feedback to the user and is the end of the control loop.
5) **Operating Protocol:** To guide the operations, each BCI has some sort of algorithm that defines how the system is turned on and off, how the system communicates (continuous or discontinuous). The algorithm manages the transmission of the message is executed by the user or by the system. The algorithm also defines the feedback provided to the user from the system.

IV. Components of IoT

To understand the functionality and significance of the IoT, it is essential to Six IoT building blocks work together and provide functionality [23] [36]:

1) **Identification Block:** The identification method is used to identify devices in the network. Devices are identified with the Object ID, which is the name of the device, and the object address, which provides the address of the device in the communication network [17]. The main addressing methods of IoT objects are IPv6 and IPv4 [23].
2) **Sensing Block:** Sensors are used for collecting the data of objects in the communication network and sending the collected data to the destination database or the cloud. The data collected is analyzed in the cloud. Actuators, i.e. hardware mechanical devices such as switches, are also used in IoT platforms and operate in the opposite way of a sensor [18], [23], [26].
3) **Communication Block:** This block contains many heterogeneous objects that exchange data and various services with each other and with the IoT platform. The communication block contains IoT communication protocols like MQTT and CoAP to connect to the objects that are connected in the IoT and to send data to the management system. The sensors and other devices are connected to the Internet by communication technologies like ZigBee, NFC, UWB, Wi-Fi, SigFox, and BLE [15], [23].
4) **Computation Block:** The computation block consists of two parts, hardware, and software. Many hardware platforms have been built to run IoT applications, for example, Intel Galileo, Raspberry PI, Gadgeteer, UDOO, and Arduino. Similarly, many software platforms are used to perform the functionalities of IoT. The main software platform is the operating system that runs throughout almost the whole activation time of the device. The cloud platform is also a computational component of the IoT; it enables small objects to send data to the cloud, it facilitates big data processing in real-time and helps the end-user to obtain knowledge extracted from the big data [15], [23].
5) **Services Block:** The IoT services aid IoT application developers by providing a starting point for development. When developers know the services available, they mainly focus on building the application rather than designing the service and architecture for supporting the IoT application. The IoT services are divided into four categories. Identity-related services can be divided into two categories, active and passive. Services that broadcast information and have a constant power or take power from the battery are active identity-related services. Active identity-related services can transmit or send information to another device. Passive identity-related services have no power source and need some external device or mechanism to transmit its identity. Passive identity-related services can only read information from devices. Information aggregation services refer to the actions of collecting data from sensors, processing that data, and transferring it to the IoT application for processing. Collaborative aware services use the data provided by the information aggregation services to make decisions and react accordingly. Ubiquitous services provide collaborative aware services anytime to anyone who needs it anywhere [18], [20], [28].
6) **Semantic Block:** IoT provides different services, for which it needs knowledge, and to get that knowledge smartly, IoT uses different machines. Knowledge extraction can include finding and using resources, modeling information, and recognizing and analyzing data to reach some decision and provide the correct service. So, it can be justifiably claimed that the semantic block is the brain of the IoT [15], [18], [23].

V. Experimental Setup

In the experiment three candidates with the average age of 30 years old participated. The procedure of the experiment was explained to them before the test and promised not to use the private information of participants anywhere. In the experiment, the same task in [33], [37] is designed and employed. In the task, a picture of hand fisting is displayed and asked the subjects to imagine it by open eyes. The identified patterns are then employed for controlling a lamp for turning on and off. The results are presented in the next part.

VI. Results

Candidates attempted to control a lamp light based on real fisting movement. Figure 2 is the spectrum of a subject's scalp before and after hand fisting for one subject based on the ERD computations. The spectrum figures are very useful for finding the source of patterns. Figure 3 is one of the computed ERDs which is generated based on the hand fisting. The obtained ERD pattern is then employed for trained a model and generating the commands for controlling a lamp application. For the training a model, the average of the obtained results was 87.33% with ANOVA $P < 0.05$. By using the trained model for real-time processing, the average results achieved at 73.33% with ANOVA $P < 0.05$. The Utilized software was Matlab 2019a.

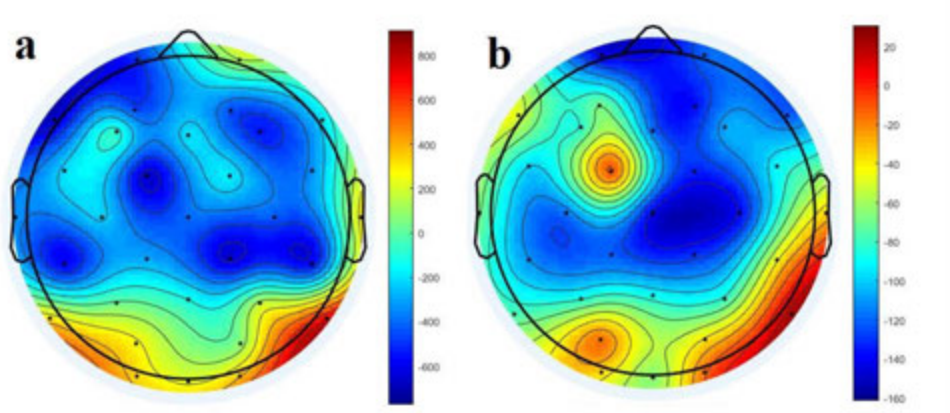


Fig. 2. The power of ERD is computed in the scalp spectrum map. a. The ERD of one participant before applying the task. b. The ERD of the same participant after applying the task.

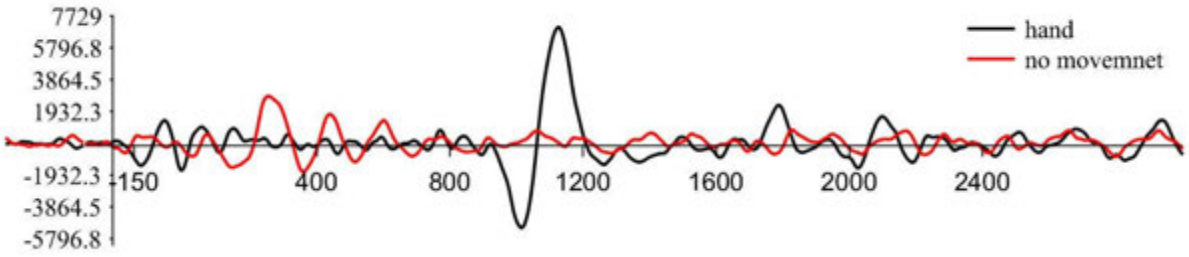


Fig. 3. A sample of identified ERD movement from the C3 channel for one subject. Vertical axes is the amplitude in microvolt and the horizontal axes is time in msec.

VII. Discussion

In this study, the research question rises as follows: is it possible to develop a personalized brain interface that is integrated into the existing IoT infrastructure? To answer the question, Figure 3 is implemented. The architecture of this system is built around a combination of the EEG device, personal computer, motion capture device named as Leap Motion Controller (LMC) [38], which is a small wearable device. The leap motion controller (LMC) is used for the detection of grayscale images of the fingers. The LMC device consist of two cameras and three infrared LEDs for the detection of infrared light. The two cameras are used for creating the grayscale stereo images using the data of infrared light. In the experiment, The device use algorithms to calculate the raw sensors data from hands. Skeletal data of the hand is captured by the LMC device. The motion tracking of this device is about 80 cm. [39].

The BCI algorithm has two steps of offline (training a model) and real-time processing. In order to train a model, the EEG signals were recorded while the same task in [9] was presented. At the same time, the LMC device capture the hand motion to synchronize the hand motion with the EEG. In the experiment 150 trials were recorded for hand fisting state and the moment of hand fisting, the EEG data is marked to find the exact moments of the ERD. To identify which area of the brain is more activated, the power of the obtained ERD before (Figure 2 (a)) and after hand fisting (Figure 2 (b)) is calculated and depicted as scalp spectrum in Figure 2. Figure 2 (b) shows that the area around the location C3 is the source of the generated ERD signals using right hand movements, which is aligned with the theories. One of the obtained ERD based on the C3 location is depicted in Figure 3. The ERD specification is a peak which is started about 850 ms to 1000 ms, and the ERS is then observable immediately from 1000 ms to 1250 ms. The pattern in Figure 3 shows the intention of the subject for movement (ERD) and turning the intention to action of fisting (ERS). Therefore, the aim is designing a model for finding patterns such as Figure 3 automatically. After collecting the EEG signals from three subjects, the FBCSP-DSLVQ features are extracted and the best features are then selected using the KLDA algorithm [11]. The selected features are then classified using the SSVM-GRBF and the results are presented based on the accuracy and validated by the ANOVA statistical analysis. For the real-time experiment, the SSVM-GRBF models are then trained for individual participants and then employed for the real-time mode with 20 trials.

One existing problem in the existing BCI systems is the delay (in some cases, up to 1 second), and wired interface.

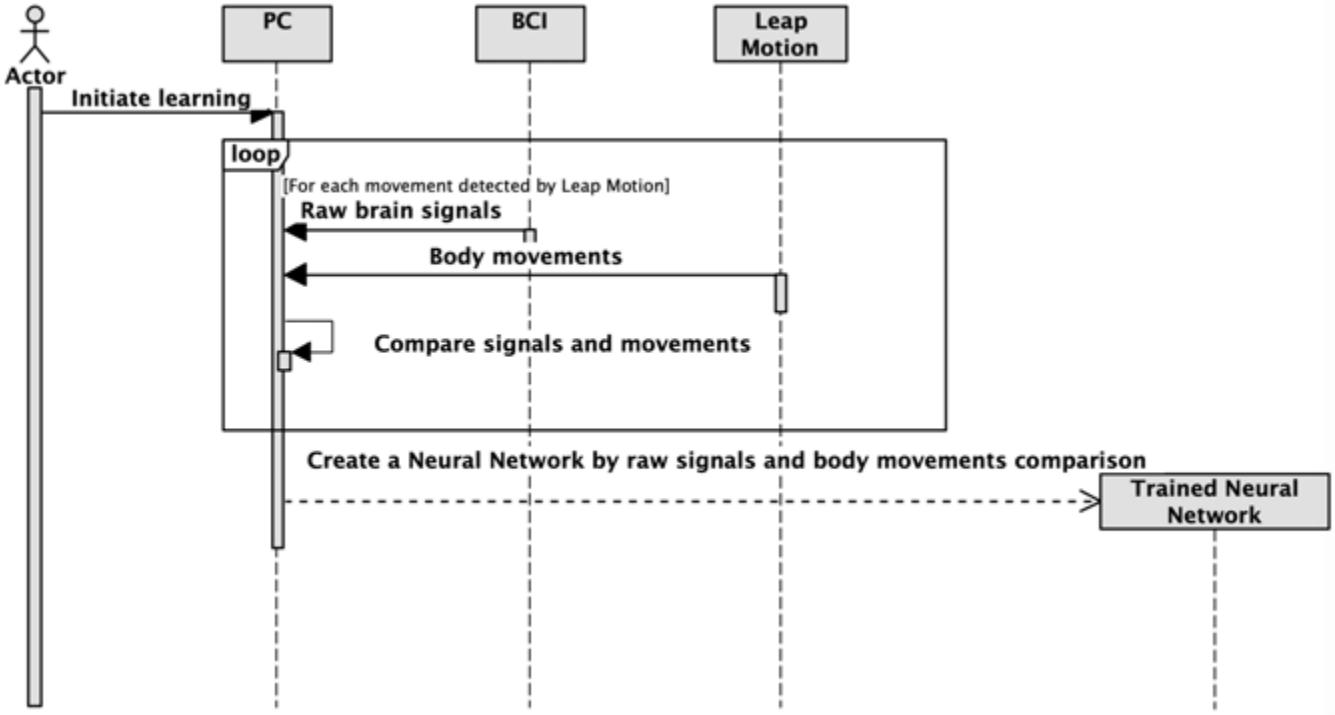


Fig. 4. BCI learning process

Both options are not suitable for wearable IoT-connected devices. To solve above-mentioned problem, a two-step integration process in Figure 4 is proposed. The first step is the learning stage, which are marked with the purple color in Figure 3. The advantages of a house appliances applications such as turn on and off a light in compare to control of a vehicle or bionic hand are only two classes are required to be identified. The SVM-based classifiers are the best choice for two classes identification that achieved high accuracy with low risk of mistakes. Moreover, the mentioned house appliance applications has lower risk of user damages in a case of error in compare with control a bionic hand/leg and vehicle, also they do not expose the users in perilous situation.

VIII. Conclusion and future work

Combination of the BCI and IoT has future capability of more convenient life for disabled patients. In the experiment, three subjects participated to control a light as a home appliances based on hand fisting movement. The algorithm for detecting the hand fisting ERD pattern was a combination of the FBCSP with the DSLVQ weighting approach with the SSVM classifier with the GRBF kernel. The average results of the offline and real-time experiments were 87.33% and 73.33% with ANOVA $P < 0.05$, respectively. Based on the obtained significant results and low level of risk in a case of error, it is concluded to develop the work by employing more participants and different home appliances in the experiments.

This research could be extended by applying the combination of the BCI and IoT techniques for providing facilities such as door opening, or even controlling electric appliances.

Acknowledgements

This paper is supported by Academy of Finland via: (a) ee-IoT project n.319009, (b) FIREMAN consortium CHIST-ERA/n.326270, and (c) EnergyNet Research Fellowship n.321265/n.328869.

References

[1] D. Castro and J. Misra, "the Internet of Things," *Center for Data Innovation*, p. 34, 2013.

[2] O. Vermesan and P. Friess, "Internet of Things Applications - From Research and Innovation to Market Deployment," *River Publishers*, p. 372, 2014.

[3] K. K. Patel and S. M. Patel, "Internet of Things-IOT: definition, characteristics, architecture, enabling technologies, application & future challenges," *International Journal of Engineering Science and Computing*, vol. 6, no. 5, pp. 6122–6131, 2016.

[4] Y. C. Lin and K. G. Marra, *Neural engineering*, 2012.

[5] B. He, S. Gao, H. Yuan, and J. R. Wolpaw, *Braincomputer interfaces*, 1st ed. Elsevier B.V., 2013, vol. 110.

[6] B. Nguyen, D. Nguyen, W. Ma, and D. Tran, "Investigating the possibility of applying EEG lossy compression to EEG-based user authentication," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2017-May, pp. 79–85, 2017.

[7] T. Nakamura, V. Goverdovsky, and D. P. Mandic, "In-ear EEG biometrics for feasible and readily collectable real-world person authentication," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 3, pp. 648–661, 2018.

[8] A. Hekmatmanesh, R. M. Asl, H. Wu, and H. Handroos, "Eeg control of a bionic hand with imagination based on chaotic approximation of largest lyapunov exponent: A single trial bci application study," *IEEE Access*, vol. 7, pp. 105 041–105 053, 2019.

[9] A. Hekmatmanesh, H. Wu, A. Motie-Nasrabadi, M. Li, and H. Handroos, "Combination of discrete wavelet packet transform with detrended fluctuation analysis using customized mother wavelet with the aim of an imagery-motor control interface for an exoskeleton," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 30 503–30 522, 2019.

[10] A. Hekmatmanesh, H. Wu, M. Li, A. M. Nasrabadi, and H. Handroos, "Optimized mother wavelet in a combination of wavelet packet with detrended fluctuation analysis for controlling a remote vehicle with imagery movement: A brain computer interface study," in *New Trends in Medical and Service Robotics*. Springer, 2019, pp. 186–195.

[11] A. Hekmatmanesh, F. Jamaloo, H. Wu, H. Handroos, and A. Kilpeläinen, "Common spatial pattern combined with kernel linear discriminate and generalized radial basis function for motor imagery-based brain computer interface applications," in *AIP Conference Proceedings*, vol. 1956, no. 1. AIP Publishing LLC, 2018, p. 020003.

[12] A. Hekmatmanesh, R. M. Asl, H. Handroos, and H. Wu, "Optimizing largest lyapunov exponent utilizing an intelligent water drop algorithm: A brain computer interface study," in *2019 5th International Conference on Event-Based Control, Communication, and Signal Processing (EBCCSP)*. IEEE, 2019, pp. 1–5.

[13] A. Hekmatmanesh *et al.*, "Sleep spindle detection and prediction using a mixture of time series and chaotic features," *Advances in Electrical and Electronic Engineering*, vol. 15, no. 3, pp. 435–447, 2017.

[14] T. O. W. a. R. D. D. Brain-computer and J. J. Vidal, "Tow ard direct brain-computer 9027 communication," *Annual review of biophysics and bioengineering*, vol. 2, pp. 157–80, 1973.

[15] F. L.A. and D. E., "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalography and Clinical Neurophysiology*, vol. 70, no. 6, pp. 510–523, 1988.

[16] N. M. A. L., "Actions from thoughts," *Nature*, vol. 409, no. January, pp. 403–407, 2001.

[17] J. R. Wolpaw and N. Birbaumer, "Brain computer interfaces for communication and control," vol. 113, pp. 602–614, 2002.

[18] J. E. Huggins *et al.*, "Workshops of the Sixth International BrainComputer Interface Meeting: braincomputer interfaces past, present, and future," *Brain-Computer Interfaces*, vol. 4, no. 1-2, pp. 3–36, 2017.

[19] S. Makeig, C. Kothe, T. Mullen, N. Bigdely-Shamlo, Zhilin Zhang, and K. Kreutz-Delgado, "Evolving Signal Processing for Brain–Computer Interfaces," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1567–1584, 2012.

[20] C. Brunner, N. Birbaumer, B. Blankertz, C. Guger, A. Kübler, D. Mattia, J. d. R. Millán, F. Miralles, A. Nijholt, E. Opisso, N. Ramsey, P. Salomon, and G. R. Müller-Putz, "BNCI Horizon 2020: towards a roadmap for the BCI community," *Brain-Computer Interfaces*, vol. 2, no. 1, pp. 1–10, 2015.

[21] Kai Keng Ang and Cuntai Guan, "Brain-computer Interface for Neurorehabilitation of Upper Limb After Stroke," *Proceedings of the IEEE*, vol. 103, no. 6, 2015.

[22] A. Nijholt, B. Reuderink, and D. O. Bos, "Turning shortcomings into challenges: Brain-computer interfaces for games," *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, vol. 9 LNICST, no. 2, pp. 153–168, 2009.

[23] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441–446, 2000.

[24] B. Blankertz, G. Curio, and K.-R. Müller, "Classifying single trial EEG: Towards brain computer interfacing," *Advances in Neural Information Processing Systems conference*, vol. 1, no. c, pp. 157–164, 2002.

[25] del R. Millán J., M. J., F. M., C. F., V. M., H. J., and B. F., "A local neural classifier for the recognition of EEG patterns associated to mental tasks," *IEEE Transactions on Neural Networks*, vol. 13, no. 3, pp. 678–686, 2002.

[26] S. Gao, Y. Wang, and X. Gao, "Visual and Auditory Brain Computer Interfaces," vol. 61, no. 5, pp. 1436–1447, 2014.

[27] R. Singla, "Ssvep-based bcis," *Evolving BCI Therapy: Engaging Brain State Dynamics*, p. 91, 2018.

[28] G. Pfurtscheller, "The hybrid BCI," *Frontiers in Neuroscience*, vol. 4, no. April, pp. 1–11, 2010.

[29] A. Teles *et al.*, "Using Brain-Computer Interface and Internet of Things to Improve Healthcare for Wheelchair Users," no. c, pp. 92–94, 2017.

[30] B. Jagadish, M. P. R. S. Kiran, and P. Rajalakshmi, "A Novel System Architecture for Brain Controlled IoT Enabled Environments," 2017.

[31] X. Zhang *et al.*, "Converting Your Thoughts to Texts: Enabling Brain Typing via Deep Feature Learning of EEG Signals," *2018 IEEE International Conference on Pervasive Computing and Communications, PerCom 2018*, no. March, 2018.

[32] ——, "Internet of Things Meets Brain-Computer Interface: A Unified Deep Learning Framework for Enabling Human-Thing Cognitive Interactivity," *IEEE Internet of Things Journal*, vol. 14, no. 8, pp. 1–8, 2018.

[33] A. Hekmatmanesh, H. Wu, F. Jamaloo, M. Li, and H. Handroos, "A combination of csp-based method with soft margin svm classifier and generalized rbf kernel for imagery-based brain computer interface applications," *Multimedia Tools and Applications*, pp. 1–29, 2020.

[34] A. Hekmatmanesh, S. M. R. Noori, and M. Mikaili, "Sleep spindle detection using modified extreme learning machine generalized radial basis function method," in *2014 22nd Iranian Conference on Electrical Engineering (ICEE)*. IEEE, 2014, pp. 1898–1902.

[35] S. M. R. Noori, A. Hekmatmanesh, M. Mikaeili, and K. Sadeghniiat-Haghighi, "K-complex identification in sleep eeg using melm-grbf classifier," in *2014 21th Iranian conference on biomedical engineering (ICBME)*. IEEE, 2014, pp. 119–123.

[36] M. Ullah and K. Smolander, "Highlighting the Key Factors of an IoT Platform," *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 901–906, 2019.

[37] A. Hekmatmanesh, "Investigation of eeg signal processing for rehabilitation robot control," 2019.

[38] UltraLeap, "Leap Motion Controller," 2020. [Online]. Available: https://www.ultraleap.com/product/leap-motion-controller/

[39] T. Guzsvinecz, V. Szucs, and C. Sik-Lanyi, "Suitability of the kinect sensor and leap motion controller-A literature review," *Sensors (Switzerland)*, vol. 19, no. 5, 2019.

The success factors of a national healthcare ecosystems maturation: preliminary results

M. Trkman[1,2], S. Bajrić[1] and R. Malkoč[3]
[1]*Jožef Stefan Institute, Laboratory for Open Systems and Networks, Ljubljana, Slovenia*
[2]*University of Ljubljana, School of Economics and Business, Ljubljana, Slovenia*
[3]*Better, Ljubljana, Slovenia*
{marina, samed}@e5.ijs.si, roko.malkoc@better.care

***Abstract*—A healthcare ecosystem is an information system at the level of a disease, a hospital, a city or a nation, consisting of an open platform and applications from different vendors that use the data in the platform. The ecosystem is able to orchestrate workflow between applications to make healthcare processes more effective and efficient. The ecosystem is gradually evolving as applications from different vendors can share the common data repository in the platform, as well as to collaborate and exchange patient data. With a marketplace of applications, the ecosystem can provide competition between applications and enable their seamless substitution. The use of a national healthcare ecosystem is still rare in practice today. Once the open platform is implemented its ecosystem needs time to evolve and demonstrate the expected benefits. Little is known about what a national healthcare ecosystem needs to mature. We present a case study of the Slovenian healthcare ecosystem, and discuss our preliminary list of success factors for the maturation of a successful healthcare ecosystem. Our findings underline the importance of the ecosystem's community and its strong role in the growth of a national ecosystem.**

***Index Terms*—open platform, ecosystem, backbone, patient.**

I. Introduction

Ecosystems are organizational networks [3]. In the field of healthcare, they can support the healthcare systems of specific diseases, hospitals, cities or even states. The core of an ecosystem is an open platform. The main purpose of an open platform is to break the traditional data lock-in between different application vendors by enabling different vendor applications to manipulate shared patient data repositories [25]. The primary role of an ecosystem is to support market functions for ecosystem services [4]. Ecosystems facilitate collaboration among the vendors and also force them to compete for quality, value, and service.

Once the open platform is implemented, its ecosystem needs time to evolve [16], [19]. New applications need to integrate into the existing ecosystem and start sharing common patient data. Many European countries have made efforts to implement an open platform and build a healthcare ecosystem around it. The Estonian ecosystem, for example, covers the whole country in a uniform way and virtually records the medical history of all citizens from birth to death [6]. Countries like Austria [15] and Sweden [23] also offer several e-Health services. It is well known how a healthcare ecosystem should be technically supported, but there are very few case studies of what helps an ecosystem to evolve.

To answer this research question, we have evaluated the following aspects of an ecosystem, namely, technical, social and contextual aspects. Different components of the ecosystem develop independently of each other. The focus of our work is to understand the level of maturity of the Slovenian healthcare ecosystem, and to present initial success factors for the maturation of a national healthcare ecosystem.

The rest of the paper is structured as follows. Section II contains basic information concerning healthcare ecosystems supported by open platforms. A case study of the Slovenian healthcare ecosystem is presented in Section III. The factors of successful ecosystem's use are further discussed in Section IV. Concluding remarks are given in Section V.

II. Background

The vision of an open platform is to create an open ecosystem that will drive competition at the application, service and platform levels [1]. A key role of an open platform is to provide a data repository of patient data in an open, shareable format so that the data is not locked into different vendor applications. The platform allows hospital/region/nation healthcare system to select a unique set of applications from different vendors, and "plug them in" the platform, which then provides services to orchestrate workflow between applications for more effective and efficient decision support.

An open platform provides an infrastructure and services that are fully based on open published standards. These open standards are critical to achieving the central goal of an open platform, namely, data portability and application substitutability. In general, there are two groups of standards: technical and content-based standards.

The most important technical standard for an open platform is the Application Programming Interface (API). An open platform appears to applications as nothing more than a set of open APIs that allow them to access the platform. An API document that conforms to the OpenAPI specification is implemented as a JSON[1] or YAML[2] document [22]. JSON

This work is supported by the research project "Technology and business aspects of the future e-health ecosystem" founded by the Ministry of Education, Science and Sport of the Republic of Slovenia.

[1]JSON is an acronym for "JavaScript Object Notation".
[2]YAML is a recursive acronym for "YAML Ain't Markup Language".

is the current choice of API technology for an open platform. Actions in API documents typically involve querying, reading, and writing patient data. The next important technical standard is FHIR , which is used to enable the automatic transfer of patient data from one healthcare application to another.

Content-based standards have the role of metadata and define the way in which patient data is uniquely represented in the platform. The standards help to describe how a specific clinical concept (e.g. a blood pressure) should be formatted, structured and coded. The definition of clinical concepts and models is a task for clinicians, who must agree on the overall meaning and the associated parameters. The clinical model platform repository consists of reference models (data representation and sharing) referenced by archetypes (theme-based models of content), referenced by templates (business-event specific data set). When the content-based standards are ready, different applications can manipulate patient data using open APIs. Metadata disclosure and sharing facilitates interoperability, improves quality and reduces duplication of effort [20]. Dictionaries also play an important role in defining clinical content. Ideally, a platform should provide terminology services that support global standard terminology, locally defined terminology, and also mechanisms to support mappings between the two of them. SNOMED[3] and dm+d[4] are examples of two very well-known dictionaries.

In addition to the standards, the ecosystem is consisted of real patient data as well. There are two types of patient data: structured and unstructured. Structured patient data can be stored using a standard called openEHR[5] . OpenEHR is the only open standard currently available for the representation of fine-grained structured clinical content that is sufficiently mature and has proven itself on a scale. The standard has been adopted in the Norwegian hospital sector and as a national standard in India, Slovenia and Brazil [1]. Recently, openEHR has also been adopted in the United Kingdom [8]. On the other hand, the IHE-XDS[6] standard can be used for unstructured data. IHE-XDS is an open standard that provides a mechanism for sharing documents and images along with relevant metadata in a health and care environment. It is primarily used for documents and images.

In summary, the open platform supports the storage of structured and unstructured patient data using open standards and dictionaries (e.g. SNOMED). Such open platforms provide support for the development of the healthcare ecosystem. A healthcare ecosystem is vendor and technology-neutral and is designed to eliminate data lock-in in different vendor applications. It facilitates innovation and forces vendors to compete on quality, value, and service [1], [2]. It enables applications from multiple vendors to work together for different users (e.g. clinicians, patients, careers, social workers, commissioners, managers, care workers, etc.) so that there is many-to-many substitutability between applications.

Ecosystems in healthcare can have different focuses, so there are different models of ecosystems. The first model focuses on geographical coverage; examples are hospital, regional and national healthcare ecosystems. The second model focuses on specific clinical needs, such is specific disease management (see example [18]). There is little literature on healthcare ecosystems for each of these focuses. In our paper, we address this gap with our case study of the national healthcare ecosystem established in Slovenia by discussing the success factors of its maturation.

III. Case study

We provide a case study of the Slovenian healthcare ecosystem, which presents a concrete model implementation: *local healthcare communities provide a platform for their population*. We conducted an interview with the head of Better Platform Pre-Sales, who gave us valuable insights into the current state of the Slovenian ecosystem. The questionnaire of our one-hour interview can be found in the Appendix. Further, we used information from Episode 40 in the podcast Faces of Digital Health.

The Slovenian national programme for improving the efficiency and effectiveness of healthcare data is called eZdravje. The programme is 85% financed by the European Union [13]. Since 2008, the programme has also funded 13 software development projects [12] to build the ecosystems open platform - the information structure of the ecosystem. The aim of the open platform was to support the integration of different healthcare applications supported by different vendors. Several vendors worked together on the open platform project. Different application vendors have provided different technical parts of the open platform. The consortium consisted of the following companies: Marand (now Better), SRC-Infonet, List, Nova Vizija, ComTrade, Pinna, Audax. The project was sponsored by the Ministry of Health of the Republic of Slovenia and was completed in February 2013. Since 2015 it is managed by the National Institute of Public Health (NIJZ), while the technical maintainer is Better [5].

The most important project was called Central Registry of Patient Data (CRPP). There were two phases of its development. First, the CRPP was deployed as an IHE based, cross-enterprise enabling document level sharing, having the ability to mobilize documents already produced by legacy systems. Second, the backbone was upgraded with a clinical data repository to store structured EHR records at the national level based on openEHR. This added the ability to share core clinical information such as basic data, referrals, prescriptions, drug lists, allergies and reactions, vaccination records, and other related data to the backbone.

Today, the Slovenian open platform fulfils the minimum core services of the open platform as defined by Apperta [1] in the following way:

[3]SNOMED is an acronym for "Systematized Nomenclature of Medicine".

[4]dm+d is an acronym for "Ddictionary for Medicines and Devices.

[5]openEHR is an open standard specification in health informatics that describes the management and storage, retrieval and exchange of health data in electronic health records

[6]XDS stands for Cross-Enterprise Document Sharing, while IHE is a non-profit organization based in America. It is an acronym for Integrating the Healthcare Enterprise (IHE) in the US.

- *Authentication service is executed in the Security platform* [26]. TThe Security platform is an application for managing user accounts with authorization and authentication, and for managing the permissions of applications. The application was developed by the company ComTrade and is now managed by the Ministry of the Public Administration, Republic of Slovenia.
- *Registry of Patients and Geodetic Units (RPPE)* is the central patient registry and is, the data source for the Patient Summary. It is also used to identify a patient. The registry combines data from three sources: CRPP, Registry of Geodetic Units, and Database of Health Insurance Institute of Slovenia (ZZZS). Therefore, the RPPE is a strong service that allows users to link the EMŠO (Personal Identification Number) with the tax identification number and the KZZ (Health Insurance Card) number.
- *Clinical data repository:*
  Patient data is stored in structured and unstructured way. For structured data openEHR is used, for unstructured IHE-XDS. An example of a document with structured data is patient summary, while an example of a document with unstructured data is a discharge letter (saved as PDF). The dictionary (terminology) server of the platform is HealthTerm [17]. The dictionary was developed by the vendor CareCom. The context of patient data in the platform repository is defined by Clinical Knowledge Manager (UKZ) [14], which is consisted of a set of clinical and demographic models.
- *Service Dictionary:* The Ministry of Health has a website eZDRAV, where information such as public APIs, organisational procedures, or instructions for developers to deploy new applications via the backbone should be exchanged. However, by this date (February 2020), the list of public APIs that provide access to the backbone is missing or could not be found. We have also not been able to find any organizational processes for: How can a vendor certify its application to show that it is compatible with the platform and can be offered on the platform market? Where is the market? How can it be proven that the application does not have security or privacy issues?
- The Slovenian open platform also has one *Electronic Registry of Vaccinated Persons (eRCO)*, where it stores surveillance and monitoring of the vaccination as well as monitoring of adverse reactions [11]. The registry was developed by the company ERC and is managed by the NIJZ.

When the open platform of the ecosystem was prepared, different vendors provided different healthcare applications (see [10]). Today, the platform supports approximately 2600 healthcare providers, including hospitals, health centers and dental studios. There are 50 million records (20 million unstructured and 30 million structured records) in the backbone. Ninety-five percent of the Slovenian population has at least one data set stored in the backbone. Applications that manipulate the patient data of the Slovenian population are collected on the portal zVem [27]:

- **ePrescription** [21] is a service that connects doctors and pharmacists. The advantage for the resident is that he/she does not need to have a physical prescription list to receive the medication, but only the Health Insurance Card (KZZ). So, doctors make prescriptions for a patient in their application. The prescriptions are stored in the backbone. Pharmacies use their application to retrieve data about the prescriptions and to issue the drugs to the patient. This collaboration is possible because the data of doctors and pharmacists is stored in the backbone in a structured way. The service was developed by the company Infonet and is currently managed by the NIJZ.

- **eAppointment** [7] is a set of web pages developed by the company IN2 and is now managed by the Ministry of Health of the Republic of Slovenia. The first advantage for the residents is that they can view the current waiting list for specialists at various locations in Slovenia. Waiting list is an application that allows patients to be informed about the expected waiting time for an appointment with a particular specialist and to make a medical appointment with a specialist, while eRefferal is an application for making a medical appointment with a selected specialist. Appointment documents are stored in a structured form in the backbone.

- **Documents** [9] is an application for accessing patient data. The advantage for the resident is that he/she can view all healthcare documents relating to him/her and his/her children. Patient summary is defined by the Ministry of Health of the Republic of Slovenia and includes twelve sections such as allergies, vaccinations, surgical procedures, diagnosis, pregnancy, medical implants, disabilities, social history, prescribed medication, etc.

IV. Discussion

Apperta has defined what it takes to implement an open platform. However, an ecosystem is more than an open platform. To get a better idea of what the success factors of an ecosystems maturation are, we have adopted the list mandatory requirements for an open platform defined by Aperta, and expanded it with insights from our following analysis of the problems of the Slovenian healthcare ecosystem. Our results are summarized in Figure 1.

There are some *core services of the open platform* that still need to be implemented in the platform defined by Apperta. The maturity level of the core services of the open platform provided by the Slovenian healthcare backbone varies. For example, the maturity level of the service dictionary is very low. Common public APIs are one of the main prerequisites for an ecosystem that could not be found for the Slovenian open platform. The Ministry of Health has a website eZDRAV, which should provide (among other things) information for developers. The service dictionary is missing because the existing APIs are not implemented in the sense of open APIs. Existing APIs are not made publicly available because they



Fig. 1. Success factors of a national healthcare ecosystem.

only allow for simple authentication (username and password). The authentication and authorization of end users with tokens should be in the domain of APIs and not of the applications. As proposed by Apperta, the HL7 FHIR standard should be used to enable proper interoperability of different vendor systems. Currently, it is not supported by the Slovenian backbone. For the development of mobile applications of the ecosystem, it is essential to have OAuth 2.0 authentication, which it is not currently supported by the open platform. There is no demo environment for developers to test new applications. Technically, it is possible for a new vendor to offer a new application, but in practice, it is very complicated to add it to the ecosystem, not only because the APIs are not public, but also because the system for positioning the new application in the ecosystem is missing - *the ecosystem's organizational processes* are not defined. For example, a process of accreditation and certification of new applications that include certain security standards is missing. eZdravje services are protected by a secure healthcare network called zNET. A system for managing information security in the Slovenian healthcare is called SUVI [24].

In Slovenia, there is a community working on defining the missing processes for positioning a new application in the backbone. In 2014, several successful health-technology companies from Slovenia started a *community of vendors* Healtday.si. The goal of the community was to help regional start-ups and established companies to become globally active and to take a new look at the Slovenian healthcare ecosystem. The community started with the vision to connect Slovenian IT stakeholders. Today it connects insurers, the pharmaceutical industry and different governmental bodies. It also connects 40-50 IT companies of different sizes and 200-300 people attending the meetings. The community helps to set up a funding mechanism, define organizational processes, introduce pilot projects, link different players, invite foreign vendors, organize workshops, etc. Not so long ago, there was a special workshop that dealt with complaints that in Slovenia, there are not enough digital solutions embedded in a healthcare system. The community has set itself a new goal: 5 new solutions should be connected to the backbone by the end of 2021. By implementing this goal, the community is discovering what is needed to connect a new application to the national backbone. The resulting insights into what is good and what is bad in this process will enable the next community members to connect their applications to the backbone more quickly.

The ecosystem should be *managed centrally*, which is currently not the case: solutions and projects are managed by the NIJZ and the Ministry of Health. As a result, information is scattered across several webpages and sub-webpages with an overwhelming structure of implemented projects that provided different types of services (e.g. application, documentation) to different beneficiaries. Grouping the results by beneficiary groups would be beneficial for residents, professionals, developers and auditors. A centrally managed ecosystem would be able to provide a service dictionary.

The next problem is the existence of *user's resistance*, which should be managed centrally. Over time, the relationship between negative and positive users of healthcare applications (such as doctors, nurses, pharmacists, insurance company) is changing in favour of digitalization. It takes time for doctors to adopt a new way of accessing patient's data using computers. In the past, patients transported their documents. The new way of retrieving documents is more troublesome for physicians, because there are authorization and authentication requirements that must be fulfilled. In Slovenia, there is a law which states that patient data must be stored electronically. However, no one monitors compliance with the law, and consequently no penalties are imposed for violations.

V. Conclusion

Realizing the vision of an open platform is not about creating an instance of a platform, but rather creating an ecosystem in which a number of application vendors compete for business. In order to maintain the growth of the Slovenian healthcare ecosystem, the support of the Ministry of Health is needed in such a way that clinicians are encouraged to use it, and not leave any alternatives. In addition, a public set of organisational process models is needed to enable seamless growth of the ecosystem. Community Healthday.si helps define these models. From a technical point of view, the APIs should be published and also structured in such a way that their reusability by external applications is promoted. The adoption of an information system should be handled with care, and the adoption of ecosystem's services is no different. We believe that centralized management of all ecosystem's services (for all stakeholders) is missing, which is one of the reasons why we miss the service dictionary. We believe that the existence of a proper service dictionary is a step in the ecosystem's development – as a result of the proper formalization of an ecosystem for the external world.

REFERENCES

[1] Aperta Foundation, "Defining an open platform," 2018.
[2] P. P. Barros, W. B. F. Brouwer, S. Thomson, and M. Varkevisser, "Competition among health care providers: helpful or harmful?", European Journal of Health Economics, vol. 17, pp. 229–233, 2016.
[3] M. Benedickt, "Modelling ecosystems in information systems-a typology approach", Multicourse Business Information Systems, Luneburg, Germany, 2018.
[4] M. Beštek, and P. Eklund, "Is national eHealth in Slovenia on track to be an open eHealth platform?", European Journal for Biomedical Informatics, vol. 15(1), 2019.
[5] Better platform. https://platform.better.care/
[6] Connected Health. https://connectedhealth.ee
[7] Čakalne dobe. https://cakalnedobe.ezdrav.si/
[8] Digital Health. https://www.digitalhealth.net/2020/06/e-prescribing-go-live-at-somerset-nhs-ft-goes-better-than-expected
[9] Dokumenti. https://zvem.ezdrav.si/ogled-dokumentov
[10] V. Dolnicar, "ITHACA case study no. 6: Slovenia", Ljubljana, 2019.
[11] eRCO. https://www.nijz.si/sl/elektronski-register-cepljenih-oseb-in-\nezelenih-ucinkov-po-cepljenju-erco
[12] Ezdravje. http://www.ezdrav.si/category/projekti/
[13] Ezdravje. http://www.ezdrav.si/ezdravje/
[14] Ezdravje. http://ukz.ezdrav.si/ckm/
[15] Gesundheit Osterreich. https://goeg.at
[16] A. Ghazawneh, and O. Henfridsson, "Micro-strategizing in platform ecosystems: a multiple case study", International Conference on Information Systems, Shangai, China, 2011.
[17] HealthTerm. https://www.carecom.com/
[18] HiGHmed consortium. https://blog.better.care/highmed-consortium-covid19
[19] E. J. S. Hovenga, and C. Lowe, "Digital health ecosystems: Use of informatics, connectivity and system interoperability". Measuring Capacity to Care Using Nursing Data, pp. 263-307, Academic Press, 2020.
[20] D. G. Katehakis, J. Goncalves, M. Massi, and S. Bittins, "Interoperability infrastructure services to enable secure, cross-border, operational eHealth services in Europe", in Proceedings of the 17th International HL7 Interoperability Conference IHIC 2017, Greece, pp. 40–51, 2017.
[21] NIJZ. https://www.nijz.si/sl/erecept-0
[22] S. Schwichtenberg, C. Gerth, and G. Engels, "From Open API to Semantic Specifications and Code Adapters", in Proceedings of the 24th International Conference on Web Services, IEEE, USA, pp. 484-491, 2017.
[23] N. Sellberg, and J. Eltes, "The Swedish Patient Portal and Its Relation to the National Reference Architecture and the Overall eHealth Infrastructure", Information Infrastructures within European Health Care: Working with the Installed Base, Springer, pp. 2225–244, 2017.
[24] SUVI. http://www.ezdrav.si/category/projekti/suvi/
[25] A. Tiwana, Platform ecosystems aligning architecture, governance, and strategy. Morgan Kaufmann, 2014.
[26] Varnostna shema. https://vs.gov.si/VS.web/
[27] zVem. https://zvem.ezdrav.si/domov

APPENDIX - INTERVIEW QUESTIONS

1) What do you mean when you say that the Slovenian healthcare system has a backbone?
2) Which Slovenian healthcare organizations are now connected to the backbone?
3) Which user profiles can use it and for what? Functionalities of the backbone (eRefferals, eAppointment, eReceipt).
4) How mature is the trust in e-documents in Slovenia? Do still some ambulants need to have patient's data on paper and on the computer?
5) Who is responsible for managing/maintaining the backbone?
6) How can a backbone support ecosystem for eHealth? Who should be the players?
7) Communities that support ecosystem in Slovenia?
8) What is a web page where developers can contribute applications to our ecosystem? Are APIs published somewhere? Why not, are there difficulties?
9) Which standards are supported by ecosystem's platform (HL7 FHIR, SNOMED CT, openEHR and IHE-XDS)?
10) What are the services: authentication services, master patient index, service dictionary, and clinical data repository? Other services?

Interface Digital Twins: Rendering Physical Devices Accessible to People who are Blind

N. Caporusso*, N. Elleman*, S. S. Cho*
* Department of Computer Science, Northern Kentucky University, Highland Heights, United States
caporusson1@nku.edu

Abstract— **In the recent decades, accessibility received increasing attention. As a result, nowadays, established principles (e.g., Universal Design) contribute to creating products that take into consideration the diverse needs and abilities of all individuals. Unfortunately, there is a gap between the digital and physical world: while websites and user interfaces are adopting accessibility principles, most devices do not support an equitable use. Specifically, individuals who are blind are not able to independently learn and operate several categories of physical products. In this paper, we propose a system for rendering the User Interface (UI) of physical devices and products accessible to individuals with disabilities, with specific regard to people who are blind. To this end, we introduce the concept of Interface Digital Twin (IDT), that is, a digital replica of the physical UI that leverages the accessibility features of mobile devices (e.g., vibration and text-to-speech) to render the interface accessible, help individuals understand its components and learn how to operate it, and support users in accomplishing their tasks on the actual device. The advantage of the proposed solution is that it can be integrated in current technology without requiring any modification. We detail the architecture of the system and we discuss its advantages and use cases.**

Keywords- accessibility; universal design; user interface, Assistive Technology.

I. INTRODUCTION

Improving product usability and, particularly, reducing access barriers is crucial for serving an increasing number of users with diverse needs and abilities, such as the elderly and people with cognitive disorders or physical impairments. Supporting individuals who are blind is especially important, because their population is growing at a fast pace: according to recent statistics [1], it is expected to double and reach 2 million people by 2060, in the United States alone. This, in turn, urges for solutions that can be adopted immediately.

Indeed, the introduction of Universal Access principles resulted in improvements in several aspects of Information Technology (IT), over the last decade, and led to a more equitable use of devices and products. As an example, increased accessibility support in smartphones and tablets, the introduction of new features in web browsers (e.g., magnification), new HyperText Markup Language (HTML) components, such as semantic tags and attributes for Accessible Rich Internet Applications (ARIA), and enhancements to Text-to-Speech (TTS) engines and screen readers make it easier for people who have severe vision conditions to navigate sophisticated web pages and applications. Also, improving IT literacy has been the focus of initiatives dedicated to rendering people with disabilities independent in accessing and using technology. As a result, nowadays a large population of individuals who are blind can independently control most applications and websites on their smartphone or desktop device. Unfortunately, there still exists a large divide between the digital and physical world in terms of equitable access: unless they are considered public accommodations under the Americans with Disabilities Act or they are targeted to specific demographics, products are not required to demonstrate any accessibility features. As a result, their interface (e.g., buttons, knobs, display) is the result of a trade-off between form, functionality, and cost, which often makes it difficult to incorporate universal design principles. Moreover, despite several guidelines have been published and enforced (e.g., Title III), regulatory gaps leave space to non-compliance.

Although there are no general statistics about the level of universal access of products, single studies on specific types of devices revealed that most of them lack accessibility to individuals with disabilities. For instance, recent research on vending machines [2] demonstrated that it is impractical for people with sensory impairments to complete a purchase on specific types of eVending. The study analyzed product selection, payment, and collection in two common models of vending machines and found that poor User Interface design, lack of tactile information, and absence of non-visual feedback prevent the majority of customers with severe visual impairments from being able to independently select, pay, and collect a product. Similarly, most physical devices, from home appliances to technology gadgets, suffer from poor accessibility, especially to individuals who are blind. Consequently, while learning a device might require time and assistance, depending on the level of accessibility of its

user interface, being able to independently use a product for the first time might be especially difficult, and, in some cases, impossible. Contributing factors include labels and messages that are not translated into tactile or auditory form (e.g., touch screens), components having counterintuitive affordances (e.g., credit card readers) or lacking touch or sound cues (e.g., capacitive keypads), missing support for error, and several other issues that also affect usability for individuals who are sighted. Several solutions have been attempted, and the promising work of ongoing research studies in multiple fields is producing new developments that have the potential to unlock access and simplify interaction. Nevertheless, the complexity of most system introduces further risks and challenges in terms of product feasibility, market penetration, and user adoption.

In this paper, we present the concept of Interface Digital Twin (IDT) and we discuss how it can be applied to create accessible replicas of the User Interface of a physical device that can be navigated with a smartphone using touch and audio. We describe the architecture of IDTs and we detail its integration in different types of existing technology using affordable resources that are already available, widespread, and easy to use.

II. Related Work

In the last decade, developers have increasingly been realizing that accessibility is an important requirement of software and especially web pages. However, prior research [2] highlighted manufacturers' reluctancy to change the design of physical products, despite many lawsuits due to their non-compliance with guidelines and regulations in terms of equitable access, such as the American Disability Act. Unfortunately, this creates a divide between digital and physical technology that ultimately impacts the quality of life of people with disabilities and, specifically, individuals who are blind.

Consequently, several research groups in the field of assistive technology are addressing the lack of accessibility of physical devices: most of the work involves the development of new systems that incorporate haptic devices and leverage sensory substitution and multi-modal feedback to render visual information in tactile or audible form. Examples include smart canes combining computer vision or ultrasound systems and Machine Learning algorithms that provide users with information about the environment and help them in indoor and outdoor navigation tasks, hand-held systems reproduce shapes and text in tactile form, and wearable input and output devices that interact with the user via touch cues and vibration [3].

In the last decade, attention has been to Tangible Interfaces [4], which refer to input and output devices that enable touch-based communication and interaction with applications on a computer, mobile device, or appliance. Indeed, they provide the benefit of translating visual information over the tactile channel, thus, making it potentially more accessible to individuals with vision impairments. Although tangible interfaces, such as the Tactile Shape Displays introduced in [5] can be utilized to represent simple shapes, more sophisticated devices are needed to support the complexity of the layout of a User Interface. Furthermore, as they require the development and production of new technology, they might not be sustainable given the numbers, market structure, and device regulations of assistive technology for people who are blind.

In [6], the authors propose the use of web-based interfaces to support people who are blind in interacting within a smart home environment. Although their solution incorporates accessible HTML tags and attributes that simplify control of basic home automation features, such as switching lights and checking their status, it might not be suitable for more sophisticated user interactions and for creating an accurate representation of the actual interface of a physical device or object.

A one-size-fits-all solution in the form of a Universal User Interface for individuals who are blind is presented by [7]: its architecture is based on a centralized database of ontologies and user models that works in combination with a dedicated mobile application. This, in turn, adapts interfaces depending on users' needs and displays them on the screen. Unfortunately, approaches that involve sophisticated architectures and complex implementation roadmaps, imply long time-to-market. Moreover, requiring end users to install additional software introduces risks in the deployment and adoption phases.

Solutions based on affordable systems that require very simple technology and little modifications to existing devices have been studied by several research groups. [8] suggests the use of digital fabrication tools to render the user interfaces of physical objects (e.g., home appliances) accessible thanks to tactile buttons and Braille labels printed in 3D. Moreover, the authors describe how leveraging crowdsourcing dynamics can enable individuals to create, share, and download models, for free, so that end users themselves can fix issues in home appliances and other devices.

Indeed, the accessibility challenge implies a trade-off between the urgency of the problem and the relevance of the solution, where the former favors low technical complexity to sophisticated technology and the latter is solely driven by the results in terms of user adoption.

III. System design

The objective of our work is to develop an affordable and immediately actionable solution to the lack of accessibility of many products currently on the market. Specifically, our goal is to empower users and organizations focusing on accessibility to implement changes without requiring any intervention from the manufacturer. To this end, we applied a participatory design approach to the development of the system discussed in this paper [2]. Specifically, we involved groups of professionals, students, end users, and organizations that have expertise in disability, and we realized several workshops and studies to define the requirements and specifications of our solution. As an outcome of our preliminary work, users prioritized timely implementation, easy-to-learn technology, and low-cost interventions. Moreover, they urged fixes and patches that

can be implemented by individual organizations and users, which they favored to more complete and powerful solutions that required sophisticated architectures, longer implementation roadmaps, and the involvement of multiple stakeholders.

As a result, the proposed solution consists of a web-based platform that enables users to design and share a digital replica of the user interface of a physical device, so that individuals who are blind can access it and use the accessibility features on their smartphones to navigate it, while receiving additional multi-modal guidance by the system. By doing this, they can build a mental model of the configuration of the physical interface of the actual device and learn how to operate it. For instance, while being in front of a vending machine, they can access the digital replica of its user interface with their mobile device: the text-to-speech features of their smartphone can provide them with a brief description of the main components; also, as they swipe their finger on the screen of their smartphone and browse the interface, tactile feedback and audio can be used to help them locate and read the label of buttons, receive instructions about how navigate the options of the payment system, simulate operations and obtain feedback that helps them learn how to make a selection, complete a purchase, and use the system, before or while they operate the physical device.

To this end, we introduce the concept of Interface Digital Twin, that is, a completely digital replica of a User Interface specifically designed to exactly mimic the form, layout, and workflow of a physical device, product, or object (e.g., a device or a paper document) to render it accessible and help users learn it. In contrast to Tangible Interfaces [4], IDTs aim at supporting multi-modal interaction via tactile icons and audible feedback without requiring a dedicated or external physical device. In addition to reproducing the original UI, the primary purpose of IDTs is to provide an additional information and interaction layer that improves usability and accessibility (e.g., by helping the user locate controls using vibration patterns and by translating visual labels in audible form) and adds user customization. As a result, its layout and features are aimed at supporting alternative multi-modal content that results in a more effective interaction with the associated object or device.

In general, IDTs are designed to be utilized when the user is in close proximity with the physical device or object they are associated with. Specifically, we expect individuals to primarily interact with them within in the Intimate (0 – 50cm), Personal (50cm – 1m), or Social (1 – 4 m) proxemic spaces. For instance, a user can download and access the IDT of the User Interface of a vending machine with their smartphone, by either scanning a tag or selecting the device among the ones available nearby (see Figure 1). To this end, IDTs should be compatible with different types of devices, screen sizes, and operating systems to account for the diverse requirements of diverse users. Simultaneously, standards must be defined to make them conveniently available for download and access, accurate in replicating the layout of the user interface of the physical device or object, and effective in making it accessible; also, they should be easy to design, publish, share, and reuse. Finally, they should enable affordable and versatile implementation with resource-effective systems that are accessible to everyone.

Therefore, the architecture of the proposed solution consists of three main components:

- a set of IDTs implemented based on standards and guidelines that ensure accessibility and functionality requirements are met;
- a system for conveniently retrieving and accessing the IDT of the physical device or object, especially in contexts of Just-In-Time interaction, such as purchasing from Vending Machines;
- a centralized repository facilitating the distribution of IDTs.

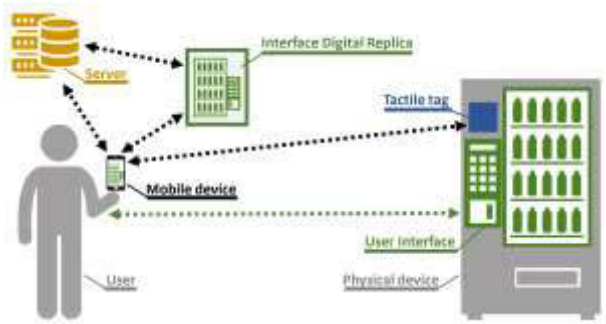


Figure 1. Example of interaction with an Interface Digital Replica: if a physical device having a User Interface that is not inherently accessible has an IDT associated to it, a tactile sign can be placed in a visible location (e.g., on its facade) to indicate the presence of the IDT. By doing this, when in proximity of the physical device, the user can sense and scan the IDT tag (e.g., QR code, NFC chip, or Bluetooth beacon). Consequently, the mobile device retrieves the Interface Digital Replica from the server and displays it on the screen of the smartphone, so that the user can navigate and understand the position and function of the components of the UI in the digital replica and operate them on the actual physical device.

A. Interface design and implementation

Several programming languages can be utilized to implement IDTs. However, based on feedback from our user-centered design processes, our system supports client-side web technologies, such as the HTML, Cascading Style Sheets (CSS), and JavaScript. Nowadays, thanks to HTML5 Application Programming Interfaces, browser integration, and external libraries and components, developers can implement the intangible version of an interface and make it interactive without requiring complex programming skills or sophisticated technical resources. By doing this, users are not required to install any specific application; also, using the web browser as a viewer increases compatibility, reduces implementation risks, provides a debugging environment, and helps troubleshooting potential issues. Modern languages enable the development of responsive web pages that dynamically adjust their content to fit any screen size, which, in turn, addresses device diversity. Existing elements, such as semantic tags and ARIA attributes, can conveniently be utilized to standardize implementation; also, currently available plug-ins for enabling touch-based interaction and gesture control can be used in combination with TTS and vibration features to provide individuals with multi-modal

feedback. Specifically, these features can be utilized for reproducing some aspects of the affordances of the physical object by representing them into visual, audio, or tactile form. For instance, the UI of a vending machine can be replicated by a webpage in which HTML elements are positioned to match the configuration of the physical device (see Figure 2), so that when users move their fingers over the screen, they can sense the presence of buttons and listen to their labels in audio format. Also, in case of small devices or very complex UIs, the user can zoom in a specific part of the Interface.

In its simplest form, the IDT can replicate the layout of the physical UI of the device. In addition, IDTs can host further information about its components, such as how to operate them properly. For instance, previous studies found that individuals who are blind might find it difficult to collect their purchase from vending machines because the product release doors have anti-theft mechanisms that affect their accessibility in case vision is impaired. Similarly, in several devices, buttons support multiple actions depending on different ways to press them (e.g., short press versus long press). Therefore, the IDT could provide users with additional information about their operational details. Moreover, they can implement step-by-step wizards that offer guidance on how to complete the task, as if there was a human attendant assisting the user.

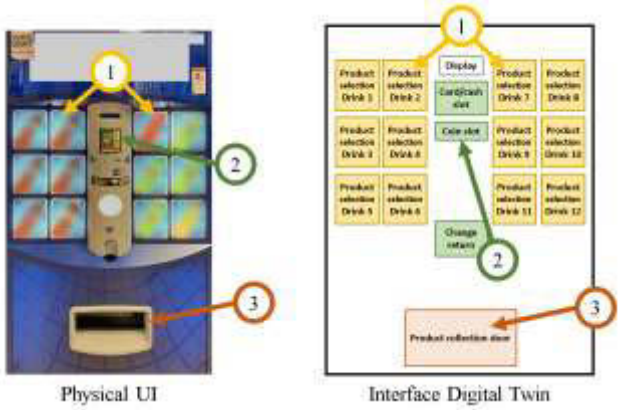


Figure 2. Example of the User Interface of a physical device, i.e., a vending machine (left), and its corresponding IDT (right). In addition to showing the location of the components of the UI, the IDT can organize them into categories that help the user understand their function, such as product selection (1), payment (2), and product collection (3). As users move their fingers on the display, the smartphone can vibrate or emit sounds to highlight the presence of an element, while its label can be displayed in audible form. Furthermore, the IDT can contain additional information on how to operate each component and can offer guidance on how to accomplish the task (i.e., complete a purchase).

In addition, custom UI building blocks and scripts can be designed to introduce more advanced features and make them uniformly available across multiple IDTs. Finally, as the languages utilized by the system inherently respond to Open source practices, developers can easily design new interfaces based on the components of existing IDTs. Specifically, in the case of devices, such as vending machines, in which models having the same physical structure and interface (invariant) are utilized to serve different products with diverse possible placement combinations (variant), the code of an IDT can be reused and modified to match the exact configuration and maintain consistency in terms of product selection.

Although for the purpose of this paper we mostly focus on accessibility aspects and on the client-side, IDTs can have a server-side component that enables additional features, such as retrieving more information about specific components or creating a Content Management System where authorized users can dynamically update the content of the interface (e.g., the labels indicating product names on the UI of a vending machine). Also, IDTs act as separated containers: they do not share any information with one another. However, client-server interaction can be designed to enable IDTs to exchange information with the platform, such as the profile and characteristics of the user, which help customize the aspect and content of the IDT to their specific needs (e.g., allergen warnings).

B. Centralized database of interfaces and guidelines

The architecture of the system consists of a web platform that acts as a centralized repository that enables developers to create and share IDTs that users can download and access on their device.

Specifically, the platform has the purpose of defining the standards and describe the requirements, specifications, and implementation guidelines for the design of IDTs; by doing this, developers can produce IDTs that adhere to a common set of rules and protocols. In addition, the platform can host wikis, discussion boards, and other communication tools that support user-centered and participatory design practices: members who want to contribute to the project can discuss changes to the standards, promote updates and improvements, and share developers' guides and tutorials.

Secondly, the platform acts as a centralized hub for sharing and accessing IDTs. The codebase of IDTs can be hosted in separated repositories on external distributed Version Control Systems (e.g., GitHub, Bitbucket). This enables programmers to fork existing interfaces and reuse and modify their components, which is especially convenient in case of small changes, such as translating labels in multiple languages, or updating the configuration of buttons (e.g., diverse product configurations on different vending machines having the same model). There are several options for publishing the client-side of the IDT: it can be hosted on the developer's server, made it public on a code versioning system (e.g., GitHub pages), or uploaded on the platform itself. The actual webpage of the IDT can be made available via a dedicated URL on the platform. Moreover, each IDT can work as an independent sandbox, which increases security and provides developers with more flexibility. The developer will be responsible for hosting the server-side component of the IDT, if any. Regardless of where the client-side IDT is hosted, its link is associated with a redirect URL on the platform of the proposed system. The URL contains the handle of the developer, the brand of the manufacturer of the physical device or object, the product name, and its model, so that users can quickly search for an interface; in addition to specifying visibility settings,

configuring accessibility criteria, and verifying that repositories meet the requirements, the platform keeps track of new versions and updates. The developer is responsible for any back-end scripts or databases that enable additional features with respect to the interactivity offered by client-side HTML, CSS, and JavaScript. This is to maximize interoperability and to avoid interfering with or over-regulating the many diverse systems utilized by different developers. Also, the platform can enable users to register an account, create their profile, and store their preferences. Also, the user can decide to share their preferences (e.g., presence and type of disability, preferred level of guidance, primary type of feedback) with IDTs, so that their components and interaction can be customized to better support their needs. Furthermore, the web platform can also store a history of devices the user interacted with, to facilitate access to frequently utilized IDTs.

Finally, the platform enables users and manufacturers to quickly associate IDTs with devices by supporting the creation of physical tags that can be placed on the physical object. In this regard, the platform can offer templates for tactile tags and automatically generate Quick Response (QR) codes or Near Field Communication (NFC) chips containing the URL of the IDT, to make it more convenient for users to access and control the interface, as discussed below. Also, it can provide developers with snippets and sample source code that help programmers quickly integrate IDTs in existing applications.

C. *Convenient access to and control of User Interfaces*

Each IDT can be associated with a unique URL that can be utilized to access their navigation. Moreover, the web address can be incorporated in a tag that enables users to easily retrieve the digital replica when they interact with the physical device. For instance, it can be in the form of a QR code or an NFC chip that can be positioned in a visible location of the physical product so that the tag can be conveniently scanned or read with a mobile phone. In order to render the tag easier to recognize to individuals who are blind, its area can be marked with a tactile pattern and it can be accompanied by a sign in Braille.

In addition, a wireless system (e.g., Bluetooth beacon) can be incorporated into the device so that users can discover the machines that are available nearby. This can be realized with dedicated mobile applications. Furthermore, physical products that are installed in a specific location, such as vending machines, can be associated with their geolocation coordinates or with their position in a building, floor or office so that people can use an app or a web browser to access a directory of the compatible devices nearby and visualize their interface via a navigation menu.

IV. Conclusions and Future Work

Despite existing accessibility guidelines require manufacturers to incorporate universal design features that enable individuals with different types of disabilities to use their products, regulatory gaps and the cost of updating existing technology make it very difficult to rapidly implement solutions that are applicable to systems currently on the market. As a result, people with disabilities are unable to independently use several devices, from basic home appliances to sophisticated eVending.

In this paper, we focused on improving the accessibility of physical devices, with specific regard to individuals who are blind. To this end, we presented a novel solution aimed at providing users with a system that can immediately render any interface easily accessible at a negligible cost, without requiring any intervention from the manufacturer. Specifically, we introduced the concept of IDTs, that is, replicas of the actual physical UI that the user can navigate thanks to the accessibility features of their smartphone. By doing this, they can understand the placement and function of the components of the interface, learn how to operate the device and accomplish their tasks. We detailed the architecture of the proposed system and we discussed how IDTs can be applied to any devices and products, thanks to its limited technical complexity.

In our future work, we will evaluate the effectiveness of the solution in rendering physical UIs accessible and supporting individuals with disabilities in independently using them.

References

[1] R. Varma, T.S. Vajaranant, B. Burkemper, S. Wu, M. Torres, C. Hsu, F. Choudhury, and R. McKean-Cowdin, “Visual impairment and blindness in adults in the United States: demographic and geographic variations from 2015 to 2050”. JAMA ophthalmology, 134(7), pp.802-809. 2016.

[2] N. Caporusso, K. Udenze, A. Imaji, Y. Cui, Y. Li, and S. Romeiser, ” Accessibility Evaluation of Automated Vending Machines.” In International Conference on Applied Human Factors and Ergonomics (pp. 47-56). Springer, Cham. 2019.

[3] A. Bhowmick, and S.M. Hazarika, “An insight into assistive technology for the visually impaired and blind people: state-of-the-art and future trends”. Journal on Multimodal User Interfaces, 11(2), pp.149-172. 2017.

[4] M.S. Baldwin, R. Khurana, D. McIsaac, Y. Sun, T. Tran, X. Zhang, J. Fogarty, G.R. Hayes, and J. Mankoff. “Tangible Interfaces”. In Web Accessibility (pp. 715-735). Springer, London. 2019.

[5] A.F. Siu, “Advancing Accessible 3D Design for the Blind and Visually-Impaired via Tactile Shape Displays”. In The Adjunct Publication of the 32nd Annual ACM Symposium on User Interface Software and Technology (pp. 146-149). 2019.

[6] M. Buzzi, B. Leporini, and C. Meattini, “Simple Smart Homes Web Interfaces for Blind People”. In WEBIST (pp. 223-230). 2018.

[7] A. Khan, S. Khusro, and I. Alam, “Blindsense: An accessibility-inclusive universal user interface for blind people”. Engineering, Technology & Applied Science Research, 8(2), pp.2775-2784. 2018.

[8] A. Guo, J. Kim, X.A. Chen, T. Yeh, S.E. Hudson, J. Mankoff, and J.P. Bigham, “Facade: Auto-generating tactile interfaces to appliances”. In Proc of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 5826-5838)

Evaluation of Resource Management System for InfaaS-adaptive Disaster Management Application Platform

Yasuhiro Watashiba*, Yuki Matsui*, and Susumu Date*
* Osaka University, Osaka, Japan
Email: watashiba-y@cmc.osaka-u.ac.jp

***Abstract* - The concept of Information-as-a-Service (InfaaS) is important for collaboration-based disaster management applications. For satisfying this concept, collaboration-based disaster management applications is required to have three functionalities: clarity, synchronization, and continuity for information-providing. Since these functionalities are needed to handle not only application-layer but also infrastructure-layer, it is difficult to develop functionalities on each collaboration-based disaster management application. From this consideration, we aim to realize InfaaS-adaptive disaster management application platform which enables to provide three functionalities. Though the clarity and the synchronization are met by functions of distributed visualization system, the continuity requires the mechanism that can tolerate partial failures of IT infrastructure components caused by the disaster events. Thus, we have been studying and developing Software-Defined IT infrastructure which enables to manage various infrastructure resources by software. Though the previous study achieved the recovery of assigned resources on each IT infrastructure component, a mechanism to handle all components of IT infrastructure is required for autonomous recovery. Therefore, we propose new resource management system which enables to control all components of Software-Defined IT infrastructure comprehensively. In this paper, we introduce the architecture of proposed resource management system and evaluate the behavior, performance, and effectiveness.**

Keywords* - *component; formatting; style; styling; insert (key words)

I. INTRODUCTION

The concept of Information-as-a-Service (InfaaS) is important for collaboration-based disaster management applications, which helps the decision makers in the disaster management. For supporting an appropriate decision making process in the disaster management, the flow of information to authorities and the synchronization of retrieved and analyzed data between different groups such as local governments and Non-Governmental Organizations (NGOs) must be maintained. Additionally, since the disaster management performed by the organizations located at different countries and areas is focused on, it become more important to provide the correct information for all organizations [1]. In such situation, there is a critical need to have a system that makes easy to understand various types of data and synchronizes between different organizations so that all involved receive the same information simultaneously. Furthermore, the data should be provided continuously to decision makers for making the best decisions in disaster progresses. In order to satisfy these requirements, the disaster management applications should be designed and developed based on the InfaaS concept.

For satisfying the InfaaS concept in the disaster management applications, there are three key functionalities: clarity, synchronization, and continuity for the information. Clarity means that the functionality makes it easy to understand the various information generated from the simulations and the analysis results of retrieved large-scale datasets. Synchronization means that the disaster management applications keep to provide the information on multiple sites concurrently. Continuity indicates that the information must be continually provided for the decision makers even when the disaster makes a part of IT infrastructure on which the disaster management applications are running inoperable. In order for a disaster management application to achieve these functionalities, the mechanisms not only on application-layer but also on infrastructure-layer are required. Thus, it is difficult to develop and operate such mechanisms in each disaster management application.

From this consideration, we aim to realize InfaaS-adaptive Disaster Management Application Platform (InfaaS-AP) as the base for the development and deployment of the disaster management applications. In the InfaaS-AP, a mechanism of the distributed visualization system is adopted for achieving the functionalities of clarity and synchronization required for InfaaS concept. Because many disaster management applications adopt the visualization technology as a comprehensible way of providing the various information. Additionally, some of the existing distributed visualization systems have a mechanism to transfer the visualized images to multiple sites concurrently. On the other hand, since continuity requires the IT infrastructure that can adapt to rapidly changing environments caused by the partial failures of resources, we have been studying and developing Software-Defined IT Infrastructure (SDITI). The SDITI is a mechanism to control IT infrastructure by software for offering the needed flexibility and resilience to the disaster management applications on the InfaaS-AP.

In the research so far, though the availability of the components in SDITI, such as computing resources and network among them, is addressed, each component is managed and controlled independently. However, current SDITI does not have a mechanism to comprehensively handle its components for autonomously recovering the service of the disaster management application on the InfaaS-AP for the various situations. In other words, current SDITI statically deals with a failure based on a certain scenario designed by an administrator. For realizing the autonomous service recovery on the InfaaS-AP, we propose a novel resource management system which enables to unify management of the SDITI components.

This paper is structured as follows. Section II briefly introduces the InfaaS-AP and clarify the research purpose. Section III analyses the required functionalities for autonomously recover from the failure of resources, and then explains the design and the implementation of proposed resource management system. In Section IV, we evaluate the behaviors of the proposed resource management system. In Section V, we conclude this paper with the future work.

II. BACKGROUND

A. *Collaboration-based Disaster Management Application*

In the collaboration-based disaster management, the disaster management application has an important role for supporting the cooperation between multiple organizations. Generally, each organization is located in a different site and makes decisions based on retrieved information. Thus, the collaboration-based disaster management application is required to provide same information for all organizations continuously.

InfaaS is an important concept for developing such a collaboration-based disaster management application. InfaaS means that any information is provided for the receivers clearly and persistently. In a disaster management, there are various types of information, such as the data transmitted by Internet of Things (IoT) devices and the result of computer simulations and data analysis. Moreover, the knowledge and skill of the receivers also diverse. For an appropriate decision making, it is essential for decision makers to properly understand the information. Thus, a collaboration-based disaster management application should have functionalities according to the InfaaS concept.

For achieving three key factors of the InfaaS concept in a collaboration-based disaster management application, the management in not only application layer but also IT infrastructure layer is required. This makes difficult to develop the functionalities in each collaboration-based disaster management application. From this consideration, we have been studying and developing InfaaS-adaptive disaster management application platform (InfaaS-AP).

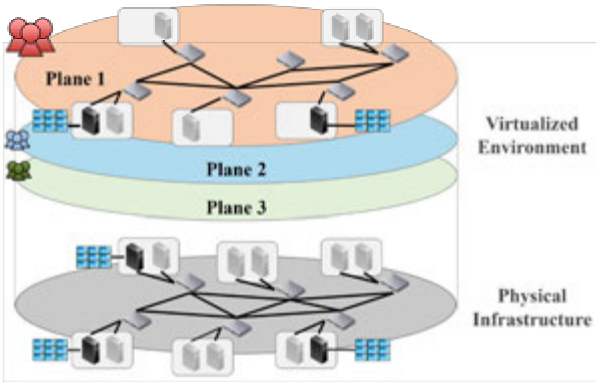


Figure 1. Conceptual design of InfaaS-AP

B. *InfaaS-adaptive Disaster Management Application Platform*

InfaaS-AP is the environment for development and deployment of collaboration-based disaster management applications. InfaaS-AP enables to implement the collaboration-based disaster management applications without any consideration of the functionalities for three key factors of InfaaS concept. The conceptual design of our InfaaS-AP is shown in Fig. 1. In this design, InfaaS-AP creates a virtualized environment to assign to each disaster management application, called plane, from a physical infrastructure and then allocate the plane to a disaster management application. A plane is composed of computing resources, network resources, storage resources, and visualization devices.

As the functionality of clarity, a method to make decision makers with different knowledge understand the information easily and correctively is required. Visualization technology is an effective method due to a fact that it helps to understand information easily and quickly by converting various types of data to image. Additionally, since decision making is generally performed by discussing based on multiple information, large-scale and high-resolution display device, such as tiled display wall (TDW), is desirable. The functionality of synchronization is provided by a mechanism of distributed visualization system. In the distributed visualization system, a server running visualization application and display device are located in different sites and visualization images are transferred via network. Since some middleware of distributed visualization system allows to transfer a visualization image to multiple display device, synchronization of information providing for multiple sites is achieved.

The functionality of continuity requires to dynamically control for IT infrastructure. When a disaster causes partial failures of IT infrastructure allocated to a disaster management application, InfaaS-AP has to rapidly change the environment for maintaining the provision of information. In order to manage IT infrastructure dynamically, we have been studying and developing Software-Defined IT Infrastructure (SDITI) which leverages Software-Defined technologies such as virtualization technology and Software-Defined Networking (SDN) in our research project of Joint

Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures (JHPCN) in Japan [10]. InfaaS-AP is realized by integrating SDITI into distributed visualization system. The components of our prototyped InfaaS-AP are as follows:

1. SAGE2 [5] with TDW
2. Docker [2]
3. OpenFlow [6]

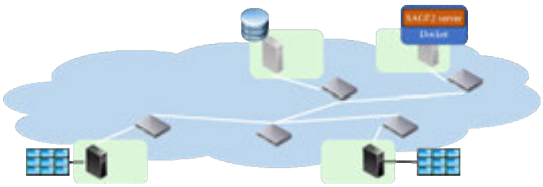
SAGE2 is a middleware of distributed visualization system. There are two types of components running on SAGE2: SAGE2 server which manages virtual display space and SAGE2 display client which is a receiver. SAGE2 allows to simultaneously transfer the contents on the virtual display space to multiple TDW in different sites. Moreover, since SAGE2 can handle TDW, it is suitable to conceptual design of InfaaS-AP. Docker, which is one of container technologies for provisioning light-weight virtual environment, is adopted for running SAGE2 server. In InfaaS-AP, the recovery of SAGE2 component is reboot on other computer due to a fact that the failures caused by a disaster event suddenly happened. Thus, the light-weight operation is adequate for the situation. OpenFlow is an implementation of SDN which enables to dynamically manage network by a controller. Since the network topology allocated to a plane illustrated in Fig. 1 has to be reconstructed according to changing computers running SAGE2 components, the functionalities for dynamically changing the allocated network paths is required. Thus, the feature of OpenFlow is suitable for the functionalities. On our prototype of InfaaS-AP, the developer can implement a disaster management application with the functionalities of InfaaS concept as an application on SAGE2.

In our research so far, the foundation of InfaaS-AP utilizing the three components has been constructed as the prototype. However, the operation of components in current InfaaS-AP is performed individually according to the assumed situation. In other words, the current InfaaS-AP do not have a mechanism to autonomously administer and operate the components of InfaaS-AP for disaster events. To appropriately handle various situations caused by a disaster, the mechanism is essential. In this research, we propose a resource manager of InfaaS-AP to manage and control the behavior of the components comprehensively.

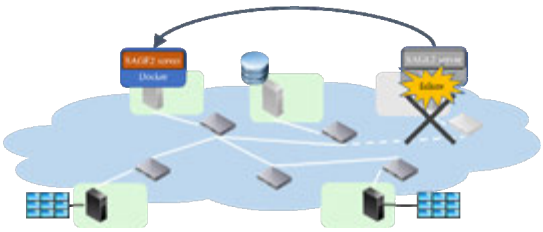
III. RESOURCE MANAGER FOR INFAAS-AP

A. *Analysis of required functionalities*

In order to design proposed resource manager, the required functionalities are analyzed based on a scenario which is assumed a situation before and after happening a failure shown as Fig. 2. In the scenario, the components of a plane, which is a set of resources allocated to a disaster management application by InfaaS-AP, before occurring a failure is illustrated in Fig. 2(a). The plane is composed of a Docker container with SAGE2 server running on a computing node, two TDWs, a network storage, and network paths connecting between them. On the plane, the



(a) A plane before occurring a failure.



(b) A plane after occurrting a failure. When a failure occurred in a site belonging to the plane, the resources assigned to the plane are reconstructed and the application is restarted.

Figure 2. Overview of assumed scenario

disaster management application generates visualized images from data in a network storage, and then transfers them to TDWs via SAGE2.

When a failure has occurred, InfaaS-AP needs to reconfigure the allocated components of the plane for continuity of the disaster management application. In this scenario, it is assumed that a failure occurs in the site where a computing node running the SAGE2 server is located as shown in Fig. 2(b). Firstly, InfaaS-AP detects the failure and then looks for available alternative computing resources for redeploying a Docker container with SAGE2 server from a physical infrastructure. After deciding alternative computing resources, InfaaS-AP removes resources in the damaged site and adds the alternative computing resources. After that, InfaaS-AP reconfigures network paths allocated to the plane due to a fact that the alternative computing resources are basically located in different site. For keeping the connection between the new SAGE2 server and TDWs, the network paths needs to be rerouted. Moreover, by accessing to the same source of information, the SAGE2 server can keep to distribute information generated by the disaster management application even after the redeployment of the Docker container. Finally, the Docker container with SAGE2 server is redeployed on the alternative computing resources and the information flow by the disaster management application is restarted.

For achieving autonomous recovery in InfaaS-AP according to this scenario, the five following functionalities are required.

(1) Collecting the statuses of components
(2) Detecting a system failure
(3) Deciding alternative resources
(4) Reconfiguring allocated network paths
(5) Redeploying Docker container

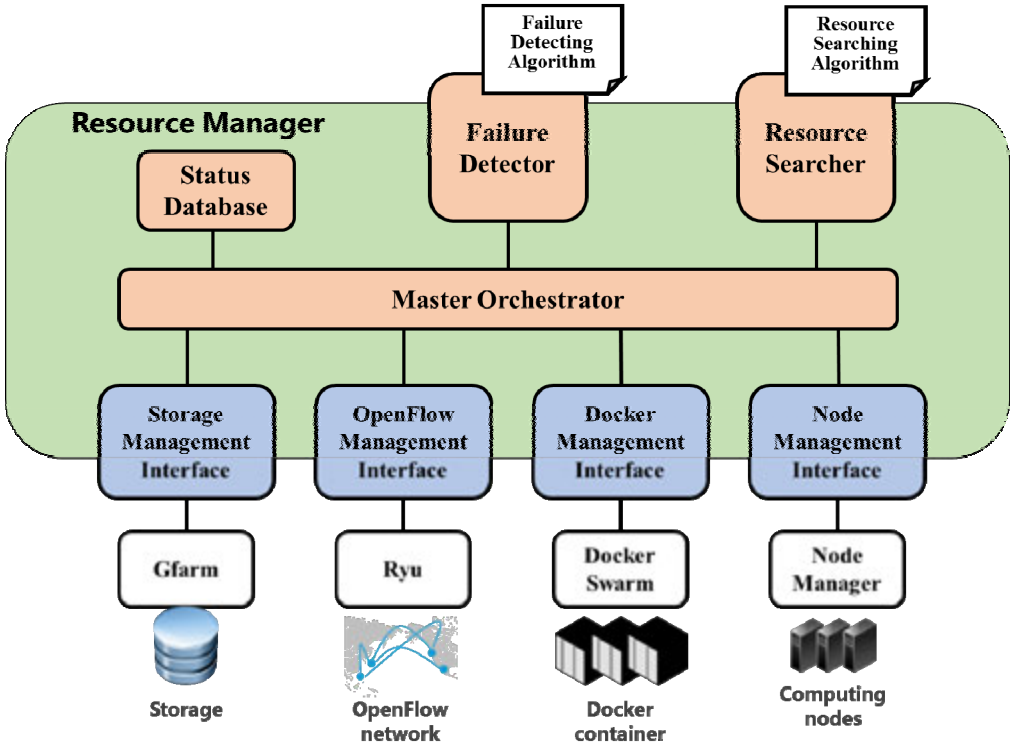


Figure 3. Structure of proposed resource management system

The functionality (1) is to collect the statuses of all components in InfaaS-AP. This is because it is necessary to constantly monitor the statuses of resources for detecting an interruption of disaster management application and searching alternative resources from physical infrastructure. As the examples of statuses collected in InfaaS-AP, the CPU usage rate on each computing node, reachability of the network endpoint and execution status of the Docker containers are raised.

The functionality (2) is required to detect a system failure based on the statuses retrieved by the functionality (1). There are various methods to judge whether a failure has occurred or not. Thus, the functionality (2) should have a flexible mechanism. The functionality (3) is also similar because there are various algorithms to select an alternative resources such as computing nodes and network paths. Additionally, since InfaaS-AP assumes the deployment in wide-area environment, the policies for resource management and security may be different in each site. Thus, it is essential to enable to adopt an appropriate method according to the situation.

The functionality (4) and (5) perform the redeployment of resources based on the result of the functionality (3). The functionality (4) replaces the network paths allocated to the plane. In this functionality, the control based on OpenFlow is required. The functionality (5) operates the management system for Docker and reruns the SAGE2 server.

B. *Resource Manager for InfaaS-AP*

From the consideration in Section III-A, we design and develop our proposed resource management system for InfaaS-AP. The structure of proposed resource management system is shown in Fig. 3. The proposed resource manager is composed of four modules: *Master Orchestrator*, *Failure Detector*, *Resource Searcher*, and *Status Database*. Moreover, since every components of InfaaS-AP have respective independent managers, the proposed resource manager needs to have management interfaces to administer and operate them in an integrated fashion.

Master Orchestrator is a core module of proposed resource manager to manage all functionalities described in Sec. III-A. The role of *Master Orchestrator* is to instruct other modules according to the situation. The functionality (1) is performed through the management interfaces so that the statuses of each component are retrieved by the respective independent managers. Based on analysis of the retrieved statuses, *Master Orchestrator* issues instructions to the other modules and the managers of components. The implementation of *Master Orchestrator* is independently developed by Python. The reason is why there is not an appropriate framework to comprehensively handle all components of InfaaS-AP. Moreover, every components of InfaaS-AP do not have an affinity each other.

Failure Detector and *Resource Searcher* have roles for the functionality (2) and (3) respectively. *Failure Detector* periodically analyzes the retrieved statuses based on an algorithm defined by an administrator. When a failure is detected, this module requests the search of alternative

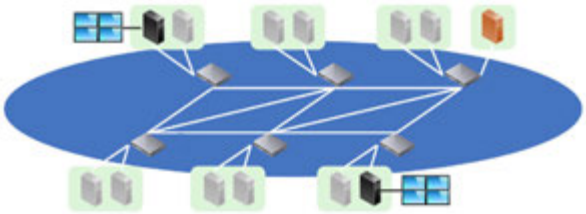


Figure 4. Physical environment for experiments

Table 1. Specification of computing node

CPU	Intel(R) Xeon(R) E5-2620 0 @ 2.00GHz
Memory	DDR3 1600 8GB DIMM x8
NIC	Intel(R) Gigabit Ethernet port (onboard)
OS	Centos 7.3

resources to *Master Orchestrator*. *Resource Searcher* looks for an appropriate alternative resources according to the request of *Master Orchestrator*. Since both functionalities needs to flexibly replace the method, these modules have a plug-in algorithm mechanism. This mechanism allows an administrator of InfaaS-AP to easily change new algorithms according to situations and policies.

Status Database is for storing the statuses retrieved by functionality (1). *Master Orchestrator* updates the information in *Status Database* for every retrieving statuses. The algorithms in *Failure Detecting* and *Resource Searching* need to access *Status Database* for the respective processes. Since these algorithm may also use past statuses, *Status Database* is incorporated into proposed resource manager. In the prototype of proposed resource manager, SQlite [8] is adopted.

In the prototype of proposed resourced manager, the respective managers of InfaaS-AP's components are adopted as follows. Node manager is independently implemented by using Linux commands. The reason is why Linux commands have rich functions and light-weight behavior. Additionally, since there are also the functions to monitor resources, it is suitable for proposed resource manager.

For managing Docker container, Docker Swarm [3] is adopted. Docker Swarm is a clustering tool for Docker. Docker Swarm allows to comprehensively manage computing nodes for deploying Docker containers and provides various functions to control them. This tool make it easy to operate Docker containers in InfaaS-AP.

As the manager for OpenFlow, Ryu [7], which is a controller for OpenFlow, is utilized in the prototype. Since Ryu is implemented by Python, it has a high affinity to *Master Orchestrator*. Furthermore, a functionality to enable external operations through REST-API is suitable for the design of proposed resource manager.

As storage manager, Gfarm [9], which is a wide area virtual file system to provide large-scale storage, is adopted. Since Gfarm has the mechanism for centralized management, it is easy to control a behavior of storage system from *Master Orchestrator*.



Figure 5. Initial structure of plane

Table 2. Result of performance measurement

Action	Time (s)
Failure detection	12.20
Searching alternative resources	0.11
Reconfiguring network	0.31
Redeploying Docker container	16.20

IV. EVALUATION

For evaluating the proposed resource manager, we conducted evaluation experiments. In this section, the experiment environment is introduced and then the experiments for performance of proposed resource manager and effectiveness of plug-in algorithm mechanism is explained.

A. Environment for Experiments

The experiment environment is composed of 13 computing nodes, six OpenFlow available switch and two TDWs as illustrated in Fig. 4. Note that OpenFlow available switches are virtually constructed on a switch (NEC UNIVERGE PF5240). System structure of all computing nodes has same and the specification of computing node is shown in Table. 1. Note that a computing node, called master node, is used for proposed resource manager. The master node accommodates proposed resource manager and all manager of InfaaS-AP's component. Initial structure of plane allocated to a disaster management application is shown in Fig. 5. In the experiments, it is assumed that only this plane is created by InfaaS-AP.

For the experiments, we developed and deployed simple Failure Detecting Algorithm and Resource Searching Algorithm. Failure Detecting Algorithm detects a failure by the health-check function of Docker container provided in Docker Swarm. Resource Searching Algorithm decides alternative computing resources from available computing nodes in random manner and calculate new network paths based on the shortest path manner.

B. Performance of proposed resource manager

In this experiment, we measured the elapsed time of each process for recovery when an intentional failure in the computing node accommodated the Docker container with SAGE2 server was generated. The failure was occurred by stopping the Docker container. The result is

Table 3. Result of file I/O performance

	Random	Bandwidth aware	Storage location aware
Write speed (Mb/s)	0.12	22.77	51.66
Bandwidth (Mb/s)	9.24	184.75	417.72
Write time(s)	195.56	11.54	5.32

shown in Table. 2. Note that the number of experiments is 10 and the values are average of all trials.

From this result, the times for failure detection and redeploying Docker container are long. However, it is caused by the process of Docker Swarm. Thus, the time of failure detection can decrease by adopting other method. On the other hand, the time of searching alternative resources depends on the scale of physical infrastructure. Additionally, the algorithm is a simple method. Thus, in real environment, the elapsed time will increase.

C. *Comparison of resource searching algorithms*

In this experiment, we compared the file I/O performance of alternative resources allocated by three different Resource Searching Algorithms. For this experiment, we additionally developed two Resource Searching Algorithms: bandwidth-aware algorithm and storage-location-aware algorithm. Bandwidth-aware algorithm looks for an alternative computing node based on available bandwidth between computing node and network storage. The status of available bandwidth in each link can be retrieved by a function of OpenFlow. By considering this information, an alternative computing node is decided. Storage-location-aware algorithm focuses on the locality of an alternative computing node and location of data. In other word, this algorithm finds an alternative computing node in the bandwidth-aware manner and additionally migrates data to other computing node in the same site as the alternative computing node. Gfarm allows to migrate the data to a computing node explicitly.

The comparison result is shown in Table 3. The data size transferred from the network storage is 1096MB and the number of files is 487. Note that the migration is performed for all data. The results is average values in ten times of measurement. In the result, write time of bandwidth-aware was 5.90% of random, and storage-location-aware was 2.72%. On the other hand, storage-location-aware algorithm has the overhead time to migrate data from original computing node to new one. The overhead time is depends on the specification of data and available bandwidth between original and new computing nodes. Therefore, to adopt an appropriate algorithm according to the environment and situation is essential for gaining high performance, and the mechanism of plug-in algorithm is effectiveness.

V. CONLUSION

In this paper, we proposed the resource manager of InfaaS-AP for enabling autonomous recovery. In this research, we have been studying and developing SDITI which enables to provide the functionality of information continuity for InfaaS-AP. For achieving autonomous control of components in InfaaS-AP, we designed and developed resource manager to comprehensively manage and control InfaaS-AP's component and provide plug-in algorithm mechanism for easily replacing the algorithm of failure detecting and resource searching. In the evaluation, we conducted the experiments for performance of proposed resource manager and effectiveness of plug-in algorithm mechanism. As the future work, we are plan to evaluate proposed resource manager on PRAGMA-ENT [4] which is a global SDN testbed.

ACKNOWLEDGMENT

This work was partially supported by JSPS KAKENHI Grant Numbers JP16H02802, JP17K00101.

REFERENCES

[1] CENTRA: CENTRA: collaborative projects, http://www.globalcentra.org/projects/, 2019.

[2] Docker Inc.: Docker: Enterprise application container platform, https://www.docker.com, 2019.

[3] Docker Inc.: Swarm mode overview, https://docs.docker.com/engine/swarm/, 2019.

[4] K. Ichikawa, P. U-chupala, C. Huang, C. Nakasan, T. Liu, J. Chang, L. Ku, W. Tsai, J. Haga, H. Yamanaka, E. Kawai, Y. Kido, S. Date, S. Shimojo, P. Papadopoulos, M. Tsugawa, M. Collins, K. Jeong, R. Figueiredo, and J. Fortes, "Pragma-Ent: an International Sdn Testbed for a Cyberinfrastructure in the Pacific Rim," Concurrency And Computation: Practice And Experience, pp.e4138, 2017.

[5] T. Marrinan, J. Aurisano, A. Nishimoto, K. Bharadwaj, V. Mateevitsi, L. Renambot, L. Long, A. Johnson, and J. Leigh, "SAGE2: A New Approach for Data Intensive Collaboration Using Scalable Resolution Shared Displays", 10th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, 2014.

[6] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, \OpenFlow: Enabling Innovation in Campus Networks," SIGCOMM Computer Communication Review, vol.38, no.2, pp.69-74, March 2008.

[7] Ryu project team: Ryu SDN framework, https://osrg.github.io/ryu/, 2019.

[8] SQlite consortium: SQLite home page, https://www.sqlite.org/index.html, 2019.

[9] O. Tatebe, K. Hiraga, N. Soda, "Gfarm Grid File System", New Generation Computing, July 2010, Volume 28, Issue 3, pp.257-275, 2010.

[10] Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures (JHPCN), https://jhpcn-kyoten.itc.u-tokyo.ac.jp/en/, 2019.

IoT Protocol Selection for Smart Grid Applications: Merging Qualitative and Quantitative Metrics

Mehar Ullah*, Syed Rameez Ullah Kakakhel**, Tomi Westerlund**, Annika Wolff*,
Dick Carrillo*, Juha Plosila**, Pedro H. J. Nardelli*
* LUT University, Lappeenranta, Finland
** University of Turku, Finland
Contact: mehar.ullah@lut.fi

***Abstract*—The concept of Internet of Things (IoT) implies that objects are connected to share information with each other and/or with a platform based on standardized communication protocols such as MQTT and HTTP. Recently, their performance has received much attention due to more widespread utilization. But, the vital aspect of the adoption of these IoT protocols is not just performance per se but also their feature set. However, comparative analysis that takes into account both the features and performance aspects are rarely presented. This paper investigates a combination of a qualitative comparison model (based on communication, operational and security attributes) with quantitative metrics in terms of CPU/memory utilization and time consumption. Quantitative metrics are collected for two of the four IoT protocols. Our analysis shows that there is no one-size-fits-all solution in terms of application layer IoT protocols. Additionally we emphasize that, on a higher level, both qualitative and quantitative measures are necessary for IoT stakeholders to consider before selecting the most appropriate protocol for the specific application to be deployed. We illustrate the results using a smart grid example simulating a collection of 1000 smart meters connected to the grid.**

***Index Terms*—IoT protocols, smart grid, MQTT, HTTP,**

I. Introduction

The Internet of Things (IoT) concept was first coined by Keven Ashton in 1999 during a presentation to Proctor and Gamble and later referenced by him in the MIT Auto-ID Center [1]. IoT is one of the fastest growing technologies that is gaining momentum in the various domains like smart grid, transportation, healthcare, industrial automation etc. Nowadays, IoT applications have already moved further away than just simple RFID tags and are incorporating different sources of data and sensors. This stream of data needs to be moved somewhere else to be processed, for example, to train machine learning models and visualization. This place is what we call an IoT platform. For companies to run their specific IoT applications, an IoT platform is thus needed. The IoT platform provides important services and features to applications: endpoint management, connectivity and network management, analysis and processing, data management, application development, security, event processing, monitoring, access control and interfacing [2], [3].

Many devices nowadays use the technologies in the IoT stack for inter-connection and sharing of information. For those devices to connect to various heterogeneous devices they need different types of protocols to overcome the problem of interoperability. The application layer is responsible for the interaction to the end user. With the advancement in technology new devices enter the market that need protocol support to perform well. [4]. With varying needs and diverse set of devices come a diverse set of protocols.

In the current scenario, there are many IoT based technologies which are available to fulfill the needs of smart grid (SG) applications. The IoT technologies are used in SG mainly for long range data transmission bi-directionally between the user side and the utility through IoT based devices like smart meters. The IoT based SG systems requires advanced wireless technologies compared to wired based technology in most of the cases to reduce the deployment and usage complexity. The information flow of data in smart grid can be divided in two channels: one is the flow between all the smart meters connected in a hub with the IoT enabled devices and appliances. The second flow is between the control centers and the utility providers. The data flow with the smart meters and the IoT enabled devices can be achieved by wireless technologies [5].

The question thus arises, which protocol will better suit the needs of these aforementioned channels. To answer that we utilize both qualitative and quantitative measures. The qualitative metrics are presented for four major IoT protocols while quantitative measures are presented for two of them (HTTP and MQTT). We then apply this knowledge to a smart grid model to identify suitable protocols for the diverse grid applications.

The rest of the paper is organized as follows. Section II contains related work. In Section III, we highlighted the four IoT protocols and smart grid attributes. Section IV contains the results, analysis and discussion. Section V concludes the paper and presents future expansion to the work.

II. Related work

In the literature some similar work is done by Niccolo et al. [6] and Edielson et al. [7]. They offer a qualitative comparison between MQTT and CoAP. The comparisons are thorough but expressive and thus not easily extendable. Their comparison is also limited to a subset of the attributes we have collected and identified (nine and seven respectively). Vasileios et al. [8] and Ala et al. [9] provide a detailed description of major application layer IoT protocols. However, they do not

provide criteria for cross comparison. Yuang et al. [10] offers a quantitative comparison of DDS, MQTT and CoAP in the context of eHealth. Paridhika et al. [11] provide a performance analysis in the context of smart parking. Antonio et al. [12], look at semantic and syntactic interoperability in the context of HTTP, CoAP and MQTT. Jens Dede and Anna Forster [13] offer a qualitative analysis of connectivity protocols such as Wi-Fi Direct and BLE for opportunistic communication. Kabeer et al. [14] provide a qualitative comparison of routing protocols.

The aforementioned is a small subset of the literature on IoT protocol comparisons. However, the theme remains the same; the studies focus on quantitative comparisons in varying contexts while offering brief qualitative comparisons. Even if provided, the qualitative assessments are limited to a subset of the protocols, expressive and not easily extendable.

IoT protocol assessments towards smart grids are primarily limited to connectivity protocols such as, WiFi, Zigbee, LoRa, LTE and 5G [15]. Another study [16] highlights an overview of the IoT elements along with architecture layers, and compared the IoT protocols like HTTP, CoAP, MQTT, and XMPP in terms of the features provided by those protocols in smart grid applications i.e. interoperability, scalability, security, latency, and provisioning. This work focuses on both features and performance of application layer protocols.

III. IoT Protocols and Smart Grid Attributes

In this study we have selected four main IoT protocols *MQTT, HTTP, CoAP, XMPP*. Similarly, smart grids also constitute different models and architectures, we present the seven main components relevant to grid communication and distribution.

A. IoT Protocols and their attributes

MQTT (Message Queuing Telemetry Transport) was introduced by IBM in 1999 and is used to collect data from remote devices. It is designed for lightweight devices that are constrained by both processing power and memory and in 2014 it became an organization for the Advancement of Structured Information Standards (OASIS) [17].

HTTP (Hyper Text Transfer Protocol) is the protocol that is used for the exchange of information between internet connected nodes that communicate via hypertext. HTTP is most commonly used application layer protocol on the internet and was first proposed in 1989.

CoAP is REST (Representational State Transfer) styled request-response protocol standardized by the Internet Engineering Task Force (IETF) [10]. CoAP can be used on devices with limited processing power, storage and bandwidth. The main difference compared to HTTP is the lower header size.It is a dual-layered protocol.

XMPP (Extensible Messaging and Presence Protocol) was originally a text-based messaging protocol for messaging between devices over a network using the XML document format, versions on XMPP are being explored for IoT by the XMPP Standards Foundation for control, sensor data and provisioning. Its strong point is its very familiar device@server naming approach. Besides that it provides end-to-end security, provisioning, and publisher-subscriber style communication.

Below is a brief of the classification model presented in [18]. There are five main categories in this section, namely, Communication Attributes, Security Attributes, Connection Attributes, Operational Attributes and Message and Payload support. All of these have sub-categories that further expand on the main characteristics. We provide a brief description of these attributes here, for deeper explanation we refer authors to [18].

1) Communication Attributes: Communication attributes constitute five main features, paradigm or architecture, i.e., is it publish-subscribe based, request-response or does it offer a message queuing architecture. Second, which underlying transport protocol is used (TCP or UDP). Sheng et al. offer a good discussion on the importance of transport protocol [11]. Third, is there multicast support? Often it is required to send messages to multiple recipients at once, a typical scenario for IoT as thousands of devices would be present in a locality. Fourth, reliability/QoS, Does the protocol offer any surety of delivery? Fifth, congestion control, when systems are overloaded and packet loss occurs, does the protocol offer any congestion control, surety or packet re-transmission mechanisms.

2) Connection Attributes/Performance: This explains the establishing and maintaining of connections. Sub categories for connection attributes include four characteristics. First, *communication complexity* in this case refers to the number of steps required to establish/setup a connection and be ready to send actual data packets. Second, *signaling traffic generated* or *frequency of updates* means besides the normal operation, how much extra traffic is generated that is not related to the user transmission, but the functioning of the protocol itself, e.g. ping requests to identify available clients or periodic updates to assess the network etc. Third, *connection establishment speed and performance*, this one deals only with the number of packets required to establish a connection. The measurements are fast, medium and slow. Slow is when a protocol has high communication complexity and requires two or more steps to establish a connection. Fourth, *session orientation* is the ability to handle sessions, this improves performance as a single user session can be used to send many updates.

3) Security Attributes: The IoT will inherit the same security issues as the current internet and will amplify them because of its deep penetration and direct connection to the real world. A major differentiating factor for the choice of the protocol is the security attributes. A point to note here is that we will be talking about the features provided by the protocol (in the protocol's draft documentation) and not the ones provided via third-party add-ons (unless officially recognized) or provided via commercial implementations of the protocol. The three different aspects of security are: First, *connection security* relates to any mechanisms provided by the protocol that enable the authentication of the nodes before any actual communication takes place. Some of the methods can include a secure key exchange between the endpoints and authentication. Second, *communication security* relates to securing communication between two nodes or endpoints.

TABLE I
QUALITATIVE COMPARISON. ADAPTED FROM [18].

		MQTT	HTTP	CoAP	XMPP
Communication Attributes	Communication pattern	Publish/Subscribe	Request/Response	Req./Resp.& Pub./Sub.	Req./Resp. & Pub./Sub.
	Transport protocol	TCP	TCP	UDP	TCP
	Multicast support	Yes	No	Yes	Yes
	Reliability/QoS	Yes	Nil/via TCP	Yes	Nil/via TCP
	Congestion Control	Yes	Relies on TCP	Yes	Relies on TCP
Connection Performance	Communication Complexity	Medium	Mininal	Minimal	High
	Signaling Traffic	Low to high	Low	Low	Low
	Communication Speed	4	3	3	5
	Session Orientation	Yes	Yes	Yes	Yes
Security Attributes	Connection Security	TLS	SSL	DTLS	TLS
	Communication Security	TLS	SSL	DTLS	TLS
	User Security	User/Password	User/Password	Nil	SASL
Operational Attributes	Operations	Centralized	Centralized	Decentralized	Both
	Discovery	No	Yes	Yes	Yes
	Message Durability	Yes	Yes	Some How	Yes
	Caching	Yes	Yes	Yes	Yes
Message/ Payload Support	Message Overhead	Minimal	High	Minimal	High
	Design Orientation	Byte-wise	Payload Specific	Payload Specific	Payload Specific
	Block Transfer	No (Max. msg256MB)	Yes	Yes	Yes

Communication security features would ensure that a third-party interception would result in no sound or identifiable information leakage. Third, *user security* in our context means whether the protocol ensures the validity/authentication of a user utilizing the system to communicate, either to the server or the broker.

4) Operational Attributes: Distributed Operation or Centralized operation refers to whether there is a control of a single node on the operation and communication of the remaining nodes or that the communication is one-to-one where both the nodes are on equal grounds. Some protocols might request that all the data go through a central point (a server or broker) and is then distributed to the intended recipients, in other cases direct messages would be the norm. Hybrid solutions can also exist. Service/node discovery defines if there are ways that the protocol offers via which machines can notice peers or associated group members. Other point is: are messages retained or discarded as soon as they are delivered? In unreliable environments or guaranteed delivery requirements, it is vital to have message retaining, although this comes at the cost of memory.

5) Messages and Payload Support: Knowing which protocol will easily integrate with the current systems is valuable and knowing the payload helps in that regard. Current web services and technologies will benefit from a protocol that offers support for web payloads (i.e., HTML, Plain text, XML). Situations that require machine-to-machine interaction will benefit from JSON or byte-wise transmission. First, Message Overhead, in the shortest form, how many more data bytes are required to make a transmission happen (in headers, checksums) for each packet to be transmitted. Second, Design Orientation, does the protocol define any document format or not? If the protocol defines a specific document format that makes it document-centric and if it does not, that will make the protocol data-centric. Third, Fragmentation/Block Transfer, the capability to divide a large data set into smaller ones and then transfer them automatically without user intervention.

B. Attributes of Smart Grid network

There are several logical networks for smart grid communication, i.e., home are network (HAN), neighborhood area network (NAN) and field area network (FAN). Various applications are expected to emerge with different requirements and features within these networks. In the section we will be discussing a few selected applications that has got significant attention from researcher in smart grid paradigm. The characteristics and requirements of those applications are highlighted below.

1) Automatic meter reading: One of the basic and simplest application of smart grid is automatic meter reading (AMR) and is used to collect meter readings, events, and alarm data from smart meters. The important published standards used in AMR are ANSI C12.1-2008, IEEE 1377 and IEC 61968-9. However, the most used one is IEC 61968-9 more general and covers most of the AMR communication, i.e., meter reading, meter connect and disconnect, meter data management, outage detection, prepaid metering etc. Both IEEE 1377 and ANSI C12.19 standards are used to provide specifications for the communication syntax for data exchange between the end device and the utility server [19].

2) Demand response: Demand response (DR) is used by the utility operators for the optimal balancing of the power generation and consumption by utilizing the implementation of various load control programs or by offering dynamic pricing. Dynamic pricing programs are used to encourage the customers to use less energy during the peak hours. The

primary role of the DR application is to communicate to the customer the price information by sending it to the smart meter in the form of a price signal. After receiving the pricing signal, the customer adjusts the energy usage to the minimum pricing period to reduce the power bill [20].

3) Electric vehicles: Electric vehicles (EV) use rechargeable batteries instead of fossil fuels and can be charged from the distributed feeder. Mostly the EV's are expected to be charged at home, however there is also the possibility to charge the EV's at the public charging stations established in parking areas [21]. Smart charging is one of the good options which enables controlled charging of the EV's using the bidirectional communication capabilities of smart grid. Centralized charging control is the device used for smart charging and allows the energy transfer session in real time. It checks the time taken by the EV's for charging and the energy available.

4) Substation automation: The substation automation mainly concerns with the monitoring, protection, and control functions that are carried out at the substation and feeder equipment. The main protocols used at the substation automation domain are IEC 61850 and DNP3/IEE1815. The DNP3 protocol is used to provide the communication specifications for low-bandwidth monitoring and control operations. IEC 61850 standard is based on interoperable intelligent electric devices (IEDs) that interacts with each other. IEC61850 standard is covering a maximum of the aspects of SAS which includes real-time, high bandwidth protection and control applications [22].

5) DER/Microgrid: Distributed energy resources (DER) are the small power sources that can generate and store power and are connected to the distributed grid. The DER can be a renewable source such as solar panels, wind turbine and battery storage system or non-renewable source such as combustion turbine and fuel cell. The DER can be implemented as a distributed generation source, distributed storage source or a combination of both. A microgrid is a small local electric power system that consist of one or more DER units and loads. In a normal operation, the microgrid is connected to the grid and performs operations in a synchronized mode. During any fault or maintenance event, the microgrid operates autonomously in an island mode and is able to support its own load [23]. Based on the purpose, there are two types of microgrids, utility microgrids that serves parts of the utility and industrial/commercial microgrids that serves customer facilities [24].

6) Wide Area Measurement: Wide area measurement system (WAMS) is an advanced sensing and measurement system that is continuously monitoring the health of the power grid and it obtains the system state and power quality information from the state modules based on phasor measurement units (PMUs). The PMUs utilise GPS to provide the accurate system state measurements in real time and provide time stamp for each measurement.The utility control center can obtain high resolution phase information because of the precise information by the PMUs and enables them to initiate proper response time in seconds and protect the whole wide area network from black out events. In smart grids the PMUs are installed at the distribution domains to monitor the overall power system in real time [25].

7) Distribution supervision: The main duty of the distribution supervision is to increase the visibility of the power distribution network, so that to take the proactive actions to prevent equipment failure and ensure pubic safety. The distribution supervision scope is distributed and include passive infrastructure like transmission lines, cables and branching points. Most of the overhead transmission lines in smart grid are equipped with the adequate sensors and actor nodes for continuous monitoring and preventive measures. There are also sensors used in the underground transmission lines to monitor the thermal condition and corrosion of the buried conductors. Also, some sensors are used in some cases to detect line faults and monitor nearby environment to prevent contact with vegetation [26]. For the weather conditions wireless sensors networks are used because the renewable energy sources are mainly dependent on the weather.

IV. Analysis and Discussion

To augment the qualitative comparison model in Table I, we performed memory, CPU utilization, latency and throughput tests for two of the aforementioned protocols (HTTP and MQTT). The test was executed on a Core i7 machine (4 cores, 8 threads), running Ubuntu 14.06 with 32 GB of RAM. Messages were routed via local-host (to mitigate network effects). HTTP webserver was implemented via mongoose, while MQTT service was provided via the Mosquitto broker. Stress testing for HTTP was conducted via ApacheBench and for MQTT via custom shell scripts utilising tools provided by Eclipse Mosquitto project[1]. To stress test the capability of the system acting as a data concentrator in a grid network we simulated 1000 clients sending a total of up-to one million messages. For clearance, the test failed at 2000 clients with messages exceeding half a million.

A. Results

Figure 1 represents the throughput and figure 2 represents latency of both protocols. The red line in figure 2, across x-axis is HTTP latency as it stayed below 1 ms (1 ms is the mean value across all concurrent requests, mean value per request was 27 ms.) MQTT latency fluctuated from 7 ms to 12 ms depending upon the total message load. The blue lines in figure 1 represents throughput and we can see that MQTT offers 2-3x the throughput of HTTP. HTTP throughput was low even though it offered mean lower latency because of the deviation in results. Under stress approx. 2 percent of the requests took more than 1s to process, highest being 3 s with a standard deviation (sd) of up-to 120 ms. MQTT latency however remained relatively consistent with an sd of 3.2 ms.

Figure 3 represents the time consumed to process one million messages from 1000 clients concurrently and the overall memory consumption. MQTT concluded the test in

[1]The tools used during the experiment are:
https://mosquitto.org/
https://github.com/cesanta/mongoose/
https://httpd.apache.org/docs/2.4/programs/ab.html.

TABLE II
TRAFFIC REQUIREMENTS FOR THE IMPORTANT SMART GRID APPLICATIONS

Application	Reference standards	Traffic characteristics	Suggested IoT Protocol	Discussion
AMR/AMI/DR	ANSI C12.19, IEEE 1377, IEC61968-9 OpenADE, OpenADR	Delay tolerant Mostly periodic/event based Small burst size Multicast/broadcast	MQTT	Consistent latency performance Better suited to telemetry Low message overhead Designed for multicast transmission Low footprint on client side
DER/Microgrid/EV	IEEE 1547.x, IEC 61850-7-420	Delay sensitive Semi-periodic/event based Multicast/broadcast	MQTT	Consistent latency performance Better suited to telemetry Low message overhead Designed for multicast transmission Low footprint on client
Substation automation	IEC 61850, DNP3/IEE1815, IEEE 1646	Extremely Delay sensitive No re-transmission Event based Reliable multicast	MQTT	MQTT QoS 3 assured delivery ensures no re-transmission
Wide area monitoring	IEEE C37.118	Delay sensitive Periodic Limited re-transmission	HTTP	Integration into area monitoring tools Varied payload support Triggers can be programmed in easily
Distribution supervision	IEEE 1451.x	Delay sensitive/tolerant Periodic/event based Random Low power consumption	HTTP	Integration into area monitoring tools Varied payload support Triggers can be programmed in easily HTTP REST API's for advanced integration i.e. Tableau for visualization Microsoft PowerBI for analysis

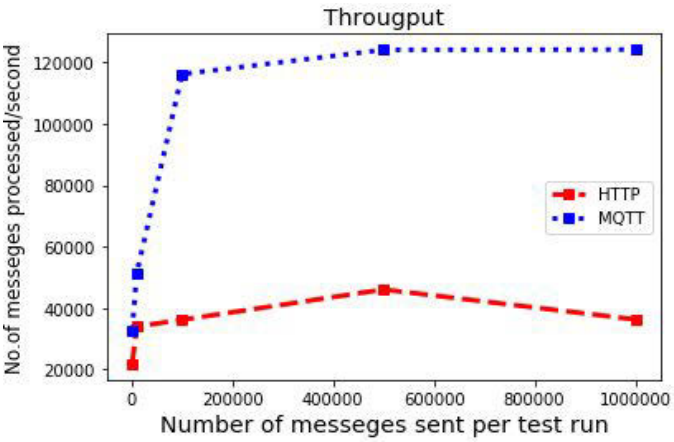


Fig. 1. Throughput

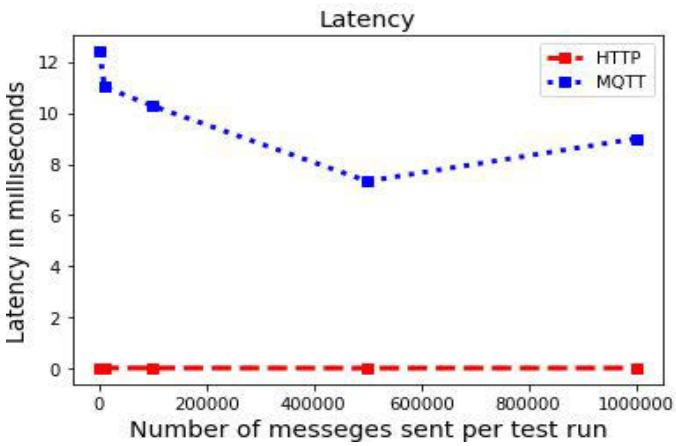


Fig. 2. Latency

approx. 3.5 seconds, consuming 123 MB of RAM. The mongoose web-server performed the tests in approx 13 seconds consuming less than 1MB of RAM. Even though application design choices, purpose, features and loaded libraries will significantly affect its memory footprint, one insight that can be derived is that MQTT memory consumption increases overtime as a result of message caching at the broker. It is a design choice enforced by the protocol (broker has to maintain the list of clients, topics, and messages; Mosquitto does so via an in-memory database). Since keep-alive for HTTP was not used as each connection terminated mongoose clears the buffers resulting in significant total memory reduction. CPU utilization was 100 percent for both protocols and all test runs, hence that graph will not be included.

B. Discussion

The test results show that MQTT provides the best throughput, up to 4 times better than HTTP. While HTTP provides ten to twenty times better latency for the same load although with a lot of disparity. MQTT broker module is less processor intensive, about four times as efficient as compared to HTTP. Reliability is much better with MQTT as well because of its intrinsic quality of service levels. While HTTP gives a significantly lower system memory footprint and easy integration into the current World Wide Web (to create Web of Things). Table II represents the smart grid traffic characteristics provided by [15]. We can now make a more informed choice towards protocol selection based on the knowledge of Section III and IV (A). The suggested IoT Protocol section of Table II offers our assessment and reasoning as to why a specific protocol would be a better choice. However, with the varied needs, requirements and compatibility with internal systems, CPU and power performance on client and server side, the final choice will always be in the hands of designers and implementers. Security (cryptography) was not included in the assessment as both rely on SSL/TLS combination. Strict security would make XMPP a contender as it enforces the use

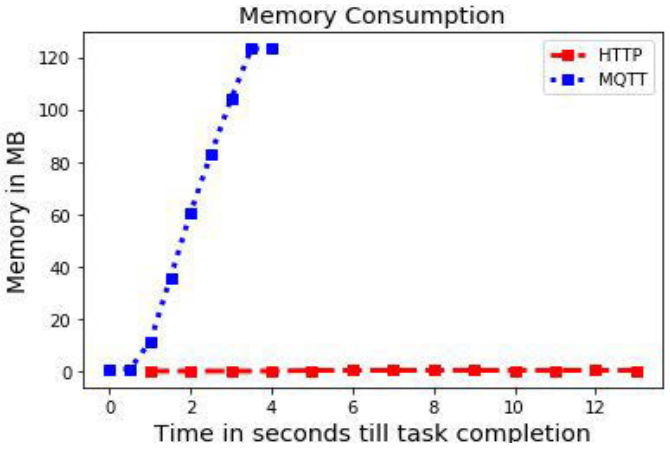


Fig. 3. Memory consumption.

of TLS and SASL, unlike HTTP and MQTT. The analysis confirms our previous notional assessment that different protocols will reside to cater to different needs. In complex systems, like smart grids, multiple protocols will be used. CoAP does try to bring the middle ground but fails in adoption and support.

V. Conclusions and Future Work

The purpose of this work was to highlight the complexity of the protocol selection problem and offer more than one tool/view towards a suitable selection. The point of doing this exercise is to highlight that mere performance results or quantitative results on their own is not an optimal strategy for protocol selection. The analysis, inevitably, also highlights that there is no one-size-fits-all IoT application layer protocol. The selection comes down to a protocol that offers the least compromise in terms of features and utilization available computational capabilities of the devices. The work can further be expanded by testing more IoT protocols. Interoperability is also a big concern, especially considering that the different IoT platforms might offer different protocol support.

Acknowledgments

This work is partly supported by Academy of Finland: FIREMAN (CHIST-ERA/n. 326270), ee-IoT (n.319009) and EnergyNet Research Fellowship (n.321265/n.328869).

References

[1] C. Greer, M. Burns, D. Wollman, and E. Griffor, "Cyber-Physical Systems and Internet of Things NIST Special Publication 1900-202 Cyber-Physical Systems and Internet of Things," 2019.

[2] M. Ullah, P. H. Nardelli, A. Wolff, and K. Smolander, "Twenty-one key factors to choose an iot platform: Theoretical framework and its applications," *arXiv preprint arXiv:2004.04924*, 2020.

[3] M. Ullah and K. Smolander, "Highlighting the Key Factors of an IoT Platform," *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 901–906, 2019.

[4] M. B. Yassein, M. Q. Shatnawi, and D. Al-Zoubi, "Application layer protocols for the Internet of Things: A survey," *Proceedings - 2016 International Conference on Engineering and MIS, ICEMIS 2016*, 2016.

[5] N. Shaukat, S. M. Ali, C. A. Mehmood, B. Khan, M. Jawad, U. Farid, Z. Ullah, S. M. Anwar, and M. Majid, "A survey on consumers empowerment, communication technologies, and renewable generation penetration within Smart Grid," *Renewable and Sustainable Energy Reviews*, vol. 81, no. April 2016, pp. 1453–1475, 2018.

[6] N. De Caro, W. Colitti, K. Steenhaut, G. Mangino, and G. Reali, "Comparison of two lightweight protocols for smartphone-based sensing," *IEEE SCVT 2013 - Proceedings of 20th IEEE Symposium on Communications and Vehicular Technology in the BeNeLux*, pp. 0–5, 2013.

[7] E. P. Frigieri, D. Mazzer, and L. F. C. G. Parreira, "M2M Protocols for Constrained Environments in the Context of IoT : A Comparison of Approaches," *XXXIII Brazilian Telecommunications Symposium (SBrT)*, no. April 2016, pp. 1–4, 2015.

[8] V. Gazis, M. Gortz, M. Huber, A. Leonardi, K. Mathioudakis, A. Wiesmaier, F. Zeiger, and E. Vasilomanolakis, "A survey of technologies for the internet of things," *IWCMC 2015 - 11th International Wireless Communications and Mobile Computing Conference*, pp. 1090–1095, 2015.

[9] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[10] Y. Chen and T. Kunz, "Performance evaluation of IoT protocols under a constrained wireless access network," *2016 International Conference on Selected Topics in Mobile and Wireless Networking, MoWNeT 2016*, 2016.

[11] P. Kayal and H. Perros, "Through a Smart Parking Implementation," *2017 20th Conference on Innovations in Clouds, Internet and Networks (ICIN)*, pp. 331–336.

[12] A. J. Jara, A. C. Olivieri, Y. Bocchi, M. Jung, W. Kastner, and A. F. Skarmeta, "Semantic Web of things: An analysis of the application semantics for the IoT moving towards the IoT convergence," *International Journal of Web and Grid Services*, vol. 10, no. 2-3, pp. 244–272, 2014.

[13] J. Dede and A. Förster, "Comparative analysis of opportunistic communication technologies," *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, vol. 190, pp. 3–10, 2017.

[14] K. Khan, A. Waris, and H. Safi, "A Qualitative Comparison of Various Routing Protocols in WSN," vol. 1, no. 1, pp. 7–13, 2016.

[15] R. H. Khan and J. Y. Khan, "A comprehensive review of the application characteristics and traffic requirements of a smart grid communications network," *Computer Networks*, vol. 57, no. 3, pp. 825–845, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2012.11.002

[16] S. K. Viswanath, C. Yuen, W. Tushar, and W.-T. Li, "SyStem DeSign of the internet of thingS for reSiDential Smart griD," *Smart Microgrids*, no. October, pp. 31–63, 2016.

[17] D. Pratikkumar, S. Amit, and A. Pramod, "Semantic Gateway as a Service architecture for IoT Interoperability," *IEEE Software*, vol. 32, no. 2, 2015.

[18] S. R. U. Kakakhel, T. Westerlund, M. Daneshtalab, Z. Zou, J. Plosila, and H. Tenhunen, "A qualitative comparison model for application layer IoT protocols," *2019 4th International Conference on Fog and Mobile Edge Computing, FMEC 2019*, pp. 210–215, 2019.

[19] P. Koponen, M. L. Pykälä, J. Peltonen, and P. Ahonen, *Interfaces of consumption metering infrastructures with the energy consumers: Review of standards*, 2010, no. 2542.

[20] G. R. Newsham and B. G. Bowker, "The effect of utility time-varying pricing and load control strategies on residential summer peak electricity use: A review," *Energy Policy*, vol. 38, no. 7, pp. 3289–3296, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.enpol.2010.01.027

[21] S. Shao, M. Pipattanasomporn, and S. Rahman, "Grid integration of electric vehicles and demand response with customer choice," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 543–550, 2012.

[22] IEEE, "IEEE standards for electric power systems communication and distribution."

[23] K. Benjamin, L. Robert, I. Toshifumi, and M. Satoshi, "Making the MicroGrid," no. june, pp. 54–65, 2008.

[24] B. Kroposki, T. Basso, and R. DeBlasio, "Microgrid standards and technologies," *IEEE Power and Energy Society 2008 General Meeting: Conversion and Delivery of Electrical Energy in the 21st Century, PES*, 2008.

[25] A. Naumann, P. Komarnicki, M. Powalko, Z. A. Styczynski, J. Blumschein, and M. Kereit, "Experience with PMUs in industrial distribution networks," *IEEE PES General Meeting, PES 2010*, pp. 1–6, 2010.

[26] M. Erol-Kantarci and H. T. Mouftah, "Wireless multimedia sensor and actor networks for the next generation power grid," *Ad Hoc Networks*, vol. 9, no. 4, pp. 542–551, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.adhoc.2010.08.005

A Survey on Parallel Architectures and Programming Models

Branimir Pervan and Josip Knezović
University of Zagreb, Faculty of Electrical Engineering and Computing
branimir.pervan@fer.hr; josip.knezovic@fer.hr

***Abstract*—In addition to large-scale computers, multicore processors have taken a significant part in all kinds of devices, from personal computers to cell phones. Although programming techniques for parallel systems exist for a while, the development of applications that can appropriately utilize multicores is still challenging in many aspects, especially for full exploitation of the computational resources. Additionally, another challenge is the efficient and easy programming of heterogeneous systems for the complete exploitation of silicon resources. Solutions to making parallel programming more developer-friendly are various programming models that abstract parallelism and concurrency. Implementations of those models need to extend even to lower layers of software parallelism and hardware parallelism as well.**

This paper gives an overview of parallel architectures and trending programming models for such processing units and systems. It also presents challenges to scalability and portability in parallel systems and presents up to date trends in heterogeneous systems that heavily exploit parallelism.

***Index Terms*—parallel architectures, parallel programming, heterogeneous systems, scalability, portability, domain-specific languages, reactive programming**

I. Introduction

Moore's Law stated that the number of transistors on an integrated circuit doubles approximately every two years. The author of the law, in a recent interview [1] said that he sees Moore's law "dying here in the next decade or so", main reason being the fact that technology has hit 3 types of "wall", namely [2]:

1) Power wall - As a result of the significant increase in frequency, the ability to dissipate heat has reached a physical limit.
2) Instruction-Level parallelism wall - ILP causes a super-linear increase in complexity and power consumption of the processing unit without linear speed up in the application performance.
3) Memory wall - The mismatch between memory speed and computation speed.

From the latter follows that new hardware has to be designed and developed in multicore or multichip fashion to deliver expected higher performance. Consequently, programs need to be developed in such a manner to exploit given multicore system.

To achieve desired speedups in multicore and potentially heterogeneous environments, one has to exploit potential parallelism on every layer, starting from Instruction-Level to upper layers. Although most of today's languages offer explicit programming at the level of threads and locks [3], without modern programming constructs, such models are unreasonably difficult to use on a wider scale and developer-unfriendly. Parallel programming models offer an abstraction over lower layers of parallelism with varying success.

The main reason for research in the domain of parallel processing is higher availability of multicores in every segment of society mainly due to lower prices and favorable cost-performance ratios. Also, most of the algorithms developed in computer science are designed for single instruction flow, and lots of existing applications aren't able to efficiently exploit underlying multicore hardware as well.

Heterogeneous parallel programming is another reason for research as it emerged with the appearance of General-purpose Graphics processing units (GPGPUs) and programming frameworks such as CUDA [4]. The presence of GPUs in computers of an everyday user enables (to a wider audience) processing of large amounts of vectorized data in parallel. Availability of Field-programmable gate array (FPGA) technology opened a way into cheap specific accelerators. Programming models and frameworks for easy and efficient development of such accelerators are yet to be explored [5]. Another challenge is algorithms for runtime management which would be able to detect which parts of code should be executed on which hardware to achieve maximum performance or cost-performance ratio. Also, High-Level synthesis presents a challenge, considering the limitations in existing frameworks (i.e. Vivado HLS [6]) especially when programming languages with higher levels of abstraction are used.

The rest of this paper is organized as follows: Architecture models are classified in section 2, while an overview of most common parallel programming models is given in section 3. Current major trends and possible research topics are listed and analyzed in section 4.

II. Architecture models

The base for any classification of any computer architecture, which implies parallel architectures as well, starts with Flynn's classification [7]. Flynn proposed four main ways in which computer approaches instruction and memory streams:

- Single instruction stream, single data stream - SISD
- Single instruction stream, multiple data streams - SIMD
- Multiple instruction streams, single data stream - MISD
- Multiple instruction streams, multiple data streams - MIMD

Although not in Flynn's classification, various sources added Single instruction, multiple threads - SIMT as an execution model where SIMD model is combined with multithreading techniques, and SPMD - Single program, multiple data streams [8].

On the other hand, Hennessy and Patterson classify parallelism, with respect to applications, on two levels [9]:

- Data-Level parallelism - This level of parallelism copes with parallelization problems by focusing on ways of distributing data across different computation nodes or devices which then operate on that data in parallel. The scale on which it can be applied varies greatly, from multicores to heterogeneous clusters. Data containers for data which is operated on can be specialized data structures or any regular data structures like arrays or matrices that are being divided and distributed.
- Task-Level parallelism - Focuses on separating and scheduling different tasks for execution on different processing units. Those tasks can operate on shared or distributed data.

With respect to ways in which computer hardware can exploit those two types of parallelism, there are four major ways [9]:

- Instruction level
- Vector architectures and GPUs
- Thread level
- Request level

A. Instruction-Level parallelism

Instruction-Level parallelism, in its most common case, implies overlapping of instructions during execution to achieve an increase in performance. There are basically two approaches to exploit ILP: Dynamic (Hardware) and Static (Software). Dynamic exploitation relies on capability of processors to detect when instructions can be executed in parallel, while software exploitation relies on compilers to detect parallelizable instructions at compile time [9].
The most common technique used to exploit Instruction-Level parallelism is Pipelining. Given that every modern processor has an *n*-stage pipeline where every stage represents one phase in execution of an instruction, exploiting would consist of overlapping instructions (coloured differently in figure 1) in different phases of execution. Keeping a pipeline full requires finding sequences of unrelated instructions either during runtime or compile time.

Loop Unrolling is another technique to exploit ILP. It is a loop transformation technique which exploits space-time tradeoff eliminating instructions that control the loop, such as pointer arithmetic and tests for loop ending on each iteration

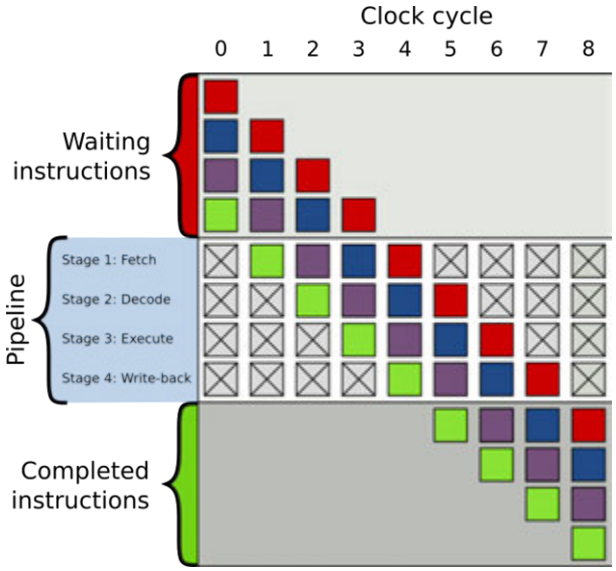


Fig. 1. 4-stage pipeline [10]

[11]. This effectively reduces branch penalties and delays in reading data from memory. Unrolling itself is performed in such way that loop is rewritten as a repeated sequence of independent statements [12].

Branch penalties can be reduced by using Branch Prediction techniques. Branch prediction aims to predict which branch the processor has to take after the execution of a conditional jump. Static branch prediction uses information collected from earlier runs. The key observation is that individual branch is often highly biased toward taken or untaken [9]. Dynamic branch prediction uses information collected on-the-fly. Such techniques include branch prediction buffers (more concretely, for example, Two-level adaptive predictor [13]) or branch history tables.
Another technique is Dynamic Scheduling. A major limitation of Pipelining is potential data dependency between instructions in various phases of execution. If currently executing instruction requires the result of an instruction that is also in pipeline, data hazard will happen and execution will be stalled. Using Dynamic Scheduling technique, processor rearranges the instruction execution while maintaining data flow and exception behavior. [9].

B. Vector architectures and Graphics Processing Unit

A prototype of SIMD architecture is a vector processor (vector architectures in general). Those are based on the fact that one instruction can be applied to multiple data in parallel exploiting data-level parallelism [9]. Vector processors can also operate on one-dimensional arrays (vectors) which enable them to process large amounts of data in parallel. Another way in which vector processor can exploit its capabilities is to grab a set of data elements scattered around memory, to organize them in large vector registers, and to process them.
On the other hand, Graphics processing units (GPU) have

always been processors with capabilities to process large amounts of data in parallel. With the appearance of interfaces and platforms which enabled easier development of programs for such devices (e.g. CUDA [4]), they emerged as purely computational devices. CUDA uses a modified version of C language to enable the programmer to utilize the full potential of GPUs and GPU-like architectures with many of parallel floating-point units.
SIMD extensions for multimedia, such as SSEx or AVX [14], can be considered as another type of vector architectures as they exploit the fact that multimedia data is usually narrower than internal processor buses, registers and computational units.

C. Thread-Level parallelism

This type of parallelism is being exploited from higher levels of abstraction relative to ILP and Vector processors. Threads, being the smallest sequence of instruction that can be managed independently [15], imply the existence of multiple program counters and hence are being exploited primarily through MIMD computers [9]. Thread-Level parallelism is being exploited on various scaled computers: from desktop computers to servers. By running many threads at once, one can decouple various parts of the program of which some could be prone to high latencies, and thus better exploit assigned processing time. On a server scale, exploitation of Thread-Level parallelism leads to increase in performance and availability when it comes to high amounts of I/O and memory latencies.

D. Request-Level parallelism

Being implemented mostly on larger scales, Request-Level parallelism exploits parallelism among largely decoupled tasks specified by the programmer or the operating system [9]. Such tasks are, e. g., user requests to a web service hosted on a server, or multiple queries to database on servers that, nowadays, handle millions of requests within hours. In general, those requests are largely independent and mostly involve read-only data operations. Computations within those requests are also easily partitioned within a request and across different requests.

III. Programming models

Programming model is an abstraction of architecture used to express programs for the abstracted platform. It exposes interfaces of some subset of its functionalities to enable the programmer to write programs for that platform. Parallel programming model abstracts parallel architecture. Parallel programming problems can be approached with respect to process interaction and with respect to problem decomposition [16]. Problem decomposition approach is analogous to already mentioned classification of parallelism with respect to applications (Data-Level and Task-Level parallelism) with the addition of implicit parallelism which resembles everything that the compiler, the runtime or the hardware are responsible for. On the other hand, the process interaction approach takes into consideration ways of unavoidable communication between parallel processes. These are:

- Message passing - Processes exchange data through messages. Communication can be either synchronous or asynchronous
- Shared memory - Processes communicate through part of memory which can be accessed by all participants. To avoid race conditions, semaphores and monitors are usually utilized
- Implicit communication - Covers every aspect of communication hidden from the programmer.

Savage's process interaction classification fits to classification of pure parallel programming models according to [17] who also proposes implementations of those models:

A. Pthreads

POSIX threads (Pthreads) is standardized programming interface for exploiting threading mechanisms implemented by various processors [18]. Although Pthreads originated as UNIX standard, it eventually settled down as an official. Pthreads have access to their private and to global (shared) process memory. API contains 4 main groups of subroutines:

- Thread management
- Mutexes
- Condition variables
- Synchronization

Unlike processes, they are more lightweight when it comes to creation and memory requirements. On the other hand, they require manual synchronization through mutexes, and can't easily be ported to non-POSIX compliant systems.

B. OpenMP

Open Multi-Processing (OpenMP [19]) is an API that supports multiprocessing on most platforms through utilization of shared memory. It is implemented as an extension for C, C++ and Fortran programming languages [20] and thus requires a compiler for the respective languages with OpenMP support . Generated code is portable and can be used on GPUs as well. Since it relies on the shared memory model, communication model is simple (no need to incorporate messaging). On the other hand, it currently runs well only in shared memory multiprocessor platforms.

C. MPI

Message passing interface implementations rely purely on message exchange to enable programming of parallel computers. Absence of using shared memory (only limited distributed shared memory in MPI-2) makes it memory independent and more scalable than OpenMP. The MPI standard includes following concepts [21]:

- Communicator
- Point-to-point message-passing
- Collective communications
- Group and communicator concepts
- Process topologies
- Environmental management

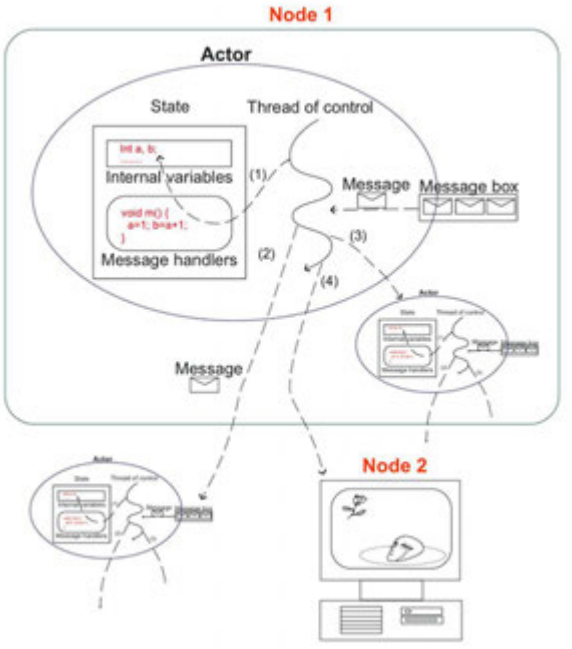


Fig. 2. Available actions for actors in the Actor model [26]

- Process creation and management
- One-sided communications
- Extended Collective operations,
- External interfaces
- I/O
- Profiling interface

To date, one of MPI implementations, namely Open MPI is used by many supercomputers from the TOP500 list [22].

D. Actors

The actor model is a mathematical model that treats actors as primitives of parallel computation. As an autonomous unit, an actor can (depicted in figure 2):

- Receive messages
- Make local decisions
- Create new actors
- Send messages
- Set up response for next message received

To avoid locks, actors can reach each other only through messages. Global state modifying is forbidden, as an actor can modify only its private state. One possible implementation of Actor model is in programming language Scala [23], [24] and Akka framework [25].

IV. Trends

A. Domain-Specific Languages

Except for the parallel programming itself, the current major trend in this field is efficient exploitation of processing power in heterogeneous environments using Domain-Specific languages. Such languages, in most cases, abstract away parallelism or concurrency and heterogeneity. Normally, Domain-Specific language is a programming language itself and it requires whole compiler infrastructure which, in addition to standard optimizations, enables Domain-Specific optimizations such as Optimizing Linear Computations [27]. To help overcome that, Delite [28], [29] has been developed (whole stack represented in figure 3). It is a framework for building custom Domain-Specific languages, compiler infrastructure, runtime and management as well. Domain-Specific languages built on Delite are embedded into Scala [24], general purpose programming language. Those languages are then transpiled to Intermediate representation and then further compiled to C++, CUDA or OpenCL [30]. Example of a Domain-Specific language built upon Delite is OptiML, used for machine learning purposes [31]. Additional example of a DSL paired with mathematically provable rewriting system is LIFT [32] Another Domain-Specific language approach is an intermediate language used to aid programmers to express transformations to better exploit parallelism. An example of such language is DMLL - Distributed MultiLoop Language [33].

B. Stream processing

Stream processing using Domain-Specific languages is another major trend. This programming languages facilitate processing of large sequences of data, such as multimedia, video, graphic or signal processing [34], [35]. One such Domain-Specific language is StreamIt [36] which is, although it originates from 2002., still actual due to regular updates and additional research [34], [37] including [38] which proposed transformations, facilitated through Intermediate language - LaminarIR in between of compilation of StreamIt to C, resulting in dramatic speedups and energy savings. Another notable project in field of stream processing is compilation flow of Brook streaming language (Brook) directly to Field-programmable gate array circuits (FPGA) [39], [40]. Some work has been done in optimization of throughput for streaming applications on CPU - FPGA systems under latency and power constraints [41]. Also, algorithms have been developed to map tasks onto the heterogeneous systems by their architectural capabilities and task characteristics. From other point of view, [37] developed a framework to map streaming applications onto GPU, while [42] developed Java programming language extension which incorporates stream processing with the language which later runs on Java virtual machine.

C. Custom runtimes

Another approach would be to create custom runtimes and management systems for parallel and heterogeneous environments. Harmony [43] is a runtime and execution model that provides parallelism management, dynamic scheduling, matching of certain classes of algorithms to specifically tailored architectures and online monitoring. Another available system is Dandelion [44] by Microsoft. It provides a unified programming model to program heterogeneous systems with data-parallel applications. By utilizing .NET framework

and LINQ, it integrates data-parallel operations into general-purpose programming languages (C#, F#). Compiling chain later cross-compiles .NET code to CUDA kernels. Data-parallel portions are distributed automatically. HPX (High Performance ParalleX) [45] is a parallel runtime system that facilitates distributed operations, fine-grained constraint-based parallelism, and supports runtime adaptive resource management. Delite [28], [29] could also be classified as a custom runtime framework.

D. Reactive programming

Another trend to consider is using Reactive programming techniques [46] and Functional programming languages. Immutability and event-driven development can make parallel programming significantly easier, more explicit and expressive. Functional languages with parallel patterns can also be used for hardware generation. Moreover, they can be useful in avoiding many of the problems faced when generating hardware from imperative languages [47]. Reactiveness can also be achieved on processor level [48]. Using parallel data structures where parallelism is abstracted away in operations on those structures can also significantly increase performance [49].

E. Programming languages

To achieve and express parallelism, various researchers proposed extensions to the existing programming languages or completely novel programming languages.

1) Extensions: Extensions to the existing general-purpose programming languages are a possibility as well. Cilk [50] and Charm++ [51] are examples of extensions of the C / C++ programming languages, while X10 [52] and Habanero-Java (successor to the aforementioned X10) [53] are extensions to the Java programming language. Intel Threading Building Blocks [54] is an example of a C++ template library used for parallel programming on multicore processors. It provides a work-stealing task scheduler and a scalable memory allocator.

2) Custom programming languages: Chapel [55] is a custom programming language for productive parallel computing with support for distributed arrays, data parallelism, and task parallelism. Legion [56] is also a custom programming model and runtime system for distributed heterogeneous computers. Legion introduces logical regions as a new abstraction for describing the structure and usage of program data.

F. Other approaches

In heterogeneous environments, especially with FPGAs, speedups can be achieved on Sliding-Window applications [57]. They can also be used in non-explicitly computing environments, like network switches in data centers to improve overall performance and reduce energy costs [58].

By extending OpenMP programming model, OmpSs [59] and OpenACC [60] provide support for heterogeneous hardware by exposing programming constructs for offloading kernels to accelerators. Other approaches include languages like C++ AMP [61] in advantage to OpenCL and CUDA claiming better portability across different architectures and generations within the same architecture [62]. Another trend are customized heterogeneous High Performance Computing systems, i.e. MANGO project [63].

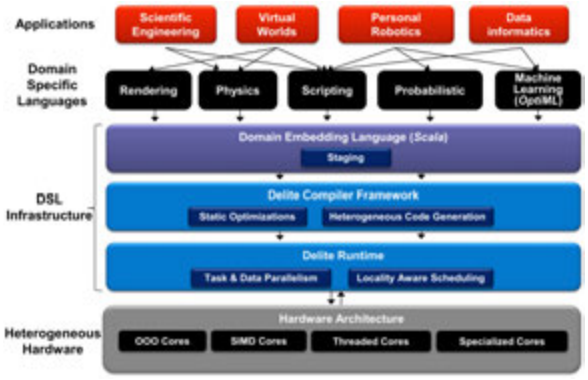


Fig. 3. The Delite framework for Domain-Specific languages [28]

V. Conclusion

Although virtually omnipresent, parallel programming and its challenges persisted from the very beginning of multiprocessing and multicore computational units. However, it is more actual than ever for a couple of main reasons:

- Affordable prices of various multicore processors in both home and enterprise environments
- Existing applications are not efficient at exploiting parallelism

Moreover, popularization of General-purpose GPUs (GPGPU) enhanced with appropriate programming models (CUDA) introduced affordable heterogeneity. Also, FPGA circuits introduced solutions for fast Domain-specific accelerations with acceptable price-performance ratio minding the expensive hardware development process.

To improve future research, an overview of general concepts of parallel architectures and its exploitations was made. The most common existing parallel programming models, for pure and heterogeneous platforms as well, with up-to-date trends, were analyzed.

Future research could be done on multiple levels:

1) Easier and more programmer-friendly parallel programming models efficient at exploiting parallelism on lower levels of abstraction
2) Explicit programming models for heterogeneous environments
3) Algorithms for automatic discovery of parallelizable code targeted for specific devices in heterogeneous environment
4) Reactive and Event-driven programming with parallelizable units of work in heterogeneous environment
5) Integration of Functional and reactive programming into heterogeneous programming models

6) Domain-Specific languages design and compiler infrastructure for optimal targeting heterogeneous systems

REFERENCES

[1] "Gordon Moore: The Man Whose Name Means Progress, The visionary engineer reflects on 50 years of Moore's Law," http://spectrum.ieee.org/computing/hardware/gordon-moore-the-man-whose-name-means-progress, Accessed: 2.6.2020.

[2] J. L. Manferdelli, N. K. Govindaraju, and C. Crall, "Challenges and Opportunities in Many-Core Computing," *Proceedings of the IEEE*, vol. 96, no. 5, pp. 808–815, May 2008.

[3] H. Sutter and J. Larus, "Software and the Concurrency Revolution," *Queue*, vol. 3, no. 7, pp. 54–62, Sep. 2005. [Online]. Available: http://doi.acm.org/10.1145/1095408.1095421

[4] "NVIDIA CUDA," https://developer.nvidia.com/cuda-zone, Accessed: 2.6.2020.

[5] D. Bacon, R. Rabbah, and S. Shukla, "FPGA Programming for the Masses," *Queue*, vol. 11, no. 2, pp. 40:40–40:52, Feb. 2013. [Online]. Available: http://doi.acm.org/10.1145/2436696.2443836

[6] "Vivado High-Level Synthesis," https://www.xilinx.com/products/design-tools/vivado/integration/esl-design.html, Accessed: 2.6.2020.

[7] M. J. Flynn, "Some Computer Organizations and Their Effectiveness," *IEEE Transactions on Computers*, vol. C-21, no. 9, pp. 948–960, Sept 1972.

[8] M. McCool, J. Reinders, and A. Robison, *Structured parallel programming: patterns for efficient computation*. Elsevier, 2012.

[9] J. L. Hennessy and D. A. Patterson, *Computer Architecture, Fifth Edition: A Quantitative Approach*, 5th ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[10] "4-stage pipeline," https://upload.wikimedia.org/wikipedia/commons/c/cb/Pipeline%2C_4_stage.svg, Accessed: 2.6.2020.

[11] W. Petersen and P. Arbenz, *Introduction to Parallel Computing : A practical guide with examples in C: A practical guide with examples in C*, ser. Oxford Texts in Applied and Engineering Mathematics. OUP Oxford, 2004. [Online]. Available: https://books.google.hr/books?id=azN8DFRtDTgC

[12] A. Nicolau, *Loop Quantization: Unwinding for Fine-Grain Parallelism Exploitation*, ser. Cornell University, Department of Computer Science, Technical report. Ithaca, N.Y.: Cornell University, Dept. of Computer Science, 1985.

[13] T.-Y. Yeh and Y. N. Patt, "Two-level Adaptive Training Branch Prediction," in *Proceedings of the 24th Annual International Symposium on Microarchitecture*, ser. MICRO 24. New York, NY, USA: ACM, 1991, pp. 51–61. [Online]. Available: http://doi.acm.org/10.1145/123465.123475

[14] Intel, *Intel 64 and IA-32 Architectures Software Developer's Manual, Volume 1: Basic Architecture*, http://www.intel.com/content/dam/www/public/us/en/documents/manuals/64-ia-32-architectures-software-developer-vol-1-manual.pdf, Accessed: 2.6.2020.

[15] L. Lamport, "How to Make a Multiprocessor Computer That Correctly Executes Multiprocess Programs," *IEEE Transactions on Computers*, vol. C-28, no. 9, pp. 690–691, Sept 1979.

[16] J. E. Savage, *Models of computation*. Addison-Wesley Reading, MA, 1998, vol. 136.

[17] J. Diaz, C. Munoz-Caro, and A. Nino, "A Survey of Parallel Programming Models and Tools in the Multi and Many-Core Era," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 8, pp. 1369–1386, Aug 2012.

[18] "POSIX Threads Programming," https://computing.llnl.gov/tutorials/pthreads/, Accessed: 2.6.2020.

[19] "OpenMP," http://www.openmp.org/, Accessed: 2.6.2020.

[20] A. Silberschatz, P. Galvin, and G. Gagne, *Operating System Concepts*. Wiley, 2012. [Online]. Available: https://books.google.hr/books?id=-qU8ygAACAAJ

[21] "MPI: A Message-Passing Interface Standard Version 3.1," http://mpi-forum.org/docs/mpi-3.1/mpi31-report.pdf, Accessed: 2.6.2020.

[22] "TOP500," https://www.top500.org/, Accessed: 2.6.2020.

[23] P. Haller and M. Odersky, "Actors That Unify Threads and Events," in *Proceedings of the 9th International Conference on Coordination Models and Languages*, ser. COORDINATION'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 171–190. [Online]. Available: http://dl.acm.org/citation.cfm?id=1764606.1764620

[24] M. Odersky and al., "An Overview of the Scala Programming Language," EPFL Lausanne, Switzerland, Tech. Rep. IC/2004/64, 2004.

[25] "Akka Framework," http://akka.io/, Accessed: 2.6.2020.

[26] "The Actor Model," https://wcl.cs.rpi.edu/salsa/tutorial/salsa1_1_2/node6.html, Accessed: 2.6.2020.

[27] W. Thies, "Language and Compiler Support for Stream Programs," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[28] K. J. Brown, A. K. Sujeeth, H. J. Lee, T. Rompf, H. Chafi, M. Odersky, and K. Olukotun, "A Heterogeneous Parallel Framework for Domain-Specific Languages," in *Proceedings of the 2011 International Conference on Parallel Architectures and Compilation Techniques*, ser. PACT '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 89–100. [Online]. Available: http://dx.doi.org/10.1109/PACT.2011.15

[29] H. Lee, K. Brown, A. Sujeeth, H. Chafi, T. Rompf, M. Odersky, and K. Olukotun, "Implementing Domain-Specific Languages for Heterogeneous Parallel Computing," *IEEE Micro*, vol. 31, no. 5, pp. 42–53, Sep. 2011. [Online]. Available: http://dx.doi.org/10.1109/MM.2011.68

[30] A. K. Sujeeth, K. J. Brown, H. Lee, T. Rompf, H. Chafi, M. Odersky, and K. Olukotun, "Delite: A Compiler Architecture for Performance-Oriented Embedded Domain-Specific Languages," *ACM Trans. Embed. Comput. Syst.*, vol. 13, no. 4s, pp. 134:1–134:25, Apr. 2014. [Online]. Available: http://doi.acm.org/10.1145/2584665

[31] H. Chafi, A. K. Sujeeth, K. J. Brown, H. Lee, A. R. Atreya, and K. Olukotun, "A Domain-specific Approach to Heterogeneous Parallelism," *SIGPLAN Not.*, vol. 46, no. 8, pp. 35–46, Feb. 2011. [Online]. Available: http://doi.acm.org/10.1145/2038037.1941561

[32] M. Steuwer, T. Remmelg, and C. Dubach, "Lift: a functional data-parallel ir for high-performance gpu code generation," in *2017 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*. IEEE, 2017, pp. 74–85.

[33] K. J. Brown, H. Lee, T. Rompf, A. K. Sujeeth, C. De Sa, C. Aberger, and K. Olukotun, "Have Abstraction and Eat Performance, Too: Optimized Heterogeneous Computing with Parallel Patterns," in *Proceedings of the 2016 International Symposium on Code Generation and Optimization*, ser. CGO '16. New York, NY, USA: ACM, 2016, pp. 194–205. [Online]. Available: http://doi.acm.org/10.1145/2854038.2854042

[34] S. M. Farhad, Y. Ko, B. Burgstaller, and B. Scholz, "Profile-guided Deployment of Stream Programs on Multicores," in *Proceedings of the 13th ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, Tools and Theory for Embedded Systems*, ser. LCTES '12. New York, NY, USA: ACM, 2012, pp. 79–88. [Online]. Available: http://doi.acm.org/10.1145/2248418.2248430

[35] G. Ofenbeck, T. Rompf, A. Stojanov, M. Odersky, and M. Puschel, "Spiral in Scala: Towards the Systematic Construction of Generators for Performance Libraries," *SIGPLAN Not.*, vol. 49, no. 3, pp. 125–134, Oct. 2013. [Online]. Available: http://doi.acm.org/10.1145/2637365.2517228

[36] W. Thies, M. Karczmarek, and S. P. Amarasinghe, "StreamIt: A Language for Streaming Applications," in *Proceedings of the 11th International Conference on Compiler Construction*, ser. CC '02. London, UK, UK: Springer-Verlag, 2002, pp. 179–196. [Online]. Available: http://dl.acm.org/citation.cfm?id=647478.727935

[37] H. P. Huynh, A. Hagiescu, O. Z. Liang, W. F. Wong, and R. S. M. Goh, "Mapping Streaming Applications onto GPU Systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 9, pp. 2374–2385, Sept 2014.

[38] Y. Ko, B. Burgstaller, and B. Scholz, "LaminarIR: Compile-time Queues for Structured Streams," *SIGPLAN Not.*, vol. 50, no. 6, pp. 121–130, Jun. 2015. [Online]. Available: http://doi.acm.org/10.1145/2813885.2737994

[39] F. Plavec, Z. Vranesic, and S. Brown, *Stream Programming for FPGAs*. Dordrecht: Springer Netherlands, 2009, pp. 241–253. [Online]. Available: http://dx.doi.org/10.1007/978-1-4020-9714-0_16

[40] ——, "Exploiting Task- and Data-Level Parallelism in Streaming Applications Implemented in FPGAs," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 6, no. 4, pp. 16:1–16:37, Dec. 2013. [Online]. Available: http://doi.acm.org/10.1145/2535932

[41] X. Wei, Y. Liang, T. Wang, S. Lu, and J. Cong, "Throughput Optimization for Streaming Applications on CPU-FPGA Heterogeneous Systems," *Power (Watt)*, vol. 16, p. 18, 2016.

[42] J. H. Spring, J. Privat, R. Guerraoui, and J. Vitek, "Streamflex: High-throughput Stream Programming in Java," *SIGPLAN Not.*, vol. 42, no. 10, pp. 211–228, Oct. 2007. [Online]. Available: http://doi.acm.org/10.1145/1297105.1297043

[43] G. F. Diamos and S. Yalamanchili, "Harmony: An Execution Model and Runtime for Heterogeneous Many Core Systems," in *Proceedings of the 17th International Symposium on High Performance Distributed Computing*, ser. HPDC '08. New York, NY, USA: ACM, 2008, pp. 197–200. [Online]. Available: http://doi.acm.org/10.1145/1383422.1383447

[44] C. J. Rossbach, Y. Yu, J. Currey, J.-P. Martin, and D. Fetterly, "Dandelion: A Compiler and Runtime for Heterogeneous Systems," in *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, ser. SOSP '13. New York, NY, USA: ACM, 2013, pp. 49–68. [Online]. Available: http://doi.acm.org/10.1145/2517349.2522715

[45] H. Kaiser, T. Heller, B. Adelstein-Lelbach, A. Serio, and D. Fey, "Hpx: A task based programming model in a global address space," in *Proceedings of the 8th International Conference on Partitioned Global Address Space Programming Models*, 2014, pp. 1–11.

[46] S. Khare and A. Gokhale, "Distributed Reactive Stream Processing."

[47] R. Prabhakar, D. Koeplinger, K. J. Brown, H. Lee, C. De Sa, C. Kozyrakis, and K. Olukotun, "Generating Configurable Hardware from Parallel Patterns," *SIGOPS Oper. Syst. Rev.*, vol. 50, no. 2, pp. 651–665, Mar. 2016. [Online]. Available: http://doi.acm.org/10.1145/2954680.2872415

[48] X. Li and R. von Hanxleden, "Multithreaded Reactive Programming the Kiel Esterel Processor," *IEEE Transactions on Computers*, vol. 61, no. 3, pp. 337–349, March 2012.

[49] A. Prokopec, P. Bagwell, T. Rompf, and M. Odersky, "A Generic Parallel Collection Framework," in *Proceedings of the 17th International Conference on Parallel Processing - Volume Part II*, ser. Euro-Par'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 136–147. [Online]. Available: http://dl.acm.org/citation.cfm?id=2033408.2033425

[50] R. D. Blumofe, C. F. Joerg, B. C. Kuszmaul, C. E. Leiserson, K. H. Randall, and Y. Zhou, "Cilk: An efficient multithreaded runtime system," *Journal of parallel and distributed computing*, vol. 37, no. 1, pp. 55–69, 1996.

[51] L. V. Kale and S. Krishnan, "Charm++ a portable concurrent object oriented system based on c++," in *Proceedings of the eighth annual conference on Object-oriented programming systems, languages, and applications*, 1993, pp. 91–108.

[52] V. A. Saraswat, V. Sarkar, and C. von Praun, "X10: concurrent programming for modern architectures," in *Proceedings of the 12th ACM SIGPLAN symposium on Principles and practice of parallel programming*, 2007, pp. 271–271.

[53] V. Cavé, J. Zhao, J. Shirako, and V. Sarkar, "Habanero-java: the new adventures of old x10," in *Proceedings of the 9th International Conference on Principles and Practice of Programming in Java*, 2011, pp. 51–61.

[54] A. Kukanov and M. J. Voss, "The foundations for scalable multi-core software in intel threading building blocks." *Intel Technology Journal*, vol. 11, no. 4, 2007.

[55] B. L. Chamberlain, D. Callahan, and H. P. Zima, "Parallel programmability and the chapel language," *The International Journal of High Performance Computing Applications*, vol. 21, no. 3, pp. 291–312, 2007.

[56] M. E. Bauer, "Legion: Programming distributed heterogeneous architectures with logical regions," Ph.D. dissertation, Stanford University, 2014.

[57] P. Cooke, J. Fowers, G. Brown, and G. Stitt, "A Tradeoff Analysis of FPGAs, GPUs, and Multicores for Sliding-Window Applications," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 8, no. 1, pp. 2:1–2:24, Mar. 2015. [Online]. Available: http://doi.acm.org/10.1145/2659000

[58] C. Kachris and D. Soudris, "A survey on reconfigurable accelerators for cloud computing," in *2016 26th International Conference on Field Programmable Logic and Applications (FPL)*, Aug 2016, pp. 1–10.

[59] A. Duran, E. Ayguadé, R. M. Badia, J. Labarta, L. Martinell, X. Martorell, and J. Planas, "Ompss: a proposal for programming heterogeneous multi-core architectures," *Parallel processing letters*, vol. 21, no. 02, pp. 173–193, 2011.

[60] "OpenACC," https://www.openacc.org/, Accessed: 2.7.2020.

[61] K. Gregory and A. Miller, *C++ AMP*. Microsoft Press, 2012.

[62] M. Daga, Z. S. Tschirhart, and C. Freitag, "Exploring Parallel Programming Models for Heterogeneous Computing Systems," in *2015 IEEE International Symposium on Workload Characterization*, Oct 2015, pp. 98–107.

[63] J. Flich, G. Agosta, P. Ampletzer, D. A. Alonso, A. Cilardo, W. Fornaciari, M. Kovac, F. Roudet, and D. Zoni, "The MANGO FET-HPC Project: An Overview," in *2015 IEEE 18th International Conference on Computational Science and Engineering*, Oct 2015, pp. 351–354.

HPC-based parallel software for solving applied Boolean satisfiability problems

V.G. Bogdanova, S.A. Gorsky, and A.A. Pashinin
Matrosov Institute for System Dynamics and Control Theory of SB RAS, Irkutsk, Russia
bvg@icc.ru, gorsky@icc.ru, apcrol@gmail.com

Abstract - **Boolean satisfiability is one of the fundamental problems in mathematical logic and the theory of computation. Many practical problems can be formulated as problems of Boolean satisfiability (solving a system of Boolean equations). These include the tasks of cryptography, qualitative research of binary dynamic systems, logical programming, robot action planning, and many others. NP-complexity of the Boolean satisfiability problem actualizes the development of parallel software for solving it in a high-performance computing environment. We focus on the development of specialized parallel solvers for Boolean satisfiability problems oriented to the application in the qualitative study of binary dynamic systems based on the Boolean constraint method. The Boolean constraint method uses a formal definition of the dynamical property, presented in the language of predicate logic with limited existential and universal quantifiers. Based on this definition, a model of dynamical property is formed in the form of Boolean constraints by a series of sequential formal transformations. The developed parallel solvers provide a conclusion on the feasibility of Boolean constraints, interpret this conclusion accordingly the tested property, and find, unlike similar software tools, all sets of values of Boolean variables that lead to this conclusion. An example of applying the developed solver to constructing cycles of a given length of stream ciphers based on shift registers is considered. The results confirming the effectiveness and scalability of the developed software are presented.**

Keywords - Boolean satisfiability problem; Boolean constraints method; parallel solver; binary dynamic systems

I. INTRODUCTION

Mathematical models described in the language of Boolean equations (constraints) are widely used in various subject areas, such as pattern recognition, coding theory, cryptography, and qualitative analysis of binary dynamic systems (BDS). Reducing exhaustive problems from these areas to a Boolean satisfiability problem requires the development of a technology for constructing Boolean formulas. The computational complexity of the satisfiability problem for Boolean constraints actualizes the development of parallel software for solving it in a high-performance computational environment. We use microservice-oriented technology HPCATAMP [1] for a qualitative study of BDS based on the Boolean constraint method (BCM, [2]). The constructing Boolean formulas technology for encoding the problems in BDS qualitative analysis to the Boolean satisfiability was considered in [3]. Using BCM, a model of the dynamical property of BDS is built. This model includes constraints, and their feasibility indicates the feasibility of this property. These constraints are Boolean while studying the dynamics of BDS in a finite time interval. The Boolean model of the BDS property takes into account both the dynamic equations of the BDS and the structural characteristics of the property expressed in the language of propositional logic or predicate logic with bounded quantifiers. Thus, checking the property satisfiability is reduced to solving an NP-complexity SAT problem (or PSPACE-complex 2QBF problem, depending on the property) with subsequent solving based on data parallelism in case of the large dimension BDS. Let us call this approach the BCM-approach.

In this study, we propose a specialized parallel solver for Boolean satisfiability problems oriented for applying the BCM-approach in the qualitative analysis of BDS. Data parallelization is performed by splitting the original Boolean model. The subtasks for verifying the satisfiability of the obtained submodels are performed as independent ones on a high-performance computational cluster. The developed parallel SAT solver provides a conclusion on the feasibility of Boolean constraints, and find unlike similar software tools, all or a given number of satisfying assignments of Boolean variables. This solver is intended for high-performance computational SMP cluster. In a hybrid heterogeneous high-performance environment, a service-oriented approach is used for solving BDS qualitative analysis problems. The parallel solver is implemented as a microservice installed on dedicated multi-core computing resources. The problem statement agent with the web interface accepts the user's request, performs splitting of the initial Boolean model at the upper level, sends out tasks to microservice-solvers, monitors their implementation, and combines the results of the work. Each microservice simultaneously determines the feasibility of a submodel. Microservices for building a Boolean model and verifying its feasibility are designed both for stand-alone and for joint work in a distributed solver [4] under the control of an active group of agents that self-organizes for the non-procedural formulation of the problem. The research described in [4] develops in the direction both for creating packages of applied microservices [1] and for implementing a composite service based on such a package. The problem statement made earlier determines the composition of this service.

The search for cycles of limited length is one of the main tasks of qualitative analysis, the computational complexity of which increases exponentially with an

increase in the dimension of the BDS state vector. This study aims to apply parallel and distributed technologies to solve the problem of finding cycles of a given length in high-dimensional BDS. The searching cycles of a given length k requires to find all solutions. The authors are aware of only sequential SAT solvers able to find all solutions. Therefore developed new parallel solver HpcAllsat is used for solving the problem of searching cycles. The testing of HpcAllsat efficiency is carried out on the example of searching the short length cycles of stream ciphers based on shift registers. The study of solving this problem based on the procedural SAT approach is given in [5].

II. Related Work

In the past two decades, different approaches [6-12] was developed to simplify solving this problem. However, this problem is not fully resolved for large-scale BDS. In [7], the following areas of development of sequential attractor search algorithms were identified:

1. The method of enumeration and simulation [8]. The dimension of the BDS significantly limits the application of this method. In addition, this method is incomplete. It cannot guarantee the presence of all attractors of a given length [5].

2. The method of encoding Boolean function using BDD (Binary Decision Diagram) [9], the disadvantage of which consists in exponential growing the runtime with increasing the dimension of the BDS.

3. Algorithms using the SAT approach, in which transitions in BDS for a bounded number of steps is represented as a propositional formula [10]. The path corresponding to these transitions is found using the SAT solver. Then it is checked whether there is a closed trajectory in this path. The process is iteratively repeated for an increased number of steps. The efficiency of the algorithm mainly depends on the number of transition steps and the BDS dimension.

4. Aggregation approach [6], based on the decomposition of the structural graph of BDS into strongly coupled components (SCC). The aggregation approach uses the decomposition of the structural graph of BDS into SCC, the effectiveness of this approach substantially depends on the structure of the BDS. The SCC decomposition approach is difficult when the graph has a large amount of SCC, which usually occurs in biological networks [11]. In some cases, the divided subnets are still too large. However, for small enough subnet, the resulted attractor states of it may differ from the attractor states of the original network [12].

Parallel versions are also developed for some of the approaches listed above (for example, [6]). The aggregation approach is parallelized. However, the number of SCCs limits parallelization depth. The SAT-based algorithm [10] cannot be parallelized due to its inherent algorithm design [6]. Thus, despite active research in the field of a qualitative analysis of the BDS dynamics, the relevance of developing new and modifying existing methods for studying the behavior of high-dimensional BDS trajectories remains.

III. BCM Approach

Let $X = B^n$ ($B = \{0, 1\}$) be the set of binary vectors of dimension n (BDS state space). By $t \in T = (1,\ldots,k\}$ we denote the discrete time (number of time steps).

For each initial state $x^0 \in X$, let us define the trajectory $x(t, x^0)$ as a finite sequence of states $x^0, x^1, \ldots, x^k$ from the set X. Next, we consider a synchronous autonomous nonlinear BDS in which each pair of adjacent states x^t, x^{t-1} ($t \in T$) of the trajectory is connected by the relation

$$x^t = F(x^{t-1}) \tag{1}$$

Here $F : X \to X$ is a vector function of logic algebra, called the transition function. Thus for any $x^0 \in X$, the system of Boolean equations (1) represents a model of the dynamics of the BDS trajectories behavior in the state space X on a finite time interval $T = \{1,2,\ldots,k\}$. Hereafter, the value of k in the definition of the set T is assumed to be a predetermined constant.

The application of the Boolean constraint method for checking the dynamical property reduces to the next scheme:

- Constructing, using the equations of dynamics (1), a function of a one-step transition L, $L(x^{t-1}, x^t) = \vee_{i=1}^{n} (x_i^t \oplus F_i(x^{t-1}))$ [1];
- Propositional encoding of this function in the DIMACS format [13];
- The building the k-step transition function $\Phi(x^0, x^1, \ldots, x^k) = \vee_{t=1}^{k} L(x^{t-1}, x^t)$ [2] based on this propositional formula;
- Constructing a Boolean model of property by supplementing the function Φ with constraints corresponding to the specification of the property;
- Determining the feasibility of these Boolean constraints using, in depending on the property, the SAT or 2QBF solver (to obtain, if needed, the constructive solutions, and post-processing results in this case).

For example (Fig. 1), the above-described scheme is used for searching cycles of a given length $k = 3$ based on the BCM approach for an autonomous synchronous BDS [14], the dynamics equations of which are of the form

$$x_1^t = x_3^{t-1}, x_2^t = x_1^{t-1} \cdot x_4^{t-1}, x_3^t = x_4^{t-1}, x_4^t = x_1^{t-1}. \tag{2}$$

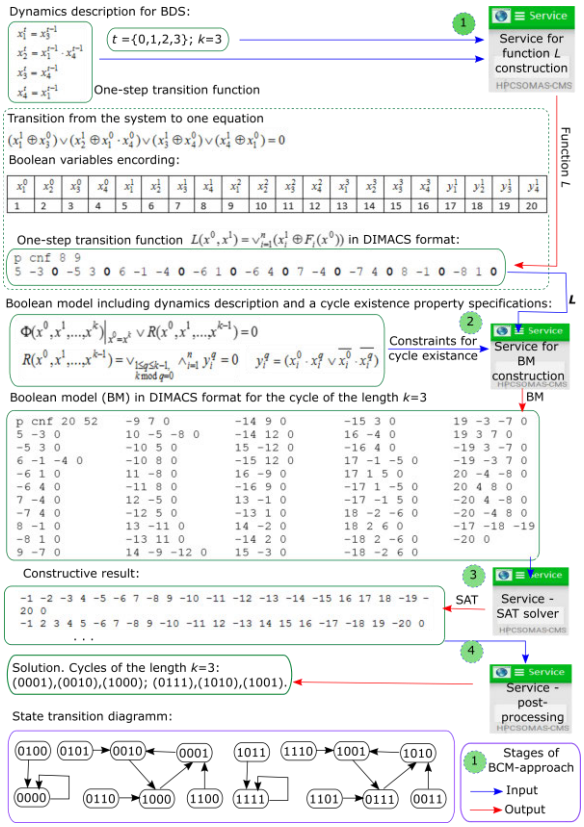


Figure 1. Search for cycles of length k=3 in BDS (2) based on the BCM approach

At the first stage of BCM-approach, the first service produces the function L of a one-step transition from state x^0 to state x^1 for the system (2), convert L to CNF, performs Boolean encoding of the function L in DIMACS format. At the second stage, the second service constructs the Boolean model by repeating function L with n-increasing indexes of variables for t from 2 to k and adding cycle existence condition R (Fig. 1). The indices $1, 2, \ldots, 16$ correspond to state variables, and indices $17, 18, 19, 20$ – to helper variables y_i^q of the condition R (Fig. 1) for $q = 1$. The automation technology for constructing a Boolean model of dynamic properties is described in detail in [3].

At the third stage, SAT-solver based service is used for verifying Boolean model satisfiability. Using the SAT solver, we make sure that the constraints are satisfiable, and the cyclic sequence of length $k = 3$ is found (Fig. 1).

A traditional SAT solver produces only one solution, a satisfying assignment of Boolean variables (constructive solution). All solutions SAT (AllSAT) is a variant of the propositional satisfiability problem [15]. In [15], advanced methods for solving the AllSAT-problem are considered, and sequential AllSAT solvers are described. Freely distributed versions of these solvers are available [16], in particular nbc_minisat_all-1.0.2 and bc_minisat_all-1.1.2. We use the sequential AllSAT solver nbc_minisat_all-1.0.2 if it is necessary to find all solutions when searching for all cycles of a given length k. However, the sequential solver can not solve this problem in an acceptable time for high-dimension BDS functioning in a long time interval k.

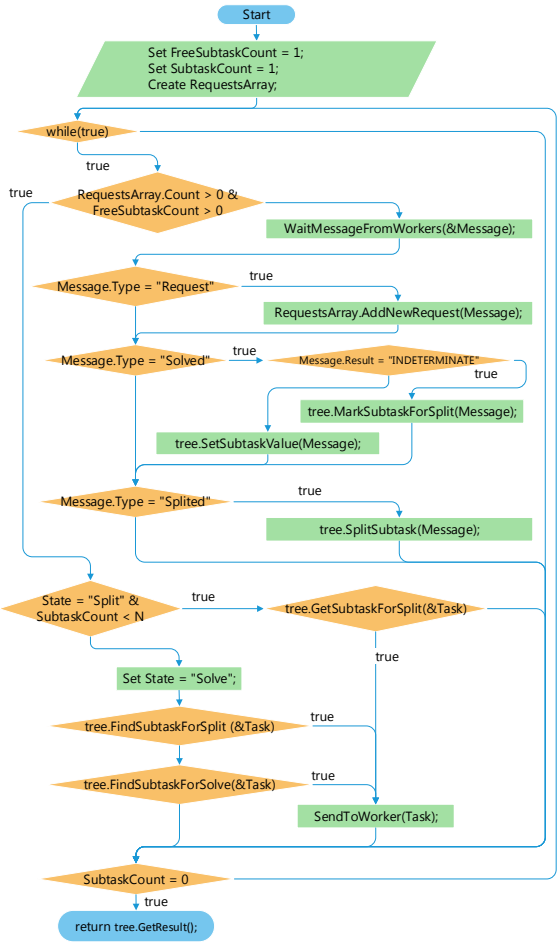


Figure 2. The algorithm of master process

IV. PARALLEL ALLSAT SOLVER

The parallel search of all cycles of a given length is performed to solve problems with a high-dimension state vector of BDS. To organize such a search, the HpcAllsat parallel solver is developed, which is a modified version of the Hpcsat author solver [17]. The solver HpcAllsat is implemented as an MPI application in C++ and is designed to find all solutions to the SAT of high dimension on computing clusters with SMP nodes running Linux. HpcAllsat is based on the search space partitioning approach between several sequential AllSAT solvers launched in slave processes. The algorithms of the master and slave processes are presented in Fig. 2 and Fig. 3.

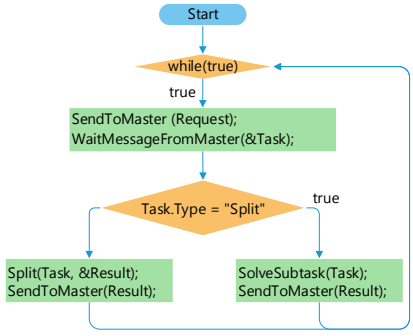


Figure 3. The algorithm of slave process

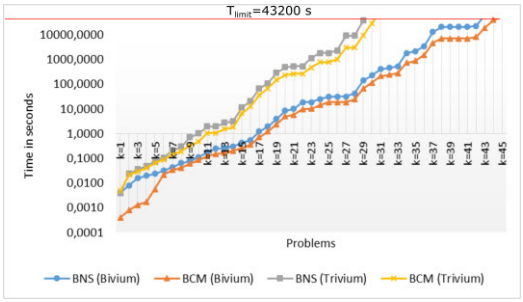


Figure 4. The comparison of BNS- and BCM-approachs

The master process organizes a dynamic queue of subtasks, forms a binary splitting tree (search tree), and adjusts the loading of slave processes. The tree node corresponds to the splitting variable and contains information about the status of the subtask solution for true and false values of this variable. These subtasks are obtained by splitting the initial Boolean property model depending on the accessibility of available computing resources. Master process alternates subtasks for simplifying submodels and verifying their satisfiability in response to the request of the slave process.

The master process goes into standby mode if free slave processes end when distributing subtasks. Else, the distribution of tasks continues, and the results are recorded in the answer file. The solver provides for the resumption of the splitting process if the sequential AllSat solver, executed in the slave process, has not completed the satisfiability test of the submodel (for example, in case of emergency termination of work or the time limit has expired). In such cases, new nodes are added to the search tree by the further splitting of this submodel.

V. Computational Experiment

The experiments are carried out on the high-performance computational cluster of the Irkutsk Supercomputer Center of the SB RAS [18]. The first experiment is carried out to compare the declarative BCM approach with the algorithm approach for the sequential solving of the problem of searching for cycles of length k in large-dimensional BDS. The Boolean model of the BDS property includes both the BDS dynamics equations and the specifications characterizing the property. The same solver can be used to verify the different BDS properties. Based on the algorithmic approach, each program is written to test only one. As a solver based on the algorithmic approach, we used the sequential BNS solver (Boolean Networks with Synchronous, freeware version 1.3 from February 2017 is available on the site [19]) based on the SAT approach [5, 10]. The search for cycles based on the BNS solver algorithm will be called the BNS approach in computational experiments. The structure of the solver algorithm, including an iterative cycle in which solver Minisat is launched, makes it difficult to parallelize [6]. Therefore, for an adequate comparison of declarative and algorithmic approaches, the computations in both cases were carried out sequentially. In the first experiment, the sequential solver nbc_minisat_all-1.0.2 is used as the ALLSAT solver in the BCM approach. The first experiment is performed on a PC with Intel Core i7-2670QM CPU 2.2 GHz×8.

As test problems for comparing the two approaches while searching for cycles of a given length k, we used BDS, which describes the dynamics of the Trivium and Bivium shift register-based stream ciphers with a high dimension of the state vector ($n = 288$ and $n = 177$, respectively). An analysis of these ciphers is given in [5]. The cycle search time for a given length k in the experiment was limited to twelve hours. For the Trivium cipher, the description of the dynamics has the form [5]:

$$x_1^t = x_{288}^{t-1} \oplus x_{287}^{t-1} \cdot x_{286}^{t-1} \oplus x_{243}^{t-1} \oplus x_{69}^{t-1}$$
$$x_{94}^t = x_{93}^{t-1} \oplus x_{92}^{t-1} \cdot x_{91}^{t-1} \oplus x_{171}^{t-1} \oplus x_{66}^{t-1}$$
$$x_{178}^t = x_{177}^{t-1} \oplus x_{176}^{t-1} \cdot x_{175}^{t-1} \oplus x_{264}^{t-1} \oplus x_{162}^{t-1}$$
$$x_i^t = x_{i-1}^{t-1},$$
$$i \in \{2,3,\ldots,93,95,\ldots,177,179,\ldots,288\}, t \in \{1,\ldots,k\}.$$

Bivium, a simplified version of the Trivium cipher, has a smaller dimension of the state vector. The dynamics equations of Bivium [5] have the following form:

$$x_1^t = x_{177}^{t-1} \oplus x_{176}^{t-1} \cdot x_{175}^{t-1} \oplus x_{69}^{t-1} \oplus x_{162}^{t-1}$$
$$x_{94}^t = x_{93}^{t-1} \oplus x_{92}^{t-1} \cdot x_{91}^{t-1} \oplus x_{171}^{t-1} \oplus x_{66}^{t-1}$$
$$x_i^t = x_{i-1}^{t-1},$$
$$i \in \{2,3,\ldots,93,95,\ldots,177\}, t \in \{1,\ldots,k\}.$$

The same sets of cycles of length k are found using both the BCM approach and the BNS solver. There are the following solutions:

- One cycle of length $k = \{1, 10, 11, 15\}$, two cycles of length $k = 12$, and 21 cycles of length $k = 3$ for Trivium;
- One cycle of length $k = 1$ and five cycles of length $k = 3$ for Bivium.

Fig. 4 shows the results (increasing total solution time) obtained for sequentially searching for cycles of length k based on the BCM approach, compared with the algorithmic one. Based on the algorithmic approach, it is not possible to exceed k =42 for Bivium and k = 29 for Trivium within the time allotted for solving all the problem (12 hours). Based on the BCM approach, we reached k = 44 for Bivium and k = 30 Trivium. New cycles for these values of k for both chippers are not found. Repeated cycles (cycles with the same set of states) are excluded during post-processing of the results of the solution in time O ($k \in K$), where K is the number of found cycles of length k. In comparison with runtime for solving problems, this time is insignificant. A further increase in the value of k leads to an excess of the time limit for sequential solving the problem.

The second experiment was to evaluate the effectiveness of the developed by the authors of the parallel ALLSAT-solver HpcAllsat. The search for cycles is carried out for Grain80, Bivium and Trivium on the first segment of the computational cluster “Akademik V.M.

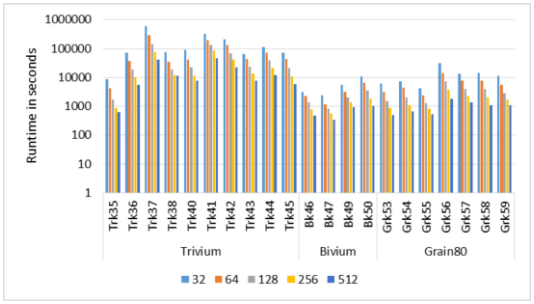


Figure 5. Runtimes of parallel solving

Matrosov", the node of which has two 16-core processors AMD Opteron 6276 " Bulldozer" / "Interlagos" with a frequency of 2.3 GHz. Description of the dynamics Grain80 has the following form [4]:

$$x_{80}^t = x_{64}^{t-1} \oplus x_{61}^{t-1} \oplus x_{53}^{t-1} \oplus x_{46}^{t-1} \oplus x_{38}^{t-1} \oplus x_{34}^{t-1} \oplus x_{29}^{t-1} \oplus$$
$$\oplus x_{22}^{t-1} \oplus x_{16}^{t-1} \oplus x_{10}^{t-1} \oplus x_{1}^{t-1} \oplus x_{64}^{t-1} \cdot x_{61}^{t-1} \oplus x_{38}^{t-1} \cdot x_{34}^{t-1} \oplus$$
$$\oplus x_{16}^{t-1} \cdot x_{10}^{t-1} \oplus x_{61}^{t-1} \cdot x_{53}^{t-1} \cdot x_{46}^{t-1} \oplus x_{34}^{t-1} \cdot x_{29}^{t-1} \cdot x_{22}^{t-1} \oplus$$
$$\oplus x_{64}^{t-1} \cdot x_{46}^{t-1} \cdot x_{29}^{t-1} x_{10}^{t-1} \oplus x_{61}^{t-1} \cdot x_{53}^{t-1} \cdot x_{38}^{t-1} \cdot x_{34}^{t-1} \oplus$$
$$\oplus x_{64}^{t-1} \cdot x_{61}^{t-1} \cdot x_{22}^{t-1} \cdot x_{16}^{t-1} \oplus x_{64}^{t-1} \cdot x_{61}^{t-1} \cdot x_{53}^{t-1} \cdot x_{46}^{t-1} \cdot x_{38}^{t-1} \oplus$$
$$\oplus x_{34}^{t-1} \cdot x_{29}^{t-1} \cdot x_{22}^{t-1} \cdot x_{16}^{t-1} \cdot x_{10}^{t-1} \oplus$$
$$\oplus x_{53}^{t-1} \cdot x_{46}^{t-1} \cdot x_{38}^{t-1} \cdot x_{34}^{t-1} \cdot x_{29}^{t-1} \cdot x_{22}^{t-1},$$
$$x_i^t = x_{i+1}^{t-1}, i \in /\{1,2,..,79\}.$$

The runtime (T), speedup (S) and efficiency (E) are calculated using the formulas:

$$S(p) = \frac{T_{32} \cdot 32}{T_p}, E(p) = \frac{S(p)}{p}, T^{av}(p) = \frac{1}{N}\sum_{i=1}^{N} T_i(p),$$
$$S^{av}(p) = \frac{1}{N}\sum_{i=1}^{N} S_i(p), E^{av}(p) = \frac{1}{N}\sum_{i=1}^{N} E_i(p)$$

where N is a number of test tasks, T_{32} and T_p are the time to solve the problem on, respectively, the 32 and p processor cores; T^{av} , S^{av} and E^{av} are the average time and speedup to solve one problem, and the average efficiency of a load of processor cores, accordingly. The 32 processor cores are taken as the basic unit since this value is the minimum of cores number, on which the most difficult test tasks could be solved within the limited time interval. A similar approach to the calculation of these



Figure 7. The service for searching k length cycles

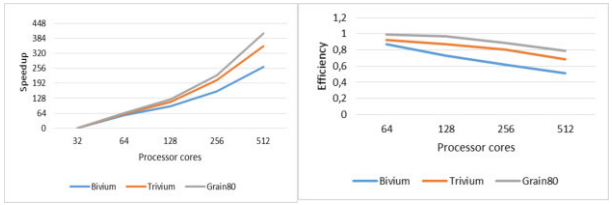


Figure 6. The average speedup and efficiency of HpcAllsat

characteristics is described in [20]. The results of evaluating the runtime with an increase in the number of processor cores are shown in Fig. 5. The decreasing runtime is close to linear for all test problems. Results of the average speedup and efficiency for solver HpcAllsat are shown in Fig. 6. Overall, the results of testing a parallel solver showed speedup close to linear, and efficiency lies within acceptable limits.

In the third experiment, a distributed solution is tested using a composite service to search for cycles of a given length k. The nodes of three computing clusters and the cloud resource [21] on which the service is installed are used as a hybrid environment (Fig. 7). In the experiment, the value "4" is set in the Split field for preliminary splitting into 16 submodels. Though the heterogeneous nature of the computational environment affects some decrease in efficiency (as a rule, accompanying a significant increase in nodes) (Fig. 8), the runtime is significant decreases with increasing the number of processor cores (Fig. 9). Experiments show that in parallel solving the problem on the SMP cluster, the k cycle length can be further increased by increasing the number of processor cores. The presented approach ensures that all cycles of a given length k are found in the state space of stream ciphers based on shift registers. As noted in [5], taking into account the fact that the state spaces in the problems being solved have dimensions 2^{288} and 2^{177}, the problem of finding all cycles of a given length is not

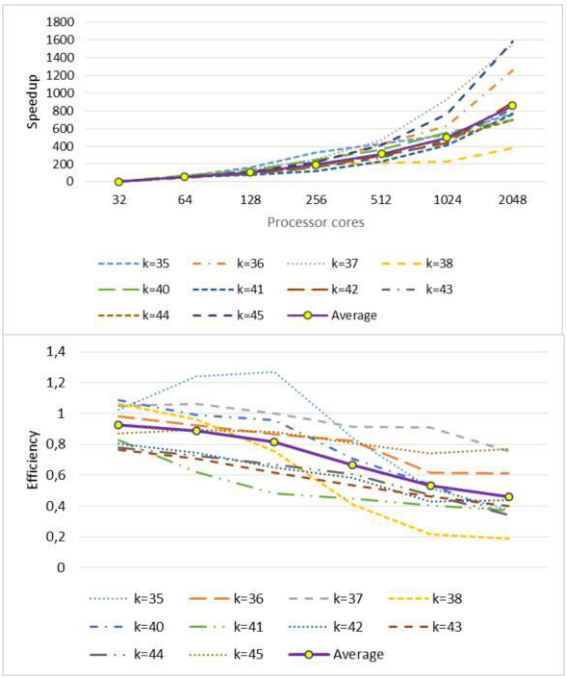


Figure 8. The speedup and efficiency of Trivium distributed solving

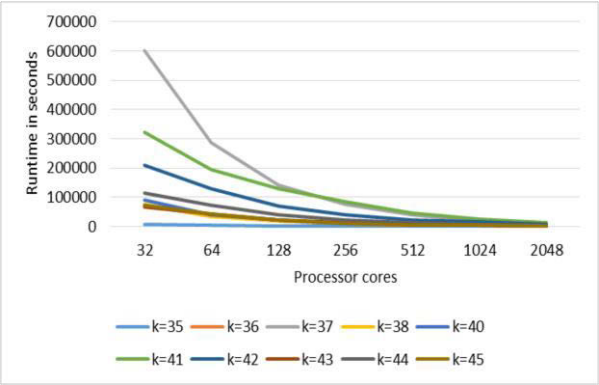


Figure 9. The runtime of Trivium distributed solving

trivial. The proposed declarative approach is also applicable to testing other dynamic properties of BDS [2] including BDS with control actions [22].

VI. Conclusion

Based on the BCM approach, a parallel solver is developed to verify the feasibility of the Boolean model of the investigated property of the binary dynamic system with the construction of all or a given number of constructive solutions. The parallel solver is designed to solve the problems of the qualitative analysis of binary dynamic system based on the Boolean constraint method and is focused on application in a high-performance cluster environment. This solver is implemented as a microservice. A composite service is developed for organizing two-level parallelization in solving problems of qualitative analysis in a hybrid environment. Computational experiments are carried out on the example of studying the properties of the presence of a cycle of a given length in a high-dimension binary dynamic system and confirmed the operability and effectiveness of the developed software tools. These tools can significantly increase the dimension of the state vector of the binary dynamic system and the time interval of its functioning when solving these problems in a high-performance computational environment.

Acknowledgment

The present investigation was supported by the Russian Foundation of Basic Research, projects no. 18-07-00596.

References

[1] G.A. Oparin, V.G. Bogdanova, A.A. Pashinin, and S.A. Gorsky, “Microservice-oriented Approach to Automation of Distributed Scientific Computations,” in Proceedings of the 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 2019, P. 236-241.

[2] G. Oparin, V. Bogdanova, and A.Pashinin, “Qualitative analysis of autonomous synchronous binary dynamic systems,” in Mathematics in engineering, science and aerospace (MESA), Cambridge, UK, I&S - Florida, USA, vol. 10, no. 3, 2019 pp. 407-419.

[3] V.G. Bogdanova and S.A. Gorsky, “Multiagent Technology for Parallel Implementation of Boolean Constraint Method for Qualitative Analysis of Binary Dynamic Systems,” in Proceedings of 42st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2019, pp. 1227-1232.

[4] Oparin G.A., Bogdanova V.G., Pashinin A.A. , Gorsky S.A. Distributed solvers of applied problems based on microservices and agent networks // 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2018, pp. 1415-1420.

[5] E. Dubrova and M. Teslenko, “A SAT-Based Algorithm for Finding Short Cycles in Shift Register Based Stream Ciphers,” in IACR Cryptology ePrint Archive. 2016, p. 1068.

[6] W. Guo, G. Yang, W. Wu, L. He, and M. Sun, “A parallel attractor finding algorithm based on boolean satisfiability for genetic regulatory networks,” in PLoS One, Vol. 9(4):e94258, 2014.

[7] A. Mizera, J. Pang, H. Qu, and Q. Yuan “A new decomposition method for attractor detection in large synchronous boolean networks,” in Dependable Software Engineering. Theories, Tools, and Applications, SETTA 2017, Lecture Notes in Computer Science. Springer, Cham, vol. 10606, 2017, pp. 232-249.

[8] R. Somogyi and L.D. Greller, “The dynamics of molecular networks: applications to therapeutic discovery,” in Drug discovery today, vol. 6(24), 2002, pp. 1267–1277.

[9] A. Naldi, D. Thieffry, and C.Chaouiya, “Decision diagrams for the representation and analysis of logical models of genetic networks,” In Computational Methods in Systems Biology. CMSB 2007. Lecture Notes in Computer Science, vol 4695, M. Calder and S. Gilmore, Eds. Springer, Berlin, Heidelberg, 2007.

[10] E. Dubrova and M. Teslenko, “A SAT-based algorithm for finding attractors in synchronous Boolean networks,” in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 5, 2011, pp. 1393–1399.

[11] C. Su, J. Pang, and S. Paul, “Towards Optimal Decomposition of Boolean Networks,” in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 14, no. 8, 2015.

[12] S. Choo and K. Cho, “An efficient algorithm for identifying primary phenotype attractors of a large-scale Boolean network,” in BMC Systems Biology, vol. 10, 2016, article number 95.

[13] “Satisfiability suggested format,” http://beyondnp.org/static/media/uploads/docs/satformat.pdf [online, accessed: 11.12.2019].

[14] O. Colon-Reyes, R. Laubenbacher, and B. Pareigis, “Boolean Monomial Dynamical Systems,” arXiv:math/0403166v1.

[15] T. Toda and T. Soh, “Implementing Efficient All Solutions SAT Solvers,” in ACM Journal of Experimental Algorithmics, vol. 21(1), 2016, pp. 1-44.

[16] “All Solutions SAT Repository,” http://www.sd.is.uec.ac.jp/toda/code/allsat.html [online, accessed: 11.12.2019].

[17] V.G. Bogdanova and S. A. Gorsky, “Scalable parallel solver of Boolean satisfiability problems,” in Proceedings of 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, pp. 0222-0227, 2018.

[18] “Irkutsk Supercomputer Centre of SB RAS,” http://hpc.icc.ru [online, accessed: January 31 2019].

[19] “Boolean Networks with Synchronous,” https://omictools.com/bns-tool (online, accessed: 30.11.2019).

[20] A. Kishimoto, A. Fukunaga, and A. Botea, “Evaluation of a simple, scalable, parallel best-first search strategy,” in Artificial Intelligence, vol. 195, 2013, pp. 222–248.

[21] https://firstvds.ru/ (accessed: 24-Jan-2020).

[22] G. Oparin, V. Bogdanova, S. Gorsky, and A. Pashinin, “The synthesis of stabilizing feedback for binary dynamic systems: a logical approach,” in Mathematics in engineering, science and aerospace (MESA), Cambridge, UK; I&S - Florida, USA, vol. 10, no. 3, 2019, pp. 477-486. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

SOFTWARE DESIGN AND AUTOMATION IN INDUSTRY

Sample Based Synthesis of Car Engine Noise

Dubravko Miljković
Hrvatska elektroprivreda, Zagreb, Croatia
dubravko.miljkovic@hep.hr

***Abstract* - A recent trend in electric cars introduces new problem due to their silent movement. Electric cars are much quieter than conventional cars equipped with internal combustion engines. This lack of noise poses a danger to other road users. To increase the safety of pedestrians, under certain conditions, artificial noise generators must be used, as required by EU regulations. Car engine noise synthesis is proposed that combines sample based synthesis and envelope scaling dependent on engine rotational speed.**

Keywords - car; engine; noise; synthesis

I. INTRODUCTION

Electric vehicles are much quieter than conventional counterparts and as such pose an increased danger to pedestrians that can no longer rely on noise to determine distance and speed of approaching vehicles, [1]. This situation is quite common at low speeds encountered mostly in town centers and car parking lots, [2]. When such vehicles operate in an urban environment at low speeds, fake noise that warns pedestrians should be added with an Acoustic Vehicle Alerting System (AVAS), Fig. 1. In order to improve pedestrian safety, recent regulations require all full-electrical (EVs), hybrid-electrical (HEVs) and plug-in hybrid electric vehicles (PHEVs) vehicles to be equipped with AVAS. From July 1, 2019, all new types of electric and hybrid vehicles in the European Union (EU) must be provided with AVAS that emits a continuous noise of at least 56 dBA at 2 m distance whenever an electric vehicle is reversing or moving at speeds under 20 kilometers per hour (12 miles per hour), [3]. Such systems should also be retrofitted to existing electric vehicles in EU by 2021. Most systems produce sounds that mimic conventional cars with internal combustion engines, however, [4-10], more complex sounds like a futuristic melody ('Sci-fi spacecraft' sounds) that change pitch in response to vehicle speed



Figure 1. Electric vehicle with AVAS (modified John Voo photo)

(Nissan) are also available, [8]. After analysis of vehicle noises, method for synthesis of engine noise is outlined.

II. CAR NOISE

Car noise originates from the following main sources: tires, powertrain (motor/engine and transmission), intake orifice and exhaust orifice. Various sub-sources of car noise are shown in Fig. 2., [11]. Contribution of sub-sources in relation to vehicle speed is shown in Fig. 3, [12]. Tire/pavement and aerodynamic noise are monotonically increasing with the speed, while propulsion noise is dependent on the engine rotational speed and gear shifting.

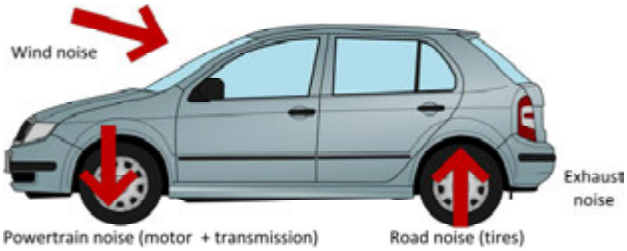


Figure 2. Main sources of car noise

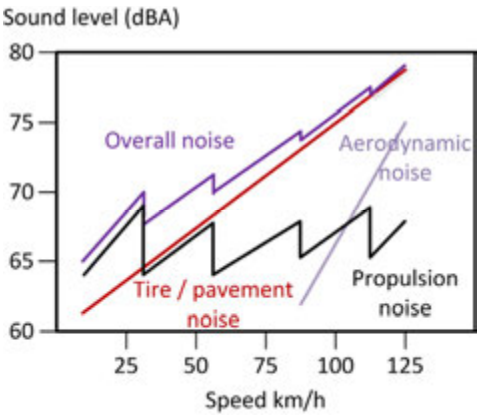


Figure 3. Contribution of sub-sources of car noise, modified from [12]

Noise produced by electric vehicles consists of all these elements with the main difference that it uses an electric motor instead of internal combustion engine. They also lack gearbox (use fixed gear instead). Due to the low noise levels electric vehicles, if not equipped with AVAS, can mask themselves among other traffic. According to measurements described in [13], the mean street noise level was 73.4 dBA, with substantial spatial variation in a range 55.8 - 95.0 dBA. One example of the recorded envelope of street noise in a busy street, [14], is shown in Fig. 4 with the corresponding spectrogram shown in Fig. 5. The sound envelope reveals the noise dynamics and spectrogram relative contributions of spectral components.

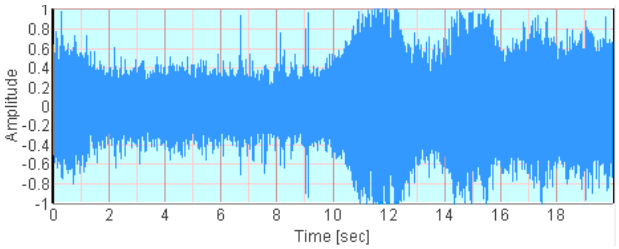


Figure 4. Envelope of street noise

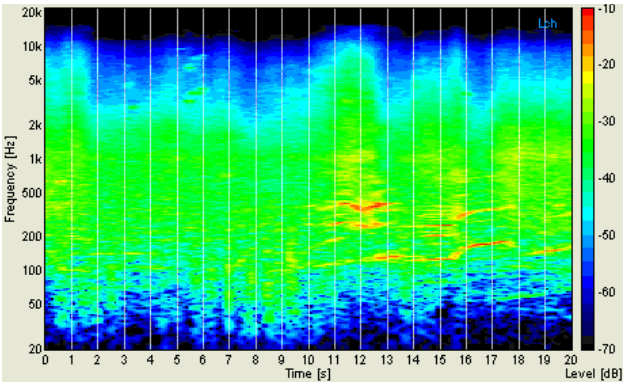


Figure 5. Spectrogram of street noise

III. ENGINE NOISE

A. *Noise of Electric Motor*

The most common electric motor used in electric vehicles is a permanent magnet synchronous motor (PMSM). It is a synchronous electric motor whose inductor consists of permanent magnets. The basic sources of noise of electric motors are electromagnetic, mechanical and aerodynamic source, [15]. However, the dominant noise is generated by the vibration of its machine parts. The sound envelope of an electric motor used in a Tesla car during spin up, rotation and spin down, [16], is shown in Fig. 6 and the related spectrogram in Fig. 7.

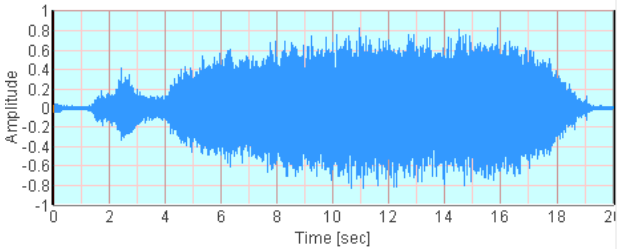


Figure 6. Tesla motor noise envelope (spin up and down)

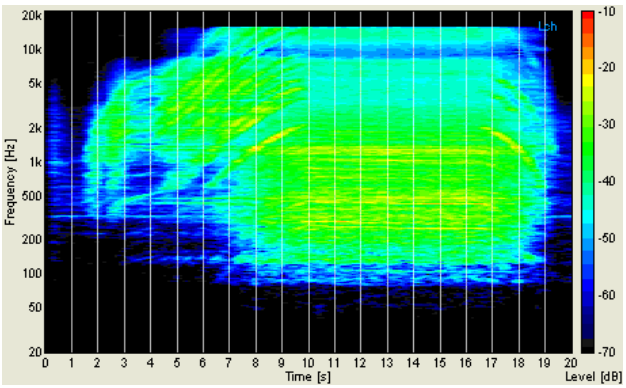


Figure 7. Tesla motor noise spectrogram (spin up and down)



Figure 8. Audi e-tron electric SUV

One example of an EV with AVAS is Audi e-tron SUV, Fig. 8. Synthesized AVAS engine noise, [17], envelope is shown in Fig. 9 and the related spectrogram in Fig. 10.

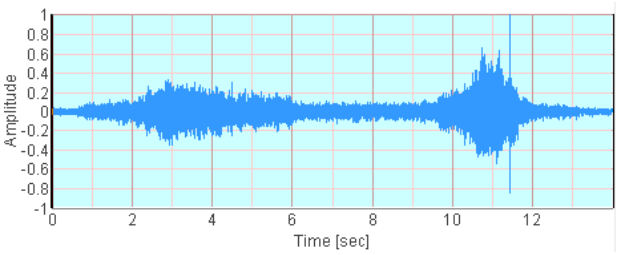


Figure 9. Audi e-tron noise envelope (slow forward motion and reverse)

Please note that in this example sound generated while moving in reverse is about 6 dB louder than during low speed forward motion.

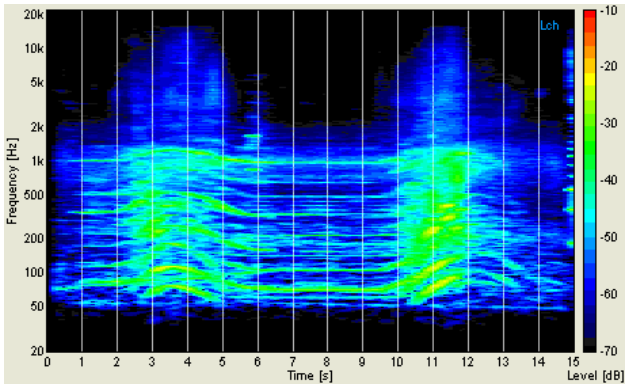


Figure 10. Audi e-tron noise spectrogram (slow forward motion and reverse)



Figure 11. Nissan Leaf electric car

Another example of AVAS synthesized engine noise is the one generated by Nissan Leaf electric car, Fig. 11. It is kind of previously mentioned futuristic sound, [8-10]. The sound envelope of intermittent noise in a reverse motion (modern variant of a back-up beeper), [18], is shown in Fig. 12 and spectrogram in Fig. 13. Pitch is not changed.

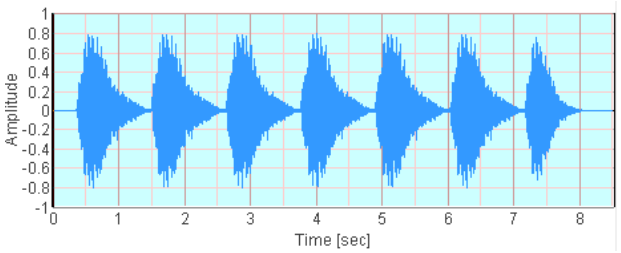


Figure 12. Nissan Leaf noise envelope (reverse)

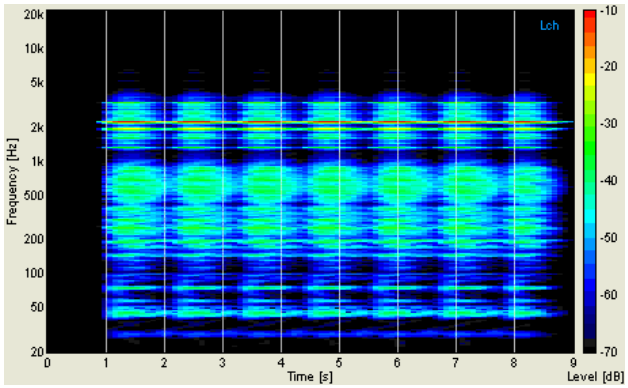


Figure 13. Nissan Leaf noise spectrogram (reverse)

Such kind of short repeating sounds may be best produced by replaying complete recorded segments, instead of messing up with a more complicated synthesis.

B. *Noise of Internal Combustion Engine*

Conventional cars use petrol and diesel engines. Most common are engines with 4 and 6-cylinders, although sometimes engines with 3, 5, 8 or even 12 cylinders may be found. Cylinder firing rate f_{CFR} is dependent on the engine rotational speed N_{RPM} (Revolutions Per Minute) and for 4-cycle engines is given by (1), [19]:

$$f_{CFR} = \frac{N_{RPM}}{120} \tag{1}$$

Engine firing rate f_{EFR} is dependent on the f_{CFR} and the number of cylinders N_C, (2):

$$f_{EFR} = N_C f_{CFR} \tag{2}$$



Figure 14. Mercedes 63 AMG

The sound envelope and spectrogram of an internal combustion motor used in Mercedes 63 AMG, Fig. 14, with twin-turbo 4.0-liter V-8 engine, [20], recorded during driving and gear changes are shown in Fig. 15 and Fig. 16.

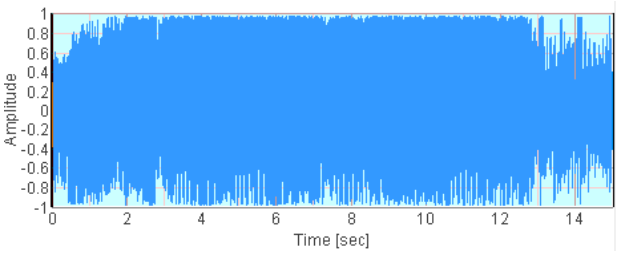


Figure 15. Mercedes 63 AMG noise envelope (gear changes)

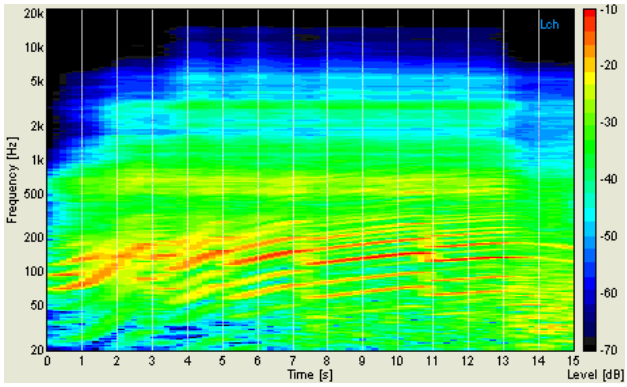


Figure 16. Mercedes 63 AMG noise spectrogram (gear changes)

- Cycle-to-Cycle Variations

Cycle-to-cycle variations of the engine combustion noise are particularly pronounced at idle and low RPMs. Envelope of such variations, [20], is shown in Fig. 17 and spectrogram in Fig. 18.

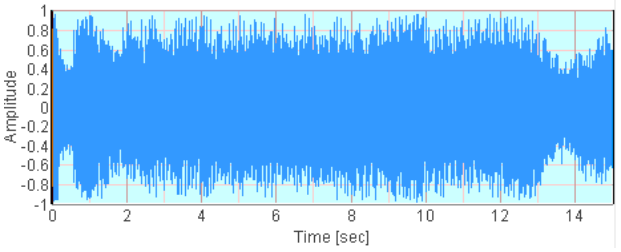


Figure 17. Mercedes 63 AMG idle noise envelope

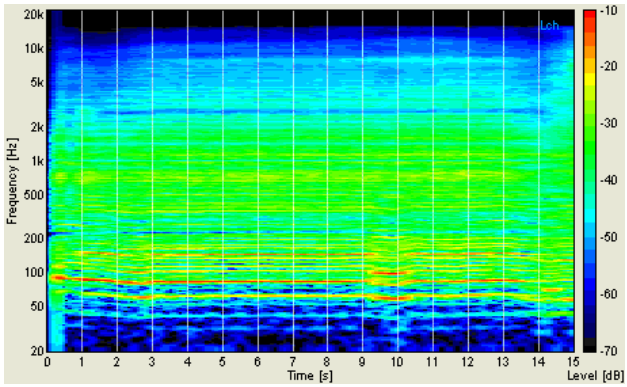


Figure 18. Mercedes 63 AMG idle noise spectrogram

IV. Sample Based Synthesis

Sample-based synthesis is a type of audio synthesis that uses samples stored in a wavetable. It uses sampled sounds in the form of very short recordings (resulting in compact storage requirements) instead of audio synthesis with oscillators producing fundamental waveforms like

square, saw or triangle, [21]. As such, it can provide a specific sound profile for each vehicle. Three main issues to address in sample-based synthesis are looping, pitch-shifting and data reduction, [21]. To reproduce the cycle endlessly, the system simply needs to loop selected stored waveform block, [22]. Sound synthesis can be performed with fixed and variable sample rate.

A. *Fixed Sampling Rates*

The fixed sample rate is simpler for simulation on a PC computer, as it will work on commonly used sound cards that support several sampling rates. It is suitable for software implementation in PC games and simulations. However, resolution in achieved pitch changes is coarse, because it involves upscaling and dropping samples from a waveform. For example, dropping every 21th sample in a recording raises the pitch for 5 %.

B. *Variable Sampling Rate*

With the variable sample rate it is much easier to achieve precise pitch in relation to the input RPM parameter, because the time separation between successive samples can be controlled to a much higher precision. The time period between successive samples T_D is given by (3) where f_c is the clock frequency.

$$T_D = \frac{f_c}{N_{RPM}} \tag{3}$$

For example, for twice the pitch frequency time period between samples is divided by two. This is suitable for use on microcontroller with D/A converter. However, this is not easy to simulate on the PC with sound card that use fixed sampling rate streams. It can be done by choosing the highest sampling rate and repeating samples in a stream. Similar coarse pitch resolution problem appears as with the fixed sample rate. Variable sample rate also complicates low-pass filtering of synthesized signal and analog switch-capacitor low pass filter solution can be used instead.

C. *Choice of Sampling Rate*

The spectral content of Mercedes twin-turbo 4.0-liter V-8 engine petrol engine noise, [20], averaged over the 10 s period is shown in Fig. 19. There is not much acoustic energy present in engine noise bellow frequency of few kHz. Hence, 8 kHz sampling rate for 0-4 kHz frequency range and 8-bit quantization for 40 dB dynamics in previous frequency range should be sufficient, particularly if amplitude normalization of the recorded signal is used.

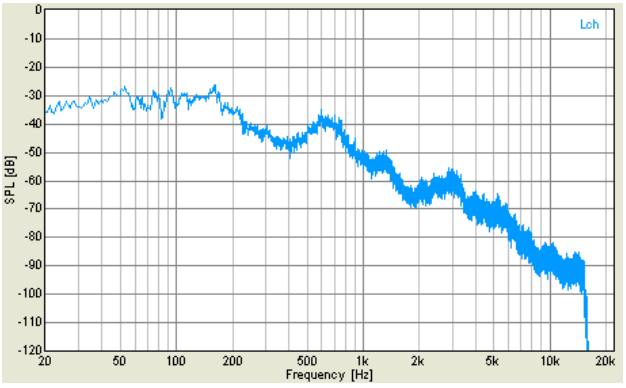


Figure 19. Engine noise spectral content

D. *Pitch Shifting Using Variable Sampling Rate*

When motor RPM changes, the pitch of the synthesized sound needs to be changed by pitch shifting, [23, 24]. The pitch shifting principle is illustrated in Fig. 20. Because electric vehicles generally don't use a gearbox, RPM scaling from the electric motor (operating in a 0 - 16000 RPM range) to conventional motor (0 - 7000 RPM) is necessary, as shown in (4), where N_{RPM} is the rotational speed of the conventional engine, N_{eRPM} of an electric motor and k is the scaling factor ($k \approx 0{,}44$).

$$N_{RPM} = kN_{eRPM} \tag{4}$$

Depending on the input RPM, time between successive samples (variable sampling rate) is calculated or fetched from the previously calculated table. These values are used in frequency divider that feeds a counter that address sample memory. When this counter reaches the number of samples of a particular recording, it is reset to zero, and the loop repeats. This is given by (5) where $w(t)$ represents samples stored in a wavetable, and N is the number of samples in wavetable for a particular recording.

$$y(t) = w(t \bmod N) \tag{5}$$

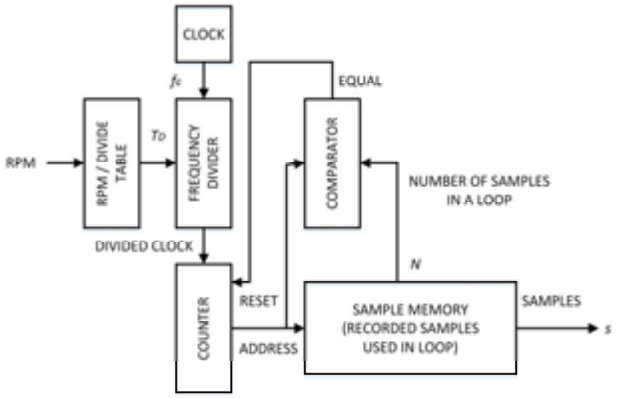


Figure 20. Block diagram describing pitch shifting

Block diagram only describes the principle used for pitch shifting. In real synthesizer, building blocks are software implemented as software in a microcontroller.

E. *Synthesis of Engine Noise Using Multiple Recordings*

Input to synthesizer is electric motor RPM, that is, due to the lack of gearbox, also related to a vehicle speed. Noise at various RPMs differs in noise level, but also in spectral content. If one uses single recording for all RPMs, unnatural sound may be produced. For more realistic engine noise, different recordings, [25], should be played in a loop, depending on the motor RPM. Proposals for RPM ranges for recordings are shown in Table I. If one needs to synthesize noise only for speeds less than 20 mph, than just two recordings may suffice. Accurate transition RPMs have to be determined from particular recordings,

TABLE I. RECORDINGS AND CORESPONDING RPM RANGES

Recording	Engine RPM
1	500 - 1000
2	1000 - 2000
3	2000 - 4000
4	4000 - 6000
5	6000+

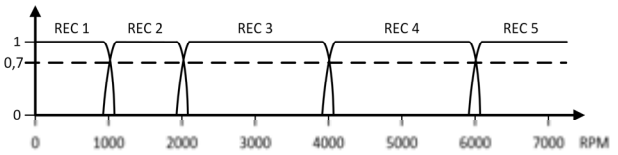


Figure 21. Transitions of noise recordings at various RPM

as it is difficult to record sound at a precise RPM. For example, instead of 1000 RPM, limit may be 988 RPM. To ensure smooth transitions between recordings, equal power crossfading is used, Fig. 21, with crossovers at 0.707 and a squared sum of fade functions of 1 due to the RMS addition of two signals. Transition regions can be stored in a table. The diagram in Fig. 22 shows principle only, with synthesis loops and fading for all recordings.

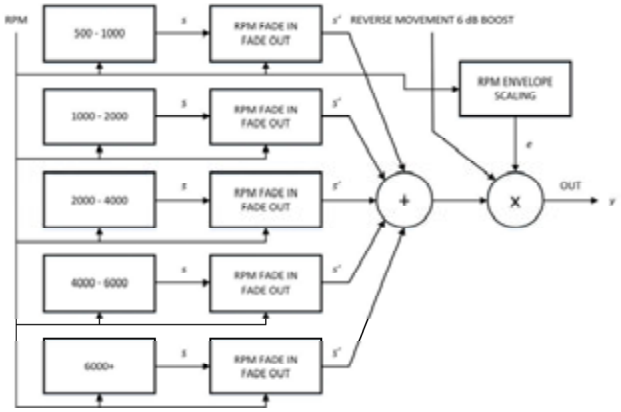


Figure 22. Engine noise synthesizer

Actually, only two synthesis channels are necessary to perform fade-in and fade-out at a given moment, Fig. 23.

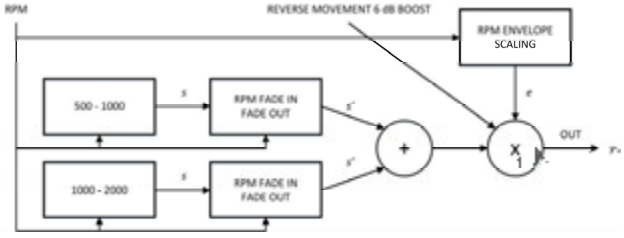


Figure 23. Fade-in and fade-out for two channels

F. *Enevelope Scaling*

Engine noise is very dependent on engine RPM. Noise level in relation to vehicle speed for conventional cars with internal combustion engine is illustrated in Fig. 24, [26]. One can note intervals corresponding to successive gear

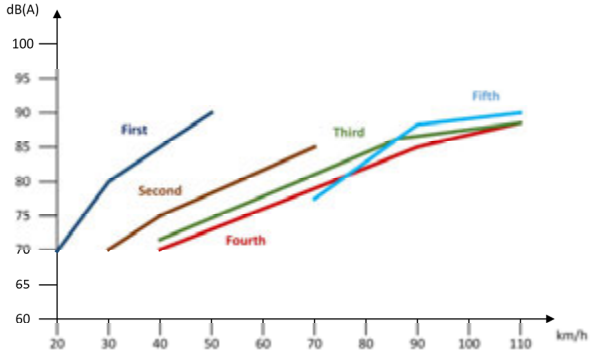


Figure 24. Noise levels vs. vehicle speed and gears, according to [26]

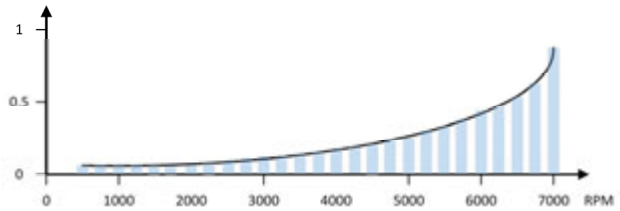


Figure 25. Engine noise envelope estimate

shifts. Electric cars generally don't use a gearbox. Hence, the relation of speed and corresponding RPM to the noise level is continuous, but nonlinear, Fig. 25. To incorporate noise level dynamics into the synthesis, envelope scaling is used. Because separate engine noise recordings are amplitude normalized, envelope scaling is used to adjust the overall noise level of the synthesized sound. Envelope scaling values are stored in a separate table with scaling values corresponding to RPM values within particular intervals, (6), where $e[l]$ represents the value of the noise envelope in l-th RPM interval. These intervals are smaller than the ones used for fade-in and fade-out of synthesized signal, e.g., every 100 RPM, hence $I_{RPM} = 100$.

$$l = \frac{N_{RPM} - N_{RPM,idle}}{I_{RPM}} \tag{6}$$

Accurate discrete values $e[l]$ for particular RPM interval are obtained by measuring noise levels. Values of $e[l]$ are then normalized by dividing with the maximal value e_{max} among all $e[l]$, (7), (8).

$$e_{max} = \max_{i=0 \ldots l-1} e[i] \tag{7}$$

$$e_n[i] = \frac{e[i]}{e_{max}} \tag{8}$$

Envelope scaling of faded samples $s'(n)$ is given by (9).

$$y(n) = e_n[l]s'(n) \tag{9}$$

Block diagram of a controller together with a power amplifier and speaker is shown in Fig. 26.

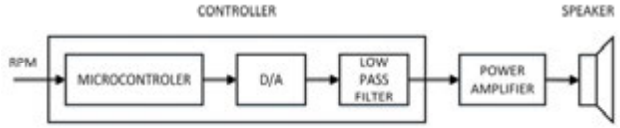


Figure 26. Controller with power amplifier and speaker

Low pass filtering can be efficiently accomplished using switch capacitor technology. Sound is synthesized by a small control unit and emitted by a speaker behind the front grille. The power required from an amplifier is dependent on the sensitivity of used loudspeaker.

V. Preparing and Editing of Recorded Sound

Recordings are, naturally, made with fixed sampling rate. When editing the recorded signal, it should be assured that the recording starts and ends with approximately zero amplitude level, i.e. with no abrupt changes in amplitude of the waveform at the beginning and the end of recording. Recording should also not be cut in a middle of a pitch period. This helps to remove or reduce discontinuities that may be pronounced when looping recorded signal, particularly in short recordings (less than a few seconds).

Also, it incorporates engine cycle-to-cycle variations that are pronounced at lower RPMs. It is wise to invest effort in editing noise recordings, to obtain recordings suitable for playing in a loop, instead of complicating synthesizer with overlapping and windowing of recordings. After selecting the appropriate recording segment, amplitude normalization is performed. In this way all recordings are of the same amplitude that is later modified by crossfading and envelope scaling.

VI. Synthesis Results

A 15 s 1000-6000 RPMs linear sweep of conventional engine is generated. Pitch shifting using a variable sampling rate was simulated by upsampling to 192 kHz, repeating samples variable number of times and down-sampling to 8 kHz. Envelope of the synthesized signal is shown in Fig. 27 and spectrum in Fig. 28. Frequency range on spectrogram is to 0-4 kHz (due to downsampled signal).

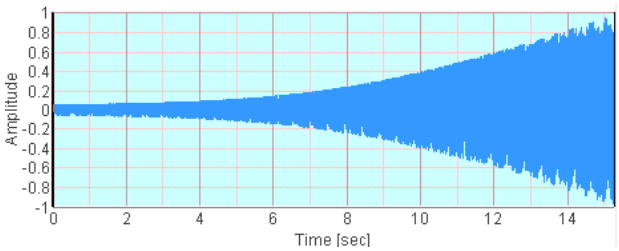


Figure 27. Envelope of synthesized engine noise

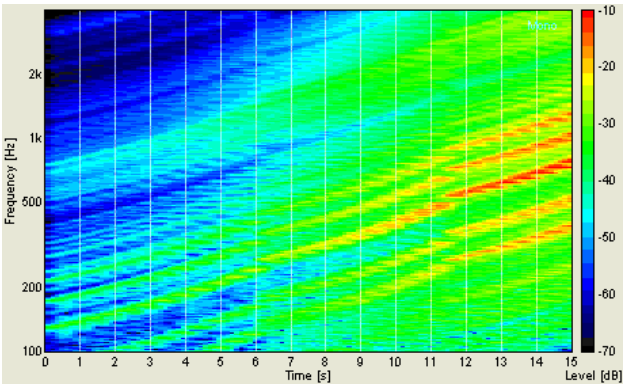


Figure 28. Spectrogram of synthesized engine noise

VII. Conclusion

Silent all-electrical and hybrid vehicles pose an increased danger to pedestrians. To alleviate this problem, new regulations require use of AVAS when vehicle is moving at slow speed or reversing. Presented AVAS that synthesize noise of a conventional engine noise makes use of multiple recordings for various RPMs to add realism to synthesized sound. Individual recordings of several seconds have been used to encompass cycle-to-cycle variations of engine noise and further aiding to realism. Variable sample rate is used that provide precise pitch shifting and is suitable for implementing on microcontrollers with D/A converter. Envelope dynamics is incorporated by RPM dependent envelope scaling. Described AVAS is sufficiently simple that, considering CPU power and memory requirements for program and stored recordings, can be realized on almost any modern 16-bit microcontroller with sufficient ROM (for recordings) and an addition of 12 bit D/A converter.

References

[1] D. Norman, "What noise does the electric car make?," MIT Technology Review, Feb 7, 2014, [Online]. Available: https://www.technologyreview.com/s/524241/what-noise-does-the-electric-car-make/ [Accessed Sept. 11, 2020].

[2] D. Mrvar, "EV Sound," YouTube, 6 Oct 2016 [Video file]. Available: https://www.youtube.com/watch?v=OMJObfCSF5M [Accessed Sept. 11, 2020].

[3] "Electric vehicle warning sounds," in Wikipedia, https://en.wikipedia.org/wiki/Electric_vehicle_warning_sounds [Accessed Sept. 11, 2020].

[4] The Wheel Network, "Mercedes Benz e-sound – Acoustic Vehicle Alerting System (AVAS) for Electric Cars," YouTube, 15 Jul 2019 [Video file]. Available: https://www.youtube.com/watch?v=k82rjRxl6Gk [Accessed Sept. 11, 2020].

[5] Harman, Halosonic, Active Noise Management Solutions, [Online]. Available: https://www.halosonic.co.uk/ [Accessed Sept. 11, 2020].

[6] TheSoundRacer, "SoundRacer Electric Vehicle Sound Module," YouTube 12 Feb 2013. [Video file]. Available: https://www.youtube.com/watch?v=nRWmuy-BOtQ [Accessed Sept. 11, 2020].

[7] D. Min, B. Park. and J. Park, "Artificial engine sound synthesis method for modification of the acoustic characteristics of electric vehicles," Hindawi Shock and Vibration, Vol. 2018

[8] Nissan, "Nissan sings with debut of 'Canto' at 2017 Tokyo Motor Show," YouTube, 24 Oct 2017 [Video file]. Available: https://www.youtube.com/watch?v=p4cjPk8mqEM [Accessed Sept. 11, 2020].

[9] Autotrader, "Here's What The Nissan LEAF PLUS Electric Motor Actually Sounds Like," YouTube, 15 May 2019 [Video file]. Available: https://www.youtube.com/watch?v=bq1tlQZJDgY [Accessed Sept. 11, 2020].

[10] Hybrid, "EV sound of Nissan Leaf," YouTube, 17 May 2019 [Video file]. Available: https://www.youtube.com/watch?v=j9zQ9vyc7kc [Accessed Sept. 11, 2020].

[11] In-Vehicle Powertrain Noise Evaluation, DISCOM Mobile, Brüel & Kjær DISCOM, [Online]. Available: https://www.bksv.com/-/media/Files/products/production-testing-systems/DISCOM/BG1958-11_Brochure_DISCOM-Mobile.ashx [Accessed Sept. 11, 2020].

[12] R. O. Rasmussen, R. J. Bernhard, U. Sandberg and E. P. Mun, The Little Book of Quieter Pavements, Rep. FHWA-IF-08-004, July 2007, [Online]. Available: https://www.thetranstecgroup.com/LittleBookQuieterPavements.pdf [Accessed Sept. 11, 2020].

[13] T. P. McAlexander, R. RM Gershon and R. L, "Street-level noise in an urban setting: assessment and contribution to personal exposure," Environ Healthv.14; 2015, PMC4350859

[14] TexasHighDef, "Street Sounds 2 "Road Nosie"", YouTube, 17 Dec 2010 [Video file]. Available: https://www.youtube.com/watch?v=jek6V2UlMUo [Accessed Sept. 11, 2020].

[15] M. Janda, O. Vitek and V. Hajek, Noise of Induction Machines, [Online]. Available: https://www.intechopen.com/books/induction-motors-modelling-and-control/noise-of-induction-machines [Accessed Sept. 11, 2020].

[16] J. Hughes, "Tesla Motor Test Bench - QA/Sale Prep (Custom control package),"YouTube, 22 Feb 2017 [Video file]. Available: https://www.youtube.com/watch?v=_Rg75JbVOpg [Accessed Sept. 11, 2020].

[17] Car Tech Connect, "Audi e-tron Artificial Engine Sound!! (Sounds like a Spaceship?)," YouTube, 18 May 2019 [Video file]. Available: https://www.youtube.com/watch?v=MUcbXplAvSc [Accessed Sept. 11, 2020].

[18] GMVolt, "Nissan LEAF Reverse Sound," YouTube, 11 Jun 2010 [Video file]. Available: https://www.youtube.com/watch?v=vZK74nUJl48 [Accessed Sept. 11, 2020].

[19] D. Miljković, "Active noise control in light aircraft cabin using multichannel coherent method," Automatika, Vol. 57, No. 4, 2016.

[20] Automann-TV, "2015 Mercedes-AMG C63 S (510hp) - pure SOUND", YouTube, 11 Jun 2010 [Video file]. Available: https://www.youtube.com/watch?v=P_KMYhD0HT0 [Accessed Sept. 11, 2020].

[21] Sample-Based Synthesis, [Online]. Available: https://soundbridge.io/sample-based-synthesis/ [Accessed Sept. 11, 2020].

[22] R. Bristow-Johnson, Wavetable Synthesis 101, A Fundamental Perspective, Audio Eng. Soc. Prerpint No. 4400, 1996

[23] J. Driedger and M. Müller, "TSM Toolbox: Matlab implementations of time-scale modification algorithms," In Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx-14), Erlangen, Germany, Sept. 1-5, 2014

[24] J. Jagla, J. Maillard and Nadine Martin, "Sample-based engine noise synthesis using an enhanced pitch-synchronous overlap-and-add method," Journal of the Acoustical Society of America, Acoustical Society of America, 2012, 132 (5), pp. 3098-3108.

[25] E. Struzynski, FS Sound Studio, 2007, [Online]. Available: http://www.fssoundstudio.com/ [Accessed Sept. 11, 2020].

[26] M. Gökdag, "Study of the road traffic noise in Erzurum-Turkey," in Iranian Journal of Environmental Health Science & Engineering 9(1):22, Dec. 2012

Fault Detection of Aircraft Piston Engine Based on Exhaustive Database Search

Dubravko Miljković
Hrvatska elektroprivreda, Zagreb, Croatia
dubravko.miljkovic@hep.hr

***Abstract* - Modern aircraft piston engine monitor collects large amounts of data representing engine parameters. Data acquired during each flight is logged to a file. From the logs acquired during numerous previous flights a database of a healthy engine is formed containing feature vectors with selected engine parameters. Due to high speed processing of even modest microcontrollers, new feature vectors acquired during a flight can be compared in real time to all previous feature vectors stored in the database. Any significant discrepancy may indicate a problem in engine operation or engine fault. Such detection may alert a pilot and help prevent engine failure, partial loss of engine power or unplanned in-flight engine shutdown.**

Keywords - aircraft; piston engine; engine monitor; fault detection; database

I. INTRODUCTION

Most general aviation aircrafts (popularly known as light aircrafts) are using aircraft piston engines. These engines provide a combination of power output and fuel consumption, suitable for such class of aircrafts [1]. However, these engines are many times less reliable than turboprop and jet engines that can be found on larger aircrafts, [2]. This fact is further aggravated because most general aviation aircrafts are single engine and engine failure or significant power loss may lead to an off airport landing that carries very high risks, particularly over inhospitable terrain (forest, water, urban areas). One way of increasing operational reliability of piston engines is application of engine monitors, also called engine analyzers and engine management system, [3]. Example of aircraft piston engine mounted on a piston engine aircraft is shown in Fig. 1. Complete or partial loss of engine power is quite a serious situation even with multi engine aircraft; because such loss severely reduces climb rate and produces asymmetric thrust that can be challenging to handle, particularly immediately after lift-off, before the aircraft reaches an altitude and speed. The idea behind the engine problem detection method described later in this paper is not to detect a particular type of engine problem, but to, in a simple way, detect engine parameters and their combinations (patterns) that are not usually encountered in a database formed from logs of previous flights. Such detection of aberrant parameters and patterns is to be used just to warn a pilot that something unusual is happening with the engine, and that adequate decisions can be made regarding the progress of flight. Accurate diagnostics is still by far in the domain of experienced maintenance personnel and is performed after the landing.



Figure 1. Piston engine mounted on a single engine aircraft (Ahunt)

II. ENGINE MONITOR

Engine monitor tracks critical engine parameters, alerting to unexpected changes using its advanced exceedance monitoring system, [4-7]. Intelligent warning messages are displayed prominently, allowing pilots to immediately recognize and interpret a critical situation. Beside presentation and logging of engine operation, most engine monitors may be also used as an aid in leaning the engine. Every new generation of engine monitors brings some improvements in number of parameters they monitor and in format how they are presented on the display (use of vertical and horizontal bars, arcs and numerical data, in recent time almost always in color). A display of one modern engine monitor, [4] is shown in Fig. 2.



Figure 2. JPI EDM-900 Engine Monitor (JP Instruments)

Many engine problems can be identified by a specific pattern that consists of CHT and EGT vertical bars. Such patterns are described in engine diagnosis charts that can be found in manuals that accompany engine monitor. Engine parameters are logged to the internal memory or memory card with the typical rate of one record every six seconds (this period is programmable in most engine monitors). Most common engine parameters that are monitored by modern engine monitors are listed in Table I, [8].

TABLE I. MONITORED ENGINE PARAMETERS

Parameter	Description
EGT	Exhaust Gas Temperature
CHT	Cylinder Head temperature
OIL TEMP	Oil Temperature [1]
OIL PRES	Oil Pressure [1]
TIT1	Turbine Inlet Temperature 1[1]
TIT2	Turbine Inlet Temperature 2 [1]
OAT	Outside Air Temperature
CDT	Compressor Discharge Temperature [1]
IAT	Intercooler Air Temperature [1]
CRB	Carburetor Air Temperature [1]
CDT - IAT	Intercooler cooling
RPM	Rotations Per Minute
MAP	Manifold Pressure
%HP	% Horse Power[4]
CLD	CHT cooling rate [2]
DIF	EGT span [3]
FF	Fuel Flow [1]

[1]optional, [2]fastest cooling cylinder, [3]difference between the hottest and coolest EGT, [4]calculated

Example of section of engine log is shown in Fig. 3.

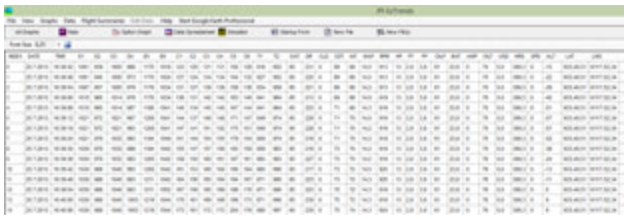


Figure 3. Small section of engine log

Recorded engine parameters that are saved in an engine monitor log can be later analyzed by specialized analysis software that is regularly provided by the producer of an engine monitor, like, e.g., JP Instruments EZTrends, [9,10]. CHT and EGT parameters from one engine monitor log are shown in Fig. 4.

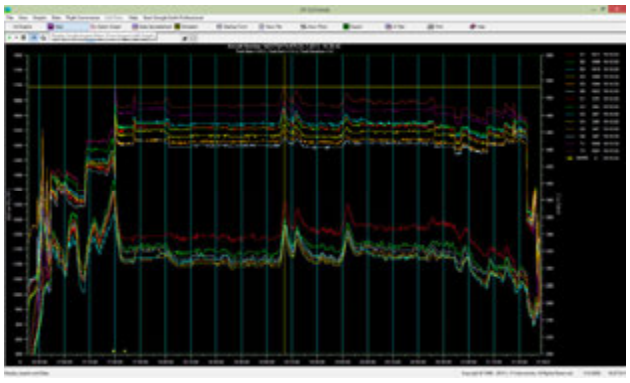


Figure 4. Engine parameters in analysis software

TABLE II. RECOMMENDED ENGINE PARAMETER VALUES SUITABLE FOR SETTING ALARM LIMITS

Parameter	Limit	
	Low	High
EGT		1550 °F 843 °C
CHT		450 °F 230 °C
TIT		1650 °F 900 °C
OIL TEMP	90 °F 32 °C	230 °F* 110 °C
DIF		500 °F 280 °C
MAP		32 inch Hg

Basic engine diagnostics involve comparing engine parameters with the predefined limits (limit checking). Default limits (could be changed) for some parameters measured by JPI engine monitors are shown in Table II.

III. REGIME SWITCHING

Engine operation during a period of flight goes through various regimes that correspond to flight regimes (idle, run-up, taxi, take-off, climb, cruise, descent and landing). Engine parameters in one regime may be quite different then in another regime. It may be difficult to compare engine parameters belonging to different regimes. For sake of comparison, it may be beneficial to separate engine operations into several regimes, [8,11]. There are few ways of separating different engine regimes, some are quite easy to determine and some are more precisely related to flight regimes and require GPS data, [11]. Among the variables that can be considered for regime separations, simple solutions are engine RPM and calculated % HP. The use of RPM is the easiest way for separating engine operation into regimes. The better way is the use of calculated percent of rated horsepower (% HP) as a variable for regime switching, [6]. It is dependent on accurate measuring of additional parameters. Calculated % HP is supplied by monitor considering Fuel Flow, RPM, OAT and MAP, but also environment temperature and altitude variations. Values of calculated % HP as a function of RPM and MAP (manifold pressure) are shown in Fig. 5.

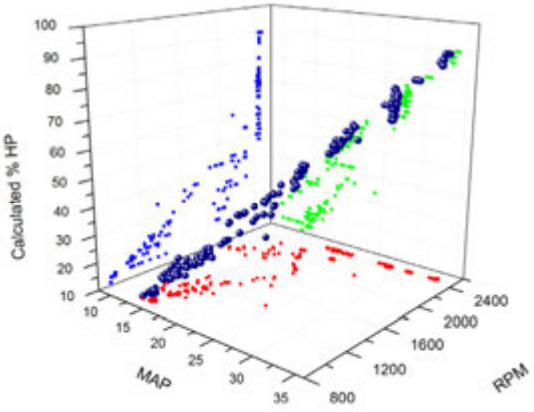


Figure 5. Calculated % HP as a function of RPM and MAP

Separation of engine operation in regimes as a function of % HP in 20 % intervals is given by (1) and listed in Table III.

$$r = \left\lceil \frac{\%\,HP}{20} \right\rceil + 1 \qquad (1)$$

TABLE III. ENGINE REGIMES AS A FUNCTION OF CALCULATED % HP

% HP	Engine regime *r*
0-19	1
20-39	2
40-59	3
60-79	4
80-99	5
100-119	6

Due to calibration errors, calculated % HP sometimes exceeds 100%, hence regime 6 is provided. Operating regimes during a typical flight are shown in Fig. 6. High power is applied during the preflight check and takeoff.

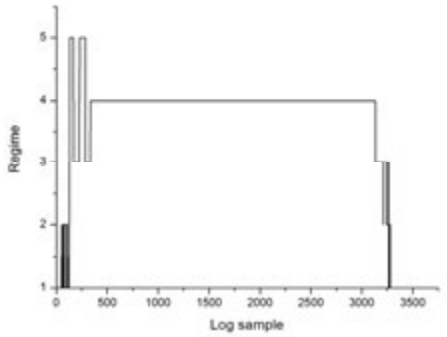


Figure 6. Engine operating regimes during a typical flight

After analyzing JPI logs supplied as samples with the EZtrends software, [10] related to Flt#70 and Flt#71 (total 5685 records, 9 h 28 min) from the six-cylinder, 209 kW (280 hp) Continental TSIO-550-G engine and belonging to Mooney 20TN Acclaim aircraft, the following distribution across regimes is achieved, as shown in Table IV.

TABLE IV. NUMBER OF RECORDS PER REGIME

Regime	1	2	3	4	5	6
Records	221	119	281	4948	114	0

All engine parameters don't react to change in calculated % HP with equal speed. Exhaust gasses, EGT and TIT, react quickly while CHT and OILT due to cylinder and engine metal mass react more slowly.

IV. FAULT DETECTION USING DATABASE SEARCH

Basic method for detection of engine problems is shown in Fig. 7. The database used in this fault detection method consists of feature vectors obtained from numerous engine logs acquired during previous flights (all flights are united in one large log) where engine problems have not been encountered. The log contains 600 records per one flight hour (depending on the selected rate). Engine parameters are in real-time compared to all previously stored feature vectors belonging to the same regime.

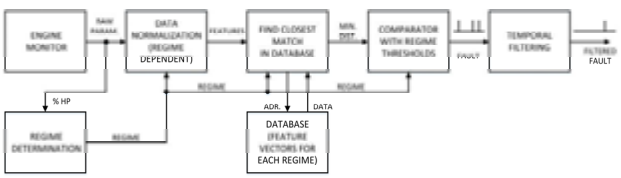


Figure 7. Normal operation without learning

If the closest vector in the database is still too distant (above specific threshold) than the test vector containing recent engine parameters is considered to belong to an engine that has developed some kind of a problem.

A. Engine Parameters Selected for Comparison

Parameters that are of most interest in interpretation and analysis of engine logs belong to one of following three groups:

- Cylinder Head Temperatures CHTs,
  Temperatures of cylinder heads for each individual cylinder
- Exhaust Gas Temperatures EGTs,
  Temperatures of exhaust gasses for each individual cylinder
- Parameters common to all cylinders like FF – Fuel Flow, OILT – Oil Temperature and OILP – Oil Pressure.

There are, naturally, many other parameters available in modern engine monitor, but they are generally of less importance and are not used for identifying engine problems in this method.

Above mentioned parameters are static, in term, that they don't reflect time dynamics of engine parameters. As such, these parameters cannot be used (without their time history) to detect problems, like EGT oscillations that precede exhaust valve problems, [12]. However, they can still be used for detection of most engine problems.

Selected engine parameters are used as elements of the feature vector **q,** (2):

$$\mathbf{q}=[CHT_1, CHT_2, \ldots, CHT_n, EGT_1, EGT_2, \ldots, EGT_n, FF, OILP, OILT] \quad (2)$$

B. Comparison Metrics

A variant of Euclidean distance is used in comparisons. The Euclidean distance, also called a Euclidean metric is the distance between two points in Euclidean space. Euclidean distance $d_E(\mathbf{p},\mathbf{q})$ between vectors **p** and **q** is given by (3), where p_i and q_i are vector components.:

$$d_E(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (3)$$

However, vector components contain EGT, CHT, FF, OILP and OILT values that belong to different ranges, have different variances and their differences will differ hugely. To balance the contributions of different engine parameters it is necessary to perform transformation called standardization, (4), that rescales data to zero mean and unit variance, where x_i' is normalized vector component.

$$x_i' = \frac{x_i - \bar{x}_i}{\sigma_i} \quad (4)$$

The average value $\bar{x}_i$ of each vector component x_i is determined for each engine operating regime, (5). The values of the particular vector component are averaged using all stored feature vectors for that engine regime.

$$\bar{x}_i = \frac{1}{N}\sum_{i=0}^{i=N-1} x_i \quad (5)$$

Standard deviation σ_i is determined according to (6).

$$\sigma_i = \sqrt{\frac{\sum_{i=0}^{N-1}(x_i - \bar{x}_i)^2}{N}} \quad (6)$$

For each regime following normalization factors are to be calculated:

- average EGT, standard deviation EGT
- average CHT, standard deviation CHT
- average FF, standard deviation FF
- average OILP, standard deviation OILP
- average OILT, standard deviation OILT

Following normalization factors have been determined using previously mentioned JPI logs, Flt#70 and Flt#71, as shown in Table V.

TABLE V. AVERAGES AND STDANDARD DEVIATIONS FOR ENGINE PARAMETERS IN VARIOUS ENGINE OPERATING REGIMES

Reg.	R1		R2		R3		R4		R5	
Par.	Av.	σ	Av.	σ	Av.	σ	Av.	σ	Av.	σ
EGT1	1103	79,1	1297	53,6	1374	112	1534	60,7	1346	25,4
EGT2	1060	84,7	1292	76,9	1412	105	1548	45,8	1405	26,1
EGT3	1086	79,2	1293	68,3	1396	98,9	1550	57,0	1364	23,5
EGT4	1066	78,3	1294	68,6	1384	102,6	1521	50,2	1352	24,2
EGT5	1176	90,8	1343	54,5	1379	88,7	1498	45,5	1351	23,0
EGT6	1096	79,0	1304	62,0	1372	102,2	1486	48,4	1324	22,4
CHT1	238	41,2	263	17,3	287	22,3	336	18,9	343	15,7
CHT2	224	42,8	257	16,7	291	23,8	323	21,0	346	19,6
CHT3	223	36,6	249	15,7	275	19,1	310	17,2	324	12,9
CHT4	226	37,2	253	15,6	280	20,9	310	21,0	330	14,2
CHT5	248	36,4	270	19,4	292	20,8	313	18,9	347	15,5
CHT6	226	37,5	255	16,4	288	20,0	318	19,3	340	16,3
FF	3,22	0,68	7,59	3,18	17,0	3,22	16,9	2,55	32,7	0,48
OILP	51,7	9,42	54,4	3,06	53,8	1,39	54,1	1,62	58,7	3,23
OILT	116	32,8	146	25,4	174	9,08	167	7,65	176	11,6

After some initial experiments, it was found that the squared difference in (3) doesn't emphasize deviations enough (has not sufficient discriminatory power), and an exponent of four is used instead of two in new distance (7):

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^4} \quad (7)$$

In real application, square root could be eliminated (with squared thresholds) to reduce computational burden.

C. Database

The database contains data for all engine operating regimes, as shown in Fig. 8. As can be seen, the database is divided into six sections corresponding to engine operating regimes. The database is filled with all normalized feature vectors (total 5685 feature vectors, 9 h 28 min) originating from JPI logs Flt#70 and Flt#71.

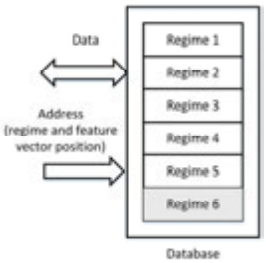


Figure 8. Database containing data for various engine operating regimes

D. Finding Closest Feature Vector in the Database

Test feature vector **t** is compared using comparison metrics (7) with all feature vectors $\mathbf{q}_{i,r}$ stored in the database that belong to the same engine operating regime r as the test feature vector and the minimum distance d_{min} is determined, where N_r is the number of feature vectors in the database that belong to engine regime r, (8).

$$d_{min} = \min_{i=1,N_r}\left(d(\mathbf{t}, \mathbf{q}_{i,r})\right) \quad (8)$$

E. Why Not Clustering?

Use of clustering (K-Means, Isodata, Kohonen SOM) could be potentially beneficial, as it would substitute a large number of feature vectors with a much smaller number of cluster prototype. This would reduce the size of the database and reducing search time. One solution with just one prototype per regime obtained by averaging is described in [8]. However, clustering is not used in this method for two reasons. First, regime switching is used and it has already partitioned database into smaller subsections. Today, with modern microcontrollers it is not difficult to compare thousands of feature vectors in a real time. Second, the precision of the comparisons is better when original feature vectors are used instead of cluster prototypes.

F. Suspicious Feature Vector Detection

Each new vector of engine parameters is compared to all previously stored vectors in the database belonging to the same engine operating regime. As the database is not too large (several thousands of feature vectors) this comparison can be accomplished in a real time. Distance between new feature vector and all stored feature vectors for a particular regime is determined. The statistical distribution of intraset distances (within the same regime) is shown in Fig. 9. Values for P_{95} and P_{99} percentiles for various engine operating regimes are shown in Table VI. The corresponding percentiles in form of Box-Jenkins plot are shown in Fig. 10. Decision thresholds for each engine regime are determined based on the statistical distribution

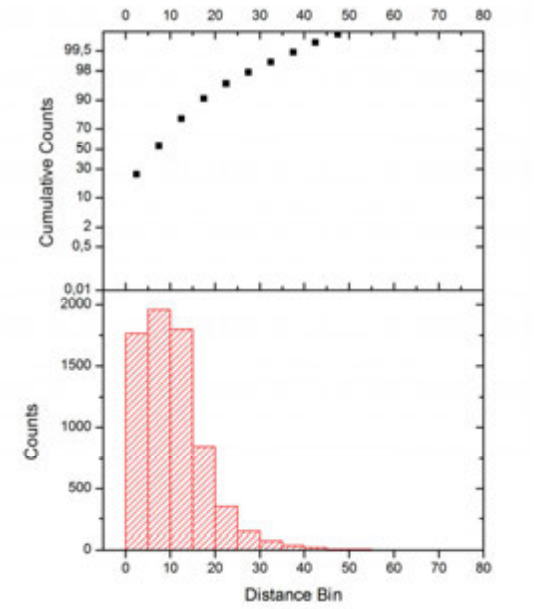


Figure 9. Statistical distribution of distances (example regime 2)

TABLE VI. REGIMES AND P_{95} AND P_{99} PERCENTILES

Regime	1	2	3	4	5
P_{95}	27,0330	24,0047	22,9221	48,8025	36,8017
P_{99}	39,5847	35,5259	34,2285	60,0759	61,2639

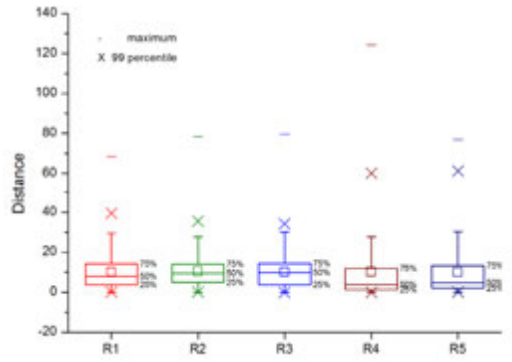


Figure 10. Percentiles of distances for each engine operating regime

of distances among feature vectors belonging to that regime. One way of determining the threshold is to consider acceptable numbers of occurrences during one flight. If the engine is operating correctly, this number should be very small or zero. However, due to lack of sufficient data necessary for determination of very high percentiles, value of 99 percentiles, may be chosen for appropriate threshold. Occasional and isolated threshold exceedances may be tolerable, particularly when they will be further reduced by temporal filtering. Different threshold $P_{99,r}$ are determined for each engine operating regime r, and if exceeded it indicates possible problem, (9).

$$d_{min} > P_{99,r} \tag{9}$$

To reduce spurious indication of engine problems, temporal threshold is also included, [13]. Temporal threshold, (6), is designed to tolerate a specific number of exceedance of a statistical threshold within a period of time. In this way error must repeat few times within a specified period, Fig. 11, otherwise is not considered and error. An alert is issued if the number of threshold exceedances in a period of time N_v is greater than N_T, (10):

$$N_v > N_T \tag{10}$$

(e.g., collected within the last minute, sliding window)

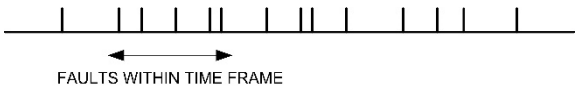


Figure 11. Permitted number of threshold violations within a specific time frame

So, if the new test vector of engine parameters is only a glitch, it would be filtered out by temporal filtering. However, if it is something more persistent, than temporal filtering will not filter it out and alert will be issued. Accurate adjustment of statistical and temporal thresholds is a tradeoff between sensitivity and false detections.

V. Proposal for the Incorporation of Learning

Previously described basic method could be further extended by introduction of automatic learning. Once engine feature vector has been identified as a normal, it can be added to the existing database, Fig. 12. However, how can we know that these feature vectors belong to a healthy engine? One way would be to wait a period of time after this feature vector **s** is collected and stored in the temporary

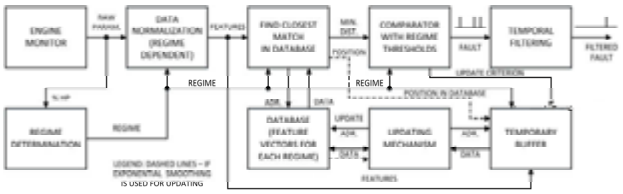


Figure 12. Incorporation of learning

buffer before it is added to the database. How long should one wait? There may be a slow drift of engine parameters toward the parameters that appear in a problematic engine.

Following requirements may be set on new feature vectors to update existing database:

- Feature vectors to update the database must be of the smaller distance to the closest stored vector than a threshold $P_{99,r}$ or using $P_{95,r}$ instead (with no K), (11).

$$d_{min} < K\, P_{99,r} \qquad \text{(e.g. } K = 0.7\text{)} \tag{11}$$

- Feature vectors are held in a temporary buffer for a specified period (e.g., duration of 20 flights). It is supposed that any failure would be already detected in a 20 hour period. So, feature vectors $\mathbf{s}(t)$ are put in a "quarantine" for 20 hours. If a fault develops in 20 hour period, stored normal feature vectors are discarded. Otherwise, they are used for update, $\mathbf{s_u}(t)$. With the data acquisition rate of one record every six seconds there will be 12000 records in 20 hours, (12):

$$\mathbf{s_u}(t) = \mathbf{s}(t - 12000) \tag{12}$$

After that, they can be used to update database in one of following ways:

- Adding a new feature vector $\mathbf{s_u}(t)$ till space is available (database size and increases comparison time). Database size has also its limits.
- Replacing older feature vectors with new ones $\mathbf{s_u}(t)$ using FIFO (First-In First-Out) principle.
- Updating closest existing feature vector with new one using exponential smoothing and a very small α. This would result in fine tuning database feature vectors.

Exponential smoothing is a method for smoothing time series data using the exponential window function, (13),(14). Smoothing is performed on all elements s_i of a vector **s**.

$$\bar{s}_i(0) = s_i(0) \tag{13}$$

$$\bar{s}_i(t) = \alpha s_i(t) + (1-\alpha)\bar{s}_i(t-1),\ t > 0 \tag{14}$$

where α is the smoothing factor, (15):

$$0 < \alpha < 1 \tag{15}$$

VI. Validation of the Method

Only basic method (for detection without learning) was tested. Testing the method with learning would require too much engine monitor data that are not easy to obtain. Due to highly reliable engines, and hence the lack of engine monitor logs that include specific engine problems, artificial faults have been used. Three artificial faults have been generated from log records belonging to a healthy

engine, by modifying engine parameters according to parameter deviations described in JPI manual containing catalogued engine failure patterns as they appear on the display of an engine monitor, [7] and as described in [11,14]. The regime used is 5, $P_{99,5} = 61{,}2639$ (Table VI).

A. *Loss of EGT for One Cylinder*

First engine problem used in a test is the loss of EGT for one cylinder engine and engine running rough. This is probably caused by stuckvalve. The corresponding catalogued diagnostic pattern (example, different cylinder may be affected) is illustrated in Fig. 13 and deviation described by (16), [14].

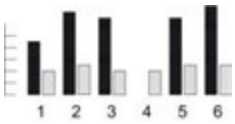


Figure 13. Catalogued engine fault pattern for fault in case A

$$EGT_{min} < 600 \tag{16}$$

Artificial test vector is $\mathbf{t_1}$ (17) and distances (18) and (19).

$$\mathbf{t_1}=[1332\ 1388\ 1359\ \underline{546}\ 1345\ 1316\ 310\ 317\ 306\ 313\ 317\ 316\ 29.4\ 53184] \tag{17}$$

$$d_{min} = 222.19200 \tag{18}$$

$$d_{min} > P_{99,5} \tag{19}$$

B. *Decrease of EGT for One Cylinder at Low RPM*

The second test problem is caused by a decrease of EGT for one cylinder. This problem appears when the intake valve is not opening fully. The corresponding catalogued diagnostic pattern (example, different cylinder may be affected) is illustrated in Fig. 14 and deviation described by (20), [14].

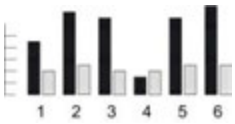


Figure 14. Catalogued engine fault pattern for fault in case B

$$DIFF > 500 \text{ and } RPM < 1500 \tag{20}$$

Artificial test vector is $\mathbf{t_2}$ (21) and distances (22) and (23).

$$\mathbf{t_2}=[1332\ 1388\ 1358\ \underline{846}\ 1345\ 1316\ 310\ 317\ 306\ 313\ 317\ 316\ 29.4\ 53\ 184] \tag{21}$$

$$d_{min} = 80.60468 \tag{22}$$

$$d_{min} > P_{99,5} \tag{23}$$

C. *EGT and CHT not Uniform*

In third test problem EGT and CHT temperatures are not uniform across all cylinders in the engine (large deviations are present). This is caused by dirty fuel injectors or fouled plugs. The corresponding catalogued diagnostic pattern (example, almost endless variations are possible within constrains) is illustrated in Fig. 15 and deviation described by (24), [14].



Figure 15. Catalogued engine fault pattern for fault in case C

$$(EGT_i / CHT_i) < 3 \text{ OR } (EGT_i / CHT_i) > 6 \text{ for any } i,\ i=1,\ldots,N_c \tag{24}$$

Artificial test vector is $\mathbf{t_3}$ (25) and distances (26) and (27).

$$\mathbf{t_3}=[1332\ \underline{1018}\ 1358\ \underline{1526}\ 1345\ 1316\ 310\ \underline{415}\ 306\ \underline{235}\ 317\ 316\ 29.4\ 53\ 184] \tag{25}$$

$$d_{min} = 64.40012 \tag{26}$$

$$d_{min} > P_{99,5} \tag{27}$$

From the previous examples, problematic test vector has exceeded the threshold value and as such examples are classified as problematic in all three cases. Due to the large number, of vector components, i.e. 15, (2), the method is sensitive only to larger variations. To increase sensitivity exponent of four in the squared difference is used in the distance (7) instead of two in Euclidean distance (3).

VII. CONCLUSION

Loss of engine power in an aircraft has always been a serious threat. To diminish the frequency of such occurrences application of digital engine monitors is recommended. An approach for detection of engine problems based on an exhaustive database search of previously stored feature vectors of selected engine parameter is described. This approach doesn't pretend to identify the cause of engine problems (this would require detailed and carefully designed expert system), but to warn pilot on unusual engine parameters that were not present in previous flights, and raise suspicion that engine reliability may be at stake. Combination of statistical and temporal threshold is used to reduce spurious alerts and hence generate more robust alerts. The described method was tested on three engine problems represented by artificially generated faults. An extension of this method that includes automatic learning and modifying database is proposed.

REFERENCES

[1] M. Bush, Mike Busch on Engines, Savy Aviator, Wayoming, 2018
[2] Aircraft Reciprocating-Engine Failure, Aviation Safety Research and Analysis Report, B2007/0191, Australia 2007
[3] D. Miljković, "Engine monitors for general aviation piston engines," CrSNDT Journal, Vol. 10, 2013, pp. 19-23
[4] Pilot's Guide, Engine Data Management, EDM-900 Primary TSO/STC, J.P. Instruments, Inc. 2011
[5] The Pilots Manual - Leaning and Diagnosing Engine Problems, Electronics International Inc., 21 April 2004
[6] Graphic Engine Monitors, Gem Series Pilot's Guide, Insight Instrument Corporation, Doc. No. 930322,Ver. 3.0, July 23, 1995
[7] JP Instruments Inc., "Pilot's Guide, Engine Data Management EDM-730 EDM-830 EDM-740 Experimental," 2010. [Online]. Available: https://www.jpinstruments.com/wp-content/uploads/2012/06/PG-EDM-730_830-Rev-A-02-JULY-2009.pdf [Accessed Sept. 11, 2020].
[8] D. Miljković, "Regime dependent aircraft piston engine monitoring," In Proc. MIPRO 2014, Opatija, Croatia, 26-30 May 2014
[9] Pilot's Guide, EzTrends, J. P. Instruments, 2006
[10] EzTrends (Download and Plotting Software), JPI, California, USA, [Online]. Available: https://www.jpinstruments.com/technical-support/software-downloads/ [Accessed Sept. 11, 2020]
[11] D. Miljković, "Engine fault detection for piston engine aircraft," In Proc. MIPRO 2013, Opatija, Croatia, 20-24 May 2013.
[12] D. Miljković, "Detecting aircraft piston engine problems by analysis of engine parameters," In Proc. MIPRO 2019, Opatija, Croatia, 20-24 May 2019.
[13] D. Miljković, "Aircraft piston engine fault detection based on uniformity of cylinder head, exhaust gas and turbochargers temperatures," In Proc. ECNDT 2014, Prague, Czech Republic, 6-10 October 2014
[14] D. Miljković, "Automatic interpretation of aircraft piston engine monitor log," In Proc. MIPRO 2014, Opatija, Croatia, 26-30 May 2014.

PROFESSIONAL PAPER IN CROATIAN

Automatizacija pročistača otpadnih voda

Marko Ilijanić*, Mario Lučan** i Goran Malčić**
* KONČAR - Elektronika i informatika d.d. , Zagreb, Hrvatska
** Tehničko veleučilište u Zagrebu, Zagreb, Hrvatska
markoilijan@gmail.hr mlucan@tvz.hr goran.malcic@tvz.hr

***Sažetak* - Otpadna voda zbog svog štetnog utjecaja na okoliš a pogotovo na čistu vodu u okolišu predstavlja veliki problem svakog naseljenog mjesta. Da bi se spriječilo zagađenje čistih voda i neželjene promjene u okolišu potrebno je otpadnu vodu prije ispuštanja u okoliš pročistiti. Pročišćenje otpadnih voda je proces u kojem se smanjuje onečišćenje do onih količina ili koncentracija s kojima pročišćene otpadne vode ispuštene u okolinu postaju neopasne za život i ljudsko zdravlje i ne uzrokuju neželjene promjene u okolišu. Pročišćenje otpadne vode odvija se u pročistaču otpadne vode.**

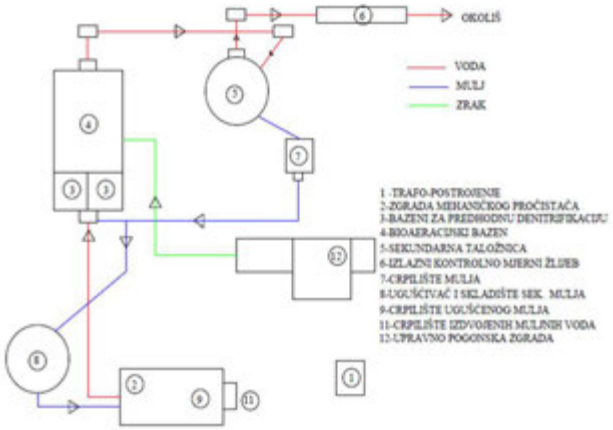
U ovom radu je obrađen rad pročistača otpadne vode kapaciteta 5250 ES (ekvivalent stanovnika). Opisani su osnovni tehnološki pojmovi otpadne vode, tehnološki procesi u predmetnom pročistaču i implementacija automatiziranog sustava za upravljanje sa predmetnim pročistačem. Prikazana su tehnička i softverska rješenja koja se koriste u automatiziranom sustavu upravljanja sa predmetnim pročistačem.

I. OPIS TEHNOLOŠKOG PROCESA

Otpadne vode nastaju upotrebom vode iz raznovrsnih vodoopskrbnih sustava za određene namjene, pri čemu dolazi do promjene prvotnih fizikalnih, kemijskih i bioloških značajki [2]. Prikupljaju se kanalizacijskom mrežom i odvode na pročišćenje prije ispuštanja u okolinu, ovisno o sastavu i količini određuje se i postupak pročišćenja. U ovome radu je obrađen pročistač otpadnih voda koji je projektiran za mjesto koje ima oko 4000 stanovnika. U njega dolazi pretežno sanitarna otpadna voda opterećena sa organskom tvari. Na temelju ulaznih i željenih izlaznih parametara otpadne vode odabran je postupak pročišćenja i dimenzije pročistača otpadnih voda.

Odabran je postupak nisko opterećenog biološkog stupnja pročišćenja uz istovremenu aerobnu stabilizaciju viška sekundarnog mulja, uključivo i prethodnu denitrifikaciju i pojačano biološko uklanjanje fosfora. Za pročišćenje otpadnih voda koristi samo bakterije i mikroorganizme koji se nalaze u otpadnoj vodi.

Sastoji od dvije cjeline, a to su mehanički i biološki pročistač otpadne vode. Mehaničkim pročišćenjem se iz otpadne vode uklanjaju krupni otpaci, brzo taložive krutine, ulja i masti. Provodi se prije biološkog pročišćenja da bi se spriječilo oštećenje biološkog pročistača sa navedenim tvarima. U ovom pročistaču otpadne vode mehanički pročistač otpadne vode izveden je kao kompaktno postrojenje. Biološko pročišćenje otpadne vode je pročišćenje otpadne vode uz pomoć mikroorganizama koji se u njoj nalaze. Primjenjuje se da bi se iz otpadne vode uklonio organski ugljik, smanjili fosforni i dušikovi spojevi te stabilizirao mulj. Biološki pročistač otpadne vode sastoji se od više različitih bazena u kojima se osiguravaju uvjeti za što kvalitetnije pročišćenje otpadne vode uz pomoć mikroorganizama.



Slika 1 Tlocrt pročistača sa ucrtanim tijekom otpadne vode, mulja i zraka

II. OPIS TEHNIČKOG RJEŠENJA

Na temelju opisa tehnološkog procesa i zahtjeva pročistača otpadnih voda projektiran je upravljački ormar. Upravljački ormar upravlja sa cijelim pročistačem otpadne vode, a sastoji se od upravljačke elektronike, korisničkog sučelja i energetskog dijela. Upravljanje je temeljeno na programibilnom logičkom kontroleru PLC (engl. *Programmable Logic Controler*), a korisničko sučelje je operatorski panel osjetljivo na dodir.

Sa cijelim procesom preko izvršnih članova upravlja PLC po upravljačkom programu na temelju ulaznih informacija. Preko operatorskog panela omogućeno je uključenje i isključenje svih crpki i miješalica. Na operatorskom panelu prikazano je stanje svake crpke i miješalice (rad/kvar) i broj radnih sati svake crpke i miješalice. U slučaju bilo kojeg kvara na uređaju PLC preko GSM modula šalje SMS poruku operateru i uključuje signalnu sirenu.

Pored svakog elementa pročistača otpadnih voda izvedeno je tipkalo za nužni stop koje prilikom aktivacije isključuje taj dio pročistača.

A. Kompaktno postrojenje za mehaničko pročišćenje otpadne vode

Svježa otpadna voda prvo ulazi u kompaktno postrojenje za mehaničko pročišćenje otpadne vode gdje se iz nje uklanjaju krute tvari, pijesak i mast. Kompaktno postrojenje za mehaničko pročišćenje sastoji se od grube mehaničke rešetke, fine automatske rotacijske rešetke, kanala aeriranog pijeskolova sa elektromotornim pužnim podnim zgrtalom istaloženog pijeska i separatorom pijeska, kanala za isplivavanje ulja, masnoća i drugih plivajućih tvari opremljenog sa elektromotornim zgrtalom ulja i masti i crpkom za crpljenje izdvojenih ulja i masti. Odabrani uređaj je kapaciteta Q_{max}= 40 l/s (114 m3/h) i nalazi se u zgradi finog mehaničkog čišćenja. Kompaktno postrojenje za mehaničko pročišćenje otpadne vode ima vlastito upravljanje pa mu je iz upravljačkog ormara omogućeno napajanje i signalizacija kvara koja preko njegovog relejnog izlaza dolazi na digitalni ulaz PLC-a.

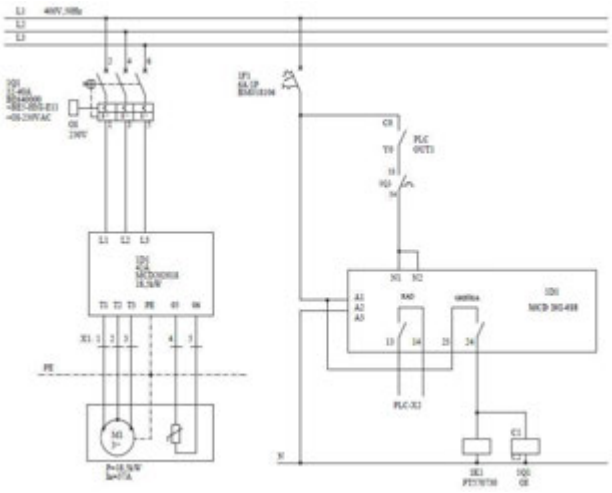
B. Bazeni za prethodnu denitrifikaciju

U dva bazena za prethodnu denitrifikaciju se tlačnim cjevovodom dovodi mehanički pročišćena otpadna voda iz kompaktnog postrojenja za mehaničko pročišćenje i povratni aktivni mulj iz crpilišta mulja. U bazenima vlada anoksično stanje. Biomasa u bazenima uzima nužni kisik iz nitrata (NO_4) čime nastaju novi spojevi nitriti (NO_3) koji se u daljnjem procesu razgrađuju na elementarni kisik i dušik. Dušik je u plinovitom stanju te preko površine bazena odlazi u atmosferu, čime se iz otpadne vode uklanja dušik. Za potrebe sprečavanja taloženja aktivnog mulja na dno bazena i potrebe pospješivanja isplinjavanja dušika u atmosferu u bazene su ugrađene odgovarajuće podvodne propelerne miješalice koje stalno rade. Smjesa otpadne vode i aktivne biomase uvode se u bioaeracijske bazene sa dna bazena prethodne denitrifikacije. Biomasa iz bioaeracijskog bazena vraća se sa gornje trećine bioaeracijskog bazena u bazene za denitrifikaciju putem cjevovoda sa nepovratnim ventilima. Preko operatorskog panela omogućeno je uključenje i isključenje svake miješalice te je na svakoj prikazan broj radnih sati.

C. Bioaeracijski bazeni

Finalno biološko pročišćavanje vrši se u bioaeracijskom bazenu pomoću aerobnih bakterija (tzv. aktivni mulj). Aerobne bakterije prerađuju organska zagađenja otpadnih voda u anorganske tvari (mulj). Osim organske tvari za svoj metabolizam i razmnožavanje potrebna im je i velika količina kisika. Potrebna količina kisika u bazen se unosi upuhivanjem komprimiranog zraka preko membranskih aeratora smještenih na dnu bazena. Komprimirani zrak proizvodi se pomoću tri kompresora. Kompresori su smješteni u pogonskoj zgradi, napajaju iz upravljačkog ormara a preko operatorskog panela moguć je odabir režima njihovog rada, automatsko ili ručno. Na operatorskom panelu se prikazuje broj radnih sati svakog kompresora i signalizacija za servis svakog kompresora nakon isteka broja radnih sati koje je deklarirao proizvođač kompresora.

Kod automatskog režima rada kompresori se uključuju ovisno o količini kisika u bioaeracijskom bazenu, sa ciljem



Slika 2 Shema spoja kompresora

održavanja količine kisika u granicama od 1-2 mg O_2/l. Kod nestanka električne energije iz distribucijske mreže i napajanja postrojenja iz lokalnog dizel-agregata smije biti uključen samo jedan kompresor. Prilikom uključenja kompresori povuku od 5 do 7 puta veću struju od nazivne što može izazvati propade u mreži i ispad zaštitnih elemenata u krugu napajanja kompresora. Kako bi se to spriječilo, kompresori se uključuju preko mekog upuštača. Odabran je meki upuštač *Danfoss MCD 202-018* koji pomoću poluvodičkih energetskih ventila ograničava poteznu struju kompresora[8].

D. Sekundarni taložnik

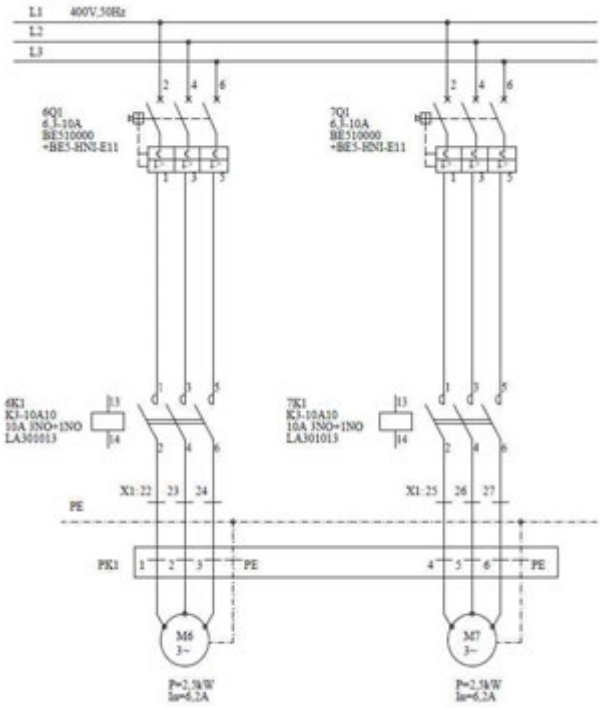
Razdvajanje biološki pročišćene vode i aktivnog mulja vrši se u kružnom sekundarnom taložniku sa pretežnim horizontalnim strujanjem od centra građevine prema obodnim preljevnim površinskim žljebovima. Horizontalno strujanje postiže se pomoću kružnog elektromotornog mosta na koji je pričvršćeno podno i površinsko zgrtalo mulja.

Površinsko/podno zgrtalo mulja je nazivne snage 0,37 kW i napajano je iz upravljačkog ormara. Na operatorskom panelu omogućeno je uključenje i isključenje površinsko/podnog zgrtala i prikazan je broj njegovih radnih sati.

E. Crpilište sekundarnog mulja

Aktivni i mineralizirani mulj iz muljnog udubljenja sekundarne taložnice preljeva se u crpilište sekundarnog mulja. Iz crpilišta sekundarnog mulja mulj se potopnim crpkama i pripadajućim cjevovodom prebacuje u bazene za prethodnu denitrifikaciju. Odnosno preko njih u bioaeracijske bazene u svrhu održavanja potrebne koncentracije aktivnog mulja u bazenima. Zbog relativno male količine mulja crpke se povremeno uključuju.

Ugrađene crpke rade u redundantnom režimu rada i nazivne su snage 2,5kW. Pomoću operatorskog panela omogućen je odabir režima rada automatsko ili ručno te je omogućeno uključenje i isključenje svake crpke. Kod automatskog režima rada crpka se povremeno uključuje. Minimalno vrijeme rada i vrijeme odmaranja crpke određuje operater. Kod automatskog rada svaki put se uključuje druga crpka zbog ravnomjernog trošenja redundantnih crpki. Kada nivo mulja u spremniku padne do zaštitnog nivoa, crpke se ne mogu uključiti. Za signalizaciju zaštitnog nivoa mulja u spremniku koristi se nivo plovak.



Slika 3 Shema spoja crpki sekundarnog mulja

F. Ugušćivač i skladište mulja

Višak sekundarnog mulja povremeno se prebacuje u bazen za ugušćivanje i skladištenje sekundarnog mulja. Tu se mulj miješa pomoću miješalice. Na površini se izdvaja višak vode iz mulja (muljna voda) koja se preko površinskog preljeva i pripadajućeg cjevovoda odvodi u crpilište izdvojenih muljnih voda. Iz ugušćivača i skladišta mulja operater po potrebi pomoću crpki prebacuju ugušćeni mulj u cisterne za odvoz mulja.

Volumen ugušćivača i skladišta sekundarnog mulja je 200 m3, dimenzija Ø=7 m , h= 5.2 m . Vrijeme zadržavanja sekundarnog mulja je 14 dana.

Montirana miješalica mulja nazivne snage 0,26kW s kojom se upravlja preko operatorskog panela isto kao sa crpkom nazivne snage 0,75kW. Kod mehaničkog kvara na miješalici zatvara se kontakt njezine krajnje sklopke pa PLC isključuje miješalicu.

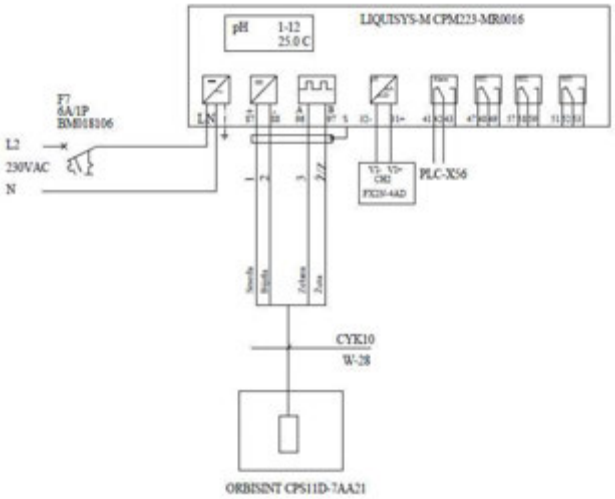
G. Crpilište izdvojenih muljnih voda

U crpilištu izdvojenih muljnih voda se skuplja otpadna voda iz sanitarnih čvorova pogonske zgrade, oborinska voda sa uređaja te višak vode iz uguščivača i skladišta mulja. Pomoću crpki sa kojima upravlja upravljački ormar otpadna voda se prebacuje na početak mehaničkog pročišćenja. Preko operatorskog panela omogućen je odabir režima rada automatski ili ručno te je prikazan broj radnih sati pojedine pumpe. Ručno uključena crpka mora se isključiti kad nivo muljne vode u crpilištu padne do nivoa isključenja. U automatskom režimu rada crpkama upravlja PLC ovisno o stanju nivo plovaka smještenih u crpilištu s ciljem održavanja određenog nivoa muljnih voda. Uključenje crpki izvedeno je tako da se svaki put uključuje druga crpka zbog ravnomjernog trošenja redundantnih crpki.

H. Izlazni kontrolni mjerni žlijeb

Nakon biološkog pročišćavanja pročišćena otpadna voda prolazi kroz kontrolni mjerni žlijeb prije ispuštanja u okoliš. U kontrolnom mjernom žlijebu postavljeni su senzori za mjerenje protoka, pH vrijednosti vode, temperature vode i mjerenje kisika u vodi. Na temelju izmjerenih vrijednosti analizira se kvaliteta pročišćavanja i provjerava se dali pročišćena otpadna voda zadovoljava zadane propise o otpadnim vodama.

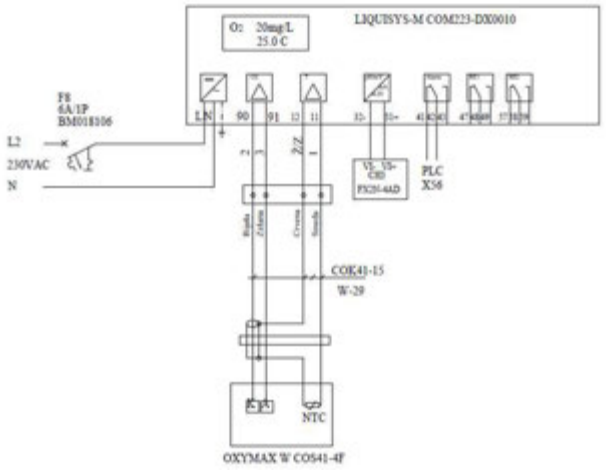
Za mjerenje pH vrijednosti pročišćene otpadne vode koristi se mjerni pretvornik *CMP223-MR0016* i mjerna sonda *CPS11D-7AA21* proizvođača *Endress+Hauser* [6]. Mjerni pretvornik konstantno vrši samokontrolu sebe i mjerne sonde te u slučaju greške ju signalizira. Montiran je u upravljačkom ormaru, a mjerna sonda ugrađena je u armaturno kućište koje se nalazi u kontrolnom mjernom žlijebu. Isti princip je primijenjen i kod mjerenja otopljenog kisika u bioaeracijskom bazenu.



Slika 4 Shema spajanja mjerne sonde i mjernog pretvornika pH vrijednosti tekućine

Mjerenje količine otopljenog kisika u pročišćenoj otpadnoj vodi vrši se sa mjernim pretvornikom *COM223-DX00100* i mjernom sondom *COS 41-4F* od proizvođača *Endress+Hauser* [5].

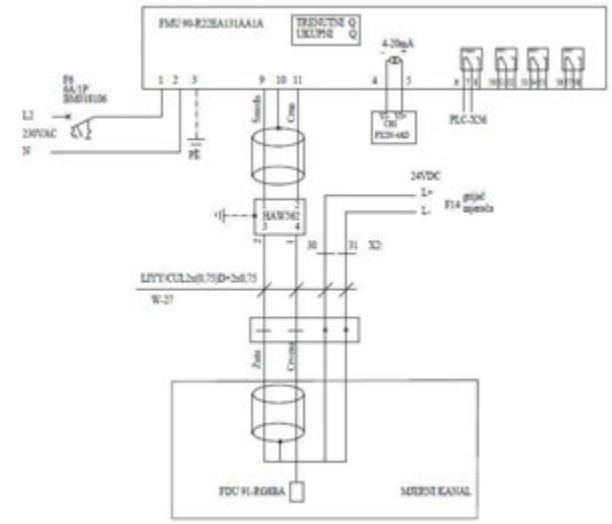
Mjerna sonda za mjerenje kisika sastoji se od plastičnog kućišta sa membranskim poklopcem u kojem se nalaze dvije elektrode i elektrolit. Kisik iz medija prolazi kroz membranu i stvara vodljivo područje između dvije elektrode čiji otpor ovisi o količini kisika. Na tom principu mjerna sonda daje strujni signal razmjeran količini kisika mjernom pretvorniku. Mjerni pretvornik iz tog signala izračunava količinu kisika u mediju. Mjerna sonda mjeri i temperaturu medija, jer temperatura utječe na rezultate mjerenja pa ju mjerni pretvornik uzima u obzir kod izračuna.



Slika 5 Shema spajanja mjernog pretvornika i mjerne sonde količine otopljenog kisika

Mjerenje protoka vrši se sa mjernim pretvornikom *FMU90-R22EA* i mjernim senzorom *FDU91-RG8BA* od proizvođača *Endress+Hauser* [7]. Mjerni senzor odašilje pakete ultrazvučnih impulsa koji se odbijaju od površine tekućine i vraćaju u senzor. Mjerni senzor mjeri vrijeme koje prođe između odašiljanja i povratka impulsa. Na temelju izmjerenog vremena i poznate brzine zvuka mjerni senzor izračunava koliki je nivo tekućine u mjernom kanalu te tu vrijednost prosljeđuje mjernom pretvorniku.

Mjerni pretvornik izračunava vrijednost protoka izlazne otpadne vode na temelju upisanih dimenzija mjernog kanala i nivoa tekućine u mjernom kanalu. U mjerni senzor ugrađen je grijač da bi mogao raditi na što nižim temperaturama.



Slika 6 Shema spajanja mjernog pretvornika i mjernog senzora protoka

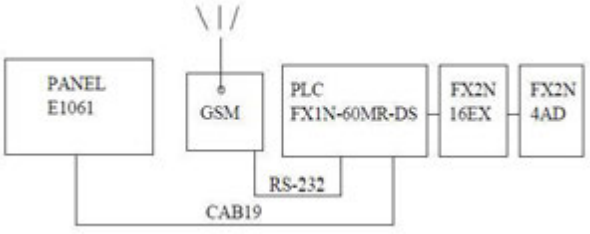
III. UPRAVLJAČKA ELEKTRONIKA

Upravljačka elektronika mora ispuniti sve zahtjeve upravljanja procesom koji su prethodno navedeni. U današnje vrijeme kod upravljanja takvim procesima gotovo uvijek se koriste programibilni logički kontroleri. PLC-i omogućuju upravljanje i nadzor nad procesom u širokom spektru aplikacija i odlikuju se sa velikom pouzdanošću u radu, fleksibilnošću, adaptivnošću, jednostavnošću programiranja i širokim aspektom komunikacije sa okolinom.

A. Programibilni logički kontroler – PLC

Odabran je PLC tip *FX1N-60MR-DS* od proizvođača *Misubishi* sa modularnom jedinicom za proširenje digitalnih ulaza *FX2N-16EX-ES*, modularnom jedinicom za proširenje analognih ulaza *FX2N-4AD*, komunikacijskim adapterom *FX1N-232-BD* i GSM modemom *MIM-G01*. Odabrani PLC spada u mikro modele PLC-a koji se obično koriste za automatizaciju manje složenih procesa. PLC se odlikuje jednostavnošću, dobrom funkcionalnošću i ima mogućnost jednostavnog proširenja sa različitim modulima[3].

Spajanje osobnog računala i PLC-a radi se sa originalnim kabelom koji se spaja na *RS-232* izlaz od računala i na konektor za programiranje od PLC-a. Komunikacijski adapter *FX1N-232-BD* je prilagodni sklop koja se montira u PLC i priključuje na konektor za dodatnu opremu te omogućava komunikaciju PLC sa različitim modulima sa *RS-232* komunikacijskim protokolom. U ovom slučaju komunikacijski protokol *RS-232* koristi se za komunikaciju PLC-a i GSM modula. Digitalni i analogni modul za proširenje spajaju se na PLC preko konektora za proširenje, sa time da se prvi spaja u PLC, a sljedeći na prvi itd. Operatorski panel je spojen na PLC s originalnim namjenskim kabelom *CAB19* preko konektora za programiranje PLC-a. Napajanje PLC-a, panela i ostale elektronike vrši se iz ispravljača koji se napaja iz uređaja za besprekidno napajanje.



Slika 7 Shema spajanja PLC-a sa ostalim modulima

B. Modul za proširenje digitalnih ulaza

Modul za proširenje digitalnih ulaza je *FX2N-16EX-ES* od proizvođača Mitsubishi i kompatibilan je sa PLC-ima serije *FX1N*. Modul se sastoji od 16 digitalnih ulaza sa naponom prorade ulaza od 24 VDC. Modul se napaja preko komunikacijskog konektora od PLC-a pa ne treba dodatno vanjsko napajanje. Jednostavno se implementira u sustav i PLC ga automatski prepoznaje kao nastavak svojih ulaza pa ne treba dodatno definirat pozicije tih ulaza.

C. *Modul za proširenje analognih uzlaza*

Modul za proširenje analognih ulaza je *FX2N-4AD* od proizvođača *Mitsubishi* i kompatibilan je sa PLC-ima serije *FX1N*. Jednostavno se implementira u sustav i PLC ga automatski prepoznaje kao svoji ulaz pa ne treba dodatno definirat pozicije tih ulaza. Modul za proširenje analognih ulaza ima velike mogućnosti prilagodbe procesu. Prilagodbe se odnose na područje mjerenja analognog ulaza, brzinu A/D pretvornika, odabira broja uzoraka za izračun srednje vrijednosti signala (filtar signala od smetnje). Odabir načina rada modula za proširenje analognih ulaz vrši se softverski i detaljnije je opisan u poglavlju softverski opis. U ovom slučaju analogni ulazi su podešeni za mjerenje signala od 4 mA do 20 mA.

D. *GSM modem*

GSM modem je *MIM-G01* od proizvođača Mitsubishi i kompatibilan je sa PLC-ima serije *FX1N*. *MIM-G01* je GSM/GPRS mobilni modem koji služi za prijenos podataka, SMS poruka i slanje e-mail-ova. Za ispravan rad potrebna mu je SIM kartica logirana ma mobilnu mrežu i vanjsko napajanje 10–40 VDC. Komunicira sa PLC-om preko *RS-232* komunikacijskog protokola. Na GSM modul montirana je antena *MAM-ANT-ANG* od proizvođača *Mitsubishi*. GSM modem jednostavno se montira na *DIN 35* montažnu šinu.

E. *Korisničko sučelje – operatorski panel*

Za korisničko sučelje koristi se operatorski panel osjetljiv na dodir tip *E1061* od proizvođača *Beijer Electronics*. Operatorski panel je kompatibilan sa *Mitsubishi*-evim PLC-ima serije *FX1N*. Komunikacija između panela i PLC-a izvedena je sa originalnim namjenskim kabelom *CAB19* koji je na operatorskom panelu priključen na 25 – pinski D ženski konektor, a na PLC-u na konektor za programiranje. Premda su PLC i operatorski panel kompatibilni automatski se povezuju i prepoznaju. Program za operatorski panel je napravljen u programu *E-Designer 7.5*.

IV. IZRADA GRAFIČKOG SUČELJA OPERATORSKOG PANELA

E-Designer 7.5. je aplikacija koja se koristi za izradu grafičkog sučelja za *Mitsubishi*-eve i njima kompatibilne operatorske panele. Jednostavna je za korištenje, ima dovoljnu veliku bazu vlastitih simbola, ima mogućnost izrade novih simbola, kompatibilna je sa *Mitsubishi*-evim PLC-ima i ima mogućnost simulacije rada operatorskog panela što olakšava i ubrzava izradu grafičkog sučelja. Grafičko sučelje operatorskog panela je koncipirano tako da se iz početnog ekrana na kojem su prikazane osnovne funkcije cijelog sustava funkcijskim tipkama ulazi u ostale ekrane koji prikazuju pojedine dijelove sustava.



Slika 8 Struktura ekrana operatorskog panela

V. IZRADA UPRAVLJAČKOG PROGRAMA

Upravljački program PLC-a i grafičko sučelje operatorskog panela izrađeni su u softverskim aplikacijama *GX Developer 8.9.* i *E-Designer 7.5.*. Aplikacija *GX-Developer 8.9.* koristi se za izradu upravljačkog za *Mitsubishijeve* PLC-e, a aplikacija *E-Designer 7.5.* koristi se za izradu grafičkog sučelja operatorskog panela[4]. *GX Developer 8.9.* omogućava izradu upravljačkih programa u ladder i SFC (engl. *Sequential Function Chart*) tehnici. U ovome radu upravljački program je izrađen u ladder tehnici.

VI. UPRAVLJAČKI ORMAR

Upravljačko ormar koncipiran je kao središnje mjesto sa kojeg operater upravlja i nadgleda radom pročistača otpadne vode. U njega je ugrađena sva opisana oprema koja se koristi za upravljanje i napajanje elektrotehničkih elemenata pročistača otpadne vode. Ostala oprema nalazi se u ormaru montirana na montažnu ploču. Meki upuštači, ulazni prekidač i UPS (engl. *Uninterruptible Power Supply*) montirani su direktno na montažnu ploču, a ostali elementi montirani su na montažnu ploču preko *DIN 35* nosača.

VII. ZAKLJUČAK

Ugradnjom upravljačkog sustava u pročistač otpadnih voda postignuti su željeni rezultati te se olakšalo upravljanje i održavanje uređaja. Upravljanje sa pročistačem se vrši sa jednog mjesta što olakšava rad operateru. Lako je pratiti alarme pomoću grafičkog sučelja koje olakšava pronalazak kvara i održavanje te u slučaju bilo kojeg kvara na uređaju PLC preko GSM modula šalje SMS poruku operateru i uključuje signalnu sirenu. Upravljački ormar je koncipiran tako da može zadovoljiti nadogradnju i poboljšanje uređaja. Moguće poboljšanje bi bila promjena načina upravljanja crpki crpilišta sekundarnog mulja koje rade s predodređenim vremenom rada i odmora. Takav način rada je rad bez povratne informacije iz sustava i ne ovisi o promjeni stanja u sustavu. Ugradnjom frekventnog pretvarača za upravljanje radom crpki i proširenjem PLC-a sa dodatnim modulom koji ima analogni izlaz moguće je upravljati sa radom crpki ovisno o izlaznom protoku otpadne vode.

LITERATURA

[1] M. Ilijanić : Automatizacija pročistača otpadnih voda, Tehničko veleučilište u Zagrebu, Elektrotehnički odjel, Zagreb 2013

[2] B. Tušar: Pročišćenje otpadnih voda, Geotehnički fakultet, Zagreb 2009

[3] Goran Malčić: Programabilni logički kontroleri, skripta za kolegij Procesna računala, Tehničko veleučilište u Zagrebu, Elektrotehnički odjel, Zagreb, 2006.

[4] Mitsubishi, Melseg FX family, PLC beginners manual, 2006 do 2020. https://si3a.mitsubishielectric.com/fa/sl/dl/2340/166388.pdf

[5] Endress + Hausser, Dissolved oxygen measurement. https://portal.endress.com/wa001/dla/5000000/6574/000/01/FA010-E.pdf

[6] Endress + Hausser, pH/ORP measurement, tehnical information. https://portal.endress.com/wa001/dla/5000319/0382/000/04/TI00194CEN_1413.pdf

[7] Endress + Hausser, Prosonic S FMU90, tehnical information. http://www.merteh.lv/eh/pdf/TI00397FEN.PDF

[8] Danfoss, MCD 200, Design Guide 2005. http://files.danfoss.com/download/Drives/doc_A_1_mg17c222.pdf

CIS

International Conference on

INTELLIGENT SYSTEMS

Steering Committee

Chairs: Slobodan Ribarić, University of Zagreb, Croatia
Andrea Budin, Ericsson Nikola Tesla, Zagreb, Croatia

Program Committee

Members: Patrizio Campisi, Roma Tre University, Italy
Bojan Cukic, UNC Charlotte, United States
Ivo Ipšić, University of Rijeka, Croatia
Marina Ivašić-Kos, University of Rijeka, Croatia
Zongmin Ma, Nanjing University, China
Neeta Nain, MNIT Jaipur, India
Nikola Pavešić, University of Ljubljana, Slovenia
Vitomir Štruc, University of Ljubljana, Slovenia
Zheng-Hua Tan, Aalborg University, Denmark

PAPERS

Automation of a Decision Tree Conversion into a Fuzzy Inference System Using ANTLR

S.S. Sosinskaya*, R.S. Dorofeev *, A.S. Dorofeev *and T.R. Usenko*
* Irkutsk National Research Technical University, Irkutsk, Russian Federation
sosinskaya@mail.ru

I. ***Abstract*** **- The paper discusses techniques of processing a sample of sets of numerical features of observations that relate to a certain subject area and belong to certain classes. Such techniques include well-known methods of constructing a decision tree and a fuzzy inference system. An isomorphism of the decision trees and a corresponding fuzzy inference system rule set is being justified. Algorithms of both methods are described in special languages. An approach to automated conversion of the decision tree to a fuzzy inference system using ANTLR, a tool for creating compilers, is proposed. The toolkit used, along with the creation of classes for lexical and parsing of a description in one language, allows generate a class for converting text from one language to another. The relevance of the approach is that with representing fuzzy classifying knowledge allows one to implement an expert system, allowing domain specialists to classify objects. Usage of the decision trees in expert systems is problematic. An example of applying this approach to a classification of leaf specimens originating from different plant species is given.**

Keywords - decision tree, fuzzy inference system, compiler, isomorphism, classification, classification quality, ANTLR

I. INTRODUCTION

Intelligent data analysis is one of the most important areas of research in the modern world. The necessity of processing large arrays of data, over several years, Data Mining is being used, in particular, for classification.

Data classification is defining rules and algorithms of their acquisition for the proper grouping of the input samples. Each group (class) is characterized by certain features, which belong to all cluster objects. The number of classes is provided in advance as a result of analysis of domain terms.

For the classification implementation, techniques for building decision trees and fuzzy inference system are used. Both techniques are implemented by the authors using the Matlab system.

The classification process consists of three stages: model training, its testing and use.

Both techniques provide testing stage, after training the model. It can be shown that the decision tree (Tree) and fuzzy inference system (FIS) are isomorphic. At present, conversion is performed with manually writing scripts. A system of automated conversion is proposed both in one and in the opposite direction based on tools for creating ANTLR compiler. The application of the proposed approach is illustrated by the classification of the Fisher's iris specimens.

….

II. METHODS

A. Binary decision tree

Binary decision trees are the most common technique of solving the problem of classification of multidimensional data. A decision tree is a graph whose nodes represent the conditions imposed on the values of the input features; the output (classification) feature is represented with values associated with the tree leaves. The branches of the tree represent paths between intermediate nodes and leaves. From each node originates two branches according to the compliance and noncompliance to the node condition [1,5,6].

The binary tree built upon the input observations and their corresponding values of the output feature can be represented as a graph, or as a set of text rules on a special language.

B. Fuzzy inference system

Fuzzy inference system is a technique of obtaining the output data about the object (for example, belonging to a specific class) based on the input data, which describe the information about the state of the object.

Let *MFc(x)* be the degree of the membership of x to fuzzy set *C*, which implements a generalization of the concept of the characteristic function of ordinary set.

The set of ordered pairs of the form

$$C = \{MF_c(x), x\},\ MF_c(x) \in [0,1] \quad (1)$$

is called a fuzzy set.

Fuzzy variables and their terms are used to describe the fuzzy sets. In the paper for the term membership functions, the triangular function is used [4,7-10].

For the implementation of fuzzy inference, a database of rules is used built on the base of terms of input and output variables; each rule has the following form:

IF a premise THEN a conclusion

In this case, the premise is a combination of terms of the input variables and conclusion is a term of the output variable. As a problem of classification is solved, a knowledge base of Sugeno-type is formed. The conclusions of the rules are set not by fuzzy terms, but a linear function of the inputs.

B. Automated conversion Tree->FIS

A decision tree is isomorphic to a corresponding fuzzy inference system, as each branch of the tree from root to leaf is a rule; each condition for input variable specified in the tree node is its term.

More strictly, let's define a decision tree as Tree={V,A,L} and a fuzzy inference system as Fis ={TIN,R,TOUT}, where

- V is set of vertices;
- A is set of arcs;
- L is set of leaves;
- TIN is set of terms of the input variables;
- R is set of rules;
- TOUT is set of terms of the output variable,

then Tree ≡Fis, as V≡ TIN, A≡ R, L ≡TOUT.

III. Related works

Many researches noticed the presence of isomorphism and used various methods and approaches to transform Tree->FIS. This issue is devoted to the work of such researchers as Bertsimas D., Dunn J., Narodytska N., Ignatiev A., Pereira F., Marques-Silva J., RAS I., Sosnowski Z., Gadomer L., Bakhshipour A., Zareiforoush H., Bagheri I., Ge D., Zeng X., Yakupov A.I., Kulchin Yu.N., Kim A.Yu., Notkin B.S., Luchter A.B. etc. Their work reflects such issues as optimal classification of trees, methods and algorithms for processing trees, the use of trees for various modeling, etc.In addition, a fairly large number of works devoted to the fuzzy inference system, This can be seen in the works of such authors as Sosnowski Z., Gadomer L., Munir K., de Ramon-Fernandez A., Iqbal S., Freitas E., Anderson P., de OliveЭти работы посвященыira Serra G., Luiz M., Radhika S., Rangarajan P. И т.д. The works of these authors are devoted to such important issues as data mining and knowledge discovery, identification using big data and fuzzy logic, state-space recursive identification algorithm based on evolving neural-fuzzy inference system, improving the lifespan of wireless sensor networks with fuzzy based cluster-ing and machine learning based data reduction

Those of them that performed the Tree-> FIS transformation performed it either mathematically or by describing fuzzy inference systems based on a specific decision tree. The novelty of the approach of the authors of this article is an attempt to automate this transformation by developing a formal grammar of the decision tree description language, on the basis of which the ANTLR system makes it possible to perform this transformation every time when necessary.

IV. Conversion of a decision tree to a fuzzy inference system

Automated conversion Tree->FIS is performed with ANTLR (Another Tool for Language Recognition), software designed to develop the compilers and interpreters for programming languages of different levels [2,3].

We described the grammar of a language for representation of a decision tree by means of ANTLR. In this grammar the rules for terminals symbols and nonterminal symbols are defined. In ANTLR rules, the terminals are marked with capital letters, the nonterminals with small ones. A fragment of the grammar is shown (see Fig. 1).

```
program  :  ( stmt ) +  ;
stmt  :  idx=INT { emitter.startRule($idx.text); } ifexpr
  |  idx=INT { emitter.startRule($idx.text); } leafexpr  ;
leafexpr  : CLASSKW LETOPERATOR v=CLASS {
emitter.processLeaf($v.text); }  ;
ifexpr  :  IFKW logicexpr THENKW nodeexpr elseifexpr
  |  IFKW logicexpr THENKW nodeexpr elseexpr
  |  IFKW logicexpr THENKW nodeexpr  ;
elseifexpr : ELSEIFKW logicexpr THENKW nodeexpr
elseifexpr | ELSEIFKW logicexpr THENKW nodeexpr
elseexpr
  |ELSEIFKW logicexpr THENKW nodeexpr  ;
elseexpr:ELSEKW v=CLASS { emitter.processLeaf($v.text); }
;
nodeexpr: NODEKW v=INT { emitter.processNode($v.text); }
;
logicexpr : f=FEATURE oper=LOGICOPERATOR v=FLOAT
{ emitter.processFork($f.text, $oper.text, $v.text); }  ;
```

Figure 1. A fragment of the ANTLR grammar for the language of decision tree description

Based on the described grammar, ANTLR generates lexical analyzer and parser classes of an interpreter for converting the description of a decision tree in a description of the fuzzy inference system.

One can supplement an ANTLR grammar text description with actions on the target programming language, Java in our case. When the grammar description is compiled, this code is copied to the target parser program.

In particular, Emitter class is included in the action part of rules. The class contains methods emitting target code of fuzzy inference system description based on the tree description analyzed by parser.

The Emitter class contains internal classes:

- Segment — a description of the attribute's value range,
- Condition — a description of the condition, and
- TreeNode — a description of the tree node.

The Emitter class has the following attributes:

- featureValues – features encountered during the analysis of the input file and their corresponding values;
- rulesForOutput – numbers of forming rules and corresponding rules texts;
- featureList – features and their areas for each rule.

Most of them are represented by the HashMap collection, which stores the key (for example, the name of the rule) and an array of instances of the Segment class).

The Emitter class has some methods to handle nonterminal characters:

- processLeaf – creating node
- processNode – filling node;
- processFork- processing a condition in a node and to create a fuzzy inference system:

• generateSegment – interval generation variable values based on values from a tree;
• addElseForkOperations - processing the else branch;
• addThenForkOperations – processing the then branch.

From the generated classes, the interpreter processing the text description of decision tree is formed, with checking lexical and syntactic errors and generating source code for Tree->FIS conversion.

The structure of the developed system is shown (see Fig. 2).

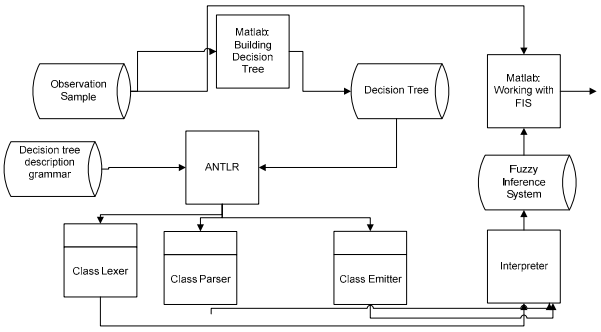


Figure 2. The structure of the system for automated Tree->FIS conversion.

For user convenience, a shell is implemented in the form of a command interface.

A. Example

As an example illustrating the efficiency of application of this approach, let us consider a set of integer features of plant leaves of 30 kinds [11]. There are 340 items of plant leaves in the set. Each of the items is described with 14 features. Features No. 1-8 represent the leave form, 9-14 represent a digital image of the leave. Names of the features and their abbreviations are presented in Table I.

We divided the set on two subsets: one for training a decision tree, and one for testing. The training subset was used to design a binary tree. The following is a graphical and a textual representation of the constructed tree (see Fig. 3-4).

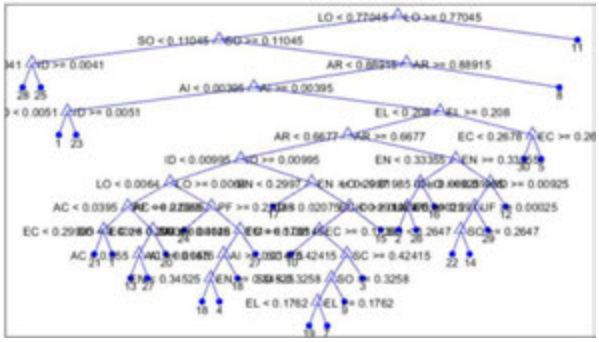


Figure 3. A decision tree for the leaves

TABLE I. NAMES OF FEATURES OF LEAVES AND THEIR ABBREVIATIONS

Names of features	*Abbreviations*
Eccentricity	EC
Aspect Ratio	AR
Elongation	EL
Solidity	SO
Stochastic Convexity	SC
Isoperimetric Factor	PF
Maximal Indentation Depth	ID
Lobedness	LO
Average Intensity	AI
Average Contrast	AC
Smoothness	ST
Third moment	TM
Uniformity	UF
Entropy	EN

```
Decision tree for classification
 1  if LO<0.77045 then node 2 elseif LO>=0.77045 then node 3 else 11
 2  if SO<0.11045 then node 4 elseif SO>=0.11045 then node 5 else 9
 3  class = 11
 4  if ID<0.0041 then node 6 elseif ID>=0.0041 then node 7 else 25
 5  if AR<0.88915 then node 8 elseif AR>=0.88915 then node 9 else 13
 6  class = 28
 7  class = 25
 8  if AI<0.00395 then node 10 elseif AI>=0.00395 then node 11 else 9
 9  class = 8
10  if ID<0.0051 then node 12 elseif ID>=0.0051 then node 13 else 23
11  if EL<0.208 then node 14 elseif EL>=0.208 then node 15 else 13
12  class = 1
13  class = 23
14  if AR<0.6677 then node 16 elseif AR>=0.6677 then node 17 else 9
15  if EC<0.2678 then node 18 elseif EC>=0.2678 then node 19 else 5
```

Figure 4. A fragment of text description of a decision tree for the leaves

In the result of the processing of the interpreter, a fuzzy inference system was generated (see Fig. 5). This system has 13 input features with different number of terms, 32 rules and output classification feature with 29 terms.

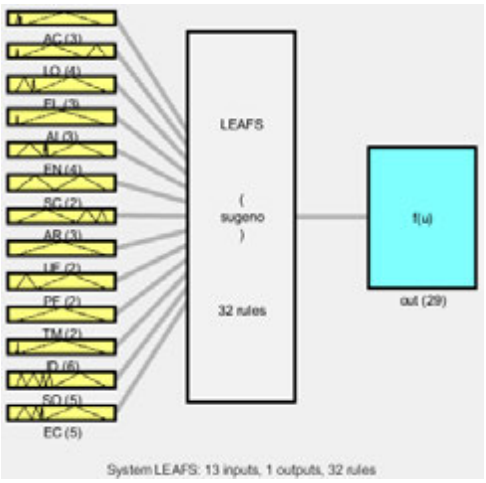


Figure 5. Graphical representation of fuzzy inference system

B. *Expert system development*

The obtained FIS was used to implement a GUI application an expert that staties questions. User answers form the values for the input variables.. Their fuzzifications are used in the fuzzy inference, resulting in espert decision, the class name of a plant. The example of the expert consulting is shown in Fig. 6.



Figure 6. The example of the expert consulting

V. CONLUSION

The work done allows us to conclude that the proposed approach of conversion automation is time efficient, as the grammar description language for decision trees does not depend on the domain requires little labor and time.

Further development of the system expected to be an application in domains, where selected observations contain enough number of items classifiers of a better quality using both decision trees and fuzzy inference system with respect to the number of classes. This allows us to construct data.

REFERENCES

[1] Bertsimas, D., Dunn, J.: Optimal classification trees. Machine Learning 106(7), 2017, pp.1039–1082

[2] Software ANTLR, http://www.antlr.org/, last accessed 2019/10/28.

[3] Cheng, Z.: A Proposal for a Generic Translation Framework for Boogie Language. 26th European Conference on Object-Oriented Programming (Doctoral Symposium) (2012), http://www.cs.nuim.ie/~zcheng/paper/ECOOP2012.pdf, last accessed 2019/10/28.

[4] Dangeti, P.: Statistics for Machine Learning. By Pratap Dangeti, 2017, https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781788295758, last accessed 2019/10/28.

[5] Narodytska, N., Ignatiev, A., Pereira, F., Marques-Silva, J., RAS, I.: Learning optimal decision trees with sat. In: IJCAI. 2018, pp. 1362–1368

[6] Sosnowski, Z., Gadomer, L.: Fuzzy trees and forests-Review WILEY INTERDISCIPLINARY REVIEWS-DATA MINING AND KNOWLEDGE DISCOVERY, 9 (6), 2019, number of article e1316

[7] Ge, D., Zeng, X.: Learning data streams online - An evolving fuzzy system approach with self-learning/adaptive thresholds INFORMATION SCIENCES Vol. 507, 2019,.pp. 172-184

[8] Munir, K., de Ramon-Fernandez, A., Iqbal, S.: Neuroscience patient identification using big data and fuzzy logic-An Alzheimer's disease case study EXPERT SYSTEMS WITH APPLICATIONS, Vol. 136, 2019, pp. 410-425

[9] Freitas, E., Anderson, P., de Oliveira Serra, G., Luiz M.: State-Space Recursive Identification Algorithm Based on Evolving Type-2 Neural-Fuzzy Inference System JOURNAL OF CONTROL AUTOMATION AND ELECTRICAL SYSTEMS, 2019

[10] Radhika, S., Rangarajan, P.: On improving the lifespan of wireless se nsor networks with fuzzy based cluster-ing and machine learning based data reduction APPLIED SOFT COMPUTING, Vol. 83, 2019,.number of article UNSP 105610

[11] Pedro F. B. Silva, Andre R.S. Marcal, Rubim M. Almeida da Silva. Evaluation of Features for Leaf Discrimination // Springer Lecture Notes in Computer Science. 2013. Vol. 7950. P 197-204.

[12] Bakhshipour, A., Zareiforoush, H., Bagheri, I.: Application of decision trees and fuzzy inference system for quality classification and modeling of black and green tea based on visual features, Journal of Food Measurement and Characterization, Vol. 14, 2020, p 1402–1416

[13] Reza, M., Asad, A., A., Borghei. A., M., Ahmadi, H., Minaei, S., Beheshti. B.: Fuzzy logic based classification of faults in mechanical differential, Journal of Vibroengineering, Vol. 17, Issue 7, 2015, p. 3635-3649

[14] Yakupov, A., I.: Application of decision trees for modeling the creditworthiness of clients of a commercial bank. // Artificial Intelligence, Vol 4, 2008, p. 208-213 translation

[15] Chaware, A., Lanjewar, U., A.: Derived Fuzzy Rules for Predicting the Students Acedemic Performance // IOSR Journal of Computer Engineering, Vol. 16, Issue 6, p 53-60

[16] Kulchin, Yu., N., Kim A., Yu., Notkin B., S., Luchter A., B .: Building a fuzzy decision tree algorithm based on experimental data when processing RVOIS signals // Informatics and Control Systems, Vol. 3, Issue 41, 2014,p. 103-114. translation

A Randomized Load Balancing Criteria using Traffic flow in SDN

Prakash Choudhary* and Pradeep Thota
Department of Computer Science and Engineering
National Institute of Technology Hamirpur, Himachal Pradesh-177005, INDIA
*Corresponding author Email: pc@nith.ac.in

Abstract— Software Defined Networks (SDN) is a trending technology in the computer science field, overriding traditional computing. It gives full-fledged control over the network and administrators to manage and update the respective network according to its needs by operating the control plane of the network. SDN helps to archive the global view of the network and the data abstraction. Today's drastically increasing network suffering from the overutilization problem of available resources. Therefore, the collisions and bottleneck bandwidth of the network is growing and hence, leading to the issue of degradation of performance and throughput. All these issues are occurring due to improper load balancing schemes and algorithms used to control the load balancing. In this paper, we are presenting the algorithm named Pathfinder, a randomized pathfinding algorithm similar to the breadth-first search (BFS) algorithm but unlike BFS, it is increasing the performance and link utilization of the system. It is dynamic traffic monitoring, and balancing algorithm will shift flow to the path which has less traffic and makes sure that almost all of the resources equally utilized.

Index Terms—Software Defined networking, OpenFlow, Routing, Controller and Pathfinder

1 INTRODUCTION

Traditional networks have limited control over networking resources, hard to configure because of predefined rules, automatic reconfiguration, and absence of virtually response mechanisms [5]. Moreover, it is also difficult to embed the high-level networking policies and any modifications demanded by the company or to integrate the new up-gradation of network software [8]. The network operator is implementing low-level commands or company-specific commands for every device which is a hectic and time-consuming process, but many protocols and routing algorithms are inventing in practice to handle these issues [1]. Each protocol and algorithm have its advantage, some arise with safety, and few emerge with efficiency. Therefore, it depends on the needs and selection made accordingly. Further, to add inventions, we have to add these all to every individual device though they are controlled in a distributed manner. Some inventions have been made by researchers, industries, and organizations towards efficient communication, which would enable logically centralized control over network devices. The advancement in the field of communication and information technology aiming to provide full-fledged control over networking devices and can be modified anytime by the programmers to provide fast, efficient, and safe communication. Software Defined Networking (SDN) is a substantial invention in the present era of networking. Unlike traditional networking, SDN gives control to the programmers to modify the existing rules and regulations as per need. Also, SDN breaks the vertically integrated nature of the networking resources by separating the network control logic from it [1] [8].

In the SDN architecture, as shown in fig.1, there are three main modules; the first one is data place, we can assume this layer as a servant because it does not have any command and control over decision making. It will just listen or implement whatever instruction given by the control plane. Second, the control plane, we can call it as the brain of the device, and it acts as a decision-maker, rules generator, and it also decides the direction to the traffic from which it has to go. In the SDN the control and data planes are decoupled, network intelligence and state are logically centralized, and the underlying network infrastructure is abstracted from the applications. Third, a centralized controller makes forwarding decisions on behalf of the network forwarding elements using a set of policies. In general, the connection among nodes (source and destination) is given by end-to-end communication, whereas the path between the nodes is provided by point-to-point or routing policy. In the SDN the interaction between the layers is managed by SouthBound Interface (SBI) and NorthBound Interface (NBI). SBI acts as an intermediary between the data plane and control plane and takes responsibility for protocol management. Whereas, NBI interface provides the API for the developers to manage and deploy their applications [8].

1.1 Load Balancing

For transferring the data from source to destination, it has to go through the links among the networking resources such as switches, routers, etc. The algorithms designed to control the flow may always choose the same path instead of the unused path. Therefore, going through the same way makes some resources hectic and the congestion of the link. Load balancing is a strategy to handle this kind of situation by avoiding the overutilization of

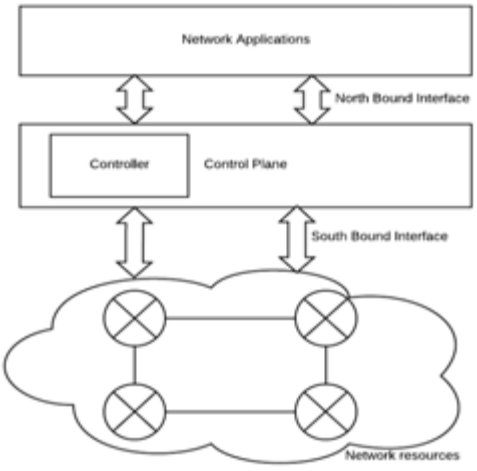


Fig. 1. The architecture of Software Defined Network

the same path multiple times. Hence, the performance of the system can be improved in load balancing. In any networking environment, the nodes have different processing speeds, and capabilities are interconnected by dedicated channels with a specific bandwidth, frequency, etc.

The flow among the nodes has to be distributed evenly based on their processing speed so that the average processing speed to run the tasks, the idle time of the network resources gets minimized. This is the main reason to go for load balancing. It plays a vital role in improving the performance speedup and efficiency. The key issue in SDN is an effective strategy to distribute traffic among multiple nodes (e.g., switches, routers). Load balancing is the process of distribution or redistribution of load among nodes so that the performance of the system will improve by maximizing the throughput. The workload has to be evenly distributed among all the nodes based on their processing speeds, availability of paths among the nodes. In bursty networks, path selection plays an important role. In this paper, we focus on providing traffic monitoring and load balancing of SDN by analyzing the most relevant research works. In the literature, many comprehensive survey papers for SDN layers and open flow focused on a single selection criterion that may either be minimum hop count or feasible bandwidth. In this paper, we are presenting the randomized algorithm called Pathfinder; it may be the first randomized algorithm that uses bandwidth, hop count, and load present at the switch as selection criteria.

The rest of the paper includes section 2, an SDN related work. Section 3, presents the proposed model along with the algorithm. Section 4, shows an implementation with analysis followed by the conclusion presents in section 5.

2 RELATED WORK

Only adding new rules to the existing schemes may increase overhead in the networking devices. Advancement in the programming area may demand additional hardware support. It is not always possible that the same resources may support modifications and solve demands [7]. Wildcard rules are not always a feasible solution because it is hard to adapt to newly added rules. And, adding of new regulations may lead to change the entire networking devices rules. Failure of root node can't make communication between any two nodes. Because it is acting as a repository for everything. Flow tables are going to be massive if the size of the network is becoming large and it also leads to increase processing time. Change of topology creates difficulty in maintaining a global view of the system, which intern leads to difficulty in pathfinding while transferring the packets and it takes considerable time to implement the path found at the device level [6] [3]. Always going with the shortest path does not make any sense though it is a dynamic environment the probability of altering the shortest path will remain less.

Moreover, leaving the less loaded path (the links which are free or not bursty at least at a given time) is not a good idea in most of the circumstances. In [6], they increased the rate of interactivity like timely changes in the system topology. Topology service discovers the topologies according to the information on the link and the best path can be computing the minimum number of hops between the switches. Hop count should not always be our selecting criteria though it works fine. If time increases, there is a chance that the flow is going through the same hops and links, leaving most of the hops idle and that is worst in practice though it is good when compared to static routing schemes but, while we are implementing in the dynamic environment the results is not appropriate[9]. The advancement in SDN demands reinvestigation of the routing algorithms due to the change in progress in technology. Most of the researchers working consistently and persistently to reduce the complexity of the algorithms. In [2], used the Shortest Feasible Open Flow Path algorithm for computing the shortest path based on the bandwidth of the links which are connected among nodes. Here, they did the calculation only based on available bandwidth. Since all the links may not have equal bandwidth. Though the algorithm is working fine even in the worst case, it is an intern giving a hint that load has to maintain separately for efficient distribution. And the bisection bandwidth of the network depends on the structure. In [4], works on reducing the Ternary Content-Addressable Memory (TCAM) space focused on minimizing the switch entries for the forwarding packets will be the same as they mentioned. The authors have done exceptionally well in reducing TCAM space of the switch, but they increased the load at the particular switch. Always finding the central node and making the flow to go through the central node leading to misbalanc-

ing of load and in the case of failure of central node missing of scalability of the network, which is the biggest problem in the field of SDN.

3 PROPOSED MODEL OF SDN

The design of Pathfinder is to minimize the TCAM entities and effectively distribute the load among the switches. The algorithm Pathfinder executed as controlling application present in the SDN management plane. However, the beauty of the proposed model lies in the load balancing module, and its flow is presented in fig 2. The algorithm will consider three parameters such as bandwidth matrix of the current network, a load of a switch and topology of the network as input parameters and decide the path based on load present in the link, hop count and bandwidth randomly but at any given time it will use only one selection criteria. Before that, when a packet comes to the switch because of requests made by the hops connected, the switch will not take any action by itself. When flow comes into it, it will just forward to the controller then; the controller will intern calls the Pathfinder to know the path based on criteria as mentioned above for this we used open flow learning switches for fine-grained control. The algorithm is a part of the SDN controller containing three parameters which hop count, bandwidth, and load present at the switch. The controller maintains these; the load balancer does not have to worry about how they are managed. This is the

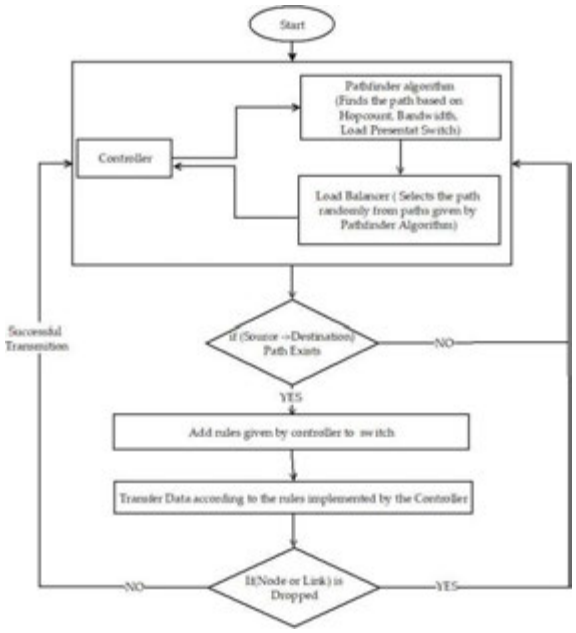


Fig. 2. Flow chart of the proposed Algorithm

best way of achieving the modularity and abstraction from the algorithm. The controller uses the advantage of implementing custom rules in open flow switches in the process of gaining control. The algorithm takes the source and destination mentioned by the controller, and from there the algorithm starts.

When the Pathfinder returns or founds its path, the load balancer module computes all three parameters and updates accordingly and finally decides its path randomly and returns the path to the controller. The controller will broadcast the path to every node to add the flow rule at each switch which is going to involve in the flow, and remaining switches drop the broadcasted rule. One of the significant advantages of this approach is that the switches which are not going to participate in the communication will never know the details of flow at any given time. Hence, it is solving the problem of security and encapsulation. Since all the nodes are connected to the controller, so there is a global view of the controller. The comprehensive look will help better in

Algorithm1 Pathfinder (G, N, S, D)

```
1:  procedure Pathfinder (G, N, S, D)
2: G is the Graph represents the topology of the network
3: N is the number of nodes in the network
4: S is the source vertex making the request
5: visited [] is a global array initialized to zero
6: Path[] is an array used to store the paths
7: The index is a global variable
8:      For < i=1; n > do
9:          visited[i] :=false;
10:         for < i=1; n > do
11:             visited[i] :=false;
12:             if visited[i] = 0 then
13:                 FindPaths(S ,D ,visited ,path ,index);
14: procedureFindPaths (S, D, visited, path, index)
15:     for all vertices adjacent from S do
16:         if visited[S] = 0 then
17:             Add S to path;
18:             visited[S] =true;
19:         if S == D  then
20:             for < i=1; index >do
21:                 return path;
22:         else
23:         if destination not yet found
24:         for all vertices adjacent to current vertex until
            we found a destination
25:         if visited[i] != true then
26:             Repeat Function: FindPaths ( S, D, visited,
                path, index)
27:         if a path is empty then
28:         return;
```

decision making while selecting the best path. In the case of failure of a node, the controller can immediately change the flow of the network dynamically.

G (N, L) is an undirected finite graph, where N= nodes of the network, L= links among the nodes. The path is

the sequence of nodes connected among themselves to connect the source to its destination. Let P1 = (sa, sb, sc,….. , si), P2 = (sa, sb, si,…., sj), P3 = (sb, sc, sd,…, sk) be the paths found by the algorithm along with three parameters mentioned above, and these paths returned to load balancer which will randomly choose the path and inform the controller. When a single selection criterion is implemented, irrespective of the number of times a host makes requests involving the paths mentioned above, the selection of path will be the same and its probability P(x) is always 1. The probability P(x) may alter if the topology of the network changes. But, in algorithm 1, we are reducing the probability of selecting the same path since we are founding many paths based on three different criteria. Let 'n' be the number of paths found by the algorithm. Here, in this case, the individual path probability P(x) = 1/n. Many load balancing methods developed are failed to reduce the individual path probability, but we tried to do in this paper. And significant fundamental steps of the Pathfinder are listed in algorithm 1.

4 Analysis and Results

The Pathfinder algorithm did not divide into any modules, but it will use the modules divided by the controller. The controller maintains topology, bandwidth, and load of the switches separately. The job of the load balancer is to choose the criteria and picks the best path randomly so that each node in the network or network resources used equally. Let T (N, L) be the maximum time taken by the Pathfinder algorithm on any network with N nodes and L edges. T = O (N + L) if the network is represented as adjacency lists. If its adjacency matric represents the network, then T(N, L) = O(N^2). If network is fully connected then the number of links are as: L = (N*(N-1)) /2 and the complexity of Pathfinder will be as: T(N,L) = O(N+(N*(N-1)) /2) then T(N, L) = O((N+ N2 – N)/2). Therefore, worst-case complexity will be O (N^2).

Simulation and Results: Mininet simulator is used to implement the pox controller and the network topology along with OpenFlow virtual switches. The Python programming language is used to design the controlling and managing routing schemes. Firstly, one python program is used to create the topology of the network. One more program is to read the bandwidth matrix, the topology of the network, and the load of the switches but it computes the path by using Pathfinder algorithms. In the next step, a Python programming module is written in the Pox controller to read the path and install the necessary rules to OpenFlow learning switches. This will parallelly update the flow and bandwidth matrices of the network.

Figure 3 and 4 represents the relationship between flows and link utilization of Round Robin, Dijkstra, and Pathfinder in different network dispensation. In all the cases maximum utilization of link is less than 0.8 while calculating the Dijkstra and Round Robin algorithms the link utilization rate is set to 1.0. In all the cases

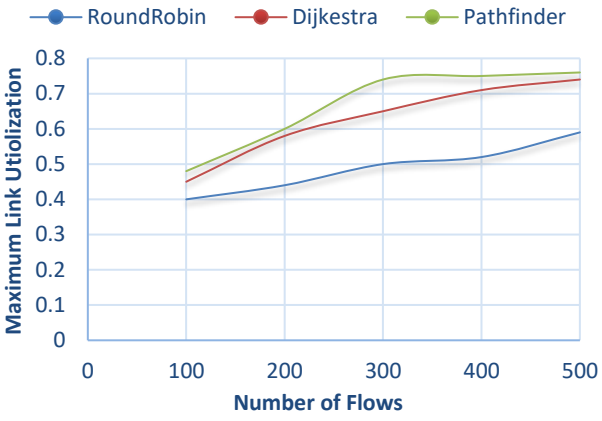


Fig. 3. Link Utilization (Abilene)

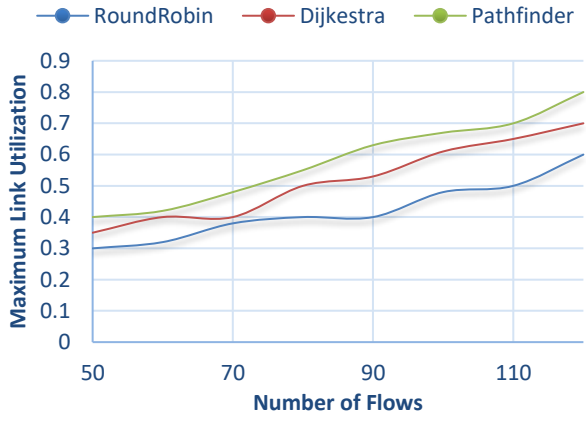


Fig 4. Link Utilization (Fat Tree)

Pathfinder is performing well when compared to above-mentioned algorithms. The maximum link utilization of Round Robin and Dijkstra algorithms are average 15-10 % less than that of *Pathfinder* in both Abilene and Fat tree networks.

5 CONCLUSIONS

In this paper, we proposed and implemented a Pathfinder, a routing algorithm to improve link utilization without originating collisions in the network. We have checked Pathfinder in both static and dynamic environments. In experimental simulations, Pathfinder shows a tremendous gain when compared to our standard algorithms such as Dijkstra and Round Robin. The preliminary simulations show that 'pathfinder' increasing the utilization of link by 10% to 15% in both static and dynamic environments.

References

[1] Bin Dai, Guan Xu, Bengxiong Huang, Peng Qin, Yang Xu," Enabling network innovation in data center networks

with software-defined networking: A survey", *Journal of Network and Computer Applications*, vol. 94, pp. 33-49, 2017.

[2] A. M. Al-Sadi, A. Al-Sherbaz, J. Xue, and S. Turner, "Routing algorithm optimization for software defined network WAN," *AIC-MITCSA*, pp. 1-6, 2016.

[3] M. Qilin and S. Weikang, "A Load Balancing Method Based on SDN," *Seventh International Conference on Measuring Technology and Mechatronics Automation*, Nanchang, pp. 18-21, 2015.

[4] Sai Qian Zhang, Qi Zhang, Ali Tizghadam, Byungchul Park, HadiBannazadeh, RaoufBoutaba, Alberto Leon-Garcia," TCAM space-efficient routing in a software defined network", *Computer Networks,* vol. 125, pp. 26-40, 2017.

[5] W. H. Liao, S. C. Kuai and C. H. Lu, "Dynamic Load-Balancing Mechanism for Software-Defined Networking," *International Conference on Networking and Network Applications (NaNA)*, pp. 336-341, 2016.

[6] M. Belyaev and S. Gaivoronski, "Towards load balancing in SDN-networks during DDoS-attacks," *International Science and Technology Conference (Modern Networking Technologies)*, pp. 1-6, 2014.

[7] ManishaChahal, Sandeep Harit, Krishn K. Mishra, Arun Kumar Sangaiah, Zhigao Zheng, "A Survey on software-defined networking in vehicular ad hoc networks: Challenges, applications and use cases", *Sustainable Cities and Society*, vol. 35, pp. 830-840, 2017.

[8] D. Kreutz, F. M. V. Ramos, P. E. Veríssimo, C. E. Rothenberg, S. Azodolmolky and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," in *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14-76, Jan. 2015.

[9] A. Chanda, C. Westphal and D. Raychaudhuri, "Content based traffic engineering in software defined information centric networks," *2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 357-362, 2013.

Data Cleaning Techniques in Detecting Tendencies in Software Engineering

P. Georgieva*, E. Nikolova** and D. Orozova***
Informatics Department in FCSE, Burgas Free University, Burgas, Bulgaria
* penka.georgieva@bfu.bg, ** enikolva@bfu.bg, *** orozova@bfu.bg

***Abstract* - The world of software engineering is dynamically changing over the last decade. Providing adequate university education is one of the key goals of the academic community for ensuring advanced and up-to-date students' training. One direction in achieving this goal is to constantly monitor the trends in the Information Technology (IT) sector. A reliable source of information is the data from the annual survey on technology and programming languages, as well as on preferred learning methods and ways to enhance competencies, conducted amongst Stack Overflow users since 2011. In processing the data from the survey, the authors have faced several problems that have provoked interest in a more general data problem – data quality and data cleaning.**

This paper looks into data quality, tools for data cleaning and the characteristics of high-quality data. A classification of data problems is proposed in the context of analyzing the information about software developers. Additionally, the proposed process of data cleaning in illustrated with data for 2018 and 2019.

Keywords - Big Data Analytics, Data quality, Data cleaning, Software engineering

I. INTRODUCTION

Constantly monitoring the changes in the development of Information Technology (IT) sector is a significant factor in the process of modernizing and providing up-to-date training to university students in Software engineering. While analyzing the data provided from Stack Overflow [1], the team of Informatics department at Burgas Free University has encountered several problems that can be classified as *data quality problems*. This motivated us to delve deeper into studying and solving problems in this area.

The data quality problems emerge in any intelligent system, regardless of its purpose or tools used. For example, for applying machine learning technics well-structured data patterns are needed. Or in modern data warehouses with the increase of data-driven decision-making processes a requirement is the storage of a huge amount of raw data. And as stated in [2], data warehouses are poised to play a leading role in next-generation initiatives, from AI to machine learning to the Internet of Things. While data warehouses do not appear frequently in the marketing literature or analytical reports on these emerging technologies, data storage will remain a critical milestone at the heart of the coming digital age.

In 2013 Bosu and MacDonell claim that the potential of provenance to assure data quality has not been exploited in the empirical software engineering community. Adopted sensibly and systematically, provenance should increase the reputation and trustworthiness of the data that is used in modelling, which will consequently result in higher quality models. Although data quality has been considered from several perspectives in empirical software engineering, as yet there is not sufficient evidence on how it influences the practice of software engineering as almost all the studies are drawn from academic institutions – a more extensive collaboration with industry is essential to understand and then improve data quality management in practice. [3]

Web-based data management needs tools to ensure secure, trustworthy performance. The Utopian future shows a semantic Web providing dependable framework that can solve many of today's data problems. However, the realistic immediate future raises several challenges, including foundational semantic Web issues, the abstract definition of data, and incomplete, evolving ontologies. In either case, the marriage of data and ontologies is indissoluble and represents the knowledge database (KDB), a basic ingredient of the semantic Web. In this article, we look closely at problems in data analysis, the first phase of data cleaning. Applying automated reasoning systems to semantic Web data cleaning and to cleaning-agent design raises many challenges. We can build trust in semantic Web logic only if it's based on certified reasoning. [4]

Despite enterprises collecting and preparing increasingly larger amounts of data for analysis, often the effectively useful data is considerably smaller than the full dataset. Due to the influx of data, data preparation becomes increasingly expensive. The execution time reported at each step is based on recent studies on loading, cleaning, and tuning. Specifically, assuming an optimistic scenario in which data cleaning corresponds to 50% of the analysis time, the rest 50% is mostly spent on loading and tuning. The loading percentage may become even higher

This research is funded by National Research Fund of Bulgaria under Contract No. KP-06-COST-8/ 06.08.2019 for providing national co-financing for the participation of Bulgarian teams in approved actions under the European program for cooperation in the field of research and technology COST under the project "Characteristics prediction and optimization of a photovoltaic system with artificial intelligence methods".

in the presence of non-relational data formats. Query execution takes 3% of the overall time. Therefore, data preparation incurs a significant overhead to data analysis. [5]

A. Data Quality

An important aspect in data collecting is the degree if its usefulness for analysis, strategic planning, and queries. Some data sets have missing values, duplicates, typos, contradictions and are considered *low-quality data*. *High-quality data* sets have the following characteristics: accuracy, completeness, currency, consistency, interpretability, accessibility.

The refinement of data quality is a process that consists of two main stages: (1) *data quality assessment*: data profiling and data analysis; and (2) *data quality improvement*: data cleaning and data enrichment. The first step includes identifying the data quality problems.

B. Taxonomy of Data Quality Problems

In the scientific publications on Data Quality, several taxonomies are proposed. More about the classification proposed by different authors can be found in [6]. In the same paper, relational, structural and time aspect of the datasets are not studied and classification of data problems is adopted and used. The classification proposed in [6] separates the data problems into two main classes, but also takes into account their intersection between them:

- ***quantitative data problems*** – missing values, domain violations, contradictions, non-significant values, outliers, redundancy, overlapping, inconsistency, different measurement units, incorrect data types, different dimensions of data, dimension violation;
- ***qualitative data problems*** – missing values, redundancy, domain violations, synonyms, homonyms, duplicates (triplicates), spelling errors, redundancy, abbreviations, domain violations, inconsistency, lexical errors, syntax errors, data dimension problem;
- ***all data types' problems*** – missing values, domain violations, redundancy, inconsistency, data dimension problem.

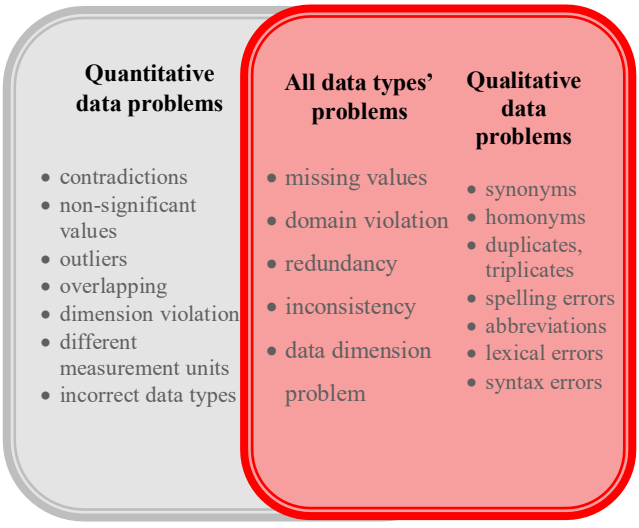


Figure 1. Taxonomy of Data Quality problems

The study presented in this paper is based on the same taxonomy. (Fig. 1)

II. Data Cleaning

All data sources eventually generate errors which can lead to wrong models, estimation bias and poor analysis and thus imply false conclusions. Fixing these intrinsic problems is known as *data cleaning*. Arthur Chapman presented the key principals of data cleaning in [7]: planning; organizing the data efficiently; error prevention; clarified responsibility; prioritization; verification; feedback; accountability, transparency, and audit-ability; documentation. The main data cleaning processes are editing, validation and imputation. Editing is used for identification of errors, validation for their correction and imputation is the replacement of missing values. More information about data cleaning can be found in [8-12].

The approach that the authors adopted from [13] is a three stages process that includes repeated cycles of screening, diagnosing and treating the datasets aiming at removing inaccurate or inconsistent data; correcting records containing incorrect formatting, typographical mistakes and/or harmonizing records so that they are consistent with each other. The screening aims at identifying four types of anomalies: 1) lack/excess of data; 2) outliers, including inconsistencies; 3) strange patterns in (joint) distributions; and 4) unexpected results. Screening methods at this stage are checking, graphing, and primary statistical tools. In the stage of diagnosis, the nature and causes of the errors are identified. Treatment deals with how to proceed the problematic data: leave the data unchanged or correct some values or delete some values or add some values or reformat the data. An additional but mandatory step is the Documentation phase when appropriate documentation for each phase of the process is provided. Documentation includes prior expectations used to screen suspect data, screening tools, diagnostic procedures used to discern errors from true values, and the decision rules that will be applied in the editing phase, diagnostic information, information on the type of editing, dates, and personnel involved.

From the taxonomy proposed in [6], two general data cleaning approaches can be derived: qualitative and quantitative. While the *qualitative approach* includes rules, limitations, and error identification patterns, the *quantitative approach* is based on statistical techniques for identifying and correcting existing errors in the data. The most important methods are:

- **statistical method** – an attribute is considered to be a random variable and if the value is not in the confidence interval, then it is wrong;
- **relationship mode recognition** - data mining and machine learning algorithms are used to find fields that do not conform to the relevant mode;
- **clustering method** - data is grouped into classes or clusters, values that do not fall into any class are considered abnormal;

- **distance method** - a metric is used to calculate the distance and thus to determine the quantitative structure;
- **classification** - a classifier is created that distinguishes "normal" from "abnormal" data;
- **dependency rules** – an attribute is associated with the variable and if some data does not meet the rule, it is considered abnormal.

Different software solutions for data cleaning exist. Excel is preferred for its potential to manipulate rows and columns, and thus perform operations for data cleaning. Relational database management systems are used for downloading and distributing tabular data for their flexibility in filtering and reducing missing data and performing complex data queries. Script languages could also be used in case of processing a large amount of data (R, Python, and Perl). Other data cleaning software systems are 1) commercial data cleaning tools: Informatica PowerCenter, IBM Information Server, Talend Open Studio, Oracle Data Integrator, SQL Server Integration Services, SAS Data Integration Studio, Pentaho Data Integration, Clover ETL and 2) research data cleaning tools: CleenexQCs, LlunaticEgds, Nadeef, Eracer.

Another approach is to choose a specific software to process and analyze the data. For this research, Orange is chosen for being open source, based on Python, its data mining and data visualization tools, and the provided tools for processing data in various formats: CSV, Excel tables, SQL tables or URLs. Orange is developed in the Faculty of Bioinformatics at the University of Ljubljana [14].

III. THE EXPERIMENT

In this research, the answers of two questions are sought:

(1) Which programming languages are associated with higher salaries? Is there a gender imbalance in salary in the IT sector?

(2) What are the programming languages that the respondents claim to use? How the results from 2019 compare with the results from 2018?

For answering these questions, the data for 2018 and 2019 from the Stack Overflow surveys is downloaded as a CSV file from [1]. Each of the columns of the dataset contains the corresponding answer to a given question, and each row represents the relevant respondent in the poll. The 2018 file, contains 98855 rows (number of respondents) and 155 columns, (number of questions asked), with no personal data, while the 2019's file has 88884 rows and 157 columns.

A. Detected and Solved Data Quality problems in the Dataset

Before proceeding and analyzing the data, the schema-related data quality problems and instance-specific problems with following scopes: attribute (field), record, record type and source have been detected. The different problems for the various cases are shown in Table I.

The dimension of the dataset – the overall number of elements in the dataset is approximately 30 million (98 855 x 155 + 88 884 x 157 = 29 277 313) and for the goals of this research, the processing of such a huge amount of data requires a transformation, so the data is reduced.

TABLE I. THE SCHEMA-RELATED DATA QUALITY PROBLEMS AND INSTANCE-SPECIFIC PROBLEMS

Problem	At schema level	At instance level
Attribute	Non-significant values	Missing values Abbreviations Misfielded values
Record	Violated attribute dependencies	Violated attribute dependencies
Record type	Uniqueness violation	Duplicated records Contradicting records
Source	Referential integrity violation	Wrong references

There are two common approaches for solving this problem – the first is to proceed without any changes, and the second – to implement reduction by one of the following (or a combination of them): reduction by removal of some dimensions; data compression; reduction by parameterization; discretization hierarchical presentation in clusters.

Missing data - there are 36.1% missing data in the files. The missing values can be classified as:

- *random missing values* which occur when the respondent has unintentionally failed to answer some questions or as input errors
- *non-random missing values* which occur when the respondent deliberately did not answer a question.

In the first case the problem can be solved by excluding either deleting all rows that miss any value or removing only specific missing values. For the second case a probabilistic approach can be applied - replacing the missing value with approximations such as mean, median or regression, as well as the maximum likelihood approaches.

In this research, a straight forward approach is applied – the rows in which the participant did not indicate their country or which programming languages he/she has worked with are eliminated.

Data dimension problem - some of the columns contain more than one value, e.g. the column *HaveWorkedLanguage*, which contains the answer to the

question "What programming languages did you work with?" The problem needs more attention because every respondent ranked the programming languages that he/she has used in descending order and thus each answer contains a different number of ordered lists. As the research is focused on the most used programming languages, column data is split into several columns in Excel so column data is converted by the *Text to Columns* and *Trim* functions into separate, consecutive columns.

At the stage of data screening the following procedures are performed in this research: validating data entries and duplicated data the columns *Sex*, "*HaveWorkedLanguage*", browsing of data tables Survey_2018 after sorting on columns *Country* and *FormalEducation*; detecting values not passing range checks and records not passing consistency checks - implemented for column *Sex*; distributions visualisation for salary distribution by gender; geographical distribution (by countries) of attribute *WorkedLanguage* and exotic programing languages in *HaveWorkedLanguage* column.

In the Diagnosis phase based on classification technique several problems are detected in the current research: the error types for data in column „Sex"; missing data (the respondent did not answer the question) in column *HaveWorkedLanguage*, Errors-data that are meaninglessly different from those surveyed: The question was not understood and another answer was given for data in column „ConvertedSalary".

In the Treatment phase is decided to correct the data if the respondent's original intent can be determined; to delete the data if it seems illogical and the value is far from normal: values for the sex attribute.

B. *Results and Analysis*

At the stage of screening data in this research the CVS data file is loaded in Excel and the necessary tools for primary data processing provided by Excel are used for lack/excess of data; inconsistencies and unexpected results. For instance, *COUNTIF* function can be used to find the answer of some questions (e.g. What is the geographical distribution of the respondents?). In the treatment phase, Excel's capabilities are used for correcting some the data values; deleting some data values; reformatting the data.

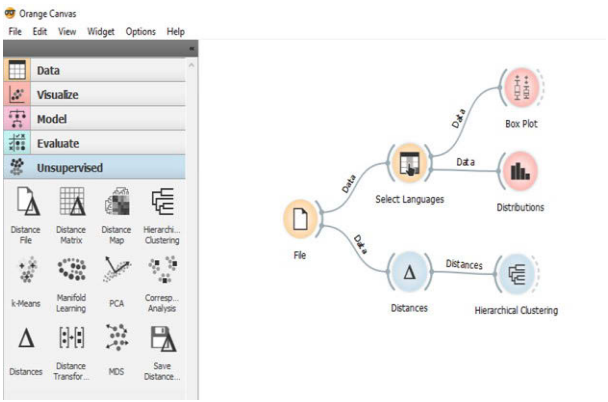


Figure 2. The workflow in the Orange system for data visualization.

The next step is to upload the data in Orange. A new process is created by loading the main data file. As it is shown on Fig. 2, the *Data Table* tool is connected to *File* to verify that the data file is properly loaded. Then, the columns of interest are selected and information about them is retrieved by *Data Info* tool.

When seeking the answer to the question: "Which programming languages are associated with higher salaries?" the task is to find evidence about the effects of the programming language on the developers' pay. Now the respondents' languages and the associated indicated incomes are taken into account and the average income is calculated for each of the programming languages with the assumption that a person's income is influenced by knowing a specific language. This is a serious data processing depending on cleaning the data. Firstly, because Orange accepts as Categorical rather than Numeric every data which differs in the number of symbols. Then all missing values, the values less or equal to 1 and the values exceeding 500 000 BGN must be removed. A very interesting pattern emerged after calculating the average years of experience. The calculations are done for 6 chosen languages - JavaScript, Java, C ++, Python, Scala, PHP. Finally, calculating the median income for the data of all programmers from the survey and the average length of service: 69449.40 and 8.3. (Fig. 3)

The average income for developers with less than 8.3 years of experience is 54627.8. These have an average number of years of experience 3.7. Considering only the programmers with experience longer than the average 8.3 separately, their income is 93549.2 and the average number of years of experience is 15.9.

The data clearly shows that income is growing strongly as traineeship increases.

For answering the second part of the first question: *"Is there a gender imbalance in terms of salary?"* two columns are used: *Gender* and *Salary*. As respondents have entered more than one gender (indicating 2 or more values) some data cleaning is to be done again. After noticing that in fact, less than 1% (28 respondents) have multiple answers it is plausible to simply ignore those rows.

This is done by the *Select Rows* tool – in this research three possible values of gender are chosen: *Male*, *Female,* and *Other*.

Using the *Distribution tool,* the absolute distribution of

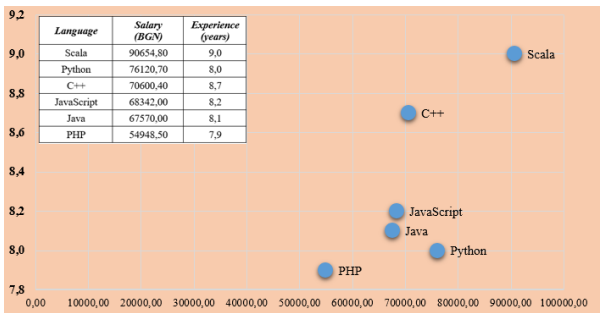


Figure 3. Programming languages, average years of experience and the corresponding average salaries earned.

salaries is obtained as shown on Fig. 4a. If only this tool is used, a wrong conclusion – women and others are underpaid could be drawn. But if the relative distribution is displayed (Fig. 4b), this bias will be removed. The comparison of the relative salary distributions for 2018 and 2019 was realized with the Kolmogorov-Smirnov test.

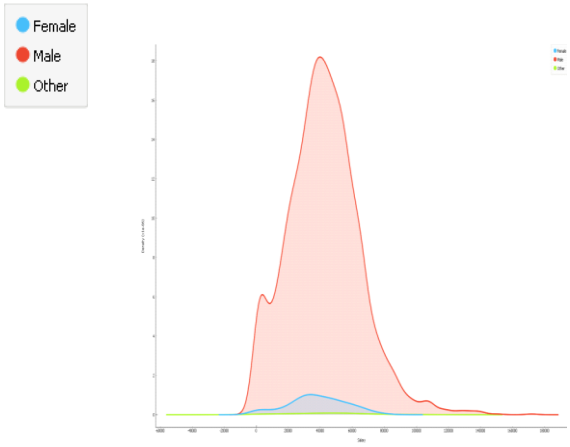


Figure 4a. Absolute distributions of salaries in 2019.

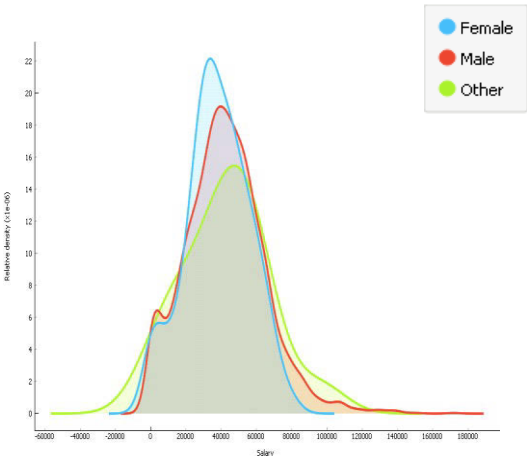


Figure 4b. Relative distributions of salaries in 2019.

Another tool for displaying the salary distribution, according to the educational level achieved by the respondents, is the *Box Plot* tool. In Fig. 5 the average values of the salaries, grouped according to the achieved degrees, are presented. Not surprisingly, respondents with lower education receive smaller salaries. For continuous attributes, the tool shows the mode (the dark blue vertical line; the standard deviation (the light blue rectangle); the median (the yellow vertical line); the continuous horizontal blue line shows the difference between the lower (25%) and the upper (75%) quartile and the dashed blue line represents the entire range of values.

In answering the question "What are the programming languages that the respondents claim to use?" the data is first reduced by selecting only the "popular" programming languages Java, C#, C, C++, SQL, Python, Javascript, PHP. As a consequence, some important tendencies might be left unnoticed. Hierarchical clustering methods are appropriate in this case. The most common way of presenting the results of these methods is the dendrogram, which graphically depicts the hierarchical structure generated by the matrix of similarity and the rules for clustering.

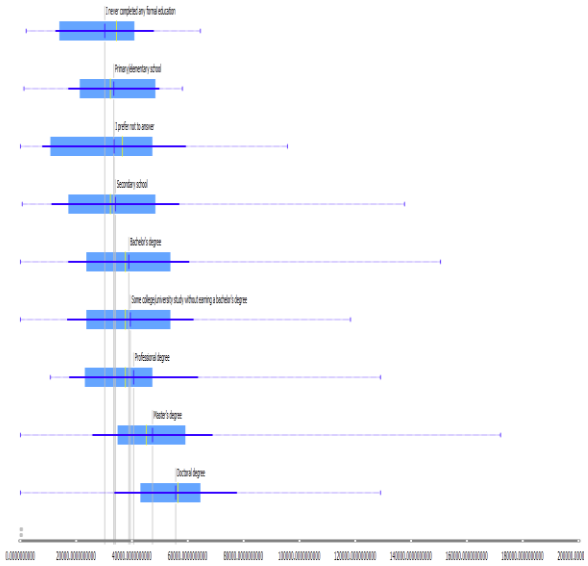


Figure 5. Salary vs. educational level visualized by the Box Plot tool.

For data analysis on the programming language utilization, a concatenation of the two files' data (Data 2018 and Data 2019) is performed and the number of occurrences among each year's survey participants is recorded for each language. The workflow in the Orange system for data concatenation and selection is presented in Fig. 6.

After that, the results for the 24 languages, mostly used in 2018 and 2019 are visualized (Fig. 7). Now the following question emerges: "Do the two histograms represent random variables with identical distributions?" and the hypothesis of homogeneity should be tested.

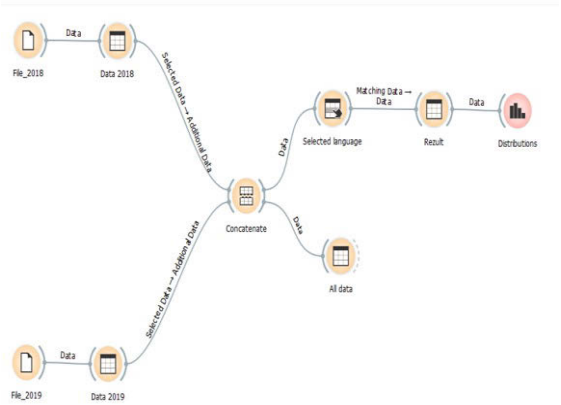


Figure 6. The workflow in the Orange system for data concatenation and selection

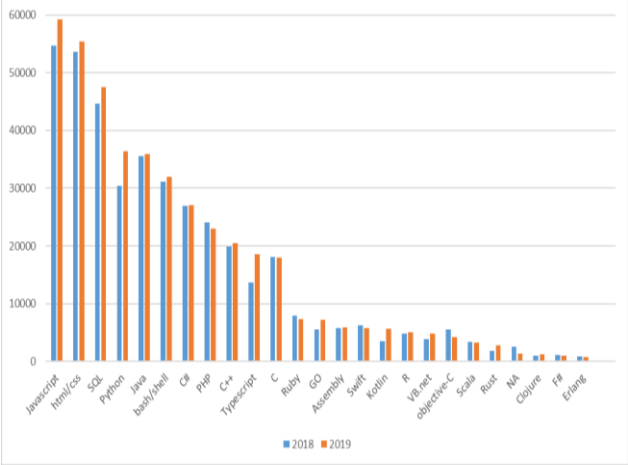


Figure 7. Most frequently used language comparison for 2018 and 2019.

For comparing the two distributions the χ^2 criteria for homogeneity is used. In our case, the p-value is 0,057, so the hypothesis that the two distributions are identical is not rejected, which means that there is no change in the programming languages that respondents claim to use.

The following conclusions can be derived from the histogram:

(1) looking at the trends, **JavaScript** again comes first, with an area of application - building user interface and mobile applications;

(2) the rate of **Python** has significantly increased in comparison with the previous year and its popularity in the scientific community seems to lead to enlarging the area of its use amongst developers;

(3) **Java**, **C#** and **PHP** as traditional programming languages still have considerable popularity, although lower in the rankings;

(4) relatively new languages as **Go**, **Scala**, **Erlang** and **Rust** attract well-paid audiences.

More conclusions about the tendencies in the IT sector for 2019 are the following:

- a significant part of the developers (about 40%), are concentrated in small and medium-sized companies (20-499 employees), and the software technology giants (Google, Facebook, Amazon, Microsoft, IBM, SAP, Oracle, etc.) make up about 14%;
- as in 2018, there is an increase in female developers finding their place in more creative or organizational positions such as front-end developers, QAs, academic researchers, data scientists/machine learning, product managers, etc., but the profession is still male dominated (nearly 90%);
- the average age of developers is increasing, with almost 20% of the survey participants now having over 15 years of experience (around 40 years old);
- according to the 2019 survey, managers, SRE/DevOps, and people involved in processing and analyzing (big) data, including machine learning, are best paid, whereas front-end and mobile developers, as well as academic researchers, seem to get the lowest salaries. [15]

IV. Conclusion

This paper outlines a major problem faced in the process of monitoring the tendencies in the IT sector – the data quality. The principles, methods, and techniques for data cleaning are presented, and authors' experience in cleaning and analyzing data from the Stack Overflow survey for 2018 and 2019 is shared. The presented study is a successive study used in the strategic planning of curricula of IT students in Burgas Free University.

The conducted research led to several changes in the curriculum and the introduction of a new bachelor's program in applied informatics. A side effect is that lecturers teach the students data cleansing techniques and argue for using quality data in all types of research and applications.

References

[1] https://insights.stackoverflow.com/survey

[2] J. McKendrick, "Rethinking the future of data warehousing," DBTA 2019, https://www.actian.com/wp-content/uploads/ 2019/ 04/Rethinking-the-Future-of-Data-Warehousing.pdf

[3] M. Bosu, and S.MacDonell, "A taxonomy of data quality challenges in empirical software engineering," Proceedings of the 22nd Australasian Software Engineering Conference (ASWEC2013). Melbourne, Australia, IEEE Computer Society Press, pp. 97-106, 2013

[4] J. Alonso-Jimenez, J. Borrego–Diaz, A. Chavez-Gonzalez and F. Jesus Martin–Mateos, "Foundational Challenges in Automated Data and Ontology Cleaning," Semantic Web Vol. 21 (1), pp. 42-52, IEEE Intelligent Systems, 2006

[5] M. Olma, S. Giannakopoulou, M. Karpathiotakis and A. Ailamaki, "Toward Intelligent Query Engines," Bulletin of the Technical Committee on Data Engineering, Vol. 42 (2), IEEE Computer Society, 2019

[6] P. Georgieva, E. Nikolova and D. Orozova, "Tendencies in Software Engineering Education," ICERI2019, 12th International Conference of Education, Research and Innovation, Seville (Spain) 11-13 November, 2019

[7] A. Chapman, "Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data.", Report for the Global Biodiversity Information Facility, Copenhagen, 2005

[8] www.geotab.com/blog/data-cleaning/

[9] https://towardsdatascience.com/data-cleaning-101-948d22a92e4

[10] www.ncbi.nlm.nih.gov/pmc/articles/PMC1198040/#pmed-0020267-b10

[11] https://hackernoon.com/data-cleaning-3c3e37f358dc

[12] H. Galhardas, "Introduction to Data Cleaning." online: https://fenix.tecnico.ulisboa.pt/downloadFile/ 563568428715919/ 7_DataCleaningIntro.pdf

[13] J. Van den Broeck, S. Argeseanu Cunningham, R. Eeckels, and K. Herbst, "Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities." PLoS Med 2 (10), 2005

[14] https://orange.biolab.si.

[15] https://insights.stackoverflow.com/survey/2019#overview

Automatic Machine Translation of Poetry and a Low-Resource Language Pair

I. Dunđer*, S. Seljan* and M. Pavlovski*
* Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
ivandunder@gmail.com, sseljan@ffzg.hr, mpavlovs@ffzg.hr

Abstract – **Automatic machine translation is gaining more and more attention in a particular segment of the research community that treats various topics from artificial intelligence, natural language processing, computational linguistics, machine learning and data science. Machine translation is a complex task in which a computer is utilised for the purpose of translating from source to one or more target languages without human involvement, or with a minimum of interventions. In general, machine translation could be implemented in higher education and academic curricula in a variety of possible fields and applications. Although there are several approaches to machine translation, two are dominant today – statistical and neural machine translation. Both are being used in this research in form of two online machine translation systems. The aim of this paper is to examine the usability of machine translation for poetry and a low-resource language pair, such as Croatian-German. The authors chose to use a data set that contained the works of a relevant contemporary poet of the Croatian language and the translations of his poems in German that were conducted by two professional literary translators. The paper demonstrates the effectiveness of machine translation of poetry with regard to special automatic quality metrics.**

Keywords - automatic machine translation; statistical and neural machine translation; quality evaluation; automatic metrics; natural language processing; low-resource; Croatian

I. INTRODUCTION AND MOTIVATION

Automatic machine translation deals with the application of computers for the purpose of (partially) automatising the process of translating from one language into another [1]. The main goal of machine translation systems is to computationally generate a vast number of translations of acceptable quality in short time and with minimal costs.

As machine translation system output is not perfect, evaluation of system performance and translation quality is of utmost importance. Although machine translation quality is a subjective term, it usually represents the similarity (overlap) between a machine translation and one or more reference (i.e. human) translations [1].

Machine translation quality has increased significantly over the last couple of decades. This is mainly due to the advances in the today's most dominant machine translation approaches: statistical (SMT) and neural machine translation (NMT). These two approaches share some common steps, tasks and techniques but at their core are fundamentally very different.

Machine translation typically heavily relies on the availability of large and high-quality data sets, which are called corpora and are used for training, tuning and performance testing of machine translation systems. Therefore, their acquisition and preparation are of great importance for the entire machine translation process.

For widely spoken languages there are lots of corpora available, mostly on the internet. However, the lack of necessary data sets for low-resource languages, such as Croatian and the majority of other less spoken languages, is a serious problem. That is why automatic machine translation cannot be applied to many language pairs (combinations) today.

Moreover, the term "low-resource" refers not only to the lack of crucial data sets, but also to the absence of necessary natural language processing (NLP) tools that are essential for the whole machine translation pipeline. They are especially employed in the data preprocessing and preparation phase, but also in later-stage evaluation. Therefore, many tools for low-resource languages have to be built first, or adopted to research-specific needs. The final machine translation system output is highly interrelated and directly dependent on the effectiveness of natural language processing tools.

This paper analyses the usability of popular and widely used automatic machine translation systems for translating poetry from Croatian into German, and German into Croatian. The evaluation of behaviour of such machine translation systems and their outputs is very important, as it examines the appropriateness of such systems for the purpose of translating domain-specific content for a given language pair.

The aim of this paper is to analyse the usability of machine translation technology for translating poetry and a low-resource language pair, in this case Croatian-German, and a data set which contained poems written by a relevant contemporary Croatian poet, Delimir Rešicki, and the translations of his poems in German that were conducted by Klaus Detlef Olof and Alida Bremer, two professional literary translators.

The poems and their translations were taken from a specialised multilingual platform called "lyrikline.org",

which in November of 2019 hosted 12578 poems written by 1403 poets with 18869 translations in 85 languages [2].

Delimir Rešicki (born in Osijek in 1960) is a Croatian poet, writer, journalist and editor with more than 15 published books of poetry, prose and essays. He won many important prizes for his literary work, including notably the Hubert Burda prize in 2008, the Goranov vijenac in 2011, and was also awarded twice the annual Vladimir Nazor prize in 2005 and 2015.

The motivation for this paper is to explore if machine translation in general is suitable for literary translation, in this particular case on the example of poetry of a relevant author, in comparison to the corresponding translations conducted by two professional literary translators.

Machine translation could be adopted in higher education and academic curricula in a variety of possible fields and applications. For instance, students could use this technology to research in their own language the body of work of a contemporary author which is not yet translated and published, allowing a more thorough understanding of the studied author. Also, students from technical disciplines could be taught how to work with automatic machine translation systems, how to prepare, handle and adjust necessary resources, how to carry out various types of quality evaluation, and how to cope with numerous challenges that come along with, especially in specific domains, such as poetry, and low-resource environments.

II. Related Work

Extensive research of statistical machine translation with focus on the Croatian language, and especially with regard to domain adaption techniques has been done by [1].

As Croatian can be considered a low-resource language with a clear lack of large high-quality data sets, one research proposed providing more resources with help of digitisation and subsequent linguistic description [3].

Web crawling for monolingual, parallel and comparable corpora is sometimes useful, but without manual inspection quality cannot be guaranteed. Therefore, human experts should look into the data, but this process is very time-consuming and expensive. This is why another research tried a gamification approach in order to increase user engagement in a specially-built crowdsourcing platform for collecting and verifying parallel corpora for Croatian [4].

In order to increase the amount of needed machine translation resources, one paper tried to combine automatic speech recognition (ASR) and machine translation in the business correspondence domain for English [5].

Evaluation of machine translation has been done in many different researches. One paper focused on applying automatic quality metrics on machine translations in the sociological, philosophical and spiritual domain [6]. Automatic and human evaluation of online machine translation systems have also been done for the English, Russian and Croatian language [7, 8].

One paper provides a systematic overview of machine translation related research with regard to literary texts [9]. Here "the dream of creating an immaculate machine translation […] system that would master literary translation is very old". The author then proceeds to cite the introduction to the anthology Machine Translation of Languages (editors Locke and Booth, 1955), which states that "[…] the question of turning a masterpiece of literature written in a foreign language into a respectable translation is one of great difficulty".

When it comes to translating poetry, a phrase-based machine translation system was built with special emphasis to rhyme and meter [10]. Another system tested online machine translation systems for Thai poetry and with regard to prosody [11]. A recent neural poetry translation system [12] has shown to be successful in generating French to English machine translations of poems taking into consideration fixed rhyme and rhythm patterns.

III. Research

This section consists of three subsections – Data Set Crawling, Preparation and Preprocessing; Automatic Machine Translation; and Quality Evaluation Approach. The first subsection deals with the process of acquiring the experimental data set, and the subsequent preparation and preprocessing phase. The second subsection discusses the applied research approach and selected machine translation tools. The third explains the automatic machine translation quality evaluation approach.

In this paper the authors decided to use automatic machine translation as a mathematical approach for quickly generating machine translations for the Croatian-German (HR>DE) and German-Croatian (DE>HR) language pairs, i.e. for both directions. Although machine translation enables cost-effective translation, it is often achieved at the evident expense of quality, which is particularly apparent when morphologically rich languages, such as Croatian, are being translated automatically.

This paper concentrates on exploring the possibilities of applying automatic machine translation systems in form of online translation services for the purpose of translating a specific data set – a collection of poems written by Delimir Rešicki, from Croatian into German, and vice versa. His poems and their translations were taken from a multilingual web platform called "Lyrikline". The paper analyses the quantitative aspects of the resulting machine translations, whereas an evaluation on a qualitative level should be done separately.

The foundations of this research originate in the studies of information and communication sciences, computer science, linguistics, and especially artificial intelligence, machine learning, probability and statistics, data science and natural language processing.

Machine translation and the related techniques are applicable in higher education and for purposes of academic curricula – first and foremost in selected courses in information and communication sciences, computer science, computational linguistics, philology and

translation studies. These courses deal among others with various aspects of natural language processing, with analysing languages computationally, with practices in teaching translation etc.

A. Data Set Crawling, Preparation and Preprocessing

The experimental data set contained several poems written in Croatian by Delimir Rešicki, which were taken from Lyrikline (https://www.lyrikline.org/en/poems/rosa-mystica-13248) – an international web platform for publishing diverse contemporary poetry. Fifteen Rešicki's poems were published online, but only 14 had human translations from Croatian into German. Therefore, only those were taken into consideration. No other data was added to the data set.

In total, the corpus consisted of 532 sentences/segments in Croatian, and 532 in German. The term "segments" refers to text chunks, i.e. text lines that do not end with a sentence delimiter.

In the Croatian part of the data set, the maximal sentence/segment size was 17 words, whereas in the German part that number was slightly lower (14 words). Minimal sentence/segment size in both languages was 1 word, with a total of 2437 words in Croatian, out of which 1414 were distinct words, and 2862 words in German (1314 distinct words). Average sentence/segment size was only 4.58 words in the Croatian and 5.38 words in German – this was expected, as the whole data set contained poems written in short verses.

Human translations by Klaus Detlef Olof and Alida Bremer were all used as reference translations afterwards in the machine translation experiments and in the evaluation phase. One or possibly more human reference translations are very important in machine translation quality evaluation, as they are considered the "gold standard" of translation quality.

All suitable poems had to be crawled from the internet, preprocessed and prepared for machine translation. The preprocessing and preparation phase encompassed various tasks: conversion of file format from HTML to plain TXT with proper character encoding (UTF-8) for correct representation of Croatian and German characters, getting rid of style and text formatting etc. Redundant and undesired characters were deleted from the data set using regular expressions. There was no need to split the data set into sentences or segments, as the poems were translated verse by verse. Then, tokenisation was performed using custom tokenisation rules.

The data set was afterwards converted into a parallel corpus. Its content was manually aligned at the sentence/segment level with 1-1 relationship, meaning that one verse from the source text corresponded to only one target verse from the human (reference) translation.

Most of the preprocessing and data set analysing was made with the use of Python and Perl.

Human inspection and correction of data before machine translation positively affects the final machine translation system quality. It is crucial that the alignments between source and reference text are correct. Therefore, the data set was manually inspected for any potential errors by two evaluators – a native speaker of the Croatian, and another of the German language. The evaluators had to check whether misalignments, orthographic mistakes, various oversights, missing or faulty characters, such as apostrophes, quotation marks, dashes, hyphens, brackets and other punctuation marks and typographic symbols were present in the data set.

Poems are oftentimes written in a limited format with short verses, which was also the case in this research, and, hence, one would expect that the corresponding translations also follow this very strict verse order. However, during the manual inspection, the evaluators found that some of the segments (verses) were not correctly aligned, obviously due to the artistic license of professional human translations, which was expected to some degree. Some of the translations were relatively freely translated, while some other, although correct, just appeared in later verses and therefore caused misalignments.

Nevertheless, for research purposes they were not fixed, i.e. they were left intact in order to respect the creativity of human translations (translations are correct but with different verse order), and to assess their impact on machine translation quality. Only some smaller segments, i.e. verses were combined where considered appropriate by the human evaluators in this experiment.

Once the parallel corpus was ready, it was used as the input for generating machine translations.

B. Automatic Machine Translation

For generating Croatian-German and German-Croatian automatic machine translations in the domain of poetry, the authors decided to employ two very popular and freely available online machine translation services that use different machine translation approaches and have different system architectures, but are trained on general data (with no dominant content) – Google Translate (https://translate.google.com) and Yandex.Translate (https://translate.yandex.com/).

Google Translate primarily uses neural machine translation [13], whereas Yandex.Translate uses a hybrid machine translation approach which combines phrase-based statistical machine translation and neural machine translation [14].

Phrase-based statistical machine translation (PBSMT) relies on non-linguistic phrases [15] and numerous model features that are trained separately and then conjointly implemented in a machine translation system model [16].

Neural machine translation (NMT) refers to a machine translation approach that is based on deep learning [17]. Here, an artificial neural network (ANN) is used for predicting sequences of words, without the need for individual machine learning of model features, such as in SMT [18].

The preprocessed data set with raw (stripped of text formatting) and cleaned text was fed to the machine translation services separately, and once the automatic translations were obtained for the whole data set and both

directions (HR>DE and DE>HR) they were tokenised as well. Then, a machine translation quality evaluation followed, in which machine translations were compared to the original human reference translations and scored with help of selected automatic quality metrics. In all machine translation quality trials, the original source texts and the initial human professional translations were considered reference texts.

C. Quality Evaluation Approach

The machine translation quality evaluation was conducted by using different automatic metrics. In total, four metrics were chosen: BLEU [19], METEOR [20], RIBES [21] and CharacTER [22]. All were calculated on the corpus level (overall, i.e. final score).

Over the time, BLEU (BiLingual Evaluation Understudy) has become the *de facto* standard for evaluation of machine translation system quality. BLEU matches machine-translated n-grams with n-grams of its reference translation, and counts the number of matches, most often for 1-4 n-grams, which are all assigned the same weights. The metric calculates the geometric mean of the n-gram precision in a machine translation based on a reference translation. Also, a brevity penalty is assigned to machine translations that are shorter than the reference translation. BLEU decreases with every rise of n-gram order, which means that the BLEU score usually drops as matching of machine-translated and reference n-grams over larger ranges becomes less frequent. BLEU score ranges from 0 (no matching of words in the machine and reference translation) to 1 (complete overlap of machine and reference translation). Scores above 0.3 represent understandable machine translation, whereas BLEU over 0.5 represents good and fluent machine translations [1].

METEOR gives more emphasis to recall than to precision, and it incorporates linguistic knowledge by taking into account word stems and synonyms (if available). At its core, METEOR computes unigram overlaps between machine translations and reference sentences. In this experiment only default parameters (for English) were used. METEOR results range from 0 to 1, and are generally higher than BLEU scores. METEOR scores above 0.5 are considered understandable, and above 0.7 as good and fluent translations [1].

RIBES (Rank-based Intuitive Bilingual Evaluation Score) focuses mainly on word order, i.e. it is more sensitive to reordering, and has shown to have better correlations with human judgements. Unlike BLEU, it considers explicit reordering of words. BLEU does that only by rewarding the higher n-gram orders, but freely permuted unigrams and bigrams matches are able to sustain a high BLEU score with little penalty caused by tri- or four-gram mismatches. To overcome reordering, RIBES adds a rank correlation coefficient prior to unigram matches without the need for higher order n-gram matches [23].

CharacTER (Character level Translation Edit Rate) is a metric based on Levenshtein's distance, and it is specified as the minimum number of character edits required to adjust a machine translation, so that it absolutely matches the reference, normalized by the length of the machine-translated sentence. The metric calculates the character level edit distance while performing the shift edit on word level. CharacTER has shown high system-level correlation with human rankings, especially for morphologically rich languages.

IV. Results and Discussion

This section deals with the experiment's results and discusses the various experimental aspects. In addition, the authors point out the different drawbacks of the conducted experiment.

Table I shows the results of applying automatic machine translation quality metrics for both translation directions and machine translation systems.

TABLE I. Automatic Quality Evaluation Results

Metric	Direction	Google Translate	Yandex.Translate
BLEU*	HR>DE	0.1765	0.1695
	DE>HR	0.1615	0.1555
METEOR*	HR>DE	0.1944	0.1920
	DE>HR	0.1938	0.1934
RIBES*	HR>DE	0.5568	0.5531
	DE>HR	0.5093	0.5152
CharacTER**	HR>DE	0.5663	0.5783
	DE>HR	0.5264	0.5309
Remarks: * higher is better, ** lower is better			

Here, best BLEU score was achieved for the Croatian>German direction and by the Google Translate system. BLEU of 0.17 indicates, overall, lower-quality machine translations. Nevertheless, during manual inspection of machine translations the authors found in lot of cases that machine-translated verses were relatively well, and sometimes almost flawless. All other BLEU scores were lower than 0.17, and here Yandex.Translate scored lower for both translation directions.

METEOR confirmed that the generated machine translations were of low quality. Still, the authors would like to point out that in this research METEOR settings for the English language were used (with no linguistic knowledge of Croatian and German) – this surely contributed to the lower scores. Best METEOR score was again achieved by Google Translate and the Croatian-German translation direction. Yandex.Translate scored again (slightly) lower for both translation directions.

Although RIBES uses rank correlation coefficients based on the word order of machine translation system output and reference translations, in this experiment it correlates positively with the BLEU scores for both machine translation systems. Here, the highest RIBES score (0.5568) was calculated for, again, the Google Translate system and the Croatian-German translation direction.

Machine translation evaluation metrics BLEU, METEOR and RIBES scored almost always better for the Croatian-German translation direction (except METEOR for Yandex.Translate), but this was not confirmed by the final metric in this research – CharacTER. Its scores were

very interesting. For both machine translation systems there were less character edits needed to adjust German-Croatian machine translations. The better system was Google Translate with a score of 0.5264. Deeper analysis and manual inspection of machine translations should yield definitive conclusions on why this metric preferred the German-Croatian translation direction.

Although machine translations generated by Google Translate appear to be of better quality than translations generated by Yandex.Translate – due to the fact that Google Translate scored better almost all times and according to all metrics (except RIBES for German-Croatian) – the authors have to point out that the metric calculations were not (necessarily) statistically significant.

Overall, from the presented metrics' results the authors can infer that the machine translations generated by both machine translation systems are of lower quality. However, one important reason why all metrics showed relatively low scores is certainly the evidenced discrepancy between the original text and the human translations in the resulting parallel corpus. Namely, the human translators translated the verses from Rešicki's poems very freely from Croatian into German, which resulted in the violation of the required 1-1 alignments in the parallel corpus. Those misalignments for sure affected the quality scores negatively, i.e. decreased the metrics' results.

The authors would also like to emphasise the main two drawbacks of this research:

- in this experiment the authors compared two machine translation systems that use different machine translation approaches, pipelines and methods;
- it is not known what data sets were used to train the chosen machine translation systems, but for sure they are not fine-tuned for translating Croatian poetry.

V. Future Research and Directions

Employing sentence/segment level quality metrics is planned. The authors would then use bootstrap resampling (with replacement) [1], due to the initial observation that the data set contained some 1-1 misalignments. This technique can mitigate the effect of statistical outliers caused by wrong alignments, reveal statistical distributions, calculate confidence intervals and provide statistically significant quality metrics' results, which are not guaranteed in this paper. The authors plan to apply word vector representations [24] and sentiment analysis [25] on machine-translated poems.

The authors intend to use different types of metrics and to conduct a metric correlation analysis.

Building own statistical and neural machine translation systems for the Croatian-German language pair is also planned. Such systems would also be trained with general domain data sets, but tuned or tested with the same poetry data set.

In order to facilitate the combination of speech synthesis and machine translation of poetry, the authors aim to work on integrating machine translation into a Croatian speech synthesis system [26] with additional word-level [27] and domain-specific evaluation [28].

Suggestions and directions for future research include also the building of rule-based or hybrid machine translation systems that are less dependent on large data sets.

VI. Conclusion

Machine translation is increasingly becoming popular in different scientific disciplines that address various NLP-related topics.

In this paper the authors decided to examine the appropriateness of machine translation in a specific and complex domain. Here the generated machine translations were evaluated through the use of four automatic machine translation quality metrics: BLEU, METEOR, RIBES and CharacTER.

The authors found that the tested machine translation systems were not really suitable for translating texts from the domain of poetry and the Croatian-German language pair, at least according to all applied evaluation metrics. The scores imply overall lower quality machine translations, but Croatian to German (HR>DE) translations were rated better in most cases.

However, this research has a few main limitations. The results should be treated as preliminary because of the relatively small testing sample. Apart from this, both machine translation systems were not domain-specific, i.e. not trained nor fine-tuned for translating literary texts (poetry-related content).

Also, comparing the utilised machine translation systems was not straightforward, as the systems use different machine translation approaches and different system architectures. Still, similar and comparable results were achieved by both systems according to the automatic metrics in this research. Moreover, comparing different architectures can give deeper insights into the overall quality perception of machine translations. But in order to provide more accurate comparisons, in future research one should compare similar architectures.

Furthermore, one metric (METEOR) was not adapted to Croatian and German. Its default settings for the English language penalised therefore even more the machine translations in question, which also contributed to the lower scores.

In addition, machine translation evaluation metrics usually favour translations that follow a strict word order when compared to the reference translations. But in this experiment the data set was relatively freely translated, which resulted in various misalignments that, for sure, skewed the results of the metrics towards lower scores. In order to provide more rigorous assessment of the level of quality of machine translations, these alignment imperfections should definitely be taken into consideration in future research.

REFERENCES

[1] I. Dunder, Statistical machine translation and computational domain adaptation (Sustav za statističko strojno prevođenje i računalna adaptacija domene) // doctoral dissertation. Zagreb: University of Zagreb, 2015, p. 281.

[2] Lyrikline.org, "About Lyrikline.org," Availabe: https://www.lyrikline.org/en/about/ (10.11.2019)

[3] S. Seljan, I. Dunđer, and A. Gašpar, "From Digitisation Process to Terminological Digital Resources," Proc. 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2013), 2013, pp. 1329–1334.

[4] R. Jaworski, S. Seljan, and I. Dunđer, "Towards educating and motivating the crowd – a crowdsourcing platform for harvesting the fruits of NLP students' labour," Proc. 8th Language & Technology Conference – Human Language Technologies as a Challenge for Computer Science and Linguistics, 2017, pp. 332–336.

[5] S. Seljan, and I. Dunđer, "Combined Automatic Speech Recognition and Machine Translation in Business Correspondence Domain for English-Croatian," Proc. International Conference on Embedded Systems and Intelligent Technology (ICESIT 2014) – International Journal of Computer, Information, Systems and Control Engineering, vol. 8, 2014, pp. 1069–1075.

[6] S. Seljan, and I. Dunđer, "Automatic Quality Evaluation of Machine-Translated Output in Sociological-Philosophical-Spiritual Domain," Proc. 10th Iberian Conference on Information Systems and Technologies (CISTI'2015), vol. 2, 2015, pp. 128–131.

[7] S. Seljan, M. Tucaković, and I. Dunđer, "Human Evaluation of Online Machine Translation Services for English/Russian-Croatian," Proc. WorldCIST'15 – 3rd World Conference on Information Systems and Technologies (Advances in Intelligent Systems and Computing – New Contributions in Information Systems and Technologies), 2015, pp. 1089–1098.

[8] S. Seljan, and I. Dunđer, "Machine Translation and Automatic Evaluation of English/Russian-Croatian," Proc. International Conference "Corpus Linguistics – 2015" (CORPORA 2015), 2015, pp. 72–79.

[9] K. Taivalkoski-Shilov, "Ethical issues regarding machine(-assisted) translation of literary texts," Perspectives, vol. 27(5), 2019, pp. 689–703.

[10] D. Genzel, J. Uszkoreit, and F. Och, ""Poetic" Statistical Machine Translation: Rhyme and Meter," Proc. 2010 Conference on Empirical Methods in Natural Language Processing (ACL), 2010, pp. 158–166.

[11] S. Waijanya, and A. Mingkhwan, "Thai Poetry in Machine Translation: An Analysis of Poetry Translation using Statistical Machine Translation," International Journal of Applied Science and Technology (AIJSTPME), vol. 5(3), 2012, pp. 49–56.

[12] M. Ghazvininejad, Y. Choi, and K. Knight, "Neural Poetry Translation," Proc. of NAACL-HLT 2018 (ACL), 2018, pp. 67–71.

[13] Y. Wu et al., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," arXiv:1609.08144 [cs.CL], 2016, p. 23.

[14] Tech.yandex.com, "About machine translation," Available: https://tech.yandex.com/translate/doc/dg/concepts/how-works-machine-translation-docpage/ (20.11.2019)

[15] P. Koehn, F. J. Och, and. D. Marcu, "Statistical phrase-based translation," Proc. 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL '03), vol. 1, 2003, pp. 48–54.

[16] P. Koehn, Statistical Machine Translation. New York, NY: Cambridge University Press, 2010, p. 488.

[17] U. Kamath, J. Liu, and J. Whitaker, Deep Learning for NLP and Speech Recognition. Berlin: Springer, 2019, p. 621.

[18] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder{Decoder Approaches," Proc. SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (ACL), 2014, pp. 103–111.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), 2002, pp. 311–318.

[20] A. Agarwal, and A. Lavie, "METEOR, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output," Proc. Third Workshop on Statistical Machine Translation (StatMT '08), ACL 2008, 2008, pp. 115–118.

[21] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada, "Automatic Evaluation of Translation Quality for Distant Language Pairs," Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing (ACL), 2010, pp. 944–952.

[22] W. Wang, J.-T. Peter, H. Rosendahl, and H. Ney, "CharacTER: Translation Edit Rate on Character Level," Proc. of the First Conference on Machine Translation: vol. 2, Shared Task Papers (ACL), 2016, pp. 505–510.

[23] L. Tan, J. Dehdari, and. J. van Genabith, "An Awkward Disparity between BLEU / RIBES Scores and Human Judgements in Machine Translation," Proc. of the 2nd Workshop on Asian Translation (WAT2015), 2015, pp. 74–81.

[24] I. Dunđer, and M. Pavlovski, "Through the Limits of Newspeak: an Analysis of the Vector Representation of Words in George Orwell's 1984," Proc. 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2019), 2019, pp. 0691–0696.

[25] I. Dunđer, and M. Pavlovski, "Behind the Dystopian Sentiment: a Sentiment Analysis of George Orwell's 1984," Proc. 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2019), 2019, pp. 0685–0690.

[26] I. Dunđer, "CroSS: Croatian Speech Synthesizer - design and implementation," Proc. 16th International Multiconference INFORMATION SOCIETY - IS 2013 / Collaboration, Software and Services in Information Society (CSS'2013), vol. A, 2013, pp. 257–260.

[27] S. Seljan, and I. Dunđer, "Automatic Word-Level Evaluation and Error Analysis of Formant Speech Synthesis for Croatian," Proc. 4th European Conference of Computer Science (ECCS '13) – Recent Advances in Information Science (Recent Advances in Computer Engineering Series 17) / Image, Speech and Signal Processing, 2013, pp. 172–178.

[28] I. Dunđer, S. Seljan, and M. Arambašić, "Domain-Specific Evaluation of Croatian Speech Synthesis in CALL," Proc. 7th European Computing Conference (ECC '13) – Recent Advances in Information Science (Recent Advances in Computer Engineering Series 13) / Language and Text Processing, 2013, pp. 142–147.

Human Quality Evaluation of Machine-Translated Poetry

S. Seljan*, I. Dunđer* and M. Pavlovski*
* Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
sseljan@ffzg.hr, ivandunder@gmail.com, mpavlovs@ffzg.hr

Abstract - **The quality of literary translation was from the beginning of literacy an important factor in publishing and, as a consequence, in research and education. The quality of literary text translation is of utmost significance for researchers and students, especially in higher education. Only complete and high-standard translations are believed to be necessary for the use in the evaluation and study of style and concepts of a given author or a literary genre. This quality verification applies even more to machine translation in general, due to the fact that such translations are deemed subpar and unsuitable for further dissemination and examination. The need for human quality evaluation of machine-translated text is therefore highly emphasised, since human translations are considered to be the "gold standard" and reference translations in the machine translation process. The aim of this paper is to explore, on the example of a data set consisting of poems written by a relevant contemporary Croatian poet, the effectiveness of applying machine translation on the Croatian-German language pair in the domain of poetry, with regard to human judgment of machine translation quality. Human evaluation in this paper is conducted by taking into account two machine translation quality criteria – adequacy and fluency, after which an inter-rater agreement analysis is performed.**

Keywords - automatic machine translation; machine translation quality evaluation; human evaluation; domain-specific evaluation; natural language processing

I. INTRODUCTION AND MOTIVATION

The quality of a literary text translation is of great importance for researchers in the field of higher education, especially those involved in studies of language and literature for the precise evaluation of an author's writing in terms of style, concepts and elements of a literary work, such as character, theme, plot, point of view, setting, conflict and tone.

In recent times the demand for generating fast and low-cost translations by applying various technological approaches – but often at the expense of quality – is becoming more and more apparent. The results of employing dominant machine translation approaches, such as statistical and neural machine translation, need to be verified and most often post-edited, i.e. manually corrected according to the user-specific requirements.

The need to demonstrate the efficacy of technology in translating literary texts rises with the ever-growing possibilities of machine translation, which could be an important tool for a tradition that, from the beginning of literacy, considered the quality of literary translation an important factor in various fields, including publishing and higher education.

The aim of this paper is to evaluate the quality of machine translations in comparison with human translations made by two professional and acclaimed Croatian-German translators, with numerous translated literary works: Klaus Detlef Olof (Oebisfelde, 1939), PhD, a professor of Slavic languages and winner of the 1991 Austrian national prize for literary translators; and Alida Bremer (Split, 1959), PhD, an accentuated promoter of Croatian and other literatures of Southeastern Europe.

This paper explores how machine translations rank when compared with translations done by professional human translators, on the example of works of Delimir Rešicki (Osijek, 1960), one of the most renowned and awarded contemporary Croatian poets. As of 2019, Delimir Rešicki published 17 books of poetry, fiction, essays and criticism, his works were translated in numerous languages including English, German, French, Italian etc., and he was awarded some notable literary prizes, including Kiklop (2005), Vladimir Nazor (2005 and 2015) and the Hubert-Burda-Preis (2008).

Although machine translations can also be assessed automatically by using quality metrics, in this paper the authors tried to examine the quality of German-Croatian machine translations in the domain of poetry with the help of human evaluators who are native speakers of the Croatian language. Here, the human quality evaluation is done with regard to two machine translation quality evaluation criteria – adequacy and fluency.

Adequacy analyses "how much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation" [1]. Fluency, however, inspects to what extent the machine translation is "one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker" [1].

The basis of this research derives from various scientific fields, such as information and communication sciences, computer science, linguistics, and especially natural language processing (NLP), and it is not surprising that the study of machine translation is already

implemented in various courses in higher education that treat and evaluate different aspects of machine translation and its resulting quality in particular.

The methods that are used for machine translation quality assessment are important not only for students in the humanities, but also in the education of students in the technical sciences that are taught how machine translation is accomplished through different phases, what limitations might occur, what quality can be expected in different scenarios, what evaluation aspects can be considered, and so on.

II. Related Work

When discussing human evaluation of machine translation, [2] states that "evaluation of segment-level machine translation metrics is currently hampered by: low inter-annotator agreement levels in human assessments; lack of an effective mechanism for evaluation of translations of equal quality; and lack of methods of significance testing improvements over a baseline".

Reference [3] concludes that "there is evidence that machine translation can streamline the translation process for specific types of texts, such as questions; however, it does not yet rival the quality of human translations, to which post-editing is key in this process".

Comparing statistical machine translation with neural machine evaluation, [4] denotes that „automatic evaluation results presented for neural machine translation are very promising, however human evaluations show mixed results". The authors report "increases in fluency but inconsistent results for adequacy and post-editing effort. Neural machine translation undoubtedly represents a step forward for the machine translation field, but one that the community should be careful not to oversell".

While identifying fluently inadequate output in neural and statistical machine translation, [5] states that „with the impressive fluency of modern machine translation output, systems may produce output that is fluent but not adequate (fluently inadequate)".

Reference [6] confirms that "due to the lack of effective control over the influence from source and target contexts, conventional neural machine translation tends to yield fluent but inadequate translations".

One research [7] denotes that "although end-to-end neural machine translation has achieved remarkable progress in the past […] years, it suffers from a major drawback: translations generated by neural machine translation systems often lack adequacy. It has been widely observed that neural machine translation tends to repeatedly translate some source words while mistakenly ignoring other words".

When it comes to quality evaluation of Croatian machine translations, [8] has done extensive research in both automatic and human evaluation methods.

Various quality aspects of Croatian machine translations have been analysed in different domains, such as sociology, philosophy, spirituality [9], business correspondence [10] or legislation [11]. Online machine translation services have also been evaluated for the Croatian language [12-14].

III. Research

This section describes the experimental data set, the applied research steps and the human machine translation quality evaluation approach.

In this paper, the authors decided to manually assess the quality of machine-translated literary texts obtained from an earlier research, in which a specific data set was used consisting of:

- a collection of poems written in Croatian by Delimir Rešicki,
- and their human translations into German.

Rešicki's poems originally written in Croatian and their corresponding professional human translations into German were crawled from a multilingual web platform called "Lyrikline" (https://www.lyrikline.org), which aims to attract diverse poets to publish their work online.

In total, 14 Rešicki's poems consisting of 532 verses, i.e. sentences or segments (chunks of text that do not end with a sentence delimiter) per language, comprised the data set which was analysed, preprocessed and prepared for machine translation.

The analysis and preparation phases were mostly accomplished with regular expressions, Python and Perl, and included various tasks, such as, applying the appropriate character encoding (UTF-8), stripping of text formatting, deleting of boilerplate and metadata, removing of redundant and unnecessary characters, tokenising, converting to a 1-1 sentence/verse-based parallel corpus etc.

Professional human translations of Rešicki's poems, which were used as "gold-standard" reference translations in the machine translation trials, were conducted by two well-known translators – Klaus Detlef Olof and Alida Bremer.

Before starting with the actual machine translation quality evaluation, a human evaluator was told to check whether misalignments, orthographic mistakes, various oversights, missing or faulty characters were present in the original data set consisting of Rešicki's poems and the corresponding human translations. The evaluator found that some of the segments (verses) were not correctly aligned, obviously due to the artistic license of professional human translations, which was expected to some degree. Some of the translations were relatively freely translated, while some other, although correct, just appeared in later verses and therefore caused misalignments. Nonetheless, for research purposes the identified imperfections were left unfixed. However, some of the smaller segments, i.e. verses were combined where considered appropriate by the human evaluators.

The machine translations were automatically generated for the Croatian-German and German-Croatian language pairs, i.e. for both directions. However, in this research the authors decided to evaluate only German-Croatian machine translations. Automatic translations were

obtained by using two freely available machine translation systems – Google Translate (https://translate.google.com) and Yandex.Translate (https://translate.yandex.com/), which are both trained on general data, and not specifically on data from the domain of poetry. Both systems are based on statistical and neural machine translation. Once all the machine translations were acquired, a human machine translation quality evaluation followed, in which machine translations were compared to the original human reference translations.

Here the authors decided to analyse the qualitative aspects of the resulting machine translations. Instead of utilising automatic machine translation quality metrics, the authors decided to perform an evaluation with the help of three human evaluators who are native speakers of the Croatian language, and who were told to rate machine translations with respect to the original reference (source) texts and by considering two quality criteria – adequacy and fluency.

When it comes to adequacy, the evaluators used a 4-point scale to rate how much of the meaning is represented in the translation: 4 (everything), 3 (most), 2 (little), and 1 (none). Fluency was also assessed on a 4-point scale. Here, the evaluators had to rate the extent to which the machine translation was well-formed grammatically, contained correct spellings, adhered to common use of terms, titles and names, was intuitively acceptable and could be sensibly interpreted by a native speaker: 4 (flawless), 3 (good), 2 (dis-fluent), and 1 (incomprehensible).

Human evaluations were made on the entire machine-translated data set. After the three evaluators finished grading all of the 532 machine translations per each machine translation system, the evaluators were asked to specify their observations during manual machine translation quality inspection. All quality scores were statistically analysed with regard to both machine translation systems, both evaluation criteria (adequacy and fluency) and with respect to each evaluator. In addition, an inter-rater agreement analysis was performed.

IV. Results and Discussion

Table I presents descriptive statistics of the whole data set calculated on the scores given by the three human evaluators. Overall, the mean scores for both adequacy and fluency are higher for Google Translate, and the standard deviations are smaller. The mean value for adequacy is lower than the mean value for fluency for Google Translate, as opposed to Yandex.Translate – here, the mean adequacy score is slightly higher than the mean fluency score.

TABLE I. Descriptive Statistics

System	Criterion	N	Mean	Std. dev.
Google Translate	Adequacy	532	2.644	0.989
	Fluency	532	2.744	1.000
Yandex.Translate	Adequacy	532	2.534	1.086
	Fluency	532	2.525	1.088

Table II shows average scores given by each evaluator during the human evaluation phase. Overall, scores for Google Translate are (in most cases) higher than for Yandex.Translate. Scores for the adequacy criterion are higher than for fluency according to two evaluators, for both machine translation systems. This is surprising, as earlier research (as stated in the Related Work section) reaffirmed that neural machine translation tends to produce machine translations that are fluent but oftentimes not adequate.

Furthermore, average scores of fluency and adequacy given by all three evaluators are higher for Google Translate.

TABLE II. Scores per Evaluator

Evaluator	1		2		3	
Criterion	A	F	A	F	A	F
Google Translate (average)	2.77	2.64	2.82	2.74	2.34	2.77
Yandex.Translate (average)	2.53	2.27	2.72	2.68	2.35	2.62
Google Translate	A (average for all evaluators): 2.64					
	F (average for all evaluators): 2.72					
Yandex.Translate	A (average for all evaluators): 2.53					
	F (average for all evaluators): 2.52					
Remarks: A = adequacy, F = fluency						

Interestingly, although the evaluators compared only the resulting machine translations with the original source text of Delimir Rešicki, the results of the human evaluation process were still skewed to the right half of the scale, i.e. all scores were above 2. This was the case for both fluency and adequacy. This implies that both machine translation systems proved to be relatively suitable for translating poetry from German into Croatian, despite the fact that the selected machine translation systems were not specially prepared and trained for that particular domain.

The inter-rater agreement, i.e. the reliability of human evaluators, was assessed with the use of Cronbach's alpha. This statistic estimates the internal consistency among evaluators. Cronbach's alpha ranges from 0.00 to 1.00, where 0.00 indicates absolute absence of agreement, and 1.00 perfect agreement among evaluators, while generally the score of 0.70 is considered as reliable. Precise interpretation of Cronbach's alpha values is as follows:

- $\alpha<0.5$ no agreement (unacceptable),
- $0.5\leq\alpha<0.6$ poor agreement,
- $0.6\leq\alpha<0.7$ acceptable agreement,
- $0.7\leq\alpha<0.9$ good agreement, and
- $\alpha\geq0.9$ excellent agreement.

In this research, the alpha values ranged from 0.84 to 0.88 (see Table III), which indicates good internal consistency in the human quality evaluation of machine translations.

TABLE III. CRONBACH'S ALPHA SCORE

System / Criterion	Adequacy	Fluency
Google Translate	0.86	0.84
Yandex.Translate	0.88	0.85

When it comes to the evaluators' observations during manual inspection, the evaluators felt that in many cases the machine translations were actually both adequate and fluent when compared to the human reference translations. This is an important finding, as this implies that machine translation can be used for, at least partially, translating texts in the domain of poetry.

The evaluators commented that in some situations the machine translations were more fluent, and they felt that this derived from the fact that German reference translations of particular verses, which were used as the input for generating Croatian machine translations, were more fluent. For instance, the original (source) text "*Pakao je , reklo bi se , bezizlazan samo za one koji ne znaju iz njega da se vrate ...*" was translated by Google Translate as "*Pakao je beznadan samo za one koji se ne znaju vratiti iz njega ...*".

They also pointed out that there were some cases they felt the professional human translators added a word or changed the sequence of words in a sentence in comparison to the original source text. E.g. the original verse "*druga vučija*" was translated by Google Translate as "*drugo vučje stopalo*" and by Yandex.Translate as "*drugi vuk šapa*". Still, both notions could not be verified by the human evaluators, as they received only the original (source) Croatian text and the two machine-translated texts generated by Google Translate and Yandex.Translate, but not the (reference) human translation into German.

The evaluators also observed that the machine translation systems were not always capable of correctly translating the grammatical case of a feminine noun. Examples of this claim are given below:

- original: "*Za Milanu Vuković Runjić*", error in translation by Google Translate: "*Za Milana Vuković Runjić*", correct translation by Yandex.Translate: "*Za Milanu Vuković Runjić*".
- original: "*Za Eriku Bók , zauvijek*"; error in translation by Google Translate "*za Erika Bóka , zauvijek*", error in translation by Yandex.Translate: "*za Erica boca , zauvijek*".

One of the evaluators denoted that in some sentences the machine translations of synonyms were not adequate, e.g.

- original: "*Ledeni nož u ledenoj torti*",
- Google Translate: "*Ledeni nož u ledenoj pita*", and
- Yandex.Translate: "*Ledeni nož u ledenom kolaču*".

In several occasions the evaluators commented that the machine translation of adjectives was not adequate nor fluent. E.g.

- original: "*jedna mu je noga bila kozja*",
- Google Translate: "*stopala su mu bila jedna od koza*", and
- Yandex.Translate: "*jedna od njegovih nogu bila je kod koze*".

Evaluators also noticed that self-standing sentences were fluent and adequate, but in terms of a stanza they were not fluent and not adequate, and seemed to be out of context, due to the fact that more lines (verses) constituted a stanza. For instance,

- original: "*U to danas vjeruju milijuni/ na šalterima kladionica/ za borbe ljudi i pasa , i jedni i drugi/ na kratkim su lancima .*",
- Google Translate: "*U to danas vjeruju milijuni/ na šalterima kladionica/ za borbe ljudi i pasa , jedna poput druge/ visi na kratkim lancima .*", and
- Yandex.Translate: "*Milijuni vjeruju u to danas/ na policama kladionica/ za borbu protiv ljudi i pasa , sami kao i drugi/ visi na kratkim lancima .*".

One of the evaluators in this experiment stated that the main limitation of the machine translation quality evaluation approach was that the human original text and the corresponding machine translation were presented in a context-free verse-based manner, i.e. only line by line.

In conclusion, according to the evaluators, the generated machine translations were comprehensible in most of the cases but should undergo a post-processing phase where all needed corrections should be made. Nevertheless, in most of the cases the generated machine translations were of relatively decent quality but still not suitable for direct publishing.

V. FUTURE RESEARCH AND ADDITIONAL DIRECTIONS

The authors plan to manually correct the verse alignments in the data set, i.e. to set it to the required 1-1 parallel corpus format, so that the misalignments do not influence the overall evaluation process. Verification of the data set alignments could be done through a specially-built NLP platform [15].

Although the data set was decent in size for evaluation purposes, it might be useful to increment it with additional data from the domain of poetry. Also, more evaluators should probably reaffirm the findings in this paper.

In spite of the fact that the human evaluation results in terms of adequacy and fluency were skewed to the higher scores, in future research native Croatian-German speakers should also be presented the reference translations made by the two human professional

translators. It would be very interesting to see if this could also improve human judgment scores.

The authors plan to repeat the human evaluation for the Croatian-German direction, and on the segment level, i.e. stanza level, instead of the verse level.

Machine translations should also be evaluated with regard to various error types and error classification methods. It might be useful to annotate the data set beforehand, statistically or linguistically [16]. Applying word embeddings could reveal interesting concept-related and semantic relationships between different unigrams [17] as well. Moreover, a sentiment analysis could detect the overall affective states in the poetry-related data set [18].

Analyses of word occurrences and their corresponding distributions [19], extracted key words [20, 21] and concordances [22, 23] might also expose author-specific writing styles and literary elements.

VI. Conclusion

With the rise of machine translation possibilities, the need to demonstrate the effectiveness of technology in translating literary texts is becoming increasingly evident.

The quality of a translation has always been essential to the full comprehension of a source text. This applies even more to the quality of machine translations of complicated, oftentimes ambiguous, or idiomatic texts, such as various literary works. Especially when it comes to studying various poets and other writers in higher education, in order to examine one author's style, concepts etc. it is very important to make sure that the automatic translations are of sufficient quality.

This can be accomplished by applying human quality evaluation of a machine-translated text in comparison to a human translation made by a professional translator.

The aim of this paper was to demonstrate, on the example of a data set containing poems written by a contemporary Croatian poet, the effectiveness of applying machine translation a specific language pair in the domain of poetry by analysing two machine translation quality evaluation criteria – adequacy and fluency. Those criteria were assessed with the help of three human evaluators. Also, the reliability of evaluators was measured as well.

This paper performed, in fact, a usability analysis in order to demonstrate the applicability of poetry-related machine translation. The authors showed that machine translation services can, to some extent, be used for translating texts from various domains. Especially Google Translate generated decent translations, and this was verified during the human machine translation evaluation process and confirmed by the good inter-rater agreement according to Cronbach's alpha.

However, results of this research are to be taken as the first insight due to several reasons. Firstly, translations from the domain of poetry are very specific and subject to personal interpretation. Secondly, the neural machine translation paradigm, which is used by both machine translation services, has introduced new algorithms when compared to statistical machine translation, and their effects on machine translation quality have not yet been explored fully, especially when it comes to flective languages, such as Croatian. Third, evaluating machine translation was done on a relatively small data set and only by three human evaluators. Therefore, the results should be treated as preliminary, and all mentioned limitations should be taken into consideration in follow-up research.

The applied research methods can easily be implemented in education. This could be useful for students in technical sciences and the purpose of learning how to conduct machine translation quality evaluation trials. When it comes to non-technical science students, machine translations could be used, e.g. for "gisting" purposes (obtaining the gist, i.e. the key semantic information about a given text), due to the fact that the elementary notion of a text can be derived from subpar machine translations.

References

[1] A. Görög, "Quality Evaluation Today: the Dynamic Quality Framework," in Proc. of the Translating and The Computer 36 Conference, 2014, pp. 155–164.

[2] Y. Graham, T. Baldwin, and N. Mathur, "Accurate Evaluation of Segment-level Machine Translation Metrics," in Proc. of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 1183–1191.

[3] J. Gutiérrez-Artacho, M.-D. Olvera-Lobo, and I. Rivera-Trigueros, "Human Post-editing in Hybrid Machine Translation Systems: Automatic and Manual Analysis and Evaluation," in Proc. of WorldCIST'18: Trends and Advances in Information Systems and Technologies, 2018, pp 254–263.

[4] S. Castilho, J. Moorkens, F. Gaspari, I. Calixto, J. Tinsley, and A. Way, "Is Neural Machine Translation the New State of the Art?," The Prague Bulletin of Mathematical Linguistics, no. 108, pp. 109–120, June 2017.

[5] M. Martindale, M. Carpuat, K. Duh, and P. McNamee, "Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation," in Proc. of Machine Translation Summit XVII volume 1: Research Track, 2019, pp. 233–243.

[6] Z. Tu, Y. Liu, Z. Lu, X. Liu, and H. Li, "Context Gates for Neural Machine Translation," Transactions of the Association for Computational Linguistics, vol. 5, pp. 87–99, 2017.

[7] Z. Tu, Y. Liu, L. Shang, X. Liu, and H. Li, "Neural Machine Translation with Reconstruction," in Proc. of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17), arXiv:1611.01874 [cs.CL], 2017, p. 7.

[8] I. Dunder, Statistical machine translation and computational domain adaptation (Sustav za statističko strojno prevođenje i računalna adaptacija domene) // doctoral dissertation. Zagreb: University of Zagreb, 2015, p. 281.

[9] S. Seljan, and I. Dunđer, "Automatic Quality Evaluation of Machine-Translated Output in Sociological-Philosophical-Spiritual Domain," in Proc. of the 10th Iberian Conference on Information Systems and Technologies (CISTI'2015), vol. 2, 2015, pp. 128–131.

[10] S. Seljan, and I. Dunđer, "Combined Automatic Speech Recognition and Machine Translation in Business Correspondence Domain for English-Croatian," in Proc. of the International Conference on Embedded Systems and Intelligent Technology (ICESIT 2014) – International Journal of Computer, Information, Systems and Control Engineering, vol. 8, 2014, pp. 1069–1075.

[11] M. Brkić, S. Seljan, and T. Vičić, "Automatic and Human Evaluation on English-Croatian Legislative Test Set," in Proc. of the 4th International Conference (CICLing 2013) "Computational

Linguistics and Intelligent Text Processing", Part I – Intelligent Text Processing and Computational Linguistics, Lecture Notes in Computer Science - LNCS, Springer, 2013, pp. 311–317.

[12] Seljan, S., Klasnić, K., Stojanac, M., Pešorda, B., and Mikelić Preradović, N, "Information Transfer through Online Summarizing and Translation Technology," in Proc. of INFuture2015: e-Institutions – Openness, Accessibility, and Preservation, 2015, pp. 197–210.

[13] S. Seljan, M. Tucaković, and I. Dunđer, "Human Evaluation of Online Machine Translation Services for English/Russian-Croatian," in Proc. of WorldCIST'15 – 3rd World Conference on Information Systems and Technologies (Advances in Intelligent Systems and Computing – New Contributions in Information Systems and Technologies), 2015, pp. 1089–1098.

[14] S. Seljan, and I. Dunđer, "Machine Translation and Automatic Evaluation of English/Russian-Croatian," in Proc. of the International Conference "Corpus Linguistics – 2015" (CORPORA 2015), 2015, pp. 72–79.

[15] R. Jaworski, S. Seljan, and I. Dunđer, "Towards educating and motivating the crowd – a crowdsourcing platform for harvesting the fruits of NLP students' labour," in Proc. of the 8th Language & Technology Conference – Human Language Technologies as a Challenge for Computer Science and Linguistics, 2017, pp. 332–336.

[16] S. Seljan, I. Dunđer, and A. Gašpar, "From Digitisation Process to Terminological Digital Resources," in Proc. of the 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2013), 2013, pp. 1329–1334.

[17] I. Dunđer, and M. Pavlovski, "Through the Limits of Newspeak: an Analysis of the Vector Representation of Words in George Orwell's 1984," in Proc. of the 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2019), 2019, pp. 0691–0696.

[18] I. Dunđer, and M. Pavlovski, "Behind the Dystopian Sentiment: a Sentiment Analysis of George Orwell's 1984," in Proc. of the 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2019), 2019, pp. 0685–0690.

[19] M. Pavlovski, and I. Dunđer, "Is Big Brother Watching You? A Computational Analysis of Frequencies of Dystopian Terminology in George Orwell's 1984," in Proc. of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018), 2018, pp. 0638–0643.

[20] S. Seljan, I. Dunđer, and H. Stančić, "Extracting Terminology by Language Independent Methods," in Proc. of the 2nd International Conference on Translation and Interpreting Studies "Translation Studies and Translation Practice" (TRANSLATA II) – Peter Lang series "Forum Translationswissenschaft", vol. 19, 2014, pp. 141–147.

[21] I. Dunđer, S. Seljan, and H. Stančić, "Koncept automatske klasifikacije registraturnoga i arhivskoga gradiva (The concept of the automatic classification of the registry and archival records)," in Proc. of the 48. savjetovanje hrvatskih arhivista (HAD) / Zaštita arhivskoga gradiva u nastajanju, 2015, pp. 195–211.

[22] R. Jaworski, I. Dunđer, and S. Seljan, "Usability Analysis of the Concordia Tool Applying Novel Concordance Searching," in Proc. of the 10th International Conference on Natural Language Processing (HrTAL2016) – Springer Lecture Notes in Computer Science (LNCS), 2016, p. 6, in press.

[23] I. Dunđer, and M. Pavlovski, "Computational Concordance Analysis of Fictional Literary Work," in Proc. of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018), 2018, pp. 0644-0648.

Using Decision Diagrams of Special Kind for Compactification of Conflict Data Bases Generated by CDCL SAT Solvers

Victor Kondratiev, Ilya Otpuschennikov, Alexander Semenov
Matrosov Institute for System Dynamics and Control Theory SB RAS, Irkutsk, Russia
Email: vikseko@gmail.com, otilya@yandex.ru, biclop.rambler@yandex.ru

***Abstract*—In the paper we propose new algorithms for constructing compact representations of databases of conflict clauses accumulated by state-of-the-art CDCL SAT solvers. These algorithms use the Decision Diagrams of a special kind (the so-called Disjunctive Diagrams). We consider several families of hard SAT instances and use them to compare the implementations of the proposed algorithms and the well-known CUDD package that uses Zero-Suppressed Binary Decision Diagrams (ZBDD) for solving similar problems. The computational experiments clearly show that our algorithm that uses Disjunctive Diagrams has better effectiveness compared to CUDD.**

***Keywords*—Boolean satisfiability problem; conflict driven clause learning; zero-suppressed binary decision diagrams; disjunctive diagrams**

I. Introduction

There exist several types of Decision Diagrams that are all widely used in various areas of computer science. One of the most well-known class of them is formed by the Binary Decision Diagrams (BDD) and their optimized form – Reduced Ordered Binary Decision Diagrams (ROBDD). These two kinds of diagrams were studied in literally thousands of papers. The key study of the algorithms for handling ROBDD [1] by R. Bryant for a long time was among the top cited papers on computer science. For many years ROBDD were used as the main algorithmic tool in symbolic verification [2]. However, starting from 1999, they are having been replaced by the solvers for the Boolean Satisfiability problem (SAT) [3]. The peculiar feature of ROBDDs consists in that they provide the maximally compact representation of a total Boolean function over a class of graphs defined using the well-known Shannon expansion [4]. Unfortunately, the problem of constructing a ROBDD representation for an arbitrary Boolean function defined by formula F has extremely high complexity. It can be shown that if there exists an algorithm for solving this problem in polynomial time (on the length of F), then $P = \#P$. Thus, generally speaking it is not realistic to construct ROBDD representations for arbitrary formulas of significant size. A large class of such formulas is formed by the databases of conflict clauses collected by state-of-the-art SAT-solvers that employ the Conflict-Driven Clause Learning (CDCL) concept [5]. The process of accumulating such a database can be organized in a parallel or distributed environment, thus producing gigabytes of conflict clauses for a single formula.

The main goal of representing Boolean formulas in the form of diagrams is to make them more compact in computer memory. A popular kind of diagrams often used today for such a purpose is called a Zero-suppressed Binary Decision Diagams (ZBDD or ZDD). They were first introduced by S.-I. Minato in 1993 [6]. Despite the fact that ZDD can be considered as a special case of BDD, they have a completely different nature compared to ROBDD, and can be used to effectively represent arbitrary sets of subsets. Therefore, it is possible to use ZDD to represent conflict databases if we view each conflict clause as a set of literals, and the Conjunctive Normal Form (CNF) - as a set of the corresponding subsets.

It might seem surprising, but the number of available software implementations of algorithms for handling ZDD is very small. The authors of the majority of papers that employ ZDD in practice use the CUDD package, introduced in [7]. The paper [8] contains a clear and useful tutorial for working with CUDD.

In [9] we introduced a class of decision diagrams that are designed to represent arbitrary Disjunctive Normal Forms (DNF) as graphs of a special kind. These diagrams, like the well-known ZDD, are constructed for an arbitrary DNF in polynomial time from the length of its description. However, unlike ZDDs, they are not binary. On the other hand, the diagrams described in [9] do not introduce additional variables to represent DNFs, while ZDDs do (see the next chapter). Article [9] contains mainly theoretical results. This work provides a number of computational experiments which show the effectiveness of our software tool for building diagrams from [9] compared to the CUDD tool used for constructing ZBDD.

Let us present a brief outline of the paper. In the next section we give the main notation and the results that are necessary for comprehension of the remaining part of the paper. In Section 3 we provide a brief description of the class of diagrams introduced in [9]. Their main goal is to compactly represent the conflict databases accumulated by modern CDCL SAT solvers. In Section 4 we first describe two developed software tools, ZDDgen and DDiagrams, that take as an input a Boolean formula in CNF or DNF and produce a ZDD or diagrams from [9], respectively. After this we perform

an extensive computational experiment aimed at comparison of the effectiveness of the developed tools with each other and also with the ZRES solver [10]. As the test instances we used the conflict databases accumulated by the MINISAT 2.2 solver on SAT instances encoding the inversion problems for several cryptographic functions. It turned out that the diagrams from [9] usually provide a more compact representation of the corresponding conflict database. In the last section we briefly conclude the obtained results and outline some directions for future work.

II. PRELIMINARIES

The variables that take the values from the set $\{0,1\}$ are called Boolean. By $\{0,1\}^n$ we denote the set of all Boolean vectors of length n. A Boolean formula F is an expression built using specific rules (see e.g. [11]) over an alphabet that contains Boolean variables, braces and Boolean connectives. An arbitrary Boolean formula F over n Boolean variables defines a total Boolean function $f_F : \{0,1\}^n \to \{0,1\}$. An arbitrary Boolean vector $\alpha \in \{0,1\}^n$ can be viewed as an assignment of variables from $X = \{x_1, \ldots, x_n\}$. Formula F over X is calles *satisfiable* if there exists such an assignment $\alpha \in \{0,1\}^n$ of variables from X that $f_F(\alpha) = 1$. Otherwise F is called *unsatisfiable*. Any assignment α, such that $f_F(\alpha) = 1$ is called a *satisfying assignment* of F. An arbitrary variable $x \in X$ or its negation $\neg x$ is called a *literal*. Two literals x and $\neg x$ are called contrary. An arbitrary disjunction of different literals among which no two are contrary is called a *clause*. A conjunction of different clauses is called a *Conjunctive Normal Form* (CNF). The Boolean Satisfiability problem (SAT) for an arbitrary Boolean formula F is then formulated as follows: for formula F to decide if it is satisfiable. SAT for an arbitrary F can be reduced to SAT for some CNF in polynomial time. SAT is the historically first NP-complete problem [12], and thus it is highly unlikely that it can be solved in polynomial time in the general case.

Despite the NP-completeness of SAT the development of computational algorithms for its solving has been relevant for a long time. At first it was justified by the interest to the automatical theorem proving methods. However, the truly intensive development of SAT-solving algorithms started once researchers began using them for symbolic verification [2]. The CDCL algorithm is the foundation of the majority of state-of-the-art solvers used in verification, combinatorics, cryptanalysis, bioinformatics and other areas. CDCL is short for *Conflict-Driven Clause Learning*. This strategy was first described in [13] and later in more detail in [14]. Its main idea is to use computer memory to store parts of the search history. In particular, the information on refuted branch of the decision tree (traversed by the DPLL algorithm [15]) that resulted in a conflict is stored in form of a new clause called *learnt clause*. The sets of learnt clauses, which are accumulated with CDCL based SAT solvers, further referred as *conflict databases*.

Modern SAT solvers can generate hundred of megabytes of learnt clauses in mere minutes. It is worth to be mentioned that to achieve high efficiency SAT solvers need to store learnt clauses in RAM. As a result of this fact, a risk of memory overflow appears. To avoid this effect CDCL SAT solvers employ special techniques that remove the learnt clauses which are viewed as irrelevant. All such decisions are heuristical, and as a result some important information can be lost. Thus, each method which represents in compact form learnt clauses databases can be viewed as valuable. Especially, it regards situations when hard SAT instances are solved in distributed computing environments: in such cases one can use compact representations of learnt clauses databases for collecting usefull information which is accumulated on different computational nodes and then is sent to some central node for additional analysis. Thus there is a need to transform a conflict base into a representation that takes sufficiently less memory compared to its original form. This is one of the use cases for the decision diagrams.

As we mentioned before, ROBDD can be used to represent Boolean functions. ROBDD is a directed acyclic graph with one root vertex and two terminal vertices associated with symbols 0 and 1. An arbitrary non-terminal ROBDD vertex has exactly two children. With any non-terminal ROBDD vertex we associate a Boolean varaiable, and different vertices can be associated with the same variable. A sequence of vertices formed by any path from ROBDD root vertex to one of its terminal vertices must follow a pre-specified order on the set of variables. For example, if the order is $x_1 < \ldots < x_n$, then in such ROBDD there is no path where a vertex associated with x_{k+1} will be met before a vertex associated with x_k. The arcs in ROBDD are usually marked by solid and dotted lines. Let v be an arbitrary non-terminal vertex and x_v be a variable associated with v. Two arcs go out from v: one is a solid line (1-edge) and one is a dotted line (0-edge). The solid line is interpreted as x_v being assigned the value of 1 and the dotted line – the value of 0. Thus, an arbitrary path in ROBDD from a root vertex to a terminal vertext $\gamma \in \{0,1\}$ is interpreted as a set of values of the corresponding variables. Therefore, an arbitrary ROBDD uniquely corresponds to some Boolean function: an arbitrary path from root to γ defines an assignment (or a set of assignments) for variables from X over which the corresponding function takes the value of γ. Using the Shannon expansion it is easy to show that an arbitrary total Boolean function can be represented as some ROBDD. Sometimes, such a representation can be extremely compact. For example, a linear function over n variables, i.e. the function defined by a formula $x_1 \oplus \ldots \oplus x_n$ has a ROBDD-representation with only $2n + 1$ vertices. However, in the general case a ROBDD representation is not obliged to be compact. Moreover, if a function is constant, then its ROBDD representation consists of a single terminal vertex, but it is impossible to effectively construct such ROBDD (at least using the algorithms known today). There exists an algorithm [1] that for an arbitrary ROBDD can count the number of the assignments on which the corresponding Boolean function takes the value of 1, and does it in time polynomial in the number of vertices. It means that the problem of constructing a ROBDD representation for a function defined by a Boolean

formula is at least as complex as $\#P$-complete problem of counting the number of satisfying assignments for an arbitrary Boolean formula. Thus, ROBDD does not suit well for representing conflict databased accumulated in the course of CDCL runtime.

Another type of diagrams that can actually be used to represent conflict databases is called Zero-suppressed Binary Decision Diagrams (ZDD). In more general case, we can use these diagrams to effectively represent sets of subsets of some finite set. By 'effectively' we here mean that the corresponding algorithm works in time limited by a polynomial in the total number of elements in a considered set of subsets. Similar to ROBDD, ZDD has two terminal vertices associated with symbols 0 and 1. Any non-terminal vertex has two children, and the edges are divided into 1-edges and 0-edges.

Let us briefly remind the idea behind the basic algorithms for handling ZDD. Consider an arbitrary set $A = \{a_1, \ldots, a_N\}$ and some set $B \in 2^A$, the elements of which are the subsets of A. With each a_i, $i = 1, \ldots, N$ we associate a Boolean variable y_i and a pair of literals $\{y_i, \neg y_i\}$. Let $b = \{a_1^b, \ldots, a_s^b\}$ be an arbitrary element from B. Let us associate with b a Boolean formula of the following kind

$$\phi(b) = l_1(b) \wedge \ldots \wedge l_N(b), \tag{1}$$

where for each $i \in \{1, \ldots, N\}$

$$l_i(b) = \begin{cases} y_i, & if \quad a_i \in b \\ \neg y_i, & if \quad a_i \notin b \end{cases}$$

With the set B we associate a Boolean formula

$$\Phi(B) = \bigvee_{b \in B} \phi(b) \tag{2}$$

The formula (2) specifies the so-called characteristic function of set B [8].

Now consider an arbitrary CNF C over a set of Boolean variables $X = \{x_1, \ldots, x_n\}$. Let us introduce a set of literals $L_X = \{x_1, \neg x_1, x_2, \neg x_2, \ldots, x_n, \neg x_n\}$. Consider an arbitrary clause from C as a subset of set L_X, and the whole CNF C as a set of the corresponding subsets. Now, introduce the set of Boolean variables $Y = \{y_1, \ldots, y_N\}$, $N = 2n$, in which y_i corresponds to literal with number i in L_X, $i \in \{1, \ldots, 2n\}$. With each clause from C we associate a formula of the kind (1), and with CNF C – a formula of the kind (2).

Example 1. Consider CNF $C = (x_1 \vee \neg x_2 \vee x_3) \wedge (\neg x_1 \vee \neg x_3) \wedge (x_2 \vee \neg x_3)$. Here $X = \{x_1, x_2, x_3\}$, $L_X = \{x_1, \neg x_1, x_2, \neg x_2, x_3, \neg x_3\}$, $Y = \{y_1, y_2, y_3, y_4, y_5, y_6\}$. Let's associate with clauses of CNF C formulas $\phi(b_1)$, $\phi(b_2)$ and $\phi(b_3)$, that look as follows:

$$\phi(b_1) = y_1 \wedge \neg y_2 \wedge \neg y_3 \wedge y_4 \wedge y_5 \wedge \neg y_6$$

$$\phi(b_2) = \neg y_1 \wedge y_2 \wedge \neg y_3 \wedge \neg y_4 \wedge \neg y_5 \wedge y_6$$

$$\phi(b_3) = \neg y_1 \wedge \neg y_2 \wedge y_3 \wedge \neg y_4 \wedge \neg y_5 \wedge y_6$$

A formula of the kind (2) corresponding to CNF C, when we view C as a set $B = \{b_1, b_2, b_3\}$ will take the following form:

$$\Phi(B) = \phi(b_1) \vee \phi(b_2) \vee \phi(b_3) \tag{3}$$

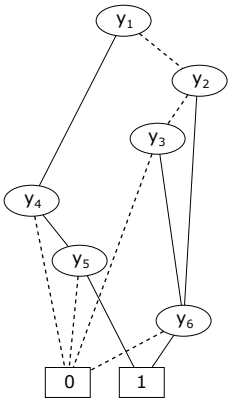


Fig. 1. ZDD-representation of formula (3) from Example 1.

Now we can construct ZDDs for formulas of the kind (2). An arbitrary formula of the kind (1) will be represented in ZDD by a path from a root to terminal vertex marked with 1. Remind that in ZDD any vertex from which the 1-edge goes to terminal 0 is deleted. Thus, for example, when representing the formula $y_1 \wedge \neg y_2 \wedge \neg y_3 \wedge y_4 \wedge y_5 \wedge \neg y_6$ the vertices associated with $\neg y_2, \neg y_3, \neg y_6$ will be removed in the final diagram.

Essentially, an arbitrary formula (2) is a Canonical Disjunctive Normal Form (CDNF) over set Y. It cah be shown that if we use a hash table to store vertices, then for an arbitrary CDNF the corresponding ZDD representation can be constructed in time linear in the number of literals in CDNF. The proof is constructed in the same manner as in [9].

Comments. Figure 1 presents the ZDD where three paths lead from the root to terminal 1-vertex. Each of these paths represents a clause as a set formed by literals. Thus, for example, the rightmost path from the root to 1 represents formula $\neg y_1 \wedge y_2 \wedge y_6$ which corresponds to clause $(\neg x_1 \vee \neg x_3)$.

As we noted above, one of the most well-known and freely available tools for constructing ZDD is the CUDD library [7], the main capabilities of which were described in [8].

III. Disjunctive diagrams and their application to representing conflict databases

The Disjunctive Diagrams were first introduced in [9]. Their goal is to represent the conflict databases accumulated by modern CDCL SAT solvers. DisJunctive Diagrams (DJD) are directed acyclic graphs that can be used to effectively represent arbitrary Disjunctive Normal Forms (DNFs). Unlike ROBDD and ZDD, DJD are not binary: an arbitrary non-terminal vertex can have up to $2n - 2$ children, where n is the number of variables in a considered DNF.

Let D be an arbitrary DNF over a set $X = \{x_1, \ldots, x_n\}$. The construction of DJD $R(D)$ that represents D is split into two stages. First, for D containing m elementary conjunctions the algorithm constructes a special disjunctive forest in time $O(n \cdot m)$. Each tree in this forest has two terminal vertices: one is marked with 1 and each path from a tree root to this vertex corresponds to a specific elementary conjunction. Another terminal vertex is marked with symbol ? and any path

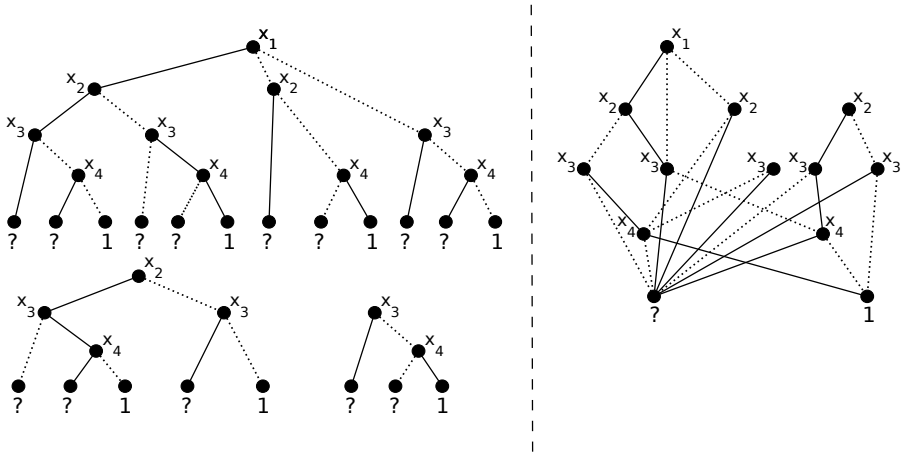


Fig. 2. Disjunctive forest $F(D)$ (left) and DJD $R(D)$ (right) for DNF D.

to ?-vertex defines a partial assignment for which the value of a function specified by a considered DNF is not known. On the second stage all terminal vertices of the disjunctive forest marked with the same symbol are glued together. Then, the constructed graph is traversed in the direction from terminal vertices to root (there can be several of them). In this process, using the rules described in [9], some of the vertices are glued together. The resulting DJD represents an original DNF. The complexity of the transition from an original DNF D to DJD $R(D)$ is $O(n^2 \cdot m)$. On Figure 2 we show the disjunctive forest and DJD representing the following DNF (with the order $\tau : x_1 < x_2 < x_3 < x_4$).

$$\begin{aligned} D = \quad & (x_1 \wedge x_2 \wedge \neg x_3 \wedge \neg x_4) \vee (x_1 \wedge \neg x_2 \wedge x_3 \wedge x_4) \vee \\ & (\neg x_1 \wedge \neg x_2 \wedge x_4) \vee (\neg x_1 \wedge \neg x_3 \wedge \neg x_4) \vee \\ & (x_2 \wedge x_3 \wedge \neg x_4) \vee (\neg x_2 \wedge \neg x_3) \vee (\neg x_3 \wedge x_4) \end{aligned}$$

In [9] we showed that DJD provide an effective apparatus for representing DNFs that are constructed by negating the CNFs formed by databases of learnt clauses. To accumulate conflict databases we used the hard SAT instances encoding the cryptanalysis problems. Note, that cryptanalysis problems for fast cryptographic functions can be effectively reduced to SAT. For this purpose one can employ various software tools. We used the TRANSALG software system [16] which was designed to take into account the cryptographic specifics.

In this paper we considered the SAT instances that encode the cryptanalysis of several keystream generators and the SAT encoding of the DES block cipher. CNFs encoding cryptanalysis problems of strong ciphers are extremely hard for all known SAT solvers. Thus we can use such CNFs to accumulate conflict databases that are as large as possible. In our experiments each such a CNF was given as an input to the MINISAT 2.2 SAT solver [17] modified to not remove any accumulated learnt clauses. Using this way we accumulated conflict databases which consisted of dozens of millions literals. We additionaly emphasize, that our goal is to construct a diagram, the number of nodes in which is substantially less than the number of literals in source database.

IV. COMPARATIVE ANALYSIS OF EFFICIENCY OF DJD AND ZDD FOR REPRESENTATION OF CONFLICT DATABASES

In our computational experiments we used several software systems: DDIAGRAMS software system[1], designed to construct DJD; ZDDGEN software system[2] that constructs ZDD-representations of arbitrary DNFs, viewing them as sets of subsets of literals; ZRES SAT solver [10], [18] which uses ZDD to represent CNFs.

Note, that to the best of our knowledge, the sources of the ZRES SAT solver are not available to general public and in our experiments we used the binary executable file for Linux operating system. Therefore, we often lacked the detailed data on the process of constructing ZDD by ZRES since if a given SAT instance is not solved within the time limit, this SAT solver outputs only the number of vertices in a constructed diagram but not the time of its construction.

The DDIAGRAMS software tool is essentially a modernized version of the program used to perform the computational experiments in [9]. The ZDDGEN is basically a software shell which under the hood employs the CUDD library which constructs ZDD. We would like to note once again, that DDIAGRAMS constructs disjunctive diagrams for an arbitrary DNF, while ZRES works only with CNFs. Actually, it is easy to see that the ZDDs for CNF C and DNF $D = \neg C$ are represented by identical graphs. The difference only is that in the case of CNF the paths in ZDD correspond to disjunctions and in the case of DNF they represent conjunctions. The ZDDGEN tool can work with both CNFs and DNFs. Below we show the results for ZDDs, for which ZDDGEN worked with DNFs of the kind $D = \neg C$, where C is the considered conflict database. The results of all computational experiments are summed up in the Table I.

Let us give some comments to Table I. As we already noted, in our experiments we considered the conflict databases accumulated by the MINISAT 2.2 SAT solver on the SAT encod-

[1]https://gitlab.com/disjunctivediagrams/
[2]https://github.com/Vikseko/ZDDgen/

TABLE I
COMPARISON OF ZRES, ZDDGEN AND DDIAGRAMS

	ZRes (time s. / nodes)	ZDDgen (time s. / nodes)		DDiagrams (time s. / nodes)	
Number of literals		**Direct Order**	**Freq. Order**	**Activity Order**	**Freq. Order**
DES_56/64 (120 vars)					
5203888	13.81 / 201594	5.02 / 201594	20.86 / 106020	11.18 / 181675	18.59 / 4055402
19468640	130.10 / 790371	21.78 / 790371	72.64 / 537117	48.77 / 517696	60.43 / 12092601
29040930	298.06 / 1241841	37.81 / 1241841	119.85 / 839367	84.11 / 909715	90.61 / 17432456
57250792	-	73.19 / 2213767	242.24 / 1379828	115.68 / 2928421	180.97 / 33422591
76290051	-	99.60 / 3043371	-	225.48 / 1740524	227.71 / 41643044
Sum_66/180 (322 vars)					
2053627	7.38 / 489462	22.28 / 489462	52.86 / 431506	7.12 / 367140	7.42 / 363827
3914149	? / 891319	46.75 / 891319	105.25 / 793655	13.36 / 668541	13.91 / 678308
7536590	? / 1569105	87.85 / 1569105	199.29 / 1360770	25.23 / 1143595	25.45 / 1178072
11170269	? / 2304183	140.38 / 2304183	290.72 / 2010567	36.71 / 1689881	37.36 / 1744536
21262312	-	335.37 / 4318608	-	65.84 / 3286418	67.90 / 3309447
Tresh_72/150 (531 vars)					
2666726	21.64 / 1188110	95.94 / 1188110	135.42 / 866387	10.70 / 818199	11.14 / 799766
5152722	? / 2163954	213.00 / 2163954	256.28 / 1547544	21.13 / 1430311	21.68 / 1432970
9870081	? / 3918591	409.99 / 3918591	516.76 / 2856377	40.72 / 2595249	40.83 / 2616915
19011085	? / 7099187	-	-	77.23 / 4747392	78.06 / 4708532
30120691	-	-	-	121.06 / 7030924	121.35 / 7070055
E0_128/128 (762 vars)					
8099227	? / 4219578	510.34 / 4219578	-	44.50 / 3582045	53.90 / 3164457
14480734	-	-	-	78.44 / 5739386	88.43 / 5303400
19468640	-	-	-	174.53 / 12861640	183.23 / 12416659
38416459	-	-	-	228.56 / 15517731	232.89 / 15457485
46161213	-	-	-	277.80 / 18228707	288.55 / 17659373
Grain_160/160 (1032 vars)					
7709217	? / 5659781	-	-	47.84 / 3689845	49.07 / 3639412
15105804	? / 11164203	-	-	95.61 / 7073861	96.65 / 7044664
29905843	-	-	-	192.81 / 12833684	196.13 / 12731346
37289442	-	-	-	246.65 / 15500971	249.55 / 15376630
44679188	-	-	-	294.93 / 18469107	299.17 / 18281523

ings for the problems of cryptanalysis of several symmetric ciphers. Headers over the sets of rows correspond to considered cryptanalysis problems. For example, `DES_56/64` means the problem of cryptanalysis of a single block of ciphertext (64 bits) of block cipher DES with the secret key of length 56 bits. `sum_66/180` corresponds to the cryptanalysis of the 180-bit keystream fragment constructed by the summation generator [19] with the secret key of size 66 bits. `thresh_72/150` means the cryptanalysis of the 150-bit keystream fragment produced by the threshold generator [19] with 72-bit secret key. `e0_128/128` corresponds to cryptanalysis of the 128-bit keystream fragment obtained from the e0 generator [20] with 128-bit secret key. Finally, `Grain_160/160` [21] means the state recovery attack on the Grain v1 stream cipher, that uses 160 bits of keystream.

In the leftmost column of the Table I the total number of literals in CNF encoding of a corresponding cryptanalysis problem is shown. In the remaining part of the Table I the parameters of the constructed diagrams are presented. Concretely, we show the number of nodes and the time it took to construct a diagram. We used three software tools for constructing diagrams. First is the ZRES SAT solver [10], which yields ZDD as a representation of the considered conflict database. Note that due to the reasons outlined above, we could not always obtain its runtime (the '?' symbol marks this fact in several cells of the table 1). Also, a surprising fact is that ZRES always builds diagrams using the order in which the variables appear in the original CNF (the so-called direct order). Meanwhile, it is a known fact that when constructing diagrams of the considered kind (ROBDD, ZDD, DJD), the order of variables can significantly influence the size of the resulting diagram.

The second and the third columns of the Table I contain the results of the application of ZDDGEN to the considered tests. In the second column we show the results for the direct order (note, that in this case the number of vertices is the same as

in the diagrams constructed by ZRES). The results in the third column correspond to the so-called frequency order, i.e. the literals that are met more frequently in the conflict database are closer to the root of the ZDD (and the root is associated with the most frequent literal).

In the last two columns of the table we show the results for the DJD constructed by DDIAGRAMS for DNFs of the kind $D = \neg C$, where C is the considered conflict database. An interesting observation consists in the fact that the direct order for DJD yields the worst results. We also show the data for the frequency order and the so-called activity order. To form the latter we use the conflict activity of the clauses from the conflict database, computed using the Variable-State Independent Decaying Sum (VSIDS) heuristic [22] used in MINISAT 2.2.

For each of the listed problems we generated 5 independent test instances. Each software tool was given 1000 seconds to construct the corresponding diagrams. In the computational experiments we used the "Academician V.M. Matrosov" computing cluster of Irkutsk Supercomputer Center [23]. In a particular experiment each tool employed a single core of Intel Xeon E5-2695v4 processor.

From Table I it is clear that DJD are better than ZDD in the majority of tests in the number of vertices. Also, DDIAGRAMS loses to ZDDGEN and ZRES in the time of diagram construction only on a single test (`DES_56/64`). It is quite surprising because in theory the time of ZDD construction has a lower growth rate compared to that for DJD. We believe that it is likely that the method which CUDD uses to construct ZDD representation of characteristic function is not optimal.

V. CONCLUSION AND FUTURE WORK

In the paper we presented the two software tools: ZDDGEN and DDIAGRAMS developed specifically for constructing effective representations of conflict databases accumulated by state-of-the-art CDCL SAT solvers, in the form of Decision Diagrams. In particular, we constructed the representations of these databases as Zero-Suppressed Binary Decision Diagrams (ZDD) introduced by S.-I. Minato in [6] and as DisJunctive Diagrams (DJD) introduced by us in [9]. In the role of test instances we considered the conflict databases accumulated by the MINISAT 2.2 SAT solver on SAT instances that encode cryptanalysis of several cryptographic fucntions. The experiments showed that DJD usually yield more compact representations of conflict databases compared to ZDD. It turned out that the performance of the CUDD library which is often used to construct ZDDs is apparently not optimal. In the nearest future we plan to study the possibility of implementing the derivation mechanisms based on DJD which are similar to those proposed in [24]. We also plan to implement a tool for constructing ZDDs without CUDD that will use the algorithms proposed in [9] to transition from tree-like structures to diagrams.

ACKNOWLEDGMENTS

The research was supported by the Program of the Fundamental Research of the Siberian Branch of the Russian Academy of Sciences, proj. AAAA-A19-119062590002-7.

REFERENCES

[1] R. E. Bryant, "Graph-based algorithms for boolean function manipulation," *IEEE Trans. Comput.*, vol. 35, no. 8, p. 677–691, Aug. 1986.

[2] E. M. Clarke, O. Grumberg, and D. A. Peled, *Model checking*. MIT Press, 1999.

[3] A. Biere, A. Cimatti, E. Clarke, and Y. Zhu, "Symbolic model checking without bdds," in *TACAS*, ser. Lecture Notes in Computer Science, W. R. Cleaveland, Ed., vol. 1579, 1999, pp. 193–207.

[4] C. E. Shannon, "The synthesis of two-terminal switching circuits," *Bell System Technical Journal*, vol. 28, no. 1, pp. 59–98, 1949.

[5] J. P. Marques-Silva, I. Lynce, and S. Malik, "Conflict-driven clause learning SAT solvers," in *Handbook of Satisfiability*, A. Biere, M. Heule, H. van Maaren, and T. Walsh, Eds. IOS Press, 2009, vol. 185, pp. 131–153.

[6] S.-I. Minato, "Zero-suppressed bdds for set manipulation in combinatorial problems," in *Proceedings of DAC '93*. ACM, 1993, p. 272–277.

[7] F. Somenzi, *CUDD: CU Decision Diagram package – release 2.7.0*, Department of Electrical and Computer Engineering – University of Colorado at Boulder, June 2018.

[8] A. Mishchenko, "An introduction to zero-suppressed binary decision diagrams," in 'Proceedings of the 12th Symposium on the Integration of Symbolic Computation and Mechanized Reasoning, Tech. Rep., 2001.

[9] A. A. Semenov and I. V. Otpuschennikov, "On one class of decision diagrams," *Automation and Remote Control*, vol. 77, no. 4, pp. 617–628, 2016.

[10] P. Chatalic and L. Simon, "Zres: The old davis–putnam procedure meets zbdd," in *Automated Deduction - CADE-17*, ser. Lecture Notes in Computer Science, vol. 1831, 2000, pp. 449–454.

[11] A. Biere, M. Heule, H. van Maaren, and T. Walsh, Eds., *Handbook of Satisfiability*, vol. 185. IOS Press, 2009.

[12] S. A. Cook, "The complexity of theorem-proving procedures," in *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing, May 3-5, 1971, Shaker Heights, Ohio, USA*, 1971, pp. 151–158.

[13] J. P. Marques-Silva and K. A. Sakallah, "GRASP: a new search algorithm for satisfiability," in *Proceedings of ICCAD '96*, 1996, pp. 220–227.

[14] J. P. Marques-Silva and K. A. Sakallah, "GRASP: a search algorithm for propositional satisfiability," *IEEE Trans. Computers*, vol. 48, no. 5, pp. 506–521, 1999.

[15] M. Davis, G. Logemann, and D. Loveland, "A machine program for theorem-proving," *Commun. ACM*, vol. 5, no. 7, p. 394–397, 1962.

[16] I. Otpuschennikov, A. Semenov, I. Gribanova, O. Zaikin, and S. Kochemazov, "Encoding cryptographic functions to sat using transalg system," in *ECAI 2016, Frontiers of Arificial Intelligence an Applications*, vol. 285, 2016, pp. 1594–1595.

[17] N. Eén and N. Sörensson, "An extensible SAT-solver," in *SAT 2003*, ser. Lecture Notes in Computer Science, vol. 2919, 2004, pp. 502–518.

[18] P. Chatalic and L. Simon, "Multiresolution for SAT checking," *International Journal on Artificial Intelligence Tools*, vol. 10, no. 4, pp. 451–481, 2001.

[19] A. J. Menezes, P. C. van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*, 1st ed. Boca Raton: CRC Press, 1996.

[20] Y. Lu and S. Vaudenay, "Faster correlation attack on bluetooth keystream generator e0," in *CRYPTO 2004*, ser. Lecture Notes in Computer Science, vol. 3152, 2004, pp. 407–425.

[21] T. J. Martin Hell and W. Meier, "Grain: a stream cipher for constrained environments," *International Journal of Wireless and Mobile Computing*, vol. 2, no. 1, pp. 86–93, May 2007.

[22] M. W. Moskewicz, C. F. Madigan, Y. Zhao, L. Zhang, and S. Malik, "Chaff: Engineering an efficient SAT solver," in *DAC'01*, 2001, pp. 530–535.

[23] "Irkutsk Supercomputer Center of SB RAS. URL: http://hpc.icc.ru." [Online]. Available: http://hpc.icc.ru

[24] A. Ignatiev and A. A. Semenov, "DPLL+ROBDD derivation applied to inversion of some cryptographic functions," in *SAT 2011*, ser. Lecture Notes in Computer Science, vol. 6695, 2011, pp. 76–89.

Detecting objects in drone imagery: a brief overview of recent progress

Saša Sambolek*, Marina Ivašić-Kos **
* High school Tina Ujevića, Kutina, Croatia
** University in Rijeka, Department of Informatics, Rijeka, Croatia
sasa.sambolek@gmail.com, marinai@uniri.hr

***Abstract* - Detecting objects on unmanned aerial vehicle (Drones) imagery is a challenging task and an under-researched problem that lately receives more and more attention in the research community. When shooting with a drone, not only the weather and the light conditions change, but the shooting height and angle change as well because the position of the camera is not fixed during shooting. The paper aims to describe the possibility of using drones in search and rescue operations and to provide a comprehensive overview of the area related to the detection of persons in drone imagery. The paper includes a description of publicly available datasets and a comparison of the state-of-the-art models for person detection in drone recordings and ends with a proposal for future research.**

Keywords – drone dataset; object detection; CNN

I. INTRODUCTION

With the development of technology, unmanned aerial vehicles (UAVs, drones) equipped with cameras find their application in industry, agriculture, surveillance, and search and rescue operations. Detecting objects on drone images is an extremely useful and still under-researched problem.

When flying at low altitudes, drone records more details on objects of interest, while larger altitudes cover a larger area. Detecting objects in drone imagery creates greater challenges than traditional panorama detection. One of the reasons is the change in shooting height which significantly affects the size of the desired object or change of the shooting angle [1]. In a single video, in a very short time, a drone can record an object from the front, side, or a bird's eye view. Changes in lighting (day, night) and weather (sunny, cloudy, foggy, or rainy) drastically affect the visibility and display of an object [1]. In addition to all of the above, the challenge in detecting and monitoring is also posed by the rapid movements of the camera, the occlusion, and the relative movement between the camera and the object.

As objects captured by a drone are often too small to be detected by the human eye on a drone control screen, object detection needs to be automated. In recent years, considerable progress has been made in detecting objects using deep learning (convolutional neural networks). The most popular deep learning-based detectors are Faster R-CNN [1], SSD [4], YOLO [5], and RetinaNet [6] trained on datasets like PASCAL VOC [7] or MS COCO [8]. However, it turns out that they are not equally successful when applied to drone-recorded images.

In order to improve the results of object detection on drone imagery, it is necessary to include these images in a training set. However, until recently, there were no publicly available datasets recorded by the drones. With datasets like Campus [9], UAV123 [10], CARPK [11], Okutama-action [12], UAVDT [13] and VisDrone [14], images taken with the drones are available, but there are tailored to specific issues such as parking control, traffic monitoring, or movement of people across pedestrian zones.

This paper aims to provide a simple but comprehensive overview of object detection on imagery recorded by drones to study existing databases and models that could be used in the detection of persons in search and rescue operations.

During the search and rescue operation, it is important to find the missing person as quickly as possible, as the survival of the missing person declines exponentially over time [15]. The weather and light conditions vary greatly between different search and rescue operations, which is an additional challenge in detecting a missing or injured person.

Today, almost all search and rescue services have integrated the use of drones in their search.

This is due to the increasing availability of drones with quality high-resolution cameras. It takes approximately 25 seconds for a video analyst to detect a victim on a drone recording [20]. The benefit of video analytics is knowing the context of the image and predicting where the person may be based on previous experiences, but the analyst focuses only on a small portion of the image so the assistance of an automated detector can be of great use.

The rest of the paper is organized as follows: in Section II. we will present public available drones datasets. An overview of methods using drones in search and rescue missions is given in Section III with an overview of the state-of-the-art object detector algorithms in drone imagery. The paper ends with a conclusion and a proposal for future research.

II. PUBLIC AVAILABLE DRONES DATASETS

A comparison of publicly available sets of images taken with a drone and prepared for deep learning tasks is given in Table 1.

TABLE 1. COMPARISON OF PUBLICLY AVAILABLE DRONE DATASETS

	Campus	UAV123	CARPK	Okutama-action	UAVDT	VisDrone
Year	2016	2016	2017	2017	2018	2018
D			x	x	x	x
S		x			x	x
M	x				x	x
Frames	929,5k	112 578	1 448	77 365	80 000	189 473
Boxes	19,5k	110k	90k	422,1	841,5k	2 500k
Categories	6	6	1	1	1	10
Resolution	1400x1904	1280x720	1280x720	3840x2160	1080x540	3840x2160

D – detection, S – single object tracking, M – multiple object tracking, 1k = 1 000

A. Campus

The Campus is a large dataset containing images and videos of different classes such as pedestrians, bicyclists, cars, skateboarders, golf carts, buses, taken inside the campus from a bird's eye view (see

Figure 1. a). The footage was taken with a 4K drone-mounted camera (3DR solo) flying at a height of approximately 80 meters. The dataset contains about 19,000 objects at a resolution of 1400 x 1904 px, [9].

B. UAV123

The UAV123 dataset contains a set of scenes ranging from urban landscapes, roads, fields, and beaches with objects such as cars, trucks, boats, and persons. Persons are additionally tagged for object tracking. Activities such as walking, cycling, swimming, car driving are also labeled.

The data are divided into 3 sub-groups [10]:

- 103 video clips taken with the DJI S1000 drone tracking different objects between 5m and 25m in height, 720p, and 4K resolution at 30 and 96 fps.
- 12 videos shot with images of lower quality and resolution
- 8 synthetic video clips recorded using a drone simulator of the Unreal4 Game Engine.

C. CARPK

The CARPK is according to authors [11], the first and largest database of drone recordings that supports object counting, and provides the bounding box annotations. More specifically, there are 89 777 tagged cars on the dataset. The cars were shot by drones in four different parking lots. The dataset is tailored to deep learning algorithms for object count and localization scenarios.



Figure 1. Examples of the scenes captured in a) Campus, b) CarPK

D. Okutama-action

The Okutama-action database contains 43 fully labeled drone video clips for training and testing models when detecting multiple simultaneous actions within different categories (reading, handling, carrying different items), [11].

The videos were recorded using two DJI Phantom 4 drones in 4K baseball court and at 30 fps at 10m to 45m height, while the camera angle is 45 or 90 degrees. The dataset for each video contains metadata such as camera angle, speed, and height. The shots were taken with two different lighting conditions (sunny and cloudy)

E. UAVDT

The UAVDT consists of 100 drone videos in multiple urban locations such as streets, squares, intersections, etc [11]. The videos were shot at 1080 x 540 px with 30 fps in different weather conditions (day, night, fog), and in three different altitude ranges (low: 10m to 30m, medium: 30m to 70m high: more than 70m) and different camera views (front view, side view, and bird's eye view). An example of the scene captured in the UAVDT dataset is given in

Figure 2.a.

F. VisDrone

VisDrone is a set of data shot in different scenes focusing on four basic problems in the field of computer vision (object detection in images, object detection in videos, single object tracking and multiple object tracking).

The dataset consists of 263 video clips and an additional 10.209 images [11]. Videos/images were recorded on different drone platforms (DJI Mavic, DJI Phantom Series 3, 3A, 3SE, 3P, 4, 4A, 4P) in 14 different cities in China. The dataset covers different weather and light conditions of maximum video resolution (3840 x 2160 px) and images (2000 x 1500 px).



Figure 2. Examples of the scenes captured in a) UAVDT, b) VisDrone

Each of the datasets presented here is important for the development of computer vision research in the field of UAV images. However, it is clear from the descriptions of each image database and examples that they are intended for a specific task and tailored to a particular problem. For a specific problem, such as searching and rescuing people, there are missing appropriate scenarios where people in non-standard poses appear (e.g. injured persons during a fall), so they need to be recorded and included in the set.

III. Computer Vision Tasks in Search and Rescue Operations

Detection of people in images and videos plays a significant role in various applications, but in this section, we focus on search and rescue applications using drone recordings.

The search and rescue problem can be divided into four application areas: in combat, on water, in urban and non-urban areas [16]. The use of drones in search and rescue operations has been discussed in [18][19][20] [17]. In this review, we will focus on non-urban areas and water areas.

In [20] image segmentation and contrast enhancement were applied and then convolution neural networks (multiple single shot detector SSD) for the detection of persons ranging from 5 to 50 px, on drone imagery. They also used a 3D game editor to generate synthetic search and rescue datasets.

The Inception model with the Support Vector Machine (SVM) classifier is used in [21] for detecting people trapped in an avalanche using drone imagery. In [23] the focus is on detecting humans on sea recorded using an unmanned aerial vehicle equipped with a multi-spectral camera. A modified MobileNet convolution neural network architecture is used for detection.

In [24], authors have developed a system for detecting people and action recognition on the Okutama-action dataset while calculating GPS locations. For object detection, a model that was upgraded to MobileNetv2 and called POINet was used. Another example of the use of GPS signals in search and rescue actions is given in [25]. The assumption is that the injured person has a mobile device switched on, so a GSM radio signal of the mobile device is used to log the position of the injured person from the strength of the single and GPS position of the drone.

A platform for the detection of persons in water with the Tiny YOLO V3 architecture integrated on the NVIDIA Jetson X1 computer was introduced at [26]. The model was trained on the COCO dataset and swimmer's custom dataset recorded with a drone equipped with a GoPro camera in HD resolution. For the detection of sea surface objects, the use of a drone thermal camera and a real-time onboard algorithm was proposed in [27] to detect and track objects on the ocean surface.

The strategy of using semi-supervised and supervised machine learning approaches for the classification of aerial imagery and object detection along with the suggestion of hardware

and software architecture for the UAV platform is given in [28]. An algorithm for planning a search path for a UAV and using unmanned ground vehicles (UVG) to verify the identity of an object detected by the UAV is given in [29]. In [30] the authors classify drone imagery on human and non-human images and provide classification results using several CNN architectures. According to [31], it was the first paper that applies multiple object visual tracking to aerial imagery for search and rescue purposes, invariant to scale, translation and rotation, and with the ability to re-identify persons. Person detection is based on color and depth data and the use of the Human Shape Validation Filter that uses the locations of human joints obtained from the Convolutional Pose Machine [32]. The purpose of the filter is to study the shape of the human skeleton on detections to avoid false detections.

According to the results of the Vision Meets Drones competition, VisDrone 2019 [33], the Cascade R-CNN [34] model and models derived from it are most commonly used to detect objects such as pedestrians, cars and bicycles in large-scale benchmark dataset covering a wide range of aspects including location (taken from 14 different cities), environment (urban and country), objects (pedestrian, vehicles, bicycles, etc.), and density (sparse and crowded scenes).

Cascade R-CNN is a multi-stage object detector framework, which aims to increase the quality of detection by constantly increasing the intersection over union (IoU) thresholds [35]. Cascade R-CNN was used in different applications including agricultural, aerial photography, fast delivery, and surveillance, followed by CenterNet [36] and RetinaNet [37].

CenterNet is a one-stage highly efficient detector for exploring the visual patterns within each bounding box. For detecting an object, this approach uses a triplet, rather than a pair, of keypoints. Paying attention to the center information, RetinaNet has a feature pyramid network (FPN) [37] attached to its backbone to generate multi-scale pyramid features. Then, pyramid features go into classification and regression branches, whose weights can be shared across different levels of the FPN. The focal loss is applied to compensate for the accuracy drop, which improves performance. The most used detectors in the VisDrone competition, as the backbone mainly use ResNet-101, ResNet 101, ResNet 50, and SEResNeXt50.

The performance that object detectors achieve on images captured with a drone is much lower than that achieved on images that are not a bird's eye view in different application domains [38, 40, 40]. The top three detectors in the VisDrone 2019 competition (DPNet-ensemble, RRNet [41], and ACM-OD) in the image detection category reach an average precision (AP) of about 29% with an IoU> 50%. For person detection a maximum of 16% AP is achieved (BetterFPN, 16.45% AP, DPNet ensemble [35], 15.97% AP, ACM-OD [35], 15.50% AP).

Slightly lower object detection results than in the case of images were achieved in the object detection category on video [44]. The three best results were achieved by the following algorithms: DBAI-Det with 29.22% AP, AF-SRNet with 24.27% AP, and HRDet+ with 23.03% AP. As in the case with the images, positive detection was counted if IoU is greater than or equal to 50%. In the case of detection of persons, the best results were achieved with DBAI-Det, VCL-CRCNN, and AFSRNet, with pedestrian detection results being different from the detection of a person in general.

The success of the best algorithms is attributed to the combination of many recently proposed powerful networks, including DCNv2 [45], FPN, and Cascade R-CNN, and detection performance is significantly enhanced by the benefits of anchor-based RetinaNet and anchorless FSAF.

IV. CONCLUSION

The paper provides an overview of the object detection on imagery recorded by unmanned aerial vehicles (drones). The first part of the paper gives an overview of the current state of publicly available datasets with their characteristics and appropriate tasks. The following section shows the research activities using drones in search and rescue operations and computer vision methods for missing person detection. Finally, the models that currently show the best detection results on images made by unmanned systems are listed.

In future work, it is necessary to create a dataset of drone recordings for better detection of injured persons. Such a dataset would contain people in atypical poses that are not contained in existing datasets. Combining knowledge transfer

from existing datasets and the new custom set, it is necessary to test the state of the art models and analyze their performance in a new set of scenes characteristic for search and rescue operations. If necessary, adaptation and enhancement of existing models will be proposed to achieve the best possible detection results for disabled and missing persons in non-urban and off-water areas.

ACKNOWLEDGMENTS

This research was supported by Croatian Science Foundation under the project IP-2016-06-8345 "Automatic recognition of actions and activities in multimedia content from the sports domain" (RAASS) and the project IP-2018-01-7619 "A Knowledge-based Approach to Crowd Analysis in Video Surveillance (KACAVIS) and by the University of Rijeka (project number 18-222).

REFERENCES

[1] S. Sambolek, M. Ivasic-Kos, Detection of Toy Soldiers Taken from a Bird's Perspective Using Convolutional Neural Networks, ICT Innovations 2019, Ohrid. Springer Communications in Computer and Information Science

[2] M. Kristo, M. Ivašić-Kos, Thermal Imaging Dataset for Person Detection; Proceedings of 42nd International ICT Convention – MIPRO 2019, Opatija, Hrvatska: Mipro, 2019. str. 1316-1321

[3] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91-99.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A. C. Berg, "SSD: Single shot multi-box detector," in European conference on computer vision, Springer, Cham, 2016, pp. 21-37.

[5] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-788.

[6] T. Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980-2988.

[7] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, "The pascal visual object classes challenge: A retrospective," International journal of computer vision, 2015, 111(1), 98-136.

[8] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In ECCV, 2014.

[9] A. Robicquet, A. Sadeghian, A. Alahi, S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in European conference on computer vision, Springer, Cham, 2016, pp. 549-565.

[10] M. Mueller, N. Smith, B. Ghanem, "A benchmark and simulator for UAV tracking," in European conference on computer vision, Springer, Cham, 2016, pp. 445-461.

[11] M. R. Hsieh, Y. L. Lin, W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4145-4153.

[12] M. Barekatain, M. Martí, H. F. Shih, S. Murray, K. Nakayama, Y. Matsuo, H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 28-35.

[13] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 370-386.

[14] P. Zhu, L. Wen, X. Bian, H. Ling, Q. Hu, "Vision meets drones: A challenge," arXiv preprint arXiv:1804.07437, 2018.

[15] R. J. Koester, Lost Person Behavior: A Search and Rescue. DBS Productions LLC, 2008.

[16] S. N. A. M. Ghazali, H. A. Anuar, S. N. A. S. Zakaria, Z. Yusoff, "Determining position of target subjects in maritime search and rescue (MSAR) operations using rotary-wing unmanned aerial vehicles (UAVs)," in 2016 International Conference on Information and Communication Technology (ICICTM), IEEE, 2016, pp. 1-4.

[17] P. Doherty, P. Rudol, "A UAV search and rescue scenario with human body detection and geolocalization," in Australasian Joint Conference on Artificial Intelligence, Springer, Berlin, Heidelberg, 2007, pp. 1-13.

[18] M. A. Goodrich, B. S. Morse, C. Engh, J. L. Cooper, J. A. Adams, "Towards using unmanned aerial vehicles (UAVs) in wilderness search and rescue: Lessons from field trials," Interaction Studies, 2009, 10(3), 453-478.

[19] S. Waharte, N. Trigoni, "Supporting search and rescue operations with UAVs," in 2010 International Conference on Emerging Security Technologies, IEEE, 2010, pp. 142-147.

[20] C. A. Baker, S. Ramchurn, W. T. Teacy, N. R. Jennings, "Planning search and rescue missions for UAV teams," in Proceedings of the Twenty-second European Conference on Artificial Intelligence, IOS Press, 2016, pp. 1777-1778.

[21] K. Yun, L. Nguyen, T. Nguyen, D. Kim, S. Eldin, A. Huyen, E. Chow, "Small target detection for search and rescue operations using distributed deep learning and synthetic data generation," in Pattern Recognition and Tracking XXX (Vol. 10995, p. 1099507), International Society for Optics and Photonics, 2019.

[22] M. Bejiga, A. Zeggada, A. Nouffidj, F. Melgani, "A convolutional neural network approach for assisting avalanche search and rescue operations with UAV imagery," Remote Sensing, 2017, 9(2), 100.

[23] A. J. Gallego, A. Pertusa, P. Gil, R. B. Fisher, "Detection of bodies in maritime rescue operations using unmanned aerial vehicles with multispectral cameras," Journal of Field Robotics, 2019, 36(4), 782-796.

[24] R. Geraldes, A. Gonçalves, T. Lai, M. Villerabel, W. Deng, A. Salta, H. Prendinger, "UAV-based situational awareness system using deep learning," IEEE Access, 2019, 7, 122583-122594.

[25] S. O. Murphy, C. Sreenan, K. N. Brown, "Autonomous unmanned aerial vehicle for search and rescue using software-defined radio," in 2019 IEEE 89th Vehicular Technology Conference VTC2019-Spring, 2019, pp. 1-6. IEEE.

[26] E. Lygouras, N. Santavas, A. Taitzoglou, K. Tarchanidis, A. Mitropoulos, A. Gasteratos, "Unsupervised human detection with an embedded vision system on a fully autonomous UAV for search and rescue operations," Sensors, 2019, 19(16), 3542.

[27] F. S. Leira, T. A. Johansen, T. I. Fossen, "Automatic detection, classification and tracking of objects in the ocean surface from UAVs using a thermal camera," in 2015 IEEE aerospace conference, IEEE, 2015, pp. 1-10.

[28] J. Sun, B. Li, Y. Jiang, C. Y. Wen, "A camera-based target detection and positioning UAV system for search and rescue (SAR) purposes," Sensors, 2016, 16(11), 1778.

[29] Z. Kashino, G. Nejat, B. Benhabib, "Aerial wilderness search and rescue with ground support," Journal of Intelligent & Robotic Systems, 2019, 1-17.

[30] T. Marasović, V. Papić, "Person classification from aerial imagery using local convolutional neural network features," International Journal of Remote Sensing, 2019, 1-19.

[31] A. Al-Kaff, M. J. Gómez-Silva, F. M. Moreno, A. de la Escalera, J. M. Armingol, "An appearance-based tracking algorithm for aerial search and rescue purposes," Sensors, 2019, 19(3), 652.

[32] S. E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, "Convolutional pose machines," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724-4732.

[33] D. R. Pailla, "VisDrone-DET2019: the vision meets drone object detection in image challenge results, 2019.

[34] Z. Cai, N. Vasconcelos, "Cascade R-CNN: Delving into high-quality object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154-6162.

[35] M. Burić, M. Pobar, M. Ivasic-Kos, Adapting YOLO Network for Ball and Player Detection; Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM, Prag, Češka: SciTePress, 2019. str. 845-851

[36] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, "Centernet: Keypoint triplets for object detection," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6569-6578.

[37] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, "Feature pyramid networks for object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117-2125.

[38] M. Ivasic-Kos, M. Krišto, M. Pobar, Person Detection in Thermal Videos Using YOLO; Proceedings of SAI Intelligent Systems Conference IntelliSys 2019: Intelligent Systems and Applications, Cham: Springer, 2019. str. 254-267

[39] M. Pobar, M. Ivasic-Kos, Detection of the leading player in handball scenes using Mask R-CNN and STIPS, Proc. SPIE 11041, Eleventh International Conference on Machine Vision (ICMV 2018), Muenchen: SPIE, 2018

[40] M. Ivasic-Kos, M. Pobar; Building a labeled dataset for recognition of handball actions using Mask R-CNN and STIPS, 7th European Workshop on Visual Information Processing EUVIP, Tampere, Finska: IEEE, 2018. str. 1-6

[41] C. Chen, Y. Zhang, Q. Lv, S. Wei, X. Wang, X. Sun, J. Dong, "RRNet: A hybrid detector for object detection in drone-captured images," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.

[42] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, Z. Zhang, "MMDetection: Open MMLab detection toolbox and benchmark, 2019, arXiv preprint arXiv:1906.07155.

[43] B. Singh, M. Najibi, L. S. Davis, "SNIPER: Efficient multi-scale training," in Advances in Neural Information Processing Systems, 2018, pp. 9310-9320.

[44] P. Zhu, D. Du, L. Wen, X. Bian, H. Ling, Q. Hu, T. Peng…, "VisDrone-VID2019: "The vision meets drone object detection in video challenge results," 2019.

[45] X. Zhu, H. Hu, S. Lin, J. Dai, "Deformable convnets v2: More deformable, better results," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9308-9316.

Individualization of Anonymous Identities Using Artificial Intelligence (AI)

T. Bronzin*, B. Prole*, A. Stipić* and K. Pap**
*CITUS, Zagreb, Croatia
**University of Zagreb/Faculty of Graphic Arts, Zagreb, Croatia
tbronzin@citus.hr, brigita@citus.hr, arian@citus.hr, kpap@grf.hr

***Abstract* - Individualization of anonymous identities using artificial intelligence - enables innovative human-computer interaction through the personalization of communication which is, at the same time, individual and anonymous. This paper presents possible approach for individualization of anonymous identities in real time. It uses computer vision and artificial intelligence to automatically detect and recognize person's age group, gender, human body measures, proportions and other specific personal characteristics. Collected data constitutes the so-called person's biometric footprint and are linked to a unique (but still anonymous) identity that is recorded in the computer system, along with other information that make up the profile of the person. Identity anonymization can be achieved by appropriate asymmetric encryption of the biometric footprint (with no additional personal information being stored) and integrity can be ensured using blockchain technology. Data collected in this manner is GDPR compliant.**

Keywords – artificial intelligence (AI), individualization, communication, biometric identity, biometric footprint, anonymous identity, blockchain, computer vision

I. PERSONAL IDENTIFICATION

A. Biometric Identity

Humans can be identified not only by their personal documents, but also by some personal characteristics that can be recorded and kept in various digital formats. And while documents can be falsified, digitally recorded data can be made much more resistant to tampering and, for the most part, can be valid for long periods of time.

There are just a few examples of such personal characteristics that can be recorded and later used to verify the identity of a person with reasonably high level of precision and confidence. Today, we can include biometric data about the fingerprints, an iris and a face in this limited set. An important point here is that all of these characteristics are being used for identification. This implies that there has to be a process in place following these basic steps:

1. Positively identify a person using some reliable official documentation
2. Record one or more personal characteristic and save it in a digital format
3. Link digital information to other reference personal documentation

This is a standard way to keep a record of all people in a country (for example) and usually only fingerprints and faces are being recorded although irises are already widely used in passports today. Once you have this information stored, it is relatively easy to get a proper identification of a person. All you need is a fingerprint (not necessarily a complete one) or a good face shot and matching those to a person in a database is a straightforward process. Obviously, we can argue that this digitized information represents a **biometric identity** of a person.

An important fact related to biometric identity is that it is highly (very close to 100%) reliable and we can clearly understand when an available fingerprint or a photo of a person's face, iris or even an ear lobe (to some extent) are good enough to establish unambiguous proof of identification.

B. Personal Data Protection and GDPR

Biometric data is clearly the most personal and private data that anyone may expose, and most of them can be collected without person's awareness, knowledge or approval. Therefore, this entire area requires the implementation of specially designed rules to protect that data from any kind of wrong handling.

The process of collecting personal biometric data may not always be so straightforward because there are some other important considerations to think about. First and foremost, there is a privacy issue and it is regulated by many laws and acts, with EU GDPR act being the most widely known one. There are at least two important issues here: collecting information and permanently saving information that already is or can be linked to a specific person.

Collecting personal information, especially biometric ones, is not always allowed and even when it is, there are rules about the measurement and recording process that have to be followed. This may become particularly important if including brain signals patterns [1][2][3] in biometric identity gains wide popularity in practical applications.

Image on Figure 1 (taken from [2]) shows an example of high-level view of a possible general EEGBS (EEG-based Biometric System) framework that uses human brain signals to identify (biometric identity) or just verify

(biometric footprint) a person based on stored biometric data.

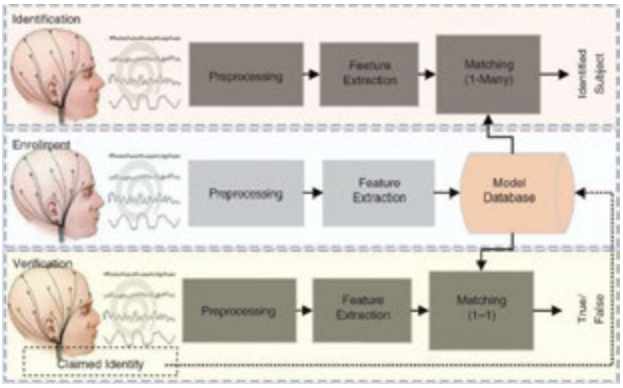


Figure 1. General EEGBS framework

Saving such information is particularly critical because, except in very specific purposes (registrations in official national police databases, for example) this can create explicit link between biometric data and a specific person which is, generally speaking, not allowed without imposing strong restrictions. For example, storing personal fingerprints in a police database is allowed, but storing the same fingerprints (or even asking to do so) in company's HR records are forbidden unless there are some exceptional but valid reasons to do so. Consequently, the same logic applies for all biometric identity data.

By the same token, biometric footprint (see below) data can, in many situations, also be considered as sensitive private information.

C. Anonymous Identities

There are also some other, non-conventional but highly interesting elements that could be used to define a **biometric footprint** of a person. Biometric footprint is not identical to biometric identity because it may not necessarily lead to a reliable identification of a person. The differences in reliability between biometric identity and biometric footprint may be too large to send a person to a court and possibly to a prison, but it could still be very useful in many other practical applications. Therefore, it is really all about the level of confidence. Whenever we have near 100% confidence, it is a biometric identity. Otherwise, it is biometric footprint and we will use it in cases where much lower (but still significant) levels of confidence are acceptable.

Biometrics is the science of analyzing physiological, behavioral, or genetic characteristics specific to each individual in order to be able to authenticate their identity [4][5]. Physiological biometrics records and/or measures inherent physiological characteristics of a person (fingerprints, iris, face, height, weight, arm and leg lengths etc.) that could be useful for determining true or **anonymous identities**. More details about specific measures can be found later in this paper.

This way, when used properly, the biometric footprint would not violate any personal data protection act, but it could still be used to identify a person, or to be more precise, identify an **anonymous person**. Although it may look at first like a kind of an oxymoron, it is really not. To understand this, we need to understand the concept of confidence.

Establishing a person's true identity using biometric identity data means comparing some externally available biometric data (fingerprints or face images, for example) and comparing them to previously created reference database that contains a mix of reliable standard and digital information about that person. When a match is found, a person is identified. This process is never 100% accurate, but it is still accurate enough to be widely accepted in highly sensitive areas such as security and criminology.

Biometric footprint may contain not only physiological data about the person, but also behavioral data. We can measure, record and permanently save derived information like voice, walk, body posture, handwriting etc. Not all of them is equally reliable but, as we shall explain in this paper, they can still be good enough to have some useful practical applications.

With biometric footprints we do not have any reference database to compare with, at least not the one with actual identities. But there is nothing to prevent us from keeping biometric footprints in a separate database and compare a biometric footprint recorded today with all other biometric footprints recorded a year ago (or any other time period), for example. This type of comparison will not tell us who the person is, but it may tell us (with a reasonable high confidence) that is was (or was not) the same person.

It is also important increase data integrity as much and possible one of the most promising potential candidates today is Blockchain technology because it can significantly reduce risks of serious damages caused by fraudulent data and software. Any unauthorized and uncontrolled tampering with personal biometric data can be extremely dangerous. However, if you decided to use Blockchain technology, it is critical to "… never store personal identification information or any other sensitive data on the blockchain; instead, consider storing a hash of encrypted data on the chain and the encrypted data either in a co-located, off-chain database or an existing source system" [15].

Recording these alternative human body characteristics can be done even without knowing who the person is being recorded. In fact, this is currently the predominant way of doing it. The rationale behind this idea is to have a reasonably reliable (with 80% of confidence, for example) information about individuals, but still not know who that individual really is. This kind of biometric footprint could be considered to be an anonymous identity. Confidence levels of anonymous identities can be further increased if we combine several (as many as possible) measurements together.

Obviously, this type of comparison is not as strong and as reliable as actual identification, but it has one huge advantage: it does not violate any personal identification laws. Therefore, it can be used for many practical purposes and some of them are described later in this paper.

D. De-Identification and Anonymization Process

Sometimes, if the existing biometric data is directly linked to identifiable person, additional "cleaning" of such

data may be required. There are two approaches to this issue:

- De-identification, and
- Anonymization

De-identification is defined as "…the process of removing or concealing personal identifiable information (shorter: personal identifiers), or replacing them with surrogate personal identifiers, in order to prevent the recognition (identification) of a person directly or indirectly, e.g. via association with an identifier, user agent or device" [6].

De-identification is **reversible** (two-directional) process of removing or obscuring any personally identifiable information from individual records in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them. It involves the provision of additional information (e.g. secrete key) to enable the extraction of the original identifiers in certain situations by an authorized body or by a trusted party [6].

Anonymization is slightly different as it is **irreversible** (uni-directional) process of deidentification that does not allow the original personal identifiers to be obtained from de-identified ones. Anonymized personal identifiers cannot be linked back to an original as they do not include the required translation variables to do so [6].

De-identification and/or anonymization of personal biometric data are extremely important. Such data could be easily misinterpreted (various profiling practices, for example) and cause direct damages to a person. Even worse, such data could be used for malicious purposes like revealing highly sensitive personal information about personal mental and health status or illegal tracking, surveillance and social control.

II. Establishing Biometric Footprint

This paper focused on physiological and behavioral biometrics that helps creating biometric footprint.

A. *Physiological Biometrics*

There are two ways
Typical physiological biometrics are:

- Fingerprint
- Finger vein patterns
- Hand geometry
- The eye
- Face shape
- Human body shape and measures

Fingerprint has the unique pattern that is very hard to copy/fake, but it is possible. Establishing biometric footprint fingerprint requires contact with the sensor/device.

Finger vein recognition looks for unique patterns of veins beneath the skin that is almost impossible to fake and usually is combined with fingerprint but, to establishing biometric footprint, fingers need to have contact with sensor/device.

Hand/fist geometry relates to the dimensions of fist, position and size of fingers, relation between fingers and requires contact with sensor/device. Arm geometry, on the other hand, can be determined with contactless approach with sensors like 3D (depth) camera and it will be discussed letter in this paper.

The eye, specifically retina and iris can also be used for establishing biometric footprint. Technology for retina/iris recognition is contactless, but it requests long period of time for acquiring the data (5-15 seconds) [7].

Facial recognition relates to specific face features, such as the shape of the face, position of eyes, nose and mouth, and distances between these features. This type of physiological biometrics uses contactless technology to establish biometric footprint. One of disadvantage of this recognition approach is that requires that face is clearly visible.

This paper proposes additional way of establishing biometric footprint by using human body shape and body measures by using contactless technology, explained later.

B. *Digital Human Body Measurements*

Digital human body measurements uses computer vision to establish, without any contact (just by "looking at"), reference points that represents human body, they are similar, but different than reference points called anthropometric points used by "classic" (manual) method defined by International Biological Programme (IBP) [8]. For that purpose, newly proposed standard protocol for digital measurement (DM-I) [9] has been defined for assessing anthropometric dimensions of human body that can be used as biometrics for identifying people. Figure 2 shows reference points for digital body measurements on human skeleton used in DM-I.

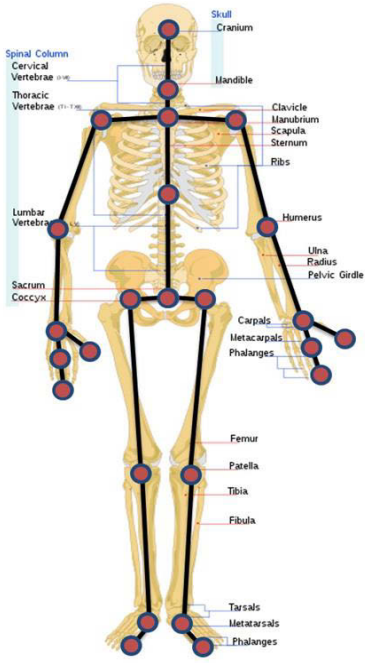


Figure 2. Reference points on human skeleton

Twenty-five reference points shown in Table I. are used as endpoints for human body measurements.

3D camera is used to calculate distance from the subject (in just one second) and locate reference points that are used for measurement calculations.

TABLE I. LIST OF REFERENCE POINTS FOR DIGITAL MEASUREMENTS OF HUMAN BODY

No	Reference point name/description
1	Left ankle
2	Right ankle
3	Left elbow
4	Right elbow
5	Left foot
6	Right foot
7	Left hand
8	Right hand
9	Tip of the left hand
10	Tip of the right hand
11	Head
12	Left hip
13	Right hip
14	Left knee
15	Right knee
16	Neck
17	Left shoulder
18	Right shoulder
19	Base of the spine
20	Middle of the spine
21	Spine at the shoulder
22	Left thumb
23	Right thumb
24	Left wrist
25	Right wrist

Physiologic biometrics of the person like height, forearm length, torso height, waist circumference, torso circumferences, shoulder circumferences, etc. [13] has been calculated by using those reference points as shown in Figure 3.



Figure 3. Example of digital body measurement tool that use DM-I

C. Reliability of digital body measurement

Recent research and studies [9][13] that used Microsoft Kinect 3D camera/sensor, bring conclusion that digital measurements with it are not appropriate for clinical trials demanding extremely high precision, but it's good enough to be used in most applications for sports and fashion. For example, reliability of Kinect sensor is excellent for measuring person's height and acceptable for left forearm length and left lower leg length. Small errors, from 3.31%-6.68% as stated in [9], occur due to clothing, illumination, sensor height and distance. This is in line with previous research [10] where results showed that the 3D scanning is highly precise for most cases.

D. Digital body measurement and anonymous identities

Previously, we have established that digital body measurement performed using today's high-resolution stereoscopic video cameras can provide a lot of additional, even unconventional data about human body's physical and behavioral characteristics. If we skip digital identities for the moment and restrict ourselves only to digital footprints, we will find a lot of specific measures that can be used to build anonymous identities and still conform to GDPR and similar regulations.

Since we only know anonymous identities and not true person's identities, we can start the process of tracking and comparing those anonymous identities without violating any personal rights. The process is simple and consist of these steps:

- Record selected (always the same) set of physiological and behavioral information, avoiding those that could directly lead to a person's identification (digital identity)
- Save the recorded information (or even better, corresponding encrypted values) in some kind of a database system for future comparison
- Continue recording additional sets of physiological and behavioral information
- Compare the newly recorded sets with those stored in a database and see if some of them can be found
- When you spot second, third (and so on) occurrence of the same digital footprint, assume that it is the same person and proceed with the appropriate actions

This simple recipe opens the path to using anonymous identities in various practical applications.

E. Human Behaviour

Typical behavior biometrics are:

- Voice
- Signature/handwriting dynamics
- Keystrokes
- Gait
- Body movement/gestures
- Posture

Voice, signature/handwriting dynamics and keystrokes are human behaviors that are used to identify person by comparing voice/signature/keystroke patterns. All these methods require that person has contact with sensors and are prone to the error due to biological reasons or light condition when sensors are gathering data. For hygienic

reasons, this may not be acceptable in many cases (like COVID-19 outbreak).

Gait, body movement/gestures and posture can be detected with computer vision, by using the same type of 3D camera/sensors like Microsoft Kinect. This process does not require any direct person to touch any sensor and can be combined with (digitally acquired) body measures to form **biometric footprint.**

III. PRACTICAL APPLICABILITY OF BIOMETRIC FOOTPRINT

A. Marketing applications

Marketing is one possible application for biometric footprint and anonymous identities.

Both privacy concerns (related to legal requirements like GDPR/ePrivacy and other) and health requirements (that requires non-touch technology) resulted in research and development of the system that allows individualization of anonymous identities using artificial intelligence (AI). Proposed solution uses **biometric footprint** constructed from persons race, age, gender, facial expressions and digital body measurement, persons gait signature, specific body movement/gestures recording and specific posture. It is engineered in a way that even in low-light conditions, algorithms can still provide high-quality, high-accuracy results.

One of the most common scenarios is implementing loyalty program that needs to comply to privacy regulation (GDPR or other). An example of a loyalty program is related to the popular soft drinks offered in shops and displayed in retail refrigerators with the glass door. Refrigerator glass door has been replaced with semi-transparent display and 3D camera/sensor Kinect is placed above the door, as shown in Figure 4.

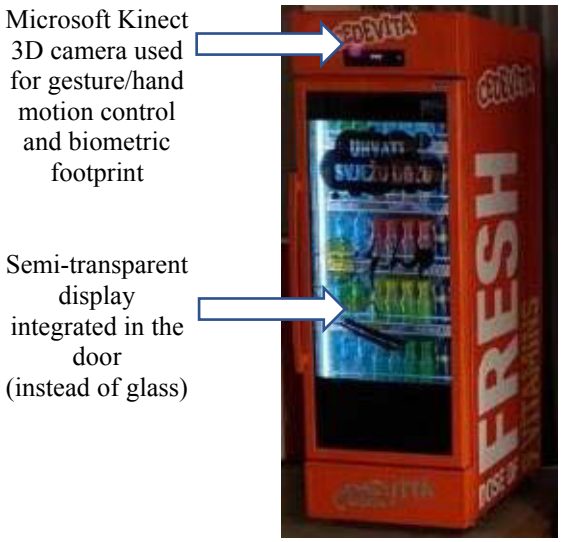


Figure 4. Smart retail refrigerator with loyalty program and gamification

The scenario goes like this: when a person approaches the refrigerator for the first time, her/his biometric footprint is taken, anonymized and recorded, after the person give her/his consent. Before even the consent is shown to the person, artificial intelligence embedded in refrigerator is analyzing the age group of the person and if it is a minor, younger than 13 years (or an equivalent number of legal years for the specific country of installation) than no consent or recording is taking place. For person older than required minimum age, initial biometric footprint is constructed, like a puzzle, from following attributes: age group, race, gender, digital body measures (starting from face to whole body figure). During the interaction between person and refrigerator the other part of the biometric footprint is constructed by analyzing persons gait, body movement/gestures and posture.

Based on the anonymous, but unique identity of the person, AI embedded in the refrigerator counts how many bottles are picked-up by that person, scanned by 3D camera, identified by computer vision and bought in predefined time (this information comes from the integration with cash registers). Because all refrigerators are connected to the central cloud system, person is recognized in any shop in the world where refrigerators are placed and enjoys the same loyalty program. For example, when person picks and buys the 5th bottle, the 6th one is free of charge.

B. Security applications

Security is the obvious suspect when it comes to practical applications of biometric footprint data and various industries have already implemented different types of solutions in this area. Solid portion of them are trying to improve the classical problem of authenticating a person by utilizing some of the biometric information. This is commonly referred to as an **adaptive authentication**.

In banking and financial business in general, for example, improving the security during the authentication process is seen as being critical by most of the users. Any flaws in this area may immediately turn many existing customers away from such faulty solution, with a good chance to lose them forever. Therefore, all Fintech industries are investing a lot of efforts and resources into this issue.

The idea behind adaptive authentication solutions is to combine different types of information in the authentication process:

- Information that a user claims to know (like username, password and PIN)
- Information that is physiologically or behaviorally (or both) inherent to the user (biometric data)

The first element is traditional but the second one reflects the effort to include biometric data and make it reliable enough to be used in these highly sensitive applications. Even more specifically, behavioral biometrics is often being used as an innovative approach to user authentication based on the creation of a unique profile for every single customer [14].

Specific behavioral measures used in these applications may come as a surprise to most people. Due to strict regulations (GDPR, among others) these solutions are collecting, analyzing and then applying information how do you press keys on a keyboard or on a mobile device, how you use your mouse and similar. This is done with a help of advanced AI models and the data is not directly associated with a person. Instead, it is locally collected, stored and analyzed but never explicitly linked to a person. At least in theory.

IV. CONCLUSION

Although it may look strange at first, the idea of having "anonymous identities" is really an interesting approach that is rapidly gaining momentum in some areas. This is not a replacement of standard "known identities" but it already has some very important practical applications today and the number will increase over time. Virtually all practical examples of collecting biometric identity (for standard identities) or biometric footprint (for anonymous identities) use artificial intelligence of some kind. Stereoscopic cameras (like Microsoft Kinect) and advanced image processing algorithms are very common in these scenarios.

Currently, the most popular areas for applications that use anonymous identities are security and marketing, but there are others as well.

These new implementations are getting increasing attention because they can also satisfy two very specific conditions:

- Compliance with GDPR (and similar) privacy protection regulation, and
- Ability to establish touchless communication between the computer (or any other piece of equipment) and the person being measured (or a person interacting with some computerized device)

The latter becomes very important in the current healthcare conditions when coronavirus disease (COVID-19) outbreak caused global pandemic in end of 2019/first half of 2020.

In spite of high reliability of recognition accuracy of fingerprints (99%) [11], palm print, hand/fist geometry [12], these technologies require contact-based approach which is not always possible to implement. This is particularly true today. Touchless measurements and user interface completely eliminate the need of any physical contact, but yet keep the functionality intact.

REFERENCES

[1] J.L. Camp, ""Digital identity" IEEE Technology and Society Magazine, vol. 23, no. 3, pp. 34-41, Fall 2004, DOI: https://doi.org/10.1109/MTAS.2004.1337889

[2] K.P. Thomas, A.P. Vinod, "Toward EEG-Based Biometric Systems: The Great Potential of Brain-Wave-Based Biometrics" in IEEE Systems, Man, and Cybernetics Magazine, vol. 3, no. 4, pp. 6-15, Oct. 2017, DOI: https://doi.org/10.1109/MSMC.2017.2703651

[3] G. Al-Hudhud, M.A. Alzamel, E. Alattas, A. Alwabil, "Using brain signals patterns for biometric identity verification systems", Computers in Human Behavior, Volume 31 /R. D. Tennyson, ScienceDirect, Pages 224-229, ISSN 0747-5632, DOI: https://doi.org/10.1016/j.chb.2013.09.018

[4] Group of authors, "Biometrics: authentication & identification (definition, trends, use cases, laws and latest news) - 2020 review" Gemalto, Seen online on 15th of May, 2020 at: https://www.gemalto.com/govt/inspired/biometrics

[5] F. Golshani, "Digital Biometrics." In: Furht B. (eds) Encyclopedia of Multimedia. Springer, Boston, MA, DOI:10.1007/978-0-387-78414-4, ISBN 978-0-387-74724-8, online: https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-78414-4_297

[6] S. Ribarić, N. Pavešić, „De-identification for privacy protection in biometrics" in User-Centric Privacy and Security in Biometrics, Vielhhauer, Claus, London: IET, pp. 293-315, 2017, online: https://bit.ly/de-indentifc_ribaric

[7] A. Hussain, A. Bhuiyan, and A. Mian, "Biometric security application for person authentication using retinal vessel feature," in Proceedings of the International Conference on Digital Image Computing: Techniques and Application (DICTA '13), Tasmania, Australia, pp. 1–8, November 2013, DOI: https://doi.org/10.1109/DICTA.2013.6691489

[8] M. Mišigoj-Duraković, "Kinanthropology: biological aspects of physical exercises", Faculty of Kinesiology, University of Zagreb, Zagreb, 2008

[9] I. Gruić, D. Katović, A. Bušić, T. Bronzin, V. Medved, M. Mišigoj-Duraković, "Construction and Validation of Protocol for Digital Measurement of Human Body", Communications in Computer and Information Science / Cabri, J. ; Pezarat-Correia, P. ; Vilas-Boas, J (eds.). Switzerland: Springer International Publishing, pp. 86-99 2019, DOI: https://doi.org/10.1007/978-3-030-14526-2_6

[10] S. Clarkson, J. Wheat, B. Heller, S. Choppin, "Assessing the suitability of the Microsoft Kinect for calculating person specific body segment parameter", In: 4th IEEE Workshop on Consumer Depth Cameras for Computer Vision, Zurich, Switzerland, Springer, Cham, 6 September, 2014, DOI: https://doi.org/10.1007/978-3-319-16178-5_26

[11] B.T. Ulery, R.A. Hicklin, J. Buscaglia, and M.A. Roberts, "Accuracy and reliability of forensic latent fingerprint decisions," Proceedings of the National Academy of Sciences of the United States of America, vol. 108, no. 19, pp. 7733–7738, 2011.

[12] N. Duta, "A survey of biometric technology based on hand shape," Pattern Recognition, vol. 42, no. 11, pp. 2797–2806, 2009.

[13] D. Katović, I. Gruić, A. Bušić, T. Bronzin, K. Pažin, F. Bolčević, V. Medved, M. Mišigoj-Duraković, "Development of Computer System for Digital Measurement of Human Body: Initial Findings", Proceedings of the 4th International Congress on Sport Sciences Research and Technology Support / Pedro Pezarat Correia and Jan Cabri (eds.). Porto: SCITEPRESS – Science and Technology Publications, pp. 147-153, 2016, DOI: https://doi.org/10.5220/0006086001470153

[14] Group of authors, "Behavioral biometrics and biometrics in payment cards: Beyond the PIN and password" Gemalto, Seen online on 15th of May, 2020 at: https://www.gemalto.com/financial/inspired/behavioral-biometrics

[15] W. Henry, B. Chew, A. Lora, H. Chae, "Developing winning blockchain applications in public sector" Deloitte Development LLC, Seen online on 15th of May, 2020 at: https://bit.ly/blokchain_deloitte

Towards the Open Ontology for IoT Ecosystem's Security

Igor Tomičić, Petra Grd
Artificial Intelligence Laboratory, Center for Forensics, Biometrics and Privacy
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42000 Varaždin, Croatia
Email: (igor.tomicic, petra.grd)@foi.hr

Abstract— **The IoT ecosystem is complex and includes hardware devices, services, software, connectivity, standards and protocols. IoT is continuously and persistently contributing to the ubiquity of technology in all aspects of human endeavours, and trends seem to indicate that myriad of IoT solutions are developed and/or bought for various purposes, from the personal home to industrial usages, with the alarming lack of security concerns, giving priority mostly to the fast and cost-effective development. Given the complex nature of such solutions, it is rarely seen that the security aspect is considered through all of the IoT ecosystem elements; also, one of the significant reasons being that the IoT solutions are not always developed in a holistic nature, within the same company, but only provide parts of the complete IoT solution. In order to enable easier and more secure development of IoT solutions, this paper presents first steps towards developing the open ontology for IoT ecosystem elements security, relating them with existing security concepts, primitives, threats, vulnerabilities and practices.**

Keywords— IoT; security; Internet of Things; Ontology

I. Introduction

The first mention of Internet of Things (IoT) was in 1999 when MIT's Kevin Ashton stated that "adding RFIDs to everyday objects would create an Internet of Things" [1]. Today, there are numerous definitions of IoT, the most common one being [2] that IoT is a heterogeneous field defined as "a variety of things or objects which, through unique addressing schemes, are able to interact with each other and cooperate with their neighbours to reach common goals". The IEEE IoT initiative [3] released a document with baseline definitions of IoT, where they separate definitions for small systems which contain uniquely identifiable things and small sensors, and complex systems that interconnect millions of things with a capacity to deliver complex services. IoT for low complexity systems is defined as [3] "a network that connects uniquely identifiable "things" to the Internet. The "things" have sensing/actuation and potential programmability capabilities. Through the exploitation of unique identification and sensing, information about the "thing" can be collected and the state of the "thing" can be changed from anywhere, anytime, by anything." IoT for complex systems is defined as [3] "a self-configuring, adaptive, complex network that interconnects "things" to the Internet through the use of standard communication protocols. The interconnected things have physical or virtual representation in the digital world, sensing/actuation capability, a programmability feature and are uniquely identifiable. The representation contains information including the thing's identity, status, location or any other business, social or privately relevant information. The things offer services, with or without human intervention, through the exploitation of unique identification, data capture and communication, and actuation capability. The service is exploited through the use of intelligent interfaces and is made available anywhere, anytime, and for anything taking security into consideration." The connected devices ("things") may include various kinds of sensors and actuators, mobile devices, car/vehicle computers, TV sets; but also non-ICT appliances like light bulbs, speakers, dishwashers, microwave ovens, refrigerators, electrical energy sources and building components [4], [5], equipped with computing, communication and sensing capabilities. The widespread and importance of current IoT technologies can be observed through its numerous application domains, which include, but are not limited to [5]: smart cities [6], smart power grids [7], smart health [8], smart transport [9], smart buildings [10], smart living solutions [11], smart human settlements [12], [13], [14], and other.

The rapid development of IoT technology and its application in different domains caused a large increase in IoT devices. According to [15] the total number of IoT connected devices is projected to amount to 75.44 billion worldwide by 2025 and global market to 1.6 trillion US dollars. This large amount of connected devices which are heterogeneous and most often poorly protected, pose a tremendous security risk. The success of IoT depends, not only on great applications, devices and analytics, but also on approach to security. When talking about IoT security, the main problem is the high interaction between humans, machines and robots and IoT technologies with constraints in terms of connectivity, computational power and energy[16], [17]. An enormous number of different devices, technologies, data consumers, agents and application scenarios makes it difficult to determine assets, threats, vulnerabilities,

TABLE I: Examples of IoT ecosystem cores [21]

Core	Hardware platform	Software platform	Standards
Connected device	Arduino, T-Mote Sky	TinyOS, Processing, Contiki OS	HGI
Connectivity	Wi-Fi or ZigBee systems-on-chip	Californium, Erbium	IPSO Alliance, ZigBee Alliance
Application services	Cloud infrastructure	Pachube	SOA, JSON, EPC
Supporting services	M2M optimized GGSN	NSN M2M suite, EDCP	ETSI M2M TC

risks and countermeasures so security can be addressed [18]. The unattainable goal is to secure every device, subnetwork, transport or core network, and each analytics, service or storage system. To address the aforementioned issues, ontologies are often used. Ontologies are defined as a "well-founded mechanism for the representation and exchange of structured information" [19], [20].

The idea behind this paper is to present first steps towards developing the open ontology for IoT ecosystem elements' security, relating them with existing security concepts, primitives, threats, vulnerabilities and practices. The ontology will try to shed the light on IoT vulnerabilities, their attack vectors, impacts on numerous security objectives and attacks which exploit such vulnerabilities. Chapter 2 gives an overview of related work on IoT ontologies and security ontologies in general, as well as analysis and comparison of IoT security ontologies developed up to date. In chapter 3, the steps towards the developing of the ontology are described. The scope, basic concepts and properties are identified and the proposed ontology is compared to existing ontologies.

II. Related work

According to [21] at this time IoT solutions range in a variety of proprietary platforms, protocols, and interfaces, making the components of solutions by different vendors barely compatible. Up to date, no single dominating set of standard protocols, interfaces and platforms has emerged. Nevertheless, a lot of progress has been made in determining the IoT ecosystem.

Mazhelis et al. [21] propose a view of IoT ecosystem that emerges around the core. They view the core as hardware, software or standards used for interconnection, more specifically: connected devices and gateways, connectivity between the devices and the Internet, application services and supporting services. Each of these include hardware, software and standards (eg. Table I).

Szilagyi and Wira [22] propose a basic architecture of the IoT (Figure 1) where they focus their attention on the "thing" aspect of the architecture. The "things" that form the IoT must do something useful in order to be considered for integration in an IoT system. They state that, when integrated into an IoT system, the "thing" can be characterized by three main capabilities: (1) communication; (2) programmability; and (3) sensing and/or actuating capabilities.

Due to the rapid development of IoT, in recent years a number of IoT ontologies has emerged. Hachem et al. [23] propose a new ontology for IoT. They identified four main

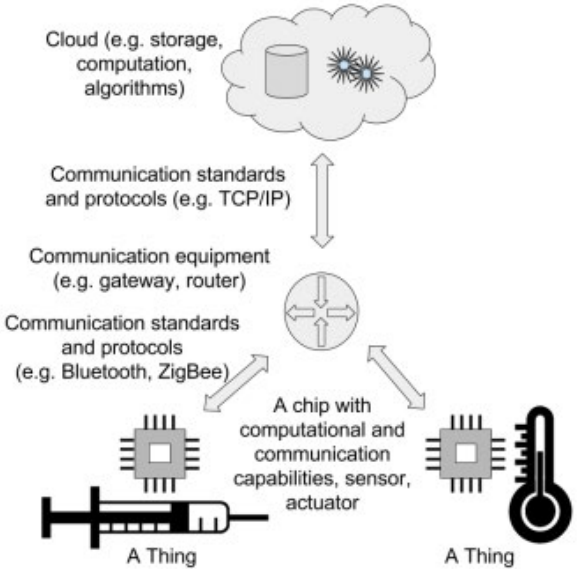


Fig. 1: Basic architecture of the IoT [22]

components: classes, relations, attributes and individuals. Their proposed global ontology models three aspects of the real world present in the IoT. The three parts of the ontology are device ontology, domain ontology and estimation ontology.

Kotis and Katasonov [24] present the creation of an IoT ontology that supports automated deployment process of applications in heterogeneous IoT environments. In addition to developing the new ontology, they gave an overview of existing ontologies in order to better position their work in the current state of the art.

Szilagy and Wira [22] reviewed the most popular ontologies for the IoT with emphasis on semantic technologies used in IoT, but they have not provided an in-depth analysis of selected ontologies.

Bajaj et al. [20] give an overview of the existing IoT ontologies based on the core concepts they identified (sensors, platform, testbed, service, location). They classified the reviewed ontologies into four classes: sensor ontologies, context-aware ontologies, location-based ontologies and time-based ontologies. Each of these classes is additionally divided into generic ontologies and domain-specific ontologies. Their conclusion is that no existing ontology is good enough and does not have all the required concepts.

A. Security ontologies

Before analysing and developing IoT security ontology, a review of existing security ontologies needs to be done. Different authors approach the problem of security ontologies from different aspects. Some of them developed general security ontologies, and some of the ontologies are domain specific. Authors have different categories in which they classify existing ontologies.

Souag et al. [25] performed a survey of security ontologies and classified them in eight families: beginning security ontologies [26], security taxonomies [27], [28], general security

ontologies [29], [30], specific security ontologies [31], [32], web oriented security ontologies [33], [34], [35], [36], risk based security ontologies [30], [27], [37], ontologies for security requirements [38], [39], [40] and security modelling ontologies [41], [42]. With their work they expanded on previous classifications

Mozzaquatro et al. [16] developed a reference ontology for the security in the IoT alongside which they analysed existing security ontologies. They categorized these ontologies into two main groups: general security ontologies [29], [30], [36], [35] and security ontologies applied to a specific domain [31], [43].

There have been other authors that analysed existing security ontologies [44], [45], [46], but most of them gave an overview of the same ontologies mentioned earlier.

B. IoT security ontologies

In order to better position our research, an overview of existing IoT security ontologies is given to provide context. Our analysis has shown that there are not many research papers on this topic.

One of the more comprehensive IoT security ontologies is the one by Mozzaquatro et al. [16]. They started the ontology development in their work "Towards a reference ontology for security in the Internet of things", where they proposed an IoT security ontology called IoTSec with security concepts of M2M communications to help find secure solutions to the IoT environment. They used systematic literature review and the MENTOR methodology in order to make an overview of existing security ontologies and knowledge of vulnerabilities and create a new IoT security ontology. They continued their work in the paper "An ontology-based cybersecurity framework for the Internet of things" [47] where they focused on improvement of their IoT ontology. They also proposed an ontology-based cybersecurity framework using knowledge reasoning for IoT, composed of two approaches: design time and run time.

Arnaert et al. [48] proposed an ontology aimed at obtaining more accurate results about vulnerable IoT devices. However, they focused only on scenarios when Censys and Shodan search engines are used.

Gonzalez-Gil and Martinez [18] proposed an ontology for context based evaluation of security for IoT devices called IoTSecEv. To that end, they used the IoTSec ontology developed by [16]. Their ontology has a relatively small number of devices, and lacks threats and vulnerabilities.

III. Towards the SecureIoT Ontology

The ontology proposed within this paper is being developed with OWL Web Ontology Language [49] and Protege software, an open source ontology editor and a knowledge management system. Being open, the choice of the software is easily argued, making further work and collaboration on the project not dependent on vendor-specific formats and interfaces. The ontology development methodology is partially relying on the METHONTOLOGY created by Fernandez-Lopez et al. [50], and on the the approach described by Noy et al.[51].

The elicitation of concepts and knowledge acquisition in the domain of IoT security was so far done through several sources:

- scientific papers on IoT and information security
- professional papers and whitepapers
- existing IoT systems
- existing IoT and security ontologies, and other types of knowledge representation within domain
- existing documentation and specification on specific IoT-related devices

Ontology consistency was tested with the Protege's in-built HermiT Reasoner. After the reasoner was initialized on the current version of the SecureIoT ontology, the reasoner showed that the ontology is formally consistent, showing no errors or exceptions.

A. The scope

The SecureIoT ontology proposed in this paper would be used in the domain of IoT systems; more precisely, in the design and development of secure IoT systems regardless of their usage domain - home, educational, health, infrastructure, traffic, environment, agriculture, industrial automation, logistics, or others. Every element of the IoT system would be viewed through security lenses in order to raise the awareness of security concepts and concerns, and ultimately to deploy more secure IoT systems. As stated by [25], "requirements engineering based on predefined ontologies can make the job of requirement engineering much easier and faster."

B. Basic concepts

The top concepts and their relationships, derived from the domain knowledge elicitation, are depicted in Figure 2, from which some are listed as follows:

- SecurityProperty. This ontology class relates to fundamental security concepts such as Confidentiality, Integrity, Availability, Authentication, and others.
- Location. For IoT systems, this concept directly relates to possible vulnerabilities and threats; mainly non-malicious ones.
- Communication relates to standards, protocols and hardware needed to transmit and receive data within the IoT ecosystem.
- Asset is either the physical entity within the IoT ecosystem such as sensor, actuator, mictrocontroller unit, electronic component and all its other tangible elements; or, it refers to server technologies, operating systems, networks, arhitectures.
- SecurityMechanism relates to all the mechanisms which can be deployed to mitigate threats.
- Severity is derived from the NVD v3.0 Vulnerability Severity Ratings[1], defining None, Low, Medium, High and Critical values.
- Vulnerability is a concept to be automatically populated by and synchronized with the individual entries from the

[1] nvd.nist.gov/vuln-metrics/cvss

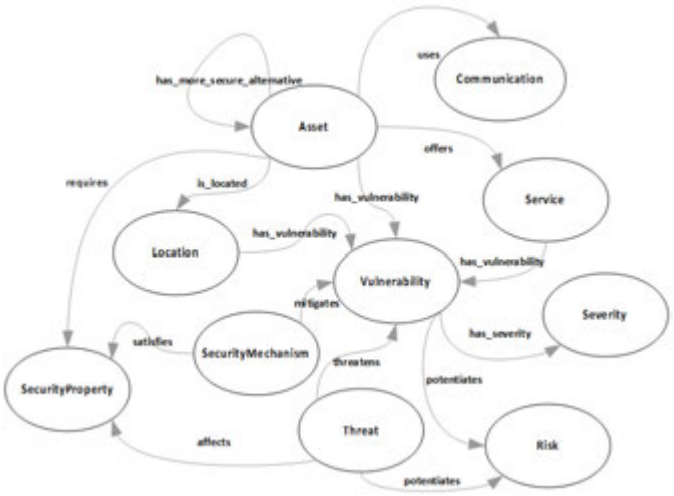


Fig. 2: Proposed top-level concepts of the SecureIoT ontology

NIST National Vulnerability Database, or the Common Vulnerabilities and Exposures (CVE).

- Threat classification was partly derived from [52], with some modifications on the concepts' hierarchy. Threats can be malicious (with human intent), and non-malicious, such as accidental, induced by extreme weather (environmental), or by some unforseen mix of chemicals (technical), etc.

The identified top-level concepts are partially derived from [16], and are depicted in Figure 2.

C. The re-use of existing ontologies and classifications

The "Severity" entity was derived from the NVD v3.0 Vulnerability Severity Ratings, defining None, Low, Medium, High and Critical values.

The "SecurityProperty" entity was derived from Alam et al. [53], where authors argue conventional security requirements for IoT: Confidentiality, Integrity, Availability, Authentication, Authorization, Access Control, Trustworthiness and Auditing. Naturally, additional security requirements for a specific IoT ecosystem may be required, but as authors note, these are the basic security properties that all IoT services should satisfy.

IoT Protocols were reused from [54], where the existing protocols were neatly organized into data protocols (such as XMPP, AMQP, MQTT, etc.), discovery protocols (such as mDNS, PhysicalWeb, etc.), insfrastructure protocols (such as 6LoWPAN, IPv6, UDP, etc.) and transport protocols (such as Bluetooth, Cellular, LoRaWAN, WiFi, WiMax, ZigBee, etc). This ontology, however, lacks general security concepts, as the existing ones are mainly taken from the OWASP IoT security guidelines[2].

Threat classification was partly derived from [52], with some modifications on the concepts' hierarchy. The existing ontologies seem to have overlooked the important aspect of this classification - the threat division on malicious and non-malicious aspects. With IoT systems' physical elements being distributed on various physical locations, susceptible to myriad of non-malicious threats (weather, chemical reactions, physical impacts, animals, accidents, floods, meteor impacts, etc.), this aspect seems crucial for the realistic settings in which IoT systems might be deployed.

The Reference Ontology for Security in the Internet of Things [16] has partially influenced this paper with the top-level concept organization and their relationships, being one of the more comprehensive ontologies in the field, but lacking some important elements which would be beneficial in engineering IoT ecosystems, such as Location, IoT Service, Physical Things, as presented and argued within the ontology proposed in this paper. From this reference ontology we would re-use Asset subclasses in the Server, Operating System, Networking and Architecture Technology categories, as they are compatible with the structure of the proposed ontology. In future research, subclasses of Vulnerabilities and Security mechanisms would also be considered for re-using, as they were elaborated in greater detail, but might need to be harmonized with existing structure.

D. Properties

The Protege software includes object properties, data properties and annotation properties. Object properties define relationships between two objects. Data properties relate individuals to concrete data, which can be of various types such as strings, integers, decimal, double, byte, rational, etc. The annotation properties provide descriptive metadata about ontology entities, do not participate in reasoning, and will be excluded from these descriptions. Some of the object properties defined in the SecureIoT ontology are the following:

- threatens_security_property. Connects Threat and SecurityProperty.
- has_connectivity. Describes the type of connectivity that a computational device can have.
- targets. Threat targets Vulnerability.
- is_mitigated_by. Threat is mitigated by a SecurityMechanism.
- is_located_in. Any physical IoT element is located in either some indoor or outdoor location.
- is_threatened_by. Some object can be threatened by a certain threat. For example, OutdoorLocation entity can be threatened by 0 or more instances of EnvironmentalThreat.
- is_compatible_with. Relates compatible technologies.
- is_used_by. Defines elements that are used by IoT subsystems.
- is_secured_by. Relates various entities with SecurityMechanism.
- has_severity. Relates vulnerabilities and NVD v3.0 Vulnerability Severity Ratings.
- can_use_actuator. Defines actuators that can be used by certain IoT devices.

[2]OWASP (2016) IoT Security Guidance. https://www.owasp.org/index.php/IoT_Security_Guidance

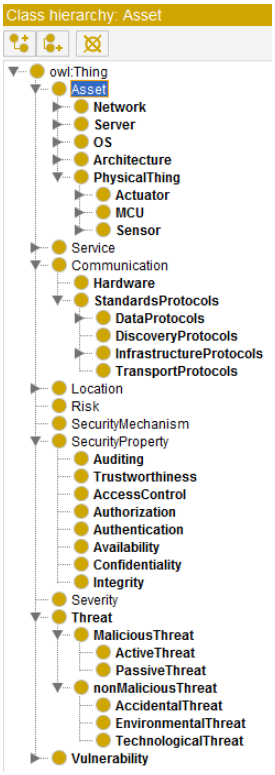


Fig. 3: A part of the SecureIoT ontology implemented in Protege

- can_use_sensor. Defines sensors that can be used by certain IoT devices.
- has_more_secure_alternative. Relates technologies and/or devices which can be used interchangeably; for example, in finding a more secure alternative for hardening the IoT system.
- has_vulnerability. Relates entities to both vulnerability classes and to specific CVE vulnerabilities.

The further ontology development is expected to derive more object properties, creating larger pool of ontology connections and expressiveness.

IV. Conclusion

In this paper we have proposed a SecureIoT ontology for the formal representation of the knowledge within the domain of IoT ecosystem's security. We have reviewed the existing body of research, pointed out the possible shortcomings and proposed a more comprehensive model, equally focusing on both IoT environment and security aspects, with the re-use of existing ontologies and harmonization of their elements within SecureIoT. The ontology, when fully developed, should be able to answer questions such as: *(1) List all the available IoT MCUs with Low or None severity ratings; (2) Which security mechanisms mitigate the violation of integrity of data?; (3) Which assets are protected by the strong encryption?; (4) Which security mechanisms thwart buffer overflow attacks?; (5) Which data protocols offer the least amount of known vulnerabilities?;* etc.

We have proposed a direct synchronization of the ontology Vulnerability class with the NIST National Vulnerability Database, or the Common Vulnerabilities and Exposures, which might prove usefull in engineering IoT systems from practical security perspective. For example, in searching for the optimal MCU, user can directly associate a concrete MCU model with current CVEs. The SecureIoT also presents an additional element which might directly relate to vulnerabilities of the IoT system - Location, with the physical devices being distributed in heterogenuous indoor and outdoor locations, being susceptible to various physical and chemical factors, weather, animals, accidents, and other non-malicious threats.

Further research will include the development of axioms as grounds for delivering rules and assumptions that enable the construction of the formal system, firstly stated in the informal way by using natural language, followed by the formalization in Protege by using Manchester OWL Syntax. Axioms would further define types of connections between concepts and cardinality constraints. Further work would also include fine-tuning of both the classes and their subclasses, object and data properties, harmonizations with other ontologies, and data population with individual instances.

References

[1] That 'internet of things' thing. [Online]. Available: https://www.rfidjournal.com/articles/view?4986

[2] A. I. L. Atzori and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, p. 2787–2805, 2010.

[3] D. R. Roberto Minerva, Abyi Biru, "Towards a definition of the internet of things (iot)," 2015.

[4] B. T. Z. Hu, S. Frénot and G. Privat, "Iterative modelbased identification of building components and appliances by means of sensor-actuator networks," 2011.

[5] P. G. I. Tomičić and M. Bača, "A review of soft biometrics for iot," in *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 1115–1120.

[6] P. Vlacheas, R. Giaffreda, V. Stavroulaki, D. Kelaidonis, V. Foteinos, G. Poulios, P. Demestichas, A. Somov, A. R. Biswas, and K. Moessner, "Enabling smart cities through a cognitive management framework for the internet of things," *IEEE communications magazine*, vol. 51, no. 6, pp. 102–111, 2013.

[7] M. Yun and B. Yuxin, "Research on the architecture and key technology of internet of things (iot) applied on smart grid," in *Advances in Energy Engineering (ICAEE), 2010 International Conference on*. IEEE, 2010, pp. 69–72.

[8] N. Bui and M. Zorzi, "Health care applications: a solution based on the internet of things," in *Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies*. ACM, 2011, p. 131.

[9] O. Vermesan, P. Friess, P. Guillemin, S. Gusmeroli, H. Sundmaeker, A. Bassi, I. S. Jubert, M. Mazura, M. Harrison, M. Eisenhauer *et al.*, "Internet of things strategic research roadmap," *Internet of Things-Global Technological and Societal Trends*, vol. 1, no. 2011, pp. 9–52, 2011.

[10] E. Welbourne, L. Battle, G. Cole, K. Gould, K. Rector, S. Raymer, M. Balazinska, and G. Borriello, "Building the internet of things using rfid: the rfid ecosystem experience," *IEEE Internet computing*, vol. 13, no. 3, 2009.

[11] A. Dohr, R. Modre-Opsrian, M. Drobics, D. Hayn, and G. Schreier, "The internet of things for ambient assisted living," in *Information Technology: New Generations (ITNG), 2010 Seventh International Conference on*. Ieee, 2010, pp. 804–809.

[12] I. Tomičić and M. Schatten, "Agent-based framework for modeling and simulation of resources in self-sustainable human settlements: a case study on water management in an eco-village community in croatia," *International Journal of Sustainable Development & World Ecology*, vol. 23, no. 6, pp. 504–513, 2016.

[13] ——, "A case study on renewable energy management in an eco-village community in croatia–an agent based approach," *International journal of renewable energy research*, vol. 6, no. 4, pp. 1307–1317, 2016.

[14] I. Tomičić, "Agent-based framework for modelling and simulation of resource management in smart self-sustainable human settlements," Ph.D. dissertation, Fakultet organizacije i informatike, Sveučilište u Zagrebu, 2016.

[15] Iot connected devices installed base worldwide from 2015 to 2025. [Online]. Available: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/

[16] B. A. Mozzaquatro, R. Jardim-Goncalves, and C. Agostinho, "Towards a reference ontology for security in the internet of things," in *2015 IEEE International Workshop on Measurements Networking (M N)*, 2015, pp. 1–6.

[17] L. a. G. S. Sicari, a. Rizzardi and a. Coen-Porisini, "Security, privacy and trust in internet of things: The road ahead," *Computer Networks*, vol. 76, 2014.

[18] P. Gonzalez-Gil, A. F. Skarmeta, and J. A. Martinez, "Towards an ontology for iot context-based security evaluation," in *2019 Global IoT Summit (GIoTS)*, 2019, pp. 1–6.

[19] J. Ye, L. Coyle, S. A. Dobson, and P. Nixon, "Ontology-based models in pervasive computing systems," *Knowledge Eng. Review*, vol. 22, pp. 315–347, 2007.

[20] P. S. N. G. V. I. Garvita Bajaj, Rachit Agarwal, "A study of existing ontologies in the iot-domain," *CoRR*, vol. abs/1707.00112, 2017.

[21] O. Mazhelis, E. Luoma, and H. Warma, "Defining an internet-of-things ecosystem," in *Internet of Things, Smart Spaces, and Next Generation Networking*, 2012, pp. 1–14.

[22] I. Szilagyi and P. Wira, "Ontologies and semantic web for the internet of things - a survey," in *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*, 2016, pp. 6949–6954.

[23] S. Hachem, T. Teixeira, and V. Issarny, "Ontologies for the internet of things," 12 2011.

[24] K. Kotis and A. Katasonov, "An ontology for the automated deployment of applications in heterogeneous iot environments 1," 2012.

[25] A. Souag, C. Salinesi, R. Mazo, and I. Comyn-Wattiau, "A security ontology for security requirements elicitation," in *International symposium on engineering secure software and systems*. Springer, 2015, pp. 157–177.

[26] J. Mylopoulos, A. Borgida, M. Jarke, and M. Koubarakis, "Telos: Representing knowledge about information systems," *ACM Trans. Inf. Syst.*, vol. 8, no. 4, p. 325–362, Oct. 1990.

[27] A. Avizienis, J. . Laprie, B. Randell, and C. Landwehr, "Basic concepts and taxonomy of dependable and secure computing," *IEEE Transactions on Dependable and Secure Computing*, vol. 1, no. 1, pp. 11–33, Jan 2004.

[28] C. E. Landwehr, A. R. Bull, J. P. McDermott, and W. S. Choi, "A taxonomy of computer program security flaws," *ACM Comput. Surv.*, vol. 26, no. 3, p. 211–254, Sep. 1994.

[29] A. Herzog, N. Shahmehri, and C. Duma, *An Ontology of Information Security*, 01 2009.

[30] S. Fenz and A. Ekelhart, "Formalizing information security knowledge," in *Proceedings of the 4th International Symposium on Information, Computer, and Communications Security*, ser. ASIACCS '09. Association for Computing Machinery, 2009, p. 183–194.

[31] J. Undercoffer, A. Joshi, and J. Pinkston, "Modeling computer attacks: An ontology for intrusion detection," in *Recent Advances in Intrusion Detection*. Springer Berlin Heidelberg, 2003, pp. 113–135.

[32] D. Geneiatakis and C. Lambrinoudakis, "An ontology description for sip security flaws," *Computer Communications*, vol. 30, pp. 1367–1374, 03 2007.

[33] G. Denker, L. Kagal, T. Finin, M. Paolucci, and K. Sycara, "Security for daml web services: Annotation and matchmaking," vol. 2870, 10 2003, pp. 335–350.

[34] G. Denker, S. Nguyen, and A. Ton, "Owl-s semantics of security web services: a case study," in *The Semantic Web: Research and Applications*, C. J. Bussler, J. Davies, D. Fensel, and R. Studer, Eds. Springer Berlin Heidelberg, 2004, pp. 240–253.

[35] G. Denker, L. Kagal, and T. Finin, "Security in the semantic web using owl," *Inf. Secur. Tech. Rep.*, vol. 10, no. 1, p. 51–58, 2005.

[36] A. Kim, J. Luo, and M. Kang, "Security ontology for annotating resources," in *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 1483–1499.

[37] A. A. Assali, D. Lenne, and B. Debray, "Ontology development for industrial risk analysis," in *2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications*, April 2008, pp. 1–5.

[38] B. Tsoumas and D. Gritzalis, "Towards an ontology-based security management," in *20th International Conference on Advanced Information Networking and Applications - Volume 1 (AINA'06)*, 2006, pp. 985–992.

[39] M. Karyda, T. Balopoulos, S. Dritsas, L. Gymnopoulos, S. Kokolakis, C. Lambrinoudakis, and S. Gritzalis, "An ontology for secure e-government applications," in *First International Conference on Availability, Reliability and Security (ARES'06)*, 2006.

[40] D. G. Firesmith, "A taxonomy of security-related requirements," 02 2020.

[41] H. Mouratidis, P. Giorgini, and G. Manson, "An ontology for modelling security: The tropos approach, knowledge-based intelligent information and engineering systems," *Lecture Notes in Artificial Intelligence*, pp. 1387–1394, 01 2003.

[42] F. Massacci, J. Mylopoulos, F. Paci, T. T. Tun, and Y. Yu, "An extended ontology for security requirements," in *Advanced Information Systems Engineering Workshops*. Springer Berlin Heidelberg, 2011, pp. 622–636.

[43] A. Gyrard, C. Bonnet, and K. Boudaoud, "An ontology-based approach for helping to secure the etsi machine-to-machine architecture," in *2014 IEEE International Conference on Internet of Things (iThings), and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom)*, 2014, pp. 109–116.

[44] D. Minoli, K. Sohraby, and J. Kouns, "Iot security (iotsec) considerations, requirements, and architectures," in *2017 14th IEEE Annual Consumer Communications Networking Conference (CCNC)*, 2017, pp. 1006–1007.

[45] F. de Franco Rosa and M. Jino, "A survey of security assessment ontologies," *CoRR*, vol. abs/1706.09771, 2017.

[46] C. Blanco, J. Lasheras, R. Valencia-García, E. Fernández-Medina, A. Toval, and M. Piattini, "A systematic review and comparison of security ontologies," in *2008 Third International Conference on Availability, Reliability and Security*, 2008, pp. 813–820.

[47] B. Mozzaquatro, C. Agostinho, D. Goncalves, J. Martins, and R. Jardim-Goncalves, "An ontology-based cybersecurity framework for the internet of things," *Sensors*, vol. 18, p. 3053, 09 2018.

[48] M. Arnaert, Y. Bertrand, and K. Boudaoud, "Modeling vulnerable internet of things on shodan and censys : An ontology for cyber security," in *SECURWARE 2016*, 2016.

[49] D. L. McGuinness, F. Van Harmelen *et al.*, "Owl web ontology language overview," *W3C recommendation*, vol. 10, no. 10, p. 2004, 2004.

[50] M. Fernández-López, A. Gómez-Pérez, and N. Juristo, "Methontology: from ontological art towards ontological engineering," 1997.

[51] N. Noy and L. Mcguinness, "Ontology development 101: A guide to creating your," rst ontology. Technical Report KSL-01-05, Stanford Medical Informatics, Stanford, Tech. Rep., 2001.

[52] M. Jouini, L. B. A. Rabai, and A. B. Aissa, "Classification of security threats in information systems." *ANT/SEIT*, vol. 32, pp. 489–496, 2014.

[53] S. Alam, M. M. Chowdhury, and J. Noll, "Interoperability of security-enabled internet of things," *Wireless Personal Communications*, vol. 61, no. 3, pp. 567–586, 2011.

[54] D. Andročec, B. Tomaš, and T. Kišasondi, "Interoperability and lightweight security for simple iot devices," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2017, pp. 1285–1291.

Decision tree algorithm for control of compressor multiset in refrigeration industry

Ivan Šulekić*, Damir Milinković ** and Tomislav Špoljarić*
* University of Applied Sciences – Department of Electrical Engineering, Zagreb, Croatia
** Ekonerg d.o.o., Koranska ul. 5, Zagreb, Croatia
isulekic@tvz.hr, tspoljaric@tvz.hr

***Abstract* - This paper describes novel control algorithm for the multi-compressor set. Algorithm has input parameters that are collected from each compressor in the set. According to these parameters, algorithm decides which combination of compressors is the best for optimum performance of the entire set**

Keywords - multi-compressor set, controls, weight factor parameters, Siemens, maximum efficiency

I. INTRODUCTION

Compressor multiset comprises of four compressors. Each of them serve to maintain pressure in the refrigeration system. Multisets control system monitors current pressure in the system and vital compressor parameters. If pressure in the system drops below referent value algorithm checks if next compressor is ready. If it is ready, control system turns it on. During exploitation, compressors parameters can dangerously exceed limit values and then control system needs to turn that compressors off. Fig. 1 shows program algorithm for control system.

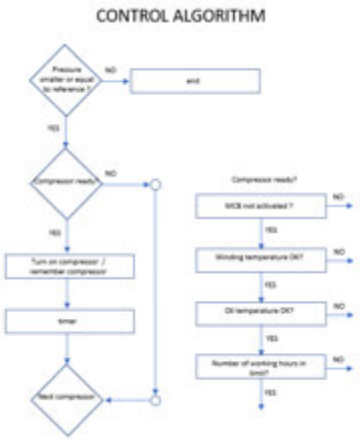


Fig. 1: control algorithm

Current control system monitors these parameters: winding temperature, oil temperature. motor overload switch state and number of working hours.

PT100 sensors are used for measuring winding and oil temperature. Motor overload switch state is obtained from NO and NC contacts. Control system software counts number of working hours. Digital outputs are used for starting and stopping compressors.

Control system assigns weight factors to all monitored parameters. With weight factors programmer can change influence of every parameter on the reliability of a system. Control system compares parameters for each compressor and decides the most optimal compressor for operation.

II. PROCESS AND CONTROL SYSTEM DESCRIPTION

Refrigeration system needs to be supplied with constant pressure. A compressor multiset is used for that purpose. Before modernization, control system multiset worked in the following way: only one compressor in the set would work until it would overload. Then another compressor would start.

New control system uses novel control algorithm for selection of the most optimal compressor. The most optimal compressor is the one with smallest total parameter value. Total parameter value is calculated from the individual parameter values of monitored compressor (1)

$$p_{tot} = \frac{\lambda p_{ot} + \lambda p_{wt} + p_{nh} + 4x\lambda p_{mcb}]}{4} \quad (1)$$

$p_{tot} - total\ parameter\ value$
$p_{mcb} - motor\ circuit\ braker\ parameter\ value$
$p_{wt} - winding\ temperature\ parameter\ value$
$p_{ot} - oil\ temperature\ parameter\ value$
$p_{nh} - working\ hour\ parameter\ value$
$\lambda - weight\ faktor$

During operation, every parameter value can be modified with weight factors. All parameter values in the algorithm are normalized .in value range from 0 to 1. Motor overload switch normalized value can only be two values: 0 or 1. Oil and winding temperature and working hour normalized values are in the range from 0 to 1.Current control system is designed for four compressors in multiset, but it can be modified for any number of compressors.

III. AUTOMATION CONTROL DESIGN

Devices used in this system are: compressor multiset, temperature sensors, motor overload switch, Siemens PLC 1200 industrial controller with analog and digital I/O module and HMI (Human machine interface). Compressor multiset consists of four compressors. Temperature sensors are realized with: PT100 sensors. Motor overload switch

have NC and NO contacts for current state signal. Siemens PLC 1200 industrial controller is used as a system regulator. Analog I/O module is used for interaction with temperature and pressure sensors. Digital I/O module is used for detecting motor overload switch states and for starting and stopping compressors. All devices are documented in technical documentation [3].

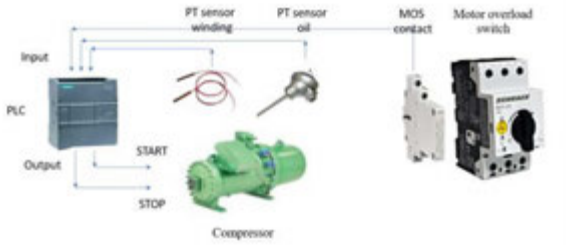


Fig 2: Graphic representation of all devices in the system

Compressor control operates according to winding and oil temperature, pump working hours and motor overload switch state. Winding and oil temperature are detected with temperature sensors. Motor overload switch state is detected with contact. Working hours are counted by control system during compressor exploitation.If control system detects higher temperature values than limit values from winding or oil that event will be logged in control system. Control system also logs every motor overload switch failure and it counts compressors working hours. All those events are normalized as values in the range from 0 to 1. The algorithm combines all those normalized values and calculates overall reliability of the compressor. If some other compressor has a better reliability, then control system will switch to that compressor. Temperature normalized values are calculated according to (2):

$$p_{ot,wt} = \frac{75 - t_{curr}}{50} \quad (2)$$

$$t_{curr} - current\ temperature$$

Working hour normalized values is calculated according to the (3) and (4).

$$p_{nh}' = \frac{p_{nhmax} - 0{,}9 p_{nhcurr}}{p_{nhcurr}} \quad (3)$$

$$p_{nh} = 1 - p_{nh}' \quad (4)$$

$p_{nhmax} - maximum\ value\ of\ working\ hours$
$p_{nhcurr} - current\ value\ of\ working\ hours$

Motor overload switch normalized value can only be 0 or 1. Overall reliability of compressor is calculated with (1):

$$p_{tot} = \frac{\lambda p_{ot} + \lambda p_{wt} + \lambda p_{nh} + 4x\lambda p_{mcb}}{4} \quad (1)$$

On Fig. 3 there is a graphic representation of overall reliability for all compressors in the set:

Current status of all compressors parameters is visualized with HMI. Operator can see the current state of all compressors. Programmer can give priority to some

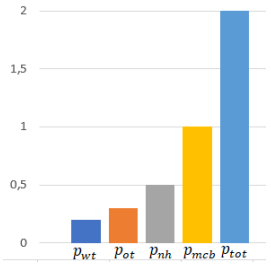


Fig 3: Graphic representation of overall liability

parameters. If some parameter has a greater impact on compressors reliability it is possible to change its weight. With this we can obtain maximum efficiency in the system.

IV. PROGRAM SOLUTION IN LADDER LOGIC

Process is organized in cycles. Number of cycles depends on the number of compressors in multiset. Each cycle has control algorithm for one compressor.

For every cycle we have three main steps:

1) *Is pressure smaller or equal to refernce ?*
2) *Check if next compressor is ready*
3) *Turn on compressor*

1)

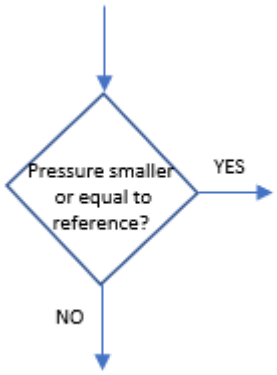


Fig 4: Program section 1: Desired pressure

2)

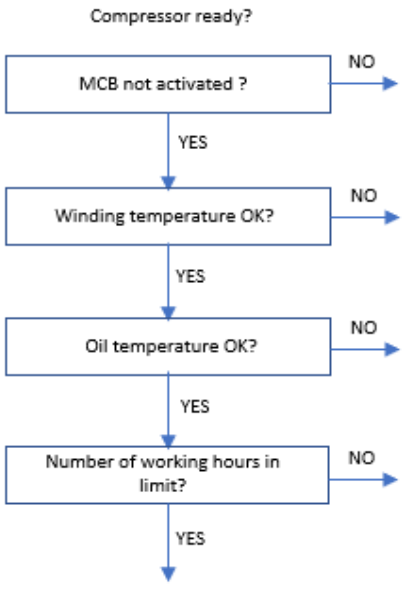


Fig 5: Program section 2: Compressors availability

3)

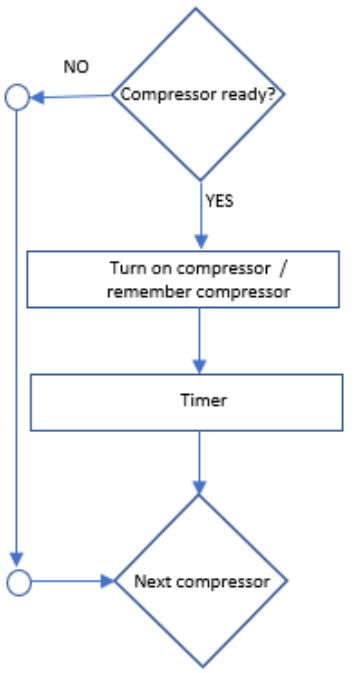


Fig 6: Program section 3: Start compressor

Before programming it is necessary to make a list of all used variables. The list includes physical and virtual I/O variables used later in the program and they are shown in table 1.

TABLE 1 LIST OF VARIABLES IN LADDER PROGRAM

Device	I/O signal	Name
digital I/O module	I1	Local start/stop
	I2	Local/remote
	I3	Compressor 1 ready
	I4	Compressor 2 ready
	I5	Compressor 3 ready
	I6	Compressor 4 ready
	I7	MOS activated
	Q1	Turn on compressor 1
	Q2	Turn on compressor 2
	Q3	Turn on compressor 3
	Q4	Turn on compressor 4
	Q6	Turn off compressor 1
	Q2	Turn off compressor 2
	Q7	Turn off compressor 3
	Q8	Turn off compressor 4
Analog I/O module	AI1	Winding temperature compressor 1
	AI2	Winding temperature compressor 2
	AI3	Winding temperature compressor 3
	AI4	Winding temperature compressor 4
	AI5	Oil temperature compressor 1
	AI6	Oil temperature compressor 2
	AI7	Oil temperature compressor 3
	AI8	Oil temperature compressor 4

The code programming is done in Siemens TIA portal with Simatic Step 7 software package [1]. It is used for communication setup, definition of variables and code writing.

In first part of the code system needs to check pressure. If the pressure is smaller than desired, control system first needs to check how many compressors are ready for activation. Winding temperature, oil temperature, number of working hours and state of motor overload switch need to be in operating limits for compressor activation (see Fig 7). When controller checks how many compressors are ready it decides which compressor to activate. It selects, according to control algorithm, the most reliable compressor. Most reliable compressor is the one that has biggest overall reliability. Example of overall reliability calculation is given in Fig. 8. Calculation is based on formulas (1), (2), (3) and (4). After selection compressor is activated. Controller waits for a specified time. If compressor doesn't reach desired pressure controller needs to start second most reliable compressor (see Fig 9). This process is repeated until desired pressure is reached. Every time when compressor stops new reliability is calculated for it. Control system has the same program cycle for every compressor.

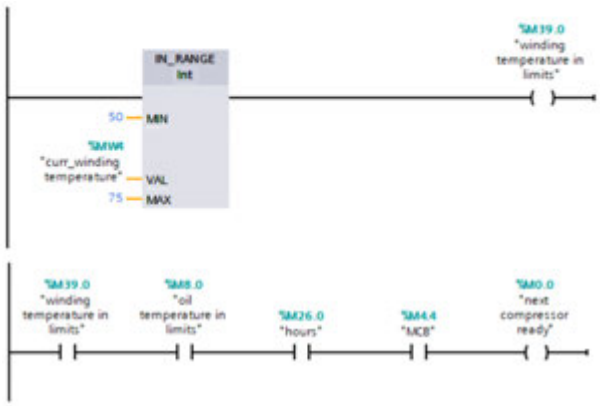


Fig 7: Example of program section 1: Compressors availability

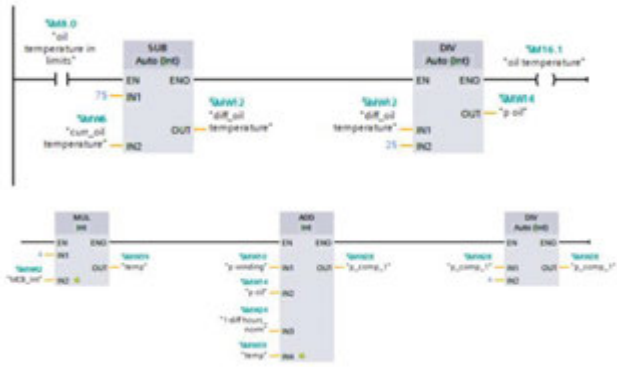


Fig 8: Example of program section 2: Overall reliability calculation

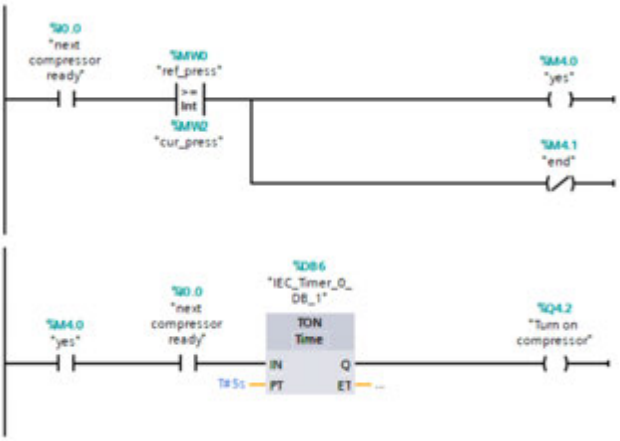


Fig 9: Example of program section 3: Turning on compressor

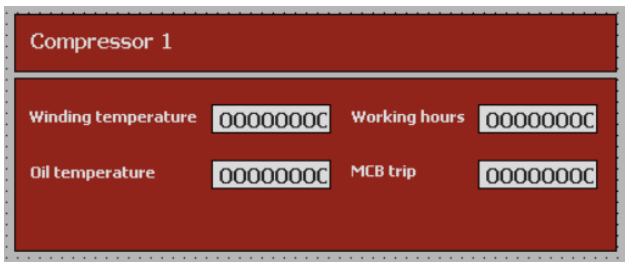


Fig. 10: Example of monitoring mask for one compressor

V. USER INTERFACE

User interface is created with Siemens TIA portal [2]. With this interface, operator can monitor operation of the entire compressor multiset. Observed parameters are visualized for every compressor. During exploitation, programmer can change weights of monitored parameters. Current values of monitoring parameters can be seen on screen (Fig. 10). Table 2 lists all variables and parameters that can be accessed through user interface. User interface is connected with Siemens PLC 1200 industrial controller.

TABLE 2 LIST OF VARIABLES IN USER INTERFACE

Name
Compressor 1 ready
Compressor 2 ready
Compressor 3 ready
Compressor 4 ready
MOS activated
Turn on compressor 1
Turn on compressor 2
Turn on compressor 3
Turn on compressor 4
Turn off compressor 1
Turn off compressor 2
Turn off compressor 3
Turn off compressor 4
Winding temperature compressor 1
Winding temperature compressor 2
Winding temperature compressor 3
Winding temperature compressor 4
Oil temperature compressor 1
Oil temperature compressor 2
Oil temperature compressor 3
Oil temperature compressor 4
Working hours compressor 1
Working hours compressor 2
Working hours compressor 3
Working hours compressor 4
Number of available compressors
Current reliability of compressor 1
Current reliability of compressor 2
Current reliability of compressor 3
Current reliability of compressor 4

Operator can only manually start compressor with user interface but only if compressor is ready. All other variables are only for monitoring purposes.

EXPERIMENTAL RESULTS

During half-year exploitation, efficiency of novel control algorithm was monitored. Monitored parameters were modified with different weight factors to achieve maximum efficiency. With maximum efficiency control system achieved maximum working hours with minimum number of failures. Number of events in that case is listed in Table 3. For comparison, number of events for previous half year without novel control algorithm is displayed in Table 4.

TABLE 3 NUMBER OF EVENTS WITH NOVEL CONTROL ALGORITHM

Parameter	**Number of events**			
	Comp 1	Comp 2	Comp 3	Comp 4
Winding temperature	19	28	10	0
Oil temperature	0	0	0	0
Compressor switch because of working hours	0	0	1	1
State of motor overload switch	5	0	2	6

TABLE 4 NUMBER OF EVENTS WITHOUT NOVEL CONTROL ALGORITHM

Parameter	**Number of events**			
	Comp 1	Comp 2	Comp 3	Comp 4
Winding temperature	45	41	17	3
Oil temperature	0	3	0	0
Compressor switch because of working hours	0	0	0	0
State of motor overload switch	11	0	3	0

Comparing Table 3 and Table 4 it can be concluded that with novel control algorithm winding overheating is reduced and working compressor can be replaced with more reliable one. Because of these improvements there are less overloads. Oil temperature was in the limits in both cases.

VI. CONCLUSION

Novel control algorithm uses compressors parameters for the optimal performance of multi-compressor set. The biggest advantage of the new algorithm is parameter adaption during exploitation. During exploitation, programmer can adapt parameter values with different weights. With this, system can be set to work on maximum efficiency. New monitored parameters can easily be added. Further possibilities of control improvement include control system with more monitored parameters.

Implementation of this algorithm resulted in the reduction of malfunctions and raise of reliability of the entire set.

REFERNCES

[1] SIEMENS SIMATIC Programing with Step 7, Function manual, V4.2.3, 08/2018

[2] TIA Portal V14 Manual, 2016

[3] Technical documentation: Špolarić T.: Manual for +ROH.CNUS control system V4.0, 2019

GIS analysis of basketball courts and healthy stores relationship for young population in the city of Skopje

N. Stojanova, R. Vignjevikj and A. Naumoski
Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University in Skopje
Skopje, Republic of North Macedonia
e-mail: {nevena.stojanova, rade.vignjevic}@students.finki.ukim.mk, andreja.naumoski@finki.ukim.mk

***Abstract* - Today we live in a world where everything that we need we can obtain from the comfort of our homes. Although this is very convenient, our physical and mental health are paying the price. Physical exercises are one of the key elements in our daily lives that keep our bodies in top shape. Therefore, in this paper we address two key health components: exercises through sport and healthy eating. For this purpose, we have mapped most of the open fields and closed basketball courts in the city of Skopje, the population numbers of young people between the age of 5 and 25, as well as the healthy food and pharmacy stores using Geographic Information Systems. Combing the information from these three sources, we obtain data fusion spatial analysis map that we use to find the best location for opening new basketball court, and in the same time providing crucial knowledge of spatial regions where young people can have healthy food or buy medicines in a case of injury.**

Keywords - geographical information systems; young people; sport; healthy food

I. INTRODUCTION

This topic has been selected for analysis because the health of the youth is jeopardized. We are living in a world where everything we need, we can obtain from the comfort of our homes. Starting from delivering food to our front door, to ordering our clothes online and even getting medications shipped to our homes, we have built a lifestyle that greatly decreases the physical activities we perform during the day. Moreover, access to fast food kiosks is easier than ever which additionally contributes to the severeness of the problem. Therefore, healthy habits and vitality need to be established and the best age to do that is while we are still developing as young individuals [1]. Researchers in the field of applying Geographical Information Systems (GIS) for solving various problems have identified some important patterns between the environment and the physical activity in the young population. The study [2], have built a GIS tool for measuring the infrastructural support for physical activity. This is very important in every city, especially for children with health issues like obesity and their relationship with the food stores. The research had investigated not only the relationship between the young population and the food retail stores, but also education, transport, recreation, social and cultural, financial, health, and other retail stores. Similar research is done by [3], where the authors have specifically discussed the relationship between food and physical activity environments as they might apply to obesity. Other research has used GIS, to find the relationship with specific health issues for the city residents, the distribution of cardiovascular risk factors [4]. This was done by onsite field study on all food stores and street segments, and that data was related to Body Mass Index (BMI) of the subjects. The results from that study have shown greater school walkability was associated with significantly lower mean BMI. Like this study, the researchers in [5], investigate the selective daily mobility bias in children and the BMI index. Another study combined GIS, GPS and Accelerometer devices to explore the young population activity and the environment [6]. As we can see from studies like [7, 8], GIS is well used to draw a connection between the health of the city young population, the physical activity as one of the key aspects of the healthy body and the distribution of the food retail stores. Therefore, it is vital to decrease the popularity of fast food – unhealthy eating, to put an accent on the stores that sell foods rich in nutrients and relate them with sport objects, like basketball courts.

The city of Skopje has 39 elementary schools and 35 high schools around, which not many healthy food stores and sports courts can be found. Also, there are so many newly built courts, that people with sports enthusiasts and any other stakeholders need a more informative way and more precise map of the locations of those courts. We find this problem very interesting to investigate further on and therefore this paper presents the analysis of the number of health stores and courts and their proximity to the schools. This analysis would also help the sportsman to find the most convenient location to exercise and to nurture the body properly.

All the conducted analysis is made with software tools provided in the ArcGIS package, version 10.4.1, or to be exact: ArcMap, ArcCatalog and ArcToolbox [9]. Using the tools built into the ArcMap package, we present the locations of the elementary and high schools, the healthy food stores and the sports courts in Skopje and their proximity to each other. Because in the last few years there has been great interest in building sports facilities, the search for business opportunities and the investments in sports facilities are generally seen as activities where detailed statistics is traditionally seen as the best way to determine where investors should put their money. However, spatial understanding can be a critical component for understanding where and how investments may occur

for given markets and this analysis could give the needed answers.

The rest of the paper is organized as follows – Section II explains the datasets we used, visualize the data and explains the process of buffering and interpolation of the target attributes. Section III presents the modeling results and a discussion for them. Finally, Section IV concludes the paper and gives direction for future work.

II. DESCRIPTION, GEO DATABASE, AND VISUALIZATION

The purpose of the paper is to display the locations of basketball courts, the locations of healthy food stores, the locations of the elementary, high schools on a map in Skopje and find out the spatial relationship information about them. Its primary usage would be as a tool for analysis and finding the most suitable location for building new basketball courts and finding the most convenient one from the existing, according to the persons' preferences. The main source of data for the preparation of this project was Google Maps [10] and the data from the state statistical office's database, [11]. They were used for the acquisition of information regarding the location of the subjects of interest and the number of pupils in the schools.

The overall data used for mapping the basketball courts and the analysis, at first was imported into excel sheets and then converted into a geodatabase, represented with tables. For each object (basketball court, school and healthy food store) we have similar attributes denoted as follows:

- **Object-ID**: Numeric identifier which represents the primary key of the object.
- **Latitude and Longitude**: Geographic XY coordinates in decimals which represent the geographic location of the object
- **Name**: Name of the object used for identification
- **Address**: Address of the object, more specifically street name and number.
- **Rate:** Numerical value obtained from the people's satisfaction with the school's educational practices and facilities.

Furthermore, for each school object, we have additional attributes like number of pupils attending the school for the academic 2017/2018 year, as well as the sea altitude of the basketball court.

In order to connect the data to the map, the map named "Macedonia_Map_GIS.mxd" was opened and in ArcMap the options File -> Add Data -> Add XY Data were selected. By doing this, the database was activated, and the corresponding coordinates were automatically parsed on the map. Further on, in the folder World, the WS 1984 geographic coordination system was chosen. The reason for choosing this one, and not any other, is that some other coordinating systems cause problems in the mapping of the coordinates. More specifically, this was achieved by choosing the option Coordinate System of Input Coordinates and afterward by clicking on the button Edit a new set of options opened. We chose Geographic Coordinate Systems -> World -> WGS 1984. Initially, we obtained the locations of the basketball courts in the form of little black dots. With further editing, the black dots were replaced by icons from different types of sports balls. Additionally, the color and size of the icons were changed. For the open basketball courts, we chose the color yellow while for the closed basketball courts we left the default color i.e. white. These changes are shown on Fig. 1.

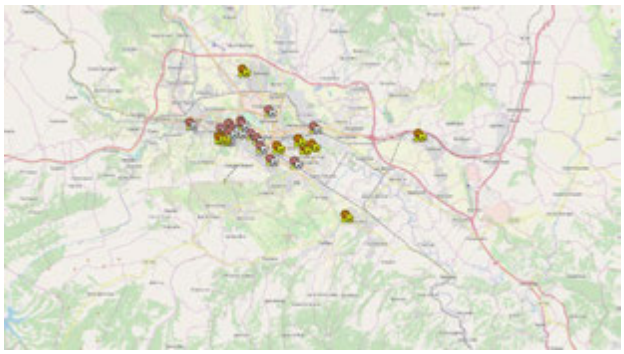


Figure 1. Mapped basketball courts

By right-clicking on any court on the map, a menu opens wherein the option "Properties", the displayed information for the basketball courts can be seen. After clicking Properties, "Display expressions" was selected which allowed us to see some data about the court when we hover over it. The expressions used are shown below, while the result from this express is presented in Fig 2.

```
Function FindLabel ( [NameOfCourt], [Address], [SeaLevel], [Rate] )
    FindLabel = [NameOFCourt] + " is located on address: " + [Address] + ". The sea level is " + [SeaLevel] + " and the rate is: " + [Rate] + " ."
End Function
Function FindLabel ( [NameOfCourt], [Address], [SeaLevel], [Rate] )
    FindLabel ="The court which is part of " + [NameOfCourt] + " is located on address: " + [Address] + ". The sea level is " + [SeaLevel] + " and the rate is:  " + [Rate] + " ."
End Function
```

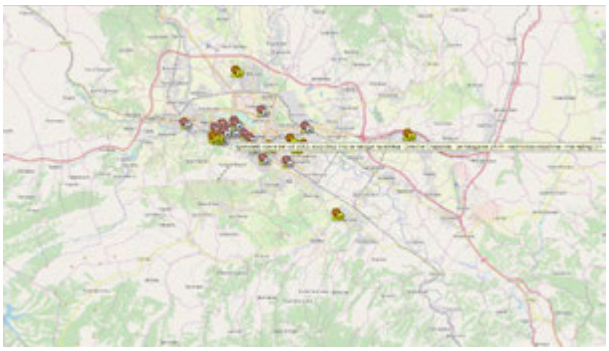


Figure 2. Display of basic information for a court on hover

The mapping for the other 2 subjects, healthy food stores and schools, was done similarly like the described procedure above. For the representation of schools, we used red house icons and for the representation of the stores, we used the icon which presents healthy food colored in green.

The visualization was done on the map of the Republic of North Macedonia, downloaded free from the web site of Geofabrik [12]. By using the corresponding Export Data function on the appropriate (gis.osm_places_a_free_1.shp) shapefile we created the map of the capital city, Skopje. With further editing and using various options in ArcMap, we removed all other borders except the borders of Skopje which can be seen in Fig. 3 marked with orange color, and the suburbs of the capital city with a blue color. The main area of our analysis focuses on the main capital area with suburbs marked with the orange color.

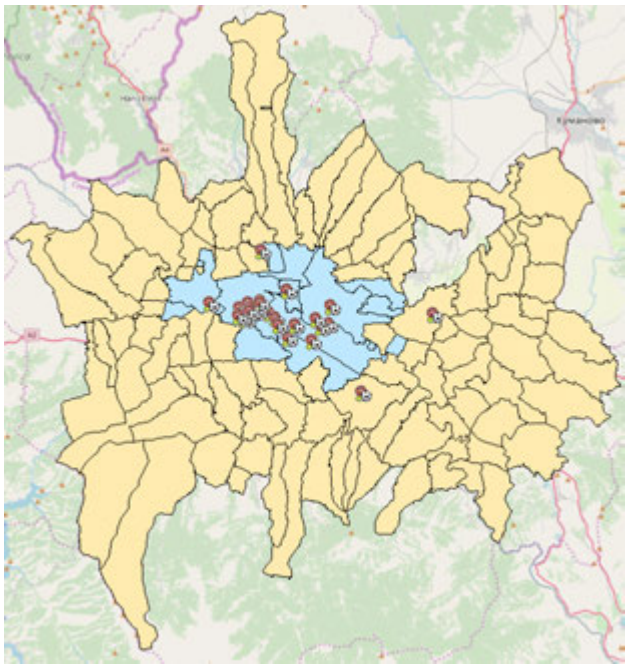


Figure 3. The borders of Skopje which are needed for analysis

Using the tool in the ArcToolbox, Multiple Ring Buffers, we were able to make a comparison between the distances of the schools to the healthy food stores and the sports courts. The main reason why this tool was used is that it provides a clearer picture of the proximity of each of the healthy food stores and basketball courts in the city of Skopje. In order to limit the area, in the Geoprocessing menu, by choosing "Environment settings" we marked the cartographic coordinate system. In Raster Analysis, the layer of Skopje was chosen for the Mask. In ArcToolBox, from the Analysis Tools, the option Proximity -> Multiple Ring Buffer was chosen which is one of the last steps in the process for configuring the Multiple Ring Buffer. Each radius is represented by a different color on every 500m distance of the schools.

III. GIS MODELLING RESULTS

In this section, the results of the conducted analysis are presented. In Fig. 4 shows the relationship between the rate of healthy food stores and the 3 subjects of interest: the schools, the sports courts, and the healthy food stores. It can be observed that there are several white regions in the area of the city of Skopje. These regions represent the stores with a high rate but there are not any schools nor sports courts. On the other hand, we have many healthy food stores with average rates concentrated in the city center, where a lot of schools and sports courts are also present. Further on, there are green colored regions. These regions represent healthy food stores at a low rate. As it can be seen, nearby these regions there are schools and courts present.



Figure 4. The effect of the healthy food stores' rate on the other subjects of interest

In Fig. 5 shows the distribution of the number of students in the schools in comparison with buffers whose centers are the sports courts within a radius of 500 meters. The red areas represent the schools that have the largest number of students.

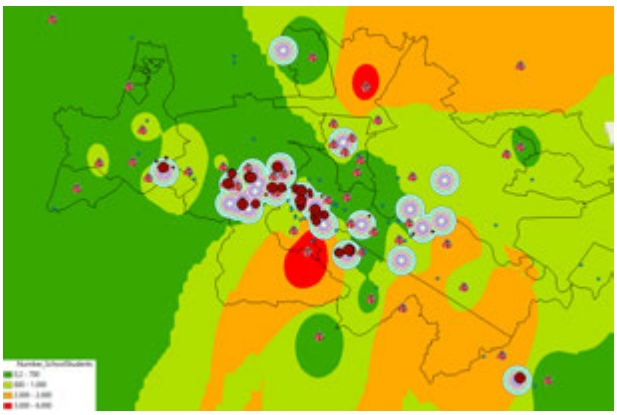


Figure 5. Relationship between the student number and the proximity of courts

The number of students in the city center is smaller than the number of students in the red areas, but in the city center, the concentration of sports courts is much larger. The buffers show that there are a lot of healthy food stores within a radius of 500 meters, but there are also stores that are represented by the small circles, that are in areas with a smaller number of students.

In Fig. 6 the distribution of the number of students in the schools and the buffers whose centers are the healthy food stores within 500 meters are shown. The distance is calculated in comparison with the schools themselves. Here, the distributions of the number of stores within a

radius of 500 meters from the schools are almost the same as the distribution of the number of stores outside the specified radius (these are marked with the yellow dots). It can be observed as well that there are schools in whose surroundings are not any healthy food stores but on the other hand, the number of students in these schools is smaller.

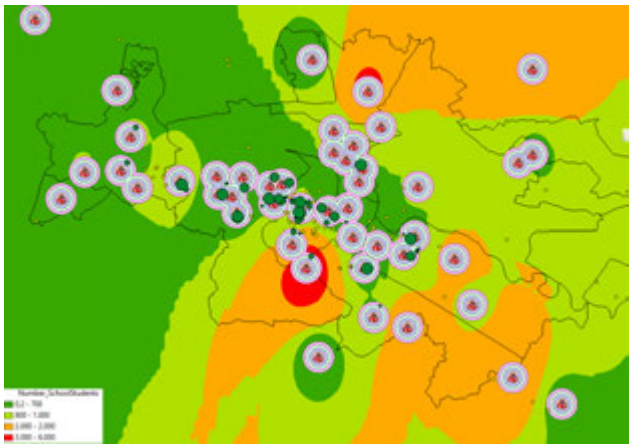


Figure 6. The relationship between the number of students and the buffers in whose center are the stores

Regarding the results represent in Fig. 4, we observed that there are stores with a high rate but there are not any schools nor stores. These are the locations that are a good opportunity to build sports courts and schools. On the other hand, there are a lot of stores nearby schools that have an average rate. These are the stores that need improvement. Regarding the Fig. 5 and 6, it can clearly be seen that the red areas are the most convenient locations for opening new healthy food stores and sports courts since the number of students in those areas is the largest and in its surroundings there are not any type of the buildings of interest. In the other parts of the city that are not close to the center, we have a lot of courts but not many healthy stores. Indeed, the density of healthy stores should expand to the other parts of the city because there are a lot of opportunities to develop their business. Additionally, it would be much more convenient for the residents of those parts of the city to have local healthy food stores and plus the number of clients could expand.

There is some additional analysis on the effect of the sea-level of the sports courts on the rate of the court. On Fig. 7, a plot representing the relationship between the rate and the sea level of basketball courts is shown.

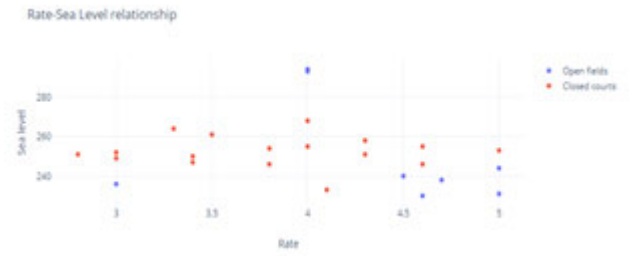


Figure 7. Rate-Sea level plot

The y-axis represents the sea level in meters while the x-axis represents the rate on a scale from 1-5. The open basketball fields are plotted with blue circles and the closed basketball courts, or more specifically basketball courts that are part of some elementary or high school, are plotted with red circles As it can be seen in Fig. 7 above, the open fields with rates higher than 4.5 tend to be on a sea-level around 240 meters but there is an exception with the field with rate 3. When talking about closed courts, it can be observed that no matter the rate of the court, they are all located on a sea-level around 250 meters. So, from this analysis, we can say that the sea level affects the rate of the open fields, but it does not have any influence on the rate of the closed courts.

From the second analysis, we were able to conclude that the closer the open field to the city center the higher the rate. Again, when it comes to the closed courts, the distance to the city center seems to have no influence on the rate of the basketball court. In Fig. 8, a plot representing the rate and the distance from the basketball courts to the center of Skopje is shown.

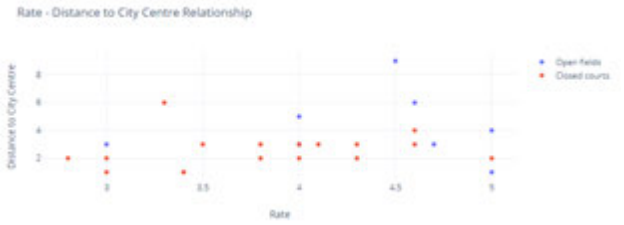


Figure 8. *Rate-Distance to City Centre plot*

The open fields are represented with blue circles while the closed courts are represented with red circles. The x-axis represents the rate and the y-axis represents the distance to the city center. The distance was calculated using the longitude and latitude values of each of the courts and the city center.

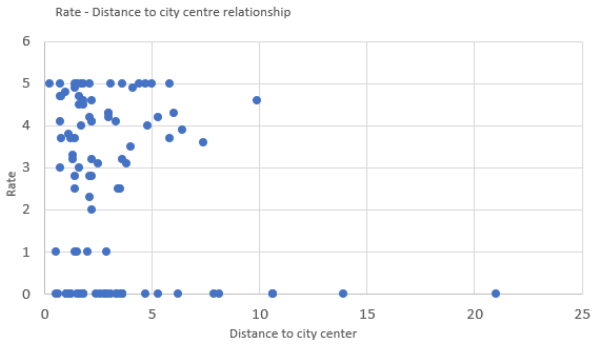


Figure 9. *Relationship between the rate and the distance to the city center of healthy food stores*

The relationship between the rate and the distance of the healthy food stores from the city center is represented in Fig. 9. The y-axis represents the rate of the store while the x-axis represents the distance in kilometers of the store to the center of the city. This relationship shows that the healthy food stores are mostly located within a radius of 5 kilometers of the center of Skopje, but the rates of the stores vary from 0-5. It is an indicator that the location of the store does not affect the rate, and it can be concluded that the stores need to be dispersed across all parts of the city.

IV. CONCLUSION

In this paper, the process of building the geodatabase and visualization of the basketball courts, schools and healthy food stores is presented. We did some analysis based on some attributes of the courts, schools and healthy food stores like rate, sea-level, distance from the courts to the other objects of interest, distance from the stores to the other two objects of interest and distance from the center of Skopje for each of the courts and stores. The data used was obtained from the services of Google and the national stastical office.

The analysis showed that the sea-level and the distance from the central area has impact on the popularity of the open basketball courts whilst the same parameters do not affect the closed courts. Further on, the areas where the healthy food stores need improvement were identified as well as the areas which are most suitable for building new schools and sports courts. This information is important because it needs to be considered when building new basketball courts, schools and healthy food stores in the future in order to have the desired results i.e. promoting healthy nutrition and sports activities to the youth for a better lifestyle. Another useful outcome is that the users of this map will have no difficulties in choosing the right basketball court for sport and recreation or selecting the nearby healthy food store for proper nutrition.

The ArcGIS packages offer many other possibilities. In future, we plan to enhance the interpolation analysis done in this paper by adding more attributes to the model such as capacity of audience at the basketball courts, information about whether an entrance fee is required, information about whether reservation is needed and working hours of the basketball court, as per the healthy food stores we plan to investigate whether the types of food they offer affects their popularity. Besides this, we also plan to implement this type of solution for the remaining cities in North Macedonia.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of Computer Science and Engineering at the Ss. Cyril and Methodius University in Skopje.

REFERENCES

[1] M.C. Gallotta, M. Chiara, S. Iazzoni, G.P. Emerenziani, M. Meucci, S. Migliaccio, L. Guidetti, and C. Baldari. "Effects of combined physical education and nutritional programs on schoolchildren's healthy habits." PeerJ, vol. 4, e1880, 2016.

[2] K. Witten, J. Pearce, and P. Day, "Neighbourhood Destination Accessibility Index: a GIS tool for measuring infrastructure support for neighbourhood physical activity," Environment and planning vol. A43, no. 1, 205-223, 2011.

[3] L.E. Thornton, R.P. Jamie, and A.M. Kavanagh. "Using Geographic Information Systems (GIS) to assess the role of the built environment in influencing obesity: a glossary," International Journal of Behavioral Nutrition and Physical Activity, vol. 8, no. 1, pp. 71, 2011.

[4] A. Cebrecos, J. Díez, P. Gullón, U. Bilal, M. Franco, and F. Escobar, "Characterizing physical activity and food urban environments: a GIS-based multicomponent proposal," International journal of health geographics, vol. 15, no. 1, pp. 35, 2016.

[5] T. Burgoine, A.P. Jones, R.J.N. Brouwer, and S.E.N. Benjamin, "Associations between BMI and home, school and route environmental exposures estimated using GPS and GIS: do we see evidence of selective daily mobility bias in children?," International journal of health geographics vol. 14, no. 1, pp. 8, 2015.

[6] P.R.W. McCrorie, C. Fenton, and A Ellaway, "Combining GPS, GIS, and accelerometry to explore the physical activity and environment relationship in children and young people-a review," International Journal of Behavioral Nutrition and Physical Activity vol. 11, no. 1, pp. 93, 2014.

[7] K. Patrice, The use of Geographical Information Systems (GIS): Mapping fast food restoraount, BMI and mental health.. Master Thesis, California State Univeriosty, Sacramento,USA, 2013.

[8] I. Kawachi, L.F. Berkman (Eds). Neighborhoods and health. Oxford University Press; 2003.

[9] ESRI. ArcGIS Desktop: Release 10.4.1. Redlands, CA: Environmental Systems Research Institute, 2018.

[10] Google Maps. "Google Maps", www.maps.google.com. Accessed 14 December 2019.

[11] State Statistical Office of North Macedonia, MakStat data for the school year 2017/2018. Source: , http://makstat.stat.gov.mk/PXWeb/pxweb/mk/MakStat, Last accessed 15 September 2019.

[12] Geofabrik official webiste. https://www.geofabrik.de/. Last accessed:10 Octomber 2020.

Decentralized trustless gossip training of deep neural networks

Robert Šajina
Faculty of Informatics in Pula
Pula, Croatia
robert.sajina@unipu.hr

Nikola Tanković
Faculty of Informatics in Pula
Pula, Croatia
nikola.tankovic@unipu.hr

Darko Etinger
Faculty of Informatics in Pula
Pula, Croatia
darko.etinger@unipu.hr

***Abstract*—Novel machine learning techniques apply decentralized model training in order to mitigate data volume and privacy issues. Current approaches assume (a) node performance homogeneity, and (b) simultaneous training. These assumptions also imply that the predictive performance of the distributed models evolves uniformly. A different approach is required since a distributed decentralized network is heterogeneous and non-stationary: nodes can join or leave the network at any point in time (churn). We propose a novel protocol for exchanging the model knowledge between peers using a gossip algorithm combined with the stochastic gradient descent (SGD). Our method has the advantage of being fully asynchronous, decentralized, trustless, and independent of the network size and the churn ratio. We validated the proposed algorithm by running network simulations in various scenarios.**

***Index Terms*—peer-to-peer network, gossip protocol, decentralized system, neural-network, machine learning**

I. Introduction

Deep neural networks have become the prominent approach in resolving classifications and predictions of unstructured and semi-structured data such as images [20], video [12], audio [21, 13] and time-series [9]. The velocity, volume and variety of available and fast changing data is making the process of training these deep neural networks a computational demanding task. Additionally, strong data privacy policies hinder this process by applying additional barriers in data sharing [25].

II. Background and Related work

Stochastic Gradient Descent (SGD), a variant of Gradient Descent (GD), is currently the most widely used optimization algorithm in deep neural network training. Models used in deep neural networks can be several hundreds of layers deep with millions of trainable parameters, creating a need for large amounts of computational resources. Distributed training algorithms are one way of tackling this challenge. Techniques such as distributed gradient optimization [27] and online learning with delayed updates [26] are just some of the methods utilized to train vast deep neural networks spanning several machines or efficiently utilizing several GPUs on a single machine [8]. Several authors propose different methods to scale deep learning algorithms. These approaches can be classified as *parameter server* based and *non-parameter server* based approaches [6]. A *parameter server* based algorithm is a learning algorithm where a central entity coordinates the learning process. The most used and widely spread parameter server algorithm is Federate Learning [5].

In this paper, we are focusing on the non-parameter server technique. There are methods [17, 2, 22, 15] that propose decentralized parallel training of connected workers. This proposed approach required ring topology [19] as a network connection structure between the workers, where each worker only communicates with its closest two neighbors. In this scenario, a worker exchanges its knowledge with its neighbors only once the synchronization barrier is met that ensures synchronous learning [17, 22, 15]. The alternative is the exchange of the model parameters with a random neighbor [2], but still, all these learning protocols are synchronous in exchanging their parameters.

GoSGD algorithm [4] uses a gossip averaging method in order to make different workers converge towards a consensus by exchanging information in a peer-to-peer manner. However, this method utilizes multiple GPUs on a single machine with identical performance, with data being separated evenly to each of the workers/threads. GossipGraD [6] modifies this algorithm by changing partner communication selection algorithm so that each worker communicates with only one partner during the next $log(n)$ steps of training, where n in the total number of peers. In [23], authors propose an approach for decentralized distributed training with data being available to all training nodes in the network using IPFS [3] decentralized filesystem and IOTA (see [1]) networking infrastructure. In [11], authors propose a model where agents connected in a peer-to-peer network are training local models. These agents are coordinated by a master node to prevent models in diverging from each other. Coordinator first pools all of the agent models, then aggregates them and sends them the final model back. Similar to this approach, the algorithm proposed in [10] trains agents in parallel, then collects their trainable variables and stores them in a global variable used for calculating updates for the optimizer.

In contrast to previous efforts, which are either synchronous or presume homogeneous nodes, in this paper, we propose an alternative technique named *PeerSGD*, which enables decentralized training in heterogeneous peer-to-peer networks without explicit data sharing.

III. PeerSGD Algorithm

We based our approach on GoSGD - gossip averaging algorithm proposed in [4] and consider a set *M* of independent nodes in a peer-to-peer network called agents. Each agent hosts a deep neural network model of the same architecture. We denote the sets of weights for a given agent as w_i. One can initialize the weights of these models with the same or a random value.

At each training iteration, all agents proceed with two steps, as depicted in algorithm 1 introduced in [4]: (1) gradient update, and (2) mixing update. At gradient update step, the agent updates its hosted network's weights with stochastic gradient descent on a random mini-batch. Mixing update step is a stage where the agent shares his knowledge with other agents in the network based on a random Bernoulli variable with expectancy *p* as proposed in [4]. At each mixing-update step, the agent can send its weights at most once but can receive several weights [4].

Our approach differentiates in proposed 2. When an agent receives weights, it can decide if he wants to accept the received weights based on the loss difference, making our approach trustless. We denote this maximum loss difference as wd. If the loss difference between local and received weights is within accepted loss diversity (dw), the agent updates its weights before performing any further gradient updates based on the algorithm. After receiving weights update, the agent performs weights averaging with local and received weights based on the losses calculated from a batch of training data — weight averaging favorites model with lower loss.

After weights averaging, the agent computes Kullback Leibler (KL) Divergence on initial predictions of both models to measure the match of the two models as proposed by [24] and used by [7]. With final loss calculated as the sum of the initial model loss and KL Divergence loss, the agent calculates and updates its model gradients.

Algorithm 1 Agents Pseudo-code

```
function TRAIN(p, w)
    p: probability of exchange
    w: initial weights (can be NULL)
    dw: diversity of accepted weights
    repeat
        TRAINONBATCH()

        if B(p) then
            j =RANDOM(M)
            SENDWEIGHTS(j)

    until Maximum iteration reached
```

The accuracy of the network is evaluated on a model created by averaging all of agents models weights as proposed by [4]:

$$w = \frac{1}{M} \sum_{i=1}^{M} W_i$$

Algorithm 2 Update Weights function

```
function UPDATEWEIGHTS(m_i, w_j)
    x, y ← NEXTTRAINBATCH()
    m_j ← MODEL(w_j)

    logits_i ← LOSS(m_i, x)
    logits_j ← LOSS(m_j, x)

    lossDiff ← ABS((loss_i − loss_j)/loss_i)
    if lossDiff > dw then return

    i ← 1 − loss_i/(loss_i + loss_j)
    j ← 1 − loss_j/(loss_i + loss_j)

    w_i ← w_i ∗ i + w_j ∗ j
    SETWEIGHTS(mi, w_i)

    loss ← loss_i+ KLLOSS(logits_i, logits_j)
    UPDATEGRADIENTS(m_i, loss)
```

Layer	Options
Conv2D	filters=28, kernel_size=(3,3), input_shape=(28, 28, 1)
MaxPooling2D	pool_size=(2, 2)
Flatten	
Dense	units=128, activation=relu
Dropout	rate=0.2
Dense	units=10, activation=softmax

TABLE I
NEURAL NETWORK LAYER STRUCTURE

IV. Evaluation and Results

We evaluate the proposed algorithm through various scenarios using different numbers of agents in the network on the MNIST [16] dataset of handwritten digits, which consists of 60000 training images and 10000 test images. We divided the data into batches, each batch containing 128 images. The probability of exchange was set to $p0.02$, and the parameter of acceptance $wd = 0.2$ in all simulations. We implemented our simulator using Tensorflow [18] in version 2.0. As an optimization method, we used Adam [14] optimizer with an initial learning rate set to 0.01. We also trained a single neural network as a baseline for evaluation using identical properties as the agents with a learning rate set to 0.001 for better convergence.

A. Simulation scenarios

Simulation 1. In this scenario, we simulated learning and interaction between 10 and 100 agents, starting with identical or random weights. In this experiment, all agents have the same computing power. The results are shown in figure 1. In this scenario, each agent received four batches of training data.

Simulation 2. This scenario is similar to the one before; the only difference being that in this scenario, each agent received a random number of training data batches in the range [1, 7]. The results are shown in figure 2.

Simulation 3. This scenario extends the one before but modifies the simulated computing power of agents. Agents with a greater amount of training data learn faster than the agents with a smaller amount of training data. Faster learning agents consequently gossip more messages to the network. The results are shown in figure 3.

Simulation 4. In this scenario, we simulated agents churn, that is, agents joining and leaving the network. Each agent was given a random number of training data batches in the range [1, 7]. When an agent is inactive, it does not train nor can receive weight updates by other agents who are active on the network. As a probability of an agent to leave or join the network, we used the same function as used to determine if an agent exchanges its weights with other agents with probability set to 0.02. The results are shown in figure 4.

Simulation 5. For our final test, we wanted to test model robustness. In this simulation, we trained the models for 1000 epochs, which results in 468750 batches iteration for all agents in total. Each agent received equal portion of data with all agents in total having access to all training data. In the ten agent scenario, each agent received 47 batches of training data, while in a 100 agent scenario, each agent received five batches of training data. In order to test the robustness of the networks, we augmented the data by adding 5000 new augmented images into the test set with the following augmentation parameters:

- Rotation range: 10%
- Zoom range = 5%
- Width shift range = 5%
- Height shift range = 5%

The results with accuracies on initial test data and augmented test data are shown in table III.

B. Results

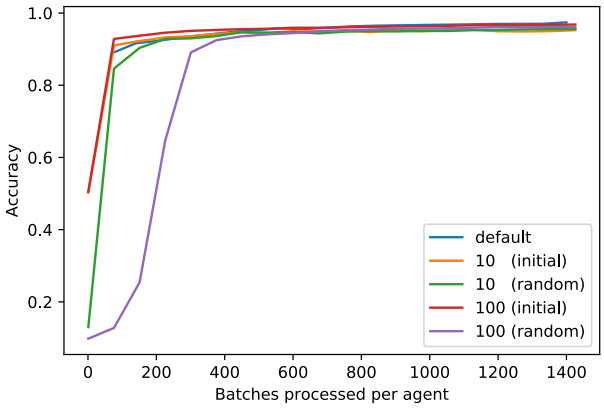


Fig. 1. Simulation 1: Convergence on test data between 10 and 100 agents each having 4 batches of training data with weights being identical (initial) or random (random) for all agents with default network trained on all training data as a reference

Based on simulation results, we observe that agents joining the network can contribute to the overall accuracy of the network, provided they have unique datasets they can use for training. The dataset size distribution over agents does not impact the overall network accuracy. When computing power is dependent on the size of available data, we observe no significant impact on model accuracies, while it has only a small impact on smaller network accuracies (around 0,50%). When the network is impacted by high churn rate we observe that choosing identical starting weights for all agents offers better convergence. When testing the robustness, all agents were trained until the total 468750 (i.e. 1000 epoch over training data) of batch iterations was reached. We compared the results with the default network. While the default network has the highest accuracy, we observe that networks with higher numbers of agents provide more robustness (around 3% in our experiment).

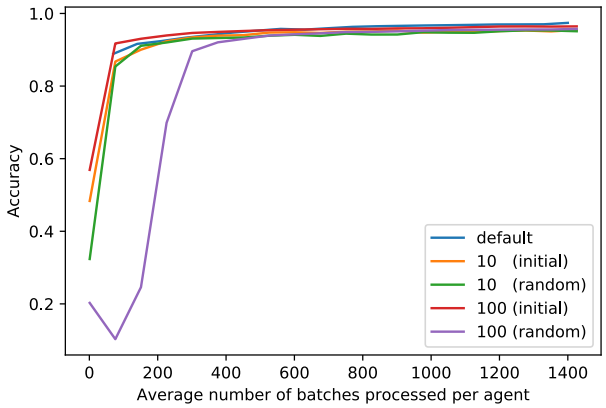


Fig. 2. Simulation 2: Convergence on test data between 10 and 100 agents each having random [1, 7] batches of training data with weights being identical (initial) or random (random) for all agents with default network trained on all training data as a reference

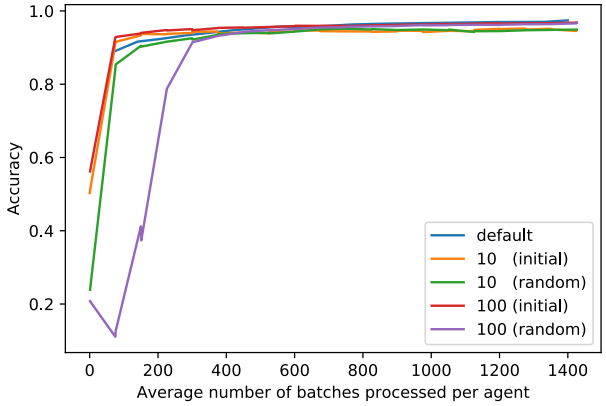


Fig. 3. Simulation 3: Convergence on test data between 10 and 100 agents each having random [1, 7] batches of training data with weights being identical (initial) or random (random) for all agents with computing power of agents based of amount of their data

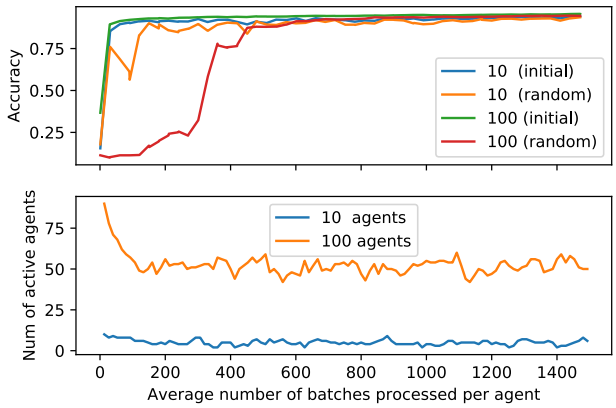


Fig. 4. Simulation 4: Convergence on test data between 10 and 100 agents each having random [1, 7] batches of training data with weights being identical (initial) or random (random) for all agents with agents being active for 50% of the time

Agents	Weights	Accuracy
1		97.41
Simulation 1		
10	initial	95.27
10	random	95.53
100	initial	96.82
100	random	96.11
Simulation 2		
10	initial	95.47
10	random	95.10
100	initial	96.44
100	random	95.74
Simulation 3		
10	initial	94.80
10	random	94.90
100	initial	96.85
100	random	96.70
Simulation 4		
10	initial	94.10
10	random	93.65
100	initial	95.64
100	random	94.50

TABLE II
SIMULATION ACCURACY

Agents	Weights	Test acc	Augmented acc
1		98.72	84.83
10	initial	98.10	85.20
10	random	97.97	86.55
100	initial	97.72	88.21
100	random	97.72	87.52

TABLE III
TESTING THE ROBUSTNESS OF THE NETWORKS ON AUGMENTED TEST DATA

C. Limitations

We found that when the data classes are segregated across multiple agents, the network cannot achieve convergence. To demonstrate that, we segregated MNIST dataset by numbers across ten agents, each of them receiving data for just one number, respectively. Overall, network accuracy was around 10%, which is equivalent to a random guess. One way of tacking this problem is by introducing similar approach as taken in [22]. Furthermore, our approach requires that all agents use identical structure and number of neural network layers, which might not be possible for all agents in the network.

V. CONCLUSION

In this paper, we have introduced PeerSGD, a Stochastic Gradient Descent (SGD) algorithm for scaling deep learning algorithms in peer-to-peer networks using the gossip communication protocol. We have evaluated our approach by simulating different scenarios and network sizes. Our approach is fully asynchronous, decentralized, and does not require data sharing between peers. Our approach performs well in heterogeneous environments with different computing powers and different amounts of local data being available to agents. We have shown that bigger networks provide more robust network models with better churn resistance. Further research should focus on eliminating the requirement of an identical network structure across agents. Each agent should have a choice in choosing the best network structure based on his requirements, performance properties, and amount of available data.

REFERENCES

[1] Roman Alexander. *IOTA - Introduction to the Tangle Technology: Everything You Need to Know about the Revolutionary Blockchain Alternative*. Independently published, 2018. ISBN: 198041887X.

[2] "Asynchronous decentralized parallel stochastic gradient descent". In: *35th International Conference on Machine Learning, ICML 2018* 7 (2018), pp. 4745–4767.

[3] Juan Benet. "IPFS - Content Addressed, Versioned, P2P File System". In: *CoRR* abs/1407.3561 (2014). arXiv: 1407.3561. URL: http://arxiv.org/abs/1407.3561.

[4] Michael Blot et al. "Distributed optimization for deep learning with gossip exchange". In: *Neurocomputing* 330.April (2019), pp. 287–296. ISSN: 18728286. DOI: 10.1016/j.neucom.2018.11.002. arXiv: arXiv:1804.01852v1.

[5] Keith Bonawitz et al. "Towards Federated Learning at Scale: System Design". In: (2019). arXiv: 1902.01046. URL: http://arxiv.org/abs/1902.01046.

[6] Jeff Daily et al. "GossipGraD: Scalable Deep Learning using Gossip Communication based Asynchronous Gradient Descent". In: (2018). arXiv: 1803.05880. URL: http://arxiv.org/abs/1803.05880.

[7] Kai Gao et al. “Deep mutual learning for visual tracking”. In: 2019, pp. 1–5. ISBN: 9781450371582. DOI: 10.1145/3321408.3323927.

[8] Otkrist Gupta and Ramesh Raskar. “Distributed learning of deep neural network over multiple agents”. In: *Journal of Network and Computer Applications* 116 (2018), pp. 1–8. ISSN: 10958592. DOI: 10.1016/j.jnca.2018.05.003. arXiv: 1810.06060.

[9] Hassan Ismail Fawaz et al. “Deep learning for time series classification: a review”. In: *Data Mining and Knowledge Discovery* (Mar. 2019). DOI: 10.1007/s10618-019-00619-1.

[10] Zhanhong Jiang et al. “Collaborative deep learning in fixed topology networks”. In: *Advances in Neural Information Processing Systems* 2017-Decem.Nips (2017), pp. 5905–5915. ISSN: 10495258. arXiv: 1706.07880.

[11] Michael Kamp et al. “Efficient decentralized deep learning by dynamic model averaging”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11051 LNAI.i (2019), pp. 393–409. ISSN: 16113349. DOI: 10.1007/978-3-030-10925-7_24. arXiv: 1807.03210.

[12] Andrej Karpathy et al. “Large-scale Video Classification with Convolutional Neural Networks”. In: *Proceedings of International Computer Vision and Pattern Recognition (CVPR 2014)*. 2014.

[13] Aditya Khamparia et al. “Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network”. In: *IEEE Access* PP (Jan. 2019), pp. 1–1. DOI: 10.1109/ACCESS.2018.2888882.

[14] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *CoRR* abs/1412.6980 (2014).

[15] Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. “Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication”. In: (2019). arXiv: 1902.00340. URL: http://arxiv.org/abs/1902.00340.

[16] Yann LeCun and Corinna Cortes. “MNIST handwritten digit database”. In: (2010). URL: http://yann.lecun.com/exdb/mnist/.

[17] Xiangru Lian et al. “Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent”. In: *Advances in Neural Information Processing Systems* 2017-Decem (2017), pp. 5331–5341. ISSN: 10495258. arXiv: 1705.09056.

[18] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[19] Sape Mullender et al. *Distributed systems*. Vol. 12. acm press United States of America, 1993.

[20] Farhana Sultana, A. Sufian, and Paramartha Dutta. “Advancements in Image Classification using Convolutional Neural Network”. In: (May 2019).

[21] Naoya Takahashi, Michael Gygli, and Luc J. Van Gool. “AENet: Learning Deep Audio Features for Video Analysis”. In: *IEEE Transactions on Multimedia* 20 (2017), pp. 513–524.

[22] Hanlin Tang et al. “D^2: Decentralized Training over Decentralized Data”. In: *Proceedings of the 35th International Conference on Machine Learning* 80.1 (2018), pp. 4848–4856. URL: http://proceedings.mlr.press/v80/tang18a.html.

[23] Shibo Yao. *Signal2Vec : Time Series Embedding*. December 2016. Springer International Publishing, 2019, pp. 80–90. ISBN: 9783030202576. DOI: 10.1007/978-3-030-20257-6. URL: http://dx.doi.org/10.1007/978-3-030-20257-6%7B%5C_%7D7.

[24] Ying Zhang et al. “Deep Mutual Learning”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2018), pp. 4320–4328. ISSN: 10636919. DOI: 10.1109/CVPR.2018.00454. arXiv: 1706.00384. URL: https://github.com/YingZhangDUT/Deep-Mutual-Learning.

[25] Jianxin Zhao et al. “Privacy-preserving machine learning based data analytics on edge devices”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 341–346.

[26] Shuxin Zheng et al. “Asynchronous Stochastic Gradient Descent with Delay Compensation”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. JMLR.org, 2017, pp. 4120–4129.

[27] Martin Zinkevich et al. “Parallelized Stochastic Gradient Descent”. In: *Advances in Neural Information Processing Systems 23*. Ed. by J. D. Lafferty et al. Curran Associates, Inc., 2010, pp. 2595–2603. URL: http://papers.nips.cc/paper/4006-parallelized-stochastic-gradient-descent.pdf.

Discrete Bat Algorithm for Event Planning Optimization

S. Delalić[1], A. Alihodžić[1], M. Tuba[2], E. Selmanović[1], D. Hasić[1]
[1] Faculty of Science, University of Sarajevo, Bosnia and Herzegovina
[2] Singidunum University, Belgrade, Serbia
delalic.sead@pmf.unsa.ba, adis.alihodzic@pmf.unsa.ba, tuba@ieee.org, eselmanovic@pmf.unsa.ba, damir.hasic@pmf.unsa.ba

***Abstract* - Many public figures, companies and associations are planning events in different cities and at the same time have active profiles on social media. The planning process requires processing a large amount of data and different parameters when choosing the best event venue. Social media captures a large number of fan actions per day. This paper describes the process of selecting the most appropriate cities to organize events, aided by data collected from social media. The problem is defined as a combinatorial optimization problem. A modified metaheuristic Bat algorithm was proposed, implemented, and described in detail to solve the problem. Although the original Bat algorithm is designed to solve continuous optimization problems, the implemented bat algorithm is adapted to solve the defined problem. The algorithm is compared to the exhaustive search method for smaller instances, and to the greedy and genetic algorithm for larger instances. The algorithm was tested on benchmark data on cities in 20 European countries, as well as on real data collected from pages on the social network Facebook. Bat algorithm has shown superior results compared to other techniques, both in time and in the quality of the solutions generated.**

Keywords - Bat Algorithm; Discrete Bat Algorithm; Event Planning; Metaheuristic

I. INTRODUCTION

The development of the Internet and information technology has changed many common human practices. Software development has simplified many processes. One of the most significant changes is the ability to easily collect and store large amounts of data. Each process can be easily monitored through information systems and online platforms. Occasionally, information for which there is no current use is also stored, but there are assumptions that this information may contain useful information for the future.

In today's world, many jobs in various fields are handled through online platforms, and many public figures and organizations have profiles on social media and online platforms. This makes it clear that a large number of different data is being collected, many of which remain unused, although it can make a significant contribution to understanding and improving business processes.

Due to the large amount of data collected, a complete area of data mining has been developed, which attempts to leverage the data and find important information to improve the business processes.

As the security of the collected user data is of great importance, data should be used rationally and without compromising user privacy. Therefore, it is preferable to use data that cannot reveal the identity of users, and some of the most important such data is accumulated data, such as data on the number of users from a city.

Public organizations and public figures organize events in different cities. Musicians organize concerts, sports clubs organize friendly matches and meetings with supporters, political and other associations organize meetings of members, etc. The choice of cities is most often guided by the experience and judgment of the managers. Therefore, it is important to utilize the data available from social networks and online platforms to enhance the organization process and the selection of

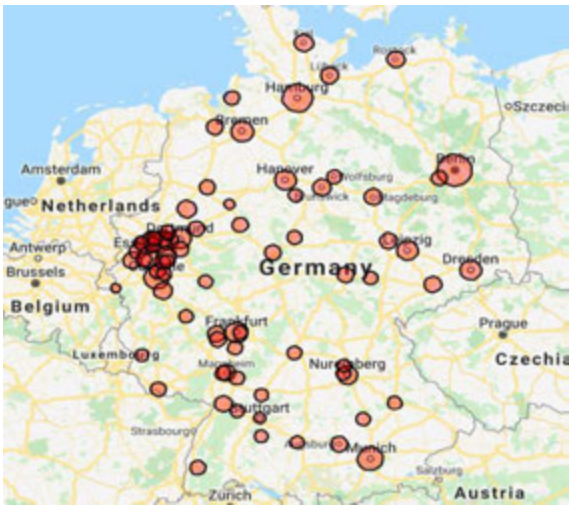


Figure 1. Visualization of users in Germany

cities and venues for event organization. A visualization of the number of fans/users on the example of cities in Germany is shown in Figure 1. The corresponding number of fans is presented for 79 cities in Germany. Every popular social network, like Facebook, offers access to similar reports to its users and site owners.

The problem can be modeled as a combinatorial optimization problem. The goal is to maximize the total

number of visitors, while respecting the fact that the number of visitors in a city depends on the number of potential visitors from other neighboring cities where events are not held.

The problem of choosing cities in their general form is a hard optimization problem, and the solution space is impossible to search in real time. Therefore, heuristic methods are applied that produce good enough results in a short enough time. The paper describes the implementation and usage of the modern metaheuristic Bat algorithm for solving the optimal event-planning and city selection problem.

The paper consists of five sections. In the current section, a brief introduction, motivation, and importance for observing the problem are given. In the second section, a literature review is provided. In the third section, a detailed description of the Bat algorithm implementation on the problem of city selection is given. In the fourth section, the obtained results are described and a comparison of the implemented algorithm with the greedy and the genetic algorithm is stated. In the last section, a brief conclusion is given and a plan for future work is outlined.

II. Related Work

In [1], the problem of event planning was observed based on data collected from online platforms. The aforementioned problem was modeled in the form of the combinatorial optimization problem, and solutions with a greedy algorithm and an adjusted genetic algorithm were implemented. In addition, a benchmark dataset was created based on cities from 20 European countries. The algorithms were compared and the results are described in detail. This paper was an inspiration for the implementation of another approach that will improve the previously obtained results.

Similar research has been conducted in the field of music industry and concert planning. In [2], the problem of concerts planning is described, which in addition to the number of visitors includes the traveling cost between cities. In addition to the process of selecting cities, the algorithm also creates a concert schedule, minimizing the total length of the route. A simulated annealing algorithm and a genetic algorithm were implemented to solve the problem of planning concerts.

Process optimizations are standard tasks in many areas of the industry. Information systems and online platforms often contain algorithms to optimize specific processes. Optimization algorithms are most commonly used to optimize the most expensive or common processes, as they result in the greatest savings. Optimization algorithms are used in the field of transport (Vehicle Routing Problem [3]), warehousing [4], online commerce and recommendations systems [5], job scheduling [6], etc. There are a number of examples where previously collected data is used for process improvements and analyzes, such as anomaly detection in GPS records [7] or improvements of the vehicle routing algorithms using GPS data [8].

In the process of solving complex problems, advanced metaheuristic algorithms are used. In addition to the aforementioned genetic algorithm and simulated annealing, there are a number of modern metaheuristic approaches.

A special group of optimization algorithms is a group of collective intelligence algorithms. These algorithms are most commonly inspired by organisms from nature and their specific behavior. In collective intelligence algorithms, each individual possesses a certain intelligence; however, the real power of the algorithm is in the collaboration of the individuals. In [9], the capabilities of the collective intelligence algorithm are described and the advantages and disadvantages of some of the most well-known algorithms are outlined. Some of the known algorithms are Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Elephant Herding Optimization (EHO), Bat algorithm (BA), Firefly algorithm (FA) etc. All the above algorithms have been used to solve optimization problems in various fields of science and industry.

This paper focuses primarily on the use of the Bat algorithm. The Bat algorithm was developed in 2010 by Xin-She Yang [10]. In [11], many applications of the Bat algorithm were given. The bat algorithm is primarily intended for continuous optimization problems. However, the algorithm has been implemented on a number of problems from combinatorial optimization. In [12], the Bat algorithm was proposed to solve the problem of a symmetric and asymmetric traveling salesman. The paper describes how to adapt the Bat algorithm to solve problems in a discrete space, and proposes a pseudocode for the discrete Bat algorithm.

The Bat algorithm was also used for the problem of handwritten digit recognition in a combination with support vector machines [13], and for the problem of multilevel image thresholding [14]. In addition to applications in the field of image processing, the Bat algorithm was used to solve vehicle routing problems and other problems finding optimal routes ([15], [16]). The Bat algorithm was also used to find the optimal problem of permutation flow shop scheduling [17], as well as the well-known multidimensional Knapsack problem [18].

III. Case Study

As mentioned in the previous section, the problem of selecting cities for event organization is modeled as an optimization problem in [1]. It is necessary to collect and define several data for the successful implementation.

- Let $C = \{c_1, c_2, \dots, c_n\}$ be the set of observed cities.
- Let p_i be the population for each city c_i, where $i = \overline{1,n}$.
- Let $d_{i,j}$ be the distance between the cities c_i and c_j, for all $i = \overline{1,n}, j = \overline{1,n}$. The distance between cities can be calculated in many ways. The simplest way is to calculate the air distance between centers. In addition, it is possible to use online services that calculate the road and time

distances between geographical coordinates. The implementation of the algorithm does not change regardless of how all distances are calculated.

This paper describes the implementation and use of the Bat algorithm to solve the modeled problem. As mentioned earlier, this algorithm was introduced in 2010 [10]. The algorithm is inspired by the behavior of microbats in nature. Microbats have an advanced echolocation capability and use it to detect prey and avoid obstacles. The BA is based on simplified echolocation rules:

- All bats know the difference between pray and obstacles;
- Bats are searching for pray and flying randomly with velocity v_i at position x_i. Their frequency is fixed to f_{min}, and wavelength λ and loudness A_0 are varying. They can adjust the wavelength and the pulse rate depending on their pray position;
- The assumption is that the loudness varies from a large A_0 to a minimum constant value A_{min}.

The bat algorithm is primarily intended to solve problems in the field of continuous optimization. The observed problem is modeled as a combinatorial optimization problem, which means that the solutions are located in a discrete space. As mentioned in the related work, the bat algorithm was adjusted to solve a discrete TSP problem in [12]. The guidelines outlined in this paper are used to solve the problem of event planning. A modified pseudocode of the Bat algorithm tailored to solve city selection problem is given in Algorithm 1.

Algorithm 1 The modified BA for Optimal City Selection Problem

```
function batAlgorithm
begin
   Define the objective function f_GOAL(x)
   Initialize the bat population X = x_1, x_2, ..., x_n
   Find the best bat in population x_best

   for each bat x_i in the population do
      Initialize the pulse rate r_i, velocity v_i and loudness A_i
   end

   repeat
    for each bat x_i in the population do
      Generate new solution
      Generate random number rand
      if rand > r_i then
        Generate solution around the best bat
      end

      if rand < A_i and f_GOAL(x_i) > f_GOAL(x_best) then
        Accept the new solution
        Increase r_i
        Reduce A_i
      end
end
```

First step of the successful implementation of the Bat algorithm is the definition of the objective function f_{GOAL}. The objective function is defined as in [1], because this makes it easier to compare the performance of implemented algorithms. It is important to note that changes in the goal function are simple, and to add real-world constraints it is sufficient to modify the goal function and add penalties if constraints are not met.

The objective function is created by estimating the total number of visitors for the given selection of cities. The goal function assumes that all visitors from selected cities will attend the event; however, at the same time it is necessary to add visitors from nearby non-selected cities.

It is likely to assume that guests from a non-selected city will attend an event at the nearest venue. At the same time, we can assume that not all visitors will attend the event unless it is held in their city.

Let X be the set of all selected cities. Let Y be the set of all non-selected cities. Then the union of sets X and Y is equal to C. The $b_{j,m}$ values are defined as 1 for if the city c_j is closest to the selected city c_m. Otherwise, the value is 0. The objective function is then calculated as

$$f_{GOAL} = \sum_{c_i \in X} p_i + \sum_{j \in X, m \in Y} x_{j,m} \cdot p_m \cdot b_{j,m},$$

where $x_{j,m}$ is the correlation factor between cities.

The correlation factor $x_{j,m}$ represents the probability that a person from city c_j will attend an event in city c_m. The factor can be calculated in many ways. Distance is the most important parameter for calculating it. The formula for the correlation factor given in [1] was also used in for easier comparison of obtained results:

$$x_{i,j} = \begin{cases} 1, & 0 \le d_{i,j} < 5 \\ 0.95 & 5 \le d_{i,j} < 10 \\ 0.6 & 10 \le d_{i,j} < 20 \\ 0.3, & 20 \le d_{i,j} < 50 \\ 0.02 & 100 \le d_{i,j} < 200 \\ 0 & 200 \le d_{i,j} \end{cases}$$

Once the objective function has been successfully defined, it is necessary to initialize the bat population. Each bat in the population represents one solution, in other words, one choice of cities to host an event. The initial bat population was selected to maximize the search space, but at the same time to reject irrational solutions. Each solution in the initial population is created by choosing the largest cities, with some variance. If it is necessary to select k cities, then the set of $c \cdot k$ largest cities is created, and one combination from the set is chosen for each bat.

The best bat in the population is the solution for which most visitors are guaranteed. This bat represents the bat with the highest value of the goal function

$$f_{GOAL}(x_{best}) \ge f_{GOAL}(x_i), \forall i = \overline{1,n}.$$

In the next step, it is necessary to define the pulse rate r_i, velocity v_i and the loudness A_i, for each bat in the population. The pulse rate r_i is chosen randomly from the segment [0, 0.4]. The loudness A_i is chosen also randomly from the segment [0.7, 1]. The velocity of each bat depends on the distance from the best bat in the population. The distance between two bats is calculated as

the number of cities selected in one and not selected in the other bat.

In the main part of the algorithm, it is necessary to define a function to generate a new solution from an existing one. This feature is implemented in such a way that when creating a new solution, it selects a random selected city and selects a random non-selected city. The previously selected city is not in the new solution, while the previously non-selected city is in the new solution. In simpler terms, a randomly chosen selected city and a randomly chosen non-selected city change their roles.

The process of increasing r_i and reducing A_i is given by standard formulas, where new r_i is calculated as

$$r = r \cdot (1 - e^{-\gamma \cdot I}),$$

while new A_i is calculated as $A_i = \alpha A_i$. In the previous formulas, I represents the iteration number, α and γ represent constants.



Figure 2. Visualization of selected cities in Croatia

IV. RESULTS

In the literature, the problem described in this paper has only been found in [1]. Apart from the datasets for the 20 European countries defined in the aforementioned paper, no datasets were found for testing. In this section, a comparison of the implemented discrete Bat algorithm with the previously implemented and used methods for a given problem will be stated: an exhaustive search for smaller instances, a greedy algorithm and a genetic algorithm.

The dataset contains the names of cities from European countries along with their populations. The disadvantage of the datasheet was noted, as it was dependent on geolocation online services that obtained the geographic coordinates of the center based on the name of the city. Therefore, different tests can give different results, because the same service is not used. Therefore, a modification of the datasheet was made, so that each city is presented by four values: city name, latitude and longitude of the center and the city population. Each file stores a list of JSON objects with the specified data. After collecting the coordinates, several changes occurred compared to earlier datasets. The dataset is available in [19]. Graphical visualization of the collected dataset and selected cities for Croatia is given in Figure 2, where six of more than 100 different cities were selected as event locations. Number six is an arbitrarily chosen constant suitable for testing. Cities were selected by maximizing the value of the goal function described in the previous section.

TABLE I. RESULTS ON THE EUROPEAN COUNTRIES DATASET

Country	Number of cities	Number of selected cities	Greedy approach	Genetic algorithm (best)	Bat algorithm (best)
Albania	74	6	1756903.56	1756903.56	1756903.56
Austria	77	6	3089728.96	3105748.76	3117662.95
Belarus	100	6	3948980.48	3948980.48	3948980.48
Belgium	100	6	2613593.5	2690237.8	2690237.8
Bosnia and Herzegovina	48	6	1000334.9	1009116.5	1009116.5
Croatia	127	6	1634131.47	1634131.47	1634131.47
	100	6	1618533.45	1618533.45	1618533.45
Czech Republic	21	6	2575935.38	2594651.78	2594651.78
Denmark	100	6	2463799.25	2551272.95	2551272.95
Finland	100	6	2105270.49	2135607.9	2137304.19
France	434	6	10226615.6	10098110.2	10325996.31
	100	6	6891700.3	6930376.18	7024254.98
Germany	79	6	10901545.76	11528898.46	11554187.86
Italy	144	6	8517525.84	8566803.56	8566803.56
	100	6	8014905.54	8033536.1	8033536.1
Montenegro	13	6	308485.3	316925.6	316925.6
North Macedonia	5	3	672191.8	672191.8	672191.8
Norway	47	6	1596130.31	1596130.31	1596130.31
Serbia	100	6	4738269.6	4738269.6	4738269.6
Slovenia	16	6	562255.2	568610.7	568610.7
Spain	717	6	17099382.9	16720003.19	17659128.45
	100	6	13192605.7	13214860	13296928.28
Sweden	61	6	3148095.68	3231208.3	3231208.3
Turkey	580	6	27715396.5	26714344.5	27715396.45
	100	6	27356808.9	26777738.7	27356808.9

The dataset consists of smaller countries, such as Slovenia, Montenegro and Northern Macedonia, with less than 20 cities listed, middle countries like Bosnia and Herzegovina with 48, Germany with 79 and Croatia with 127 cities, and large countries such as France with 434, Turkey with 580 and Spain with more than 700 cities.

Distance between cities is calculated using the well-known haversine formula for calculating distances between two points on the sphere.

Test results are listed in Table 1. The table shows the comparison of the algorithms for the 20 countries in the observed dataset. For countries with larger number of cities, two tests were performed. The first test uses all the cities in the list, while the second test is done for the first 100 cities. Six cities are selected for each test, while in Northern Macedonia, which contains only five cities, three cities have been selected.

The parameters for the Bat algorithm are set to the following values:

- Population size is set to 20.
- Initial loudness is set as a random number in segment [0.7, 1]. Initial pulse rate value is set as a random number in segment [0, 0.4].
- α and γ are constant values 0.98.
- Maximal number of iterations is set to 1000.

The genetic algorithm was tested also on 1000 iterations. Genetic and Bat algorithm were executed 10 times for each instance of the dataset. The best and the average value were taken into consideration. Only best results are shown.

As can be seen, the Bat algorithm produced the best results for all countries. In some of them, genetic algorithm and greedy approach gave the same results. Each instance was tested smaller number of cities (15 cities, 6 selected). The Bat algorithm gave the same results as the exhaustive search algorithm, which confirms the successful application of the bat algorithm to the observed problem. During the testing phase, it was observed that a genetic algorithm requires approximately 20% more iterations to reach at the best solution than the Bat algorithm. The BA showed extremely small variations in the results of individual testing (only for large instances, such as Turkey). For the genetic algorithm, larger deviations were observed in the individual tests.

The bat algorithm gave better results than the genetic algorithm in six of the 20 instances tested (30%). In other instances, the algorithms gave the same results. The BA has shown better performance on instances with more cities. Greedy algorithm gave equal results to the genetic and bat algorithm for some instances. However, there are a number of cases where this algorithm does not produce the best result. These cases occur for clusters of cities, where all cities within a group may be smaller, but their number increases the cost-effectiveness of choosing one of them.

The algorithm has been tested on multiple instances collected from Facebook pages. Facebook offer a large number of reports and detailed information about the users who visit the page. They are only available to page owners and administrators. One of these reports is a set of pairs where for each city, a number of Facebook page fans from that city are assigned.

In practice, several important limitations significantly affect the results. When selecting cities, it is necessary to provide a space where the event can be organized. Often, some cities are excluded. Such a case occurs if a similar event in the city has taken place in the recent past. All these constraints can be easily included in the definition of a goal function, and do not change the implementation of the bat algorithm.

The algorithm was tested on data from multiple real Facebook pages. Additional restrictions were included to represent the time of organization of the last event in that city, the exclusion of certain cities from observation and the capacity of the venue. Restriction data was manually collected. The bat algorithm was run over the same settings, and it gave significantly better results compared to other approaches. Each of these constraints is simply included by modifying the goal function. Violating each condition adds a penalty value to the function. No additional modifications to the algorithm were made.

V. CONLUSION

The observed problem is an interesting example of the combination of modern technologies and the advantages of the modern world with an algorithmic approach to problem solving. The paper demonstrates the use of the Bat algorithm in solving a practical problem that can improve business processes. The algorithm was tested on data containing cities from 20 cities in European countries. The Bat algorithm has been compared to a greedy algorithm and a genetic algorithm. The Bat algorithm has been shown to produce significantly better results than other methods.

In a future research, additional restrictions will be added, such as the cost of transport between cities and the cost of renting a venue. Other modern metaheuristic algorithms will be implemented.

The idea and concept of using data collected from online platforms and information systems to optimize the event planning process can be of great importance. Apart from organizing events, a similar problem can be formulated to solve a number of other practical problems, such as determining a city to expand a business or building a new office, shop, etc. It can also bring great savings and increase planning efficiency and speed, and is therefore significant for research.

REFERENCES

[1] Delalić, S., Alihodžić, A., & Selmanović, E. (2019, October). Innovative Usage of Online Platforms Analytics on Event Planning based on the Genetic Algorithm with Greedy Approach. In *2019 XXVII International Conference on Information, Communication and Automation Technologies (ICAT)* (pp. 1-6). IEEE.

[2] Delalić, S., Chahin, M., & Alihodžić, A. (2019, October). Optimal City Selection and Concert Tour Planning Based on Heuristic Optimization Methods and the Use of Social Media Analytics.

In *2019 XXVII International Conference on Information, Communication and Automation Technologies (ICAT)* (pp. 1-6). IEEE.

[3] Demir, E., Huckle, K., Syntetos, A., Lahy, A., & Wilson, M. (2019). Vehicle Routing Problem: Past and Future. In *Contemporary Operations and Logistics* (pp. 97-117). Palgrave Macmillan, Cham.

[4] Žunić, E., Delalić, S., Hodžić, K., Beširević, A., & Hindija, H. (2018, November). Smart Warehouse Management System Concept with Implementation. In *2018 14th Symposium on Neural Networks and Applications (NEUREL)* (pp. 1-5). IEEE.

[5] Bobadilla, J., Ortega, F., Hernando, A., & Alcalá, J. (2011). Improving collaborative filtering recommender system results and performance using genetic algorithms. *Knowledge-based systems*, *24*(8), 1310-1316.

[6] Lin, T. L., Horng, S. J., Kao, T. W., Chen, Y. H., Run, R. S., Chen, R. J., ... & Kuo, I. H. (2010). An efficient job-shop scheduling algorithm based on particle swarm optimization. *Expert Systems with Applications*, *37*(3), 2629-2636.

[7] Žunić, E., Delalić, S., Hodžić, K., & Tucaković, Z. (2019, July). Innovative GPS Data Anomaly Detection Algorithm inspired by QRS Complex Detection Algorithms in ECG Signals. In *IEEE EUROCON 2019-18th International Conference on Smart Technologies* (pp. 1-6). IEEE.

[8] Žunić, E., Hindija, H., Beširević, A., Hodžić, K., & Delalić, S. (2018, November). Improving Performance of Vehicle Routing Algorithms using GPS Data. In *2018 14th Symposium on Neural Networks and Applications (NEUREL)* (pp. 1-4). IEEE.

[9] Yang, X. S. (2010). *Nature-inspired metaheuristic algorithms*. Luniver press.

[10] Yang, X. S. (2010). A new metaheuristic bat-inspired algorithm. In *Nature inspired cooperative strategies for optimization (NICSO 2010)* (pp. 65-74). Springer, Berlin, Heidelberg.

[11] Yang, X. S. (2013). Bat algorithm: literature review and applications. *arXiv preprint arXiv:1308.3900*.

[12] Osaba, E., Yang, X. S., Diaz, F., Lopez-Garcia, P., & Carballedo, R. (2016). An improved discrete bat algorithm for symmetric and asymmetric traveling salesman problems. *Engineering Applications of Artificial Intelligence*, *48*, 59-71.

[13] Tuba, E., Tuba, M., & Simian, D. (2016). Handwritten digit recognition by support vector machine optimized by bat algorithm.

[14] Alihodzic, A., & Tuba, M. (2014). Improved bat algorithm applied to multilevel image thresholding. *The Scientific World Journal*, *2014*.

[15] Osaba, E., Carballedo, R., Yang, X. S., Fister Jr, I., Lopez-Garcia, P., & Del Ser, J. (2018). On efficiently solving the vehicle routing problem with time windows using the bat algorithm with random reinsertion operators. In *Nature-Inspired Algorithms and Applied Optimization* (pp. 69-89). Springer, Cham.

[16] Osaba, E., Yang, X. S., Fister Jr, I., Del Ser, J., Lopez-Garcia, P., & Vazquez-Pardavila, A. J. (2019). A discrete and improved bat algorithm for solving a medical goods distribution problem with pharmacological waste collection. *Swarm and evolutionary computation*, *44*, 273-286.

[17] Luo, Q., Zhou, Y., Xie, J., Ma, M., & Li, L. (2014). Discrete bat algorithm for optimal problem of permutation flow shop scheduling. *The Scientific World Journal*, *2014*.

[18] Sabba, S., & Chikhi, S. (2014). A discrete binary version of bat algorithm for multidimensional knapsack problem. *international journal of bio-inspired computation*, *6*(2), 140-152.

[19] Delalić S, Alihodžić A, Tuba M, Selmanović E, Hasić D. European city repository. URL: bitbucket.org/delalics/cities/master/; 2020.

A Conceptual Network Analysis of Gamification Practices in Primary and Secondary Education

Igor Tomičić, Markus Schatten
Artificial Intelligence Laboratory, Center for Forensics, Biometrics and Privacy
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42000 Varaždin, Croatia
Email: (igor.tomicic, markus.schatten)@foi.hr

***Abstract*— Gamification is not a new concept, but it has gained a significant momentum in the past years for several reasons, including but not limited to the ubiquity of technology, the growing ease in the use of the game design applications and the growing number of technologically savvy individuals. Such individuals are routinely using smartphones and other computing devices for various tasks, including playing games and using e-learning platforms. Gamification is mostly defined as the use of game design elements and game mechanics in a non-game context, with the main objective of engaging and motivating users. The existing research points to the possibility of improving the performance of students in the learning process, and this paper presents a state of the art, conceptual network analysis of gamification practices in learning processes with the aim to better understand game-based learning in primary and secondary education. The results from this paper would be used for the development of a novel learning platform based on artificial intelligence techniques.**

Keywords*— *gamification, concaptual network analysis, primary education, secondary education

I. Introduction

Gamification is defined by Deterding et al. [1] as the use of game design elements in non-game contexts, and is generally considered as a fairly rapidly growing field.

Whilst gamification can be used in almost any social context, in education some of the main research questions according to [2] are: *1) What educational contexts does gamification apply to, and 2) What game elements are used in gamification of educational systems?* For game elements in gamification, authors propose the principles of gamification design (goals, challenges and quests, customizations, progress monitoring, feedback, competition and engagement, social aspects and engagement, visible status, content unlocking, freedom of choice, freedom of failure, storytelling, etc.) and game mechanics (points, badges, levels, flowcharts, leaderboards, virtual currencies, avatars). The authors emphasize that there are limitations in research efforts which deal with the effectiveness of the use of game elements in an educational environment, and that the development of software tools to support gamification in an educational context would positively contribute to such research efforts in the field. Figures 1 and 2 respectively represent an overview of the number of papers that deal with individual elements of game mechanics (showing which elements are the most popular), and principles of gamification design. "Engagement" as a concept appeared with relatively strong connections within our conceptual network analysis.

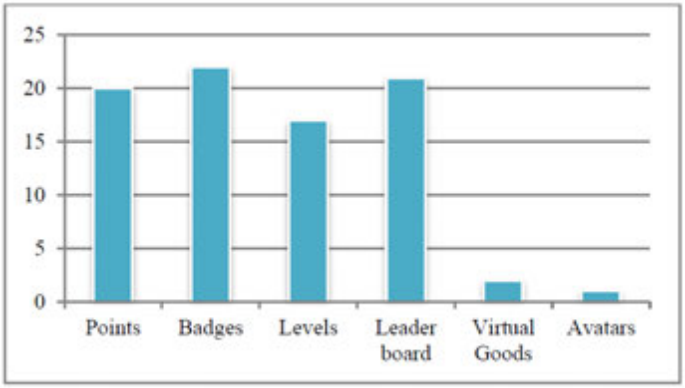


Fig. 1: Elements of game mechanics [2]; points, badges, levels and scoreboards stand out as the most frequent elements in research papers

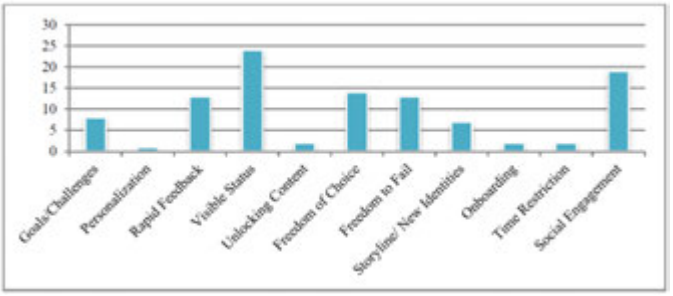


Fig. 2: Principles of gamification design and distribution according to the occurrence rate in scientific papers [2]

Gamification has largely been applied to approaches to student learning in higher education, as can be observed from

Target audience	Number	Frequency (%)
Higher Education	12	46.15%
Non-specific context/level	6	23.08%
Training and Tutorials	3	11.54%
Languages	2	7.69%
Elementary Education	2	7.69%
Lifelong Education	1	3.85%

TABLE 1: Target audience distribution [3]

the TABLE 1, showing the distribution of the gamification target users [3].

Also, the authors identified 7 goals of gamification that represent "motivation" as a major factor in gamification research efforts:

- Mastering skills, thus enhancing certain student competences;
- Challenges that add value to the learning process;
- Student involvement in the learning process through a simpler and more interesting learning process;
- Improving learning while maximizing knowledge acquisition;
- Behavioral changes by rewarding desirable and punishing undesirable behaviors;
- Socialization through various mechanisms of socialization and group learning;
- Guidelines - discussing the benefits of gamification as a means of motivation and coping with learning difficulties.

Within the paper [4] the authors give an overview of the fields in which the principles of gamification have been adopted: science, mathematics, foreign languages, cultural heritage, health, computer sciences, software engineering, business and logistics. The authors find that gamification encourages collaboration between students, independent learning, homework completion, facilitates assessment, integrates exploratory approaches, and enhances student creativity. Improving motivation and involvement are emphasized by the authors as the main motivation for adopting gamification techniques - to make learning more attractive, "engaging" and effective. Our research shows that there are indeed strong connections involving a concept "motivation" within the resulting network analysis graph.

The goal of [5] is to discover how to implement game elements within the course Computer Games, and to observe students' perceptions of implementation. It has been noted that elements such as goals, scoreboards, and big boss fights have been accepted favorably by students, which is consistent with the results of other papers such as [6]. The element that received the most student approval was the fight against the "big boss" enemy, which actually represented the final exam in front of a panel of experts in the field of computer games. Students also positively evaluated the use of the element of happiness, that is, randomness. Students were not satisfied with the way in which collaborative teams were formed (randomly created student groups), and with some other aspects of implementation.

In a report describing gamification experiences in higher education, the authors of [7] report correlating gamification with college pass rates, and participation in voluntary activities and challenging tasks. The authors also associate gamification with a greater level of interaction within the course and an increased level of focus on a course design. The paper summarizes the result of a three-year study of gamification by a Dutch technical university over technically challenging courses such as computer organization and cloud computing, and over a total of over 450 students.

The authors of [8] describe one significant drawback of the body of work in the field of gamification in educational processes, namely the relatively small number of empirical studies and the large number of theoretical papers lacking experimental moments to describe the experiences of the students. Our own analysis confirms this claim, as there are relatively weak connections on the concept "empirical research" in the resulting conceptual network graph. Nevertheless, researchers conclude that gamification has a positive impact on the education of engineering topics by making topics more manageable, increasing intrinsic motivation, scientific knowledge, collaboration, interest, and reducing or better managing the amount of work [8].

In the context of empirical research, the authors of [9] describe an experiment in which game elements such as grades, levels, scoreboards, challenges, and badges were used to enhance teaching in a graduate course. Using a variety of metrics, the authors compared the results of a gamified course with the same course that was performed without using gamification. The results showed significant increases in attendance at lessons and online participation, proactive behaviors, and study of course reference materials. According to the authors, the students found that the gamified instance of the course was more motivating, interesting and easier to learn than other courses.

Villagrasa and other authors [10] provided an overview and conceptualization of gamification for the 3D Art course in higher education. They discuss the importance of introducing story flow and game mechanics into the course, such as tracking development, credits, and virtual currency, and state that the use of gamification increases student involvement and motivation, as well as the overall effectiveness of the learning process, compared to conventional non-gamification teaching methods. However, as in most conceptual papers, in conclusion, the authors do not base their conclusions on a specific case study or experiment conducted.

Hamari and others [11] raise the question "does gamification work" in their review of 24 empirical studies of gamification. The most common elements of gamification through the aforementioned research are points, scoreboards and badges. According to most of the cases investigated, gamification has contributed to the positive effects and benefits of the learning

process, but with certain limitations, which should be further explored.

A model for introducing gamification into the field of e-learning within higher education is presented in [12], where authors state that the successful integration of gamification leads to greater levels of student satisfaction, motivation and involvement. The authors state that e-learning should include elements of gamification, and the main stages of development include analysis, planning, development, implementation and evaluation, with particular emphasis (through all stages) on customer experience. The authors also emphasize the importance of personalization within the e-learning system.

In 2012, the Rochester Institute of Technology's School of Interactive Games and Media developed an actual game system called "Just Press Play" to help students actively engage with the learning environment and to foster positive social behaviors that can lead to academic success. In this way, connections between students from different ages and with the faculty are encouraged, attending workshops, suggesting ideas for course development, etc. The leading goal of the game was thus to increase students' involvement in the life of the University for the sake of academic and social success [13].

"Who is Herring Hale?" is an alternative reality game developed in 2006 at Brighton University to assist new students in orienting and adjusting to student life, and to improve student inclusion [14]. The game included points and prizes, a series of 10 tasks (one per week), with a background story about time travel and the support of an online community on the campus social network. The assignments were based on physical and online services available to students. The game included hidden clues, codes, and encrypted content contained within the usual information about a particular service. The results of applying this method showed that the format of the game was not liked by all students, but also proved to be very effective for those students who accepted the game.

SCOOT [15] is an interactive mixed-reality game created for Australian universities and museums between 2004 and 2009, using a variety of platforms (mobile, web, public displays) as tools to guide players to their target locations and tasks. The game uses the real and virtual worlds as the backdrop of gameplay, and focuses on the Role Play Game (RPG) genre, which is chosen to be largely compatible with the intentions of the game's design, with elements such as roles, missions, exploration, narrative.

Star Question is a system developed to help students test the knowledge learned on a course, and includes grading, commenting, and badge systems [16]. Students involved with Star Question reportedly found it helpful to experience increased levels of understanding, repeat lessons, and reduced stress when repeating lessons and preparing for tests.

Multiplayer Classroom is developed for course "Theory and Practice of Game Design" at Indiana University [17], and included a game-based rating system and experience points, a final grade level course evaluation in game, guilds (groups within the classroom) created by students based on shared interests and ideas, and collaboration and localization within guilds to achieve a common group goal.

Classcraft is also inspired by RPG games, and the first version of the system was developed by Canadian teacher Shawn Young in 2011. The system's official release was launched in 2014 and is now available as a web and mobile application. It is a system that allows management and collaboration among students through playing RPG games, and is used actively from the lower grades of primary school to higher education institutions in a large number of countries. The system includes player role selection (such as wizard, warrior or healer), avatars, health points, action points, random events, and other elements of RPG games. Students progress through levels by solving quests and thus gaining "powers" they can use during class - such as skipping a question on a test or consulting with a classmate on how to solve a task. The analytical part of the platform allows teachers to observe and analyze student behaviors. "Classcraft" as a concept appeared within our conceptual network analysis with several connections, marking its appearance within primary and secondary schools.

SimSchool (www.simschool.org) is a virtual classroom platform available as a web-based "populated" application with so-called "simStudents" who have artificial emotional intelligence (they can laugh, cry, be frustrated, raise their hand, seek attention, etc.). The system is based on theories of cognition, emotion, social behavior, for the purpose of providing an "authentic experience" for teachers who aim to understand students and deliver effective teaching.

Labster (https://www.labster.com/) provides a virtual lab environment for courses such as biology and chemistry, contains graphics and scenarios which can be perceived as a game.

CodeCombat is a web application that allows you to learn programming languages (Python, JavaScript, CoffeeScript, Lua) in a playful RPG environment. The player can select a character in the game and must write code to perform tasks and accomplish goals in the game. CodeCombat contains game elements such as progress tracking, damage level, health level, speed and weapons of players, clan formation, content unlocking, missions, rewards, points, virtual currencies, etc.

Plantville is a simulation game developed by Siemens that enables new and potential employees to understand the management of industrial plants and technologies.

The Radix Endeavor (https://www.radixendeavor.org/) is a MMOPG (massively multiplayer online game) developed in 2013 at the Massachusetts Institute of Technology for middle and high school students learning statistics, proba- bility, genetics, algebra, geometry. etc. The system contains narratives, graphic elements of the game, character selection and customization, collaboration, tasks and goals, monitoring progress, overcoming problems, etc.

In the next sections we will introduce our methodology, containing the conceptual network analysis of gamification practices in order to better understand game-based learning in primary and secondary education.

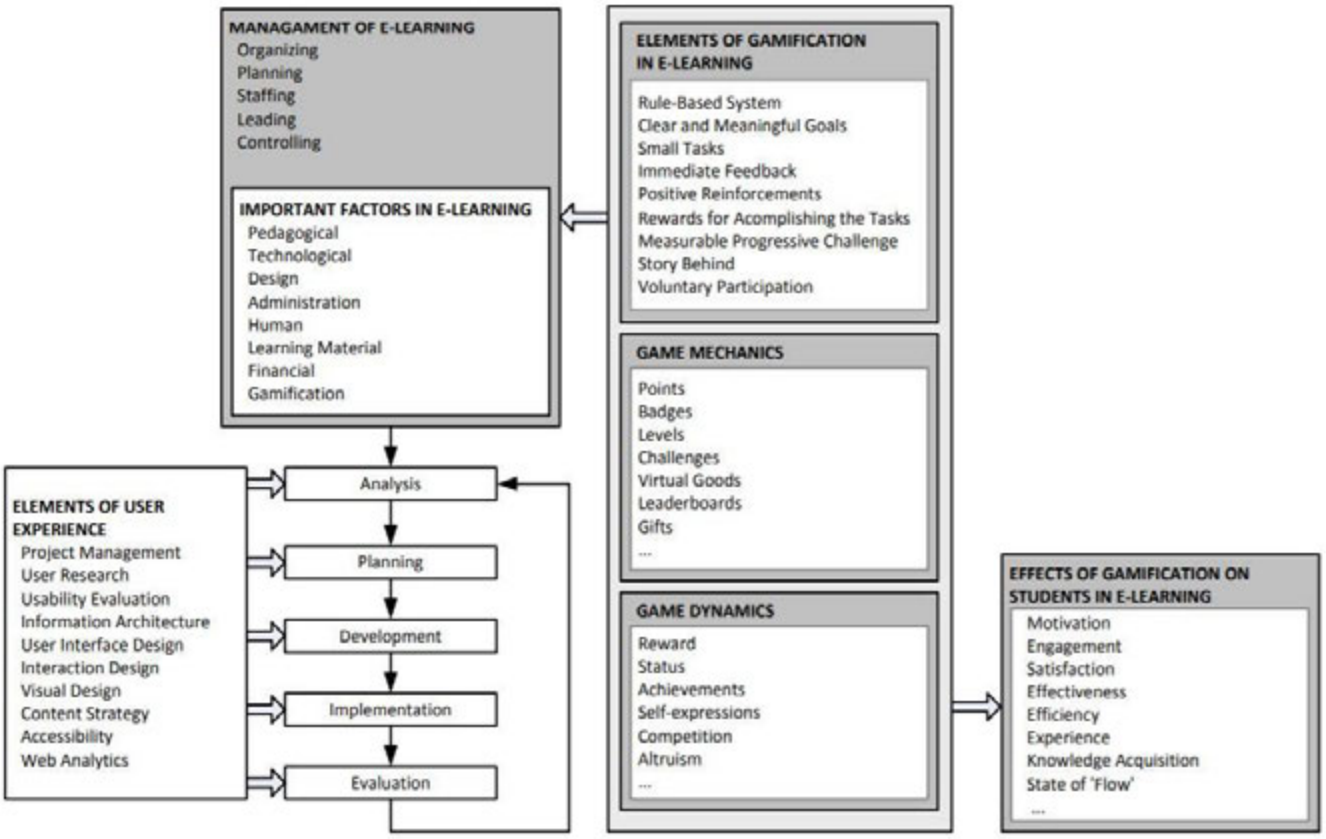


Fig. 3: A model for introducing gamification into an e-learning system [12]

II. METHODOLOGY

Herien, in order to identify most important topics in the domain of gamification in primary and secondary education we have adopted a methodology similar to the one used in previous papers [18], [19], [20]. In order to construct a conceptual network based on publications we use a bipartite graph constructed through co-occurence of two keywords on a particular paper. In other words, we consider to concepts (keywords) which represent the nodes of the graph to be connected by an edge if they occur on the same paper.

Firstly, to find the relevant publications we have systematically searched through more than 400 papers which turned out on the identified keyword combinations consisting of mostly "gamification", "education", "primary", "elementary", "middle", "secondary" in common bibliographic databases. Because a number of papers had not made clear about the school context on which the gamification was applied within the paper title, nor within the abstract, manual research and context understanding of such papers was unavoidable in order to properly include papers within this work. The final set of papers which were relevant to our study had 41 entries of which 5 didn't have specified keywords. Thus the keywords of 36 papers were analyzed and visualized using Gephi. The visualised network is shown on figure 4.

In the visualization nodes (concepts) are represented as red rectangles with the actual keyword written inside and edges are displayed as curved lines. The thickness of the edges's line indicates the weight of the edge - i.e. if there is a connection between two concepts on more papers, the line will be thicker. The Fruchterman Reingold algorithm has been used to construct the graphs layout.

III. DISCUSSION

As one can see from figure 4 there is one main well connected cluster (formed around the expected concept "gamification") as well as two smaller clusters on the left hand side of the visualization. Additionally, there are six weaker connected clusters at the edges of the graph.

The concept "gamification" in the main cluster is the most central node and shows strongest connections to the concepts: "primary school", "children", "student", "secondary education"; "game based learning" and interestingly "mobile learning".

Connected concepts relating to specific subjects and learning fields include "physics", "elementary mathematics", "natural sciences", "algebra", pointing to possible domains of gamification implementations. Also, there is an indication of using gamification for students with disabilities: concepts

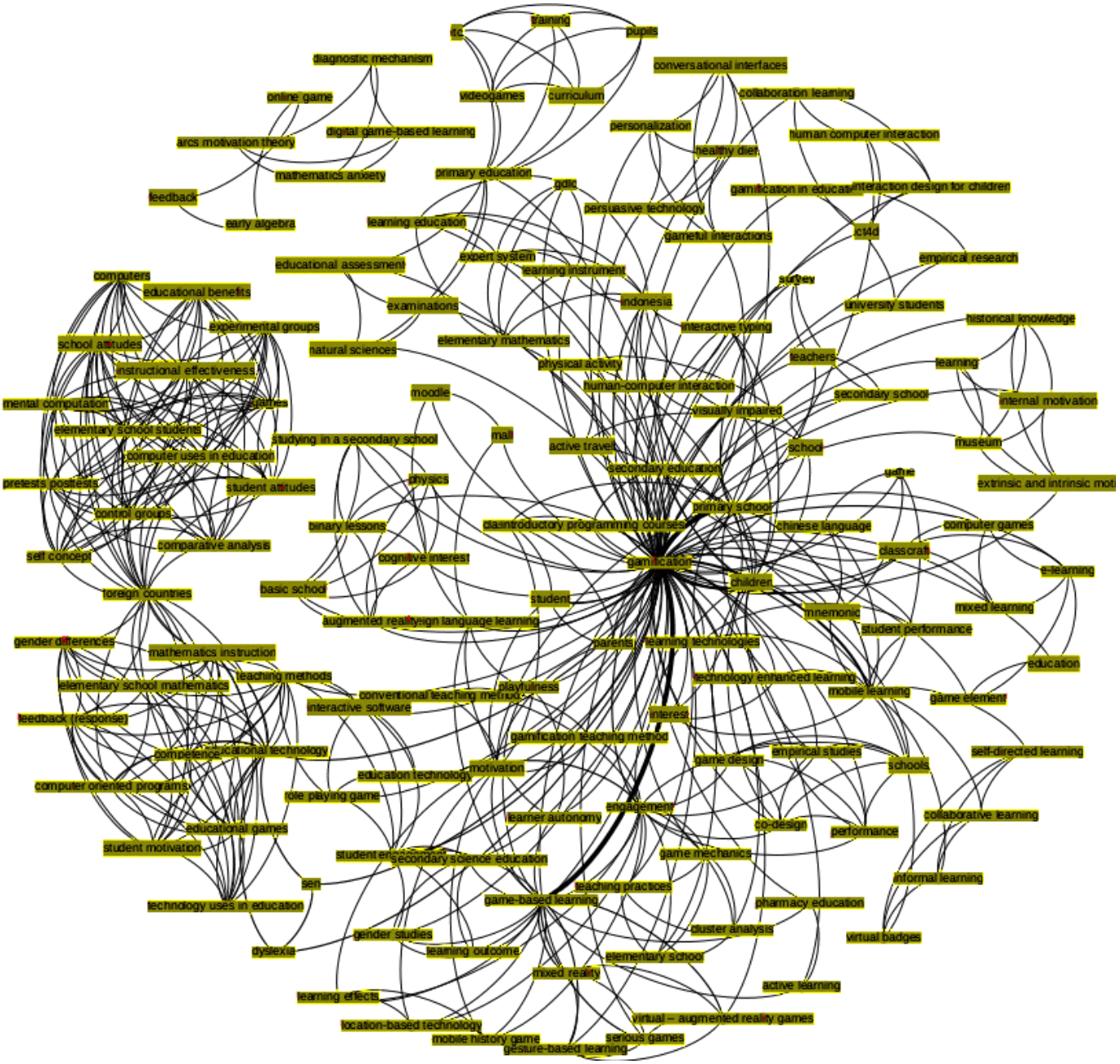


Fig. 4: Conceptual Network visualized using Gephi

such as "dyslexia" and "visually impaired" showed up within the network with significant connections.

The two smaller clusters have formed due to papers that used procedurally generated keywords, and as can be seen from the visualization are only weakly connected to the rest of the graph. This in a way indicates that procedurally generated keywords are quite different from author picked ones.

IV. Conclusion

In this paper we have provided a conceptual network analysis of most relevant gamification publications in primary and secondary education. After a literature review and systematic publication search we have extracted the keywords used and constructed a conceptual network visualization. The constructed network is well connected with one main cluster which formed around the most central node "gamification", two smaller clusters which formed due to automated keyword generation techniques as well as 6 smaller clusters with weaker connection to the main body of the graph.

The topology of the graph indicates a cohesive field of research since only one smaller cluster is disconnected from the rest of the graph meaning that all other clusters of papers usually have some keywords used in common.

V. REFERENCES

[1] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: defining" gamification"," in *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*, pp. 9–15, 2011.

[2] D. Dicheva, C. Dichev, G. Agre, G. Angelova, *et al.*, "Gamification in education: A systematic mapping study," *Educational Technology & Society*, vol. 18, no. 3, pp. 75–88, 2015.

[3] S. de Sousa Borges, V. H. Durelli, H. M. Reis, and S. Isotani, "A systematic mapping on gamification applied to education," in *Proceedings of the 29th annual ACM symposium on applied computing*, pp. 216–222, ACM, 2014.

[4] I. Caponetto, J. Earp, and M. Ott, "Gamification and education: A literature review," in *European Conference on Games Based Learning*, vol. 1, p. 50, Academic Conferences International Limited, 2014.

[5] M. Sillaots, "Gamification of higher education by the example of computer games course," in *The Seventh International Conference on Mobile, Hybrid, and On-line Learning (eLmL)*, pp. 62–58, 2015.

[6] K. M. Kapp, The gamification of learning and instruction. Wiley San Francisco, 2012.

[7] A. Iosup and D. Epema, "An experience report on using gamification in technical higher education," in Proceedings of the 45th ACM technical symposium on Computer science education, pp. 27–32, ACM, 2014.

[8] A. P. Markopoulos, A. Fragkou, P. D. Kasidiaris, and J. P. Davim, "Gamification in engineering education and professional training," International Journal of Mechanical Engineering Education, vol. 43, no. 2, pp. 118–131, 2015.

[9] G. Barata, S. Gama, J. Jorge, and D. Gonçalves, "Engaging engineering students with gamification," in 2013 5th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES), pp. 1– 8, IEEE, 2013.

[10] S. Villagrasa, D. Fonseca, E. Redondo, and J. Duran, "Teaching case of gamification and visual technologies for education," Journal of Cases on Information Technology (JCIT), vol. 16, no. 4, pp. 38–57, 2014.

[11] J. Hamari, J. Koivisto, H. Sarsa, et al., "Does gamification work?- a literature review of empirical studies on gamification.," in HICSS, vol. 14, pp. 3025–3034, 2014.

[12] M. Urh, G. Vukovic, E. Jereb, et al., "The model for introduction of gamification into e-learning in higher education," Procedia-Social and Behavioral Sciences, vol. 197, pp. 388–397, 2015.

[13] R. Martinez, C. Martin, S. Harris, K. Squire, E. Lawley, and A. Phelps, "Just press play: Design implications for gamifying the undergraduate experience," Proceedings of the Games+ Learning+ Society, vol. 8, 2012.

[14] K. Piatt, "Using alternate reality games to support first year induction with elgg," *Campus-Wide Information Systems*, vol. 26, no. 4, pp. 313–322, 2009.

[15] D. Polson and C. Morgan, "The 10% rule. maximising learning through collaborative game design," in *Proceedings of IADIS International Conference Mobile Learning 2010*, pp. 197–204, IADIS Press, 2010.

[16] S. Kim, "Star question: Gamification of a reviewing process using self-setting question and game mechanism in undergraduate education," *The Social Sciences*, vol. 9, no. 6, pp. 437–441, 2014.

[17] L. Sheldon, *The multiplayer classroom: Designing coursework as a game*. Nelson Education, 2011.

[18] M. P. Bach, M. Schatten, and Z. Marušic´, "Data mining applications in tourism: A keyword analysis," in Central European Conference on Information and Intelligent Systems (24; 2013), 2013.

[19] M. Schatten, "What do croatian scientist write about? a social and conceptual network analysis of the croatian scientific bibliography," Interdisciplinary Description of Complex Systems: INDECS, vol. 11, no. 2, pp. 190–208, 2013.

[20] M. Schatten, J. Rasonja, P. Halusek, and F. Jakelić, "An analysis of the social and conceptual networks of ceciis 2005-2010," in Central European Conference on Information and Intelligent Systems, 2011.

VI. REVIEWED PUBLICATIONS

[1] L. Ardoiz García *et al.*, "Gamification in english teaching in primary education," 2017.

[2] T. Jagušt, I. Boticki, V. Mornar, and H.-J. So, "Gamified digital math lessons for lower primary school students," in *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 691–694, IEEE, 2017.

[3] I. Boticki, J. Baksa, P. Seow, and C.-K. Looi, "Usage of a mobile social learning platform with virtual badges in a primary school," *Computers & Education*, vol. 86, pp. 120–136, 2015.

[4] V. Marín, M. López, and G. Maldonado, "Can gamification be introduced within primary classes?.," *Digital Education Review*, vol. 27, pp. 55–68, 2015.

[5] M. D. Kickmeier-Rust, E.-C. Hillemann, and D. Albert, "Gamification and smart feedback: Experiences with a primary school level math app," *International Journal of Game-Based Learning (IJGBL)*, vol. 4, no. 3, pp. 35–46, 2014.

[6] Y. Udjaja, A. C. Sari, *et al.*, "A gamification interactive typing for primary school visually impaired children in indonesia," *Procedia computer science*, vol. 116, pp. 638–644, 2017.

[7] T. A. Papp and A. Theresa, "Gamification effects on motivation and learning: Application to primary and college students," *International Journal for Cross-Disciplinary Subjects in Education*, vol. 8, no. 3, pp. 3193–3201, 2017.

[8] T. Halloluwa, D. Vyas, H. Usoof, and K. P. Hewagamage, "Gamification for development: a case of collaborative learning in sri lankan primary schools," *Personal and Ubiquitous Computing*, vol. 22, no. 2, pp. 391–407, 2018.

[9] T. Jagušt, I. Botički, and H.-J. So, "Examining competitive, collaborative and adaptive gamification in young learners' math learning," *Computers & education*, vol. 125, pp. 444–457, 2018.

[10] D. Gooch, A. Vasalou, L. Benton, and R. Khaled, "Using gamification to motivate students with dyslexia," in *Proceedings of the 2016 CHI Conference on human factors in computing systems*, pp. 969–980, 2016.

[11] G. Dodero, R. Gennari, A. Melonio, and S. Torello, "Towards tangible gamified co-design at school: two studies in primary schools," in *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play*, pp. 77–86, 2014.

[12] P. Zikas, V. Bachlitzanakis, M. Papaefthymiou, S. Kateros, S. Georgiou, N. Lydatakis, and G. Papagiannakis, "Mixed reality serious games and gamification for smart education," in *European Conference on Games Based Learning*, p. 805, Academic Conferences International Limited, 2016.

[13] K. C. Lim and K. E. Leong, "A study of gamification on geogebra for remedial pupils in primary mathematics," in *Asian Conference on Technology in Mathematics (ATCM)*, pp. 222–228, 2017.

[14] A. A. De Freitas and M. M. de Freitas, "Classroom live: a software-assisted gamification tool," *Computer Science Education*, vol. 23, no. 2, pp. 186–206, 2013.

[15] I. Ioannou and E. A. Kyza, "The role of gamification in activating primary school students' intrinsic and extrinsic motivation at a museum," in *Proceedings of the 16th World Conference on Mobile and Contextual Learning*, pp. 1–4, 2017.

[16] C. Hursen and C. Bas, "Use of gamification applications in science education," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 14, no. 01, pp. 4–23, 2019.

[17] E. S. Rivas, J. R. Palmero, and J. S. Rodríguez, "Gamification of assessments in the natural sciences subject in primary education," *Educational Sciences: Theory & Practice*, vol. 19, no. 1, 2019.

[18] D. Rawendy, Y. Ying, Y. Arifin, and K. Rosalin, "Design and development game chinese language learning with gamification and using mnemonic method," *Procedia Computer Science*, vol. 116, pp. 61–67, 2017.

[19] D. J. Miller and D. P. Robertson, "Educational benefits of using game consoles in a primary classroom: A randomised controlled trial," *British Journal of Educational Technology*, vol. 42, no. 5, pp. 850–864, 2011.

[20] J. Huizenga, W. Admiraal, S. Akkerman, and G. t. Dam, "Mobile game-based learning in secondary education: engagement, motivation and learning in a mobile city game," *Journal of Computer Assisted Learning*, vol. 25, no. 4, pp. 332–344, 2009.

[21] A. Kolovou and M. V. D. Heuvel-Panhuizen, "Online game-generated feedback as a way to support early algebraic reasoning," *International Journal of Continuing Engineering Education and Life-Long Learning*, vol. 20, no. 2, pp. 224–238, 2010.

[22] A. Fadhil and A. Villafiorita, "An adaptive learning with gamification & conversational uis: The rise of cibopolibot," in *Adjunct publication of*

the 25th conference on user modeling, adaptation and personalization, pp. 408–412, 2017.
[23] Y.-M. Huang, S.-H. Huang, and T.-T. Wu, "Embedding diagnostic mechanisms in a digital game for learning mathematics," *Educational Technology Research and Development*, vol. 62, no. 2, pp. 187–207, 2014.
[24] Y. Udjaja, V. S. Guizot, and N. Chandra, "Gamification for elementary mathematics learning in indonesia," *International Journal of Electrical and Computer Engineering*, vol. 8, no. 5, p. 3859, 2018.
[25] E. Coombes and A. Jones, "Gamification of active travel to school: a pilot evaluation of the beat the street physical activity intervention," *Health & place*, vol. 39, pp. 62–69, 2016.
[26] G. Y. Hong and M. Masood, "Effects of gamification on lower secondary school students' motivation and engagement," *International Journal of Social, Education, Economics and Management Engineering*, vol. 8, no. 12, pp. 3483–3490, 2014.
[27] A. Khan, F. H. Ahmad, and M. M. Malik, "Use of digital game based learning and gamification in secondary school science: The effect on student engagement, learning and gender difference," *Education and Information Technologies*, vol. 22, no. 6, pp. 2767–2804, 2017.
[28] S. Papadakis and M. Kalogiannakis, "Using gamification for supporting an introductory programming course. the case of classcraft in a secondary education classroom," in *Interactivity, Game Creation, Design, Learning, and Innovation*, pp. 366–375, Springer, 2017.
[29] L. Sera and E. Wheeler, "Game on: The gamification of the pharmacy classroom," *Currents in Pharmacy Teaching and Learning*, vol. 9, no. 1, pp. 155–159, 2017.
[30] M. Sanmugam, Z. Abdullah, H. Mohamed, B. Aris, N. M. Zaid, and S. M. Suhadi, "The affiliation between student achievement and elements of gamification in learning science," in *2016 4th International Conference on Information and Communication Technology (ICoICT)*, pp. 1–4, IEEE, 2016.
[31] A. A. Carvalho, N. Zagalo, and I. Araújo, "From games played by secondary students to a gamification framework," in *Society for Information Technology & Teacher Education International Conference*, pp. 737–744, Association for the Advancement of Computing in Education (AACE), 2015.
[32] D. Gooch, A. Vasalou, and L. Benton, "Exploring the use of a gamification platform to support students with dyslexia," in *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–6, IEEE, 2015.
[33] V. L. Buzko, A. V. Bonk, and V. Tron, "Implementation of gamification and elements of augmented reality during the binary lessons in a secondary school," in *Proceedings of the 1st International Workshop on Augmented Reality in Education Kryvyi Rih, Ukraine, October 2, 2018*, no. 2257, pp. 53–60, CEUR-WS. org, 2018.
[34] B. Pardoel, "Gamification and its potential for foreign language learning-lessons from a six-week gamified moodle course for german as a foreign language at secondary school level," 2018.
[35] I. Yildirim, "The effects of gamification-based teaching practices on student achievement and students' attitudes toward lessons," *The Internet and Higher Education*, vol. 33, pp. 86–92, 2017.
[36] C. Cruaud, "The playful frame: gamification in a french-as-a-foreign-language class," *Innovation in Language Learning and Teaching*, vol. 12, no. 4, pp. 330–343, 2018.
[37] E. Sanchez, G. Kalmpourtzis, J. Cazes, M. Berthoix, and R. Monod-Ansaldi, "Learning with tactileo map: From gamification to ludicization of fieldwork," 2015.
[38] L. da Rocha Seixas, A. S. Gomes, and I. J. de Melo Filho, "Effectiveness of gamification in the engagement of students," *Computers in Human Behavior*, vol. 58, pp. 48–63, 2016.
[39] M. Otero-Agra, R. Barcala-Furelos, I. Besada-Saavedra, L. Peixoto-Pino, S. Martínez-Isasi, and A. Rodríguez-Núñez, "Let the kids play: gamification as a cpr training methodology in secondary school students. a quasi-experimental manikin simulation study," *Emergency Medicine Journal*, vol. 36, no. 11, pp. 653–659, 2019.
[40] M. Schatten and M. Schatten, "Gamification of game programming education: A case study in a croatian high school," in *Proceedings of the Central European Conference on Information and Intelligent Systems*, pp. 13–18, 2019.

Adaptive rolling window selection for minimum variance portfolio estimation based on reinforcement learning

Bruno Gašperov*, Fredi Šarić, Stjepan Begušić, Zvonko Kostanjčar
* Laboratory for Financial and Risk Analytics, Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia

***Abstract* - When allocating wealth to a set of financial assets, portfolio optimization techniques are used to select optimal portfolio allocations for given investment goals. Among benchmark portfolios commonly used in modern portfolio theory, the global minimum variance portfolio is becoming increasingly popular with investors due to its relatively good performance which stems from both the low-volatility anomaly and the avoidance of the estimation of first moments i.e. mean returns. However, estimates of minimum variance portfolio weights significantly depend on the size of the rolling window used for estimation, especially considering the non-stationarity of the underlying market dynamics. In this paper, we use a model-free policy-based reinforcement learning framework in order to directly and adaptively determine the optimal size of the rolling window. Training is done on a subset of trading stocks from the NYSE. The resulting agent achieves superior performance when compared against multiple benchmarks, including those with fixed rolling window sizes.**

***Keywords* – reinforcement learning; portfolio optimization; covariance estimation**

I. INTRODUCTION

Despite the implications of the Capital Asset Pricing Model (CAPM), multiple empirical studies show that the global minimum variance (GMV) portfolio yields surprisingly high average returns [1] [2] [3]. This is related to the general stylized fact of stock markets - the low volatility anomaly – an observation that low-beta (low-volatility) stocks overperform high-beta (high-volatility) stocks [4]. Furthermore, calculation of the GMV portfolio weights does not require the estimation of first moments (expected stock returns) which is known to be notoriously difficult [5]. It is in part due to the aforementioned advantages that investors have lately been increasingly turning to the GMV portfolio [6]. (For performance of the GMV portfolio on a subset of considered stocks see Figure 1.)

However, it should be noted that estimates of minimum variance portfolio weights significantly depend on the size T of the rolling window used for estimation of the sample return covariance matrix. The optimal size of the window can be reasonably expected to depend on the presence of correlation changes in the time series of returns, with shorter optimal window sizes possibly corresponding to periods that include correlation change points (and vice versa). To the authors' knowledge, the amount of research done in this area is quite scarce with only a few relevant studies [7]. Motivated by this, our goal in this paper is to develop a method for adaptive rolling window selection based on reinforcement learning. The use of reinforcement learning is warranted due to intrinsic sequentiality and non-stationarity of the problem as well as the absence of data labels (as that the optimal choice of the rolling window size at each time step is not known).

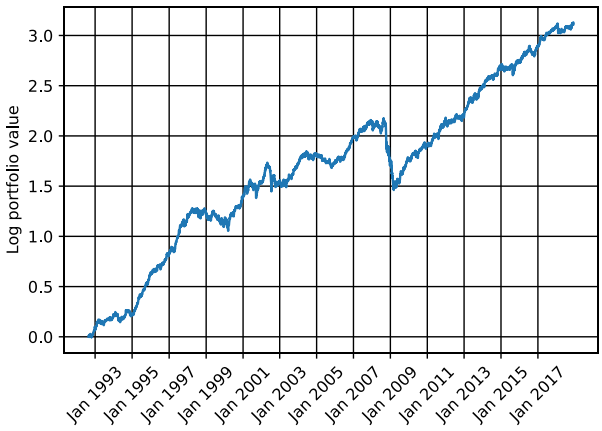


Figure 1: Global minimum variance portfolio log value for the considered 20 NYSE stocks on the whole dataset for a fixed size of the rolling estimation window (one trading year i.e. T=252)

II. PROBLEM DEFINITION

Under the minimum variance portfolio optimization strategy, historical data is firstly used to obtain the sample covariance matrix $\hat{\Sigma}$ which is given by the following expression:

$$\hat{\Sigma} = \frac{1}{T-1}\sum_{t=1}^{T}(\boldsymbol{r_t} - \bar{\boldsymbol{r}})(\boldsymbol{r_t} - \bar{\boldsymbol{r}})^{\mathrm{T}}$$

where T is the size of the estimation window, $\boldsymbol{r_t}$ return at time t and $\bar{\boldsymbol{r}}$ the sample mean of the returns.

The GMV portfolio w^* is then given as the solution to the following optimization problem:

$$\min_{\boldsymbol{w}} \frac{1}{2}\boldsymbol{w}^T\hat{\boldsymbol{\Sigma}}\mathbf{w}$$

under the constraint:

$$\mathbf{1}^{\mathrm{T}}\boldsymbol{w} = 1$$

The analytical solution exists and is given by:

$$w^* = \frac{\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}}{\mathbf{1}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{1}}$$

Given that the only constraint is that portfolio weights should sum up to 1, short selling is allowed. Note that the sample covariance matrix depends on T, i.e. the size of the rolling window used for estimation. Considering that the sample covariance is used as an input to the optimization step, the obtained portfolio weights consequently also depend on T.

Despite the unbiasedness of the sample covariance matrix, it is corrupted by a significant degree of estimation error, especially when the number of assets N is commensurate with the size of the estimation window T. Various classes of methods have been proposed to ameliorate this problem, including factor model estimators and shrinkage estimators [8] [9] [10]. Unlike such methods, in this paper we do not attempt to directly „clean“ the covariance matrix from noise but instead use reinforcement learning as an auxiliary tool for adaptive rolling window selection. However, it is possible to combine such methods with our adaptive framework.

Under our framework, at each day (t) the trader makes the choice of the size (length) of the rolling window used for the estimation of the sample return covariance matrix based upon the current state of the environment. The resulting window size is then used to obtain the sample covariance matrix from which the weights of the GMV portfolio (as explained above) are analytically calculated and the investment is made accordingly. To further simplify the problem we make two additional assumptions. Specifically, the assumption of complete market liquidity which implies that each trade can be completed at the last observed price, and the assumption of zero trading costs.

III. LITERATURE REVIEW

Current research on the selection of covariance matrix estimation window size in the context of portfolio optimization is scarce, especially if adaptive approaches are considered. Bayraktar and Bilge [7] study optimal estimation widow sizes using classical Markowitz portfolio optimization, albeit in a non-adaptive fashion. Golosnoy [11] uses a time-varying window size for covariance matrix estimation in order to find the optimal trade-off between estimation error and bias. Optimal portfolio weights are monitored via the use of control charts that give an alarm when the control statistic leaves the acceptance area, indicating changes in the means of the optimal portfolio weights. The authors empirically show that time-varying window size strategies outperform the alternatives in most cases. Härdle et. al. [12] use a data-driven approach to adaptively vary the size of the window over which a local constant-parameter model is estimated for high-frequency financial variables. It is found that this adaptive approach yields significantly better forecasts than alternative approaches that use fixed-size estimation intervals. Finally, the authors conclude that adaptive approaches enable both better high-frequency forecasts and insights into local variations of the parameters.

IV. CONTRIBUTIONS

The main contributions of this paper include the following:

- We sucessfully use deep reinforcement learning for dynamic determination of the optimal size of the rolling window used for estimation of the sample return covariance matrix. To the best of the authors' knowledge, similar studies have not been previously done. The resulting agent achieves superior performance over various benchmark agents.
- We construct a novel state-space representation based on Frobenius norms of the differences between consecutive correlation matrix estimates in order to alleviate detection of correlation changes in the time series of financial returns. Although such approaches are commonly used for break-point detection in multivariate time series, this is, to the best of the authors' knowledge, the first attempt to combine such an approach with reinforcement learning to tackle the problem of rolling window length selection for portfolio optimization.

V. DATA AND MODEL DESIGN

A. Data

We consider a dataset comprised of daily closing stock prices from the New York Stock Exchange (NYSE), obtained via Yahoo! Finance. Firstly, we remove all stocks that include at least one missing value and randomly select a fixed subset of 20 stocks. The price data has a time span of nearly 29 years (from 1990-Jan-02 until 2018-Dec-04). The dataset is divided into three (training, validation and testing) sets with the approximate ratio of 0.60:0.20:0.20. After preprocessing, the data is fed into a neural network that acts as a policy network which should ideally output optimal actions for each set of inputs.

B. Reinforcement learning setting

We can represent the problem as a Markov Decision Process (MDP) where the agent (the investor) at each time step (a day) receives a trading signal (a function of current and past returns) from the environment (the market) and makes actions (the choice of the size of the rolling window used for estimation) based upon it. The environment then rewards the agent with a numerical reward (proportional to the negative variance of the realized portfolio returns). The sole goal of the agent is to find the optimal policy which maximizes the expected sum of discounted cumulative rewards (i.e. to minimize the variance of realized portfolio returns).

State space – In order to facilitate learning we first use raw returns to calculate sample correlation matrix estimates M_t with a fixed-size rolling window of 70 days for each day t from 1990-May-29 until 2018-Dec-04. We proceed by calculating the Frobenius norm F_t of the difference between matrices M_t and M_{t-1} for each day t from 1990-

May-30 until 2018-Dec-04. Finally, for each day t, we calculate the means of the sets $\{F^i\}_{i=0,\dots,29}$ where:

$$F_t^i = \{F_{t-21i}, F_{t-21i-1}, F_{t-21i-2}, \dots, F_{t-21i-20}\}$$

The obtained values $F^0, F^1, \dots, F^{29}$ are used to construct the state representation $S_t = \{F_t^0, F_t^1, \dots, F_t^{29}\}$ for each time t. Finally, we perform Z-score normalization on each of the state representation features. Intuitively, large entries in the state representation can be understood to imply correlation changes in the corresponding time periods which subsequently may lead to smaller optimal rolling window sizes.

Action space – Each action A corresponds to a certain choice of the size of the rolling window T used for estimation. To limit the number of available actions and hence enhance the training process, the rolling window size is given by $21(k+6)$ where $k \in \{0,\dots,24\}$. Therefore at any time t there are exactly 25 actions available to the agent, each one representing a rolling window of length $k+6$ in months (considering that the trading month consists of 21 days). Since t is in days, weights are rebalanced on a daily basis.

Rewards – All rewards R_i are set to 0 expect for the reward at the end which is proportional to the negative variance of the realized returns, thereby encouraging variance minimization. Therefore:

$$R_{T_e} = -\sum_{t=1}^{T_e} (r_t - \bar{r})^2$$

where T_e is the length of the episode, $\bar{r}$ the average portfolio return and r_t portfolio return at time t. Obviously, this is a case of sparse rewards. One should note that, since there is only one non-zero reward (precisely at the end of the episode), the discount rate is not relevant and can be set to 1.

Episodes – Each episode can be represented as a complete trajectory:

$$(S_0, A_0, R_0 = 0, \dots, S_{T_e}, A_{T_e}, R_{T_e} = -\sum_{t=1}^{T_e} (r_t - \bar{r})^2, S_{T_e+1})$$

Note that since we are backtesting (i.e. using historical data), transitions between states are deterministic and it generally holds that A_t does not influence S_{t+1}.

C. Policy (neural) network

The reinforcement learning agent is represented by a neural network. Two fully connected neural networks with different architectures are considered in model selection. Architecture A consists of only two hidden layers (with 512 and 256 neurons, respectively) while architecture B consists of five hidden layers (each one with 64 neurons). The ReLU function is used as an activation function in all hidden layers. The output layer in both architectures consists of a single neuron with a sigmoid activation function, and the final output is passed into to the appropriately scaled floor function. Finally, architecture A is chosen due to higher performance on the validation set. It should be noted that under this approach the neural network maps states directly into deterministic (optimal) actions instead of probability distributions over the action space. This is somewhat similar to the approach taken by Jiang and Liang [13] and has the advantage of eschewing the problem of high variance associated with more conventional approaches.

D. Optimization procedure

In order to train the weights of the neural network, a simple genetic algorithm is used. Genetic algorithms are a class of optimization algorithms suitable for „black-box" function optimization. Recently, they were shown to be competitive with gradient-based methods for deep reinforcement learning problems [14] [15]. Under such an approach, the function that maps the weights of the neural network to the expected reward (proportional to negative variance) is directly maximized. At the very start, a certain number of random agents are generated and evaluated. A number of top-performing agents are chosen as parents who yield offspring with the same weights up to a random perturbation. This process is then repeated for a large number of generations. To prevent overfitting, a variant of early stopping is employed. The exact hyperparameter values used for training are given in Table 1.

TABLE 1: GENETIC ALGORITHM - HYPERPARAMETER VALUES

Hyperparameter	Hyperparameter value
Number of agents per generation (population size)	50
Number of generations	300
Mutation rate	Decreases from 0.02 to 0.002
Mutation rate change frequency	once per 15 generations
Truncation size (number of top agents used as parents)	10

VI. RESULTS AND DISCUSSION

A. Simulated data

As a brief proof of concept, we train a neural network with a single hidden layer consisting of 10 neurons on a time series (length L=600) of returns generated by a five-dimensional multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}(t))$. Initially, $\boldsymbol{\Sigma}(t) = \boldsymbol{\Sigma}_1$. At the break point $t_1 = 100$ the covariance matrix abruptly changes to $\boldsymbol{\Sigma}_2$ ($\boldsymbol{\Sigma}_2 \neq \boldsymbol{\Sigma}_1$) and finally at $t_2 = 400$ it reverts back to $\boldsymbol{\Sigma}_1$. The mean $\boldsymbol{\mu}$ is constant throughout the episode. After training with multiple random seeds, the model is evaluated on a testing dataset consisting of a time series (again of length L=600) of returns generated by a different multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}'(t))$ with break points happening at $t_1' = 200$ and $t_2' = 500$. Figure 2 shows the agent's actions in the testing set. The agent seems to be able to react to both break points and adjust the estimation window size accordingly although with a significant lag ($\Delta t \approx 50$). Since the positions of break points were altered in the testing set, the results indicate that such an agent is, with the use of proper agent architecture and feature preprocessing, potentially capable of generalizing to unseen environments.

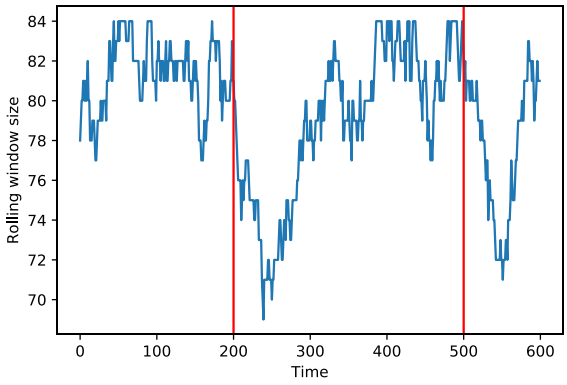


Figure 2: The agent's actions (choices of the rolling window sizes) in the testing set. Change points are denoted by vertical red lines.

B. *Real-world dataset*

After training, the model is evaluated on a testing dataset spanning the time period which lasts approximately 6 years. Figure 3 shows the agent's actions (chosen rolling window sizes) in the testing set as a function of time. The performance of the agent (with the ex-post annualized volatility used as the metric) is compared against traditional benchmarks (different choices of fixed sizes for rolling windows) and given in Table 2. Graphical comparison is made in Figure 4. The results clearly indicate that the agent successfully generalizes and manages to outperform the benchmarks, albeit quite modestly. Target downside deviations for the three portfolios are given in Table 3.

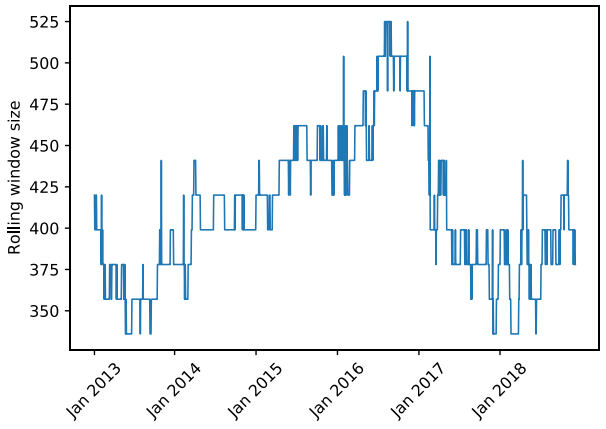


Figure 3: The agent's actions (choices of the rolling window sizes at each day t) in the testing set

TABLE 2: PORTFOLIO PERFORMANCE RESULTS MEASURED VIA ANNUALIZED EX-POST VOLATILITY

	Annualized ex-post volatility (assuming 252 trading days a year)
Our agent	10.91%
Best fixed window size (588, chosen *a posteriori*)	10.98%
Fixed window size (252, length of a trading year)	11.00%
Average random policy (100 iterations)	11.07%

TABLE 3: PORTFOLIO PERFORMANCE RESULTS MEASURED VIA TARGET DOWNSIDE DEVIATION

	Target Downside Deviation
Our agent	7.57%
Best fixed window size (588, chosen *a posteriori*)	7.59%
Fixed window size (252, length of a trading year)	7.63%
Average random policy (100 iterations)	7.65%

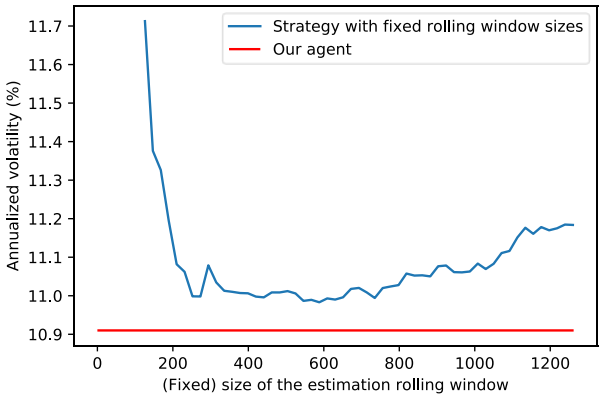


Figure 4: Comparison of annualized volatility achieved by our model versus annualized volatilities achieved by strategies using fixed rolling window sizes (in the testing set)

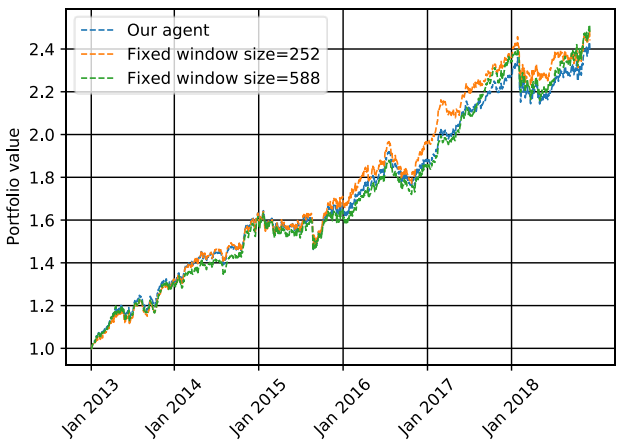


Figure 5: Portfolio value for both our agent and strategies using fixed rolling window sizes (in the testing set)

TABLE 4: PORTFOLIO PERFORMANCE RESULTS MEASURED VIA THE SHARPE RATIO

	Sharpe ratio (assuming 252 trading days a year)
Our agent	1.41
Best fixed window size (588, chosen *a posteriori*)	1.45
Fixed window size (252, length of a trading year)	1.43
Average random policy (100 iterations)	1.42

Additionally, Sharpe ratios are compared in order to determine if the obtained strategy is favorable when both risk and realized returns are accounted for. Note that the risk-free rate is set to 0 for the sake of simplicity. The performance results (shown in Table 4 and also in Figure 5) indicate that the resulting agent fails to attain a superior Sharpe ratio.

Although this is mostly unsurprising, considering our reward function is designed to only take variance into account, it is interesting to note that the results do not cast further doubt on the implications of the CAPM but are rather more aligned with them.

VII. LIMITATIONS AND FUTURE WORK

The main limitations include the assumption of no trading costs and the inclusion of only one type of state features. Furthermore, testing is performed on a period that is temporally distant from the training period, which implies that emphasis is given to long-term patterns between the state features and the reward, while shorter-term patterns are not captured by the model.

Among possible extensions of our work we mention enrichment of the state space with additional features (possibly macroeconomic indicators, interest rates data, yield curve data etc) and consideration of a wider spectrum of neural network architectures (especially convolutional neural networks). Additionally, instead of dynamically determining optimal discrete rolling window sizes, it would be interesting to perform a dynamic determination of optimal weighting schemes for covariance estimators (for instance, finding optimal smoothing factor values for exponential smoothing).

VIII. CONCLUSION

In this paper, we introduce an adaptive rolling window selection approach for minimum variance portfolio estimation and show that it manages to outperform the classical benchmarks (with fixed rolling window sizes) on a real-world dataset. The results suggest that using reinforcement learning based methods as an auxiliary tool in conjunction with more classical approaches, such as the analytic estimation of the GMV portfolio, may represent a feasible step toward better portfolio management strategies. It should also be noted that this paper is partly inspired by the line of research which includes alternative approaches to deep reinforcement learning based on genetic algorithms and evolutionary strategies which have been gaining traction lately due to their multiple favorable properties. Lastly, we emphasize that the outlined framework can be used with different policy architectures and is generally easily amenable to future enhancements.

IX. ACKNOWLEDGEMENTS

This work has been supported in part by Croatian Science Foundation under the project 5241 and in part by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS).

LITERATURE

[1] J. Philippe, “Bayesian and CAPM estimators of the means: implications for portfolio selection,” Journal of Banking and Finance 15, no. 3, pp. 717–727, June 1991.

[2] R. Clarke, H. De Silva, and S. Thorley, "Minimum-variance portfolio composition," The Journal of Portfolio Management 37, no. 2, pp. 31-45, January 2011.

[3] R. Clarke, H. De Silva, and S. Thorley, "Minimum-variance portfolios in the US equity market," The Journal of Portfolio Management 33, no. 1, pp. 10-24., October 2006.

[4] M. Baker, B. Bradley and J. Wurgler, „Benchmarks as limits to arbitrage: Understanding the low-volatility anomaly,“ Financial Analysts Journal 67, no. 1, pp. 40-54., January 2011.

[5] R. C. Merton, „On estimating the expected return on the market: an exploratory investigation,“ Journal of Financial Economics 8, no. 4, pp. 323-361., December 1980.

[6] B. Scherer, “A new look at minimum variance investing,” Journal of Empirical Finance 18, no. 4, pp. 652-660., July 2010.

[7] E. Bayraktar and A. H. Bilge, “Determination the parameters of Markowitz portfolio optimization model,” International Conference on Mathematical Finance and Economics (ICMFE) 2011.

[8] O. Lediot and M. Wolf, „Improved estimation of the covariance matrix of stock returns with an application to portfolio selection,“ Journal of Empirical Finance 10, no. 5, pp. 603-621, December 2003.

[9] E. Pantaleo, M. Tumminello, F. Lillo and R. N. Mantegna, „When do improved covariance matrix estimators enhance portfolio optimization? An empirical comparative study of nine estimators,“ Quantitative Finance 11, no. 7, pp. 1067-1080., July 2011.

[10] S. Begušić and Z. Kostanjčar, „Cluster-based shrinkage of correlation matrices for portfolio optimization“, 11th International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 301-305., IEEE, September 2019.

[11] V. Golosnoy, „Sequential monitoring of minimum variance portfolio,“, AStA Advances in Statistical Analysis 91, no. 1, pp. 39-55., March 2007.

[12] W. K. Härdle, N. Hautsch, and A. Mihoci, "Local adaptive multiplicative error models for high‐frequency forecasts," Journal of Applied Econometrics 30, no. 4, pp. 529-550., June 2015.

[13] Z. Jiang and J. Liang, „Cryptocurrency portfolio management with deep reinforcement learning,“ 2017 Intelligent Systems Conference (IntelliSys), pp. 905-913.. IEEE, September 2017.

[14] F. Such, et al., „Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning,“, NeurIPS Deep Reinforcement Learning Workshop, 2018.

[15] A. Sehgal, H. La, S. Louis and H. Nguyen, „Deep reinforcement learning using genetic algorithm for parameter optimization“, 2019 Third IEEE International Conference on Robotic Computing (IRC), pp. 596-601., IEEE, 2019

Synthetic astronomical image sequence generation

Anita Gribl, Davor Petrinović
Faculty of Electrical Engineering and Computing, University of Zagreb
Zagreb, Croatia
anita.gribl@fer.hr, davor.petrinovic@fer.hr

***Abstract*—Astronomical image reconstruction from multiple observations requires detailed knowledge of the imaging system. The reconstruction accuracy is determined by the imaging model correctness. Numerous parameters determine the model, and their precise estimation is a precondition for successful reconstruction. The best way to measure the model sensitivity on the parameter accuracy is a detailed simulation of the whole imaging process that takes into account all of the model parameters. The simulation enables an artificial manipulation of these parameters to reach the required accuracy. This paper describes a novel approach to generate a synthetic astronomical image sequence from the chosen star catalogs. Synthesis assumes point sources extracted from the catalogs and described through their angular position and magnitude in different spectral bands. The model parameters shall be tuned to the actual raw image captured by the SLR camera and chosen lens. The sequence of synthetic images of approximately the same part of the sky, but slightly shifted, is generated. The small shifts result that the light field integrates into different spatially displaced sampling grids, thus facilitating high-resolution image reconstruction from spatially subsampled projections.**

***Index Terms*—synthetic astronomical image generation, super-resolution, compressed sensing, geometric transformation, optical aberrations modeling, PSF, MTF**

I. Introduction

In the area of image processing, researchers put a lot of effort into increasing the spatial imaging system resolution. The resolution is limited by the sensor spatial resolution and optics. Super-resolution methods enable the reconstruction of the high-resolution (HR) image from a single or multi-frame low-resolution (LR) image [1], [2]. A subset of these algorithms, known as compressed sensing algorithms, assume signal sparsity and apply convex optimization for signal reconstruction [3]. Independently of the applied method, the reconstruction accuracy can be verified by comparing the reconstructed image with the ground-truth one.

This paper proposes the method for generating the astronomical synthetic image sequence, which can be used for objective evaluation of reconstruction techniques. It is important to understand the physical aspect of the imaging system parameters to make a mathematical model that incorporates all of them. The model parameters shall be tuned to the actual raw image captured by the camera and chosen lens. If the model parameters are precisely determined, the synthetic LR images will match those captured by the camera. Only precise determination of model parameters enables accurate image reconstruction. The parameters can be further manipulated for their sensitivity analysis. The same imaging system model enables HR image synthesis, for the same part of the sky, thus representing a ground-truth for verification of the reconstruction algorithm accuracy. Astronomical images natively satisfy the requirement of signal sparsity in their spatial domain due to the point-source nature of star objects. Consequently, they are a suitable 2D signal for application of compressed sensing algorithm for their representation and reconstruction. The star objects are point sources in the spatial domain, respectively Dirac Delta impulses, spatially spread in the image plane due to optical aberrations, electrical sensor properties, motion blur, etc. To make a synthetic astronomical image, point sources expected within the observation field of view, are extracted from the chosen star catalogs found in the publicly available VizieR catalogs database [4]. The database contains almost 20000 different catalogs divided by missions, target astronomy objects, and wavelength regime. The objects in catalogs are typically described through their angular position and magnitude in different spectral bands. The proposed method can be applied to any set of point sources described in the chosen catalog.

The synthetic image is determined by external camera parameters (viewpoint - azimuth, elevation, and rotation), by exposure time, by the selected optics (focal length which specifies the view angle), by the sensor dimension and its resolution, by the lens f-number which determines the diffraction pattern on the circular aperture, by the various optical aberrations and the unwanted residual motion of the imaging system during exposure what cause additional spreading of the object image on the sensor (PSF), by the lens vignetting, by photon sensitivity of the sensor, sensor gain (ISO) and noise, color filter array in front of the sensor, and finally by the portion of the sky enclosed in the frame, which depends on the observation location, date, exact time and external camera and/or telescope mount parameters.

II. Related work

The methods for high spatial resolution image reconstruction incorporate the upsampling techniques and image restoration techniques. An overview of super-resolution techniques, which try to increase an upsampling rate, are given in [5], [1], [6], [7]. An overview of super-resolution techniques based on compressed sensing algorithms is given in [8]. These algorithms assume that a target signal is sparse in a certain representation domain and that the number of samples needed for its reconstruction is much lower than required by the Nyquist-Shannon theorem. Some researchers also propose

nonregular sampling sensor as a solution for high spatial and temporal resolution from multi-frame image [9], [10], [11]. Captured images are typically corrupted by noise, motion blur, and camera defocus. The degraded images require image restoration, which implies the application of deconvolution and noise reduction techniques. The deconvolution implies an application of inverse blurring process, respectively an application of inverse system Point Spread Function (PSF). The example of fast shift-variant blur approximation considering the physical aspect of those approximations is given in [12]. An overview of blind deconvolution algorithms that assume unknown blur is given in [13]. Color filter arrays in cameras decrease the number of samples and make the reconstruction of high-resolution color images difficult. The missing pixels are estimated using different interpolation techniques, known as demosaicing. Super-resolution techniques, which also include demosaicing, are given in [14], [15]. The quality of the imaging system is typically described by the system Modulation Transfer Function (MTF). How to read and measure system MTF is explained in [16] and [17]. The modeling of optical aberrations using point spread functions is given in [18], [19], [20], [21] and [22]. The detailed explanation of imaging system aberration and their mathematical formulation using Zernike polynomials are given in [23].

III. Proposed method

The observation model of astronomical images can be written as:

$$\mathbf{y} = \mathbf{WCx} + \mathbf{n} \tag{1}$$

where $\mathbf{y}$ is a measurement vector that consists of captured LR images, $\mathbf{W}$ is a measurement matrix, $\mathbf{C}$ is convolution matrix, $\mathbf{x}$ is an unknown sparse signal of point sources and $\mathbf{n}$ is additive noise. Captured images are corrupted with noise and blur. The blur is a consequence of optical aberrations, electrical sensor properties, motion effects, and the presence of anti-aliasing filters. The blur caused by optical aberrations is modeled inside the convolution matrix. Image degradations caused by motions or the presence of anti-aliasing (AA) filters in the camera are modeled inside the measurement matrix to separate the optical effects from the motion effects.

A. Image registration

The image registration process includes the star objects annotation and the alignment of each image with the chosen referent one by finding the linear and nonlinear transformations between catalog and the referent image, and between each image and the referent one. The star objects in raw images are detected as local extrema taking into account that images were blurred before detection. The blur assures that potentially wrong sensor pixels or white pixels caused by salt and pepper noise would not be recognized as local extrema. The star objects are detected in the further steps as centers of the 2D Gaussian profile, since the catalog magnitudes are proportional to the log-integral of estimated 2D Gaussian profile, but not individual pixel intensity. The list of N brightest objects is generated for each image. The brightest N stars, according to visual magnitude, enclosing the expected field of view are extracted from the star catalog and projected to the sensor plane. To annotate the coincident star objects between each image and catalog, the plate solving method was implemented by applying the triangle similarity method. The method is based on the fact that the relative relations of the triangle edges are preserved invariant of the linear transformations: translation, rotation, and scaling. To represent triangles, the feature vectors are formed by selecting all combinations of two out of $p \cdot N$ closest neighbors for the N brightest stars, where p is the suitably chosen percentage (e.g. with $N = 100$ and $p = 7\%$, seven neighbors are considered). The feature vector is a complex number with magnitude equal to the ratio of two triangle edges and angle equal to the angle between them. Percentage p is selected as a compromise between the complexity (the total number of neighbor combinations), geometric locality, robustness, and accuracy of the method. Each hypothetically matched feature vector pair from the image and catalog, resulting from their local geometric similarity, infers a set of global mapping parameters. Only matching pairs that infer similar global transformation parameters are searched among the stars in images and stars from the catalog, thus yielding linear transformation parameters for each input image. Fig. 1 shows the image registration process and geometric transformations used for mapping the catalog to the referent image and all input images to the referent one. 2D mapping

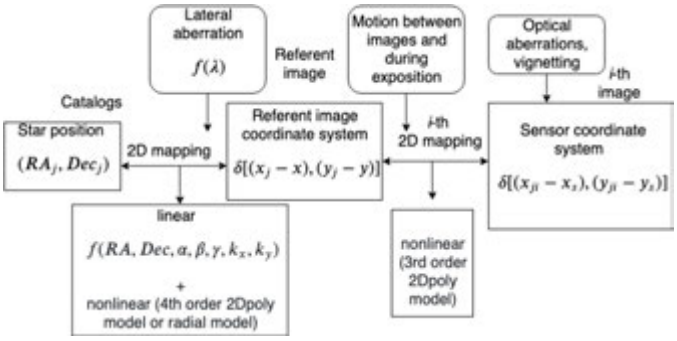


Fig. 1: Geometric transformation and image registration

of star positions from catalogs to the referent image consists of two steps. The first step is a coarse approximation using a linear transformation of the catalog star positions to the image plane. It is a function of celestial coordinates (RA, Dec), camera rotation angles (α, β, γ), and sensor angular gains (k_x, k_y). The linear transformation includes the Euler's rotations and gnomonic projection of the star positions to the referent image plane. The linear transformation model is a good fit in the image center, but can not approximate distortions at the edges. Nonlinear geometrical deformations at the edges are the result of barrel and pincushion lens distortion and projection of the 3D sphere to the 2D image plane. Consequently, a nonlinear transformation is applied. Transformations using 3rd or 4th order 2D polynomial model are sufficiently accurate for prime lenses with low distortion, while higher order polynomial models or radial models can be used for wide angle lenses

with high distortion [24], [25]. The mapping error can be also caused by lateral chromatic aberrations. Depending on the light wavelength, those aberrations cause spatially dislocated projections of the same star position in the image plane. This distortion can be accounted for by determining the nonlinear mapping for each wavelength of interest. Star positions in the referent image coordinate system are mapped to the *i*-th image coordinate system using a nonlinear polynomial model. The least mean squares method is applied to find the linear and nonlinear model parameters. The accuracy of image registration is verified by measuring the distances of the expected object positions from the actual positions. The linear model contains an error of uncompensated nonlinearity. The nonlinear polynomial models compensate the nonlinearity error.

B. Motion estimation

The motion blur effects are the consequences of long exposure, object motions, as well as unwanted residual motions of the imaging system during exposure. For the astronomical images, objects are static and blur is caused by the residual tracking error of the motorized telescope mount (e.g. declination drift, RA periodic and non-periodic error). The motion trajectory for each pixel position on the sensor plane can be constructed from the image registration step (i.e. from the identified 2D mapping model between the referent and *i*-th image) and from the time stamps of each image of the stack. Since the spatial displacement is estimated from discrete time samples corresponding to time instances of individual input images, continuous time motion trajectory is created by the cubic spline interpolation through these samples. This continuous motion model is then used to synthesize motion blur using the temporal integration of its PSF.

C. Devignetting

Vignetting is a reduction of brightness in the corners and at the edges of the image frame. The vignetting function can be estimated directly from the pre-processed image stack under the assumption of homogenous sky illumination i.e. by relying on the skyglow that is the diffuse luminance of the night sky, commonly caused by light pollution. The model assumes a radially symmetric smooth two-dimensional function. The center of the function is initially placed at the image center. The medians of pixels intensities for different radial distances from the center are calculated. These medians are used for fitting a synthetic vignetting model to the image stack. Radial intensity drop-off function is modeled by a sum of cosine functions (i.e. by its Fourier representation). By retaining only low-order (low spatial frequency) coefficients, the smoothness of the two-dimensional radially symmetric function can be ensured. The center of the estimated vignetting function can be fine-tuned in the iterative optimization process by minimizing the fitting error to the input stack. Additionally, the 4th order 2D polynomial function is also estimated and applied to fit the non-radially symmetric vignetting effects of the lens. The final model is the cascade of these two functions. To correct the images, the inverse function is simply applied to input images, thus flattening the image background.

D. Denoising

The important cause of noise in astronomical images is the long exposure time. The noise, which increases with exposure time and sensor temperature, is called dark current noise. The dark frame can be estimated once, prior to observations, as a part of sensor calibration procedure and subtracted from the raw input images in the post processing steps. Alternatively, dark frame can be captured by the camera itself after each individual light frame and subtracted immediately from the captured light frame. The advantage of the latter approach is better temporal tracking of the dark current variability (primarily due to increasing sensor temperature), but the effective exposure time is halved due to necessity of capturing dark frames with the same time duration. Poisson noise is another type of noise distributed across the whole frame caused by the quantum nature of the light. Unlike the additive Gaussian noise, Poisson noise is signal-dependent noise and it is difficult to separate it from a useful signal. It is proofed [26] that its standard deviation is proportional to the square root of pixels intensities. The transformation like Anscombe transformation transforms Poisson noise to Gaussian noise and enables easier Poisson noise removal. In our approach, instead of Anscombe transformation, the spatial 2D model of temporal noise standard deviation is directly estimated from the registered stack images after vignetting compensation. It is approximately equal to the square root of the reciprocal vignetting model. The processed stack with the removed flat background can then be divided by the estimated noise model, thus yielding the final image that is slightly attenuated at the edges but has a constant value of noise standard deviation across the whole frame. Such image can then be non-linearly mapped in pixel magnitudes to mask the noise bellow the chosen threshold.

E. Imaging system modeling

Point spread function (PSF) is the impulse response of an imaging system to a point object. Optical Transfer Function (OTF) represents the Fourier Transform of PSF. OTF is a good measure of imaging system quality. It is a complex function composed of Modulation Transfer Function (MTF) and Phase Transfer Function (PSF). To find the impulse response of the whole imaging system, the convolution of individual PSFs is necessary. Different physical aberrations that occur during imaging are modeled by their PSFs. Convolution of PSFs in the spatial domain represents the multiplication of their OTFs in the frequency domain, but only if the PSFs are shift-invariant. The convolution of the point object and PSF produces the effect of blurring on the image plane.

1) Diffraction PSF: Diffraction PSF is caused by Fraunhofer (far-field) diffraction of the light on the lens aperture. It represents the dispersion of the light intensity in the spatial domain. The shape and size of the diffraction PSF depend on the aperture size and shape. For a circular aperture, light diffraction produces an Airy pattern [22]. The width of the

Airy pattern is determined by the wavelength of the light, lens f-number, and pixel pitch. It is shift-invariant and represents the upper limit of imaging system resolution. The normalized Airy pattern function can be written as:

$$PSF_{Airy} = \left(\frac{2J_1(x)}{x}\right)^2 \tag{2}$$

where J_1 is the Bessel function of the first kind. The example of the Airy pattern is shown in Fig. 2.

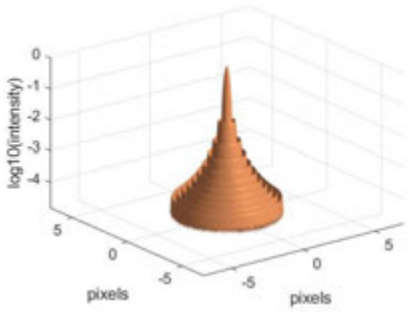


Fig. 2: 3D Airy pattern with parameters: N_{rings} = 10, f-number = 2.8, PixelPitch = 6.5 μm, λ = 650 nm

2) Bivariate Gaussian PSF: Gaussian distribution coarsely approximates all other physical effects present in the imaging process. It includes monochromatic aberrations such as defocus, spherical aberrations, field curvature, coma, and astigmatism, as well as longitudinal chromatic aberrations, electrical cross-talk in the sensor, etc. The estimation of the Gaussian parameters should be made separately for each of three color channels. The uncorrelated form of bivariate Gaussian model is used. It can be calculated as:

$$\mathbf{R} = \begin{bmatrix} cos(\theta) & sin(\theta) \\ -sin(\theta) & cos(\theta) \end{bmatrix} \quad \mathbf{S} = \begin{bmatrix} \lambda_1^2 & 0 \\ 0 & \lambda_2^2 \end{bmatrix} \quad \mathbf{C} = \mathbf{RSR}'$$

$$f(\mathbf{v}) = \frac{1}{2\pi\sqrt{\det(\mathbf{C})}} \exp\Big[-\frac{1}{2}\Big((\mathbf{v}-\mu)^T\mathbf{C}^{-1}(\mathbf{v}-\mu)\Big)\Big] \tag{3}$$

where λ_1 and λ_2 are standard deviations in directions of major and minor semiaxes, θ is rotation angle, $\mathbf{v} = [\mathbf{x}; \mathbf{y}]$ is vector of positions, and $\mu = [\mu_x; \mu_y]$ is peak position. The parameters (λ_1, λ_2, θ) are estimated from the point source objects across the whole image frame by optimization algorithms. These parameters are spatially variable and depend on the part of the frame. The example of Gaussian is shown in Fig. 3.

3) Optical low pass filter: The ideal low-pass filter prevents aliasing and Moiré effect by completely blocking the signal above the Nyquist spatial frequency and passing through the whole signal below it. The ideal brick-wall filter would cause the undesirable ringing at the Nyquist frequency in the spatial domain. The alternative filter usually found in digital cameras is a beam splitter. The beam splitter consists of two birefringent plates, which split the light spot in the horizontal and vertical direction by the known distance d [21] resulting with four spots at the distance of d in both directions. The PSF of such splitter is a sum of four Dirac Delta impulses separated in both spatial directions. The magnitudes of spots are decreased to one-quarter of their initial intensities. The cutoff frequency where MTF hits the first zero depends on the displacement d.

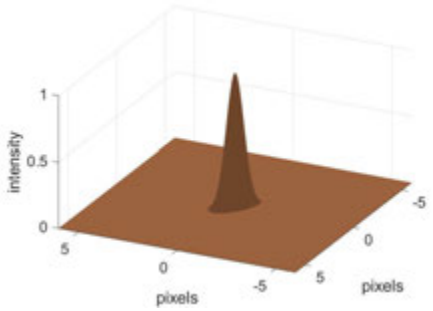


Fig. 3: 3D Gaussian of the width $k_\lambda \cdot \max(\lambda_1, \lambda_2)$ with parameters: λ_1 = 3 μm, λ_2 = 1.5 μm, $\theta = \pi/3$, k_λ = 3, PixelPitch = 6.5 μm

4) Box function: The pixel width box function is used. This function integrates spatially continuous light field into the pixel intensity. The implementation of this convolution can be numerically efficient performed by accumulation and differentiation of the over-sampled light field representation (2D CIC filter). Fig. 4 shows the observation model and the process of obtaining the system PSF which is the convolution of all PSFs in the imaging system.

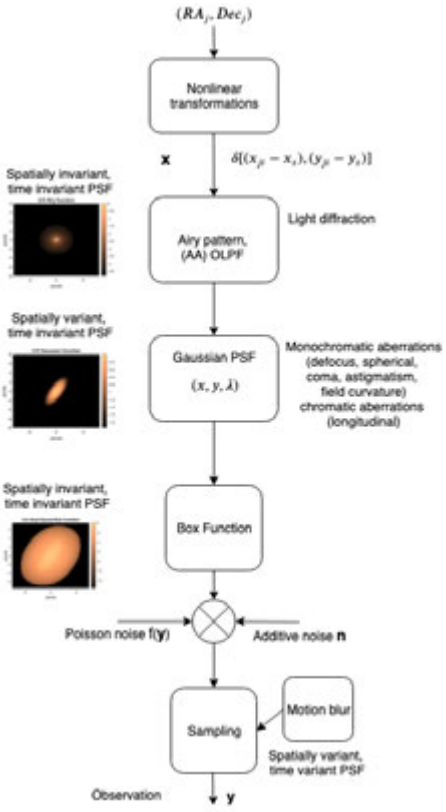


Fig. 4: The observation model

IV. Experiments and result

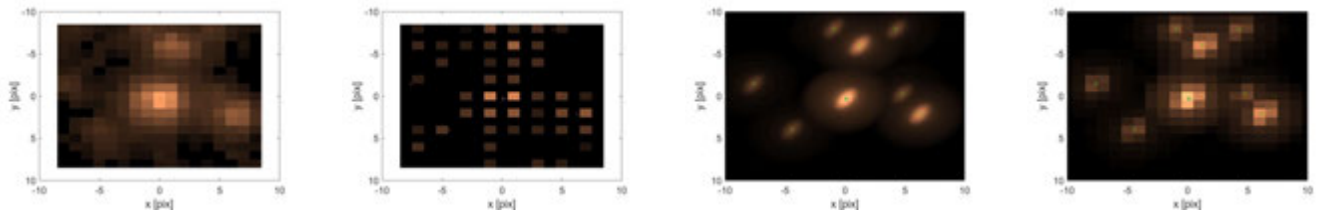
The method follows the conventional simple astronomic stack processing pipeline and consists of the following steps:

1) Plate solving of individual images in order to annotate the N brightest star contained in the image and identify geometric transformation for each particular image
2) Residual sensor motion estimation from identified geometric transformation (declination drift and right ascension periodic and non-periodic errors)
3) Spatial image registration of all input images to a chosen referent image by using identified coincident pairs found in plate solving step, thus producing a sequence of geometrically aligned images which can be stacked together with spatial transformation models
4) Initial stacking resulting with the processed image with reduced noise which is used for vignetting and noise model estimation
5) Construction of the inverse vignetting model and its application to input frames
6) New stacking of flattened input frames by using the pixel noise model for inverse variance weighting
7) Flat background removal
8) Lateral chromatic aberrations compensation by spatial registration of RGB channels of the final stack
9) Noise spatial flattening by estimation of spatially smooth temporal noise model and application of the inverse noise model to the stacked image
10) Identification of the final linear and non-linear geometric 2D mapping between stacked image and used catalog
11) Detailed analysis of all of the expected point sources in the processed image for the given catalog resulting with the spatially variant PSF model
12) Stack deconvolution using smooth spatially variant PSF model

The images were not taken from some of the available datasets but were captured by the SLR camera in a two-hour observation set. Camera parameters and full-stack information are given in Table 1. The total time for capturing one image was 40 s including exposure time, dark frame subtraction and image saving, thus the total number of the captured images was 180. The subset of images used in stack processing was obtained by rejecting the highly illuminated and blurred images (caused by the external light sources, the appearance of clouds, the guide system errors, gusts of wind). The advantages of the image stack processing versus the individual image processing are simply outliers removal (wrong pixels and individual objects (satellites, airplanes) removal) and noise reduction by the factor of the square root of the number of averaged frames [27]. The star positions are extracted from the HR, TYCHO-2 and GSC2.2 catalogs which consist of 9000, 2.5 million and 455 million stars respectively. Input images of the stack are approximately centered at NGC7635 (also known as Bubble Nebula) and the initial plate solving was performed using expected star positions from the HR and TYCHO-2 catalogs of the same sky region. For the synthetic image generation, only the narrow central image region around the TYC4279-1488-1 target star was considered in experiments. Positions of all point sources in this region were then extracted from the GSC2.2. catalog with 6 arc minute radius (658 objects). In GSC 2.2. catalog object magnitudes are defined in photographic red, blue and infrared passbands (R_F, B_J, I_N). Objects, falling within the rectangular region of 10 pixels around target star with a magnitude brighter than $R_F = 15$ and $B_J = 16$, were retained for the synthetic image generation. The system PSF, which is a convolution of Airy function, Gaussian and box function obtained by the spatial integration of the continuous light field on the interpolated upsampled grid and sampled into sensor grid, was placed at the fractional position of the point source on the interpolated sensor grid of each image. The PSF parameters are estimated on the image stack separately for each of the three color channels. The quality of the synthesized image is evaluated only by visual inspection. Fig. 5a shows the blue channel of the processed stack image previously demosaiced in the stack processing pipeline. Fig. 5b shows blue channel values of one raw image selected from a series of input images. Fig. 5c shows the synthesized continuous light field on the interpolated upsampled grid and finally Fig. 5d shows the sampled light field in the sensor image plane. The processed stack image in Fig. 5a has been already demosaiced, so it is can not be directly compared with the synthesized image. The second problem for synthetic image quality evaluation is different spectral responses of the space telescope and the camera sensor. The solution might be achieved by a predictive model for expected red, green and blue magnitudes derived from the available catalog magnitudes. To sum up, it can be seen that the proposed method can generate a good visual representation of the real image using only the estimated parameters. However, more precise synthesis can be achieved

TABLE I: Project parameters

Target Name	NGC7635
Exposure Time (EXIF)	15.0 s
f-number (f/Av)	1.4
ISO Speed	800
Camera focal Length	105.0 mm
Sensor Resolution	5505 x 3670 pix
Sensor pixel pitch	6.50 μm
Sensor crop factor	1.01 (35 mm equivalent)
Camera focal Length (35mm equivalent)	105.5 mm
Expected horizontal angle of view	19°20.9'
Expected vertical angle of view	12°59.1'
Expected diagonal angle of view	23°10.1'
Minimum required catalog radius	695.0 min
Image Resolution	5472 x 3648 pix
Expected image scale	12.775 arcsec/pix
Reference star Catalog radius	765.0 min
Number of reference stars in Catalog	62327
Expected number of reference extracted radial segment Catalog stars in frame	30665
N number of extracted stars from frame for plate solving	1000
Catalog star magnitude	< 15
Number of good light frames in stack	177



(a) Demosaiced processed stack image

(b) Blue channel values of one raw image selected from a series of input images

(c) Synthesised image, continuous light field on upsampled grid

(d) Synthesised image sampled on sensor grid

Fig. 5: The extracted region around the TYC4279-1488-1 - Blue channel

if the model parameters are tuned to the parameters of the real image.

V. Conclusion

The paper describes the method for synthetic astronomical image generation that includes all expected image degradations of a typical imagining system. The current implementation includes optical distortion, aberrations and vignetting, camera motion effects, electrical sensor properties, sensor noise, and spectral response of the optics and sensor. Parameters of individual degradation models are derived from a typical astronomical stack processing pipeline and are fed to the synthesis algorithm. Parameters need to be finely tuned to the actual imaging parameters for accurate modeling. Experimental results show that the synthesized image is close to the real one, but further parameter manipulation with objective similarity measures are necessary to obtain even more precise results. Such a synthetic image will represent the ground-truth for further research in image reconstruction and restoration using a compressed sensing concept.

VI. Acknowledgement

This work was partly supported by: Croatian Science Foundation under the project number IP-2014-09-2625 and DOK-2018-01-2976; DATACROSS project under number KK.01.1.1.01.009. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Croatian Science Foundation.

References

[1] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine vision and applications*, vol. 25, no. 6, pp. 1423–1468, 2014.

[2] S. Zhao, R. Jin, X. Xu, E. Song, and C.-C. Hung, "A variational bayesian superresolution approach using adaptive image prior model," *Mathematical Problems in Engineering*, vol. 2015, 2015.

[3] R. G. Baraniuk, "Compressive sensing [lecture notes]," *IEEE signal processing magazine*, vol. 24, no. 4, pp. 118–121, 2007.

[4] F. Ochsenbein, P. Bauer, and J. Marcout, "The vizier database of astronomical catalogues," *Astronomy and Astrophysics Supplement Series*, vol. 143, no. 1, pp. 23–32, 2000.

[5] J. Yang and T. Huang, "Image super-resolution: Historical overview and future challenges," *Super-resolution imaging*, pp. 20–34, 2010.

[6] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE signal processing magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[7] J. Tian and K.-K. Ma, "A survey on super-resolution imaging," *Signal, Image and Video Processing*, vol. 5, no. 3, pp. 329–342, 2011.

[8] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling [a sensing/sampling paradigm that goes against the common knowledge in data acquisition]," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[9] M. Schöberl, J. Seiler, S. Foessel, and A. Kaup, "Increasing imaging resolution by covering your sensor," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 1897–1900.

[10] J. Seiler, M. Jonscher, M. Schöberl, and A. Kaup, "Resampling images to a regular grid from a non-regular subset of pixel positions using frequency selective reconstruction," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4540–4555, 2015.

[11] M. Jonscher, J. Seiler, D. Lanz, M. Schöberl, M. Bätz, and A. Kaup, "Dynamic non-regular sampling sensor using frequency selective reconstruction," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[12] L. Denis, E. Thiébaut, F. Soulez, J.-M. Becker, and R. Mourya, "Fast approximations of shift-variant blur," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 253–278, 2015.

[13] P. Campisi and K. Egiazarian, *Blind image deconvolution: theory and applications*. CRC press, 2017.

[14] B. K. Karch and R. C. Hardie, "Adaptive wiener filter super-resolution of color filter array images," *Optics express*, vol. 21, no. 16, pp. 18 820–18 841, 2013.

[15] S. Farsiu, M. Elad, and P. Milanfar, "Multiframe demosaicing and super-resolution of color images," *IEEE transactions on image processing*, vol. 15, no. 1, pp. 141–159, 2005.

[16] M. Wilcox, "How to measure mtf and other properties of lenses," *Optikos Corporation, Cambridge, MA, USA, Tech. Rep*, pp. 4–04, 1999.

[17] H. Nasse, "How to read mtf curves," *Carl Zeiss, Camera Lens Division*, 2008.

[18] F. van den Bergh, "On the rendering of synthetic images with specific point spread functions," in *Twenty-Third Annual Symposium of the Pattern Recognition Association of South Africa*, 2012, p. 75.

[19] K. Fliegel, "Modeling and measurement of image sensor characteristics," *RADIOENGINEERING-PRAGUE-*, vol. 13, no. 4, pp. 27–34, 2004.

[20] G. D. Boreman, *Modulation transfer function in optical and electro-optical systems*. SPIE press Bellingham, WA, 2001, vol. 4.

[21] "A SIMPLE MODEL FOR SHARPNESS IN DIGITAL CAMERAS – AA," https://www.strollswithmydog.com/resolution-model-digital-cameras-aa/, 2016.

[22] J. W. Goodman, *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.

[23] J. C. Wyant and K. Creath, "Basic wavefront aberration theory for optical metrology," *Applied optics and optical engineering*, vol. 11, no. part 2, pp. 28–39, 1992.

[24] D. C. Brown, "Decentering distortion of lenses," *Photogrammetric Engineering and Remote Sensing*, 1966.

[25] P. Drap and J. Lefèvre, "An exact formula for calculating inverse radial lens distortions," *Sensors*, vol. 16, no. 6, p. 807, 2016.

[26] S. W. Hasinoff, "Photon, poisson noise." 2014.

[27] G. Sluder and D. E. Wolf, *Digital microscopy*. Academic Press, 2013.

Novel Class Detection in Non-stationary Streaming Environment with a Discriminative Classifier

R. Šajina*, N. Tanković* and D. Etinger*

*Faculty of Informatics, Juraj Dobrila University of Pula, Croatia
{romeo.sajina, nikola.tankovic, darko.etinger}@unipu.hr

Abstract - In a data streaming environment, one of the biggest challenges for a machine learning classifier is to detect the changes in the concepts that the data corresponds to - a phenomenon called Concept Drift. It can manifest in different modes: existing classes can continually evolve, experience a sudden shift, or a novel class can emerge. Algorithms for novel class detection using cluster-based techniques and statistical approaches applied to the model outputs rely on the assumption that the feature space of the data posses some distance metric governing the class affiliation. Therefore, the novel class will correspond to a significant distance from the known clusters. Most of the time, these assumptions are correct, but the resulting algorithms are challenging to apply on higher-dimensional data such as images. In this paper, we present a novel approach called Discriminative Classifier Detector (DCD) for detecting concept evolution. DCD trains alongside the classification model. The primary model used to evaluate our approach is a Convolutional Neural Network (CNN) classifier for which the proposed DCD is a densely layered neural network. DCD is applied to the outputs of CNN's penultimate layer to achieve its independence against the number of output classes, which enables the underlying CNN model to evolve independently. DCD requires no structural change when the detected novel class is added to the CNN classifier. We demonstrate the effectiveness of our approach on several well-known datasets.

Keywords - ***Covolutional Neural Network, Novel Class Detection, Concept Evolution, Discriminative Classifier***

I. Introduction

Classification models (CM) in a non-stationary environment yield many challenges that are largely tied to the variety and unpredictability of data. Most importantly, these challenges involve detecting and handling these changes continuously, a phenomenon called *Concept Drift*.

CMs should operate in a semi-supervised mode while working with streaming data that is subject to concept drift. CM must predict labels on-the-fly and in parallel keep track of the fact whether data is experiencing any changes. Depending on the employed change detection algorithm and incremental learning approach, the model will be continually retrained with additional labeled data, but these labels will originate from model predictions. Change detection algorithms are the critical component to prevent drastic performance degradation of CMs.

Besides concept drift, where instances from known classes deviate from concepts in the training dataset [5, 6], we can also observe ***Concept Evolution***, in which a new class previously unknown to the model can emerge [2, 7, 8, 12].

In this paper, we present the Discriminative Classifier Detector (DCD) that detects *Concept Evolution* in data streams. The fundamental idea of DCD is to exploit model output behavior when feeding an instance of a novel class, as shown in [9], i.e., the model will behave differently when classifying instances of a known or unknown class. Our approach assumes that the model will extract a strong feature representation from the training dataset. Our approach can easily be integrated with any existing classification models that possess an internal feature representation. Additionally, our approach does not require any structural changes to the originating model.

We can summarize the contributions of this paper as follows:

1) a class detector architecture that enables easy integration with new or existing classification models,
2) a domain-independent data augmentation procedure that retrains the DCD with positive and negative examples.

II. Related work

Early research on concept evolution in a neural network focused on analyzing probabilities produced by the softmax function [9] where the probability can be exploited to determine if the instance is of known or novel class.

Research by [7, 4, 2] focused on detecting concept evolution by employing the clustering algorithm and declaring

an instance being of novel class if it is proclaimed as an outlier by clusters.

Research by [12] shows that the classification network can create a robust internal feature representation of the training data. Furthermore, a strong feature representation can be observed in the output layer, but even stronger representation in the penultimate layer. This idea is applied in [12] to create an ensemble of *prototypes* where they represent properties of the observed class. Furthermore, instances with the same properties are joined in the same *prototype* using the clustering algorithm.

One of the intuitive solutions for novel class detection is training an independent network directly on training data and noise. Then, a new instance is first fed into this network where it is decided if the instance is of known or novel class. Based on a similar principle, [9] proposes to use an auxiliary decoder, where a layer is added to the classification network after softmax output, to reconstruct the input. The reconstructed image is then fed into a smaller network, called abnormality module, where the decision is made whether an instance is of known or novel class. While training, the abnormality module sees the reconstructed image from clean instances and, as negative examples, sees the reconstructed image from distorted instances with either white noise, brown noise, or pink noise [9].

III. Methodology

To demonstrate the superiority of our approach, we will establish a Baseline for novel class detection, which will be discussed in section III-B.

Both approaches followed the same evaluation principle, using the same CNN network and dataset in the process. The CNN network is a simple network with two convolution layers, two fully connected layers each followed by a dropout layer with a dropout rate of 50%, and a fully connected output layer counting altogether approximately 28000 parameters. Three datasets will be used in the evaluation: MNIST [1], Fashion-MNIST [10] and Kuzushiji-MNIST [11]. MNIST is a dataset of ten handwritten digits, Fashion-MNIST is a dataset of ten articles of clothing, and Kuzushiji-MNIST is a dataset of ten Kuzushiji (cursive Japanese) characters. Figure 1 shows examples instances from all datasets.

All datasets are containing around 60000 instances of 10 different classes. For the simulation of the streaming environment, we will randomly select 50% of classes (i.e., five classes) and for each selected class randomly select between 400 and 1000 instances, thus selecting up to 5000 instances in total. Unselected data will be used to simulate incoming data in a streaming environment. With previously mentioned selected instances, we will train the initial CNN model, after which a chosen novel class detector will be notified.

We will simulate a streaming environment by randomly selecting classes for each next iteration where each class has a 70% probability of being selected. There is special behavior for the first few iterations which will include only initially selected classes, thus not selecting any novel classes. For each selected class in the iteration, we will randomly select between 200 and 500 instances that will be introduced to the detector. The detector will then classify given instances as either instance of a known class or instance of a novel class. Instances that are classified as being from a known class are labeled by predicted class from the CNN model, and instances that are classified as instances of a novel class are labeled with a new label. In a given iteration of the simulation, instances of multiple novel classes can be introduced to the model, meaning that a clustering algorithm must be applied to properly separate instances of different novel classes. When creating clusters for instances of multiple novel classes we use the approach described in [12], i.e. we use CNN model feature representation to create clusters of novel instances, which are then accordingly labeled with a new label. All new instances are added to the training dataset with their respective labels, and the CNN model is retrained once there are more than 10000 new instances, after which the chosen detector is notified.

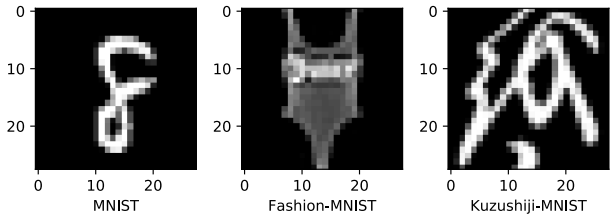


Fig. 1. Example of an instance from each dataset. We used three datasets to simulate streaming data and compare the performance of the Baseline and DCD approach.

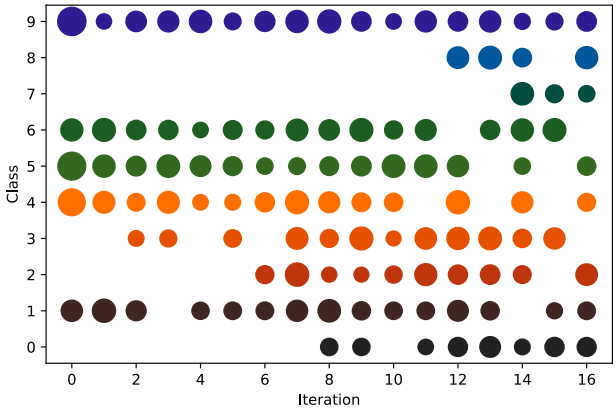


Fig. 2. Distribution of class instances over time. At the beginning of the streaming data simulation, we randomly select 50% of available classes with which the model is trained. For each selected class, we randomly select up to 1000 instances and introduce them to the simulation. The next few iterations include only initially selected classes, i.e. no novel classes will be selected into the iteration. In each upcoming iteration, we randomly select classes that will be used in the evaluation where each class has a 70% of being included in the iteration. For each selected class we randomly select up to 500 instances which are shown on the image as circle size.

For each simulation process, we repeat the process of randomly selecting classes, selecting respective instances, and evaluating those instances to determine if they are of known or novel class. Simulation process is shown in figure 2. We repeat the simulations ten times and report the average performance results.

Both models were trained on 100 epochs while using Adam optimizer with a learning rate set to 0.005 and using *categorical cross-entropy* as a loss function.

A. Evaluation metrics

The evaluation of detection performance will produce a confusion matrix where FP denotes the total novel class instances misclassified as an existing class, FP denotes the total existing class instances misclassified as a novel class.

We use two metrics to describe the performance of novel class detector: F_{new} as metric to describe the percentage of existing class instances misclassified as a novel class and M_{new} as metric to describe the percentage of novel class instances misclassified as an existing class. They are calculated as follows:

$F_{new} = \frac{FP*100}{N-N_c}$

$M_{new} = \frac{FN*100}{N_c}$

with N being the number of all instances, and N_c is the number of novel class instances.

Finally, we calculate the overall accuracy of a detector $Accuracy = \frac{A_{known}+A_{new}}{n}$, where A_{known} is the number of known class instances identified correctly, A_{new} is the total number of novel class instances classified correctly, and n is the number of evaluated instances. The described metrics are based on the previous research [12, 7].

B. Baseline

When working with neural network classification, a Softmax function is used to convert network output to class probabilities. In the process of training, the goal is to achieve as highest target class probability as possible. In the example of the MNIST dataset, an average output probability of a trained network is 99% and may slightly decrease on an unseen dataset (98%). When the network is classifying an unknown input, output probabilities can significantly drop. In the paper [9], authors exploit feeding different unseen inputs to a trained network to analyze network output probabilities. In the example of the MNIST trained network, feeding a uniformly generated input will produce average probabilities of 83%, while feeding a notMNIST [3] dataset will produce average probabilities of 92%. Based on the analyzed difference of the output probabilities in scenarios where the network is classifying known data and unseen data, we can establish a Baseline for detecting novel classes. Baseline detector will calculate average $\overline{p} = avg(max(softmax(x)))$ and standard deviation $\sigma_p = \sigma(max(softmax(x)))$ of probabilities generated by x, where x is training dataset. That information will be used while detecting novel classes in such a way that the class will be declared novel if the probability of the evaluated instance x' meets the condition $max(softmax(x')) < (\overline{p} - \sigma_p)$, otherwise, it will be declared as a known class.

IV. Discriminative Classifier

In this paper, we propose a Discriminative Classifier Detector (DCD) to determine if an instance is of novel or known class.

DCD is based on the assumption that the model will behave differently when it sees an instance of a novel class versus when it sees an instance of a known class. The analyzed model output probabilities in section III-B confirmed this assumption by outputting significantly lower probabilities when classifying instances of a novel class.

DCD is a small neural network model with two fully connected layers using ReLU as the activation function. The first layer contains 512 neurons; the second layer contains 256 neurons, while the last layer has two neurons to describe known or novel class. The input to the DCD will be the CNN model feature representation, i.e., outputs from the penultimate layer. DCD will be trained after the CNN network completed its training. DCD will be trained with the feature representations of the original training dataset produced by the CNN model, and they will be labeled with a class 0, therefore classifying the inputs as known classes.

The problem emerges when we need to generate a dataset to accommodate novel classes feature representation by the CNN model. Simply generating a random noise for that purpose would not bring expected results, one reason being that random noise may have a different distribution of pixels. Alternatively, random noise can be sampled from the training dataset pixels distribution, but the produced result will be an image with scattered pixels that differs from the domain.

A more robust approach to exploit CNN model behavior, when a new class emerges from a specific domain, is to use training data to generate a augmented dataset. While generating the augmented dataset we used the following parameters for data transformation:

- rotation scale = 90
- width shift range = 30
- height shift range = 0.3
- brightness range = (0.76, 1.3)
- shear range = 0.3
- zoom range = 0.2

An example of an original MNIST instance alongside augmented instance is shown in figure 3.

Generated augmented dataset is then fed into the CNN model to produce their respective feature representations which will then be labeled as a novel class (label: 1) when training DCD. Finally, the DCD can be trained and is expected to learn how CNN will behave when it sees instances from know or novel class. The process of generating a dataset for DCD and model structures is shown in figure 4.

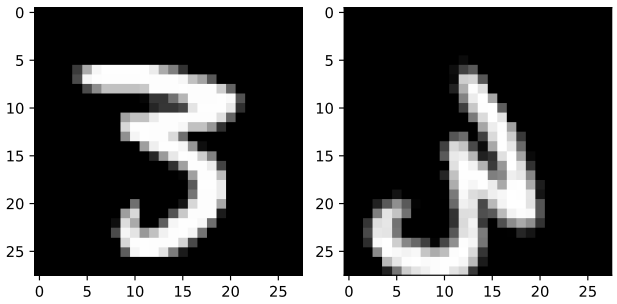


Fig. 3. Example of the original instance from MNIST dataset on the left and the same instance transformed on the right. The transformed dataset is used to produce positive examples, while the original dataset is used to produce negative examples in the process of creating a dataset for training DCD.

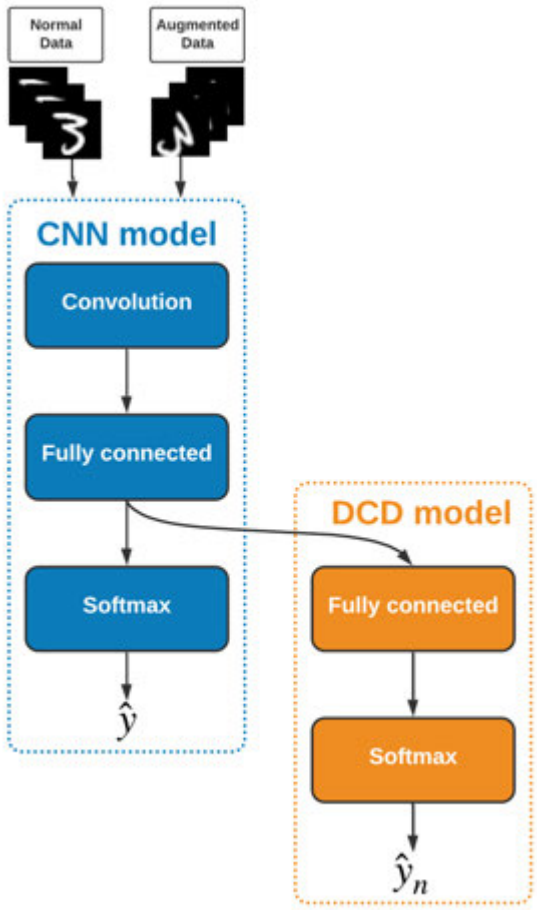


Fig. 4. CNN and DCD model architecture. DCD model is connected to the CNN model's penultimate layer and for training using layer outputs that are collected by feeding generated positive and original negative examples.

V. Evaluation

Simulation results of Baseline and DCD method across datasets separately observing F_{new} and M_{new} metrics are shown in table I.

Comparing the methods we see that DCD outperforms Baseline on all metrics on all datasets. Furthermore, DCD provides 2.27% lower F_{new} on MNIST dataset, 12.91% lower F_{new} on Fashion-MNIST dataset and 6.90% lower F_{new} on Kuzushiji-MNIST dataset over Baseline. Comparing methods on M_{new} metric shows smaller performance margin of DCD over Baseline, i.e. DCD provides 1.32% lower M_{new} on MNIST dataset, 2.06% lower M_{new} on Fashion-MNIST dataset, and 0.78% lower M_{new} on Kuzushiji-MNIST dataset over Baseline.

To confirm the superiority of the DCD method we will compare their respective $Accuracy$ metrics as discussed in section III-A. A comparison of results achieved on $Accuracy$ metric on all datasets is shown in table II. Comparing the results shows that the DCD method outperforms Baseline on all datasets, i.e. DCD provides 2.13% higher accuracy on MNIST dataset, 12.2% higher accuracy on Fashion-MNIST dataset, and 6.36% higher accuracy on Kuzushiji-MNIST dataset. Note that in the streaming process there is a significant class imbalance as most of the instances will be of known class, therefore the accuracy will be mostly dependent on the F_{new} score.

Finally, DCD outperformed the Baseline detector on all metrics and is generally a more stable solution as it provides a significantly lower standard deviation for all metrics.

VI. Limitations and future work

Most of the approaches presented in section II assumes the appearance of a single novel class at a time which can not always be expected in real-world applications. It is worth noting that much research has been dedicated to multiple class novelty detection, but mostly on clustering based classifiers, rather than neural networks. Future work should consider extending on multiple class novelty detection when using a neural network as a classifier.

The newly proposed approaches should also try to be non-invasive on the classification model, i.e., integrate detection algorithm into the existing model structure rather than adapting the model to enable novel class detection. This statement is especially directed towards models that are already in production and are designed to work well in their domains. That way, models are able to continue operating swimmingly on their specific task but extended with independent novel class detection algorithm. In future work, we should consider comparing our approach with the state of the art approaches to evaluate if demonstrated superiority holds truth. In our approach, we used data augmentation to generate novel class examples, but this can vary across domains. In that regard, we should focus future work on creating a more general approach that will provide better and more consistent results. In the experiments, we used three datasets, all of them being black and white. In future work, we should consider

Method	MNIST		Fashion-MNIST		Kuzushiji-MNIST	
	F_{new}	M_{new}	F_{new}	M_{new}	F_{new}	M_{new}
Baseline	8.53 ± 1.10	28.26 ± 8.71	20.31 ± 3.53	62.99 ± 11.18	18.71 ± 1.86	25.57 ± 4.62
DCD	**6.26 ± 0.52**	**26.94 ± 7.79**	**7.40 ± 0.93**	**60.93 ± 6.60**	**11.81 ± 1.15**	**24.79 ± 5.32**

TABLE I
NOVEL CLASS DETECTION PERFORMANCE OVER DATA STREAMS. ALL VALUES ARE EXPRESSED IN PERCENTAGES

Method	MNIST	Fashion-MNIST	Kuzushiji-MNIST
	Accuracy		
Baseline	89.92 ± 1.03	76.49 ± 4.15	80.85 ± 4.62
DCD	**92.05 ± 0.96**	**88.69 ± 1.28**	**87.21 ± 0.90**

TABLE II
NOVEL CLASS DETECTION ACCURACY. ALL VALUES ARE EXPRESSED IN PERCENTAGES

exploiting model behavior and evaluating our approach to more complex and colored datasets.

VII. CONCLUSION

This paper proposes a novel class detector: Discriminative Classifier Detector - DCD. It exploits the assumption that the classification model will behave differently when it sees the instances of a novel class versus the instances of a known class. DCD uses the penultimate layer representation of the classification network model. We provided the procedure for creating a training dataset for DCD: the original dataset is used to produce the negative instances, and an augmented version of the original dataset is used to produce the positive instances. By comparing the DCD results with the Baseline model, we demonstrated the effectiveness of our approach on several performance metrics: DCD outperformed the baseline on accuracy metric with an average 6.89% improvement. DCD is working independently of the classification model, which makes it easy to apply on any new or existing models, regardless of the model architecture and configuration.

REFERENCES

[1] Yann LeCun, Corinna Cortes, and CJ Burges. "MNIST handwritten digit database". In: *ATT Labs [Online]. Available: http://yann. lecun. com/exdb/mnist* 2 (2010).

[2] Mohammad Masud et al. "Classification and novel class detection in concept-drifting data streams under time constraints". In: *IEEE Transactions on Knowledge and Data Engineering* 23.6 (2010), pp. 859–874.

[3] Yaroslav Bulatov. "Notmnist dataset". In: *Google (Books/OCR), Tech. Rep.[Online]. Available: http://yaroslavvb. blogspot. it/2011/09/notmnist-dataset. html* 2 (2011).

[4] Satyen Abrol et al. "Design and implementation of SNODSOC: Novel class detection for social network analysis". In: *2012 IEEE International Conference on Intelligence and Security Informatics*. IEEE. 2012, pp. 215–220.

[5] João Gama et al. "A survey on concept drift adaptation". In: *ACM computing surveys (CSUR)* 46.4 (2014), pp. 1–37.

[6] Brandon Shane Parker and Latifur Khan. "Detecting and tracking concept class drift and emergence in non-stationary fast data streams". In: *Twenty-ninth AAAI conference on artificial intelligence*. 2015.

[7] Ahsanul Haque, Latifur Khan, and Michael Baron. "Sand: Semi-supervised adaptive novel class detection and classification over data stream". In: *THIRTIETH AAAI conference on artificial intelligence*. 2016.

[8] Ahsanul Haque et al. "Efficient handling of concept drift and concept evolution over stream data". In: *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*. IEEE. 2016, pp. 481–492.

[9] Dan Hendrycks and Kevin Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks". In: *arXiv preprint arXiv:1610.02136* (2016).

[10] Han Xiao, Kashif Rasul, and Roland Vollgraf. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms". In: *arXiv preprint arXiv:1708.07747* (2017).

[11] Tarin Clanuwat et al. *Deep Learning for Classical Japanese Literature*. Dec. 3, 2018. arXiv: cs.CV/1812.01718 [cs.CV].

[12] Zhuoyi Wang et al. "Robust high dimensional stream classification with novel class detection". In: *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE. 2019, pp. 1418–1429.

Evaluation of diatoms biodiversity models by applying different discretization on the class attribute

A. Naumoski, G. Mirceva and K. Mitreski
Faculty of computer science and engineering, Ss. Cyril and Methodius University in Skopje, North Macedonia
{andreja.naumoski, georgina.mirceva, kosta.mitreski}@finki.ukim.mk

***Abstract* - One of the main goals of knowledge discovery from environmental data is through data analysis to find the relationship between the living organisms, represented with the diversity of the diatoms community members, and the characteristics of the environment. This is very important information for both ecologists and decision makers. Therefore, in this paper we apply various machine learning algorithms for revealing this relationship by using different number of discretization levels for the target attribute. The target attribute represents the biodiversity index of the community and it is calculated based on the abundances of the diatoms. For building models, different types of machine learning algorithms are considered including decision trees, rule induction algorithms, neural networks and Naïve Bayes. The obtained models are also examined regarding resistance to over-fitting, as well as statistical significance.**

Keywords – ecological modelling; biodiversity indices; diatoms; machine learning algorithm.

I. INTRODUCTION

Biodiversity indices (diversity and evenness indices) are quantitative mathematical measures that are quantifying how many different species are in the community. These indices provide richer information for decision makers than the abundances of the species. If we use these mathematical terms and correlate with the measured environmental parameters at the same time, they can provide very important information for both ecologists, biologists and decision makers [1]. The focus of habitat suitability modelling is modelling the relationship between the abiotic factors and the species' community expressed with the biodiversity indices of the species that inhabit that environment [2]. This is more evident when we use species that are bio-indicators of the environment, like diatoms. Because these organisms react rapidly with the environmental changes, they have a strong relationship with the environment, and that relationship can be modeled. However, this non-linear relationship between the diatoms community expressed with the biodiversity indices and the environment is not easy to be found. There are various approaches that can be applied to solve this problem. Therefore, the aim of this paper is to find the most suitable algorithm for that purpose.

The biodiversity indices as a mathematical construction have been adapted for other scientific disciplines and later in ecological studies [3, 4, 5]. They have been used in many ecological studies for different environments like lakes, mountains, oceans and woodlands [6, 7, 8, 9]. Other studies point out the importance of the differences and the common points of the diversity and evenness indices [10, 11, 12, 13]. Calculating the biodiversity indices is done by taking the raw abundance data for every specie in the community, and then each index is calculated separately. If the measured data is composed of information regarding the environmental stress factors, then we can relate the biodiversity indices with these abiotic factors, as it is done in habitat suitability modelling. Machine learning algorithms in many cases are used for the purpose of analyses such as clustering and classification [14, 15, 16, 17, 18, 19].

In this direction, as well as to further improve the set of learning tools used for biodiversity modelling, in this work we apply different types of classification methods, like rule induction, neural network and Bayesian algorithms, as well as the typical decision tree modelling algorithms. In [20, 21, 22], we made similar studies regarding the environmental influence on the diatoms' biodiversity indices in Lake Prespa. The results from these studies provide interesting results and information regarding the relationship between the abiotic factors of the lake ecosystem and the diatoms' biodiversity. One downside of these models is that many of them obtained medium classification accuracy and for many biodiversity indices the algorithms were not able to build model. The results of the paper [22] point out that further investigation of the biodiversity categorization is needed. So, besides testing other machine learning algorithms and how they influence on the biodiversity modelling, the paper should also focus on investigating the influence of using different number of discretization levels for the target attribute (biodiversity indices). The categorization of certain physico-chemical parameters is done similarly as for the water quality and trophic state index classes found in the literature.

The datasets used in this paper consists of 10 input parameters detailing the measured abundances of the top 10 most abundant diatoms represented with 10 biodiversity indices and several physico-chemical parameters that are measured during the monitoring program. In the evaluation analyses of the machine learning algorithms, we consider two different biodiversity models obtained by using different number of

discretization levels (3 and 5 levels) to investigate the models' accuracy and resistance to over-fitting. The over-fitting resistance is estimated by calculating the Root Mean Squared Error (RMSE) between the models' descriptive and predictive classification accuracy. The accuracies of the obtained models' are tested for statistical significance, in order to ensure that the improvement in performance is achieved not by chance or random fluctuation. For this reason, we use the two-stage procedure described in [23], combined with Aligned Friedman test [24] and Hommel test [25]. The two tests, Aligned Friedman and Hommel test, form a non-parametric statistical analysis multiple comparison procedure. During these tests, the datasets and the algorithms used to analyze that dataset are examined and ranked based on their performance. The best performing algorithm in each experiment is taken as a control algorithm and the rest of the algorithms are tested with respect to the control algorithm.

The rest of the paper is organized as follows: Section II presents the dataset description and the experimental setup, while in section III the experimental results and the discussion regarding the results are presented. Finally, Section IV concludes the paper and the research direction is outlined.

II. Data Description and Experimental Setup

The dataset used for evaluation of the algorithms consists of the data that were collected in 16 months period as part of the EU project TRABOREMA [26]. All the measurements were done according to the measurement protocol, and the diatom samples were later counted under microscope [26]. Using the data from these measurements, the ecologists and the decision makers could use the abiotic information contained in the physico-chemical measurements, as well as the biological information from the diatom measurements to set new protocols and directives. Many physico-chemical parameters were part of the water properties analytics together with the 116 different diatom species. The 116 different diatom species compromising the diatom community are converted into 10 biodiversity indices. In our experiments the following indices are considered: Brilouni [27], $E_{1\text{-}D}$ [28], $E_{1/D}$ [28], $E_{\text{-ln(D)}}$ [28], Heip [29], Hill [30], Hurlbert [31], Mcintosh [32], Simpson [33] and Shannon [34].

Since it is very difficult to know the exact value of each measured parameter in the water ecosystem, and it is even more difficult to know the number of diatoms in each time and space in the lake, the biologists and ecologists use the categorization of the given parameter, like water quality classes. In this direction, each calculated biodiversity index is more useful and expresses a more realistic picture of the lake condition if it is categorized. Therefore, each calculated biodiversity index is transformed or categorized in three (low, medium, high) and five (very low, low, medium, high, very high) categories. The biodiversity indices are divided into three and five equal intervals between the lowest and highest values for each index. Doing this, as we show in [25], will affect the accuracy of the models produced by the machine learning algorithms.

If we add these ten biodiversity indices as a target or output attributes together with the 21 measured input attributes (the physico-chemical parameters), together form the evaluation dataset. The models are evaluated by using the AUC-ROC evaluation measure. The descriptive performaces of the models are obtained by using the entire dataset for training and testing. On the other hand, the predictive performaces are obtained by using the standard 10-fold cross-validation procedure. For this evaluation, we use several types of classification methods: Rule induction (OneR [35], NNge [36, 37], PART [38] and Ripper [39]), Bayesian modelling (Naïve Bayes [40]), Neural Network algorithms (RBF Network [41] and Multilayer Perceptron [42]) and decision tree modelling algorithms (BF Tree [43, 44], FT [45], C4.5 [46], LAD Tree [47], LM Tree [48], NB Tree [49] and REP Tree [50]).

For further confirmation of the achieved results from the algorithms, we examine them using the two-stage procedure proposed in [23] for statistical significance. This evaluation is done for the results obtained for both the descriptive and predictive power of the models. In the two-stage procedure, the first part is using the non-parametric Aligned Friedman procedure [24] is used due to the fact that the number of methods (variants) is small, and their ranks should be obtained using the aligned procedure. It is important to note that the rejection of the null hypothesis by the Aligned Friedman procedure means that there is a significant difference among the datasets, not the performances of the models obtained from these datasets. The level of significance is set to 0.05 in our experiments. In case when the null hypothesis is rejected, then the two-stage procedure, starts with the post hoc Hommel test [25]. Since, this procedure, changes the level of significance in each iteration, the results that we have considered are the adjusted p-values to make a fair comparison between each p-value. Considering this that we discussed here in this paragraph, in this paper we present the average and the highest rank as well as the rejected hypothesis of each experiment. For more information on how the two-stage procedure works in details, we kindly ask the reader to read the paper presented in [23].

III. Experimental Results

A. Performance Analysis

In this section, we present the experimental results for the descriptive and predictive performances of the models by using various classification algorithms and biodiversity indicies by using 3 and 5 categories for the class attribute.

The AUC-ROC values obtained in the analysis of the descriptive performances when using 3 categories for the class attribute are presented in Table 1. From these results, it is clear that we have variable AUC-ROC values for the models for the biodiversity indices. On the other side, the best algorithm to build a model is NNge algorithm, which is a nearest-neighbor like algorithm.

TABLE I. THE AUC-ROC VALUES OBTAINED IN THE EVALUATION OF THE DESCRIPTIVE PERFORMANCES OF THE CLASSIFCATION ALGORITHMS BY USING 3 CATEGORIES FOR THE BIODIVERSITY INDICES. THE HIGHEST VALUE PER ALGORITHM IS BOLDED.

Algorithm / Biodiversity index	BFTree	FT	C4.5	Ripper	LADTree	LMT	Multilayer Perceptron	NaiveBayes	NBTree	NNge	OneR	PART	RBFNetwork	REPTree
Brilouni	0.50	0.81	0.50	0.50	0.88	0.69	**1.00**	0.65	0.50	**1.00**	0.64	0.50	0.74	0.69
$E_{1\text{-}D}$	**0.72**	0.80	**0.99**	0.57	0.83	0.61	0.96	0.66	**0.71**	**1.00**	0.66	0.93	0.76	0.69
$E_{1/D}$	0.50	0.64	0.61	**0.74**	0.82	0.66	0.98	0.67	0.61	**1.00**	0.69	0.61	0.75	0.66
$E_{-ln(D)}$	0.50	0.79	0.96	0.61	0.80	0.64	0.97	0.68	0.64	0.99	0.66	0.63	0.71	0.50
Heip	0.63	0.87	0.92	0.58	0.84	0.58	0.93	0.64	0.65	0.99	0.64	0.81	0.71	0.50
Hill	0.50	**0.95**	0.50	0.50	**0.97**	0.67	0.89	**0.79**	0.50	**1.00**	0.53	0.50	**0.87**	0.50
Hurlbert	0.50	0.86	0.98	0.59	0.82	**0.99**	0.98	0.66	0.50	0.99	0.67	0.60	0.65	**0.81**
Mcintosh	0.50	0.82	0.85	0.61	0.83	0.65	0.99	0.67	0.63	**1.00**	**0.70**	0.78	0.68	0.64
Simpson	0.61	0.85	0.96	0.68	0.91	0.74	0.93	0.69	0.66	**1.00**	0.58	**0.94**	0.77	0.61
Shannon	0.50	0.79	0.72	0.62	0.87	0.68	0.97	0.65	0.50	**1.00**	0.66	0.72	0.75	0.50

The AUC-ROC values obtained in the evaluation of the predictive performances of the models when using 3 categories (see Table 2) present different situation. Table 2 shows that the AUC-ROC values of the various biodiversity indices are comparable, where the Simpson and the Shannon biodiversity indices show slightly better results. Regarding the algorithms, we can note that LADTree and LMT algorithms achieve better results compared with the rest of the analyzed algorithms. This means that the decision trees class of methods here achieves slightly better results compared to the rule induction and neural network algorithms. Furthermore, the RMSE between the descriptive and predictive performances (presented in the last column in the table) shows that the models with the Shannon biodiversity index have highest resistance to over-fitting (lowest RMSE of 0.25), while most of the other models obtain comparable RMSE (between 0.26 and 0.28).

Table 3 and Table 4 depict the influence of using different number of categories for the class attributes. If we compare Table 1 and Table 3, which present the descriptive performances by using 3 and 5 categories, we can see similar results in both cases. Slightly worse and slightly better results are noticeable through the tables, but it is very clear again that the NNge algorithm is the best. However, if we continue to investigate the predictive power of the models on unseen data, presented in Table 2 and Table 4, this algorithm is among the worst. Comparing the other biodiversity models, we have diverse AUC-ROC values across the tables. Compared with Table 2, where we have pinpointed some of the best biodiversity indicies (Simpson and Shannon), here it is hard to say for the predictive performances. The same conclusion is backed up by the RMSE values calculated for these experiments. As a conclusion of these experiments, we can say that categorization of the class attribute with a higher number of levels does not always lead to better results in a case of diatoms biodiversity datasets.

TABLE II. THE AUC-ROC VALUES OBTAINED IN THE EVALUATION OF THE PREDICTIVE PERFORMANCES OF THE CLASSIFCATION ALGORITHMS BY USING 3 CATEGORIES FOR THE BIODIVERSITY INDICES. THE HIGHEST VALUE PER ALGORITHM IS BOLDED. THE HIGHEST RESISTANCE TO OVER-FITTING IS UNDERLINED.

Algorithm / Biodiversity Index	BFTree	FT	C4.5	Ripper	LADTree	LMT	Multilayer Perceptron	NaiveBayes	NBTree	NNge	OneR	PART	RBFNetwork	REPTree	RMSE
Brilouni	0.49	0.52	0.42	0.46	0.48	0.55	0.47	0.51	0.45	0.45	0.45	0.43	0.53	0.47	0.27
$E_{1\text{-}D}$	0.45	0.52	0.54	0.49	0.44	0.54	0.52	0.53	0.44	0.52	0.49	0.45	0.47	0.49	0.32
$E_{1/D}$	0.52	0.49	0.47	0.42	0.51	0.44	0.54	**0.56**	0.46	0.45	0.47	0.50	**0.58**	0.52	0.26
$E_{-ln(D)}$	0.50	0.54	0.49	0.45	0.49	0.50	0.46	0.52	0.51	0.51	0.49	0.50	0.53	0.51	0.27
Heip	0.53	0.47	0.56	0.50	0.50	0.45	0.46	0.44	0.52	0.52	0.50	0.51	0.50	0.49	0.28
Hill	0.43	0.47	0.43	0.44	0.41	0.54	0.56	0.54	0.44	0.49	0.50	0.43	0.54	0.44	0.28
Hurlbert	0.51	0.49	0.46	0.53	0.49	0.49	0.45	0.53	0.46	0.46	0.51	0.45	0.49	0.50	0.33
Mcintosh	0.54	0.56	0.48	0.52	0.53	0.50	0.53	0.51	0.47	**0.53**	0.49	0.48	0.53	0.49	0.27
Simpson	**0.60**	**0.59**	0.52	**0.60**	**0.65**	0.64	**0.57**	0.46	**0.58**	**0.53**	**0.56**	**0.63**	0.50	**0.60**	0.26
Shannon	0.47	0.53	**0.57**	0.49	0.48	**0.66**	0.52	0.50	0.46	0.46	0.49	0.53	**0.58**	0.52	0.25

TABLE III. THE AUC-ROC VALUES OBTAINED IN THE EVALUATION OF THE DESCRIPTIVE PERFORMANCES OF THE CLASSIFCATION ALGORITHMS BY USING 5 CATEGORIES FOR THE BIODIVERSITY INDICES. THE HIGHEST VALUE PER ALGORITHM IS BOLDED.

Algorithm / Biodiversity Index	BFTree	FT	C4.5	Ripper	LADTree	LMT	Multilayer Perceptron	NaiveBayes	NBTree	NNge	OneR	PART	RBFNetwork	REPTree
Brilouni	0.50	0.84	0.97	0.50	0.81	0.59	0.95	0.63	0.56	**1.00**	0.64	0.96	0.68	0.74
E_{1-D}	0.50	0.84	0.98	0.50	0.83	0.65	0.95	0.68	0.88	**1.00**	0.63	0.98	0.74	0.74
$E_{1/D}$	0.50	0.86	0.88	0.50	0.84	0.67	0.95	0.68	0.60	**1.00**	0.61	0.93	0.77	0.50
$E_{-ln(D)}$	0.59	0.86	**0.99**	0.50	0.81	0.66	0.96	**0.70**	**0.91**	**1.00**	0.62	**0.99**	**0.78**	0.63
Heip	0.94	0.85	**0.99**	0.66	0.83	0.73	0.94	0.68	0.86	**1.00**	0.65	0.97	0.75	**0.77**
Hill	0.75	**0.89**	0.79	0.69	**0.88**	**0.97**	0.96	0.69	0.50	**1.00**	**0.70**	0.74	0.75	0.50
Hurlbert	0.50	0.83	0.95	0.50	0.82	0.62	0.96	0.68	0.74	**1.00**	0.59	0.94	0.74	0.69
Mcintosh	0.97	0.81	0.98	**0.70**	0.81	0.62	0.95	0.66	0.68	**1.00**	0.63	0.81	0.71	0.50
Simpson	0.50	0.70	0.97	0.52	0.81	0.59	**0.97**	0.66	0.64	**1.00**	0.63	0.90	0.72	0.50
Shannon	**0.95**	0.75	0.96	0.68	0.80	0.62	**0.97**	0.67	0.83	**1.00**	0.63	0.98	0.71	0.67

B. Ranking and Statistical Significance

As we noticed in the previous section, the results are showing some interesting patterns for the descriptive performances when the same dataset is used for both training an testing. However, we also try to see if there is some underlying patterns in the results for the predictive performances when the model is tested on unseen data. This is done using the two-step procedure consisting of the Aligned Friedman test [24] as well as the post-hoc Hommel test [25] to determine the models with statistical difference when the number of categories for the class attribute is 3 and 5, as well as a combination of both these datasets. The results from the evaluation are presented in Table 5 and revealed some interesting patterns that are discussed further in this section. According to the analysis for the descriptive performances for the biodiversity datasets with 3 and 5 categories for the class attribute (C3_Train and C5_Train), the models build by the NNge algorithm have best average rank (lowest value) and they are taken as control models. Most of the classification algorithms achive statistically significant results compared to the control model.

That is not a case for the predictive performances, where the classification model with the best average rank is generated with the RBF Network algorithm by using 3 categories for the class attribute (C3_Test) and with the LMT algorithm when using 5 categories (C5_Test). However, the majority of the models obtained with the other algorithms do not have statistically significant results compared to these control models.

The last analysis is a combination of both cases when using 3 and 5 categories for the class attribute. The analysis showed that the NNge algorithm has best descriptive performances (C3C5_Train), while it is hard to say which algorithm performs the best for making predictions on unseen data (C3C5_Test).

TABLE IV. THE AUC-ROC VALUES OBTAINED IN THE EVALUATION OF THE PREDICTIVE PERFORMANCES OF THE CLASSIFCATION ALGORITHMS BY USING 5 CATEGORIES FOR THE BIODIVERSITY INDICES. THE HIGHEST VALUE PER ALGORITHM IS BOLDED. THE HIGHEST RESISTANCE TO OVER-FITTING IS UNDERLINED.

Algorithm / Biodiversity Index	BFTree	FT	C4.5	Ripper	LADTree	LMT	Multilayer Perceptron	NaiveBayes	NBTree	NNge	OneR	PART	RBFNetwork	REPTree	RMSE
Brilouni	0.48	0.56	0.49	0.49	0.50	0.49	0.52	0.48	**0.53**	**0.52**	0.52	0.48	0.48	0.50	0.29
E_{1-D}	0.48	0.53	0.52	0.47	0.51	0.50	0.50	0.47	**0.53**	0.51	0.52	0.47	0.51	0.50	0.32
$E_{1/D}$	0.47	0.54	0.51	0.49	0.52	**0.62**	0.54	0.46	0.46	0.51	**0.56**	0.47	0.47	0.48	<u>0.28</u>
$E_{-ln(D)}$	0.50	0.51	**0.52**	0.51	0.56	0.57	0.54	0.51	0.51	0.51	0.51	0.48	0.55	0.55	0.31
Heip	**0.56**	0.56	**0.52**	0.46	0.55	0.56	0.53	0.51	**0.53**	0.50	0.51	0.49	0.49	0.54	0.33
Hill	0.45	**0.59**	0.48	**0.59**	**0.56**	0.55	0.60	**0.56**	0.51	0.58	0.44	0.50	**0.53**	0.50	<u>0.28</u>
Hurlbert	0.47	0.52	0.48	0.48	0.55	0.53	0.50	0.51	0.50	0.49	0.51	0.50	0.52	**0.57**	0.30
Mcintosh	0.45	0.50	0.50	0.58	0.53	0.53	0.48	0.50	**0.53**	0.48	0.49	0.50	0.52	0.48	0.32
Simpson	0.46	0.48	0.51	0.47	0.51	0.48	0.48	0.50	0.49	**0.52**	0.51	**0.53**	0.47	0.45	<u>0.28</u>
Shannon	0.48	0.50	0.48	0.50	0.51	0.50	0.50	0.51	0.44	0.51	0.49	0.49	0.52	0.49	0.34

TABLE V. THE AVERAGE RANK OF THE CLASSIFICATION ALGORITHMS. THE BEST AVERAGE RANK IS BOLDED. THE MODEL WITH BEST AVERAGE RANK IS TAKEN AS CONTROL MODEL FOR THE SECOND PART OF THE PROCEDURE. THE MODEL VARIANTS THAT ARE REJECTED BY THE HOMMEL TEST ARE UNDERLINED. THESE MODELS HAS STATISTICALLY SIGNIFICANT DIFFERENCES WITH THE CONTROL MODEL.

	BFTree	**FT**	**C4.5**	**Ripper**	**LADTree**	**LMT**	**Multilayer Perceptron**	**NaiveBayes**	**NBTree**	**NNge**	**OneR**	**PART**	**RBF Network**	**REPTree**
C3_Train	12.3	5.20	5.60	11.2	4.00	8.30	2.70	8.50	11.2	**1.10**	9.10	8.45	6.40	10.9
C3_Test	6.45	5.70	8.75	8.30	8.35	5.95	7.25	6.00	9.80	8.20	8.50	9.20	**5.35**	7.20
C5_Train	10.1	5.80	3.10	12.2	6.40	10.3	3.30	10.2	8.45	**1.00**	11.4	3.80	7.95	10.9
C5_Test	11.5	5.15	8.20	8.45	3.95	**5.10**	6.35	8.40	7.70	6.90	7.50	9.90	7.95	7.95
C3C5_Train	11.1	5.5	4.45	11.8	5.14	9.37	2.95	9.35	9.8	**1.05**	10.2	6.10	7.17	10.8
C3C5_Test	9.07	**5.12**	8.45	8.6	6.15	5.47	7.02	7.35	8.75	7.57	7.92	9.55	6.4	7.58

The analysis for the statistical significance revealed that there is a statistical significance in the descriptive performances of the algorithms, which is mainly not a case for the predictive performances.

IV. CONCLUSION

In this paper, we made experimental evaluation of models for determining the relationship between the diatoms' biodiversity of an ecosystem and the abiotic factors in that environment. For that purpose, we used different groups of machine learning algorithms (decision trees, rule induction and neural networks), and we investigated the influence of the number of discretization levels used for the class attribute. The performances of the models were measured using the AUC-ROC evaluation measure. Furthermore, the results of the models are examined for their statistical significance. Based on the results, we can make several conclusions. The best descriptive performance is obtained with the NNge algorithm. Moreover, the models obtained with the other algorithms have statistically significant differences compared to the model with the NNge algorithm. We get a different picture if we look at the predictive performances of the models on unseen data. In these experiments, the model with the best rank cannot be decided, different algorithms build best models in different cases, and many of the remaing algorithms do not have statistically significant differences with the best model in the particular case. Furthermore, the increase in the number of categories of the class attribute does not improve the models' performances, and as the results show, in many cases the performances are decreased.

The results from this paper can give more insides in the research of diatoms' biodiversity modelling since biologists are more willing to use classes instead of numbers to evaluate the ecological status of the lake ecosystem. Further research could be make by using additional biodiversity indices. Also, different classification algorithms for fuzzy modelling can be applied.

ACKNOWLEDGMENT

This work was partially financed by the Faculty of computer science and engineering at the Ss. Cyril and Methodius University in Skopje.

REFERENCES

[1] S. Villéger, N.W.H. Mason, and D. Mouillot, "New multidimensional functional diversity indices for a multifaceted framework in functional ecology," Ecology, vol. 89, pp. 2290–2301, 2008.

[2] F. Recknagel, and W. Michener, "Ecological Informatics. Data management and knowledge discovery," 3rd Edition. Springer International, pp. 1-482, 2018.

[3] J.A. Hendrickson, and P.R. Ehrlich, "An expanded concept of "species diversity"," Not. Nat. Phila. vol. 439, pp. 1-6, 1971.

[4] E.C. Pielou, "Shannon's formula as a measure of specific diversity," Its use and misuse, American Naturalist, vol. 100, pp. 463-465, 1966.

[5] E.C. Pielou, "Species-diversity and pattern diversity in the study of ecological succession," Journal of Theoretical Biology, vol. 10, pp. 370-383, 1966.

[6] C. Tang, Y.Yi, Z.Yang, Y. Zhou, T. Zerizghi, X. Wang, X. Cui, and P. Duan. "Planktonic indicators of trophic states for a shallow lake (Baiyangdian Lake, China)." Limnologica, vol. 78, 2019.

[7] R.H. Zimmerman, "The ecology of mosquito larvae (Diptera: Culicidae) in the subalpine zone of the eastern Sierra Nevada Mountains, California, USA." Journal of Vector Ecology, vol. 44, no. 1, 2019.

[8] N.K.P. Udaya, N.A. Karthick, D. Poomagal, M. Susithra, M. Chandran, and S. Bhuvaneswari. "Fungal endophytes of an aquatic weed Marsilea minuta Linn." Current Research in Environmental & Applied Mycology (Journal of Fungal Biology), vol. 8, no. 1, pp. 86-95, 2018.

[9] T. Yamada, S. Aiba, Y. Kubota, K. Okubo, I. Miyata, E. Suzuki, H. Maenaka, and M. Nagano, "Dynamics of species diversity in a Japanese warm-temperate secondary forest," Ecosphere, vol. 2, 2011.

[10] E. Magurran, "Ecological diversity and its measurement," Princeton University press, Princeton. NJ, 1988.

[11] E.C. Pielou, "Ecological diversity," Wiley Interscience, New York, 1975.

[12] A. E. Magurran, "Measuring biological diversity," Blackwell, Oxford UK, 2004.

[13] M.O. Hill "Diversity and evenness: a unifying notation and its consequences," Ecology, vol. 54, pp. 427-431, 1973.

[14] A.H. Fielding, "Cluster and lassification techniques for the biosciences," Cambridge: Cambridge University Press, 2007.

[15] J. Levatić, D. Kocev, M. Debeljak, and S. Džeroski, "Community structure models are improved by exploiting taxonomic rank with predictive clustering trees," Ecologiccal Modelling, vol. 306, pp. 294-304, 2015.

[16] C. Kampichler, R. Wieland, S. Calmé, H. Weissenberger, and S. Arriaga-Weiss, "Classification in conservation biology: A comparison of five machine-learning methods," Ecological Informatics, vol. 5, no. 6, pp. 441-450, 2010.

[17] C.A. Drew, Y.F. Wiersma, and F. Huettmann, “Predictive species and habitat modelling in landscape ecology,” Springer, 2011.

[18] S. Oppel, A. Meirinho, I. Ramirez, B. Gardner, A.F. O'Connell, P.I. Miller, and K. Louzao, “Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds,” Biological Conservation, vol.156, pp. 94-104, 2012.

[19] A. M. Prasad, L.R. Iverson, and A. Liaw, “Newer classification and regression tree techniques: bagging and random forests for ecological predictions,” Ecosystems, vol. 9, pp.181-199, 2006.

[20] A. Naumoski, and K. Mitreski, “Influence of climate change on diatoms diversity indices in Lake Prespa,” Journal of E-Analytical - Energy and Climate Change - Southeast Europe in Focus, vol. 2, no. 1, 2009.

[21] A. Naumoski, “Multi-target modelling of diatoms diversity indeces in Lake Prespa,” Applied Ecology and Environmental Research, vol. 10, no. 4, pp. 521-529, 2012.

[22] A. Naumoski, “Learning models of abiotic influence on the biodiversity indices in Lake Prespa,” Proceedings of the 12th International Conference for Informatics and Information Technologies (CIIT 2015), pp. 257-261, Bitola, Macedonia, 2015.

[23] S. García, A. Fernández, J. Luengo, and F. Herrera, “Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power,” Information Sciences, vol. 180, pp. 2044-2064, 2010.

[24] M. Friedman, “The use of ranks to avoid the assumption of normality implicit in the analysis of variance,” Journal of the American Statistical Association, vol. 32, pp. 674–701, 1937.

[25] G. Hommel, "A stagewise rejective multiple test procedure based on a modified Bonferroni test". Biometrika, vol. 75, pp. 383–386, 1988.

[26] S. Krstić, “Description of sampling sites,” TRABOREMA Project, FP6, 2006.

[27] L. Brillouin,” Science and information theory,” Academic Press. New York. pp. 320, 1956.

[28] B. Smith, and J.B. Wilson, “A consumer’s guide to evenness indices,” Oikos, vol. 76, pp. 70–82, 1996.

[29] C. Heip, “A new index measuring evenness,” J. Mar. Biol. Assoc. vol. 54, pp. 555–557, 1974.

[30] R.V. Alatalo, “Problems in the measurement of evenness in ecology, “ Oikos, vol. 37, pp. 199–204, 1981.

[31] S. H. Hurlbert, “The nonconcept of species diversity: a critique and alternative parameters,“ Ecology, vol. 52, pp. 577–586, 1971.

[32] R.P. Mcintosh, ”An index of diversity and the relation of certain concepts to diversity,“ Ecology, vol. 48, pp. 392–404, 1967.

[33] E. H. Simpson, “Measurement of diversity, “ Nature, vol. 163, pp. 688, 1949.

[34] C.E. Shannon, and W. Weaver, “The mathematical theory of communication. University Illinois Press, Urbana, 1963.

[35] R.C. Holte, “Very simple classification rules perform well on most commonly used datasets,” Machine Learning, vol. 11, pp. 63-91, 1993.

[36] M. Brent, “Instance-Based learning: Nearest Neighbor with generalization,” Hamilton, New Zealand, 1995.

[37] R. Sylvain,”Nearest neighbor with generalization,” Christchurch, New Zealand, 2002.

[38] F. Eibe, and H.I. Witten, “Generating accurate rule sets without global optimization,” In: Fifteenth International Conference on Machine Learning, pp. 144-151, 1998.

[39] W.W. Cohen, “Machine learning,” In: proceedings of the Twelfth International Conference on Machine Learning (ICML 1995), pp. 115–123, 1995.

[40] J.H. George, and P. Langley, “Estimating continuous distributions in bayesian classifiers,” In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, pp. 338-345, 1995.

[41] D.S. Broomhead, and D. Lowe, "Multivariable functional interpolation and adaptive networks," Complex Systems, vol. 2, pp. 321–355, 1988.

[42] J.L. McClelland, “Parallel distributed processing”, MIT Press, 1986.

[43] H. Shi, “Best-first decision tree learning,” Hamilton, NZ, 2007.

[44] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: A statistical view of boosting,” Annals of statistics, vol.28, no. 2, pp. 337-407, 2000.

[45] J. Gama, “Functional trees,” Machine Learning, vol. 55, no. 3, pp. 219-250, 2004.

[46] J.R. Quinlan,”C4.5: Programs for Machine Learning,” Morgan Kaufmann, Sanateo, CA, 1993.

[47] G. Holmes, B. Pfahringer, R. Kirkby, E. Frank, and M. Hall, “Multiclass alternating decision trees,” European Conference in Machine Learning, pp. 161-172, 2001.

[48] N. Landwehr, M. Hall, and F. Eibe, “Logistic Model Trees,” Machine Learning, vol. 95, no. 1-2, pp. 161-205, 2005.

[49] R. Kohavi,”Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid,” In: Second International Conference on Knowledge Discovery and Data Mining, pp. 202-207, 1996.

[50] J.R. Quinlan, “Simplifying decision trees,” International Journal of Man-Mahine Studies, vol. 27, no. 3, pp. 221-248, 1987.

Classifying Protein Structures by Using Protein Ray Based Descriptor, KNN and FuzzyKNN Classification Methods

G. Mirceva, A. Naumoski and A. Kulakov
Faculty of computer science and engineering, Ss. Cyril and Methodius University in Skopje, Skopje, Macedonia
georgina.mirceva@finki.ukim.mk

***Abstract* - Bioinformatics community continuously studies the protein molecules and the processes in which they are involved. To understanding these processes, the researches need to understand the characteristics and functions of proteins. Therefore, classification of protein structures is one of the most studied problems in proteomics. Although many methods are developed for this purpose, still there is a need for fast and accurate methods that would provide classification of protein structures. In this paper, our aim is to develop a method that would provide accurate classification of proteins. First, we extract feature vectors with the geometrical characteristics of the protein structures. For that purpose we use our protein ray based descriptor, which represents how the amino acid residues are placed with respect to the center of mass. After feature extraction, next we apply two instance-based learning methods for making predictions. Besides the KNN classifier, which is based on classical set theory, we also use the FuzzyKNN classification method that is based of fuzzy set theory. For evaluation, we use a part of the SCOP database. The paper presents the experimental results achieved by using the two classification methods by considering various number of nearest neighbors.**

Keywords - protein structure; protein classification; protein ray based descriptor; KNN; FuzzyKNN

I. INTRODUCTION

Protein molecules are very important in human bodies because they take part in various processes that happen in the cells of the organisms. Proteins have many different functions including transport of oxygen, enzyme in biochemical reactions, defensive enzymes, antibodies that bind to foreign molecules in order to neutralize them. They also may be contractile proteins that afford movements in the muscles. Proteins also may be signaling proteins, like insulin that signals to enable the uptake of glucose from the blood that is provided by proteins with glucose transport function. Due to the importance of understanding the protein molecules, there is a huge interest from many research groups to analyze the protein molecules in order to understand their functions in the human body.

As the technology advances, different techniques are introduced that afford determination of the structure of proteins. In order to share the data about the protein structures, the Protein Data Bank (PDB) [1], [2] is formed as a main repository for proteins. Fig. 1 shows how the number of structures stored in PDB rises through the years. However, the data about the structure of the proteins are not very useful if we are not able to understand what they do in the human body. Therefore, it is very important to determine the functions of the proteins.

There are many methods that aim to classify the protein molecules, and they base on the premise that the proteins that are part of the same class share the same functions. Although a great effort is made for providing fast classification of proteins, the methods are not able to follow the speed of discovering new protein molecules whose structure is determined but their functions are not discovered yet. Therefore, the proteomics research community needs fast and accurate computational methods for classification of protein structures.

Among the most important methods for protein classification are: SCOP (Structural Classification Of Proteins) [3] and CATH (Class, Architecture, Topology and Homologous superfamily) [4]. With the SCOP method, the classification of proteins is done manually by inspecting the proteins visually. Due to the manual inspection of the proteins, the SCOP method takes long time to classify an inspected protein structure. Therefore, in the literature there are also automatic and semi-automatic methods for classification of proteins. For example, the CATH method is a semiautomatic method where proteins are classified in an automatic manner, and if a given protein is not able to be classified automatically, then it is classified in manual way.

There are methods where the classification of proteins is made by aligning protein sequences, like Needleman–

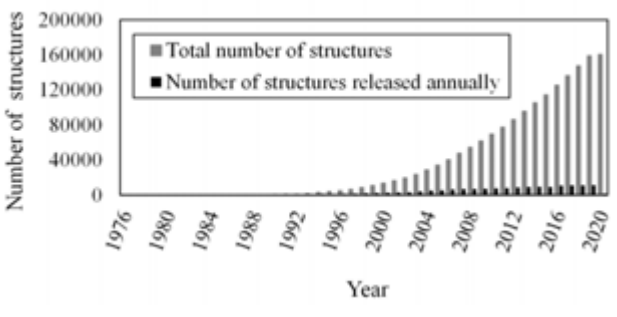


Figure 1. Number of structures released in PDB.

Wunch [5], BLAST [6] and PSI-BLAST [7]. These methods are able to recognize proteins with similar sequences, but they may dismiss proteins whose structures are similar while hawing sequences that are not very similar. Due to this, similarity between proteins is better discovered by making alignment of their structures instead of their sequences. For example, CE [8], MAMMOTH [9] and DALI [10] make alignment of protein structures. Also, there are methods that combine sequence and structure alignment (SCOPmap [11], FastSCOP [12]).

The alignment of protein sequences or structures is a time-consuming task, thus these methods take too long. Thus, many methods are introduced where the proteins are represented with their feature vectors, and later the comparison is made by finding the similarity between the feature vectors of the inspected proteins. For that purpose, first, the features of the sequences [13] or structures [14] of the proteins should be extracted, and then prediction model could be generated for classifying proteins by applying some classification algorithm from machine learning. In this way, instead of using all data about the proteins, only the main information about the proteins is preserved in the feature vectors, thus reducing the amount of data that are processed further. The classification stage is performed over the reduced data containing the extracted feature vectors, so in that way the training of the prediction model is faster as well as the classification of a given protein that is examined.

In this study, we put our focus on this class of methods where the classification of the proteins is made by using the feature vectors as their representatives. Previously, we have done many studies for extracting various features of the proteins, most of which focus on their geometrical features of the tertiary structure of the proteins. In our research work presented in [15], we made an intensive analysis of various approaches that we have proposed that could be used to extract feature vectors for the proteins. In [15], we mainly focused on developing feature vectors that could be used to find the proteins that are similar with an inspected protein, which is actually solving a protein retrieval task. Besides protein retrieval, we may continue the research in a direction of applying the proteins' feature vectors for solving a task of protein classification, which is the focus of this study.

In this paper, we develop a method that could be used to classify protein structures. First, we extract the geometrical features of the proteins by using the protein ray based descriptor [15]. With the analysis made in [15], we proved that although the protein ray based descriptor is very simple, it provides an accurate retrieval of similar protein structures. This feature vector holds features that show how the protein backbone is positioned with respect to the center of mass. In this work, we use this feature vector for solving a protein classification task. In order to do that, we use two instance-based learning methods. Besides the well-known KNN (k nearest neighbors) method [16], we also use the FuzzyKNN classification method [17]. Both methods make decision based on the class labels of the nearest neighbors of the examined sample. However, the classical KNN method relies on the classical set theory, while FuzzyKNN is its adaptation that is applicable for fuzzy sets.

This research paper is organized in the following manner. Section 2 gives detailed description of the proposed method for protein classification. In this section, the protein ray based descriptor, as well as the KNN and FuzzyKNN classification methods are described. Section 3 shows the experimental results that are obtained. In the analysis, there are experiments by using these two classifiers, and also, we make experiments by considering different number of nearest neighbors in the decision process. Finally, the main conclusions and some possible future enhancements are discussed in Section 4.

II. THE PROPOSED METHOD FOR PROTEIN CLASSIFICATION

In this study, we propose a method that provides classification of protein structures based on their geometrical features. First, we extract the features of the protein structures, and then by using corresponding classification algorithm, we classify the unknown protein into a corresponding class. Regarding the first step, we extract the features from the protein ray based descriptor [15]. In the second step, we use two classification methods for instance-based learning, where the decision about the class of the inspected sample is made based on the classes of its nearest neighbors. For that purpose, besides the KNN (k nearest neighbors) classifier [16] that is based on the classical set theory, we also use the FuzzyKNN classifier [17] that is based on the theory of fuzzy sets. In the following subsections, the protein ray based descriptor is described, as well as the KNN and FuzzyKNN classifiers.

On Fig. 2, we give illustration how the training and testing is made. In order to classify protein structures, first we need the data about the structures of the proteins that are stored in their PDF files, which are taken from the PDB database [2]. The PDB files hold information about the primary, secondary and tertiary structure of proteins. During training, based on the data from the PDB files of

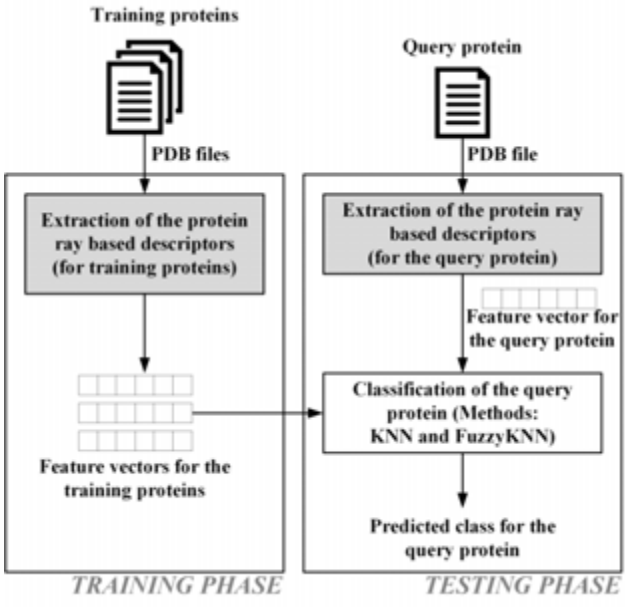


Figure 2. Ilustration of the proposed mehod for protein classification.

the training proteins, we extract their protein ray based descriptors, which are stored as their representatives. During testing, first the protein ray based descriptor for the examined query protein is extracted, and then the query protein is classified by using the KNN or FuzzyKNN classifier. For that purpose, we use knowledge from the SCOP database, where the proteins are classified in hierarchical manner. The most important level in the SCOP hierarchy is the domain level, so we directly classify the proteins on this level. For each training protein we take the information from the SCOP database in which SCOP domain the protein belongs to. The predicted class of the query protein indicates in which SCOP domain the inspected query protein is expected to belong to.

A. Protein Ray Based Descripor

Before giving description how the protein ray based descriptor is extracted, first we will describe which are the counterparts of the proteins molecules. The protein molecule is formed from one or several chains. The SCOP database [3] contains knowledge about the classification of such protein chains, so each sample in the dataset that we use corresponds to a given protein chain. The protein chains are defined by the amino acid residues sequence, which correspond to its protein sequence. These sequence folds in some particular way in the space, thus defining the secondary and tertiary structure of the protein chain. The amino acid residues contain several atoms that are connected in a particular way based on that what are the amino acids that they correspond to. The most important atoms of the amino acid residues are the C_α atoms. The residues are connected through these C_α atoms, and they form the skeleton of the protein chain, or also known as protein backbone. In this study, we consider only the information about the positions (3D coordinates) of the C_α atoms of the protein chains, thus we know the exact way how the protein skeleton is folded in the space. In our previous studies [15], in some of the approaches for protein retrieval besides the C_α atoms we also considered the remaining atoms of the amino acid residues, but it resulted with lower accuracy in the retrieval.

As it was mentioned above, in this research we consider only the C_α atoms of the protein chain, and we take into account the information how these atoms are positioned in the space, thus forming the 3D model. In order to provide scale invariance, we scale the model so the Euclidean distance between the center of mass and the most distant C_α atom equals 1. Besides invariance to scaling, we also should provide invariance to translation and rotation, but we will deal with that later in the step of extracting the final features in the feature vector. Another very important issue that we need to deal with is to provide descriptors with fixed length. Namely, different protein chains have different number of C_α atoms, so we have to find a unique way how to represent the skeletons of the protein chains with equal number of features although they may have different number of C_α atoms. We solve this by making interpolation of the skeleton of the protein chain.

In [15], we presented two approaches for interpolation of the protein skeleton, i.e. uniform and non-uniform interpolation based on that whether the representative points of the skeleton are equidistant along the skeleton or not. The analysis made in [15] proved that the uniform interpolation of the protein skeleton provides better protein retrieval. Therefore, in this study we interpolate the skeleton of the protein chain uniformly.

The interpolation of the skeleton is done in the following way. First, we calculate the length of the skeleton of the protein chain as

$$L = \sum_{i=1}^{N_\alpha - 1} d_{Euclidean}(i, i+1), \tag{1}$$

where N_α is the number of C_α atoms in the protein chain and $d_{Euclidean}(i,i+1)$ corresponds to the Euclidean distance between the i-th and (i+1)-th C_α atoms. Before extracting the feature vectors, first we need to decide how many interpolation points we want to use. The number of interpolation points is denoted with N. The interpolation of the skeleton is made by determining the 3D coordinates of N interpolation points, which are equidistant along the skeleton on distance $L/(N-1)$. In [15], we also analyzed the influence of the length of the feature vectors over the accuracy of the retrieval, and the results showed that by using 64 interpolation points we obtain comparable accuracy with the cases when the number of interpolation points is higher. According to the results presented in [15], it is optimal to interpolate the protein skeletons with 64 interpolation points, thus in this study we use N=64 interpolation points.

After interpolation of the skeleton, next the features of the feature vector are extracted. For that purpose, we use the same idea that is used in the extraction of the ray descriptor [18], which is proposed for retrieving similar 3D objects. This is done by emitting rays from the center of mass to the interpolation points. The i-th feature of the feature vector is calculated as Euclidean distance between the center of mass and the i-th interpolation point. In this way we provide invariance to translation and rotation of the feature vector. The extracted feature vector shows how the protein skeleton is positioned compared to the center of mass.

B. KNN and FuzzyKNN Classification Methods

In this study, for making decisions about the classes for the unknown proteins we use two classification methods that perform instance-based learning. These methods do not build a particular model, because during testing the query sample is compared with all training samples and the decision is made by voting. In this research we use the KNN (k nearest neighbors) method [16] that is applicable for classical sets, and the FuzzyKNN [17] that relies on the theory of fuzzy sets.

Since KNN is a well-known classification method, we will give just a short description of this method. The query sample, the query protein chain in our case, is compared with all training samples. The distance between a given training sample x and the query q is denoted as $D(x,q)$. In this study we use the Euclidean distance as a distance measure. Next, the k nearest neighbors are identified by considering the training samples with lowest distance to

the query. In the description of the methods, k stands for the number of nearest neighbors that are considered. After finding the k nearest neighbors, then the final decision is made by voting. One possible approach is to make majority voting, where each neighbor has equal vote in the voting. However, by making majority voting, the majority class could be chosen based on the votes of more distant neighbors. To overcome this, it is better to perform voting where the weight of the vote of a given neighbor depends on the similarity between that neighbor and the query. The distance weighted voting could be made by giving the following voting weights to the nearest neighbors $S(x,q)=1-D(x,q)$ or $S(x,q)=1/D(x,q)^n$. By using distance weighted voting, the closer neighbors have higher weights in the voting. For the KNN method, we use the following weights in the voting: 1 (majority voting), $1-D(x,q)$ (marked as 1-D) and $1/D(x,q)$ (marked as 1/D). Besides the voting scheme that is used, it is also very important to define how many nearest neighbors would be considered, because it has a great influence on the decisions.

Besides the KNN classification method, in this paper we also use FuzzyKNN [17], which is inspired by KNN. In FuzzyKNN, similarly as in KNN, for a given query sample q, first the k nearest neighbors are identified, and they form the set of nearest neighbors NN. The decision about the class for the query sample q is made by maximizing

$$\frac{\sum_{x\in NN} S(x,q)M_c(x)}{\sum_{x\in NN} S(x,q)}, \tag{2}$$

where $M_c(x)$ is the membership function for a given class c. $S(x,q)$ is the similarity between a given nearest neighbor x and the query q and it is calculated as $S(x,q)=1/D(x,q)^{2/(m-1)}$, where $D(x,q)$ is the distance between x and q. The parameter m defines the weights of the votes in the voting. In this paper m is set to 2, so $S(x,q) = 1/D(x,q)^2$.

The membership function $M_c(x)$ for a given class c could be crisp or gradual. The crisp function is defined as

$$M_c(x)=\begin{cases}1, & x\in C\\ 0, & x\notin C\end{cases}, \tag{3}$$

where C denotes the set of samples from the inspected class c. However, because we use a method that relies of fuzzy set theory, in this study we use a gradual function [17] that is defined as

$$M_c(x)=\begin{cases}0.51+0.49\frac{n_C}{k}, & x\in C\\ 0.49\frac{n_C}{k}, & x\notin C\end{cases}, \tag{4}$$

where $n_c=|C|$ is the number of samples that belong to the class c.

III. Experimental Results

In this study, we used the knowledge about the protein structures stored in PDB [2] in order to obtain information about their tertiary structure. We also used knowledge from the SCOP database [3] about the classification of the protein chains according to the SCOP method. The dataset that is used in this study contains 6145 samples from 150 classes. The samples correspond to the protein chains, while the classes correspond to the SCOP domains from the SCOP hierarchy. The samples are almost uniformly distributed in the classes. The training set has 5531 samples, while the test set has 614 samples. On this link we provide details regarding the dataset that is used. As evaluation measure we use the classification accuracy, which is appropriate when the dataset is balanced as in our case. In the analysis, we made tests by considering different number of nearest neighbors (k = 1, 2, 3, 4, 5, 10, 15, 20, 30, 40, 50). On Fig. 3, the experimental results that are obtained with the KNN and FuzzyKNN classification methods are presented. Regarding KNN, we made experiments by using different voting schemes: 1 (majority voting), 1-D, 1/D. From the results presented on Fig. 3, it can be seen that as the number of nearest neighbors increases, the classification accuracy declines with both methods. Regarding the three voting schemes for KNN that were examined, it is evident that by using majority voting lowest results are obtained, then the voting scheme 1-D follows, while the voting scheme 1/D showed best performance. Regarding these three voting schemes, it is also interesting to note that by using the first two voting schemes (majority voting and 1-D), the classification accuracy starts to declines much earlier and it declines much faster as we increase the number of nearest neighbors that are considered. On the other hand the performances declines slower with FuzzyKNN and KNN(1/D). This suggests that with FuzzyKNN and KNN(1/D), the number of neighbors that are considered has lower influence on the results than in the first two cases. For k<=5, the results with both methods are comparable. As k goes up to 20, there is slight decrease for KNN(majority voting) and KNN(1-D). Much significant decrease for these two schemas is remarkable for k>=30, which is due to the fact the classes are almost that size, so by using larger value for k, among the nearest neighbors there will be more samples from the other classes.

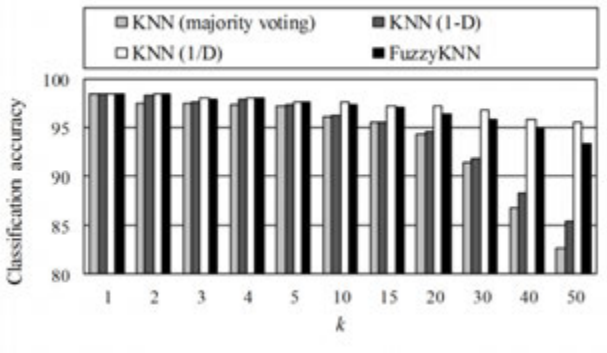


Figure 3. Experimental results obtained with the KNN and FuzzyKNN classification methods. For KNN, we use the following voting schemes: majority voting, 1-D, 1/D.

In our previous study [15], we made an extensive evaluation where several protein descriptors, including the protein ray based descriptor, were compared with several existing approaches. The results from that study showed that the protein ray based descriptor is very simple, but provides an accurate retrieval of protein structures, and it is comparable with the existing approaches.

IV. Conclusion

In this paper, we proposed a method that provides classification of protein structures based on their geometrical features. First, we extracted feature vectors for the protein chains in the training set. For that purpose, we used our protein ray based descriptor, which showed as very simple and also accurate in our previous studies for protein retrieval. During testing, first the feature vector for the inspected protein chain was generated. Then, we applied two instance-based learning algorithms that could be used for solving classification task, i.e. KNN and FuzzyKNN.

The evaluation of the proposed method was made by using knowledge from the PDB database about the tertiary structure of proteins. The class labels for the samples in the dataset were taken from the SCOP database based on that in which SCOP domain a given protein chain belongs to. For the KNN method, we made experiments by using three different voting schemes by giving the following weights to the nearest neighbors: 1 (majority voting), 1-D and 1/D. The results showed that KNN(majority voting) and KNN(1-D) obtain lower results, and it is evident that with the rise of the number of nearest neighbors, there is significant decrease with these two schemes. On the other hand, with KNN(1/D) and FuzzyKNN, the classification accuracy declines much slower as the number of nearest neighbors increases. This means that KNN(1/D) and FuzzyKNN are less sensitive to the choice of the optimal number of nearest neighbors.

This work could be further continued in several possible directions. Instead of the protein ray based descriptor, also some other feature vectors could be applied. In the choice of possible feature vectors, we may consider feature vectors that contain not only geometrical features, but we may also consider some biological features of the proteins. Regarding KNN and FuzzyKNN, besides the Euclidean distance, we may also try some other distance measures. For FuzzyKNN, besides the membership function that is used, also some other membership functions could be considered, which may be more suitable for these data. Regarding the choice of classification method, we plan to extend the study by applying other classifiers that are based on classical sets, as well as classifiers that are adapted for fuzzy sets.

Acknowledgment

This work was partially financed by the Faculty of computer science and engineering at the "Ss. Cyril and Methodius University in Skopje", Skopje, Macedonia.

References

[1] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," Nucleic Acids Res., vol. 28, no. 1, pp. 235–242, January 2000.

[2] RCSB Protein Data Bank, http://www.rcsb.org, 2019.

[3] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "Scop: a structural classification of proteins database for the investigation of sequences and structures," J. Mol. Biol., vol. 247, no. 4, pp. 536–540, April 1995.

[4] C. A. Orengo, A. D. Michie, D. T. Jones, M. B. Swindells, and J. M. Thornton, "CATH – a hierarchic classification of protein domain structures," Structure, vol. 5, no. 8, pp. 1093–1108, August 1997.

[5] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," J. Mol. Biol., vol. 48, no. 3, pp. 443–453, March 1970.

[6] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," J. Mol. Biol., vol. 215, no. 3, pp. 403–410, October 1990.

[7] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Res., vol. 25, no. 17, pp. 3389–3402, September 1997.

[8] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path," Protein Eng., vol. 11, no. 9, pp. 739–747, September 1998.

[9] A. R. Ortiz, C. E. Strauss, and O. Olmea, "Mammoth: an automated method for model comparison," Protein Sci., vol. 11, no. 11, pp. 2606–2621, November 2002.

[10] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," J. Mol. Biol., vol. 233, no. 1, pp. 123–138, September 1993.

[11] S. Cheek, Y. Qi, S. S. Krishna, L. N. Kinch, and N. V. Grishin, "SCOPmap: automated assignment of protein structures to evolutionary superfamilies," BMC Bioinformatics, vol. 5, pp. 197–221, December 2004.

[12] C. H. Tung and J. M. Yang, "fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies," Nucleic Acids Res., vol. 35, W438–W443, July 2007.

[13] K. Marsolo, S. Parthasarathy, and C. Ding, "A multi-level approach to SCOP fold recognition," IEEE Symposium on Bioinformatics and Bioeng., pp. 57–64, October 2005.

[14] P. H. Chi, Efficient protein tertiary structure retrievals and classifications using content based comparison algorithms, PhD thesis, University of Missouri-Columbia, 2007.

[15] G. Mirceva, I. Cingovska, Z. Dimov, and D. Davcev, "Efficient approaches for retrieving protein tertiary structures," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 9, no. 4, pp. 1166–1179, July/August 2012.

[16] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," Mach. Learn., vol. 6, no. 1, pp. 37–66, January 1991.

[17] J. M. Keller, M. R. Gray, and J. R. Givens, "A fuzzy k-nearest neighbor algorithm," IEEE Transactions on Systems, Man, and Cybernetics, vol. 15, no. 4, pp. 580–585, July/August 1985.

[18] D. V. Vranic, 3D Model Retrieval, Ph.D. Thesis, University of Leipzig, 2004.

An Analysis of Early Use of Deep Learning Terms in Natural Language Processing

B. Dalbelo Bašić*, M.P. di Buono **
* Faculty of Electrical Engineering and Computing – TakeLab – University of Zagreb, Zagreb, Croatia
** UNIOR NLP Research Group – University of Naples "L'Orientale", Naples, Italy
bojana.dalbelo@fer.hr
mpdibuono@unio.it

***Abstract* - In this paper, we present the preliminary results on the analysis of deep learning terms used for natural language processing (NLP) tasks. We propose a statistical analysis of papers published from 2012 to 2015 in the main ACL conferences. Our aim is investigating which DL term, and consequently which DL method, is mostly used for each specific NLP task, since its introduction in the field. In order to do this, our first contribution is the development of two terminological lists, referring respectively to DL methods for text analysis and NLP tasks. The list of deep learning terms contains 41 terms and acronyms, as well as a NLP term list contains 145 terms and acronyms. From our corpus, the frequencies of various terms have been extracted with respect to the ACL conference and the publication year. After the preliminary data analysis, we decided to restrict the extraction process to abstract texts. We applied multivariate techniques called correspondence analysis in order to visualize and evaluate the joint behavior of our variables.**

Keywords* - *Correspondence Analysis; Deep Learning; NLP

I. INTRODUCTION

The use of deep learning (DL) architectures and algorithms has been initially introduced mainly for computer vision and pattern recognition by [1]. Due to the massive advances in these fields since their introduction, new DL methods have been become focus for most recent natural language processing (NLP) researches. As highlighted by [2], for decades, shallow models, e.g., SVM and logistic regression, trained on high dimensional and sparse features, have been the basis of machine learning approaches for most of NLP tasks. In the last few years, this trend changed and these tasks have been dealt by means of neural networks based on dense vector representations, suitable to improve NLP results. The mas-sive use of such models for addressing NLP problems is justified by the success of word embeddings [3-4] and DL methods [5]. In 2011, [6] showed that a simple DL framework outperforms most state-of-the-art approaches in several NLP tasks such as named-entity recognition (NER), semantic role labeling (SRL), and part-of-speech (POS) tagging. From this, numerous complex deep learning based algorithms have been proposed to solve difficult NLP tasks. In this paper, we propose a preliminary analysis on the use of DL terms in NLP tasks, investigating their evolution, and consequently the evolution of DL methods used, since the introduction in this field. The general aim is providing a more comprehensive idea of DL trends in NLP and language-related tasks. The structure of the paper is as follows: in Section II, we introduce some related work; Section III describes the terminological resources and the corpus we used in our analysis and the extraction process; in Section IV the analysis and its results are presented and in Section V preliminary conclusions are discussed.

II. RELATED WORK

In this section we introduce some of recent works focusing on analyzing the use of DL terms and methods in NLP tasks. The massive use of these methods is nowadays so spread to represent a phenomenon drawing the attention of several and different researches. For instance, [2] propose a review of significant DL related models employed in NLP tasks, providing information on their evolution together with their comparison in different tasks. Within a monographic overview on general DL methods and its applications, [7] provide a variety of examples in different application domains, related to signal and information processing tasks, including language modeling and NLP. Starting from the undeniable success of DL in the major NLP conferences since 2015, other authors [8] propose an analysis on how DL can help computational linguistics, considering specific phenomena in language (e.g., non-categorical cases in POS tagging for some words).

Identify applicable sponsor/s here. If no sponsors, delete this text box.

TABLE II. A SAMPLE OF NLP AND DL TERMINOLOGICAL LISTS

NLP Terms	DL Terms
Discourse analysis segmentation	Sequence-to-sequence model
Distributional similarity	Recurrent neural network (RNN)
Entity disambiguation	Generative adversarial network
Entity Linking	Convolutional neural network (CNN)
Error correction	Long short-term memory (LSTM)
Event extraction	Boltzmann machine
Frame identification	Autoencoder
Genre classification	Deep Belief Network

III. LINGUISTIC RESOURCES AND CORPUSEXTRACTION

Before proceeding with the extraction process we create two terminological lists, referring respectively to DL terms, indicating methods applied for text analysis and NLP tasks, and terms indicating NLP tasks (from now on we short them into DL methods and NLP tasks). Such resources have been created by means of a manual annotation by two experts. The list of deep learning terms contains 41 terms and acronyms, as well as the NLP term list contains 145 terms and acronyms (Table 1).

All About NLP (AAN) Corpus[1], created by [9], is a database containing 24,627 papers presented at NLP conferences and published between 1965 and 2015. A total of 12449 valid papers belong to these conference codes, along with an additional 619 discarded due to missing metadata. Statistics of the number of publications per year can be seen in Fig. 1. From this corpus a subset corresponding to premier ACL conferences has been extracted. The number of papers for each conference venue belonging to the aforementioned conferences in Table 2. Before the extraction process, we proceed with the normalization of our corpus, then the frequency of a term within an abstract has been counted by string matching each of the terms in our terminological lists with the text of each paper. After the preliminary analysis of our data, we decided to restrict the extraction process to abstract texts. This choice has been motivated by the need of reducing the noise existing in the whole text. Thus, we assume that the method and the general task of the pa-per are well-represented in the abstract of each paper, and these data can be used for our analysis. After this first step, for both DL and NLP extraction results we keep the data spanning from 2012 to 2015 to conduct the analysis. The rationale of creating such a subset, which does not include minor conferences and it is limited to that timespan, has two main motivations: (i) before 2012, papers about DL-based NLP tasks have been published mainly in journals (e.g., [6]) and in other venues rather than NLP conferences (e.g., [3]); and (ii) minor venues, such as student and satellite work-shops, are not characterized by the presence of such innovative approaches, at least at the beginning of their spread. The final step in the development of our data is combining the information provided by these two different data-frames, selecting the papers that present at least a term from the DL list together a term from the NLP list. During this process, we also merge together some terms from both terminological lists, in order to reduce the data according to the presence of upper class and more generic terms.

TABLE I. NUMER OF PUBLICATIONS CONSIDERED FOR EACH CONFERENCE

Conference	#Publications
COLING	3653
ACL	3324
EMLNP	1121
NAACL	1009
EACL	983
CL	763
HLT	462
ANLP	334
StarSEM	239
IJCNLP	231
CoNLL	131
TACL	61
AJCL	48
SIGDIAL	20
Proceedings of the ACL 2010 Student Research Workshop	19
HLT-NAACL Companion Volume: Student Research Workshop	17
ACL-IJCNLP: Student Research Workshop papers	12
Proceedings of the NAACL HLT 2010 Student Research Workshop	11
Proceedings of the Student Research Workshop at EACL 2009	11

IV. DATA ANALYSIS

In this section we present the methodology used to analyze the data and the interpretation of the obtained results. The result of the preprocessing phase were several incidence matrices showing the relation between pairs of categorical variables (e.g. DL terms and years or DL terms and NLP terms or DL terms and ACL conferences). Thus, the natural way of analyzing the association between pairs of categorical variables is the use of the correspondence analysis technique. So, we start with presenting correspondence analysis, as exploratory technique, and then we focus on the preliminary results of this analysis.

[1] http://aan.how/

A. Correspondence analysis

Correspondence analysis (CA) is an exploratory, descriptive, multivariate, statistical technique very similar to a much more popular and much more used principal component analysis (PCA). While PCA is suitable for dimensionality reduction and structure analysis of a number of numerical variables, P CA can reveal the relation (correspondence) between two categorical variables and their discrete values. The starting point for CA is the incidence matrix used to represent the structure or so called two-way contingency tables. For example, the incidence matrix of two categorical variables, DL terms and NLP terms. Each element of such a matrix represents the number showing how many abstracts a particular DL term appeared together with a particular NLP term. The measure of correspondence be-tween the row and column values is given as frequencies of items for a cross-classification of two categorical variables [10]. The first step towards revealing the association between variables and the structure in the incidence matrix is to transform it to a matrix with the measure of distance that is more related to the association be-tween two variables. This measure of distance is the chi-squared distance that is based on the traditional test for association between two categorical variables. CA simplifies the complex data and provides a detailed visual display of data, where each row and each column (i.e., each value of both categorical variables) is simultaneously represented as a single point in some low dimensional space (regularly two dimensional), thus revealing the structural relationship between these two categorical variables. CA has developed its own terminology. The fundamental concepts of CA are the vector of relative frequencies, called row and column profile, and variance called inertia. The objective of correspondence analysis is to reduce the dimensionality of the initial table by decomposing the total variance (inertia) in an n-dimensional space, where n is the minimum of (no. of rows, no. of columns) 1, and thus discovering the subspaces S where the clouds of projected row (respectively column) profiles have the largest inertia. This can be interpreted as finding the low-dimensional subspace S which is as close as possible (measured by chi-squared distance) to the set of profile points in the original high-dimensional space. Like in other multivariate statistical techniques, such as factor analysis, latent semantic analysis or principal component, in correspondence analysis this is done using a singular value decomposition. The use of CA to analyze linguistic phenomena was firstly proposed by [11] and, since then, is often used for visualizing and analyzing the structure of huge corpora, such as in [12].

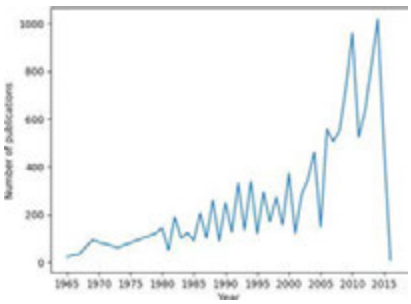


Figure 1. Number of publications per year

B. Preliminary Results

We applied CA to our subset of ANN corpus, obtaining some useful information about the first period in which DL have been applied to NLP tasks. It is worth stressing that the subset we selected for the analysis can be representative only of the earliest introduction of these methods. In fact, according to Manning (2015), 2015 can be considered the year when DL emerges powerfully, as demonstrated by the over-focusing on attempts of beating the state of art in terms of result performance at major NLP conferences. In fact, the first CA plot, which depicts the spread of DL terms over the time span 2012-2015, shows that in2015, the only year value on the positive side, the use of DL terms considerably changes with reference to the previous three years. The different status of 2015 is also proved by the fact that the first dimension (x-axis), the ones in which this year value situates, explains 72% of variability over the years. 2012 and 2014 values are similar in their being ont he same axis, but the DL terms dominating these years differ, namely recursive neural network and recurrent neural network, respectively. Finally, 2015 is dominated by CNN. From these results it seems that there are few reasons to analyze data before 2015. Nevertheless, the analysis of the data available before 2015 may provide an interesting description of pioneer applications of DL methods, as shown in Fig. 2. The second CA plot (Figure 3) presents the correspondence between DL and NLP terms in our corpus, describing several subspaces. The closest point to the origin is represented by deep neural network, which presents the highest frequency in our data. The considerable use of such a term in the earliest publications on DL and NLP is justified by the presence of experiments and applications with less specialized DL methods. Among the subsets that we can identify, the first one is formed by CNN (Convolutional Neural Network), which attracts three NLP tasks, namely text classification, question answering and entity identity. A second cloud of points, situated on the positive side of the second axis, holds three DL methods, i.e., RNN (recur-rent neural network), LSTM, recursive neural network. Ma-chine translation and speech recognition are typical tasks for recurrent neural network, while sentiment analysis is associated with methods like autoencoder and deep belief net-works. Another subgroup in this area contains LSTM and recursive neural network that are used for language modeling and relation classification. The last group is formed by NLP tasks with low frequency and not connected very strongly with some of the DL methods, on one hand these are paraphrase identification, STS (semantic text similarity), semantic parsing, event extraction, and, on the other hand, pos tagging, which is connected to feedforward neural network. These elements are far from the origin due to their frequency and belong to the same point because they present the same profile.

V. CONCLUSION

In this paper, we present our preliminary analysis on the pioneer use of DL terms to describe methods used in

NLP tasks. We analyze papers presented at the major NLP conferences from 2012 to 2015, using CA to evaluate the joint behavior of our variables. The results show that the introduction of these terms and the related methodologies is characterized by a shallow application of less specialized terms, and, consequently, methodologies to a broad NLP tasks. This can be explained by the need of experimenting these methods in order to prove their benefits in the field. As already stated, from 2015 DL applications change completely and papers are over-focused on attempts of beating the state of art in terms of result performance. We intend to extend our analysis, including more data from the most recent conferences and providing more information about the temporal evolution of each DL term and method in relation to each specific NLP task. This type of analysis will allow to describe deeply the advances in NLP provided by DL, a tsunami bringing a rundown of successes [8].

ACKNOWLEDGMENT

We would like to thank Martin Tutek for his precious help in the extraction task. This research has been partly supported by the European Regional Development Fund under the grantKK.01.1.1.01.0009 (DATACROSS).

REFERENCES

[1] K. Fukushima and M. Sei, Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position, Pattern recognition 15, no. 6, 1982.

[2] T. Young, H. Devamany, P. Soujanya, and E. Cambria. "Recent trends in deep learning based natural language processing." *ieee Computational intelligenCe magazine* 13, no. 3, 2018.

[3] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model", In Eleventh annual conference of the international speech communication associatio, 2010.

[4] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", In Advances in neural information processing systems, 2013.

[5] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank." In Proceedings of the 2013 conference on empirical methods in natural language processing, 2013.

[6] R.cCollobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch." Journal of machine learning research 12, 2011.

[7] L. Deng, and Y. Dong, "Deep learning: methods and applications." Foundations and Trends® in Signal Processing 7.3–4, 2014.

[8] C.D. Manning, "Computational linguistics and deep learning." Computational Linguistics 41.4, 2015

[9] A.R. Fabbri et al. "Tutorialbank: A manually-collected corpus for prerequisite chains, survey extraction and resource recommendation." arXiv preprint arXiv:1805.04617, 2018.

[10] M. Greenacre, Correspondence analysis in practice. Chapman and Hall/CRC, 2017.

[11] J.P. Benzécri, "Histoire et préhistoire de l'analyse des données. Partie V L'analyse des correspondances", Cahiers de l'analyse des données 2.1, 1977.

[12] A. Šilić, A. Morin, J.H. Chauchat, & B. Dalbelo Bašić, Visualization of temporal text collections based on Correspondence Analysis, Expert Systems with Applications, 2012

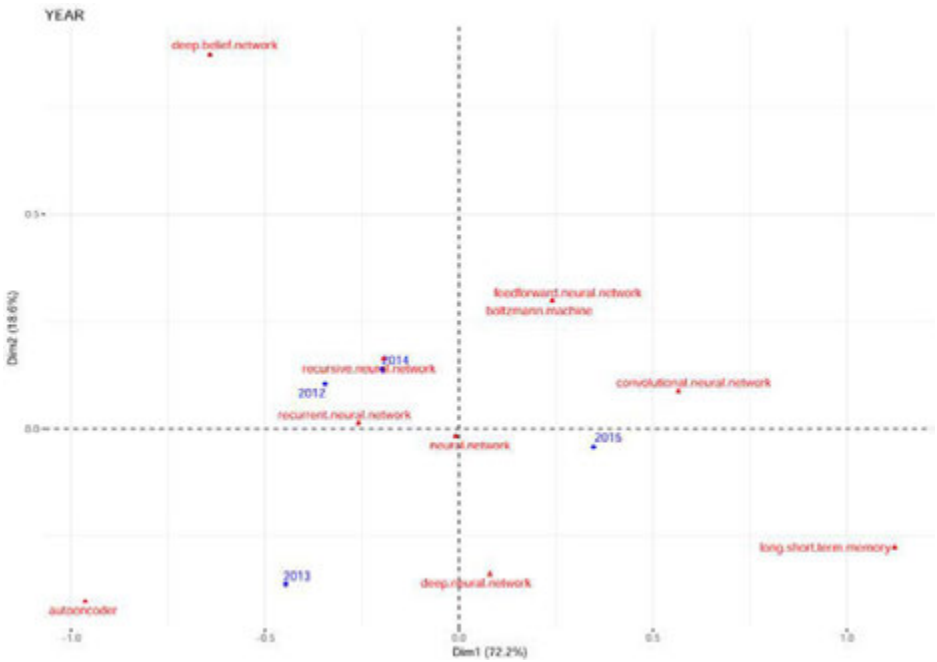


Figure 2. CA Plot for DL and years (2012-2015)

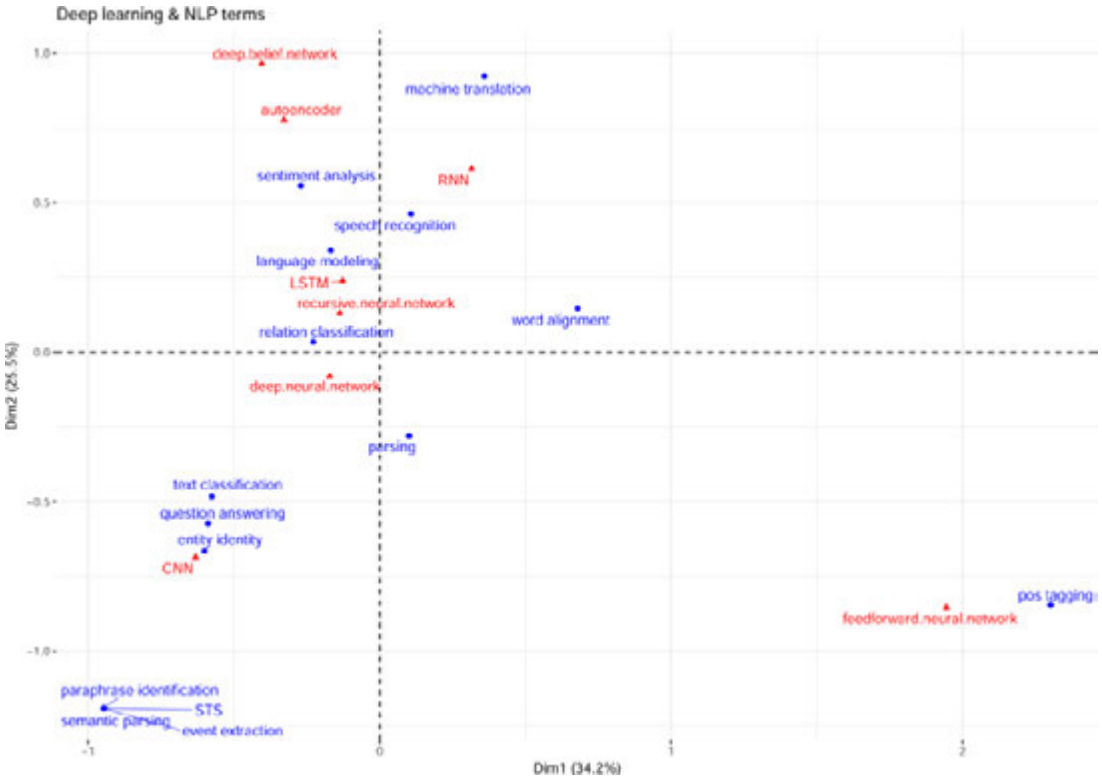


Figure 3. CA Plot for DL and NLP terms

On automated workflow for fine-tuning deepneural network models for table detection in document images

1st Igor Cherepanov
Matrosov Institute for System Dynamics and Control Theory, Siberian Branch of the Russian Academy of Sciences
134 Lermontov St., Irkutsk, Russia, 664033
Institute of Mathematics and Information Technologies of Irkutsk State University
20 Gagarin Blvd., Irkutsk, Russia, 664003
zigorewslike@gmail.com

2nd Andrey Mikhailov
Matrosov Institute for System Dynamics and Control Theory, Siberian Branch of the Russian Academy of Sciences
134 Lermontov St., Irkutsk, Russia, 664033
mikhailov@icc.ru

3rd Alexey Shigarov
Matrosov Institute for System Dynamics and Control Theory, Siberian Branch of the Russian Academy of Sciences
134 Lermontov St., Irkutsk, Russia, 664033
Institute of Mathematics and Information Technologies of Irkutsk State University
20 Gagarin Blvd., Irkutsk, Russia, 664003
shigarov@icc.ru

4th Viacheslav Paramonov
Matrosov Institute for System Dynamics and Control Theory, Siberian Branch of the Russian Academy of Sciences
134 Lermontov St., Irkutsk, Russia, 664033
Institute of Mathematics and Information Technologies of Irkutsk State University
20 Gagarin Blvd., Irkutsk, Russia, 664003
slv@icc.ru

***Abstract*—Nowadays methods and software for extracting tables from document images and portable documents (PDF) continue to be actively developed. One of the promising approaches to this task is the usage of fine-tuned object detection models. However, this approach involves many manipulations with data preparation and training process configuration. This paper proposes an automated workflow for fine-tuning deep neural network models for the table detection in document images. It enables us to automate two sub-tasks: (i) preparing a training dataset in the PascalVOC format with image transformation and augmentation; (ii) training a table detection model by using the well-known Faster R-CNN architecture. Implementation of the workflow design simplifies the use of the approach proposed by decreasing the number of required manipulations.**

***Index Terms*—Document analysis, Table extraction, Table detection, Deep learning, Fine-tuning.**

I. Introduction

Document images store a large amount of valuable tabular data that can be demanded by various scientific and practical applications. To be used in further processing, such information should be represented in a textual and tagged form. Document Engineering community considers the (*table extraction*) task as a task of extracting tables from a document and representing them in a spreadsheet-like format. This task consists of two consecutive sub-tasks: *table detection* and *table structure recognition*. The *table detection* extracts positions of a table in a document, i.e. either geometric positions of a bounding box of the table or a set of all text chunks placed inside this bounding box. The *table structure recognition* recovers positions and the cells content in a table.

Nowadays methods and software for extracting tables from document images and portable documents (PDF) continue to be actively developed. Many table extraction methods are traditionally heuristics-based [11], [13], [14], [19]. Some of them combine heuristics and machine-learning approaches [2], [3], [12]. The current trend involves deep learning techniques: binary classification based on convolutional neural networks [8], fine-tuned object detection models [1], [5], [16], [18], and semantic segmentation [9], [10]. One of the promising approaches to this task implies the use of fine-tuned object detection models. The contemporary deep neural network models for the table detection [1], [5], [16], [18] show a high accuracy on the competition datasets such as UNLV [17], Marmot[1], ICDAR 2013 [7], and ICDAR 2017 [4]). The generic approach to develop such models is based on fine-tuning pre-trained models for object detection on images.

In practice, the complexity of a document layout often hinders the use of research models trained on the competition datasets. To be efficient in real life, the models based on this approach should be trained on task-specific datasets. The training process of such models involves many manipulations with data preparation, such as combination of several heterogeneous datasets that are differentially annotated and their

[1]http://www.icst.pku.edu.cn/cpdp/sjzy/index.htm

conversion to a training dataset represented in a unified format. Moreover, various configurations of training parameters should be tested to reach a sufficient quality (precision and recall) of the fine-tuned model. Therefore, this process typically requires multiple repetition.

We view this problem as the development of an automated workflow for fine-tuning deep neural network models designed for the table detection in document images. It enables us to automate two sub-tasks: (i) preparing a training dataset in the Pascal VOC format with the image transformation and augmentation; (ii) training a table detection model by using the well-known Faster R-CNN architecture [15] with ResNet-101, the pre-trained convolutional neural network model for the feature extraction. The design is implemented as a Python *control script*[2]. It utilizes Tensorflow Object Detection API[3] as a deep learning platform. Implementation of the workflow design simplifies the use of the approach proposed by decreasing the number of required manipulations.

We tested this design and its implementation by creating our own model for table detection. The ANN model was trained on three existing datasets of documents (UNLV, Marmot, and ICDAR 2017) containing 2800 samples total. We adopted and used the image transformation technique from [5]. Additionally, the training data were augmented by affine transformations. This allowed us to improve the accuracy of table detection by 5%. The performance evaluation of the developed model showed a high recall and precision, comparable with the well-known solutions.

II. Design of workflow

This paper describes the object detection method based on TensorFlow Object Detection API for recognizing tables in a document. The TF Records format data [4] should be provided to train a model. To obtain this data we had to carry out a series of manipulations. All workflow can be described by Fig. 1.

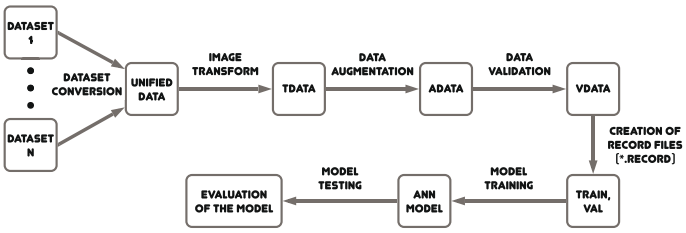


Fig. 1. Work-flow from datasets to ANN models. TData - Transformed data, AData - Augmented data, VData - Validated data

A. Data preparation

Tensor Flow Object Detection API requires a dataset of images and ground truth boxes, packaged as TFRecords of TFExamples. We developed a *control script* for preparing data in this format. It comprises the following steps:

[2]https://https://github.com/tabbydoc/dl4td

[3]https://www.tensorflow.org/lite/models/object_detection/overview

[4]https://github.com/tensorflow/models/tree/master/research/object_detection

Dataset conversion. Many datasets have different formats and should be unified. We developed scripts for transforming data from the following datasets: UNLV Marmot, ICDAR 2017. These scripts convert data to the PASCAL VOC format and have union it to one unified dataset.

Image transform. Pretrained Faster R-CNN models used natural images to train. However, tables are largely different to natural images. To transform document images to natural images, we used a special technique described in [6]. It calculates the precise distance between text regions and white spaces in the document image, which provides a good estimate about the table location (Fig. 2). This technique uses different types of the distance transform, it stores difference features from all image channels. The image transformation is performed in accordance with:

```
img = cv2.imread(path)
#Binarized image
bimage = cv2.cvtColor(img,
  cv2.COLOR_RGB2GRAY)
#DistanceTransform function
b = cv2.distanceTransform(bimage,
  distanceType=cv2.DIST_L2, maskSize=5)
g = cv2.distanceTransform(bimage,
  distanceType=cv2.DIST_L1, maskSize=5)
r = cv2.distanceTransform(bimage,
  distanceType=cv2.DIST_C, maskSize=5)
bimage = cv2.merge((b, g, r))
```

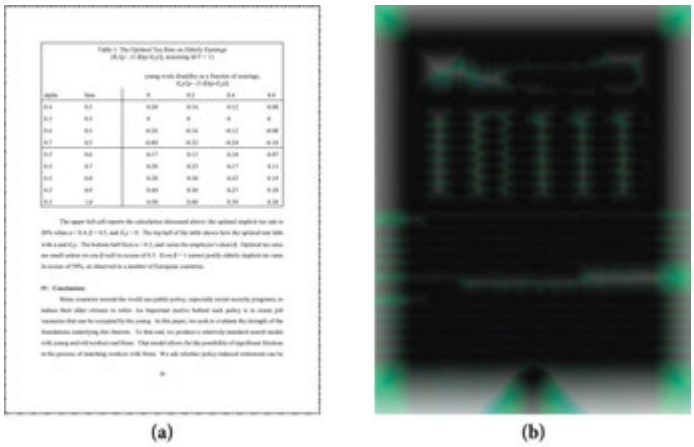


Fig. 2. Sample image conversion: a - Binary image, b - Perform the distance transform algorithm

Data augmentation. To avoid overfitting, Deep CNN heavily depends on big data. Unfortunately, our dataset is small. To increase the training dataset volume, we use data augmentation, which occurs due to a change in the width and the height of images and allows us to get 9 times more data. The following code illustrates the approach to the data conversion and augmentation:

```
scales = [(1, 1), (0.8, 1), (0.9, 1),
          (1.1, 1), (1.2, 1), (1, 0.8),
```

```
          (1, 0.9), (1, 1.1), (1, 1.2)]
image = cv2.imread(image_name)
for scale in scales:
resize_image = cv2.resize(image, None,
  fx=scale[0], fy=scale[1])
```

Figure 3 shows an example of the image transformation and augmentation.

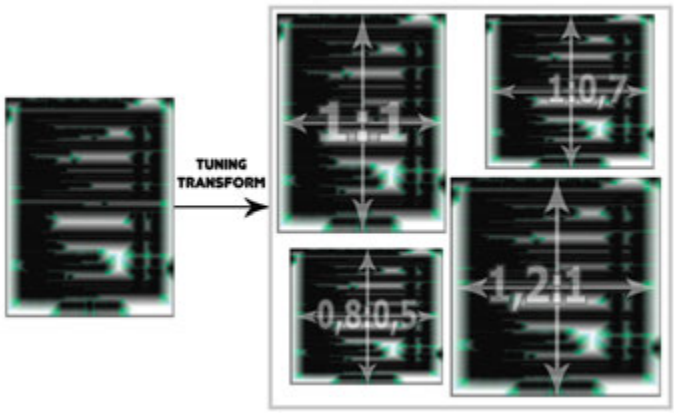


Fig. 3. Image transformation

Data validation. Before creating files in the TF Record format, data should be validated. This should be done because datasets may contain duplicates or XML files with classes that do not participate in the training process. These data are excess. Also, some datasets may contain images without a corresponding XML file, because there is no classified data in these images. These data are excess too.

Create TF Records. Tensor Flow Object Detection API requires a dataset of images and ground truth boxes, packaged as TFRecords of TFExamples. To create such a dataset for table detection, we use the *create_tf_record.py* script from this framework. It can convert data from Pascal VOC format to TF Records. This script derives *train.record* and *val.record* files. The first one is required to train a model, and the second one is used for validation.

B. Model training

After we created the valid dataset, it is possible to run the training model process. This can be done using Google Colab, Anaconda, or simply a Python development environment. TensorFlow and Object Detection API should be installed to train the model.

We use the common Faster R-CNN architecture shown in Fig. 4. The CNN (Convolution Neural Network) processes images and extracts features. The last convolution layer of the CNN is passed to the RPN (Region Proposal Network). RPN finds bounding boxes RoIs (Regions of Interest). These RoIs are combined in the RoIs pool (Region of interest pooling) and, together with the features, are fed to two fully connected layers. The result is an estimate (FC Softmax Classifier) of the model and rectangles (FC BBox repressor).

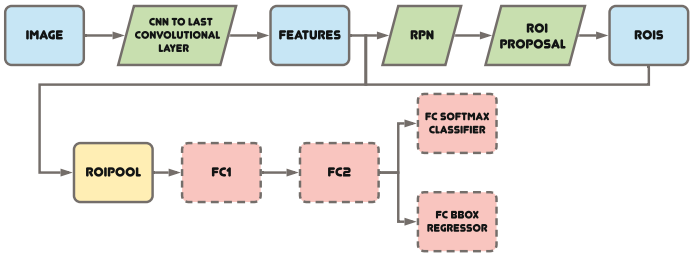


Fig. 4. Faster R-CNN Network Architecture. RPN - Region proposal network, RoIs - Regions of interest, RoIPool - Region of interest pooling, FC - Full connected

C. Case study

To test the trained model, we use TabbyPDF2 (the tool that we had previously developed[5]). It employs the model for prediction tables and writes the result of detection to XML files in the ICDAR 2013 Competition format. It allows us to use the existing script [7] to calculate *Precision* and *Recall*. Precision is defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

Mathematically, Recall is is expressed by the formula:

$$Recall = \frac{TP}{TP + FN}$$

Where TP is True Positives (predictions that we expected to see and received at the output), FP is False Positives (predictions, which should not be output and which the analyzer mistakenly returned at the output), and FN is False Negatives (predictions that we expected to see, but the analyzer did not identify them).

Using this metrics, a model was trained for detecting tables in documents, which gave an estimate of Precision **86,5136%** and Recall **97,9531%**. The model was trained on ICDAR 2017, MARMOT, and UNLV.

D. How to use

To run the *control script*, it is required to specify parameters in the configuration file(*config.ini*). Its content is divided into the following sections:

- *datasets* – include *output_path* and *local_path* parameters. The *output_path* parameter stores the path to data (converted, transformed, augmented, validated data). The *local_path* parameter is the path to the temporary folder.
- *data_name* – includes four parameters: *name*, *path_to_dataset*, *script_to_convert* and *enabled*. The *name* parameter is required to display a message on the console indicating exactly what data is converted (this parameter is necessary for the convenience of reading logs). The *path_to_datasets* parameter include s the path to dataset. The *script_to_convert* contains the path to the script that converts datasets. The *enabled* parameter indicates whether to use this dataset.

[5]https://github.com/tabbydoc/tabbypdf2

- *image_transform* – includes *script_to_transform* that contains the path to script perform image transform.
- *tuning_transform* – contains parameter for running a script that will perform data augmentation.
- *records* – has three parameters: *script_to_create_tf_records*, *path_to_output*, *path_to_label_map*. The *script_to_create_tf_records* parameter includes the path to the script that creates input files of the record type, which will be saved to the *path_to_output* path. The *path_to_label_map* stores the path to *label_map.pbtxt* file. The *label_map.pbtxt* file contains the name and Id of classes.

The implementation of the developed source code is available at github https://github.com/tabbydoc/dl4td.

III. Conclusion

The result of this work is the *control script* that allow to automate data preparation process, namely datasets conversion, image conversion, data augmentation, creation of TF Records files. Our approach can help train table detection models faster and easier.

IV. Acknowledgments

This work was financially supported by the Russian Scientific Foundation (Grant No. 18-71-10001).

References

[1] Arif, S., Shafait, F.: Table detection in document images using foreground and background features. In: Digital Image Computing: Techniques and Applications. pp. 1–8 (2018). https://doi.org/10.1109/DICTA.2018.8615795

[2] Bansal, A., Harit, G., Roy, S.D.: Table extraction from document images using fixed point model. In: Indian Conf. on Computer Vision Graphics and Image Processing. pp. 67:1–67:8. ICVGIP '14 (2014). https://doi.org/10.1145/2683483.2683550

[3] Fan, M., Kim, D.S.: Table region detection on large-scale PDF files without labeled data. CoRR **abs/1506.08891** (2015), http://arxiv.org/abs/1506.08891

[4] Gao, L., Yi, X., Jiang, Z., Hao, L., Tang, Z.: Icdar2017 competition on page object detection. In: 14th IAPR Int. Conf. on Document Analysis and Recognition. vol. 01, pp. 1417–1422 (2017). https://doi.org/10.1109/ICDAR.2017.231

[5] Gilani, A., Qasim, S.R., Malik, I., Shafait, F.: Table detection using deep learning. In: 14th IAPR Int. Conf. on Document Analysis and Recognition. vol. 01, pp. 771–776 (2017). https://doi.org/10.1109/ICDAR.2017.131

[6] Gilani, A., Qasim, S.R., Malik, I., Shafait, F.: Table detection using deep learning. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 771–776. IEEE (2017)

[7] Göbel, M., Hassan, T., Oro, E., Orsi, G.: Icdar 2013 table competition. In: 12th Int. Conf. on Document Analysis and Recognition. pp. 1449–1453 (2013). https://doi.org/10.1109/ICDAR.2013.292

[8] Hao, L., Gao, L., Yi, X., Tang, Z.: A table detection method for pdf documents based on convolutional neural networks. In: 12th IAPR Workshop on Document Analysis Systems. pp. 287–292 (2016). https://doi.org/10.1109/DAS.2016.23

[9] He, D., Cohen, S., Price, B., Kifer, D., Giles, C.L.: Multi-scale multi-task fcn for semantic page segmentation and table detection. In: 14th IAPR Int. Conf. on Document Analysis and Recognition. vol. 01, pp. 254–261 (2017). https://doi.org/10.1109/ICDAR.2017.50

[10] Kavasidis, I., Palazzo, S., Spampinato, C., Pino, C., Giordano, D., Giuffrida, D., Messina, P.: A saliency-based convolutional neural network for table and chart detection in digitized documents. CoRR **abs/1804.06236** (2018), http://arxiv.org/abs/1804.06236

[11] Perez-Arriaga, M., Estrada, T., Abad-Mota, S.: TAO: System for table detection and extraction from pdf documents (2016), https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS16/paper/view/12916

[12] Rashid, S.F., Akmal, A., Adnan, M., Aslam, A.A., Dengel, A.: Table recognition in heterogeneous documents using machine learning. In: 14th IAPR Int. Conf. on Document Analysis and Recognition. vol. 01, pp. 777–782 (2017). https://doi.org/10.1109/ICDAR.2017.132

[13] Rastan, R., Paik, H.Y., Shepherd, J.: Texus: A unified framework for extracting and understanding tables in pdf documents. Information Processing & Management **56**(3), 895–918 (2019). https://doi.org/https://doi.org/10.1016/j.ipm.2019.01.008

[14] Rastan, R., Paik, H.Y., Shepherd, J., Ryu, S.H., Beheshti, A.: TEXUS: Table extraction system for pdf documents. In: Wang, J., Cong, G., Chen, J., Qi, J. (eds.) Databases Theory and Applications. pp. 345–349 (2018)

[15] Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. CoRR **abs/1506.01497** (2015), http://arxiv.org/abs/1506.01497

[16] Schreiber, S., Agne, S., Wolf, I., Dengel, A., Ahmed, S.: DeepDeSRT: Deep learning for detection and structure recognition of tables in document images. In: 14th IAPR Int. Conf. on Document Analysis and Recognition. vol. 01, pp. 1162–1167 (2017). https://doi.org/10.1109/ICDAR.2017.192

[17] Shahab, A., Shafait, F., Kieninger, T., Dengel, A.: An open approach towards the benchmarking of table structure recognition systems. In: 9th IAPR Int. Workshop on Document Analysis Systems. pp. 113–120 (2010). https://doi.org/10.1145/1815330.1815345

[18] Siddiqui, S.A., Malik, M.I., Agne, S., Dengel, A., Ahmed, S.: DeCNT: Deep deformable cnn for table detection. IEEE Access **6**, 74151–74161 (2018). https://doi.org/10.1109/ACCESS.2018.2880211

[19] Tran, D.N., Tran, T.A., Oh, A., Kim, S.H., Na, I.S.: Table detection from document image using vertical arrangement of text blocks. International Journal of Contents **11**(4), 77–85 (2015). https://doi.org/http://dx.doi.org/10.5392/IJoC.2015.11.4.077

StimSeqOnt: An ontology for formal description of multimedia stimuli sequences

Marko Horvat*
* University of Zagreb, Faculty of Electrical Engineering and Computing
Department of Applied Computing
Zagreb, Croatia
marko.horvat3@fer.hr

Abstract - **Sequences of multimedia documents are successfully used in laboratory settings and in practice to deliberately elicit specific emotional reactions. To ensure a successful experiment the emotion provoking stimuli must be selected carefully and have a specific order in which they are presented to the participants. Temporal aspect – duration of individual stimuli within sequences, duration of whole sequences and pauses between stimuli and sequences – must also be chosen with great care. Construction of effective sequences is a delicate and time consuming activity which requires significant group manual effort from domain experts. To facilitate this task we propose a new ontology called StimSeqOnt for formal description of stimuli sequences. The ontology is written in OWL DL language and provides formal and sufficiently expressive representation of affective concepts, high-level semantics, stimuli documents, multimedia formats and repositories used. In StimSeqOnt all relevant metadata about stimuli sequences may be stored as formal concepts. If available, elicited physiological data of previously exposed participants are available for comparison thereby enabling prediction of emotional responses. The StimSeqOnt is designed in compliance with ontology guidelines to facilitate sharing and reuse of expert knowledge.**

Keywords - ontology; multimedia; stimulus; emotion; affective computing; psychophysiology

I. INTRODUCTION

Cognitive science is a compound term for several fields of study which are concerned with understanding the human mind and its processes [1]. As an interdisciplinary field cognitive science brings together philosophy, anthropology, psychology, linguistics, neuroscience, and artificial intelligence [2]. This research is concerned with attaining a comprehensive understanding of intelligence and behaviour, which include mental faculties such as language, perception, memory, attention, reasoning, and emotion [3]. Apart from artificial intelligence, within computer science another field of study is especially close to the goals of cognitive science, and that is affective computing.

Affective computing is a complex interdisciplinary field in computer science. It was defined as a distinct study back in 1997 by Rosalind Picard [4]. Primary goals of researchers in this field are understanding, description, transfer and generation of emotions in computer systems. As such it is closely related to and draws from several subareas in computer sciences like machine learning, signal processing, speech processing, information retrieval, computer vision, knowledge representation and automated reasoning, to name just the most important ones [5].

Elicitation and measurement of emotional reactions is an important area of scientific interest which is aligned both with cognitive science and affective computing. Apart from gaining a deeper insight into how emotions are created, modulated, how they influence perception, attention, thinking, memory or behaviour, such scientific investigations are typically undertaken to study psychological and neurological mechanisms explaining a wide range of behavioural and cognitive phenomena.

It has been experimentally demonstrated that visual stimuli produce measurable physiological responses in skin conductance, startle reflex, breathing and heart rate [6][7][8]. In this regard, other physiological responses have also been intensively explored, as well as multimedia formats for provoking reactions. Per example [9][10][11]. The relationship between multimedia content, emotion, physiology and neuroanatomical correlates is a very complex and intensively researched subject [12].

In any case, for the emotion elicitation using multimedia exposure to be successful, the sequence of the presented content has to have optimal order, duration, pauses, semantic meaning, as well as specific visual and auditory characteristics. The documents in stimuli sequences must be prepared well before an emotion experiment starts and annotated according to semantic and affective models, in order to accurately predict their impact on participants. Because of their use, such multimedia documents are called stimuli. The construction of stimuli sequences, for any particular task such as diagnostics or treatment of mental disorders, is always a complicated task that demands a great deal of expert work [13].

This paper proposes an ontology-based approach for a formal and comprehensive description of affective multimedia sequences for provoking emotions. Such ontology helps to reuse and share expert knowledge, thus contributing to standardization and interoperability in the field of emotion elicitation.

The remainder of the paper is organized as follows; related work is described in the next section. The model of the developed ontology is introduced in the Section 3. This includes core concepts and their relationships. The following section describes in detail the involved ontological concepts, class associations and datatypes. The intended usage of the ontology in a real emotion elicitation

experiment is presented in Section 5. Finally, Section 6 discusses various benefits of the proposed ontology and provides insight into future work plans.

II. Related Work

In recent years, researchers in affective computing, psychology and neuroscience have shown an ever-increasing interest in computer tools for detection and identification of emotions. A large amount of literature exists about each of these topics. Perhaps the best recent insights into this topic come from [14][15].

So far attempts to computationally model construction of emotion have been mostly applied to the appraisal process and cognitive models, including their relationship with stimulation of emotional reactions and the wider phenomena of emotion, i.e. covering all aspects of emotional and affective mental functioning. Per example [16][17][18]. Here the appraisal is defined as the continuous, recursive subjective evaluation of events for their pertinence, as well as the coping potential of the individual. The outcome of the appraisals from these different criteria is predicted to directly drive response patterning of physiological reactions, motor expression, and action preparation [19].

Previously it has been established that ontologies should be utilized to describe higher-semantics and affect in multimedia stimuli [20]. The benefits of formalizing knowledge descriptions are significant and outweigh foreseeable shortfalls [21].

Several miscellaneous ontologies have already been developed especially for the description of emotion. Representative examples of such ontological models are the Emotion Ontology (EMO) [22] and other diverse models for describing and reasoning on emotion context in order to improve emotion detection based on bodily expression [23], describing emotions and their detection [24], and tag-based music recommendation [25].

Finally, the most comprehensive effort into the systematization of emotional models for the usage in computer data processing systems represents the EmotionML standard recommendation developed under the umbrella of W3C Consortium [26]. Although not an ontology, EmotionML is written in XML and serves as an annotation language for multimedia. It currently offers the most sophisticated emotion glossary which includes 5 category vocabularies, 4 dimension vocabularies, 3 appraisal vocabularies and 1 action tendency vocabulary.

III. The StimSeqOnt Model

The StimSeqOnt ("Stimuli Sequence Ontology") is designed to − as stated previously − provide a formal and comprehensive, yet simple and manageable, description of sequences containing emotion-provoking multimedia documents content. Such sequences are designed to intentionally provoke targeted emotional reactions in exposed subjects. The ontology model is shown in Figure 1.

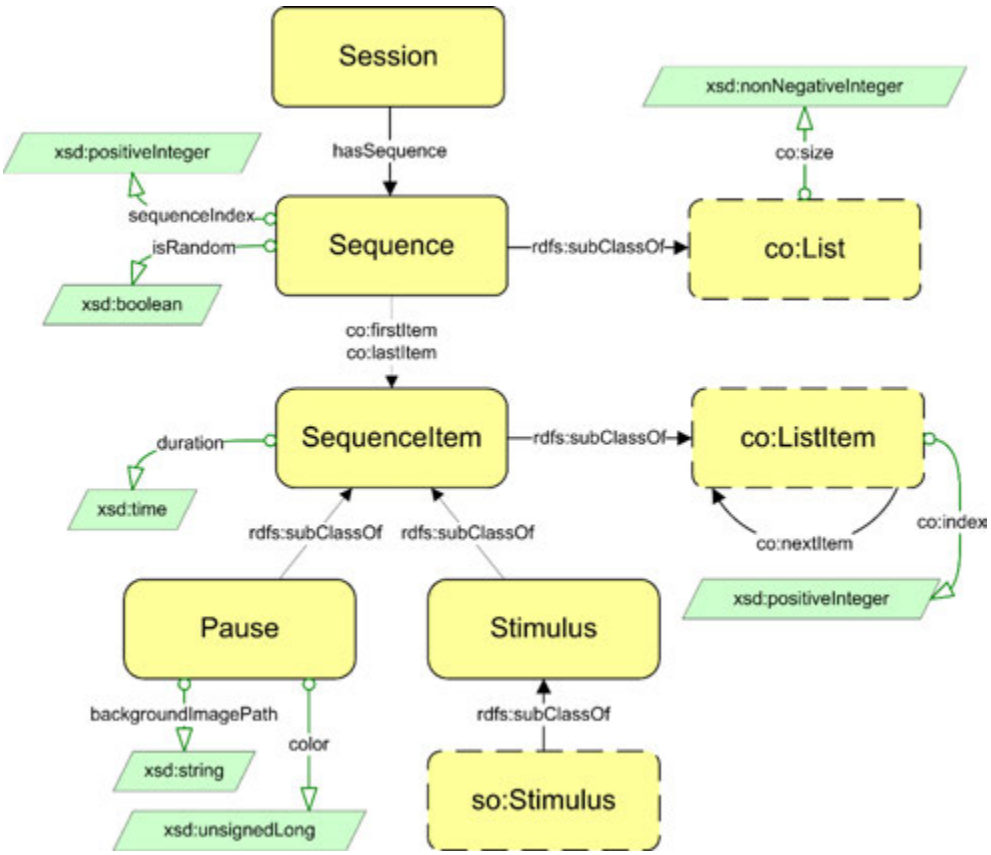


Figure 1. The concepts and relations in the StimSeqOnt model. Data properties (owl:DataProperty) are indicated with green lines while class inheritance (rdfs:subClassOf) and object properties (owl:ObjectProperty) are denoted with black lines. External ontology concepts have dashed borders.

The StimSeqOnt enables a common understanding about the perceived meaning of multimedia stimuli which can be easily shared between different researchers and computer systems. The StimSeqOnt ontology is designed to facilitate querying and retrieval of audio-visual stimuli and stimuli sequences that contain them. The ontology is written in OWL DL and, consequently, is decidable which provides for its practical use. When used with appropriate software tools it may be used to alleviate the construction of stimuli sequences.

In accordance with the most important ontology designing objectives the StimSeqOnt reuses two existing ontologies. They are utilized for formal representation of collections of emotionally annotated multimedia documents: *i*) STIMONT and *ii*) Collections Ontology. Both ontologies have necessary and sufficient expressiveness and decidability and can be used in inferences and SPARQL queries. The STIMONT (prefix "stim") [27] is an upper core ontology designed to provide an integrated and formal description of emotion, high-level semantics, context and physiology content of a multimedia stimulus. The most important feature of the STIMONT is that it provides a formal framework for supporting the explicit, human and machine-processable definition of affective multimedia content. This ontology also facilitates the storage of stimuli in emotionally-annotated databases, stimuli querying and retrieval and construction of stimuli sequences. Its model relies on W3C EmotionML format and extends it with additional emotion vocabularies. High-level semantic expressivity is made possible by reuse of SUMO common sense ontology and SUMO to WordNet mappings. The STIMONT is written in OWL DL.

The Collections Ontology (prefix "co") [28] is an OWL 2 DL ontology developed for creating sets, bags and lists of resources, and for inferring collection properties even in the presence of incomplete information. Although RDF data can be used to define collections and containers to group resources as one entity, this important feature has not been included in OWL and even in OWL 2 DL specifications. The Collections Ontology has been created to address this issue.

The StimSeqOnt is a formalism for the representation of all relevant knowledge about any type of multimedia such as images, sounds, video and text in stimuli sequences. This is achieved by reusing OWL constructs from the STIMONT model that helps to convey high-level semantics interpretation, and even induced emotional states, document metadata and emotion-related physiology.

IV. StimSeqOnt Concepts

The top concept in the StimSeqOnt ontology (prefix "sso") is sso:Session which is directly subsumed by owl:Thing class. An instance of sso:Session is related to one or more instances of sso:Sequence class with hasSequence object relationship:

$$Session \equiv \exists_{\geq 1} Sequence$$

$$hasSequence.Session \sqsubseteq Sequence$$

Sessions can only to be labeled because other important information is stored in other classes and their mutual relationships. The model was intentionally made simple and compact as to be easier to use by experts and in software tools.

In the StimSeqOnt model one session is divided into one or more sequences which contain particular stimuli. Sequences have their index and can be numbered. Each sequence must contain at least one stimulus, and every session consists of at least one sequence. Item in the sequence can either be a pause or stimulus. Pauses can be introduced between adjacent stimuli.

Stimuli are hierarchically organized into sequences. Sequence is a type of List from the Collections Ontology [28]. The co:List cannot be empty, i.e. data property co:index of the class co:ListItem is a positive integer. Therefore, at least one instance of sso:SequenceItem class must exist in ABox.

Each multimedia stimulus or pause is represented in the ABox as exactly one instance of sso:SequenceItem concept. Concept sso:pause can have a background image or a color, and sso:Stimulus is specified by so:Stimulus from STIMONT which provides for a rich description of emotion, high-level semantics, context and physiology content [27].

The order of stimuli in the sequence is defined as in a linked list. In this scheme the first sequence member in the list is attached to a sso:Sequence individual using co:firstItem object relation. The second member in the list is linked to the first with co:nextItem object relation and so on, until, eventually, the last member is denoted using co:lastItem object relation Thus the sequence can be easily traversed by following all sso:Stimulus individuals in the chained list. Duration of the exposure in seconds is captured by sso:SequenceItem duration property. Also, each list item has its index that uniquely identifies it.

V. Formal Modelling of Stimuli Sequences Example

As a realistic example how the StimSeqOnt ontology is supposed to be used, it will be demonstrated how to encode emotionally annotated multimedia stimuli sequences from a previously published experiment [29]. This study was performed at University of Zagreb, Faculty of Electrical Engineering and Computing in cooperation with experts from Department of Psychology, Faculty of Humanities and Social Sciences. In total N=10 participants where exposed to 20 optimal pictures and 20 sounds manually selected by experts for inclusion in the elicitation sequences. The aim of the experiment was to objectively compare emotion elicitation strength between two sequences consisting only of static images and only of videos, respectively. Video clips were constructed using congruent affective pictures and sounds. The required files were selected from the International Affective Picture System (IAPS) [30] and the International Affective Digital Sounds System (IADS) [31] affective multimedia databases which were designed specifically for this type of experiments [32]. A congruent picture from IAPS and a sound from IADS were combined to make one video-clip. Identical pictures were used in both sequences.

A sample of ten IAPS pictures used in the experiment are displayed in Figure 3.



Figure 2. A sample of ten IAPS pictures used as emotion eliciting video clips and images. Happiness (left) and fear dominant emotion stimuli (right column).

The dominant emotions provoked in the experiment purposely had opposite polarity: happiness and fear. The selected stimuli were considered the most likely to induce measurable emotional responses in the participants' population.

Each participant was exposed in two separate sessions or series. Each session consisted of one happiness sequence, one fear inducing sequence, and also of one neutral sequence. A single sequence consisted of either 10 pictures or 10 video-clips (Fig. 3). The length of each stimulus was exactly 15 seconds after which the participant was shown a blank neutral screen and had to write down his affective judgments in a self-assessment questionnaire (SAQ). The neutral blank screen only showed teal color which has an optimal ratio of stimulating positive and negative emotions [29]. When the participant was finished he could resume the sequence by himself (i.e. with a mouse click).

Immediately before the start of the experiment each participant was separately introduced to the stimulation protocol with a neutral sequence. The neutral sequence consisted of one low arousal and valence picture and one video-clip without dominant emotions.

The experimental protocol demanded that half of the participants were first exposed to happiness sequences, and then fear sequences, and the other half of the participants the opposite: they first watched still images and then videos. To prevent the unwanted drift of physiological signals (cardiac and respiratory) before nonneutral sequences participants were exposed to a neutral stimulus until their baseline response was established.

The participants' emotional responses were recorded by two methods: 1) self-assessment responses in pauses between stimuli i 2) real-time monitoring of physiological signals during exposure to stimuli. After each exposure to a stimulus participants filled out a self-assessment questionnaire. The physiological signals – skin conductance, electrocardiogram (ECG), respiration and skin temperature – were monitored using the AcqKnowledge software connected with BIOPAC MP150 physiology acquisition system. The whole setup was synchronized with SuperLab tool for presentation of stimuli to the participants. The tool uses a propriety format to store information about the presentation sequence.

An unspecified long rest period followed each exposure session during which participants relaxed. This was verified by examining the physiological signal parameters that were visualized in real time. The exposure could resume only after the baseline signal levels were reestablished.

The StimSeqOnt individuals stored in ABox for modelling of the described presentation sequence are depicted in the Figure 3. The sequence is designed for stimulation of happiness.

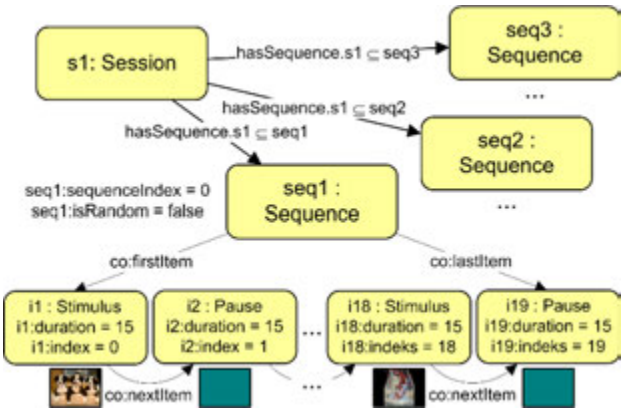


Figure 3. An illustration of real multimedia presentation sequences modelled in StimSeqOnt for the emotion provoking experiment in [32]. Significant individuals stored in ABox are shown.

In the figure above seq1 is the individual representing the happiness provoking sequence, seq2 fear provoking and seq3 the neutral stimuli list. All sequences are related to the single session individual s1. Order of items in the sequences is not randomized. Individuals seq1 and seq2 each include 20 items: 10 emotion provoking multimedia documents and 10 pauses with emotionally neutral images (blank screen in teal color). The first stimulus in seq1 can

be fetched using seq1.firstItem data relationship, and the last with seq2.lastItem or, alternatively, by following co:nextItem relationship starting from the first item in the list. All sequences and items have zero-based indices.

VI. Conlusion and Future Work

The StimSeqOnt ontology presented in the article offers a number of advantages over other possible formats for knowledge representation of emotionally annotated multimedia sequences. We demonstrated how the proposed ontology model can be applied to capture all metadata relevant for construction of a stimuli sequence to be used in an emotion elicitation experiment. This formal approach should facilitate an increase in efficiency of multimedia information retrieval. Although ontologies require more complicated prerequisites, such as reasoning engines, than markup languages or keyword-based annotations, they are more advantageous than the simpler methods. The StimSeqOnt's properties enable a formal, consistent, systematic and expressive model of the affective multimedia sequences. The ontology enables DL-based reasoning about the aggregated content and document metadata.

The StimSeqOnt is specifically designed to interlink with the previously developed STIMONT and Collection Ontology. These models together enable formal representation of high-level semantics, emotion and related states, document context and stimulated physiology that collectively define multimedia stimuli sequences. Both ontologies facilitate knowledge reuse, interoperability and formalization of stimuli information that are superior to the contemporary methods for representation of affective multimedia documents.

In the future, we would like to develop a more expressive upper ontology for the complete description of emotion elicitation experiments. Such even more expressive model would incorporate STIMONT and StimSeqOnt. Additionally, the ontology would add new vocabularies for formal representation about experimentation protocols, employed instrumentation and their settings, recorded and processed psychophysiological signals. This ontology would also need to enable automated reasoning about experimentation results using captured categorical and numerical time-series data. We hope that the presented StimSeqOnt is a step toward this larger goal.

References

[1] A. Collins and D. G. Bobrow, Representation and understanding: Studies in cognitive science. Elsevier, 2017.

[2] G. A. Miller, The cognitive revolution: a historical perspective. Trends in Cognitive Sciences, 2003.

[3] P. Thagard, "Mind: Introduction to cognitive science," Thompson, Cambridge, MA: MIT press, vol. 17, pp. 811−834, 2005.

[4] R. W. Picard, Affective computing. MIT press, 2000.

[5] S. Brave and C. Nass, "Emotion in human-computer interaction," In The human-computer interaction handbook, CRC Press, pp. 103−118, 2007.

[6] M. M. Bradley, M. Codispoti, B. N. Cuthbert, and P. J. Lang, "Emotion and motivation I: defensive and appetitive reactions in picture processing," Emotion, vol. 1(3), pp. 276, 2001.

[7] M. M. Bradley, M. Codispoti, D. Sabatinelli, and P. J. Lang, "Emotion and motivation II: sex differences in picture processing," Emotion, vol. 1(3), pp. 300, 2001.

[8] D. Kukolja, S. Popović, M. Horvat, B. Kovač, and K. Ćosić, "Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications, "International journal of human-computer studies, vol. 72(10), pp. 717−727, 2014.

[9] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," IEEE transactions on affective computing, vol. 3(2), pp. 211–223, 2012.

[10] C. G. Courtney, M. E. Dawson, A. M. Schell, A. Iyer, and T. D. Parsons, "Better than the real thing: Eliciting fear with moving and static computer-generated stimuli," International Journal of Psychophysiology, vol. 78(2), pp. 107−114, 2010.

[11] J. Rottenberg, R. D. Ray, and J. J. Gross, "Emotion elicitation using films," Handbook of emotion elicitation and assessment, Oxford University Press, New York, pp. 9−28, 2007.

[12] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," Biological psychology, vol. 84(3), pp. 394−421, 2010.

[13] K. Ćosić, S. Popović, M. Horvat, D. Kukolja, B. Dropuljić, B. Kovač, and I. Fabek, "Multimodal paradigm for mental readiness training and PTSD prevention," In New tools to enhance posttraumatic stress disorder diagnosis and treatment: invisible wounds of war. IOS Press, 2013.

[14] K. Boehner, R. DePaula, P. Dourish, and P. Sengers, "How emotion is made and measured," International Journal of Human-Computer Studies, vol. 65(4), pp. 275−291, 2007.

[15] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J., Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," In Proceedings of the 19th ACM international conference on multimodal interaction, pp. 524−528, 2017.

[16] G. L. Clore and A. Ortony, "Psychological construction in the OCC model of emotion," Emotion Review, vol. 5(4), pp. 335−343, 2013.

[17] P. C. Ellsworth and K. R. Scherer, "Appraisal processes in emotion," Handbook of affective sciences, vol. 572, V595, 2003.

[18] S. Marsella and J. Gratch, EMA: A computational model of appraisal dynamics, 2006.

[19] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," Consciousness and cognition, vol. 17(2), pp. 484−495, 2008.

[20] M. Horvat, S. Popović, N. Bogunović, K. Ćosić, "Tagging multimedia stimuli with ontologies," Proceedings of the 32nd International Convention MIPRO 2009, Croatian Society for Information and Communication Technology, Electronics and Microelectronics – MIPRO, Opatija, Croatia, pp. 203–208, 2009.

[21] M. Horvat, Generation of multimedia stimuli based on ontological, affective and semantic annotation (Doctoral thesis), University of Zagreb, 2013.

[22] J. Hastings, W. Ceusters, B. Smith, and K. Mulligan, "Dispositions and processes in the Emotion Ontology," 2011.

[23] F. Berthelon and P. Sander, "Emotion ontology for context awareness," In 2013 IEEE 4th International Conference on Cognitive Infocommunications, IEEE, pp. 59−64, 2013.

[24] J. M. López, R. Gil, R. García, I. Cearreta, and N. Garay, "Towards an ontology for describing emotions," In World Summit on Knowledge Society, Springer, Berlin, Heidelberg, pp. 96−104, 2008.

[25] H. H. Kim, "A semantically enhanced tag-based music recommendation using emotion ontology," In Asian Conference on Intelligent Information and Database Systems. Springer, Berlin, Heidelberg, pp. 119−128, March 2013.

[26] F. Burkhardt, C. Pelachaud, B. W. Schuller, and E. Zovato, "EmotionML," In Multimodal interaction with W3C standards. Springer, Cham, pp. 65−80, 2017.

[27] M. Horvat, N. Bogunović, and K. Ćosić, "STIMONT: a core ontology for multimedia stimuli description," Multimedia tools and applications, vol. 73(3), pp. 1103−1127, 2014.

[28] P. Ciccarese and S. Peroni, "The Collections Ontology: creating and handling collections in OWL 2 DL frameworks," Semantic Web, vol. 5(6), pp. 515−529, 2014.

[29] M. Horvat, D. Kukolja, and D. Ivanec, "Comparing affective responses to standardized pictures and videos: A study report," In

MIPRO, 2015 Proceedings of the 38th International Convention, IEEE, pp. 1394−1398, May 2015.

[30] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," Technical Report A−8, University of Florida, Gainesville, FL, 2008.

[31] P. J. Lang and M. M. Bradley, "The International Affective Digitized Sounds (2nd Edition; IADS-2): affective ratings of sounds and instruction manual," Technical report B-3, University of Florida, Gainesville, FL, 2007.

[32] M. Horvat, "A Brief Overview of Affective Multimedia Databases," In Central European Conference on Information and Intelligent Systems (CECIIS 2017), pp. 3−11, 2017.

Machine learning model for detecting high school students as candidates for drop-out from a study program

Đani Pašić, Danijel Kučak
University College Algebra
Ilica 242, Zagreb, Croatia
dani.pasic@racunarstvo.hr, danijel.kucak@algebra.hr

***Abstract* - Transition from high school to university is not successful for all students. Before enrolling in a program, the Admission Office tries to help those students to decide whether that program is best suited for them. They do so by using collected empirical data. The challenge is to classify which students would successfully finish the program and which would not. The students who are most likely not to finish the program successfully should be warned that their decision is not necessarily the best decision for them, and they should consider some other possibilities. This paper proposes one possible criterion for the given challenge: we will develop a machine learning model with the collected data from high schools which the students have attended prior to Algebra and calculate the probability of not finishing the program successfully. Students who would be classified by the model as unsuccessful will receive recommendation for another study program.**

Keywords - machine learning; logistic regression; higher education student; student retention

I. INTRODUCTION

Transition from high school to university can be difficult for some students. It is especially true for students who enroll in courses in engineering fields.

Algebra University College (Algebra) is an example of an engineering university focused on software, system and multimedia engineering. The passing rate for the undergraduate study programs at Algebra within four years of study is between 50% and 60% depending on a study program which is notably less than the approximate 70% reported in [8]. This paper will investigate some reasons why that is so and how can the Admissions Office at Algebra help potential students who would have problems in a study program based on the collected data from their previous education.

Different approaches can be used to find those students, but we can distinguish a few categories of the measured success in a high school. Those categories are collected grades throughout high school, type of high school and additional tests used for measuring students' success.

The problem that we are dealing with can be described as improving student retention. Many papers have been written about that topic, but in the last couple of years, the most notable solutions are rested on machine learning models in higher education. By identifying "at-risk" students early, schools can detect and contact those students and help them to be more successful [3].

In this paper, the machine learning model for detecting high school students as candidates for drop-out from a study program will be introduced. The model will be trained with the data from six generations of Algebra Software and System Engineering students. The probability of not finishing the study program will be calculated and the students who are categorized as the ones who will fail the program would be further considered for admission.

II. MODEL SELECTION

All the steps regarding choosing the appropriate machine learning model will be described in this section. Later, we will be able to continue with the desired implementation of the model.

A. The concept of undergraduate study programs at Algebra

The undergraduate study program at Algebra lasts three years. Each year is divided into two semesters. There are six different study programs. For this research, System Engineering and Software Engineering programs will be studied for collecting the data. We have two types of students: full-time students and part-time students. There is not much difference between those two types except part-time students are employed and full-time students came directly from high schools or they have attended another university before Algebra. Only full-time students will be considered for the research.

B. Choice of input features

After we have specified the categories for measuring students' results in their high school, the categories will be further explained. The first category is students' grades throughout high school. The grades can be prioritized based on the university which the student is interested in. Grade point average (GPA) can be taken as a generic indicator of students' success. Also, high schools can be divided into different types based on the field of their interest. Furthermore, in the Croatia education system, a national scholastic aptitude test is used to measure students' knowledge in some major disciplines of study. Some universities have an additional entrance examination which helps them decide which students have desired skills for entering their university program.

The categories which will be viewed in this paper are students' grades in high school and the type of the high school because these are the only categories which were not changed a lot in the previous decade and because of that they should be the best measures to determine students' success in high school. After we are familiar with the basic concept of undergraduate study programs, we can proceed with choosing the best possible features for defining the model. It is important to notice that not all the data from students' prior education can be collected, so we must limit ourselves to a set of variables that can best be used.

We are focused on engineering study programs, so it is important to know did the students finish technical, mathematics or grammar high school and are they properly prepared for enrolling in this type of program. The high school grades that will be taken into consideration for each student are from these courses: mathematics and informatics. Also, the overall GPA will be used as a common indicator of measuring a student's success in high school.

The types of high school will be marked as 1 (finished mentioned type of school) or 0 (finished any other type of school). Grades have to be taken as an average throughout the years of study. Grades and GPA can have values from 1 to 5, where 5 is the best possible grade a student can get and 2 is the worst grade which guarantees to pass a course.

TABLE I. INPUT FEATURES FOR THE MODEL

Input feature	Description
x1	Finished technical, mathematics or grammar high school
x2	The average grade in mathematics
x3	The average grade in informatics
x4	Overall GPA

C. Defining prediction

We have to denote the output variable for creating the model. As we said in the first subsection, our undergraduate program lasts three years, so we have to decide what does it mean to fail finishing the program. In theory, a student can be a part of the university for an infinite number of years. Because of that, it is decided that only students who have graduated no later than four years after entering the university are classified as the ones who have successfully finished the undergraduate program. Based on that, for training our model, we can only take students who have enrolled in the program at least five years before the current moment to be sure that they fit the request for the output variable.

The output of our model will be 1 or 0. Regarding the fact that we are interested in students who will probably fail, 1 will classify the students who will fail and 0 will classify the students who will successfully finish the undergraduate program. It will help us make a suitable conclusion about the results we get.

D. Machine learning model

Based on the defined prediction, we have to choose an appropriate machine learning model which will be applied to make predictions. Logistic Regression model designed for classification problems will be used [2]. In this paper, we have only two possible outcomes, so we are dealing with a binary classification problem. We aim to find a probability that our output is 1 (student will fail). We can interpret our hypothesis as follows:

$$h_\theta(x) = P(y = 1 \mid x;\ \theta)$$

where P stands for the probability that our output variable will be 1, in the set of input features x1, x2, x3 and x4, where θ is the set of parameter values that will be used to get the best possible classification. We can say that the result of the hypothesis is the probability that a student will fail the undergraduate study program. The result of our hypothesis will be a decimal number between 0 and 1, but the possible values of our output variable are integers 0 and 1. That means we have to provide a threshold that will say what is the boundary between failing students and the students who will successfully finish the study program.

To recap our model's procedure, we first need to calculate the probability of the input features that will result in student failure. Then, we have to apply a threshold to the probability result to classify the student as the one who will or will not fail the undergraduate study program.

III. IMPLEMENTING THE MODEL

In this section, data used for our model will be explained and then, all the steps in the model creation will be described in detail. Completing this section, we will have the prepared machine model for identifying students who are likely to fail the university program based on the collected data from their high school.

A. Data explanation

We have collected the data (input features and output variables) for 161 students. They started their higher education between 2010 and 2015. 70 students from the data set (43%) finished the study program successfully and 91 of them (57%) failed. Success and failure are defined based on the description in the previous subsection. From these numbers, we can notice that the success rate is slightly lesser than was defined for the undergraduate study programs overall. The reason could be the fact that two study programs used for creating our model are harder to be completed than other programs based on the subjective estimates from the students.

B. Steps of model creation

To get some results and make decisions based on that results, the next steps for creating the model have to be defined:

- Scale input features
- Divide the data into three data sets: training set, cross-validation set and test set
- Prepare hypotheses
- Calculate training, cross-validation and test cost results
- Add regularization to the cost function

C. Scale input features

All the input features are not on the same scale. Feature x1 can have values 0 or 1 but features x2, x3 and x4 can range from 1 to 5. To help the Gradient Descent optimization algorithm to find a local minimum faster [9], all the input features will be scaled using the Z-Score method [1]. It centers all the values at zero and scales them so most of the values fall into an inclusive range between -1 and 1. The formula for calculating Z-Score:

$$x_i = \frac{x_i - \mu_i}{s_i}$$

where μ_i is the average of all the values for feature (i) and s_i is the standard deviation of x_i in the training set. The used algorithm for scaling those features can be well extended to support additional input features. We will talk more about it in the part of evaluating the model.

D. Divide the data into three data sets

Before we start with creating our hypotheses, we have to prepare training data and data which will be used for testing our model. The easiest way for us is to use all the data about the students we have to train our model. The problem with this approach is called overfitting [4] and it refers to a model that trains data too well for the specific input features. To avoid this problem, we can divide our training set into two data sets. One for training data, and the second one for testing the model. This is also not the best solution because we will find our test errors applying the test set data and it will determine the error value to be larger for any other set of data. To solve all the mentioned problems, we can introduce an intermediate set called Cross-Validation Set and with using the intermediate set, the test set will give us a precise error result. The data sets for our model will look like this:

- 97 students (60% of the existing data) will be used as the training set
- 32 students (20% of the existing data) will be used as the cross-validation set
- 32 students (20% of the existing data) will be used as the test set

It is important to note here that the data will be shuffled before dividing it into these data sets. Shuffling is an important procedure to avoid setting up the test data to get "optimistic" results.

E. Prepare hypotheses

The fundamental part of the logistic regression model is selecting the best hypothesis. Following the model selection in the previous section, we can make some assumptions which students should fail based on their high school success. Students who have not finished the appropriate high school and have had very bad grades are most likely to fail the study program. But, without having the concrete mathematical results, it is hard to say where is the solid boundary of being a student who will most likely fail or succeed.

To make the selection possible, a few hypotheses have to be defined first. They will be categorized based on their degree of a polynomial. For this paper, degrees 1, 2 and 3 will be used which correspond to linear, quadratic and cubic functions. The hypothesis functions which will be applied:

- Linear function:

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4)$$

- Quadratic function with only quadratic members:

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_1^2 + \theta_6 x_2^2 + \theta_7 x_3^2 + \theta_8 x_4^2)$$

- Quadratic function with all members:

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_1^2 + \theta_6 x_1 x_2 + \theta_7 x_1 x_3 + \theta_8 x_1 x_4 + \theta_9 x_2^2 + \theta_{10} x_2 x_3 + \theta_{11} x_2 x_4 + \theta_{12} x_3^2 + \theta_{13} x_3 x_4 + \theta_{14} x_4^2)$$

- Cubic function:

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_1^2 + \theta_6 x_2^2 + \theta_7 x_3^2 + \theta_8 x_4^2 + \theta_6 x_1^3 + \theta_7 x_2^3 + \theta_8 x_3^3 + \theta_9 x_4^3)$$

In all the proposed functions, function g is the Logistic or Sigmoid Function [10]:

$$g(z) = \frac{1}{1 + e^{-z}}$$

The function maps any decimal number to an interval between 0 and 1 which makes it suitable for classification problems. When z is equal to 0, the function returns 0.5.

As mentioned in the previous section, θ is the set of parameter values that we can describe as the weights assigned to a particular input feature. θ values will be set up in the next subsection regarding calculating cost functions.

F. Calculate cost functions

We need to measure our hypotheses to define which one will be used to predict the outcomes. For this research, the cost function will be used. It takes an average of all the solutions of the hypothesis with the input features compared to the actual outcomes. A vectorized implementation of the function used in this paper:

$$J(\theta) = \frac{1}{m}(-y^T \log(h) - (1-y)^T \log(1-h))$$

where J(θ) is a standard cost function, m is a number of examples, y is the output data and a superscript T means transpose of a matrix. h indicates the following hypothesis:

$$h = g(X\theta)$$

where X is the input data multiplied with the vector θ and g is the sigmoid function. To get the smallest number as a cost function error result, θ values need to be minimized first. Gradient Descent algorithm is used (in the form of Octave's function fminunc) to find the minimized values for the weights. The idea behind using the cost function is to find the hypothesis with the smallest cost on the cross-validation data set. The cross-validation (CV) set is used instead of the training set to get more realistic results by testing the hypothesis with previously unseen examples. It will be applied to all the functions mentioned in the

previous subsection. The results of the cost functions are shown below:

TABLE II. COSTS FOR ALL THE USED FUNCTIONS

Hypothesis function	Training cost	CV cost
Linear function	0.54748	0.62852
Simple quadratic function	0.51346	0.58581
Full quadratic function	0.44772	0.68997
Cubic function	0.49143	0.62741

Based on the results shown in Table II, the simple quadratic function (with only quadratic members) will be used because it has the lowest cost error. The table also shows the costs of the training data set. We can see the trend of training cost becoming smaller by adding more complex functions, but cross-validation cost has the lowest error in simple quadratic function which means that adding too much complexity results in overfitting problem.

G. Add regularization to the cost function

After we have found the hypothesis function with the lowest cost error, we will try to get even lower cost error by adding regularization to the function. Regularization is the process of adding information to the function to reduce the parameters θ. It is designed to address the problem of overfitting.

For this research, we will use 8 different lambdas (λ) starting from 0.01 to 0.64, where each of the following values is multiplied by 2. Also, lambda 0 will be used to show that the result will not be changed compared to the cost error we got without using regularization. A vectorized implementation of the function by adding regularization:

$$J(\theta) = \frac{1}{m}(-y^T \log(h) - (1-y)^T \log(1-h)) + \frac{\lambda}{2m}\theta^T\theta$$

where the first part is the same as we use for the implementation without regularization and the regularization part is added to the result. θ vectors used in this function are cut by the first value because we should not regularize the first parameter in θ vector. The results of the regularized logistic regression function by presenting the cross-validation costs are shown below:

TABLE III. CROSS-VALIDATION COSTS FOR SIMPLE QUADRATIC FUNCTION WITH DIFFERENT LAMBDAS

λ	CV cost
0	0.58581
0.01	0.58630
0.02	0.58657
0.04	0.58717
0.08	0.58763
0.16	0.58922
0.32	0.59443
0.64	0.60634
0.08	0.58763

From the results shown in Table III, it is obvious that regularization has not helped us because the cost errors are slightly getting worse on each lambda increase. We will stay with the function where lambda value is equal to 0 and with these θ parameters:

TABLE IV. WEIGHTS FOR PARAMETER SET Θ

Parameter	Weight
θ_0	-0.38004
θ_1	-0.52652
θ_2	-0.35741
θ_3	0.26973
θ_4	-1.16590
θ_5	0.23344
θ_6	0.12676
θ_7	0.28080
θ_8	0.44463

Table IV shows used weights for parameters set to θ in quadratic function with only quadratic members.

IV. VALIDATING THE MODEL

The logistic regression model has been successfully implemented in the previous section. Now, we need to validate it by using different calculations and by plotting the appropriate graph.

A. Validation steps

To validate the model, the next steps need to be done:

- Calculate precision, recall [5], F1 score and misclassification error [6]
- Plot train error and cross-validation error learning curves [7]

B. Finding the threshold and calculating misclassification error

To interpret the results the right way, we need to find the threshold which will be the most suitable for our classification problem. In the process of finding the threshold, precision, recall and F1 score need to be calculated first. Definitions for calculating precision and recall:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Precision tells us how many positive outcomes (student fails the study program in our case) have been guessed from the number of predicted positive outcomes. On the other hand, recall tells us how many positive outcomes have been guessed from all the positive outcomes that occurred. For this research, a more important measure is the recall because we do not want to miss students who will actually fail. Precision is a minor problem in our case if it is not too low. It is not easy to say what does it mean "too low" for us, but we should be satisfied if at least two of three students guess correctly.

To find the most suitable threshold, we need to iterate through an array of predefined thresholds and see which combination should suit us best. We will use thresholds from 0.3 to 0.7 in a step of 0.05. The third metric that will be applied in this paper is called F1 score. It will help us

decide which sequence of precision and recall is optimal for us. Definition of F1 score:

$$\mathrm{F}_1 \text{ score} = 2\frac{Precision * Recall}{Precision + Recall}$$

The threshold which will have the highest F1 score will be used for students' classification. The results of the tested thresholds are written below:

TABLE V. PRECISION, RECALL AND F1 SCORE FOR DIFFERENT THRESHOLDS

Threshold	Precision	Recall	F1 score
0.3	0.66667	1	0.80000
0.35	0.76190	1	0.86486
0.4	0.78947	0.93750	0.85714
0.45	0.78947	0.93750	0.85714
0.5	0.77778	0.87500	0.82353
0.55	0.86667	0.81250	0.83871
0.6	0.85714	0.75000	0.80000
0.65	0.92308	0.75000	0.82759
0.7	0.91667	0.68750	0.78571

Based on the results shown in Table V, our pick will be threshold 0.35. It has a perfect recall result, where precision is satisfying if we interpret it like three of four students will be guessed correctly. As we said earlier, the recall result is more important to us, so we can say that the results are more than satisfying to us.

Additional validation is calculating the misclassification error based on the chosen model. Definition of misclassification error (ME):

$$ME = \frac{False\ Positives + False\ Negatives}{Number\ of\ test\ examples}$$

It tells us what the number of misclassified students is or how many wrong predictions do we have. For our model with threshold 0.35 being used, we got a misclassification error of 0.15625 which is a quite good result.

C. Plot learning curves

After we have found the threshold and the misclassification error, we will try to anticipate the performance of our model by adding more training examples. To achieve that, a learning curve graph is plotted and shown below:

Figure 1. shows the learning curve graph for our model. After about sixty training examples, the error curves tend to be very close to each other. It means that our model is suffering from high bias. It should benefit from adding additional input features and increasing the complexity of the hypothesis function. Adding more training data will not help us much by itself.

V. CONLUSION

Transition from high school to university is not an easy process but it can be improved for both students and universities by using machine learning model like the one explored in this paper. Each university has its own rules for defining criteria for enrolling students. Those rules can be defined by this kind of model which helps us to determine which students would be successful. By implementing and validating the model, we have the required information

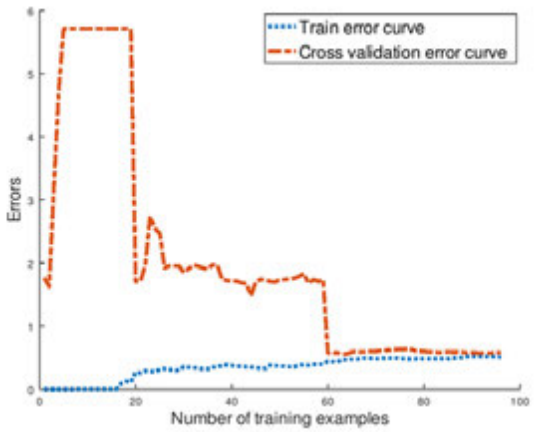


Figure 1. Learning curve graph

which will classify a student as the one who will fail or will finish the study program successfully. Students who will have 35% or more chances of failure are most likely to drop-out from the study program and this allows the Admissions Office to help those students to make a proper decision about their education.

The next step is to create an application that will make it easier for any user to calculate their success in higher education for the chosen study program. The application should be updated each time after the logistic regression model is updated to be able to calculate the most accurate predictions.

REFERENCES

[1] Brase, Charles Henry, and Corrinne Pellillo Brase. Understanding Basic Statistics, Enhanced. Cengage Learning, 2016.

[2] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. Vol. 398. John Wiley & Sons, 2013.

[3] Kučak, Danijel, Vedran Juričić, and Goran Đambić. "MACHINE LEARNING IN EDUCATION-A SURVEY OF CURRENT RESEARCH TRENDS." Annals of DAAAM & Proceedings 29 (2018).

[4] Burnham, K., and D. Anderson. "Model Selection and Multi-model Inference ‘. 2nd edn.(Springer: New York.)." (2002).

[5] Buckland, Michael, and Fredric Gey. "The relationship between recall and precision." Journal of the American society for information science 45.1 (1994): 12-19.

[6] Ounpraseuth, Songthip, et al. "Estimating misclassification error: a closer look at cross-validation based methods." BMC research notes 5.1 (2012): 656.

[7] Anzanello, Michel Jose, and Flavio Sanson Fogliatto. "Learning curve models and applications: Literature review and research directions." International Journal of Industrial Ergonomics 41.5 (2011): 573-583.

[8] “Universities are committed to improving study success rates” VSNU, Association of Universities in the Netherlands, vsnu.nl/en_GB/study-success-rates-universities.

[9] Chi, Jocelyn T., Eric C. Chi. “Getting to the Bottom of Regression with Gradient Descent” Statistics Views, Wiley, 10 Jan 2014, statisticsviews.com/details/feature/5722691/Getting-to-the-Bottom-ofRegression-with-Gradient-Descent.html.

[10] Weisstein, Eric W. "Sigmoid function." (2002).

Evaluation of Structural Hyperparameters for Text Classification with LSTM Networks

M. Frković*, N. Čerkez**, B. Vrdoljak* and S. Skansi***
*University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
**College of Information Technologies (VSITE), Zagreb, Croatia
***University of Zagreb Faculty of Croatian Studies, Zagreb, Croatia
marija.frkovic@fer.hr
ninoslav.cerkez@vsite.hr
boris.vrdoljak@fer.hr
sskansi@hrstud.hr

***Abstract* - In natural language processing, most problems can be interpreted as text classification tasks, which makes this type of task a central one. A natural subdivision of more complex types of text classification can be made according to whether the classification is multilabel or multiclass, where both of these can be tackled with either a multiclass classifier or with a combination of several binary classifiers. An example of the problem which offers a natural way of comparing multiclass and binary classification on the same data, which makes the results comparable, is a classification of text author MBTI personality type. The dataset used is PersonalityCafe MBTI. We focus our comparison on structural hyperparameters, which are the hyperparameters pertaining to network structure. Structural hyperparameters are necessary for specifying the network itself, which makes them a primary concern in its construction. The hyperparameters investigated in this paper are the number of hidden layers and layer size. Through a number of experiments, we demonstrate the choice of hyperparameters and conclude with general hyperparameter selection recommendations based on our results.**

***Keywords* - natural language processing; LSTM; structural hyperparameters; multiclass classification; binary classification.**

I. INTRODUCTION

Deep neural networks can discover representations needed for detection or classification on multiple levels, starting from simple raw input to higher and more abstract level [1] An important application of deep neural networks can be found in natural language processing (NLP), which is an interdisciplinary field of computer science, artificial intelligence, and linguistics, based on tools and algorithms for understanding and processing naturally spoken and written language [3].

Long Short-Term Memory networks (LSTM networks) [11] are a type of recurrent neural network capable of learning long-term dependencies, initially invented to address the vanishing gradient problem [8]. The LSTM networks have a structure of linked memory cell blocks, where the information is stored. The memory cell has a recurrent connection to the previous time step, which is multiplicatively gated by another cell that learns to decide when to clear the content of the memory [4].

Problems in NLP, which can be interpreted as text classification tasks, have been successfully addressed with LSTM networks. A natural subdivision of more complex text classification tasks can be made according to whether the classification is multilabel or multiclass. Multiclass text classification can be tackled with either a multiclass classifier [14] or with a combination of several binary classifiers [12]. An example of the problem which offers a natural way of comparing multiclass and binary classification on the same data, which makes the results comparable, is a classification of text author MBTI personality type.

The primary motivation behind this paper is to make a clear comparison and guideline for hyperparameter selection based on a few experiments on the same dataset. This motivation has been inspired by the complexity of the neural networks, where a significant number of elements can affect the result of training and their ability to classify data correctly. We focus our comparison on structural hyperparameters, which are the hyperparameters pertaining to network structure. Structural hyperparameters are necessary for specifying the network itself, which makes them a primary concern in its construction. The hyperparameters considered in our paper are the number of hidden layers and layer size.

We approach the personality type classification problem in two different ways: through multiclass and binary classification. Through several experiments, we demonstrate the choice of hyperparameters and conclude with general hyperparameter selection recommendations based on our results. The obtained results will be used as a basis for later hyperparameter optimization to achieve higher accuracy in personality type prediction.

The paper is organized as follows: Section II describes text processing referring to the chosen dataset; Section III presents experimental setup with the emphasis on hyperparameters; The results and discussion for multiclass and binary classification are shown in section IV; and, section V concludes this paper.

II. TEXT PREPROCESSING

Before building an LSTM neural network model and performing hyperparameter analysis, we preprocess the data in the standard way.

A. Dataset

The dataset used for this research is an MBTI Type dataset from Kaggle[1] which consists of 8.675 rows containing (a) textual posts from PersonalityCafe forum and (b) labels, which are MBTI personality types of their respective authors. Each author has an average of 50 posts. An MBTI dataset was selected because MBTI labels offer a rather complex yet well-defined multiclass text classification task, with 16 overlapping classes, which can also be broken down to binary classifications in a very natural way with four well-defined classifiers: E/I, N/S, T/F, P/J. Because the tasks are equally well defined for both the multiclass and binary setting, MBTI datasets are an excellent domain for comparing these two approaches. For more on the theory of MBTI personality types we refer the reader to [2]. The dataset used is quite large but imbalanced. Since our research does not focus on surpassing a benchmark but rather on hyperparameter evaluation of two approaches, no measure was taken to correct the imbalance as it would not influence the results of our analysis.

B. Word Embeddings

Data from the dataset was prepared and cleaned by removing links, numbers, punctuation, and special characters. Contractions were expanded to their full form. All uppercase characters were converted to lowercase. After the tokenization of sentences, where they were transformed into a list of words, stop words and one-letter words were removed. Since the posts contain a lot of MBTI type code words, they were also removed from the text, to prevent unnecessary noise in the text data. Lemmatization was performed using WordNetLemmatizer. The text was finally converted into word embeddings, using the Glove 100-dimensional word, pretrained on Twitter corpus of 2 billion tweets [9] .

III. Experiment Setup

The focus of this research is to compare different LSTM models based on structural hyperparameters.

Hyperparameters are parameters that are not learned during training but must be defined before training. As such, research in hyperparameters forms the only viable basis for selecting good hyperparameters. For the purpose of this paper, we will provisionally define two non-disjoint classes of hyperparameters. Backpropagation hyperparameters are hyperparameters controlling the optimization of the network parameters. These are, among others, regularization, learning rate, number of epochs, momentum, etc. In this paper, we are interested in a second class, namely the structural hyperparameters.

As the name suggests, these are parameters governing the neural network structure. The structural parameters we explore are the hidden layer size and the number of hidden layers.

Structural hyperparameters can be considered the dominant hyperparameters in learning since they are a necessary part of creating internal representations in neural networks, as opposed to optional parameters such as regularization.

In our experiments, we trained models of varying numbers of hidden layers and hidden layer size, which we will discuss below. To keep the comparison clear, we have opted not to use any regularization. Other necessary (non-structural) hyperparameters were kept invariant in all models: sigmoid activation for binary and softmax activation for multiclass classification, the batch size was set to 64, the number of epochs to 30, the optimizer was ADAM [10] with the learning rate of 0.001, and no regularization or dropout [5] was used. We have trained all models with both cross-entropy loss (C-E) and mean squared error loss (MSE). Modifying these parameters might improve results even further, but this will be left as a topic for future research. The architecture of the individual layers was Keras/Tensorflow CuDNNLSTM[2], trained on CUDA 10.1 GPUs.

By fixing the dataset and epochs, we deliberately limit the hyperparameters in question and focus only on the combination of layer size and number of neurons per layer (with two different loss functions), which to the best of our knowledge was not systematically explored for multi-layered LSTM networks with textual embeddings for multiclass text-based document classification. Results are evaluated by comparing multiple metrics, such as accuracy, F1-score, precision, recall, confusion matrix.

IV. Results

The results we obtained for multiclass text classification from the experiments conducted, in terms of accuracy, recall, precision, and F1 score are presented in TABLE I (we have bolded the best three results per column).

The results show that when it comes to accuracy, deeper neural networks tend to give better results. When looking at the average macro precision, recall, and F1 score, networks with 3 layers give similar or slightly worse results than the ones with 1 or 2 layers.

When it comes to the number of neurons per layer, in the networks with a single LSTM layer, a small number of neurons gives better results. In fact, models with 100, 150 and 500 neurons per layer have shown the signs of overfitting (Fig.1a), because after around 15th epoch there is a steep increase in training accuracy and decrease in training loss, but at the same time the validation accuracy is not changing or getting slightly worse, and validation loss is starting to increase as well, as shown in Fig.1b, Fig.1c and Fig.1d.

[1] https://www.kaggle.com/datasnaek/mbti-type

[2] https://keras.io/

TABLE I. MULTICLASS CLASSIFICATION RESULTS

Model	NL	NN	Accuracy [%]		Precision		Recall		F1 score		Training time [s]	
			CE	*MSE*	*CE*	*MSE*	*CE*	*MSE*	*CE*	*MSE*	*CE*	*MSE*
model1	1	5	27,419	26,114	0,063	**0,099**	0,095	0,090	0,073	0,067	76,73	65,70
model2	1	10	19,201	**28,111**	0,039	**0,099**	0,060	**0,104**	0,036	**0,091**	78,80	64,67
model3	1	25	24,347	20,891	0,072	0,060	0,085	0,072	0,071	0,059	87,13	65,77
model4	1	50	22,657	19,201	0,093	0,048	0,080	0,063	0,068	0,048	94,49	74,46
model5	1	100	18,049	21,736	0,068	0,083	0,070	0,088	0,058	0,083	105,49	85,38
model6	1	150	20,046	20,046	**0,098**	0,080	**0,114**	0,082	**0,101**	0,078	143,38	104,37
model7	1	500	16,820	24,654	0,068	0,076	0,071	0,092	0,066	0,080	297,81	337,24
model8	2	5, 5	25,346	26,344	0,053	0,059	0,085	0,087	0,062	0,049	148,56	115,01
model9	2	10, 10	24,117	19,432	0,070	0,085	0,089	0,060	0,075	0,033	153,46	116,14
model10	2	25, 25	**28,034**	27,650	**0,106**	0,084	**0,110**	**0,101**	**0,101**	**0,086**	164,91	118,16
model11	2	50, 50	24,962	24,270	0,080	0,068	0,093	0,091	0,078	0,071	183,36	127,76
model12	2	100, 100	24,654	21,889	**0,136**	**0,098**	0,091	0,091	0,080	**0,088**	208,68	145,71
model13	2	150, 150	19,662	18,126	0,081	0,083	0,084	0,082	0,078	0,077	294,21	196,66
model14	2	500, 500	20,891	24,654	0,083	0,082	0,085	0,093	**0,081**	0,074	643,09	805,44
model15	3	5, 5, 5	21,121	20,968	0,054	0,042	0,067	0,069	0,036	0,050	183,54	158,84
model16	3	10, 10, 10	27,650	27,189	0,048	0,049	0,092	0,088	0,061	0,050	212,75	161,01
model17	3	25, 25, 25	22,657	21,813	0,029	0,014	0,066	0,062	0,030	0,022	234,74	157,26
model18	3	50, 50, 50	**28,725**	24,424	0,045	0,032	0,093	0,074	0,053	0,041	248,66	174,92
model19	3	100, 100,100	**29,416**	**29,339**	0,037	0,063	**0,096**	0,097	0,053	0,063	299,14	209,79
model20	3	150, 150, 150	27,727	**29,032**	0,061	0,066	0,095	**0,098**	0,073	0,066	416,82	277,95
model21	3	500, 500, 500	21,045	21,582	0,025	0,073	0,061	0,083	0,027	0,073	962,75	1245,97
model22	3	25, 50, 100	16,743	23,118	0,025	0,035	0,063	0,068	0,022	0,035	272,30	198,35
model23	3	100, 50, 25	27,573	24,424	0,041	0,070	0,088	0,088	0,053	0,070	276,73	186,82

NL-Number of Layers, NN-Number of Neurons per Layer, CE – Cross Entropy

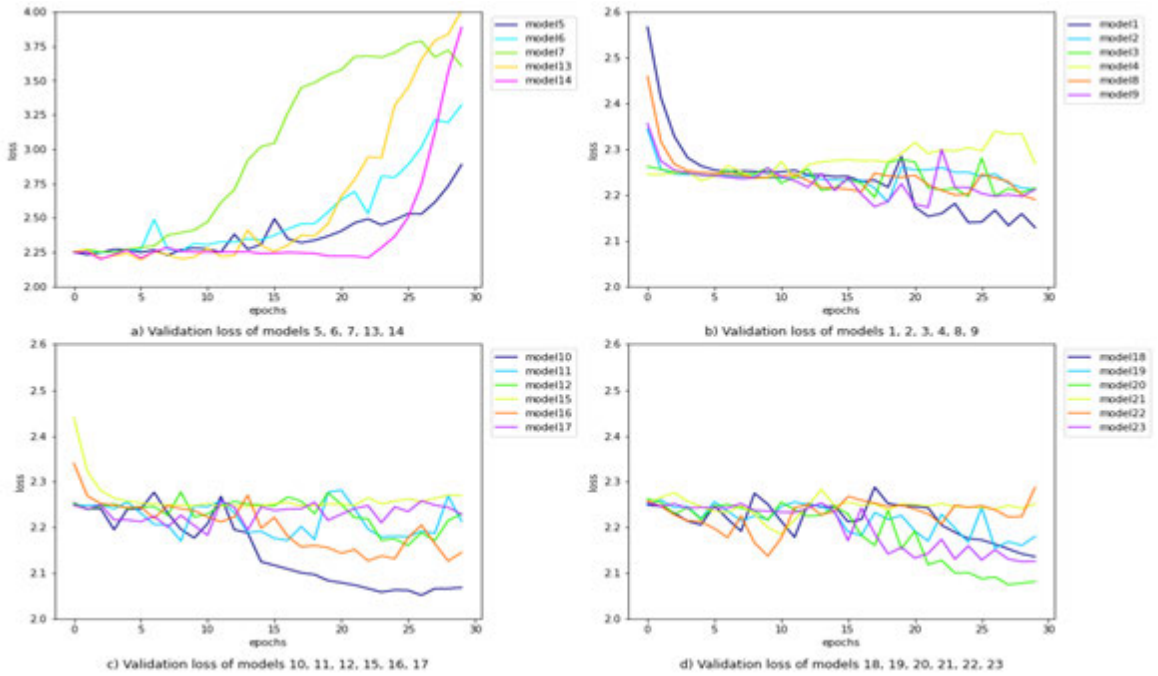


Figure 1. Validation cross-entropy loss

Networks with two LSTM layers also do not predict well with a high number of neurons. Unlike single-layer LSTM networks, they start to overfit when having 150 or 500 neurons. Networks with three LSTM layers have shown signs of overfitting when having 500 neurons per layer. Three-layered LSTMs with five neurons per layer produced poor results in terms of accuracy, but for the all layer sizes between 5 and 500 they gave satisfactory results (especially when considering training time)When comparing results for MSE loss and Cross-entropy loss, there is no significant difference. This can be seen in Fig. 2 and Fig. 3, which show 5-period simple moving averages (5-SMA) of validation accuracies of models trained with both cross-entropy (C-E) and mean squared error (MSE) loss functions. Two interesting exceptions are models 7 and 14 (one layer with 500 neurons, and two layers with 500 neurons each) where models trained with mean squared error performed quite well and have not shown any significant signs of overfitting, unlike models trained with cross-entropy.

In Fig. 4 both loss functions are shown for model 1, normalized to the interval between 0 and 1, in order to give better insight into the difference (or similarity) between them.

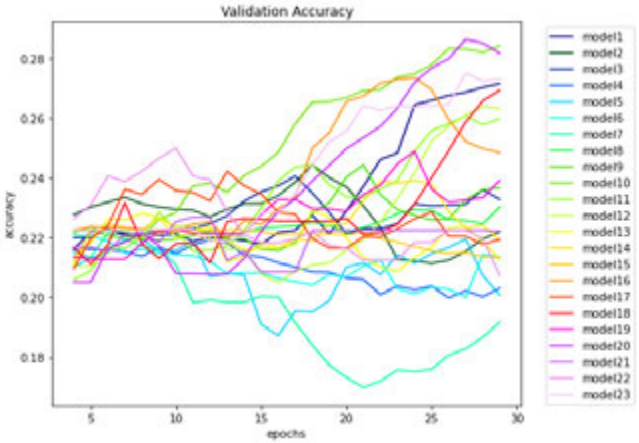


Figure 2. 5-SMA of validation accuracy with C-E loss

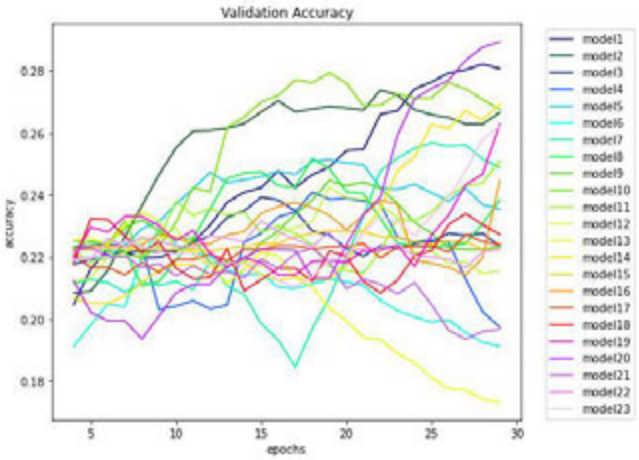


Figure 3. 5-SMA of validation accuracy with MSE loss

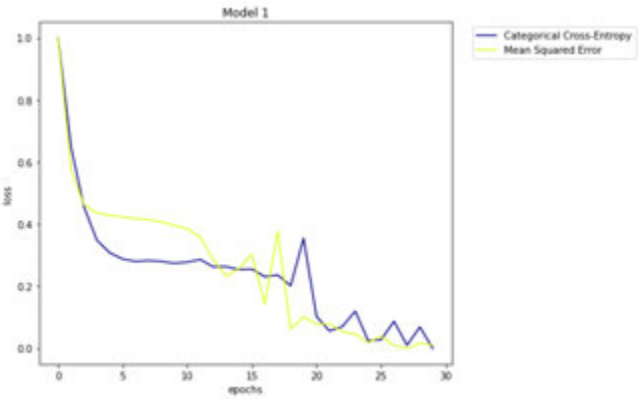


Figure 4. Comparison of C-E and MSE for Model 1

When it comes to average metrics values, they are very similar and we cannot determine which one is better, which is unexpected given the fact that cross-entropy is considered better performing in classification tasks.

As already mentioned, multiclass text classification can be tackled also with a combination of several binary classifiers. We have trained several models which use only binary classifiers for comparison with the multiclass models. Each model consists of four binary classifiers, one for each of personality preferences: introversion(I)/extraversion(E), sensing(S)/intuition(N), thinking(T)/feeling(F) and judging(J)/perceiving(P).

Across all models, the classificator T/F outperformed other classificators (I/E, S/N, J/P) which had problems with overfitting from epoch 5 onwards, as shown on Fig. 5 for model 8. For this reason, we have reduced the learning rate from 0.001 to 0.0001. Models with a higher number of layers as well as models with a higher number of neurons per layer showed a better F1 score.

LSTM network models for T/F prediction outperformed other models, but it is worthwhile to note that for T/F the dataset was quite balanced (3.981 posts for T and 4.693 posts for F), whereas all other classes were imbalanced[3] This is the reason the T/F classifier was chosen as the representative for binary classification models. Based on the results for T/F, we can presume all other classifiers would work similarly well on a larger and more balanced dataset in terms of relative number of binary labels.

The experiment results are shown in Table II. Models with 2 LSTM layers had the highest average accuracy, but models with only 1 layer showed the highest average F1 macro score. On average, simpler models outperformed other binary classificators. As an illustration we showcase the accuracy (Fig. 5) and loss (Fig. 6) for model 8 as an example of the typical behavior of binary classifiers.

However, the model which gave the best results on the test dataset was model 18, with F1 macro score of 0.716 and accuracy of 71.81%, followed by the model 7, with F1 macro score of 0.712 and accuracy of 72.04%. The number of neurons per layer had no significant impact. The 5-SMA of the loss for all T/F models is shown in Fig.7.

[3] I 6.675, E 1.999, P 5.240, J 3.434, N 7.477, S 1.197

TABLE II. BINARY CLASSIFICATION RESULTS

Model	Number of layers	Number of neurons	Accuracy [%]	Precision	Recall	F1 score	Training Time [s]
model1	1	5	56,605	0,559	0,559	0,559	80,22
model2	1	10	66,052	0,656	0,653	0,653	74,35
model3	1	25	64,900	0,644	0,641	0,641	76,65
model4	1	50	53,763	0,509	0,506	0,469	84,01
model5	1	100	55,991	0,555	0,554	0,555	99,37
model6	1	150	58,986	0,675	0,621	0,567	135,18
model7	1	500	**72,043**	**0,720**	**0,711**	**0,712**	295,60
model8	2	5, 5	62,366	0,628	0,629	0,624	123,39
model9	2	10, 10	70,276	0,700	0,700	0,700	127,06
model10	2	25, 25	56,912	0,584	0,526	0,447	130,03
model11	2	50, 50	**70,353**	0,704	**0,706**	**0,703**	143,91
model12	2	100, 100	55,223	0,526	0,500	0,359	173,72
model13	2	150, 150	68,356	0,695	0,693	0,683	256,33
model14	2	500, 500	58,909	0,585	0,563	0,545	607,40
model15	3	5, 5, 5	66,206	0,680	0,675	0,661	177,29
model16	3	10, 10, 10	52,995	0,678	0,571	0,470	181,15
model17	3	25, 25, 25	55,376	**0,777**	0,502	0,360	183,98
model18	3	50, 50, 50	**71,813**	0,716	**0,717**	**0,716**	199,23
model19	3	100, 100, 100	56,068	0,547	0,541	0,532	255,29
model20	3	150, 150, 150	66,743	0,670	0,651	0,649	377,69
model21	3	500, 500, 500	**70,353**	**0,722**	0,683	0,681	945,07
model22	3	25, 50, 100	55,223	0,276	0,500	0,356	217,40
model23	3	100, 50, 25	59,601	0,652	0,555	0,492	234,87

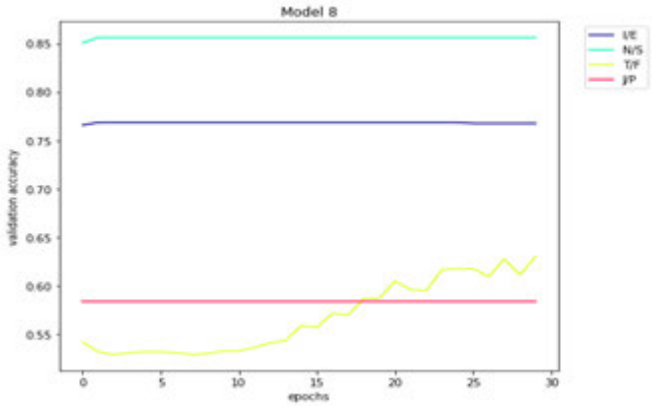


Fig 5. Validation accuracy of model 8 for 4 binary classification problems

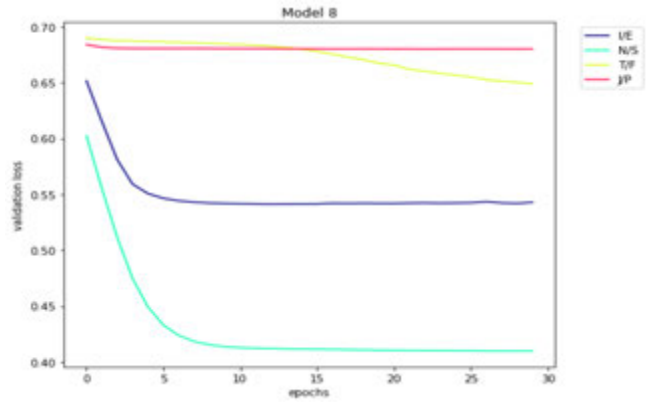


Fig 6. Validation loss of model 8 for 4 binary classification problems

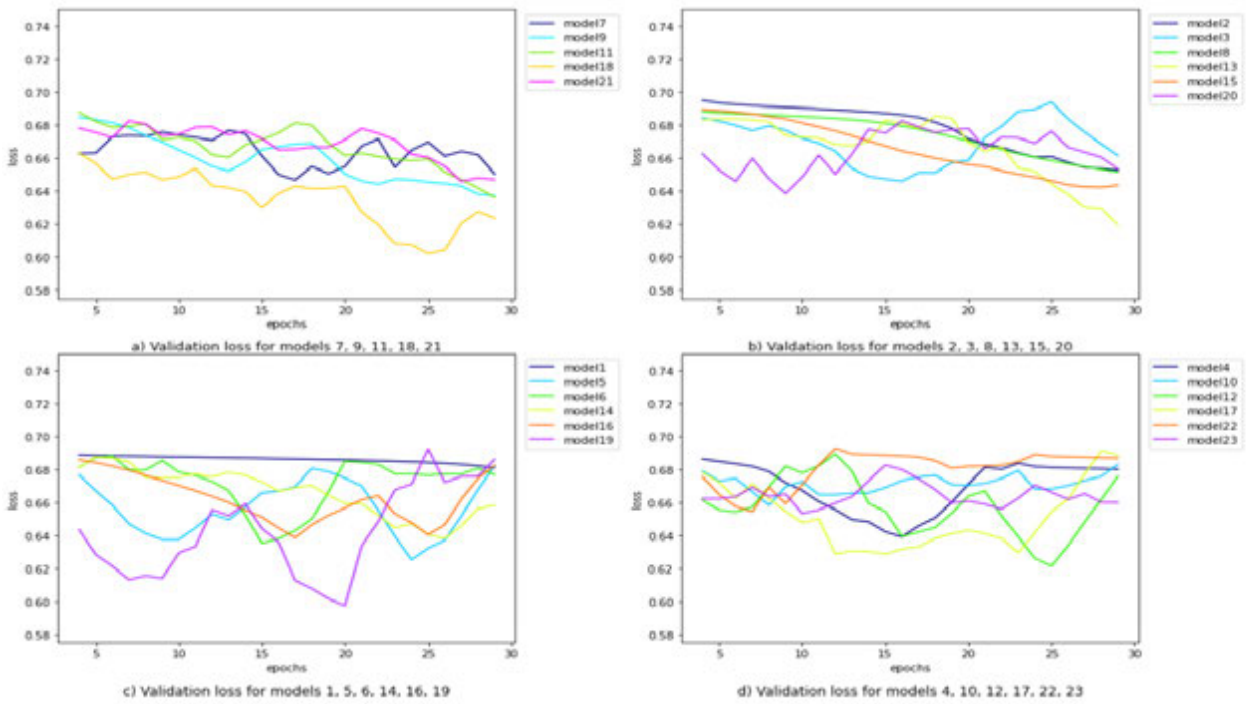


Fig. 7. 5-SMA of validation loss of different models for T/F prediction

V. CONCLUSION

Comparing the binary models with multiclass models, in both cases, simpler models gave better F1 scores, whereas models with a higher number of neurons had a higher tendency to overfit.

It is in line with our expectations that models with more layers and an increasing number of neurons per layer outperform models with a decreasing number of neurons per layer. However, model 22 did not behave according to that expectations, and such behavior can be a topic for future research.

The major issue we have detected was the imbalance of the dataset. A larger and more balanced dataset would lead to better results. However, the annotation of such a dataset would incur a significant cost as MBTI annotation requires highly skilled annotators, unlike sentiment analysis or similar tasks. Our results show that a more balanced dataset would lead to better results, as it was also emphasized by [6]. This was showcased by the T/F binary classifier, which operated on a well-balanced subset and achieved good accuracy results.

An alternative approach to the dataset imbalance problem would be to use class weights, as suggested in [7], [13]. We leave these ideas as topics for future research.

REFERENCES

[1] Y. LeCun, Y. Bengio and G. Hinton, "Deep Learning" in Nature, vol. 521., pp. 436-444, 2015.

[2] B. R. Strickland, "The Gale Encyclopedia of Psychology", 2nd edition, Gale Group, 2001

[3] D. Liu, Y. Li and M. A. Thomas, "A Roadmap for Natural Language Processing Research in Information Systems", Hawaii International Conference on System Sciences, 2017.

[4] A. Shertinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network" in the Elsevier journal "Physica D: Nonlinear Phenomena", vol. 404, 2020.

[5] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting" in Journal of Machine Learning Research, vol. 15, pp. 1929-1958, 2014.

[6] Z. Zhou and X. Liu, "Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem", IEEE Transactions on Knowledge and Data Engineering, vol. 18, pp. 63-77, 2006.

[7] A. U. Rehman, A. K. Malik, B. Raza and W. Ali, "A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis", in Multimedia Tools and Applications, 2019.

[8] Y. Bengio, "Practical Recommendations for Gradient-Based Training of Deep Architectures", version 2., 2012.

[9] J. Pennington, R. Socher and C. D. Manning, "GloVe: Global Vectors for Word Representation", 2014.

[10] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization", 2015.

[11] Sepp Hochreiter and J¨urgen Schmidhuber. 1997. Long Short-Term Memory.Neural Computation, 9(8):1735–1780, November

[12] R. Hernandez and I. S. Knight, 'Predicting Myers-Briggs Type Indicator with Text Classification', https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6839354.pdf

[13] M. Ren, W. Zeng, B. Yang, R. Urtasun, "Learning to Reweight Examples for Robust Deep Learning,". https://arxiv.org/pdf/1803.09050.pdf

[14] A. Moreno, A. Esuli, F. Sebastiani, "Word-Class Cmbeddings for Multiclass Text Classifictaion", https://arxiv.org/pdf/1911.11506.pdf

On Checking Controllability of Specification Languages for DES

Artem Davydov, Aleksandr Larionov, and Nadezhda Nagul
Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences
134 Lermontov str., 664033 Irkutsk, Russia
E-mails: artem@icc.ru, bootfrost@zoho.com, sapling@icc.ru

***Abstract*—The paper provides further development of the authors' original approach to the representation and properties checking of discrete event systems. The approach suggested is based on the automated inference in the calculus of positively constructed formulas (PCF). Discrete event system is supposed to be modeled in the form of finite automata within the framework of Ramage–Wonham supervisory control theory. It is shown how constructive inference helps to build the product of two finite automata which results in the automaton with accessible states only. Based on the nonmonotonic logical inference of PCF, a new method is presented for checking the controllability of formal languages describing specifications on the functioning of discrete event systems.**

***Index Terms*—discrete event system, supervisory control, controllability, first-order logic, automatic theorem proving**

I. INTRODUCTION

Being a part of the series of works aimed to develop a new approach to deal with controlled discrete event systems (DES), this paper presents a new method for checking the controllability of specifications on DES functioning. The theory of supervisory control (SCT) of DES gains more and more industrial applications nowadays, and the literature on DES is immense. Recent trends in SCT are overviewed in [1] while for a comprehensive survey of SCT see [2], [3], [4]. Different approaches are employed to study DES and design supervisors, including algebraic and logical ones. However, there are still open problems in SCT which require new instruments and techniques. We suggest automatic theorem proving technique based on the original calculus of positively constructed formulas to study supervised DES.

This paper presents the further development of the approach based on the application of automated first-order logic theorem proving and software system (a prover) to the representation of DES and supervised DES, which are formalized in the form of finite state machines. Previously it was presented in [5], and observability and co-observability checking was shown in [6], [7]. The logic underlying the presented approach is based on the calculus of positively constructed formulas (PCFs). It was developed in [8], [9] and further developed as a complete method for automated theorem proving with functional symbols in [10] and [11]. Thanks to its features, it is well suited and has already shown itself well in solving the above problems. In this paper, the improved formalization from [12] is used to check the controllability of specification language, i.e., if imposed constraints on system behavior may be satisfied by means of a supervisor.

The rest of the paper is organized as follows. At first, necessary preliminaries on DES and SCT are provided. The third section contains a brief description of PCFs and PCF calculus. In Section 4, a product of automata is built with the help of PCF calculus. In Section 5, a method of the controllability checking is presented. Further work is reported in Conclusion.

II. SUPERVISORY CONTROL OF DES

Consider discrete-event system (DES) in the form of a generator of a formal language [13] $\mathcal{G} = (Q, \Sigma, \delta, q_0, Q_m)$, also called a *plant* in the automatic control theory. Here Q is the set of states q; Σ is the set of events; δ: $\Sigma \times Q \to Q$ is the transition function; $q_0 \in Q$ is the initial state; $Q_m \subset Q$ is the set of marked states. Denote the function $\Gamma : Q \to 2^{\Sigma}$ as $\Gamma(q) = \{e \in \Sigma : \delta(w, q) \text{ is defined } \}$. Set $\Gamma(q)$ is called the *active event set* of $\mathcal{G}$ at the state q. Let Σ^* denote the set of all strings over Σ, including the empty string ε. δ is easily extended on strings from Σ^*. Language generated by $\mathcal{G}$ is $L(\mathcal{G}) = \{w : w \in \Sigma^* \text{ and } \delta(w, q_0) \text{ is defined}\}$, while language marked by $\mathcal{G}$ is $L_m(\mathcal{G}) = \{w : w \in L(\mathcal{G}) \text{ and } \delta(w, q_0) \in Q_m\}$. For any $L \subset \Sigma^*$ a *closure* of L is the set of all strings that are prefixes of words of L, i.e. $\overline{L} = \{s | s \in \Sigma^* \text{ and } \exists t \in \Sigma^* : s \cdot t \in L\}$. Symbol $\cdot$ denotes string concatenation and is often omitted.

Suppose some events of $\mathcal{G}$ are controllable, i.e. may be prohibited from occurring. Let Σ_c be a controllable event set, $\Sigma_{uc} = \Sigma \setminus \Sigma_c$, $\Sigma_c \cap \Sigma_{uc} = \emptyset$. The Ramadge–Wonham supervisory control framework assumes the existence of a means of control $\mathcal{G}$ presented by a *supervisor* [13]. The supervisor switches control patterns in such a way that the supervised discrete event system achieves a control objective described by some regular language K. Denote $L(\mathcal{J}/\mathcal{G})$ a language generated by the closed-looped behavior of the plant and the supervisor. Let $L_m(\mathcal{J}/\mathcal{G})$ denotes the language marked by the supervisor: $L_m(\mathcal{J}/\mathcal{G}) = L(\mathcal{J}/\mathcal{G}) \cap L_m(\mathcal{G})$. The main goal of supervisory control is to construct such supervisor that $L(\mathcal{J}/\mathcal{G}) = \overline{K}$ and $L_m(\mathcal{J}/\mathcal{G}) = K$.

Usually a supervisor is designed as another automaton, and control actions are achieved by constructing the parallel composition of these automata. In the parallel composition of two automata, events in the intersection of the respective event

sets must occur simultaneously in both automata. If an event is not shared by the automata, it may occur independently. Thus automata are synchronized by common events.

The definition of controllability plays a crucial role in characterizing those languages that can be generated by the closed-loop structure plant–supervisor.

Definition 1. *[13] K is controllable (with respect to $L(\mathcal{G})$ and Σ_{uc}) if*

$$\overline{K}\Sigma_{uc} \cap L(\mathcal{G}) \subseteq \overline{K}.$$

If K represents the admissible behavior of the system, K is controllable if occurring of any uncontrolled event after a prefix of a word from K leads to a word from K, i.e., still admissible. Only controllable languages may be exactly achieved by the joint behavior of the plant and supervisor.

The standard algorithm for verifying controllability uses composition operation on automata called the product of automata. In the product automata only shared events are present and occur simultaneously. The formal definition is as follows.

Definition 2. *If two generators $\mathcal{G}_i = (Q^i, \Sigma_i, \delta_i, q_0^i, Q_m^i)$, $i = 1, 2$, are given then their* product *is defined as $\mathcal{G}_1 \times \mathcal{G}_2 := Ac(Q^1 \times Q^2, \Sigma_1 \cup \Sigma_2, \delta, (q_0^1, q_0^2), Q_m^1 \times Q_m^2)$ where*

$$\delta(\sigma, (q^1, q^2)) = \begin{cases} (\delta_1(\sigma, q^1), \delta_2(\sigma, q^2)), & \text{if } \Gamma_1(q^1) \cap \Gamma_2(q^2) \\ \text{undefined} & \text{otherwise.} \end{cases}$$

Here $Ac(\mathcal{G})$ denotes an operation of taking accessible part of the automaton that is deleting all states which are not reachable from q_0 by some string from $L(\mathcal{G})$ and transitions attached to these states. Using the same principle, above definition may be extended on any number of automata.

Example 1. *The automaton in Fig. 3 is the result of product of the automata in Fig. 1 and Fig. 2. Here an arrow denotes initial state while the double circles denote the marked states.*

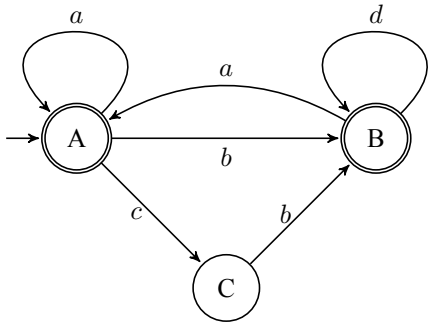


Fig. 1. Generator $\mathcal{G}$.

Let $\mathcal{H}$ be an automaton that generates K. To check controllability of K, build an automaton $\mathcal{G} \times \mathcal{H}$ and compare the active event set of each state of $\mathcal{G} \times \mathcal{H}$ with the active event set of the corresponding state of $\mathcal{G}$. K is not controllable if there exists an uncontrollable event in the latter that does not appear in the former. The worst-case computational complexity of this procedure is $O(E \cdot m \cdot n)$, where E is the number of events in Σ, m is the number of states of $\mathcal{H}$, and n is the number of states of $\mathcal{G}$ [3].

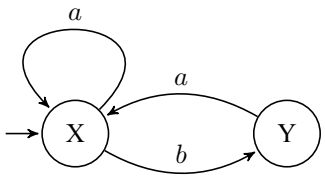


Fig. 2. $\mathcal{H}$ as recogniser for K

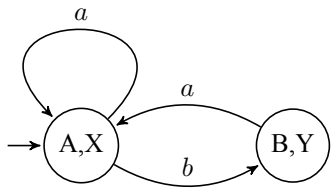


Fig. 3. Product $\mathcal{G} \times \mathcal{H}$

Example 2. *Suppose for $\mathcal{G}$ in Fig. 1 $\Sigma_c = \{b, c\}$ and $\Sigma_{uc} = \{a, d\}$. Then specification K depicted in Fig. 2 is not controllable since the uncontrollable event d cannot be disabled at the state B.*

III. THE CALCULUS OF POSITIVELY CONSTRUCTED FORMULAS

In the process of reasoning in the natural language, one often proves some statement by refuting the contradictory statement. We intend to proceed similarly. The main idea of the PCF calculus is to refute the negation of a formula $\mathcal{F}$ to prove that $\mathcal{F}$ is valid.

Let us consider a language of first-order logic that consists of first-order formulas (FOFs) built out of atomic formulas with $\&, \vee, \neg, \rightarrow, \leftrightarrow$ operators, $\forall$ and $\exists$ quantifier symbols and constants $true$ and $false$. The concepts of term, atom, literal we define in the usual way. Hereafter, non-atomic formulas and subformulas will be denoted by capital calligraphic letters ($\mathcal{F}, \mathcal{P}, \mathcal{Q}$, etc.), possibly with indices. Sets of formulas will be denoted by Greek capital letters (Φ, Ψ, etc.), possibly with indices.

Let $X = \{x_1, \ldots, x_k\}$ be a set of variables, $A = \{A_1, \ldots, A_m\}$ be a set of atomic formulas called *conjunct*, and $\Phi = \{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$ be a set of FOFs. The following formulas $\forall x_1 \ldots \forall x_k (A_1 \& \ldots \& A_m) \rightarrow (\mathcal{F}_1 \vee \ldots \vee \mathcal{F}_n)$ and $\exists x_1 \ldots \exists x_k (A_1 \& \ldots \& A_m) \& (\mathcal{F}_1 \& \ldots \& \mathcal{F}_n)$ are denoted as $\forall x_1, \ldots, x_k A_1, \ldots, A_m \{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$ and $\exists x_1, \ldots, x_k A_1, \ldots, A_m \{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$. They can be abbreviated as $\forall_X A\ \Phi$ and $\exists_X A\ \Phi$ respectively, keeping in mind that the $\forall$-quantifier corresponds to $\rightarrow \Phi^{\vee}$, where $\Phi^{\vee}$ means disjunction of all the formulas from Φ, and $\exists$-quantifier corresponds to $\& \Phi^{\&}$, where $\Phi^{\&}$ means conjunction of all the formulas from Φ. Any of sets X, A, Φ may be empty, and

in this case they could be omitted in formula formalization. Thus, if $Q \in \{\forall, \exists\}$ then $Q_X A \varnothing \equiv Q_X A$, $Q_X \varnothing \Phi \equiv Q_X \Phi$ and $Q_\varnothing A \Phi \equiv Q A \Phi$. Since empty disjunction is identical to $false$, whereas empty conjunction is identical to $true$, the following equivalences are correct: $\forall_X A \varnothing \equiv \forall_X A \rightarrow false \equiv \forall_X A$ and $\exists_X A \varnothing \equiv \exists_X A \& true \equiv \exists_X A$ and $\forall \varnothing \Phi \equiv true \rightarrow \Phi \equiv \forall \Phi$ and $\exists \varnothing \Phi \equiv true \& \Phi \equiv \exists \Phi$.

Definition 3. *Let X be a set of variables, and A be a conjunct, both can be empty.*

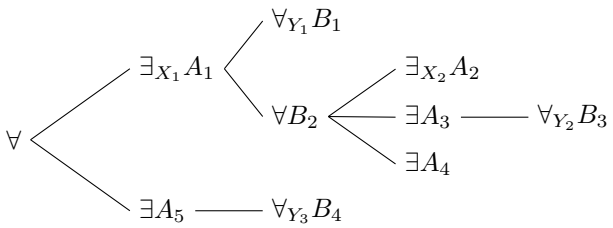
1) *$\exists_X A$ is an $\exists$-PCF,*
2) *$\forall_X A$ is a $\forall$-PCF,*
3) *If $\Phi = \{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$ is a set of $\forall$-PCFs, then $\exists_X A \Phi$ is an $\exists$-PCF,*
4) *If $\Phi = \{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$ is a set of $\exists$-PCFs, then $\forall_X A \Phi$ is a $\forall$-PCF,*
5) *Any $\exists$-PCF or $\forall$-PCF is a PCF,*
6) *There are only PCFs of a form $\exists$-PCF and $\forall$-PCF.*

The term "positively" comes from the fact that according to Definition 3 PCFs contain no negation operator $\neg$.

For the sake of readability we represent PCFs as trees whose nodes are type quantifiers, and we use corresponding notions: *node, root, leaf, branch*. For example, a PCF

$$\forall \{\exists_{X_1} A_1 \{\forall_{Y_1} B_1,\ \forall B_2 \{\exists_{X_2} A_2,\ \exists A_3 \{\forall_{Y_2} B_3\},\ \exists A_4\}\},\ \exists A_5 \{\forall_{Y_3} B_4\}\}$$

is represented as a tree as follows:



Given PCFs $\mathcal{P} = \forall\{\mathcal{F}_1, \ldots, \mathcal{F}_n\}$ and $\mathcal{F}_i = \exists_{X_i} B_i \{\mathcal{Q}_{i1}, \ldots, \mathcal{Q}_{im}\}$, $i = \overline{1,n}$, then $\mathcal{F}_i$ is called *base subformula* of $\mathcal{P}$, B_i is called *base of facts* or just *base*, $\mathcal{Q}_{ij}$ are called *question subformulas*, and roots of question subformulas are called *questions* to the base B_i, $i = \overline{1,n}$. A question of a form $\forall_X A$ (without any children) is called *goal question*.

Inside each of the base subformulas, any variable cannot be free and bound simultaneously. Furthermore, it cannot be bound by different quantifiers simultaneously.

Definition 4 (Answer)**.** *Consider some base subformula $\exists_X A \Psi$ of a PCF. A question of the subformula $\mathcal{Q} = \forall_Y B \Phi$, $Q \in \Psi$ has an answer θ if and only if θ is a substitution $Y \rightarrow H^\infty \cup X$ and $B\theta \subseteq A$, where H^∞ is Herbrand universe based on constant and function symbols that occur in the corresponding base subformula.*

Definition 5. *Let $\mathcal{P}_1 = \exists_X A \Psi$ and $\mathcal{P}_2 = \exists_Y B \Phi$, then $merge(\mathcal{P}_1, \mathcal{P}_2) = \exists_{X \cup Y} A \cup B \Psi \cup \Phi$*

Definition 6. *Consider some base subformula $\mathcal{B} = \exists_X A \Psi$. A question subformula $\mathcal{Q} \in \Psi$ has the form $\forall_Y D \{\mathcal{P}_1, \ldots, \mathcal{P}_n\}$, where $\mathcal{P}_i = \exists_{Z_i} C_i \ \Gamma_i, i = \overline{1,n}$, then $split(B, Q) = \{merge(B, \mathcal{P}'_1), \ldots, merge(B, \mathcal{P}'_n)\}$, where $'$ is a variable renaming operator. We say that $\mathcal{B}$ is split by $\mathcal{Q}$, and $split(\mathcal{B}, \mathcal{Q})$ is the result of the split of $\mathcal{B}$. Obviously, $split(\mathcal{B}, \forall_Y D) = \varnothing$.*

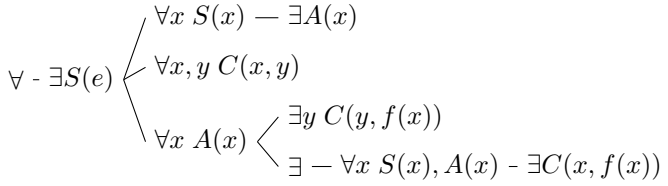
Definition 7 (The inference rule ω)**.** *Consider some PCF $\mathcal{F} = \forall \Phi$. If there exists a base subformula $\mathcal{B} = \exists_X A \Psi$, $\mathcal{B} \in \Phi$ and there exists a question subformula $\mathcal{Q} \in \Psi$, and the question of $\mathcal{Q}$ has an answer θ to $\mathcal{B}$, then $\omega(\mathcal{F}) = \forall \Phi \setminus \{\mathcal{B}\} \cup split(\mathcal{B}, \mathcal{Q}\theta)$.*

Note, that when the set Φ becomes empty after applying an ω rule, and the PCF becomes just $\forall$, then it can be concluded that the negation of an original formula is unsatisfiable.

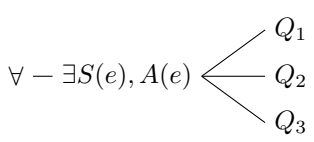
Any finite sequence of PCFs $\mathcal{F}, \omega\mathcal{F}, \omega^2\mathcal{F}, \ldots, \omega^n\mathcal{F}$, where $\omega^s\mathcal{F} = \omega(\omega^{s-1}\mathcal{F}), \omega^1 = \omega, \omega^n\mathcal{F} = \forall$, is called *an inference* of $\mathcal{F}$ in PCF calculus (with the axiom $\forall$).

Suppose that a search strategy does not use repeated application of ω to a question with the same θ (question-answering method of automated inference).

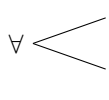
Example 3 (A refutation in PCFs)**.** *Consider the PCF $\mathcal{F}_1$*



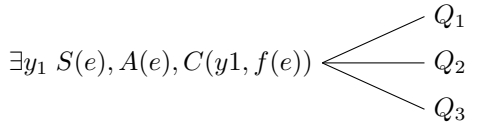
At the first step there is only one answer $\{x \rightarrow e\}$ to the first question (top to bottom numbering, name it Q_1). After applying rule ω with this answer to the only base of $\mathcal{F}_1$, the formula will be converted to the following form, $\mathcal{F}_2$:



At the second step there is also only one answer $\{x \rightarrow e\}$ to question Q_3. After applying ω with this answer, formula is split, because Q_3 has disjunctive branching. The formula will have the following form, $\mathcal{F}_3$:



where F_3^1 is:



and F_3^2 is:

$$\exists y_1\ S(e), A(e) \begin{cases} Q_1 \\ Q_2 \\ Q_3 \\ \forall x\ S(x), A(x) \text{ —— } \exists C(x, f(x)) \end{cases}$$

At the third step the first base can be refuted by answering on Q_2 (the goal question) with $\{x \to y_1; y \to f(e)\}$. The refuted base (and its whole base subformula) is to be deleted from the list of the base subformulas.

At the fourth step there is the answer $\{x \to e, y \to e\}$ to the fourth new question of the base F_3^2 which adds the atom $C(e, f(e))$ to it.

At the fifth step, the only base can be refuted by answering on Q_2 (the goal question) with the answer $\{x \to e; y \to f(e)\}$, finishing refutation because all the bases were refuted.

One of the important features of the calculus of PCFs which will be used later is that we can build *a nonmonotonic* inference by only slightly adjusting the definition of the inference rule. For this, we introduce the operator *, which will mark the atoms in the questions. Now, if a question with atoms marked with the * operator has an answer, then after applying the inference rule, the atoms in the base that participated in the matching search with the marked atoms should be removed from the base. In general, the operator * affects the property of completeness of the PCF calculus, but for the problem considered in this paper, thanks to a proper formalization, the inference always ends.

IV. CONSTRUCTING PRODUCT OF AUTOMATA USING PCF CALCULUS

To formalize a finite state machine as a generator of a formal language, the following predicates will be used. Let $L(s, S)$ denote "s is a current sequence of events in the state S" and $L_m(s, S)$ denote "s is a current sequence of events in the state S, and s is a marked string". The first arguments of these atoms will accumulate the strings of languages generated and marked by the automaton. Predicate of the form $\delta(S_1, \sigma, S_2)$ will be interpreted as the automaton transition from a state S_1 to a state S_2 with an event σ. If the target state of a transition is marked, then delta atoms with an index m are used, i.e., $\delta_m(S_1, \sigma, S_2)$ if S_2 is a marked state. The predicate $I(_)$ denotes the initial state of the automaton. Controlled and uncontrolled events will be represented in the base by separate atoms using the predicates $\Sigma_c(_)$ and $\Sigma_{uc}(_)$, respectively. As usual, the function symbol "$\cdot$" denotes strings concatenation, and the "ε" symbol corresponds to the empty string.

The general form of PCF representing some automaton consists of the single base $\mathcal{B} = \{I(S), L(\varepsilon, S), L_m(\varepsilon, S), \delta(S_1^i, \sigma^i, S_2^i), \delta_m(S_1^i, \sigma^i, S_2^i), \Sigma_c(\sigma^j), \Sigma_{uc}(\sigma^j)\}$, $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, m\}$, n is the number of transitions, m is the number of events, and two questions shown on Fig. 4. Applying the inference rules to this PCF, the strings of the languages generated and marked by the automaton will be built as the first arguments of the atoms $L(s, S)$, $L_m(s, S)$ in the base.

$$\exists\ \mathcal{B} \begin{cases} \forall \sigma, s, \sigma', s'\ L(\sigma, S), \delta(s, \sigma', s') - \exists L(\sigma \cdot \sigma', s') \\ \forall \sigma, s, \sigma', s'\ L(\sigma, S), \delta_m(s, \sigma', s') \cdot \exists L_m(\sigma \cdot \sigma', s') \end{cases}$$

Fig. 4. General form of PCF representation of some automaton

Suppose that behavior of $\mathcal{G}$ should be constrained within the specification language K, and let automaton $\mathcal{H}$ generates, or recognizes, K. In order to check the controllability of K, we first construct a product of automata $\mathcal{G}$ and $\mathcal{H}$, on the basis of which further inference will be constructed. The PCF representation of constructing the product of two automata may also be written using a single formula where atoms corresponding to different automata are distinguished by indices. The inferences will be based on the reasoning about the structures of automata which are involved in the product. The PCF $\mathcal{F}_{\mathcal{G}\times\mathcal{H}}$ constructing the product of automata will consist of one base subformula, which base conjunct is $B_{\mathcal{G}\times\mathcal{H}} = \{I_1(S_0), I_2(P_0), \delta_1(S_1^i, \sigma^i, S_2^i), \delta_2(S_1^k, \sigma^k, S_2^k)\}$, containing atoms for transitions δ_1, $i = \overline{1, n_1}$, of the first automaton and transitions δ_2, $k = \overline{1, n_2}$, of the second one. The questions of $\mathcal{F}_{\mathcal{G}\times\mathcal{H}}$ are as follows:

$$1 : \forall s, p\ I_1(s), I_2(p) - \exists I_3(s \cdot p)$$
$$2 : \forall s, p\ I_1(s), I_2(p) - \exists \delta_3(\varepsilon, \varepsilon, s \cdot p)$$
$$3 : \forall \sigma, s_1, p_1, s_2, p_2, \sigma', s'\ \delta_1(s_1, \sigma, s_2), \delta_2(p_1, \sigma, p_2), \delta_3(s', \sigma', s_1 \cdot p_1) - \exists \delta_3(s_1 \cdot p_1, \sigma, s_2 \cdot p_2)$$

Index 1 in the subscripts of atoms of this formula corresponds to the atoms of the automaton $\mathcal{G}$, 2 to the $\mathcal{H}$, and 3 to the atoms corresponding to the product of $\mathcal{G}$ and $\mathcal{H}$. Here, the functional symbol $\cdot$ is used to trace the pairs of states in the product automaton. The first question adds the initial state of the product automaton to the base. The second question adds an undetermined transition atom corresponding to the third automaton, thanks to which the connectivity of the product automaton is controlled in the third question. The latter adds the transition atoms of the product automaton to the base.

Example 4. *The steps of the inference that constructs the automaton depicted in Fig. 3 is shown in Table I. It is assumed that the base contains atoms corresponding to transitions of the automata depicted in Fig. 1 and Fig. 2.*

V. CHECKING CONTROLLABILITY USING THE PCF CALCULUS

According to the paradigm of the PCF calculus, we will prove the negation of the property that must be verified, i.e., uncontrollability of K. Note that the language expression for the controllability condition can be written as follows: $\forall s \in \overline{K}, \forall \sigma \in \Sigma_{uc}$, if $s\sigma \in L(\mathcal{G})$ then $s\sigma \in \overline{K}$. It remains for us to

TABLE I
Inference of $\mathcal{F}_{\mathcal{G}\times\mathcal{H}}$ constructing $\mathcal{G}\times\mathcal{H}$

Q	Base atoms	Substitution	Atoms added
1	$I_1(A), I_2(X)$	$\{s \to A,\, p \to X\}$	$I(AX)$
2	$I_1(A), I_2(X)$	$\{s \to A, p \to X\}$	$\delta^3(\varepsilon,\varepsilon,AX)$
3	$\delta^3(\varepsilon,\varepsilon,AX)$, $\delta^1(A,a,A)$, $\delta^2(X,a,X)$	$\{s_1 \to A, p_1 \to X,$ $\sigma \to b, s_2 \to A, p_2 \to X\}$	$\delta^3(AX,a,AX)$
3	$\delta^1(A,b,B)$, $\delta^2(X,b,Y)$	$\{s_1 \to A, p_1 \to X,$ $\sigma \to b, s_2 \to B, p_2 \to Y\}$	$\delta^3(AX,b,BY)$
3	$\delta^3(AX,b,BY)$, $\delta^1(B,a,A)$, $\delta^2(Y,a,X)$	$\{s_1 \to B, p_1 \to Y,$ $\sigma \to a, s_2 \to A,$ $p_2 \to X\}$	$\delta^3(BY,a,AX)$

prove whether occurring of an uncontrolled event results in a string that belongs to the language $L(\mathcal{G})$ but is not a prefix of a word from K.

Since the product of automata representing the plant and the specification generates only strings belonging to both $L(\mathcal{G})$ and K, to check if K is not controllable we employ the PCF $\mathcal{F}_{\mathcal{G}\times\mathcal{H}}$ with additional rules $C1$–$C6$ below. Auxiliary predicates will be used, meaning of which will be explained on their appearance.

$C1$. In the subsequent rules, states which are simultaneously achieved from the states of $\mathcal{G}$ and $\mathcal{G}\times\mathcal{H}$ by the same events will be called the *neighbouring* states and stored as arguments of the predicate $N(_,_)$. The first rule adds to the base the initial states of $\mathcal{G}$ and $\mathcal{G}\times\mathcal{H}$ as a first pair to be checked for uncontrollable transitions:

$$\forall s_1, s_2, p_1, p_2, \sigma\ I_1(s_1), I_3(p_1), \delta_1(s_1,\sigma,s_2), \delta_3(p_1,\sigma,p_2) - \\ - \exists\ N(s_1,p_1).$$

Although such s_1, p_1 are not neighboring states in our sense, $N(s_1,p_1)$ is necessary for further inference.

$C2$. This rule checks if there is an uncontrollable transition from some state of the automaton corresponding to the plant:

$$\forall s_1, s_2, p_1, \sigma\ N(s_1,p_1), \delta_1(s_1,\sigma,s_2), E_{uc}(\sigma) \\ - \exists\ Chk(p_1,\sigma,0).$$

Upon successful answer to this question, the atom $Chk(p_1,\sigma,0)$ is added to the base. The first argument of this atom denotes a state p_1 of $\mathcal{H}$, in which the transition on the uncontrollable event σ should exist for K to be controllable. By default we suppose that the transition does not exist so the third argument in $Chk()$ is 0 before checking by the next rule.

$C3$. This rule is aimed to check if both neighboring states share the same uncontrollable event:

$$\forall p_1, p_2, \sigma\ Chk^*(p_1,\sigma,0), \delta_3(p_1,\sigma,p_2) \\ - \exists\ Chk(p_1,\sigma,1).$$

A successful answer to this question means that the uncontrollable event is legal, i.e., allowed by the specification, since a transition with it exists in both $\mathcal{G}$ and $\mathcal{H}$.

We need to delete the atom $Chk(p_1,\sigma,0)$ with the help of * operator from the base, to ensure that there will not be additional answers to this question with the same pair of states.

$C4$. This rule checks if the atom $Chk(p_1,\sigma,0)$ is present in the base:

$$\forall p_1, \sigma\ Chk(p_1,\sigma,0) - \exists\ UC(p_1,\sigma).$$

The presence of $Chk(p_1,\sigma,0)$ in the base means that an uncontrollable event is not allowed by the specification, i.e., K is not controllable. So the atom $UC(p_1,\sigma)$ is added to the base upon a successful answer. It contains information about the state and the event that violate the controllability condition. This atom is further used in the goal question to complete the inference.

It should be noted here that the inference search strategy must be configured in such a way that this question is checked for an answer immediately after the previous one.

$C5$. This is the goal question that ends the inference:

$$\forall p_1, \sigma\ UC(p_1,\sigma).$$

$C6$. This rule adds the next checked pair of states to continue the inference search:

$$\forall s_1, s_2, p_1, p_2, \sigma\ N(s_1,p_1), \delta_1(s_1,\sigma,s_2), \delta_3(p_1,\sigma,p_2) - \\ - \exists\ N(s_2,p_2).$$

A pair of states s_2, p_2 which are simultaneously achieved from the previous checked states of $\mathcal{G}$ and $\mathcal{G}\times\mathcal{H}$ by the same event is added to the base.

If the inference has ended with the exhaustion of the search options for substitutions and the answer to the goal question has not been found then the specification language is controllable. This fact requires explanation. We formalized the problem in such a way that we want to find a set of variables violating the controllability of the specification. If the specification is actually not controllable then successful completion of the inference provides this set within an answer $UC(p_1,\sigma)$ to the question $C5$. Since the formalized automata are finite with finite sets of events, then the search space for the inference in this formalization is finite. Therefore, the inference ends in both cases – either the specification is controllable or not.

Example 5. *Table II shows the inference of the PCF, which establishes the uncontrollability of the specification represented by the automaton shown in Fig. 2 with respect to the plant shown in Fig. 1. It is assumed that the inference of the product automaton (Fig. 3) is already built, and the base from which the presented inference starts includes the next subset of atoms:*

$$\{I_1(A), I_3(A\cdot X), \Sigma_{uc}(a), \Sigma_{uc}(d), \\ \delta_1(A,a,A), \delta_1(A,b,B), \delta_1(B,a,A), \delta_1(B,d,B), \\ \delta_3(A\cdot X,a,A\cdot X), \delta_3(A\cdot X,b,B\cdot y), \delta_3(B\cdot Y,a,A\cdot X)\}$$

TABLE II
Inference for the controllability checking

Q	Base atoms	Substitution	Atoms added
C1	$I_1(A), I_3(A \cdot X)$, $\delta_1(A, a, A)$, $\delta_3(A \cdot X, a, A \cdot X)$	$\{s_1 \to A, s_2 \to A$, $p_1 \to A \cdot X, p_2 \to$ $A \cdot X, \sigma \to a\}$	$N(A, A \cdot X)$
C2	$N(A, A \cdot X)$, $\delta_1(A, a, A)$ $\Sigma_{uc}(a)$	$\{s_1 \to A, s_2 \to A$, $p_1 \to A \cdot X, \sigma \to$ $a\}$	$Chk(A \cdot X, a, 0)$
C3	$Chk(A \cdot X, a, 0)$ $(deleted)$, $\delta_3(A \cdot X, a, A \cdot X)$	$\{p_1 \to A \cdot X$, $p_2 \to A \cdot X$, $\sigma \to a\}$	$Chk(A \cdot X, a, 1)$
C6	$N(A, A \cdot X)$, $\delta_1(A, b, B)$, $\delta_3(A \cdot X, b, B \cdot Y)$	$\{s_1 \to A, s_2 \to B$, $p_1 \to A \cdot X, p_2 \to$ $B \cdot Y, \sigma \to b\}$	$N(B, B \cdot Y)$
C2	$N(B, B \cdot Y)$, $\delta_1(B, a, A)$ $\Sigma_{uc}(a)$	$\{s_1 \to B, s_2 \to A$, $p_1 \to B \cdot Y, \sigma \to a\}$	$Chk(B \cdot Y, a, 0)$
C3	$Chk(B \cdot Y, a, 0)$ $(deleted)$, $\delta_3(B \cdot Y, a, A \cdot X)$	$\{p_1 \to B \cdot Y$, $p_2 \to A \cdot X$, $\sigma \to a\}$	$Chk(B \cdot Y, a, 1)$
C2	$N(B, B \cdot Y)$, $\delta_1(B, d, B)$ $\Sigma_{uc}(d)$	$\{s_1 \to B, s_2 \to B$, $p_1 \to B \cdot Y, \sigma \to d\}$	$Chk(B \cdot Y, d, 0)$
C4	$Chk(B \cdot Y, d, 0)$	$\{p_1 \to B \cdot Y$, $\sigma \to d\}$	$UC(B \cdot Y, d)$
C5	$UC(B \cdot Y, d)$	$\{p_1 \to B \cdot Y$, $\sigma \to d\}$	$\varnothing$

VI. CONCLUSION

We presented a new method for checking the controllability of formal languages describing specifications on the functioning of DES modeled in the form of finite automata. Though this approach may seem sophisticated compared to traditional methods to deal with automata-based DES, its advantages will be apparent in solving such problems as designing modular supervisors and building abstractions of automata. To choose a proper order of consideration of a plant and specification modules or proper alphabet for building abstraction of automata, additional knowledge may be used. This knowledge may be written in a PCF base. One of the further directions of research on this work‘s subject is the study of the possibility of automatically constructing a controllable sublanguage of a specification if it turned out to be uncontrollable. Indeed, the advantage of the method for checking the controllability described above is the non-monotonicity of the inference mechanism used and the fact that upon completion of the inference, we have information about the state and event that interfere with controllability. At the same time, we can test the idea of automatic reduction of the automaton during successive launches of the inference of the controllability checker until we get a controlled subset of transitions.

The prover used to solve the examples is written in the Rust programming language. It is optimized for processing PCF without unconfined variables (variables that are controlled by the quantifier but are absent in its typical condition). It implements strategies for the sequential search for answers to questions, processing formulas with disjunctive branching, state rollback possibility, memory saving strategy in form of sharing same data structures representing the PCF base and questions, and some strategies for processing the equality predicate.

The procedure for finding answers to questions finds in one step at once all possible answers to all questions of the formula. Found answers are added to the main answers pool along with information at what step they were found. Next, the answer selection procedure selects sequentially answer from this pool. The answer selection procedure prohibits the reuse of identical answers, as well as answers that lead to the addition of only existing facts to the PCF base. It is possible to customize the order of answer selection. A detailed description of the prover is omitted due to the lack of space and will be presented in some future paper.

ACKNOWLEDGMENT

The research is supported by the Russian Science Foundation (project no. 16-11-00053).

References

[1] S. Lafortune, “Discrete event systems: Modeling, observation, and control,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, no. 1, pp. 141–159, 2019. [Online]. Available: https://doi.org/10.1146/annurev-control-053018-023659

[2] W. M. Wonham and K. Cai, *Supervisory Control of Discrete-Event Systems*. Springer International Publishing, 2019.

[3] C. G. Cassandras and S. Lafortune, *Introduction to Discrete Event Systems*. Springer US, 2008.

[4] C. Seatzu, M. Silva, and J. H. van Schuppen, Eds., *Control of discrete-event systems*. Springer London, 2013.

[5] A. Davydov, A. Larionov, and N. Nagul, “The formal description of discrete-event systems using positively constructed formulas,” in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2017.

[6] A. Davydov, A. Larionov, and N. Nagul, “The formal logic approach for checking the observability of a specification language on des functioning,” in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2018, pp. 0938–0943.

[7] A. Davydov, A. Larionov, and N. Nagul, “A first-order logic based approach to problems of decentralized supervisory control of des,” *AIP Conference Proceedings*, vol. 2046, no. 1, p. 020021, 2018. [Online]. Available: https://aip.scitation.org/doi/abs/10.1063/1.5081541

[8] S. N. Vassilyev, “Machine synthesis of mathematical theorems,” *The Journal of Logic Programming*, vol. 9, no. 2-3, pp. 235–266, 1990.

[9] A. K. Zherlov, S. N. Vassilyev, E. A. Fedosov, and B. E. Fedunov, *Intelligent control of dynamic systems*. Moscow: Fizmatlit, 2000, in Russian.

[10] A. Davydov, A. Larionov, and E. Cherkashin, “On the calculus of positively constructed formulas for automated theorem proving,” *Automatic Control and Computer Sciences*, vol. 45, no. 7, pp. 402–407, 2011.

[11] A. Larionov, A. Davydov, and E. Cherkashin, “The calculus of positively constructed formulas, its features, strategies and implementation,” *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija*, vol. 2013, pp. 1023–1028, 2013.

[12] A. Davydov, A. Larionov, and N. V. Nagul, “Pcf-based formalization of the parallel composition of automata,” in *Proceedings of the 1st International Workshop on Information, Computation, and Control Systems for Distributed Environments, ICCS-DE 2019, Irkutsk, Russia, July 8-9, 2019*, ser. CEUR Workshop Proceedings, I. Bychkov and A. Tchernykh, Eds., vol. 2430. CEUR-WS.org, 2019, pp. 29–41. [Online]. Available: http://ceur-ws.org/Vol-2430/paper3.pdf

[13] P. J. Ramadge and W. M. Wonham, “Supervisory control of a class of discrete event processes,” *SIAM Journal on Control and Optimization*, vol. 25, no. 1, pp. 206–230, 1987.

A Note on Geometric Calibration of Multiple Cameras and Projectors

Tomislav Petković*, Simone Gasparini**, and Tomislav Pribanić*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, Zagreb, Croatia
** University of Toulouse, Toulouse INP – IRIT, Toulouse, France
tomislav.petkovic.jr@fer.hr

Geometric calibration of cameras and projectors is an essential step that must be performed before any imaging system can be used. There are many well-known geometric calibration methods for calibrating systems comprised of multiple cameras, but simultaneous geometric calibration of multiple projectors and cameras has received less attention. This leaves unresolved several practical issues which must be considered to achieve the simplicity of use required for real world applications. In this work we discuss several important components of a real-world geometric calibration procedure used in our laboratory to calibrate surface imaging systems comprised of many projectors and cameras. We specifically discuss the design of the calibration object and the image processing pipeline used to analyze it in the acquired images. We also provide quantitative calibration results in the form of reprojection errors and compare them to the classic approaches such as Zhang's calibration method.

Keywords; geometric calibration; camera calibration; projector calibration; system of multiple cameras and projectors; reprojection error

I. INTRODUCTION

Both geometric calibration and photometric calibration of cameras and projectors have many applications including remote sensing [1], photogrammetry [2,3], 3D surface imaging and profilometry [4,5], motion capture [6,7,8], tiled displays [9,10,11], and many others. The task of geometric calibration [12,13,14] is to determine intrinsic and extrinsic parameters [12,13] of cameras and projectors, while the task of photometric calibration [15] is to establish an exact relation between measured values and physical units. In geometric calibration intrinsic parameters describe the internal properties of a camera or a projector, while extrinsic parameters describe their spatial position w.r.t. some reference coordinate system and are usually expressed in terms of a 3D rotation matrix and a translation vector. Geometric calibration is therefore an essential step which must be performed before any imaging system can be used as a measurement tool. It should be as simple as possible while retaining required accuracy in estimating imaging system parameters, especially when a system of multiple cameras and projectors is considered (Fig. 1).

The usual approach to geometric calibration is to record a calibration object such as the ones shown in Fig. 2 [12,13]. Relevant parameters of the imaging system are then estimated from the data by minimizing the reprojection error between the acquired images and the expected appearance of the calibration object obtained through an imaging model.



Figure 1. A 3D surface scanning system comprised of three projectors and six cameras



Figure 2. Various calibration objects: calibration wands (left), calibration boards (center), and a 3-axis spatial calibration object (right)

Differences between various proposed calibration methods stem from the fact that many different calibration objects are proposed. One amongst many possible classifications [12] is based on the degrees of freedom of the calibration object: (a) one dimensional objects such as calibration wands (Fig. 2 left) are most often used for simultaneous calibration of multiple cameras [6,8,16,17]; (b) two dimensional objects such as planar calibration boards (Fig. 2 center) are currently considered the state of the art in calibrating single cameras and stereo pairs [14,18,19,20]; and (c) three dimensional calibration objects such as calibration cages are mostly limited to laboratories for precise calibrations [8].

The most significant limitation of existing calibration techniques is that the calibration object must be completely visible for it to be a valid input for a calibration program; examples include standard implementations available in OpenCV [19] or in Matlab [20] where the chessboard pattern is rejected if partially visible. This presents a significant limitation for the calibration of a system com-

prised of many cameras and projectors (Fig. 1) as it is difficult or even impossible to position the calibration object so that it is entirely visible in all the views.

In this work we describe a calibration procedure routinely used in our laboratory [22] to calibrate an imaging system comprised of many cameras and projectors. We have introduced a first implementation of this method in [23], on which we here expand by describing the used image processing chain and by providing quantitative results. The proposed method uses a double-sided planar calibration board. A calibration pattern on both sides is comprised of white circles on dark background arranged in a hexagonal lattice and includes additional marking to easily identify which side of the board is observed [23]. The proposed image processing pipeline is able to extract reliable calibration points even if the calibration board is partially visible which presents a significant improvement compared to standard calibration procedures which require that the whole calibration object is visible [19, 20], and which discard partial views.

This work is organized as follows: in Section II. we give a detailed overview of existing camera and projector calibration methods. The proposed calibration method is presented in section III. The results and discussion are presented in section IV. We conclude in Section V.

II. BACKGROUND

When considering geometric calibration of multiple cameras and projectors one must note that a projector may be modelled as an inverse camera [24] and that existing cameras in the imaging system may be used to acquire images as seen by the projector [24,25]. Therefore, we will first give a brief overview of geometric calibration methods for cameras followed by methods for projectors.

A. Camera Calibration

A most commonly used imaging model is the pinhole camera model [12] which is almost always augmented by adding a lens distortion model [12,14,19], although sometimes other imaging models such as fisheye camera model [26,27] may be used. Regardless of the model which is used all camera calibration procedures share the same steps: (a) acquire many images of a calibration object; (b) process acquired images to extract the required data such as image coordinates of the calibration object or its characteristic points; and (c) minimize the reprojection error and find the optimal parameters of the imaging model through optimization. The first two steps determine the ease of use in real-world scenarios. In practice, most popular approaches include wand calibration [6,8,16,17] and calibration using a planar calibration board [14,28,29].

In wand calibration we use a wand containing several spherical markers made of reflective material (Fig. 2). An important property of spherical markers is that their center can be easily extracted by detecting circle centers in 2D images of each camera separately. This enables fast acquisition of calibration data simply by waving a calibration wand in a form of wand dance [8].

In planar calibration we use a planar calibration board containing a known calibration pattern (Fig. 2). A commonly used pattern is the checkerboard [19,20] from which corner points can be easily detected. Other patterns such as circles arranged in a regular lattice are also used [8,25]. All planar patterns can be easily printed on plain paper which makes them a method of choice for calibrating single cameras and stereo pairs. The main problem when using a planar calibration object for calibrating a system of multiple cameras is that it is visible from one side only, unlike a calibration wand which is visible from all sides simultaneously. This limitation may be improved upon by using a double-sided calibration board [23] where a rigid transformation between patterns on each side is known. This, however, comes at the cost of more complex and expensive manufacturing which may require an external contractor.

For completeness here we also mention auto- or self-calibration [13,31] which does not require a calibration object and which is often used in photogrammetry or in virtual environments.

B. Projector Calibration

A major obstacle in projector calibration is that the projector cannot acquire images. Therefore the cameras of the imaging system must be used to acquire the required data instead [24,25]. This is usually achieved by having a projector project some coded pattern which enables us to find a mapping of camera images into the coordinate system of projector's image.

Early approaches to projector calibration assumed a calibrated camera is used [32,33,34,35,36]. Then some code is projected onto planar surfaces to directly find a homography or to indirectly get the required data for calibration: fiducial markers are used in [32,33,36], chessboard pattern in [34], and a specially designed pattern in [35]. Such approaches require twice the effort and do not scale well when calibrating systems with many cameras and projectors.

Modern approaches [24,25,37,38,23] follow the proposal of Zhang and Huang [24] and employ ideas from structured light (SL) scanning [4]. A projector should project a code which directly embeds projector's row and column coordinates. Decoding the projected code then provides a direct mapping between camera's and projector's image coordinate systems. A significant advantage of directly measuring the mapping is that camera does not need to be calibrated in advance. Differences in listed approaches stem from the used code: in [24,25] a standard multiple phase shifting is used, [37] uses Gray code, and [38] a combination of them. Note that phase shifting is a better choice as it inherently provides subpixel precise mappings while Gray code requires interpolation to achieve subpixel precision. Another significant difference is the choice of the planar calibration pattern: some works use a standard checkerboard [24,37] while others use a regular lattice of bright circles on dark background [25,23] which makes the projected code easier to decode.

Other less common approaches to projector calibration include self-calibration [39] where multiple cameras are used to calibrate a single projector, and the use of a semi-transparent sheet [40] to make the code projected by the projector simultaneously visible by multiple cameras in an imaging system.

III. Calibration Procedure

We propose a calibration procedure for multi-projector multi-camera 3D scanner which is based on the following principles: (1) a calibration object must be usable even if it is only partially visible; (2) calibration points on the calibration object should be automatically detected and annotated; and (3) cameras will provide a means for projector to indirectly capture an image of the calibration board using a method proposed by Zhang and Huang [24]. We use a planar calibration object so that Zhang's planar method [14] can be used to obtain a quick initial calibration for each camera and projector. However, Zhang's method provides decoupled calibration parameters therefore the final calibration must include some form of simultaneous minimization of the reprojection errors for all cameras and projectors using bundle adjustment [21].

A. Calibration Pattern

We propose to use a calibration pattern comprised of bright circles on a dark background arranged in a regular hexagonal lattice (Fig. 3). The pattern may be used on a single-sided calibration board, but we chose a double-sided calibration board to speed up calibration and to reduce the number of acquisitions. Such board may be professionally manufactured, or it may be self-constructed by printing out the pattern and carefully gluing it to both sides. If printed and glued some additional method of measuring the transformation between sides must be used; for the board shown in Fig. 1 we have used MicroScribe G2LX [30] digitizing system to find the transformation.

To make the calibration simpler we also add marks in a form of dark rings which are used to identify the center and the side of the calibration board. To allow for any spatial orientation of the calibration board the markings are limited to an elemental hexagon and are selected in a way which makes all possible planar rotations of the board unique. Consider a hexagonal cell (Fig. 4) where the central marker **a** is surrounded by six neighbors which are arranged clockwise starting from **b**. If the orientation of the calibration board is arbitrary (we do not know it) then any of the surrounding elements **bcdefg** may take the starting role of **b** when identifying the cell in the acquired image. Therefore, the codes for identifying sides must be selected so all rotational permutations of the surrounding six elements **bcdefg** are mutually different thus making all used codes uniquely decodable. We have selected four codes which are shown in Fig. 4.

B. Recording the Calibration Pattern

The calibration board should be recorded in many positions with respect to the cameras; a minimum is two different positions for each camera and projector [14]. For each position of the calibration board then we have to acquire as many images as used SL code requires. We propose to use multiple phase shifting patterns which enable simultaneous acquisition in a system comprised of an arbitrary number of projectors [41]. For P projectors this requires $2P + 1$ phase shifted images per frequency and at least two frequencies are required. For each position of the calibration board from the acquired $2\cdot(2P + 1)$ images we then compute images as they would have been acquired if each projector was a camera. The image where all projectors are turned off is also computed for each camera. Another advantage of this approach is that if more than one camera observes the illuminated part of the board for some projector then the number of data points for calibrating that particular projector is increased, i.e. measurements are repeated up to the number of used cameras.

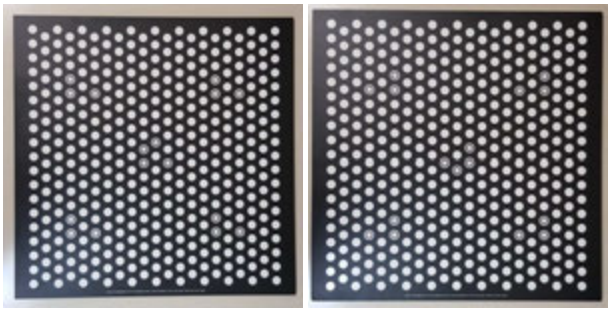


Figure 3. Two sides of a double sided planar calibration board (front and back). Note the markers in forms of inscribed black rings which identify the board side and the central point.

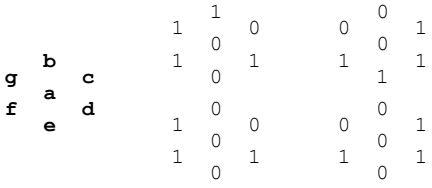


Figure 4. One hexagonal cell and corresponding values for markings denoting the central points on front and back side and sided (left column is front, right column is back). 1 indicates the ring is present and 0 it is not. Note that all planar rotations of four used codes are mutally distinguishable.

C. Pattern Processing

Once the images are acquired and the SL code is decoded using the approach from [41,42] we have to identify the circle centers and to construct a hexagonal lattice of calibration points. To make this fully automated we propose to use multi-scale analysis where on each scale the eigenvalues of the Hessian matrix are used to construct a *circleness* measure in a similar way as was proposed by Frangi et al. [43] for ridge detection by using a *vesselness* measure.

Let $I(x,y)$ denote the input image and let $G_\sigma(x,y)$ be a Gaussian kernel at scale σ. The Hessian matrix for each pixel (x,y) is then

$$H_\sigma(x,y) = \begin{bmatrix} I * \frac{\partial^2 G_\sigma}{\partial x^2} & I * \frac{\partial^2 G_\sigma}{\partial x \partial y} \\ I * \frac{\partial^2 G_\sigma}{\partial y \partial x} & I * \frac{\partial^2 G_\sigma}{\partial y^2} \end{bmatrix}, \quad (1)$$

where (x,y) dependencies are omitted from I and G_σ. For the Hessian matrix given by Eq. (1) we compute the eigenvalues $\lambda_{1,2}$ for each pixel (x,y) and order them so that

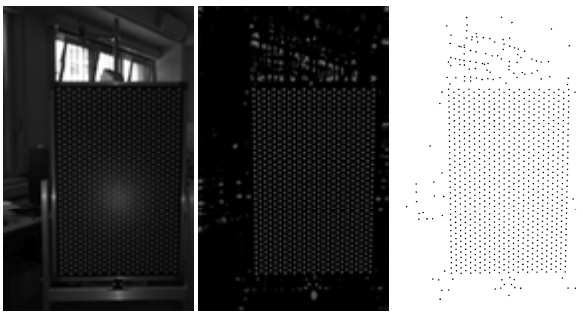


Figure 5. Detection of circles using circleness measure: input image (left), circleness (center), detected centers (right).

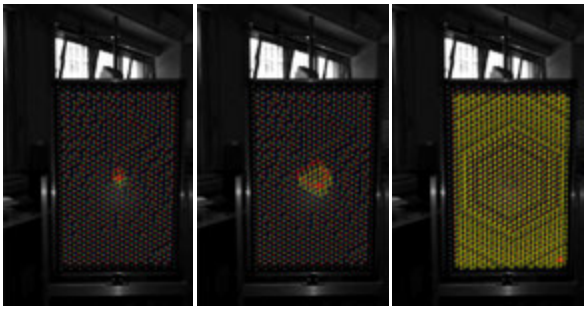


Figure 6. Grid construction: first cell (left), grid after 27 cells are processed (center), and final grid (right).

$$|\lambda_1| \geq |\lambda_2|. \tag{2}$$

The circleness measure to detect bright circles on dark background is then

$$C_\sigma(x,y) = \left(1 - \exp\left(-\alpha \frac{\lambda_2^2}{\lambda_1^2}\right)\right)\left(1 - \exp\left(-\beta(\lambda_1^2 + \lambda_2^2)\right)\right), \tag{3}$$

for $\lambda_{1,2} < 0$ and is zero otherwise. In Eq. (3) we have fixed the parameter α to $\frac{1}{4\log 2}$ and β is determined using the ridge model from [44] under the assumption of an ideal circle instead of an ideal ridge (see Eqs. (3) and (14) in [44]). Note that other choices of parameters are possible as explained in [43,44]. The circleness measure of Eq. (3) is by its definition limited to the [0,1] interval and is easily integrated over all scales using

$$C(x,y) = \max_{\text{over all scales } \sigma} C_\sigma(x,y), \tag{4}$$

In Eq. (4) scales are selected to encompass all expected circle sizes. The final result of this analysis is an image (Fig. 5 center) where circle's centers are local maxima. Note that non-maximum suppression must be used to suppress multiple detections, especially if the image exhibits strong perspective distortion: from the scale-space analysis radii for all detections are known so we delete all weaker circle centers having smaller vesselness value and falling within the radius of the locally strongest one (Fig. 5 right).

Once all centers are extracted a hexagonal grid must be constructed. We propose to do this in a two-pass procedure. In the first pass we construct all viable local elementary hexagons (see Fig. 4). In the second pass all constructed hexagons are assembled into a complete grid. This two-pass approach allows us to construct a usable grid even if the calibration board is only partially visible.

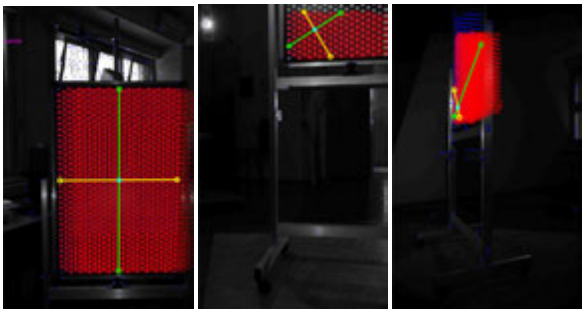


Figure 7. Constructed calibration grids: regular grid (left), and partial grids where markers are not visible or cannot be decoded (center and right) so coordinate systems are arbitrary oriented. Note the extreme perspective distortion in the rightmost example.

In the first pass a local hexagon is constructed around each detected circle center as follows: (1) find opposing pairs for the six closest neighboring circles; (2) for each opposing pairs compute the normalized distance to the central point; (3) select three pairs which have the shortest normalized distance. The key point of this pass is distance normalization: for each opposing pair the line connecting them should pass through the center of the central circle and should intersect the edge of that circle thus forming a chord, more specifically a diameter. The length of this chord changes depending on the amount of perspective distortion. If we divide the length between opposing points by the chord length, then distances are approximately normalized and may be compared to sort the points. Without this normalization the step (3) could produce a deformed local lattice where we select an opposing pair which is closest in the image space, but not in the real space. Once the hexagon cell is successfully constructed the embedded code is decoded using a simple heuristic inscribed ring detection procedure. Thus, we identify if a cell is a central point or if a cell is located on front or back side. Note that the first pass also acts as a natural filter: all found cells which cannot be matched to a hexagon are discarded.

In the second pass we start from the best local hexagonal cell. In order of desirability the starting cell is one containing the board's central point, then a cell indicated by repeated side marks, and finally any cell if no marks are visible (or if marks conflict). We then push all unprocessed neighboring cells whose centers are at **bcdefg** points (Fig. 4) into a FIFO queue. The constructed grid is expanded when a cell is popped from the queue, but only if the cell can be properly fitted to previously placed cells. The fitting includes planar rotation of each cell; for an example see Fig. 6 left and center where it is visible that local coordinate systems are not initially aligned (left and center images, direction vectors to **bcd** are shown in red, green and blue), but they become aligned after all cells are processed. This behavior is expected as the basic hexagonal element of the proposed grid is highly symmetrical making it impossible to determine the global orientation from local data (see Fig. 7).

Once both steps are complete, we have obtained corresponding coordinate pairs comprised of ideal planar coordinates on the calibration grid and of camera or projector coordinates (in subpixel accuracy). This pairs can then immediately be used in Zhang's calibration method [14]

TABLE I. REPROJECTION ERRORS FOR A SIX CAMERAS THREE PROJECTORS IMAGING SYSTEM. ALL VALUES ARE IN PX.

ID	Zhang's method	Bundle adjustment
CAM1	0.21 ± 0.30	2.87 ± 2.28
CAM2	0.21 ± 0.15	4.35 ± 2.25
CAM3	0.23 ± 0.25	1.18 ± 0.80
CAM4	0.23 ± 0.16	1.81 ± 1.08
CAM5	0.20 ± 0.16	2.74 ± 1.96
CAM6	0.21 ± 0.25	2.36 ± 1.53
PRJ1	0.22 ± 0.15	1.44 ± 0.97
PRJ2	0.22 ± 0.15	0.64 ± 0.50
PRJ3	0.26 ± 0.36	1.64 ± 1.27

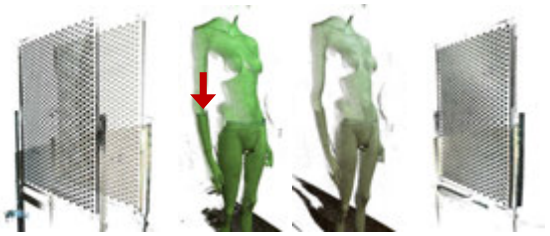


Figure 8. Point-cloud reconstructions: On the left are reconstructions using calibration data of Zhang's method when a single position of the board was selected as referenced. On the right are same reconstructions using calibration data from bundle adjustment.

to obtain an initial calibration, however note that Zhang's calibration method [14] is intended for calibration of a single camera and it cannot provide a reliable calibration (extrinsic parameters) for a multi-view system.

D. Calibration

After a pattern is processed all found corresponding point pairs are not necessarily expressed in the same coordinate system: (a) if the central marker is not visible then coordinate axes may be arbitrarily set (Fig. 7); and (b) if a side markers are not visible then differentiating between front and back side is impossible. To solve these issues, we propose to use projector data to find correspondences which then enable computation of 2D rigid registrations to align found points to the ideal grid. In other words, extracted projector coordinates are used as point descriptors and all data from partially visible boards is automatically registered to valid data for which both the center location and the side of the board are known. Sometimes even after this procedure there are leftover unaligned partial views: we leave the choice to the user to either discard them or to manually realign the grid coordinates. The data obtained in this way is finally fed into the bundle adjustment procedure to minimize the reprojection error

$$\mathbf{e} = \mathbf{x}_{px} - \mathbf{x}\left(X_{3D}, K_c, R_c, t_c, S_k, u_k\right), \qquad (5)$$

where K_c are intrinsic parameters (camera matrix and distortion parameters), R_c and t_c are extrinsic parameters (rotation and translation) of each camera and projector, and S_k and u_k (rotation and translation) are positions of the calibration board. If the system has N cameras and M projectors and if K positions of the calibration board are recorded then there are $6(N+M)$ intrinsic parameters K_c, while the total number of extrinsic parameters R_c, t_c and S_k, u_k is reduced to $6(N+M+K-1)$ one as one position must be taken as a reference.

IV. RESULTS AND DISCUSSION

The presented calibration procedure is used in our laboratory to routinely calibrate systems of up to six cameras and three projectors such as one shown in Fig. 1. Reprojection errors for one calibration of such system are shown in Table I: errors are in the order of pixels, which is sufficient for everyday use. In this example three positions of a double-sided calibration board were used which gives three planar grids for calibration, each containing 838 circles if fully visible. Note that if more than one camera views the board then the number of data points for a particular projector is increased as there are multiple simultaneous measurements. If a more precise calibration is required then more positions should be recorded. Note that calibration errors are smaller for Zhang's calibration method. This is due to the fact that Zhang's method allows more degrees of freedom than bundle adjustment so model is easier to fit to the data, i.e. in Zhang's method each grid position has its own rotation and translation parameters while in bundle adjustment these are shared. This is important in practice as only complete calibration allows us to perform reconstructions, e.g. in Fig. 8 when recording a double sided calibration board using opposing cameras reconstruction is impossible without full bundle adjustment. Even when a common a trick is used to express all extrinsic camera parameters using one selected position of the calibration board point clouds obtained from different viewpoints are not perfectly aligned as indicated by the red arrow in Fig. 8. In other words, a better reprojection error for Zhang's method is paid for with the higher uncertainty of the extrinsic parameters which is visible in reconstructed 3D point clouds.

V. CONCLUSION

We have presented a practical approach to calibrating a system comprised of multiple cameras and projectors. Presented approach is easy to use in practice and allows us to calibrate a complex imaging system quickly and efficiently.

ACKNOWLEDGMENT

This work has been supported by the Croatian Science Foundation under the grant number HRZZ-IP-2019-04-9157 (3D-CODING).

REFERENCES

[1] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," in ISPRS J. Photogramm. Remote. Sens., Vol. 115, 2016, pp. 22-36, https://doi.org/10.1016/j.isprsjprs.2015.10.004

[2] T. Luhmann, C. Fraser, and H.G. Maas, "Sensor modelling and camera calibration for close-range photogrammetry," in ISPRS J. Photogramm. Remote. Sens., Vol. 115, 2016, pp. 37-46, https://doi.org/10.1016/j.isprsjprs.2015.10.006

[3] C. S. Fraser, "Digital camera self-calibration," in ISPRS J. Photogramm. Remote. Sens., Vol. 52, No. 4, 1997, pp. 149-159, https://doi.org/10.1016/S0924-2716(97)00005-1

[4] Jason Geng, "Structured-light 3D surface imaging: A tutorial," in Adv. Opt. Photonics, Vol. 3, No. 2, 2011, pp. 128-160, http://dx.doi.org/10.1364/AOP.3.000128

[5] S. Van der Jeught, J. J. J. Dirckx, "Real-time structured light profilometry: A review," in Opt. Lasers Eng., Vol. 87, 2016, pp. 18-31, https://doi.org/10.1016/j.optlaseng.2016.01.011

[6] K. Shin, J. H. Mun, “A multi-camera calibration method using a 3-axis frame and wand,” in Int. J. Precis. Eng. Manuf., Vol. 13, 2012, pp. 283-289, https://doi.org/10.1007/s12541-012-0035-1

[7] N. Hasler, B. Rosenhahn, T. Thormahlen, M. Wand, J. Gall and H. Seidel, “Markerless Motion Capture with unsynchronized moving cameras,” in 2009 IEEE CVPR, Miami, FL, pp. 224-231, https://doi.org/10.1109/CVPR.2009.5206859

[8] T. Pribanić, S. Peharec, and V. Medved, “A Comparison Between 2D Plate Calibration and Wand Calibration for 3D Kinematic Systems,” in Kinesiology, Vol. 41, No. 2, 2009, pp. 147-155, https://hrcak.srce.hr/45835

[9] R. Juang and A. Majumder, “Photometric Self-Calibration of a Projector-Camera System,” in 2007 IEEE CVPR, Minneapolis, MN, pp. 1-8, https://doi.org/10.1109/CVPR.2007.383468

[10] D. Hrenek et al., “Tiled 360° panoramic projection system,” in 2012 Proceedings of the 35th International Convention MIPRO, Opatija, pp. 1793-1797

[11] B. Trubić et al., “System for multiple projector-based tiled video playbacks,” in The 33rd International Convention MIPRO, Opatija, 2010, pp. 1374-1378.

[12] J. Salvi, X. Armangué, J. Batlle, “A comparative review of camera calibrating methods with accuracy evaluation,” in Pattern Recognition, Vol. 35, No. 7, 2002, pp. 1617-1635, https://doi.org/10.1016/S0031-3203(01)00126-1

[13] R. Hartley and A. Zisserman, “Multiple view geometry in computer vision,” Cambrige University Press, 2003

[14] Z. Zhang, “A flexible new technique for camera calibration,” in IEEE Trans. Pattern Anal. Mach. Intell., Vol. 22, No. 11, pp. 1330-1334, 2000, http://dx.doi.org/10.1109/34.888718

[15] M. Sirianni et al., “The Photometric Performance and Calibration of the Hubble Space Telescope Advanced Camera for Surveys”, in in Publ. Astron. Soc. Pac., Vol. 117, No. 836, pp. 1049-1112, https://doi.org/10.1086/444553

[16] F. Qi, Q. Li, Y. Luo, D. Hu, “Constraints on general motions for camera calibration with one-dimensional objects,” in Pattern Recognition, Vol. 40, No. 6, 2007, pp. 1785-1792, https://doi.org/10.1016/j.patcog.2006.11.001

[17] N. Alberto Borghese, P. Cerveri, “Calibrating a video camera pair with a rigid bar,” in Pattern Recognition, Vol. 33, No. 1, 2000, pp. 81-95, https://doi.org/10.1016/S0031-3203(99)00033-3

[18] J.-Y. Bouguet, “Camera Calibration Toolbox for Matlab”, [Online, June 2020], http://www.vision.caltech.edu/bouguetj/calib_doc/

[19] OpenCV, “Camera Calibration and 3D Reconstruction”, [Online, June 2020], https://docs.opencv.org/master/d9/d0c/group__calib3d.html

[20] Mathworks, “Camera Calibration and 3-D Vision”, [Online, June 2020], https://www.mathworks.com/help/vision/camera-calibration-and-3-d-vision.html

[21] B. Triggs, P. F. McLauchlan, R. I. Hartley, A. W. Fitzgibbon, “Bundle Adjustment – A Modern Synthesis,” in Vision Algorithms: Theory and Practice IWVA 1999, LNCS Vol. 1883, https://doi.org/10.1007/3-540-44480-7_21

[22] Advanced Shape Reconstruction and Registration Laboratory, [Online, June 2020], https://www.fer.unizg.hr/en/sharklab

[23] T. Petković et al., “Multi-Projector Multi-Camera Structured Light 3D Body Scanner”, in Proc. of 3DBODY.TECH 2017 - 8th Int. Conf. and Exh. on 3D Body Scanning and Processing Technologies, Montreal QC, Canada, 11-12 Oct. 2017, pp. 319-326, https://doi.org/10.15221/17.319

[24] S. Zhang and P. S. Huang, “Novel method for structured light system calibration,” in Optical Engineering, Vol. 45, No. 8, Aug 2006, http://dx.doi.org/10.1117/1.2336196

[25] B. Li, N. Karpinsky, and S. Zhang, “Novel calibration method for structured-light system with an out-of-focus projector,” in Appl. Opt. 53, 3415-3426 (2014), https://doi.org/10.1364/AO.53.003415

[26] S. Shah and J.K Aggarwal, “Intrinsic parameter calibration procedure for a (high-distortion) fish-eye lens camera with distortion model and accuracy estimation*,” in Pattern Recognition, Vol. 29, No. 11, 1996, pp. 1775-1788, https://doi.org/10.1016/0031-3203(96)00038-6

[27] J. Kannala and S. S. Brandt, “A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses,” in IEEE Trans. Pattern Anal. Mach. Intell., Vol. 28, No. 8, pp. 1335-1340, 2006, https://doi.org/10.1109/TPAMI.2006.153

[28] P. F. Sturm and S. J. Maybank, “On plane-based camera calibration: A general algorithm, singularities, applications,” in 1999 IEEE CVPR, Fort Collins, CO, USA, 1999, pp. 432-437 Vol. 1, https://doi.org/10.1109/CVPR.1999.786974

[29] Ueshiba and Tomita, “Plane-based calibration algorithm for multi-camera systems via factorization of homography matrices,” Proceedings Ninth IEEE ICCV, Nice, France, 2003, pp. 966-973 Vol. 2, https://doi.org/10.1109/ICCV.2003.1238453

[30] MicroScribe Portable CMM [Online, June 2020] https://revware.net/revware-products/microscribe-portable-cmms

[31] T. Svoboda, D. Martinec, and T. Pajdla, “A Convenient Multi-camera Self-Calibration for Virtual Environments”, in PRESENCE: Virtual and Augmented Reality, Vol. 14, No. 4, 2005, pp. 407-422, https://doi.org/10.1162/105474605774785325

[32] M. Fiala, “Automatic Projector Calibration Using Self-Identifying Patterns,” in 2005 IEEE CVPR Workshops, San Diego, CA, USA, pp. 113-113, https://doi.org/10.1109/CVPR.2005.416

[33] A. Griesser and L. Van Gool, “Automatic Interactive Calibration of Multi-Projector-Camera Systems,” in 2006 CVPR Workshops, New York, NY, USA, https://doi.org/10.1109/CVPRW.2006.37

[34] M. Kimura, M. Mochimaru and T. Kanade, “Projector Calibration using Arbitrary Planes and Calibrated Camera,” in 2007 IEEE CVPR, Minneapolis, https://doi.org/10.1109/CVPR.2007.383477

[35] J. Liao and L. Cai, “A calibration method for uncoupling projector and camera of a structured light system,” in 2008 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Xian, 2008, pp. 770-774, https://doi.org/10.1109/AIM.2008.4601757

[36] S. Audet and M. Okutomi, “A user-friendly method to geometrically calibrate projector-camera systems,” in 2009 IEEE CVPR Workshops, Miami, FL, 2009, pp. 47-54, https://doi.org/10.1109/CVPRW.2009.5204319

[37] D. Moreno and G. Taubin, “Simple, Accurate, and Robust Projector-Camera Calibration,” in 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission, Zurich, 2012, pp. 464-471, https://doi.org/10.1109/3DIMPVT.2012.77

[38] F. Ke, J. Xie, Y. Chen, “A flexible and high precision calibration method for the structured light vision system,” in Optik, Vol. 127, No. 1, 2016, pp. 310-314, https://doi.org/10.1016/j.ijleo.2015.09.178

[39] C. Yuan, Q. Guan, X. Wang and T. Fang, “A Multiple View Self-Calibration and Metric Reconstruction Method for Structured Light System,” in 2010 Chinese Conference on Pattern Recognition (CCPR), Chongqing, 2010, pp. 1-5, https://doi.org/10.1109/CCPR.2010.5659143

[40] R. R. Garcia and A. Zakhor, “Geometric calibration for a multi-camera-projector system,” in 2013 IEEE Workshop on Applications of Computer Vision (WACV), Tampa, FL, 2013, pp. 467-474, https://doi.org/10.1109/WACV.2013.6475056

[41] T. Petković, T. Pribanić, M. Đonlić, and P. Sturm, “Efficient Separation between Projected Patterns for Multiple Projector 3D People Scanning,” in 2011 IEEE ICCV Workshops, Venice, October 22-29, 2017, https://doi.org/10.1109/ICCVW.2017.101

[42] T. Petković, T. Pribanić, and M. Đonlić, “Temporal phase unwrapping using ortho-graphic projection,” in Opt. Lasers Eng., Vol. 90, 2017, pp. 34-47, http://dx.doi.org/10.1016/j.optlaseng.2016.09.006

[43] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, “Multiscale vessel enhancement filtering,” in MICCAI’98, https://doi.org/10.1007/BFb0056195

[44] C. Steger, “An unbiased detector of curvilinear structures,” in IEEE Trans. Pattern Anal. Mach. Intell., Vol. 20, No. 2, pp. 113-125, 1998, https://doi.org/10.1109/34.659930

RTA

International Conference on

ROBOTICS TECHNOLOGIES AND APPLICATIONS

Steering Committee

Chairs: Bojan Jerbić, University of Zagreb, Croatia
Karolj Skala, Ruđer Bošković Institute, Zagreb, Croatia

Members: Tadej Bajd, University of Ljubljana, Slovenia
Ante Bakić, Inetec, Lučko, Croatia
Ricardo Branco, Portugal
Zlatko Katalenić, Robotics, Novo Mesto, Slovenia
Igor Kotenko, St. Petersburg Institute for Informatics and Automation, Russia
Zdenko Kovačić, University of Zagreb, Croatia
Danica Kragić Jensfelt, KTH, Royal Institute of Technology, Stockholm, Sweden
Jadran Lenarčič, Jožef Stefan Institute, Ljubljana, Slovenia
Duc Truong Pham, University of Birmingham, UK
Vincenzo Piuri, University of Milan, Italy
Ioan Sacala, Faculty of Automatic Control and Computers, Bucharest, Romania
Bruno Siciliano, University of Naples Federico II, Italy
Uroš Janez Stanič, Biomedical Research and Innovative Society; Kosezi d.o.o., Ljubljana, Slovenia
Zorislav Šojat, Ruđer Bošković Institute, Zagreb, Croatia

PAPERS

Estimating Robot Manipulator End-effector Forces using Deep Learning

Stanko Kružić*, Josip Musić*, Roman Kamnik†, Vladan Papić*

*University of Split, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture
†University of Ljubljana, Faculty of Electrical Engineering
E-mail: {*skruzic, jmusic, vpapic*}*@fesb.hr*, *roman.kamnik@fe.uni-lj.si*

***Abstract*—The measurement of the robotic manipulator end-effector interaction forces can in certain cases be challenging, especially when using robots that have a small payload (and consequently not capable of using wrist-mounted force sensor), which is often case with educational robots. In the paper, a method for estimation of end-effector forces using measurements from the base-mounted force sensor and deep neural networks is presented. Several deep architectures were trained using data collected on real 6-DOF robot manipulator (Commonplace Robotics Mover6 robot) using custom-made interaction object operated by a human. The obtained results show that when using appropriate deep architecture promising estimates can be achieved (with an RMSE metric on test set which was 16%, 12% and 6% of maximum force in respective directions of x, y and z axes). This makes this approach suitable for use in a variety of applications, including but not limited to usage with haptic feedback interfaces for robot control.**

***Index Terms*—robotics, force estimation, deep learning**

I. Introduction

Robotic manipulators are nowadays used for a variety of different tasks and are usually programmed to execute a set of predefined trajectories to complete a given task. This is usually referred to as position control. However, in some applications, this may be inappropriate, when feedback only about current position does not guarantee successful task completion. In these cases, force control (which also may be thought of as interaction control [1]) is used, and therefore forces on robot end-effector need to be measured. This is especially true when robot actions involve physical contact between the robot and the environment.

In most scenarios, a force sensor is mounted on the robot wrist. However, when using manipulators with small payload it may be inconvenient or even impossible to use the wrist-mount sensor. This is often the case with educational robotic manipulators or other small robots. This disadvantage can be overcome by estimating those forces instead of measuring them directly. One common method of force estimation is one using force observers, as in [2], [3], [4]. In [2] the estimation is done using model-based observers, but it requires an accurate dynamic robot model. In [3] both linear and non-linear force observers were presented (the latter applied in robot force control scheme), while in [4] control error from joint control was used to estimate forces. End-effector forces can also be estimated from joint motor torques, as in [5], which requires that torque measurements are available for each joint motor and that robot dynamics model is known. If measured joint motor torques are not available, they can be estimated using measured motor currents and in turn, be used to estimate end-effector forces [6]).

Neural networks have been, since their recent re-emergence, widely used in many fields, including robotics. They have been applied to grasping (using vision) [7], control [8], domain adaptation [9], etc. However, the force estimation in robotics is a less studied topic but does exist. In [10] neural networks are applied as an extension to force observers, which require full knowledge of system dynamics. They are used there to overcome complicated dynamics or not-fully-known dynamics model. There also exist application-specific approaches, e.g. in robotic surgery [11], [12] where estimation is done without a force sensor, but using some kind of visual feedback and neural networks.

There have been some recent applications of neural networks for learning robot (inverse) dynamics [13], [14], [15]. In [13] Lagrangian equation is implemented and used in a neural network for learning inverse dynamics. [14] uses some previous knowledge about the system to build a network. In [15] the learning of the inverse model is done using Long Short-Term Memory (LSTM) network which outperforms the Gaussian processes approach.

In the paper, we demonstrate the possibility of estimating the end-effector forces by measuring forces with a force sensor mounted under the robot base. The benefit of this method is that it does not rely on measurements of joint motor currents, because forces can be directly measured using force sensors, which are generally reliable and provide very accurate measurements. In our approach to the problem, the force sensor is mounted under the robot base and the estimation of the end-effector force is accomplished using neural networks which may prove beneficial for the task because it requires no knowledge about the robot dynamics (which is inferred implicitly, i.e. the neural network is used to approximate it).

The rest of the paper is structured as follows. In Section II, an introduction and overview of the neural network architectures are given, while Section III describes data collection process of the ~~performed~~ experiment and reports the trained NN architectures and accompanied hyperparameters. In Section IV results are presented and discussed while in Section V

conclusions are drawn based on obtained results and possible directions of future research on the topic are presented.

II. Materials and Methods

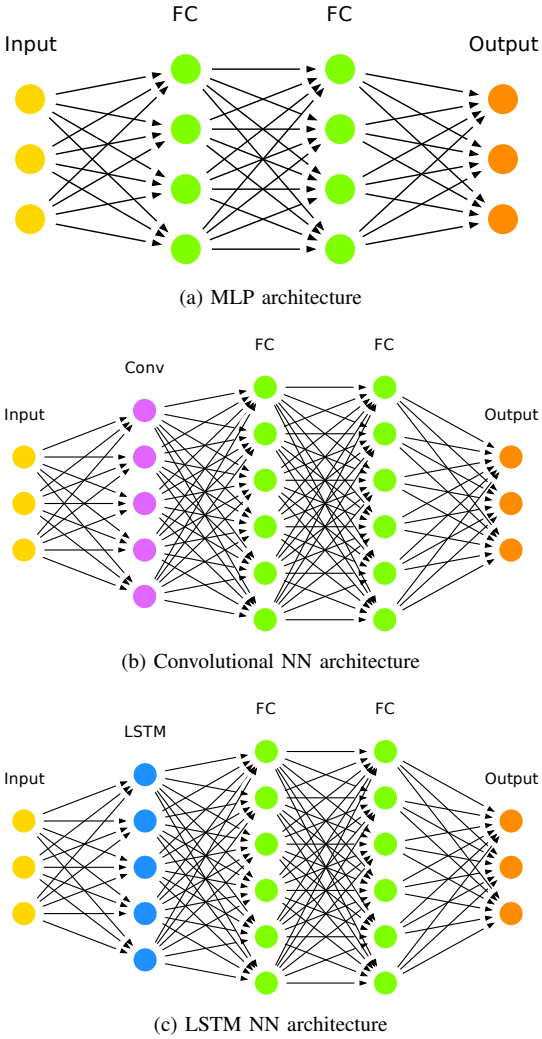
The proposed approach aims to predict end-effector forces by measuring forces on the robot base. Since it is virtually impossible to predict those values with absolute precision, the aim is to identify which NN architecture (and with what hyperparameters) will produce best results for use in a specific application.

Several NN architectures were used for the task: multilayer perceptron (MLP), convolutional NN and LSTM network, and are shown in Fig. 1. However, please note that these network graphs are only conceptual presentations of the architectures since graphs with real numbers of neurons per layer used in the experiment are inconvenient to draw. Thus, actual numbers of layers and neurons are summarised in Table I. Those architectures were used in the experiment with different numbers of layers and neurons per layer to assess which of these can produce best estimates of robot wrist force by measuring force on robot base. Networks are usually trained using back-propagation and non-linear optimisation algorithms which optimise given loss function (mean square error and mean average error are commonly used).

MLP is the simplest NN architecture which features only densely connected layers which are optimised to map inputs to outputs. On the other hand, convolutional networks use convolutional layers to extract features from input. Convolutional layers can be either 1-dimensional, which used to extract temporal features (i.e. from time sequence input), 2-dimensional, which are used for extracting spatial features (i.e. to process images) and 3-dimensional, for spatiotemporal features (sequence of images – video). LSTM [16] networks consist of LSTM layers and are a sub-type of recurrent neural networks. LSTM layers are used for processing sequential data which feature feedback connections (i.e. memory). Both aforementioned architectures have densely-connected layers to do the learning from features extracted using convolutional and LSTM layers, respectively. The MLP architecture was chosen based on the reasoning that the inputs are going to be of relatively low complexity (i.e. low number of input features), while the convolutional and LSTM architectures were chosen to assess if treating inputs as time sequences can produce better force estimates. The details about actual hyperparameters used in the experiment are presented in Section III.

III. Experimental Setup

The experiment was conducted on Commonplace Robotics Mover6 robot, a six-degrees-of-freedom educational robot. The robot manufacturer did not provide the dynamics model nor any relevant data that makes the computation of the dynamic model possible. Furthermore, the robot has a payload of only 0.4 kg, which makes the use of wrist-mounted force sensors impossible. Hence, for this purpose, an (auxiliary) interaction tool was devised to enable direct measurement of end-effector forces, as depicted in Fig. 2. The tool was



(a) MLP architecture

(b) Convolutional NN architecture

(c) LSTM NN architecture

Fig. 1. NN architectures used in the research

held by the experimenter who applied force on robot end-effector, as illustrated in Fig. 3. The robot executed random (but feasible) trajectories in the process. The other force sensor was mounted below the robot base. Both sensors were JR3 90M40 6-axis force-torque sensors, and data acquisition was done using Mathworks Simulink Real-Time software which also synchronised sensors at a rate of 100 Hz. Also, Optotrak Certus optical motion capture system was used for assessing the positions and orientations of the robot base and the interaction tool. The position data measurements were synchronised with force measurements. Robot joint positions that are provided by encoders on each of the joints were also recorded.

The data were collected as follows. The experimenter holding the interaction tool applied forces to the robot by making a contact between the interaction tool dome and the robot end-effector while the robot was either in motion or at a standstill. During the robot motion, random (but feasible) goal positions and orientations were generated and trajectories executed. Between two successive motion trajectories, the robot was,

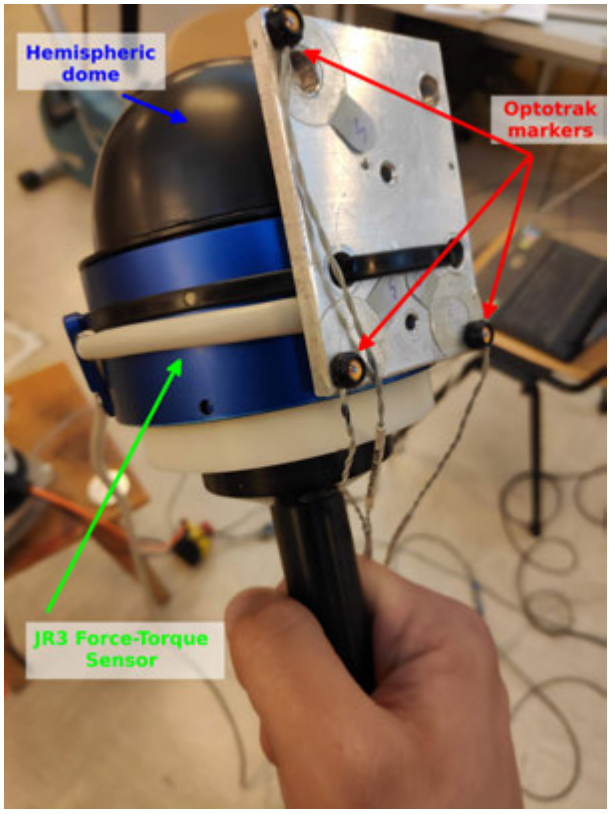


Fig. 2. Custom-built interaction tool used in the data collection with Optotrak markers. Please note that the hemispheric dome placed on top of the force sensor to focus forces.

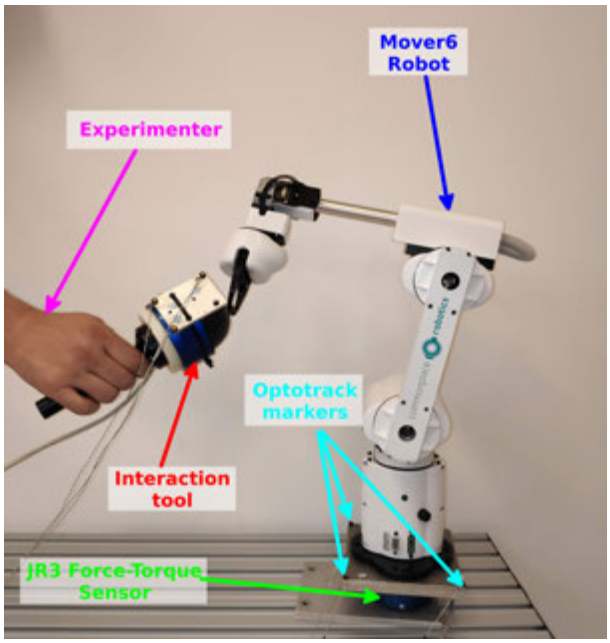


Fig. 3. The process of collecting measurements.

for a short period of time, at a standstill. A random number of up to six contacts were made in a single measurement instance, and a total of 800 measurement instances of random length, ranging from approx. 10 s to approx. 30 s were recorded, resulting in a total of 1,803,875 samples. All the measurement instances were preprocessed in such a way that both robot base and interaction tool forces were expressed in the same reference frame. This was achieved by using positions obtained from the Optotrak markers, three positioned on the robot base and three on the interaction tool, which was used to define a reference frame corresponding to the robot base and the interaction tool, respectively. Following this, the transformation between the two frames was obtained, which then was used to express both measured forces in a common reference frame so that the principal axes from both force measurements match. Finally, the obtained forces were low-pass filtered to remove noise.

From obtained measurement data, two datasets were built: one with filtered data with mean removed from data and scaled to unit variance (i.e. normalised); the other was same except that recorded joint positions were also included as inputs along with the measured forces. Due to the fact that multiple robot poses may result in the same measured forces, we included this in the dataset to assess if it will help neural networks to produce better estimates. The datasets were then divided into training, validation and test set as follows. The test set used random 20% data, the validation set used further 20% of what's left and the remaining data were used as a training set. The forces measured on the robot base-mounted sensor served as inputs to the NN (possibly along with joint positions), while the measured interaction tool forces were used as NN targets (outputs).

Several networks were trained for each of the architectures introduced in Section II with varying hyperparameters. The overview of the used architectures with their respective hyper-parameters is summarised in Table I, while obtained results are presented and discussed in the following section. Please note that MLP networks have a higher number of neurons per layer than other architectures, which is based on authors' previous experience in the field and on the fact that MLP does the learning on raw input data rather than temporal features (which is the case with convolutional and LSTM architectures, and thus need more neurons to generalise properly. In convolutional and LSTM networks, a single 1-d convolution / LSTM layer was used at the beginning of the network, reasoning that for the data of relatively low dimensionality and complexity as our is (only forces along three principal axes), a single layer is enough to extract temporal features. Those layers accept sequences of fixed length N as inputs. Thus, the sequences in the dataset had to be split into sequences of length N, and that was done such that first sequence consisted of samples of 1 to N, then from samples 2 to $N+1$, etc. from the originally recorded sequence. In the research, sequence lengths used were 5 and 10, while the number of LSTM and convolutional units in the layer was fixed to 8.

All the networks were initialised with a Glorot uniform

initialiser [17], and trained using Adam optimiser [18], starting with a learning rate of 0.001 (the optimiser is adaptive, thus the learning rate may change during the training). In the training process, the mean absolute error (MAE) was used as a loss function, while for activating neurons, the ReLU function was used. In all instances, the training lasted until there was no improvement in validation loss for 10 consecutive epochs. The training was conducted using Tensorflow 2.0 and Keras.

IV. Results and Discussion

Networks presented in Table I were trained, with the results being presented in Table II. As a measure of network fitness, the validation loss and test loss (obtained using MAE loss function) were used. As a metric of fit between targets and predictions, root-mean-square errors (RMSE) are also reported. Please note that all reported loss values are computed for all three principal axes together and are a ℓ^2 norms of vector losses along principal axes, while RMSE values are reported on the test set along each of principal axes separately.

From the results, it is not immediately observable which architecture is optimal, because all of them, at first sight, perform similarly, with no significant differences between various architectures. However, there appear some interesting observations when looking at end-effector force predictions on the test set. Examples of predictions using different trained architectures are shown in Fig. 4 (please note that Figs. 4a, 4c and 4e show estimated from same test case; similar is true for Figs. 4b, 4d and 4f, but another test case is used). Obtained test results partly contain both "good" and "bad" predictions (i.e. in one part of a single test case, predictions are fair, while on the other part are not). But, the general observation is that obtained predictions suggest that MLP architecture performs marginally worse than others, and by visual inspection of the obtained predictions, it seems that they perform much worse than that measured network fitness in Table II suggest. It was also observed that predictions made using MLP networks perform slightly worse and oscillate slightly more than predictions made using the other two architectures.

When looking at RMSE obtained on the test set along each of the principal axes, it may be concluded that networks with a smaller number of total trainable parameters generally perform better (not depending on architecture). This is likely due to simple input data (i.e. a small number of features). By using this metric, the best architecture is, still by a small margin, convolutional network (the smallest in term of the number of trainable parameters, among those trained). Also, on most of the trained networks, RMSE along z axis is the smallest one, which is encouraging, due to fact that force component along z axis is usually a dominant component of the force vector. For the architecture with the best obtained RMSE, 6% RMSE with respect to the maximum force along z axis was achieved, which is significantly better than 16% and 12% achieved along axes x and y).

Based on the results, there is no clear-cut decision which kind of architecture is optimal for the task. However, architectures that consider input (measured) forces as time-series data have marginally better predictions. Please note that hyperparameter tuning for each of the architectures may provide somewhat better results, but that would likely require a grid search approach to identify optimal hyperparameters and also a significant amount of time to train networks with all possible values of all chosen hyperparameters (i.e. the number of layers, number of neurons per layers, activation function, optimiser, loss function, etc.).

TABLE I
Trained architectures used in the experiment

#	Arch.	Joint pos.	Layers (neurons)[1]	Seq. length
1.	MLP	No	3 (64)	/
2.	MLP	Yes	3 (64)	/
3.	MLP	Yes	2 (32)	/
4.	Conv	No	3 (32)	10
5.	Conv	Yes	2 (16)	10
6.	Conv	Yes	2 (16)	5
7.	LSTM	No	2 (16)	10
8.	LSTM	Yes	2 (16)	10
9.	LSTM	Yes	2 (16)	5

[1] Layers pertain to number of densely connected hidden layers, while neurons pertain to number of neurons in each hidden layer.

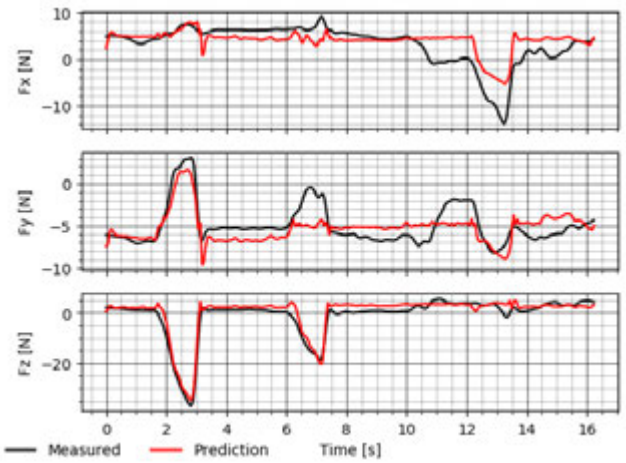
TABLE II
Network fitness and RMSE metric along principal axes (row numbers are corresponding to architectures enumerated in Table I. All quantities are expressed in Newtons [N]

#	Val. set	Test set	RMSE_x	RMSE_y	RMSE_z
1.	1.92130	3.5392	2.5623	3.0335	2.7934
2.	2.10563	3.2132	2.8725	3.0540	2.7845
3.	1.95203	3.4192	2.6007	2.8154	2.5659
4.	2.02529	3.4508	2.8554	2.9534	2.6193
5.	1.92932	3.5604	2.5468	2.7671	2.4515
6.	**1.88883**	3.3298	**2.5386**	**2.7577**	**2.3894**
7.	2.02308	3.3601	2.5757	2.9328	2.6488
8.	1.96090	**3.0722**	2.5637	2.8173	2.5134
9.	1.97469	3.5950	2.6007	2.8573	2.6239

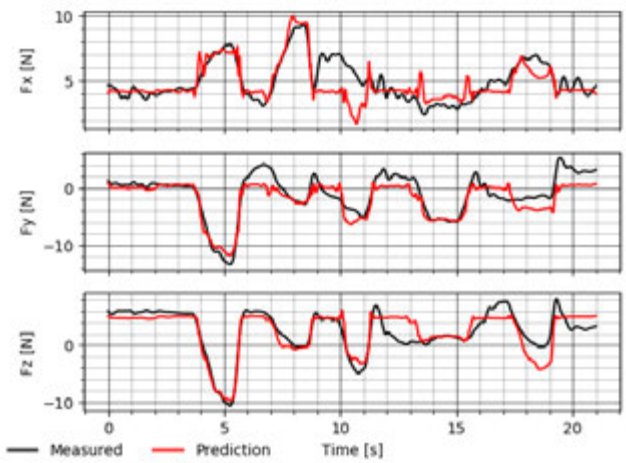
V. Conclusion

In the paper, the method for estimation of robot manipulator end-effector forces is presented. The forces are measured with a force sensor mounted under the robot base, and end-effector forces are estimated using deep neural networks. It is shown that less complex networks (in the number of parameters, which corresponds to the lower number of layers and/or lower number of neurons per layer) perform better, likely due to low complexity (and dimensionality) of input data.

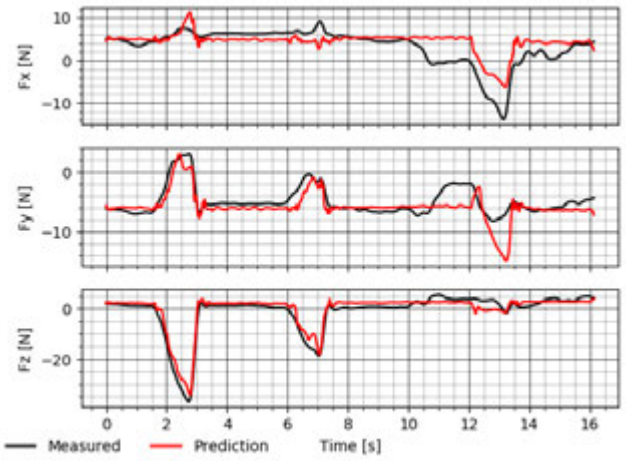
Based on obtained results, it might be concluded that this approach provides good estimates and that can be used in applications not requiring a high degree of precision. However, if high precision is required, this approach needs to be extended. One of directions of future development of this approach might be to also include joint velocities and accelerations in the



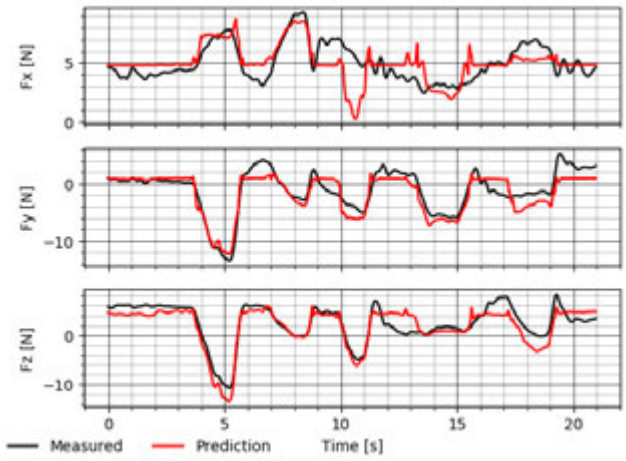
(a) Arch. #3



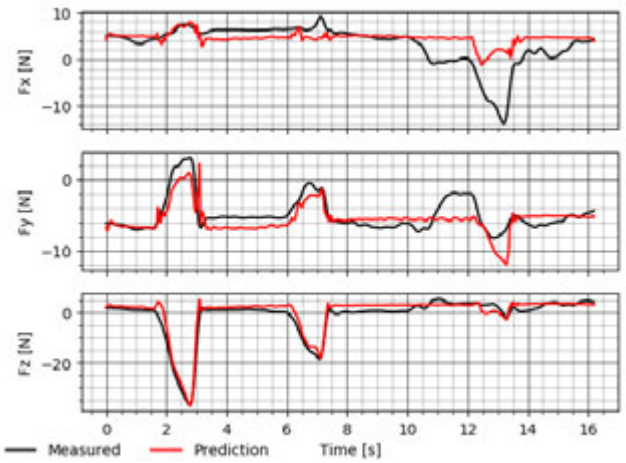
(b) Arch. #2



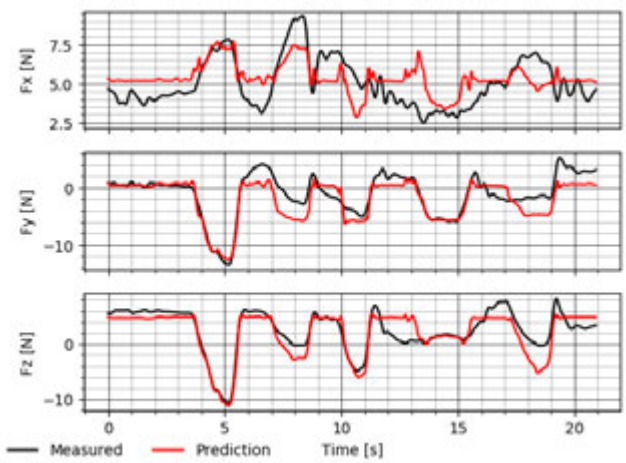
(c) Arch. #4



(d) Arch. #6



(e) Arch. #9



(f) Arch. #8

Fig. 4. Prediction examples on test set for trained networks.

dataset possibly leading to better predictions, reasoning that those are used for computations of the inverse dynamics analytically, and hence might prove useful for better generalisation of the trained neural networks. The other direction might be incorporating knowledge about the robotic system into the model, leading to the better capture of the physical model of the robot. This may be tackled by introducing additional layer(s) in neural networks that implement robot physics [13], [14] and consequently leading to better prediction of contact forces, due to physical constraints those new layers eventually imposing.

REFERENCES

[1] B. Siciliano and L. Villani, *Robot force control*, vol. 540. Springer Science & Business Media, 2012.

[2] P. J. Hacksel and S. E. Salcudean, "Estimation of environment forces and rigid-body velocities using observers," in *Proceedings of the 1994 IEEE International Conference on Robotics and Automation*, pp. 931–936 vol.2, May 1994.

[3] A. Alcocer, A. Robertsson, A. Valera, and R. Johansson, "Force estimation and control in robot manipulators," *IFAC Proceedings Volumes*, vol. 36, no. 17, pp. 55 – 60, 2003. 7th IFAC Symposium on Robot Control (SYROCO 2003), Wroclaw, Poland, 1-3 September, 2003.

[4] A. Stolt, M. Linderoth, A. Robertsson, and R. Johansson, "Force controlled robotic assembly without a force sensor," in *2012 IEEE International Conference on Robotics and Automation*, pp. 1538–1543, May 2012.

[5] M. Van Damme, P. Beyl, B. Vanderborght, V. Grosu, R. Van Ham, I. Vanderniepen, A. Matthys, and D. Lefeber, "Estimating robot end-effector force from noisy actuator torque measurements," in *2011 IEEE International Conference on Robotics and Automation*, pp. 1108–1113, May 2011.

[6] A. Wahrburg, J. Bös, K. D. Listmann, F. Dai, B. Matthias, and H. Ding, "Motor-current-based estimation of cartesian contact forces and torques for robotic manipulators and its application to force control," *IEEE Transactions on Automation Science and Engineering*, vol. 15, pp. 879–886, April 2018.

[7] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 421–436, 2018.

[8] L. Jin, S. Li, J. Yu, and J. He, "Robot manipulator control using neural networks: A survey," *Neurocomputing*, vol. 285, pp. 23 – 34, 2018.

[9] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, S. Levine, and V. Vanhoucke, "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4243–4250, May 2018.

[10] A. C. Smith and K. Hashtrudi-Zaad, "Application of neural networks in inverse dynamics based contact force estimation," in *Proceedings of 2005 IEEE Conference on Control Applications, 2005. CCA 2005.*, pp. 1021–1026, Aug 2005.

[11] A. Marbán, V. Srinivasan, W. Samek, J. Fernández, and A. Casals, "A recurrent convolutional neural network approach for sensorless force estimation in robotic surgery," *CoRR*, vol. abs/1805.08545, 2018.

[12] A. I. Aviles, S. Alsaleh, P. Sobrevilla, and A. Casals, "Sensorless force estimation using a neuro-vision-based approach for robotic-assisted surgery," in *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 86–89, April 2015.

[13] M. Lutter, C. Ritter, and J. Peters, "Deep lagrangian networks: Using physics as model prior for deep learning," *CoRR*, vol. abs/1907.04490, 2019.

[14] F. Díaz Ledezma and S. Haddadin, "First-order-principles-based constructive network topologies: An application to robot inverse dynamics," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pp. 438–445, Nov 2017.

[15] E. Rueckert, M. Nakatenus, S. Tosatto, and J. Peters, "Learning inverse dynamics models in o(n) time with lstm networks," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pp. 811–816, Nov 2017.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Y. W. Teh and M. Titterington, eds.), vol. 9 of *Proceedings of Machine Learning Research*, (Chia Laguna Resort, Sardinia, Italy), pp. 249–256, PMLR, 13–15 May 2010.

[18] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

Prediction of the behavior of a pneumatic soft robot based on Koopman operator theory

E. Kamenar [1, 2], N. Črnjarić-Žic [1], D. Haggerty [3], S. Zelenika [1, 2], E. W. Hawkes [3] and I. Mezić [2, 3]

[1] University of Rijeka, Faculty of Engineering, Vukovarska 58, 51000 Rijeka, Croatia
[2] University of Rijeka, Centre for Micro- and Nanosciences and Technologies, R. Matejčić 2, 51000 Rijeka, Croatia
[3] UC Santa Barbara, Department of Mechanical Engineering, Santa Barbara, CA 93105, USA

ekamenar@riteh.hr

***Abstract* – Thanks to their flexibility, soft robotic devices offer critical advantages over rigid robots, allowing adaptation to uncertainties in the environment. As such, soft robots enable various intriguing applications, including human-safe interaction devices, soft active rehabilitation devices, and soft grippers for pick-and-place tasks in industrial environments. In most cases, soft robots use pneumatic actuation to inflate the channels in a compliant material to obtain the movement of the structure. However, due to their flexibility and nonlinear behavior, as well as the compressibility of air, controlled movements of the soft robotic structure are difficult to attain. Obtaining physically-based mathematical models, which would enable the development of suitable control approaches for soft robots, constitutes thus a critical challenge in the field. The aim of this work is, therefore, to predict the movement of a pneumatic soft robot by using a data-driven approach based on the Koopman operator framework. The Koopman operator allows simplifying a nonlinear system by "lifting" its dynamics into a higher dimensional space, where its behavior can be accurately approximated by a linear model, thus allowing a significant reduction of the complexity of the design of the resulting controllers.**

Keywords – soft robots, Koopman operator, nonlinear lifting

I. INTRODUCTION

Owing to their compliant behavior, when compared to traditional rigid (or, as sometimes referred to in literature, hard [1]) robots, soft robots offer important advantages. They are made of highly compliant (soft) materials, which allow them to adapt their shape to environmental constraints and obstacles, making possible their efficient use in cluttered surroundings [2, 3]. Although such devices are currently still mostly employed by researchers in experimental systems, the number of examples of their practical use is increasing. Soft robots are, hence, used in various interesting applications, including human-safe interaction devices and soft grippers for pick-and-place tasks in industrial environments [2–4]. Furthermore, soft robotic devices represent a promising technology in medical applications, where they can be used as actuating (exoskeletal) components in active rehabilitation devices for human motion assistance [5], particularly for stroke patients [2, 4, 6].

Pneumatic actuation with air used as the working fluid, is often employed to inflate the hollow compliant structure, which enables the motion of the end-effector of the soft robot [3]. The high compliance, which represents one of the main advantages of these devices, at the same time makes their dynamical behavior highly nonlinear. In fact, the inherent nonlinear behavior, high dimensionality, as well as the compressibility of air, makes pneumatic soft robots very complex to model. On the other hand, mathematical modelling is necessary to synthetize an appropriate control system that enables the controlled movements of this class of devices. The modeling of soft robotic systems is furthermore an important subject because it shortens the time used for their design and deployment.

The Koopman operator, in turn, allows a linearized representation of nonlinear dynamical systems, exhibiting excellent performances. This mathematical tool is, therefore, increasingly suggested for modelling and control in e.g. biology, power grids, fluid dynamics, traffic or DC motors controls. What is more, the Koopman operator was recently successfully applied to model and control a factual complex positioning system, aimed at nanometric precision positioning, characterized by multiple sources of intricate mechanical nonlinearities induced by friction [7-10].

The aim of this work is, therefore, to provide an advanced tool for modelling the behavior of a pneumatically-driven soft robotic device, available at the premises of the University of California Santa Barbara (UCSB), by employing a data-driven machine learning approach based on Koopman operator theory [7-10]. The developed model creates the preconditions to synthetize a motion controller allowing to attain a precise tracking control of the end-effector of the used soft robotic device.

The remainder of this paper is organized as follows: in Section II a brief introduction to the Koopman operator theory, with application to nonlinear dynamical systems, is provided, followed in Section III by the description of the devised experimental set-up. The modelling procedure based on the experimentally acquired motion data and the thorough analyses of the obtained results are described in Section IV. In Section V, the main conclusions are finally drawn, and an outlook to future work is provided.

II. KOOPMAN OPERATOR THEORY IN SYSTEMS' MODELLING AND CONTROL

The Koopman operator theory allows a linearization of highly-nonlinear dynamical systems, exhibiting, in terms of accuracy, performances superior to other linear predictors, such as those based on local linearization or the so-called Carleman linearization [7]. The respective method for calculating linear predictors is relatively simple and, contrary to physically-based models grounded on knowledge of material properties and first-order principles, it is data-driven, i.e., it depends on experimental input-output data only, while it provides a global and smooth view of systems' behavior even for longer prediction intervals [7, 10, 11]. The Koopman operator represents, thus, a machine learning data-driven method that can be used to obtain a state-space representation of the model of the studied device [11, 12]. Its numerical approximations allow "lifting" the nonlinear dynamics of the considered factual device (i.e., of its state-space model) into a higher dimensional space of so-called observables (i.e., simple scalar functions), where the behavior of the considered system can be predicted by a simpler yet accurate linear (though higher dimensional [12]) model [7, 10].

What is more, the nonlinearities pertaining to many physical systems would often imply the need to use nonlinear control approaches that are associated with large computational times, as well as stability and robustness issues, especially in higher-dimensional problems [8]. Koopman operator theory, recently introduced to the control community as well [7, 13], allows, in turn, a significant simplification of the model-based design of the controller and reduces the computational complexity in real-time applications [8]. In other words, the obtained Koopman-based linear predictors can be used to design controllers that can cope with the nonlinear dynamics of the controlled nonlinear system, although they rely on linear control design approaches such as the Model Predictive Controller (MPC), the Linear Quadratic Regulator (LQR), the H-infinity method (H_∞) or similar [7, 10].

In the herein considered case, a suitable modelling of soft robot is particularly important, as it allows developing feedforward predictors that should, as precisely as possible, predict the behavior of the physical system. Feedforward controllers are complemented in this case with a feedback term that compensates for the eventual minor model uncertainties and external disturbances only [11]. If, in turn, the control system would rely mostly on the feedback term, the advantages induced by the compliance of the soft robot would be highly reduced, as it is well-known that the feedback term tends to increase the overall stiffness of the controlled system [14].

Taking into account the stated advantages, Koopman operator theory is applied to an innovative soft robotic device described in the following section.

III. EXPERIMENTAL SET-UP

The experiments conducted in the framework of this study are performed on a soft robotic experimental device developed at UCSB and depicted in Fig. 1 [15]. The device consists of a flexible air-tight envelope ("skin") and two "muscles", inflated each via an independent pressure input

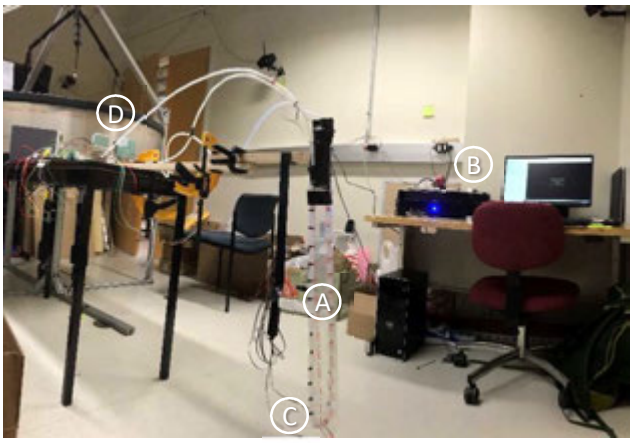


Figure 1. Experimental set-up at UCSB: soft robot (A), motion capture device (B), LED tracker (C) and electromagnetic pressure valves (D)

unit (designated in Fig 1 with "A"). A PhaseSpace Inc. Impulse X2E Motion Capture System (indicated with "B"), consisting of six cameras and eight active LED trackers attached to the characteristic points at the surface of the soft robot (see label "C" in Fig 1), is used to collect motion data. The used motion tracking device enables a sampling frequency of up to ~1 kHz and a sub-millimeter motion resolution [16]. Given its high sampling rate, the same motion capture system will likely be used in the second phase of the work as the displacement feedback for the closed-loop position control of the robot's end-effector.

The herein considered open-loop experiments are conducted by exciting the system via two randomized (white-noise) inputs in the form of PWM voltage signals applied to the proportional electromagnetic valves (indicated in Fig. 1 with "D") that generate the respective pressure output on each of the two "muscles", while the "skin" is inflated to a constant pressure. The same controller utilized for commanding the control signals to the muscles, and recording the real-time pressure in each of them, is also used to trigger the motion capture system and inherit its sampling frequency, thus synchronizing the input-output data. The acquired data consists, then, of time-related sets of *x-y* (planar) coordinates relating to the motion of each of the LED trackers. This information is used next to obtain the corresponding data-driven model of the soft robot by employing Koopman operator theory.

IV. MODELLING OF THE RESPONSE OF THE SOFT ROBOT

As pointed out, the infinite-dimensional Koopman operator is an emerging tool for the advanced modeling of highly-nonlinear dynamical systems. Identifying a nonlinear model by using experimental data implies, in turn, solving non-linear optimization problems based on an extensive knowledge of system's behavior [11, 17].

A. Application of Koopman Operator Theory to the Soft Robotic Device

It is supposed here that the discrete-time representation of the used soft robotic dynamical system in the time step t_{i+1} can be defined in the state space as [7, 11]:

$$x_{i+1} = f(x_i, u_i), \tag{1}$$

where $x_i = x(t_i) \in X \subseteq \mathbb{R}^n$ is the state of the system, while $u_i \in \mathbb{R}^m$ is the control input in the discrete time step t_i. It is assumed next that, by applying to the state space the vector of lifting functions $\Psi(x) = \big(\psi_1(x), \dots, \psi_N(x)\big)^T = z \in M \subseteq \mathbb{R}^N$, where $\psi_i: X \to \mathbb{R},\ i = 1, \dots, N$ are the chosen observable functions, the evolution in the lifting space *M* has the form:

$$\begin{aligned} z_{i+1} &= \mathbf{A}z_i + \mathbf{B}u_i \\ x_{i+1} &= \mathbf{C}z_{i+1} \end{aligned} \tag{2}$$

where **A** is system's matrix, **B** is the control matrix, **C** is the projection operator from the lifting space to the state space, while $z_0 = \Psi(x_0)$.

The aim of the work is, therefore, to predict the motion behavior of the soft robotic device based on the standardized state-space model defined by (2).

The Koopman operator $\mathcal{K}$ associated to the original system is the infinite-dimensional linear operator acting on the generalized observables $\phi: X \to \mathbb{R}$ as $\mathcal{K}\phi(x) = \phi\big(f(x,u)\big)$. Since the aim here is to obtain the time-domain prediction of the behavior of the studied dynamical system, the Extended Dynamic Mode Decomposition (EDMD) algorithm, which relies mainly on least square regression, is adopted next to construct the finite-dimensional approximations of the Koopman operator [17, 18]. By applying EDMD on the observable functions $\{\psi_1(x), \dots, \psi_N(x)\}$, the finite-dimensional approximation of the compression of the Koopman operator on the linear subspace, spanned by these functions, is hence obtained [17, 18]. Furthermore, with the assumption that in the lifted space the dynamical system evolves according to (2), in the considered problem the representation of this finite dimensional approximation will be of the form [**A**, **B**], and can be obtained as the solution of the minimization problem [7]:

$$\min_{\mathbf{A},\mathbf{B}} \sum_{i=0}^{K-1} \|\Psi(x_{i+1}) - \mathbf{A}\Psi(x_i) - \mathbf{B}u_i\|_2^2 . \tag{3}$$

In the data-driven framework, by denoting $\mathbf{X}_{lift} = [\Psi(x_0), \dots, \Psi(x_{K-1})]$, $\mathbf{Y}_{lift} = [\Psi(x_1), \dots, \Psi(x_K)]$ and $\mathbf{U} = [\mathrm{u}_0, \dots, \mathrm{u}_{K-1}]$, the matrices **A** and **B** can be obtained by finding an analytical solution to (3) by using [7]:

$$[\mathbf{A}, \mathbf{B}] = \mathbf{Y}_{lift}\big[\mathbf{X}_{lift}, \mathbf{U}\big]^{\dagger} \tag{4}$$

where † denotes the Moore-Penrose pseudoinverse of a function. It has to be noted here that matrix **U**, which contains the input control signals, remains non-lifted.

The projection operator relating to the data vector **X** is determined next as [7]:

$$\mathbf{C} = \mathbf{X}\,\mathbf{X}_{\boldsymbol{lift}}^{\dagger}. \tag{5}$$

Based on the obtained model, the p step horizon prediction of the behavior of the system $\hat{z}_{j+p}$, from an initial state x_j can be computed recursively, for $i = 1, \dots, p$, where the initial state of recursion is given by $\hat{z}_j = \Psi\big(x_j\big)$, as [7, 11]:

$$\begin{aligned} \hat{z}_{j+i} &= \mathbf{A}\hat{z}_{j+i-1} + \mathbf{B}u_{j+i-1} \\ \hat{x}_{j+i} &= \mathbf{C}\,\hat{z}_{j+i}. \end{aligned} \tag{6}$$

B. Results and Discussion

In the considered case the open loop planar motion coordinates x and y of the end-effector of the used soft robot, as measured at discrete time points for determined input pressures (i.e., control inputs u_i) at each of the two muscles, represent the attained data points x_i.

The sampling time in the numerical experiments is, in turn, set to 1 ms. Two sets of lifting functions are then considered in the EDMD algorithm. The first set consist of first-order monomials (r=1), where the lifting space coincides with the original state-space. On the other hand, the second batch of lifting functions is enlarged with second-order monomials (r=2). Taking into account the number of data points, the dimensions of matrices $\mathbf{X}_{lift}$ and $\mathbf{Y}_{lift}$ are $N \times 160\,000$.

The preliminary experimental motion data of a specific LED tracker on the end effector of the studied UCSB soft robotic device, overlaid by the respective learned responses for different prediction horizons, as obtained by applying the above Koopman operator-based approach, are thus compared in Figures 2 – 4. The depicted responses are given for an approximate 1 s time-frame, randomly chosen in the experimental data set. From the presented results, it can be deduced that the modeled responses follow very well the experimental data for a 10 ms prediction step. For larger prediction steps, the delay between the modeled and the experimental responses increases. However, it can be noticed that even for a 30 ms prediction time, the modeled response follows still the

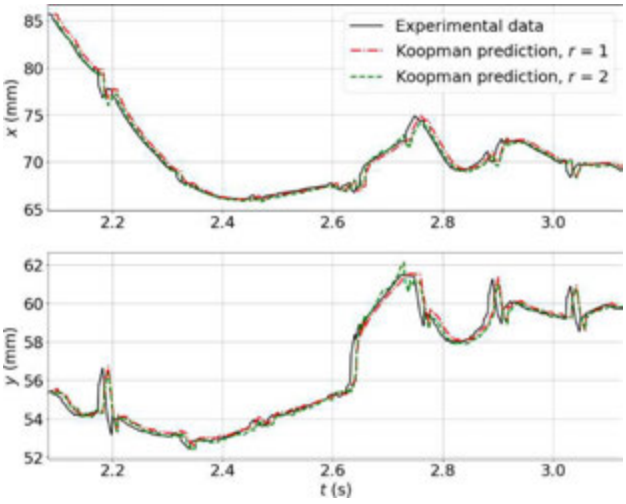


Figure 2. Comparison of the measured and predicted planar motion of the end effector of the soft robot for 10 time steps equaling a 10 ms prediction interval

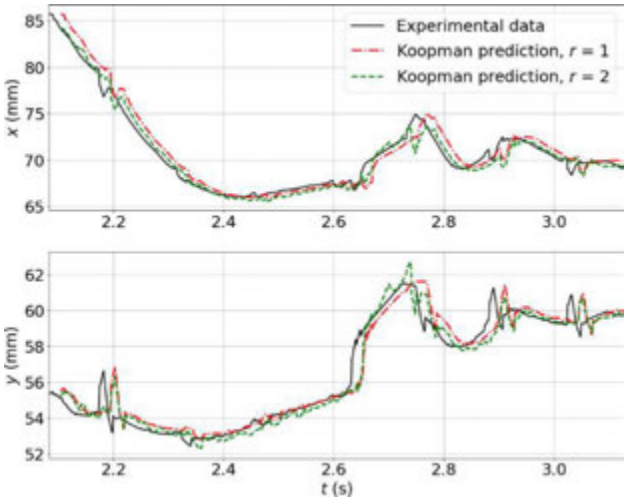


Figure 3. Comparison of the measured and predicted planar motion of the end effector of the soft robot for 20 time steps equaling a 20 ms prediction interval

experimental data with good accuracy. A further mathematical analysis of the goodness of fit of the modeled vs. experimental data is performed next, based on the order of monomials and the length of the prediction horizon, using as the relevant indicator the Mean Absolute Error (MAE) normalized over the mean value of the measured data points (the normalized MAE indicator is hence designated as NMAE):

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i| \qquad \mathrm{NMAE} = \left(\frac{\mathrm{MAE}}{\bar{x}}\right). \tag{7}$$

The calculated NMAE values for the considered experimental set-up are hence reported in Table 1. It can be seen that the deviation between the modeled and the experimental response slightly decreases when higher order monomials are used as the lifting functions while,

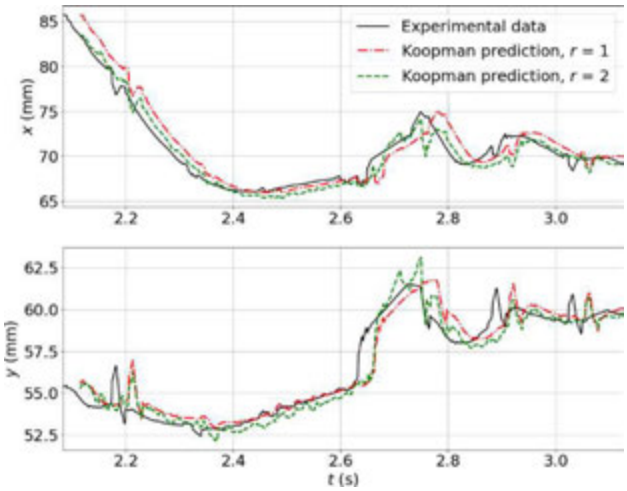


Figure 4. Comparison of the measured and predicted planar motion of the end effector of the soft robot for 30 time steps equaling a 30 ms prediction interval

TABLE I.
CALCULATED NMAE (%) FOR DIFFERENT PREDICTION HORIZONS AND USED LIFTING FUNCTIONS

τ (ms)	r=1		r=2	
	error x (%)	error y (%)	error x (%)	error y (%)
10	1.9	1.55	1.7	1.5
20	3.55	2.84	3.22	2.75
30	5.2	4	4.7	3.94
40	6.86	5.24	6.16	5.16
50	8.43	6.42	7.63	6.38

evidently, the value of the NMAE indicator increases with longer prediction step.

The obtained results represent, therefore, a good basis for the development of controllers that would enable the precise tracking of the end-effector of soft robots.

V. CONCLUSIONS AND OUTLOOK

The experimental identification of the behavior, and the respective Koopman operator-based modelling of an innovative soft robotic experimental device, is performed in this work. The acquisition of experimental motion data is based in this frame on the excitation of the system by random input signals. Based on the attained experimental data, a state-space model of the system is built by employing a finite-dimensional Extended Dynamic Mode Decomposition Koopman operator approximation. The comparison of the experimental responses with those obtained from simulations on the obtained Koopman-based model, for different prediction steps, allows establishing that the modelled responses follow accurately the experimental data.

Based on the proposed modelling approach, in the next phases of the work further experimental tests will be conducted to acquire larger set of experimental data with the goal of extending the accurate prediction window. Furthermore, advanced control typologies, e.g. based on MPC, LQR, H_∞ or similar linear control design methods, will be synthetized with the aim of attaining the precise tracking control of soft robotic devices.

ACKNOWLEDGMENTS

The work performed by E. Kamenar and S. Zelenika is made possible by using the equipment funded via the ERDF project RC.2.2.06-0001 “Development of Research Infrastructure at the University of Rijeka Campus - RISK”, and supported by the University of Rijeka grant uniri-tehnic-18-32 “Advanced mechatronics devices for smart technological solutions”.

REFERENCES

[1] R. V. Martinez, C. R. Fish, X. Chen, and G. M. Whitesides, “Elastomeric origami: Programmable paper-elastomer composites as pneumatic actuators”, *Adv Funct Mater*, vol. 22(7), pp. 1376-84, 2012.

[2] C. Majidi, “Soft robotics: A perspective - Current trends and prospects for the future”, *Soft Robot*, vol. 1(1), pp. 5-11, 2014.

[3] D. Rus, and M.T. Tolley, "Design, fabrication and control of soft robots", *Nature*, vol. 521(7553), pp. 467-75, 2015.

[4] P. Polygerinos, S. Lyne, Z. Wang, L. F. Nicolini, B. Mosadegh, G. M. Whitesides, and C. J. Walsh, "Towards a soft pneumatic glove for hand rehabilitation", in *2013 IEEE/RSJ Int. Conf. Intel. Robots & Sys.*, pp. 1512-17, November 2013.

[5] S. Li, D. M. Vogt, D. Rus, and R. J. Wood, "Fluid-driven origami-inspired artificial muscles", *P Natl Acad Sci USA*, vol. 114(50), pp. 13132-7, 2017.

[6] C. Majidi, "Soft-Matter Engineering for Soft Robotics", *Adv Mater Technol*, vol. 4(2), pp. 1800477, 2019.

[7] M. Korda, and I. Mezić, "Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control", *Automatica*, vol. 93, pp. 149-60, 2018.

[8] E. Kaiser, J. N. Kutz, and S. L. Brunton, "Data-driven approximations of dynamical systems operators for control", *arXiv preprint arXiv:1902.10239*, 2019.

[9] I. Abraham, G. De La Torre, and T. D. Murphey, "Model-based control using Koopman operators", in *Robotics: Science and Systems XIII*, Cambridge, USA, January 2017.

[10] S. Zelenika, E. Kamenar, M. Korda, and I. Mezic, "Application of Koopman-Based Control in Ultrahigh-Precision Positioning," in *The Koopman Operator in Systems and Control: Concepts, Methodologies and Applications*, vol. 484. A. Mauroy, I. Mezic, Y. Susuki, Eds., Springer Nature, 2020 (in press).

[11] D. Bruder, C. D. Remy, and R. Vasudevan, "Nonlinear system identification of soft robot dynamics using Koopman operator theory", *arXiv preprint arXiv:1810.06637*, 2018.

[12] D. Bruder, B. Gillespie, C. D. Remy, and R. Vasudevan, "Modeling and control of soft robots using the Koopman operator and model predictive control", *arXiv preprint arXiv:1902.02827*, 2019.

[13] H. Arbabi, M. Korda, and I. Mezic, "A data-driven Koopman model predictive control framework for nonlinear flows", *arXiv preprint arXiv:1804.05291*, 2018.

[14] C. Della Santina, M. Bianchi, G. Grioli, F. Angelini, M. Catalano, M. Garabini, and A. Bicchi, "Controlling soft robots: Balancing feedback and feedforward elements", *IEEE Robot Autom Mag*, vol. 24(3), pp. 75-83, 2017.

[15] J. D. Greer, T. K. Morimoto, A. M. Okamura, and E. W. Hawkes, "Series pneumatic artificial muscles (sPAMs) and application to a soft continuum robot", in *2017 IEEE International Conference on Robotics and Automation*, pp. 5503-10. May 2017.

[16] "The Impulse X2E Motion Capture System". Internet: http://phasespace.com/x2e-motion-capture/, 2019.

[17] C. Folkestad, D. Pastor, I. Mezic, R. Mohr, M. Fonoberova, and J. Burdick, "Extended dynamic mode decomposition with learned Koopman eigenfunctions for prediction and control", *arXiv preprint arXiv:1911.08751*, 2019.

[18] N. Črnjarić-Žic, S. Maćešić and I. Mezić, "Koopman operator spectrum for random dynamical system", arXiv preprint arXiv:1711.03146, 2019.

Conceptual design of an autonomous rover with ground penetrating radar: application in characterizing soils using deep learning

P. Linna[a], T. Aaltonen[b], A. Halla[a], J. Grönman[a], N. Narra[a]
[a]Computing Sciences Unit, ITC Faculty, Tampere University, Pori, Finland
[b]Satakunta University of Applied Sciences, Pori, Finland
petri.linna@tuni.fi

Abstract - **In the pursuit to make agricultural production efficient, the earliest farmers used data in the form of notes of observations. In the current age of data, it has become easier to collect data over a wide spectrum of parameters. There are numerous sensing technologies for measuring processes on and over the surface of a field, typically mounted on satellites, aerial vehicles (drones), ground vehicle and static platforms. Recently, soil has been gaining increasing attention and recognition for its significance in not only increasing productivity but also stabilizing the environment. However, characterizing soil in a field is not trivial, especially when required to access the deeper layers and quantifying the essential contents – water, nutrients and organic matter.**

This paper presents a short review of applications of ground penetrating radars (GPR) in measuring soil content and structure. The focus is on deep learning constructs that have been used for interpreting and establishing correlations in the characterization of soils. The review serves to inform design considerations for a planned autonomous rover that will be used for surveying field soils in the Satakunta region of Finland. After the review, a conceptual design of an autonomous GPR rover is presented.

Keywords - autonomous rover, ground penetrating radar, soil, deep learning

I. INTRODUCTION

The amount of agricultural data that is collected is increasing continuously and big data analytical techniques are central in extracting information from them [1]. A growing density of satellites deployed with special and varied sensors collect increasing amounts of data. Additionally, there are ever increasing instruments and farm machines that are equipped with various sensors. Some of these data are openly available, some are commercial, and some are farmers' own private data of the farm. The challenge is to find a way for the farmers to utilize these different sources effectively and put them into practice at the farm.

A. MIKÄ DATA project

In the MIKÄ DATA project, which belongs to the EIP-AGRI group and funded by the European Agricultural Fund for Rural Development, started to build intelligent data service for the farmers [2]. The Data oriented field management will help farmers to do more accurate growing planning. The project focused on the variations of soil and nutrients for visualizing and developing of the algorithms of yield forecast based on convolutional neural networks (CNN) technique [3]. Data was collected from 100 hectares, using Veristech company's a special soil electrical conductivity, optical, and pH scanning equipment, a drone with a multispectral camera and laboratory testing of soil samples. Significant and promising results to utilize new technologies like artificial intelligence techniques (AI) were made in the project. The large data amounts from the fields mean that they must be more effectively analyzed with the help of the AI. It is not possible for the farmers to analyze all that data.

B. Data steps of field health

The health of a farm field is a vague concept that we want to be able to characterize more exactly. The field's behavior regarding water is of central importance. Often in the discussions with farmers, a starting point has been evaluating surface gradients of a field. Ideally, water should not remain standing on the field, neither should it flow too quickly along the surface. Tile drainage setups are often used to solve water logging problems, but it doesn't fix all moisture related problems. So, it is essential to get exact information of field surface gradients, so farmers can make the necessary corrections. Second essential thing is to know the location of the existing subsurface drainage (if there are no exact maps available or they are not even found anywhere) and its functionality. Sentinel-2 satellite data can used to get moisture maps which can help to evaluate the moisture variation inside the field and find moisture problems. A drone with a hyperspectral camera can collect even more accurate moisture information, but it requires expensive special equipment. In both cases only surface moisture, not inside of soil, is measured.

C. PeltoAI and BioEväät projects

The determination of the condition of the field should begin with analyzing the surface gradients, subsurface drainage and moisture structure. These and many other following stages are studied in the PeltoAI project. In the parallel BioEväät project, corrective actions to problems that have been found will be made. The goal is to understand the functioning of the soil more exactly via data. In the near future, it is also more important to get more exact data to follow the amount of carbon as an additional

factor in field health as well as due to climate change related carbon sequestration considerations.

D. Paper objectives

At a very general level, the vegetation of the field needs water, heat, light and nutrients. At the next much more exact level, the soil type, the condition of drainages, the condition of the granular structure, pH-values and humus content need to be known. From the point of view of abovementioned projects, it is still unclear what is the order of importance. It is very difficult to measure all these values cost-effectively with a adequate precision in a practical way. The motivation of this paper was that with GPR one can locate drainages and their condition and the soil compaction values, and data analyses can done using artificial intelligence paradigms.

The objective of this paper is to survey the agricultural applications where GPR has been employed and highlight successful soil applications. Importantly, the radargrams produced by GPR's are not trivial to interpret and often require expert inputs. With the proliferation of deep learning methods developed for the tasks of extracting and interpreting information from big data, our interest is in employing such frameworks to work with GPR data.

II. Ground-penetrating Radars

Ground-penetrating Radar (GPR) is an established technique that has been utilized also in the analysis of the condition of farm fields. GPR uses electromagnetic radiation and microwave band, which range are 10 MHZ to 2.6 GHz. GPR data is very open to various interpretations because of its analysis have been made by hand, which is laborious, particularly in the big data amounts.

A. GPR in agriculture

The GPR technique has been researched and utilized in farm fields in many ways. It has the advantage of being a non-invasive method of gathering data from below the soil surface. Illawathura et al. has studied soil moisture with GPR data [4]. The drainage pipes have been examined for example by Allred et al. [5], but drainage pipes can also be found with an unmanned aircraft [6] and with thermal infrared imagery [7]. De-Ville et al has studied soil compaction [8]. Akinsunmade et al. [9] has studied the correlations of agrotechnical properties of the different soil more widely. New interesting research openings are the measurement of the fine roots of agricultural crops [10]. With the measurement of the roots, the evaluation of the amount of the carbon would be achieved. Shen et al. [11], did quantification for soil organic carbon with GPR. The roots change to carbon during time and improve soil that way, depending on soil type and depending on the cultivation methods. The carbon will improve the growth and condition of the field.

Thus, it appears that GPR has widely been used in agriculture for moisture measuring, soil stratigraphy, layers' thicknesses and calculating root biomass [12].

Interestingly, Rodriguez et al. mounted a GPR measuring device mounted on a drone to measure soil moisture [14].

Surprisingly our survey did not find any study which uses the deep learning framework for the analysis of GPR data in the agriculture sector.

III. Review: Artificial Intelligence for GPR data

The table below lists the papers containing the GPR and AI terminology. The findings of papers are briefly discussed below.

TABLE I. STUDIES THAT EMPLOY ARTIFICIAL INTELLIGENCE TO ANALYSE GPR DATA.

Paper	Target	Data	AI-tech	Year
[15]	Layer properties: thickness, permittivity	gprMax	ML, NN	2009
[16]	simulate GPR data	A-scan, gprMax	NN	2018
[17]	cover depth and diameter of rebar, moisture content	simulated	ML, forward solver, 120 layers	2019
[18]	corrosion, infrastructure monitoring	B-scan, gprMax	DL, u-net, cgan, encoder-decoder	2018
[19]	Bridge deck evaluation (thickness, moisture, rebar locations, corrosion)	many GPR equipments	HOG features	2016
[20]	detection of hidden surface crevasses on glaciers	GPR images from Antarctica	Support vector machine	2012
[21]	objects	B-scan, gprMaxToolBox, Citar-10database	Faster-RCNN	2018
[22]	objects	GPR	support vector ML and H-alpha decomposition	2018
[23]	landmines	B-scan	CNN	2018
[24]	Water content and its prediction	A-scan	CNN	2019
[25]	Small targets, Holographic subsurface penetration	?	DL, SR-CNN, bicubic, PDM, WIM, NSRCNN	2017
[26]	mines	UWB-SAR, simulated SAR images	CNN (two subnetwork decomposition, classification)	2018
[27]	debonding, pavement monitoring	A-scan	Support vector ML	2017
[28]	landmines	L-band	DL, CNN	2019
[29]	objects, tightness	rail	Wavelet NN, De-noising, 3 layers feedforward NN	2015
[30]	underground utilities	B-scan/A-scan	not AI yet	2018
[31]	pipes, cables and those track course	B-scan? Data generator, Ultra wideband GPR	ML	2018
[32]	landmines	B-scan, real-data	CNN	2017
[33]	deep and shallow crevasses	GPR dataset	ML, HOG	2019
[34]	objects	FDTD-simulated	DL, 9 layers CNN	2018
[35]	Explosive objects	A- and B-scan, test data: Grayscale images	CNN	2017
[36]	objects	Ultrawideband	CNN, 3 layers	2019
[37]	objects	3D, including B,C and D-scan	DL underground object detection, triplanar DCNN	2019

A. Targets

In the above table I, in the target-column, are target areas identified in each research paper. An exact target was not identified in some of the papers. In those cases the column is labelled generally just "objects". Many different targets were found. The most essential findings, from the point of view of the developing of the robot, were the moisture and the tightness. The measurement of the moisture with the help of GPR is an interesting application which has surely an advantage in the agriculture. With the help of this technique a more exact understanding of the moisture structure of the whole field would be obtained. The measurement of the tightness is an important finding because it can be used to estimate the excessive tightening of the field which affects the absorption of the moisture and limits the growth of roots. Furthermore, there were the papers which focused on finding drainage pipes and their condition, but an AI had not been utilized.

B. Data

Many of the researched papers utilized simulated data. The gprMax tool, which is open source software, was the most superior of these. It can be used to produce the GPR pictures without real equipment. This makes the data production very quick.

The data generated with gprMax tool is already classified. It is essential that the data is classified and high quality correct so it can then be used as learning material of AI.

C. AI

A large amount of work of deep learning applications on GPR data has been in the detection of buried objects – ranging from utility pipes to mines (munitions). The primary task in such applications is classification and not a recreation of images from the radargram. Image synthesis networks have been used in other fields and Alvarez & Kodagoda [18] have explored its application to recreating sub-surface permittivity maps from b-scans. They have also hypothesized on the usage of LSTMs in combination with their generative network to leverage the sequential nature of the scans. Detecting objects is not the primary concern in soil characterization apart from the detection and tracing of sub surface drains in the fields. Of greater interest are – water and roots.

IV. AUTONOMOUS GPR ROVER: CONCEPTUAL DESIGN

Different kinds of mobile robots/rovers are more generally used for data collection in a variety of environmental applications, as the robot is seen as a promising tool for enhancing environmental data collection. Our goal was to design all-terrain rover, which is based on custom-made chassis, various sensors (GPS, GPR, LiDAR and RGB-D cameras) and ROS (Robot Operating System) as software which connects all hardware/software-connections. Prototype rover model was created in Solid Works modelling software, which can be seen Fig. 1.



Figure 1. Prototype rover model.

A. Hardware

The body of the rover is made and welded of structural steel. The model has a diameter of 1200 mm x 800 mm x 500 mm. The tires are designed for off-road vehicles. Like all-terrain vehicles, the rover also uses steering and suspension.

For drive motors, the vehicle uses 48V 1kW servomotors in the gearbox for every 4 tires. A smaller 300W servo motor is used to control the vehicle. The voltage system was chosen to be 48V to keep the current proportional efficiency low, but in the area where ready battery systems are available, the voltage is within the safe voltage range. The battery was selected for reliability and voltage near the motors rated voltage. The LG Chem RESU Lithium-ion battery system has IP55 protection and is intended for outdoor use. It has a capacity of 13 kWh to operate the vehicle for approximately 2.5 hours with all engines running at full speed.

The safety system consists of an industrial class safety relay and an overlapping safety circuit. The circuit consists of an emergency stop button and a safety buffer at the front and rear. The rover can also be stopped by remote control, but it is not permanently connected. The safety relay switches off the main power of all motors and actuators and sends a remote alarm. The operation can only be resumed when the safety relay is reset by pressing a button physically located in the rover.

There were a couple of options for computer hardware, either switching to ARM-based single-board computers (SBC), such as Raspberry Pi or Jetson Nano, or alternatively using x86-based computers. For each of these options, there are examples of robots and easy-to-assemble ROS packages. In our experience, single-board computing power is still limited when running a full on-board navigation and observation stack, and cloud-based systems still have the same problems as latency.

The main reason to go for single-board computers is power consumption, which is notably smaller than x86-based computers. If single-board computers use 20-40W power, a full-size x86 computer with medium-sized GPU consumes 300-500W. Other reason is the cost. Where Raspberry Pi can be bought for 40 euros and Jetson Nano for 100 euros, the full x86 computer setup with battery powered power supply and the average consumer grade GPUs are about 500-1000 euros. In this case, with a planned platform cost is 20,000 euros, so the price difference is not the main consideration. The drive motors are 1 kW and at peak power, with all engines running at full power, the power consumption is 4 kW while running, so the power consumption of the x86 computer is not a major limitation of the application. The connections between sensors, computer, main and safety circuits can be seen in Fig. 2.

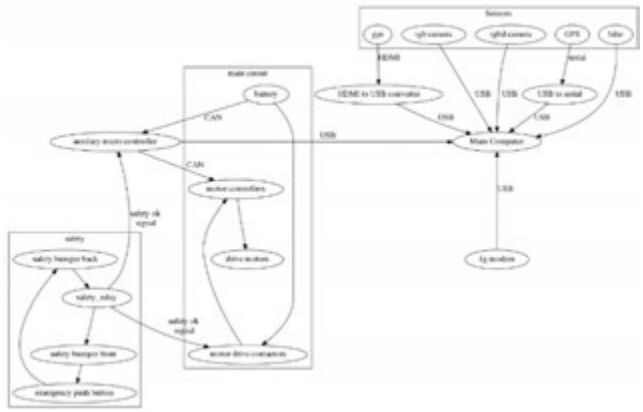


Figure 2. Connections between components.

B. Software

ROS was chosen as the software platform, as there are already several implementations of ROS-based mobile robots in both research and production. Many open source software components and drivers have already been implemented in ROS. Another reason was authors previous experience about the ROS system. The ROS system is modular in nature and consists of different nodes, and communication between nodes takes place through topics, services and actions.

Localization is divided into two scales, global and local. Globalization is done by GPS, and this is used more for mission-level navigation. Localization is used to detect unknown terrain and obstacles. This is done with Depth Cameras (RGBD), LiDAR and the Inertia Measurement Unit (IMU). The rover monitors the tilt of the IMU to handle the tilt on inclines. The biggest challenge here is isolating the driven vegetation, but the LiDAR may seem like a solid barrier, as the vehicle in question probably needs a camera and a neural network-based object detection. For new environments, data must be collected, and the identification network trained several times before a proper generalization can be achieved. Global location senses are fused with the robot_localization package, which is a universal sensory fusion package that includes an extended Kalman filter and non-scented Kalman filter nodes.

The controller is using an open source implementation of a single robust integral of sign the error (RISE) algorithm developed at the University of Florida [38]. The path designer calculates the desired direction and velocity from the material position and the desired position. The engine mapping calculates the required forces for different engines and can tolerate engine failures of 1-2 depending on faulty engines. Missions are performed asynchronously, and multiple tasks can be executed simultaneously. One task always detects and avoids dangerous objects. One checks the vehicle's stability and slows down the vehicle if the posture is unstable.

GPR data is captured from HDMI image using HDMI to USB capture card. The capture card appears on the Linux system as a webcam and is connected to the ROS via the usb_cam package. Analysis of GPR data can be done in almost real-time in rover or logged for offline use. Data is logged in rosbag, which is a handy ROS tool that can store messages on the screen so all sensory data and locations can be played back in order or used for analysis. The Rosbag video stream can also be easily cut into image sequences, and time stamps and image numbers can be converted to a csv file. The vehicle can also store GPR and RGB images from the same location for analysis. Connections can be seen in Fig. 3.

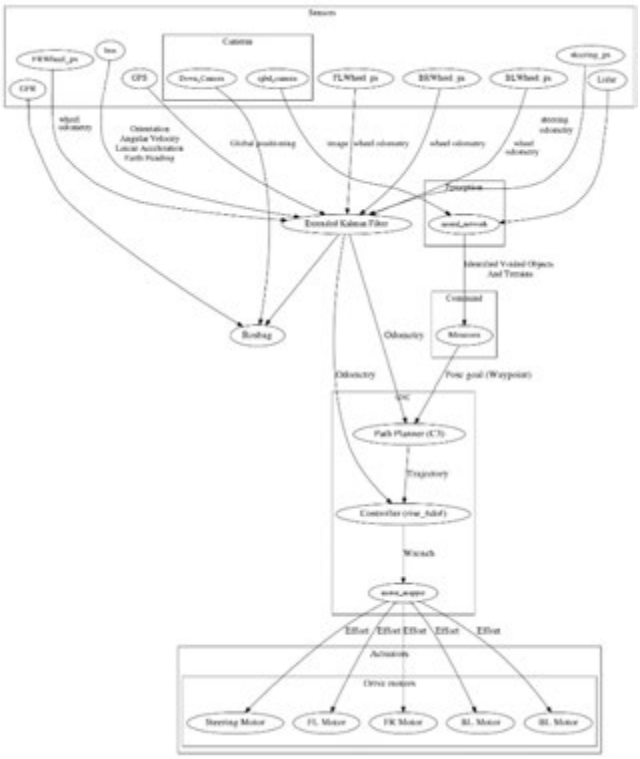


Figure 3. Connections between software components.

V. CONLUSION

In this paper, first described many GPR using targets in agriculture, the most important them were moisture and tightness measurements. This research finds out that data can be collected also by simulating like with gprMax tool. Classified and large amount of data is very important for CNN development and simulated data will have remarkable help for that work. In this paper was made short review of over twenty papers included GPR and AI keywords. So far, there wasn't any papers from agriculture sector including CNN technique, so this sector will need deeper paper search and research. According to this review, we noticed that CNN technique is not very largely used in GPR data.

In this paper a ROS robot architecture was also described. The goal is to develop a fully functional robot

and use it in the fields next year. Before the robot is ready, CNN can be tested with simulated data.

These findings will help our projects (PeltoAI and BioEväät) to understand field health more deeply and to recommend improvement actions for farm fields if needed.

A very interesting finding was that drones are a potential vehicle to carry GPR equipment and enable making rapid field measurements. Second finding, root scanning with GPR technology, will help understand the field more deeply and assist in climate change control, because there is also a need to evaluate the amount of biomass in the ground, not just over the field surface.

REFERENCES

[1] S. Ge. L. Wolfert, C. Verdouw and M.-J. Bogaardt, "Big Data in Smart Farming – A review" 2017, Agricultural Systems, 153, 69–80.

[2] P. Linna, N. Narra and J. Grönman, "Intelligent data service for farmers," 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 1072-1075.

[3] N. Narra, P. Nevavuori, P Linna and T Lipping, "A Data Driven Approach to Decision Support in Farming", EJC 2019, pp. 175-185

[4] Illawathura, Chameera, Parkin, Gary, Lambot, Sébastien, Galagedara, Lakshman, "Soil Moisture Estimation from GPR Hyperbola Fitting in the Agricultural Root Zone". CSBE/SCGAB AGM and Technical Conference 2019 (Vancouver, British Columbia, Canada, du 14/07/2019 au 17/07/2019). In: Book of Abstracts, 2019

[5] Allred, B., Wishart, D., Martinez, L., Schomberg, H., Mirsky, S., Meyers, G., Elliott, J., Charyton, C., "Delineation of Agricultural Drainage Pipe Patterns Using Ground Penetrating Radar Integrated with a Real-Time Kinematic Global Navigation Satellite System". Agriculture 2018, 8, 167.

[6] Robert Freeland, Barry Allred, Neal Eash, Luis Martinez, DeBonne Wishart, "Agricultural drainage tile surveying using an unmanned aircraft vehicle paired with Real-Time Kinematic positioning—A case study", Computers and Electronics in Agriculture, Volume 165, 2019, 104946, ISSN 0168-1699

[7] https://www.ocj.com/2018/07/using-drones-to-find-drainage-pipes/, refered 21. february 2020.

[8] De-Ville S. Pattison I. and Frost M., "COMPACT: A frequency analysis of Ground Penetrating Radar to identify agricultural soil compaction.", AGU Fall Meeting Abstracts, 2018, eid NS13B-0596

[9] Akinsunmade, A., Tomecka-Suchoń, S. & Pysz, P. Correlation between agrotechnical properties of selected soil types and corresponding GPR response. Acta Geophys. 67, 1913–1919 (2019).

[10] Liu, X., Dong, X., Xue, Q. et al. "Ground penetrating radar (GPR) detects fine roots of agricultural crops in the field". Plant Soil 423, 517–531 (2018).

[11] Shen, X., Foster, T., Baldi, H., Dobreva, I., Burson, B., Hays, D., Tabien, R., Jessup, R., "Quantification of Soil Organic Carbon in Biochar-Amended Soil Using Ground Penetrating Radar (GPR)". Remote Sens. 2019, 11, 2874.

[12] Katerina Zajícová, Tomas Chuman, "Application of ground penetrating radar methods in soil studies: A review", Geoderma, Volume 343, 2019, Pages 116-129, ISSN 0016-7061

[13] Triven Koganti, Ehsan Ghane, Luis Rene Martinez, Bo V. Iversen and Barry J. Allred, "Mapping Subsurface Drainage in Agricultural Areas Using Unmanned Aerial Vehicle Imagery and Ground Penetrating Radar ", 2019, Indian Near Surface Geophysics Conference"

[14] Wu, K., Rodriguez, G. A., Zajc, M., Jacquemin, E., Clément, M., De Coster, A., & Lambot, S., "A new drone-borne GPR for soil moisture mapping", 2019, Remote Sensing of Environment, 235, 111456.

[15] S. Caorsi and M. Stasolla, "A machine learning algorithm for GPR sub-surface prospection," 2009 Mediterrannean Microwave Symposium (MMS), Tangiers, 2009, pp. 1-5.

[16] I. Giannakis, A. Giannopoulos and C. Warren, "A Machine Learning Approach For Simulating Ground Penetrating Radar," 2018 17th International Conference on Ground Penetrating Radar (GPR), Rapperswil, 2018, pp. 1-4.

[17] I. Giannakis, A. Giannopoulos and C. Warren, "A Machine Learning-Based Fast-Forward Solver for Ground Penetrating Radar With Application to Full-Waveform Inversion," in IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 7, pp. 4417-4426, July 2019.

[18] J. K. Alvarez and S. Kodagoda, "Application of deep learning image-to-image transformation networks to GPR radargrams for sub-surface imaging in infrastructure monitoring," 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), Wuhan, 2018, pp. 611-616.

[19] P. Kaur, K. J. Dana, F. A. Romero and N. Gucunski, "Automated GPR Rebar Analysis for Robotic Bridge Deck Evaluation," in IEEE Transactions on Cybernetics, vol. 46, no. 10, pp. 2265-2276, Oct. 2016.

[20] R. M. Williams, L. E. Ray and J. H. Lever, "Autonomous robotic ground penetrating radar surveys of ice sheets; Using machine learning to identify hidden crevasses," 2012 IEEE International Conference on Imaging Systems and Techniques Proceedings, Manchester, 2012, pp. 7-12.

[21] M. Pham and S. Lefèvre, "Buried Object Detection from B-Scan Ground Penetrating Radar Data Using Faster-RCNN," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, 2018, pp. 6804-6807.

[22] H. Zhou et al., "Combination of Support Vector Machine and H-Alpha Decomposition for Subsurface Target Classification of GPR," 2018 17th International Conference on Ground Penetrating Radar (GPR), Rapperswil, 2018, pp. 1-4.

[23] F. Picetti, G. Testa, F. Lombardi, P. Bestagini, M. Lualdi and S. Tubaro, "Convolutional Autoencoder for Landmine Detection on GPR Scans," 2018 41st International Conference on Telecommunications and Signal Processing (TSP), Athens, 2018, pp. 1-4.

[24] J. Zheng, X. Teng, J. Liu and X. Qiao, "Convolutional Neural Networks for Water Content Classification and Prediction With Ground Penetrating Radar," in IEEE Access, vol. 7, pp. 185385-185392, 2019.

[25] Y. Zhang, Z. Xiao, L. Wu, X. Lu and Y. Wang, "Deep learning for subsurface penetrating super-resolution imaging," 2017 10th UK-Europe-China Workshop on Millimetre Waves and Terahertz Technologies (UCMMT), Liverpool, 2017, pp. 1-4.

[26] T. H. Vu, L. Nguyen, T. Guo and V. Monga, "Deep network for simultaneous decomposition and classification in UWB-SAR imagery," 2018 IEEE Radar Conference (RadarConf18), Oklahoma City, OK, 2018, pp. 0553-0558.

[27] S. S. Todkar et al., "Detection of debondings with Ground Penetrating Radar using a machine learning method," 2017 9th International Workshop on Advanced Ground Penetrating Radar (IWAGPR), Edinburgh, 2017, pp. 1-6.

[28] F. Giovanneschi, K. V. Mishra, M. A. Gonzalez-Huici, Y. C. Eldar and J. H. G. Ender, "Dictionary Learning for Adaptive GPR Landmine Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 12, pp. 10036-10055, Dec. 2019.

[29] Xiao-Li Chen, Mao Tian and Wen-Bing Yao, "GPR signals de-noising by using wavelet networks," 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 2005, pp. 4690-4693 Vol. 8.

[30] E. Skartados et al., "Ground Penetrating Radar Image Processing Towards Underground Utilities Detection for Robotic Applications," 2018 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), Prague, Czech Republic, 2018, pp. 27-31.

[31] D. Seyfried, A. Busche, R. Janning, L. Schmidt-Thieme and J. Schoebel, "Information extraction from ultrawideband ground penetrating radar data: A machine learning approach," 2012 The 7th German Microwave Conference, Ilmenau, 2012, pp. 1-4.

[32] S. Lameri, F. Lombardi, P. Bestagini, M. Lualdi and S. Tubaro, "Landmine detection from GPR data using convolutional neural networks," 2017 25th European Signal Processing Conference (EUSIPCO), Kos, 2017, pp. 508-512.

[33] B. Walker and L. Ray, "Multi-Class Crevasse Detection Using Ground Penetrating Radar and Feature-Based Machine Learning," IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 2019, pp. 3578-3581.

[34] J. Sonoda and T. Kimoto, "Object Identification form GPR Images by Deep Learning," 2018 Asia-Pacific Microwave Conference (APMC), Kyoto, 2018, pp. 1298-1300.

[35] D. Reichman, L. M. Collins and J. M. Malof, "Some good practices for applying convolutional neural networks to buried threat detection in Ground Penetrating Radar," 2017 9th International Workshop on Advanced Ground Penetrating Radar (IWAGPR), Edinburgh, 2017, pp. 1-5.

[36] M. Elsaadouny, J. Barowski and I. Rolfes, "The Subsurface Objects Classification using a Convolutional Neural Network," 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 2019, pp. 0874-0877.

[37] N. Kim, S. Kim, Y. An and J. Lee, "Triplanar Imaging of 3-D GPR Data for Deep-Learning-Based Underground Object Detection," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 11, pp. 4446-4456, Nov. 2019.

[38] N. Fischer, D. Hughes, P. Walters, E. M. Schwartz and W. E. Dixon, "Nonlinear RISE-Based Control of an Autonomous Underwater Vehicle," in IEEE Transactions on Robotics, vol. 30, no. 4, pp. 845-852, Aug. 2014. doi: 10.1109/TRO.2014.2305791

Increasing the Accuracy of Robotic Neurosurgical Procedures Through Robot Calibration

L. Drobilo*/**, M.Švaco*, B. Jerbić*
* University of Zagreb / Faculty of Mechanical Engineering and Naval Architecture, Zagreb, Croatia
** University hospital Dubrava / Department of Neurosurgery, Zagreb, Croatia
luka.drobilo@fsb.hr

Abstract – **One of the main concerns in neurosurgical procedures, besides safety and achieving sterile environments, is procedure accuracy. Although robots are well known for their reliability, absolute accuracy has always been an issue, especially for configurations with serially linked revolute joints. This issue can be addressed through various localisation and positioning strategies which significantly increase procedure accuracy but also increase overall procedure duration and can never achieve the full potential of the system. To address this issue robot calibration is performed to better fit the model used for positional calculations to the robot. In this article an experimental setup based on the neurosurgical robotic system RONNA is presented. A robot tool with highly precise sensors combined with a precisely measured reference phantom is used to perform measurements as well as for validation, and robot parameters are calibrated using a Sequential Quadratic Programming-based algorithm. A method for implementing the calibrated model into the RONNA workflow is presented and positioning accuracy achieved using the calibrated model is compared with a model obtained through the commercial calibration software RoboDK. Validation results show positioning accuracies comparable to the commercially calibrated model, with additional improvements yet to be implemented into the model and calibration setup. Experimental testing shows a potential for decrease in localisation time and increase in overall procedure accuracy of neurosurgical procedures, which can be very beneficial, especially for more complex procedures.**

Keywords – robot calibration; RONNA; medical robotics; neurosurgery, accuracy

I. INTRODUCTION

Neurosurgery is regarded as one of the most challenging fields of medicine, due to the highly sensitive nature of the human brain. Planning and performing a neurosurgical operation requires utmost precision and concentration, since even the smallest error can lead to life threatening complications. Besides avoiding parts of the brain critical for various bodily functions, great care must be taken to avoid the large number of blood vessels going through and around the brain. Besides safety reasons, certain procedures target brain cores that are very small and difficult to reach or require precise electrode and/or probe placement for a successful outcome of the procedure. All of this emphasizes utmost accuracy as one of the primary requirements for neurosurgical procedures.

Minimally invasive surgery holds great advantages compared to classic open surgery and those advantages are especially pronounced in neurosurgery. A great challenge for this type of surgery is enabling the surgeon to navigate to a precisely defined spot with only a drill hole or small trepanation opening to access it. To enable vision and localization of the brain without opening the skull, CT imaging is used in combination with external physical references visible both in the CT imaging and to the surgeon. Before the procedure, the surgeon marks two points in the CT image: one defining the entry point where the skull is going to be drilled, and the other defining the target point which the surgeon wishes to reach. These two points define a trajectory through which the surgical tool will reach the target point and depending on the type of surgery there can be one or several trajectories. The positions of these points, as well as positions of external reference markers, also visible in the CT imaging are then recorded as the surgical plan, which is used during the procedure to navigate the surgeon.

A widely used tool, enabling the most demanding modern neurosurgical procedures, are stereotactic frames fixed to the patient's skull through fixation screws with sharp conical tips, providing a stable reference to the CT imaging data and precise guidance platform for the surgical tools [1]. They are a state-of-the-art solution enabling the most precise neurosurgical procedures performed today, however they also come with several disadvantages like being bulky and uncomfortable for the patient, requiring a lot of manual work by the surgeon introducing more room for human error, and being impractical for imaging and procedure planning just to name a few. An alternative to stereotactic frames, enabling more flexibility, reducing the strain both on the surgeon and the patient, as well as bringing other benefits are surgical robots [2]–[5]. RONNA [6]–[8] is a surgical robot designed for stereotactic neurosurgical procedures (third generation setup is presented on Fig.1) while the system is currently preparing for upgrade to the fifth generation. It replaces the stereotactic frame with a set of small bone screws with screw-on reference markers, which can be removed when the patient is not undergoing a CT scan or surgery, significantly lowering the discomfort level for the patient and easing the scheduling and logistical operations of the hospital. At the same time, since all the positioning is performed automatically by the robot, much less pressure and fatigue is exerted on the surgeon, leaving him more focused on the medical aspects of the surgery instead of



Figure 1 Render of the RONNA G3 system with components: (1) Master robot, (2) Assistant robot, (3) Universal platform, (4) Optical tracking system , (5) Control and planning software interface [8]

performing unnecessary mechanical adjustments. By introducing the robot into the operating room an additional increase in flexibility can also be achieved, as well as a drastic reduction in procedure duration for more complex procedures requiring several trajectories [3].

Although robots are well known for their reliability, absolute accuracy has often been an issue, especially for serially linked robot arms. This is caused due to the serial linkage of revolute joints, increasing deviations present in each joint and link exponentially along the kinematic chain from the base of the robot to the end effector tool. In neurosurgical applications this issue is addressed by applying tools and strategies to precisely localize the patient in the robot workspace. The localisation setup used for the RONNA system is presented in Fig.2 [9]. An infrared (IR) stereovision camera (OTS frame) is used to precisely position the centre of the camera tool (TCP frame) on each of the external reference markers located on the patient's head ($y_1 \cdots y_4$ points). This is done by localising the camera markers (M frame) and connecting them using the transformation to the TCP frame. After being centred on each marker the robot tool centre point (TCP) coordinates are recorded and used to corelate with marker positions in the CT imaging data (CT frame). This procedure is called registration and is used to find the best

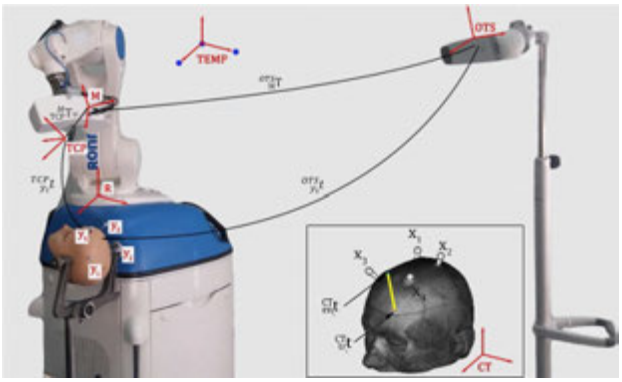


Figure 2 RONNA localization setup with coordinate frames used in the localization procedure [9]

positional fit between the markers in physical space and CT imaging space so the planned trajectory information can be transferred to the robot's coordinate system for tool positioning. This procedure is performed for each of the planned trajectories, rotating the camera so that the axis of the robot tool is aligned with the trajectory during the localisation of each of the markers. This way only a small change in robot configuration needs to be made to position the surgical tool into the required position, introducing a minimal error from the deviations in the robot model. However, this approach has a cost in time since this type of correction must be performed for each of the trajectories if the error wishes to be minimized. For a brain biopsy with usually only one trajectory this is not an issue, but e.g. in a Stereo Electroencephalography (SEEG) where several electrodes are implemented in one surgery the difference in time required for localisation is quite significant [5].

The main causes of the error are imperfections present in the robot mechanism, i.e. discrepancies between the robot mechanism and the model used for positional calculations and control. To address this issue robot calibration is performed to fit the kinematic and additional non-kinematic parameters of the model to the robot as close as possible. This way the number of localisations can be reduced, and positioning accuracy increased compared to the non-calibrated model. Three calibration levels are defined according to the source of the error [10]: first level errors are caused by poor joint mastering, i.e. deviation of the joint encoder zero position from the zero position of the joints; second level errors are kinematic errors – caused by geometric deviations in the robot links and joints (usually due to imperfections in part manufacturing and assembly) and third level are non-kinematic errors caused in varying degree by material elasticity, effect of gravity, thermal dilatations, excessive bearing clearance, gear backlash, etc.

There are many approaches to serial robot calibration [11], however in this article only model-based calibration approaches shall be discussed, which can further be divided into open loop and closed loop approaches. Open loop approaches [12]–[18] are more flexible, allowing the user to cover a much larger portion or the entirety of a robot's workspace with the measurements, however a downside is the required expensive and highly precise measurement equipment, most often laser trackers [10], [13], [15], [17], [18]. Closed loop calibration [19]–[23] is much simpler and uses the robot's internal joint position measurements combined with additional sensors and/or measurement tools, as well as external geometric references to perform the required measurements. These solutions range from very simple and cost efficient, to more complex and expensive. [19]–[23] Collected data is then run through an optimization algorithm fitting the robot parameters to the real-life robot as best as possible and reduce the deviation from the measured references as much as possible. Most often least squares [11], [12], [15], [19], Nelder-Mead [19], Levenberg-Marquart [17], [21] or genetic algorithm-based [11] algorithms are used.

In section II robot calibration procedure is presented in more detail, including the experimental setup, main components of the calibration model, and the optimisation and verification procedures. Section III explains the integration of the calibrated model into the RONNA

workflow. Finally, experimental results are analysed in chapter IV and conclusions presented in chapter V.

II. ROBOT CALIBRATION

Four main elements are required to perform robot calibration: a robot model, experimental setup to perform the necessary measurements, a parameter optimization method, and a validation method.

A. Robot Model

The most widely used kinematic definition of a robot mechanism is the approach defined by Denavit and Hartenberg [24] through the Denavit-Hartenberg model (DH) or one of its modifications. It describes the geometric configuration of the robot using a series of rotations (Rot) and translations ($Trans$) around/along the X and Z axis of the coordinate frame of each of the robot links. Besides the original definition introduced by Denavit and Hartenberg, newer modifications of the method also exist and depending on the method used (original DH definition or one of Modified Denavit-Hartenberg definitions – MDH) a different transformation sequence is used and/or an additional parameter(s) added. In this case an MDH model based on the formulation in [25] is used applying the following coordinate frame transformation for each of the links:

$$\boldsymbol{T}_i = Rot(\boldsymbol{x}, \alpha_i) \rightarrow Trans(\boldsymbol{a}_i, 0,0) \rightarrow Rot(\boldsymbol{z}, \theta_i) \rightarrow Trans(0,0, \boldsymbol{d}_i) \quad (1)$$

By sequentially multiplying transformations for all of robot links the transformation of the arm mechanism is obtained. Pre-multiplying the transformation with the robot base transformation ($\boldsymbol{T}_{base}$) and multiplying it with the robot tool transformation ($\boldsymbol{T}_{tool}$) results in the complete solution of the robot kinematic chain as follows:

$$\boldsymbol{T}_{robot} = \boldsymbol{T}_{base}\, \boldsymbol{T}_1\, \boldsymbol{T}_2\, \boldsymbol{T}_3\, \boldsymbol{T}_4\, \boldsymbol{T}_5\, \boldsymbol{T}_6\, \boldsymbol{T}_{tool} \quad (2)$$

In addition to the MDH definition, which includes parameters necessary for first and second level calibration, six more stiffness parameters [$s_1 \cdots s_6$] are added. These parameters define the joint stiffness (or the lack of it) and provide joint position adjustments depending on the distribution of the robot mass in the workspace, reducing the error caused by the elasticity present in the joints. This is achieved by using a simplified version of the Newton-Euler iterative dynamic algorithm [25], omitting the dynamic elements and simplifying the formulation for the static case. Value of the stiffness parameter s_i for each joint is multiplied with the torque value τ_i calculated through the Netwon-Euler iterative algorithm, resulting in the offset correction of the joint angle $\Delta\theta_i$ for that specific joint configuration. The process of including the joint stiffness correction into the direct kinematic solution is shown in Fig.3.

B. Measurement

The experimental setup, shown in Fig.4, is based on the RONNA secondary platform, used for new application and

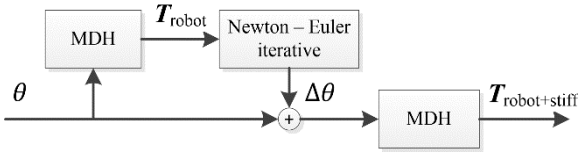


Figure 3 Implementation of the Newton-Euler iterative algorithm into the MDH direct kinematic model

tool testing before introducing them into clinical trials. A KUKA Agilus KR6 R900 sixx robot is mounted on the platform (instead of the UR robot for the secondary platform as depicted on Fig.1) and will be used as the basis for the calibration. The robot is equipped with a calibrating tool based on a tool first presented by Gaudreault et al. [20], with some additional modifications introduced, and is used for robot position referencing and measuring positional deviations. The tool, shown on Fig.5, has three precise Mahr Marcator digital linear indicators mounted on it, each with a resolution of 0.0005 mm and flat tips mounted on the indicators with the surface perpendicular to the measurement axis. Wireless emitters integrated into the indicators are used for establishing a wireless connection to the PC through a single wireless USB dongle and a zeroing tool is used for setting the zero position for the three indicators. The zeroing tool uses three small steel spheres (centres forming vertices of an isosceles triangle) for repeatable positioning onto the three V-grooves on the calibrating tool and attaching it through magnets placed under the V-grooves. Additionally, a steel sphere with 20 mm diameter and geometric quality grade G200 is mounted on the tool and used for zeroing the indicators, defining the tool TCP. The tool TCP frame is positioned in the same position (relative to the robot flange) as the point of the neurosurgical probe.



Figure 4 Experimental setup: RONNA platform, KUKA Agilus robot, calibration tool and phantom on a tripod

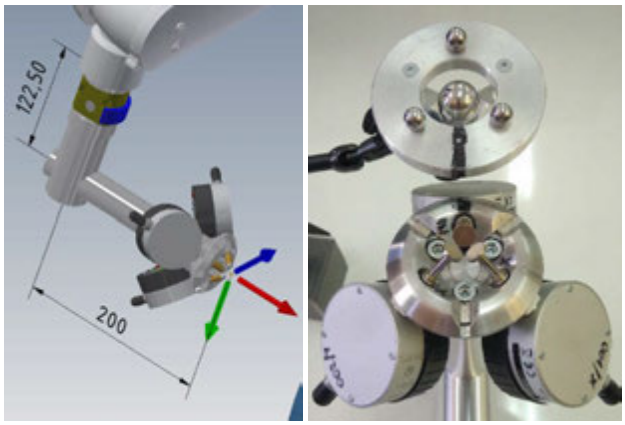


Figure 5 a) Calibration tool with TCP frame position
b) Calibration tool tip and indicator zeroing tool

Besides the calibrating tool, a reference phantom (Fig.6) is used both for collecting the measurements needed for calibration and for validating measurements after calibration. The reference phantom consists of 11 steel spheres, also with a diameter of 20 mm and G200 geometric quality class, with under 5 µm deviation both from the nominal diameter and spherical form. All the spheres are mounted on aluminium rods of varying lengths, fixed to an aluminium plate. Before the calibration procedure the phantom was measured on a Coordinate Measurement Machine (CMM), assuring high measurement accuracy, to determine precise positions of the 11 spheres with regards to each other and the four IR reflective markers mounted on the aluminium plate, which are used for coarse localisation of the phantom in the robot's coordinate system. The reference phantom is mounted on a stable base and is used as a fixed reference to acquire measurement data necessary to perform the calibration.

The calibration setup utilises a closed loop calibration approach with fixed measurement references and the robot itself as the measuring tool. Out of the eleven 20 mm steel spheres mounted on the phantom, three are chosen and used as measurement references and multiple random orientations around each of them are defined. The



Figure 6 Reference phantom with four steel marker spheres (red), seven target spheres (blue) and four IR markers for rough localization (green)

stereovision camera NDI Vicra is used for coarse localisation of the phantom, using the relative position of the IR markers mounted on the phantom to the ones present on the calibration tool and adding the difference the position of the robot TCP. The robot is then roughly positioned on the first sphere in the first orientation.

Using the information transmitted to MATLAB by the linear indicators, the TCP position is iteratively corrected until a combined deviation of the TCP from the sphere centre of 3 µm or lower is achieved. After centring the TCP withing the allowed tolerance, sensor states on each of the joints for the tool orientation and robot configuration are read and recorded. The robot is then moved into the next orientation and the process is repeated until joint states are recorded for all orientations and chosen spheres. Final measurement data for each of the orientations consist of the six joints states, the sphere identification number, and the sphere's X-Y-Z coordinates, as measured on the CMM.

All the robot cartesian and joint positions, as well as necessary positional corrections are calculated in MATLAB using the direct and inverse kinematic functions. Communication to the linear indicators is achieved through the MarComProf software. Reading the measured joint values from the robot, as well as all of the robot positioning and position corrections are performed online through RoboDK, a commercial software for robot simulation, programming and calibration.

C. *Parameter Optimisation*

If the robot mechanism, sensors, and reference spheres were ideal and a direct kinematic function is calculated for joint configurations recorded while orienting the tool around a single point in space, every calculated TCP value would be the same. However, since various imperfections and deviations are present in the kinematic chain from the robot and other components, the result will not be a single point, but a point cloud. Since it is known that the tool did in fact rotate around a single point in space, the goal therefore is to fit the robot parameters as close as possible to the real robot geometry to reduce this dissipation and bring the results as close as possible to a single point. Parameter optimisation is performed using MATLAB's *fmincon* function and its Active Set algorithm based on Sequential Quadratic Programming (SQP), solving Quadratic Programming subproblems at each major iteration. A function is defined which connects model parameters designated for calibration with measurement data and outputs a single value evaluating how the input parameters fit the robot, which will be minimised through *fmincon* iterations. Starting values of the parameters to be calibrated are also included in the initial *fmincon* call, as well as optional upper and lower limit values for each of the parameters. The starting values are the default MDH0 parameter values and allowed parameter deviations are set as ±1 mm for length parameters, ±0.5° for angle parameters and 0 to +0.1 × 10^6 Nm/rad for stiffness parameters.

Since no reference to the robot base is made, nor is it included in the measurement process, the transformation from the base to the first joint is removed and the first four MDH parameters (α_1, a_1, θ_1 and d_1), defining the relative

position from the robot base mounting flange to the first joint, will not be calibrated. Also, axes of second and third joints are parallel, causing a parameter redundancy, therefore, one of the d parameters of the parallel axes (d_2 or d_3) is also removed from the calibrated parameter set [22], [26]. The robot is mounted vertically with no significant load in the horizontal plane to stress the first joint, so the joint 1 stiffness parameter s_1 can also be excluded from the calibration, as can s_6 due to a small influence the weight of the tool has on the final joint. In total 21 parameter is selected for parameter optimisation.

Some closed loop approaches use a tool with previously known dimensions [26], while others include the tool data to the calibrated parameters set [20]. Using a previously precisely measured tool has a benefit of calibrating the parameters up to the tool flange (including all four link 6 parameters), enabling exchanging tools between several robots without the need to recalibrate them for each robot if an adequately repeatable tool interface is used. However, to simplify the process, since only one robot is used in this case, the tool will not be previously calibrated, and two of the four parameters defining the last link (θ_6 and d_6) will also be removed from the calibrated parameter set. The error present in these parameters will be included and corrected in the tool calibration when the tool is finally calibrated using the calibrated robot model.

First step in each call of the evaluation function for the parameter set of the $i-th$ iteration ($\boldsymbol{p}_i$), whose flowchart is presented in Fig.7, is tool calibration. Since only the TCP information is acquired through the point constraint and there is no measured information referencing the orientation, only the translational component of the tool transformation $tool_j$ will be considered and solved using least squares, with j marking the sphere. As model parameters change, the tool calibration will be different as well in each iteration, increasing the visibility of the effect the parameter change has on the result. Next, the calibrated tool and joint configurations recorded for each of the measured spheres ($\boldsymbol{\theta}_{set}$) are input into the direct kinematic function, calculating the TCP position for each recorded configuration, forming the $\boldsymbol{TCP}_{\text{set(j)}}$. Using the TCP values an average value $\boldsymbol{pt}_{\text{avg(j)}}$ is calculated, representing a point at the centre of the point cloud for each of the spheres. Distance from each of the points for the sphere in $\boldsymbol{TCP}_{\text{set(j)}}$ is calculated, forming the $\boldsymbol{\Delta TCP}_{\text{set(j)}}$. If only one point is used for measurement these deviations are then added together and used by the optimisation algorithm as the function value to be minimised. If more than one point is used, the CMM data for the measured spheres ($\boldsymbol{pt}_{\text{CMM}}$) is used to calculate relative distances between them and the same is also done using the average values of the point clouds $\boldsymbol{pt}_{\text{avg}}$. The two sets of values are then compared, and deviations added for each of the spheres. The sum of distance deviations for a sphere is then added to all the TCP deviations the orientations around that sphere. The same is done for all the measured points, i.e. point clouds around other spheres and finally all the deviations are added together forming the function value, i.e. the measure of parameter fitness.

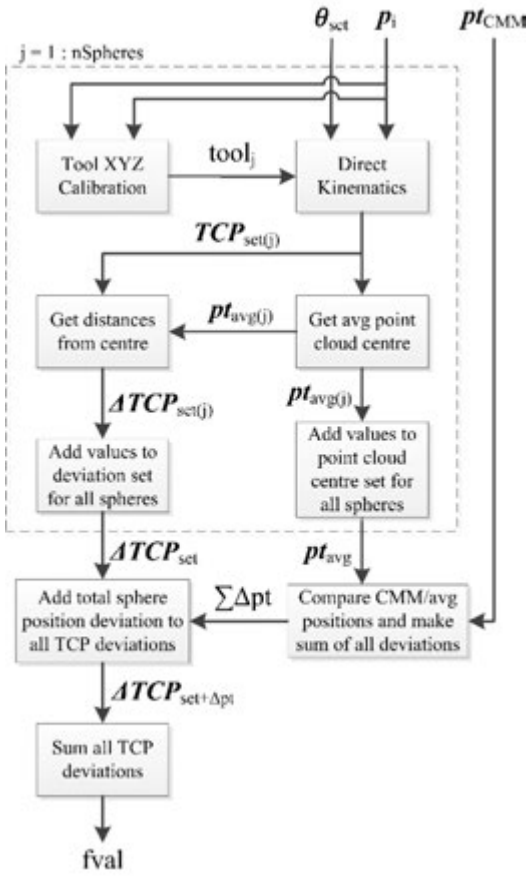


Figure 7 Diagram of the parameter fitness evaluation function

A number of numerical tests were performed to verify the optimisation procedure before performing the measurements: Here random parameter values were generated, as well as random orientations using the relative sphere locations as measured on the phantom. If only one sphere was used overall results were very good and almost all parameters can be fitted within 10^{-3} [mm,°,10^6 Nm/rad] from the reference, however parameters a_3 and d_4 would regularly deviate around 0.1 mm and a_2 and a_4 around 0.05 mm. If several spheres of unknown position were used the situation was not much different, however, if three or more spheres of known relative position were used all the parameters can be fitted within 10^{-3} mm to the reference value or even better.

After performing a number of simulations while varying the number of spheres and number of measurements per sphere used, it showed using three spheres and 15-20 measurements per sphere to be an optimal combination of measurements. If generated parameters had a highest deviation from a point of not more than 8 mm (which is a very large error for a robot of this size) and no noise was introduced into the simulation, difference of the acquired results from the reference would always be under 10^{-3} [mm,°,10^6 Nm/rad].

D. Application accuracy validation

After parameter optimisation is completed validation is performed to verify the effect calibrated parameters have in practice. In this case the same phantom is used, but this

time it is used to simulate neurosurgical localisation and positioning procedures as would be done during an operation using RONNA. Four spheres in the middle of the phantom (marker spheres) represent the IR reference markers that would be fixed to the patient's skull and the remaining seven (target spheres) represent target points defined by the surgeon. The calibration tool is used instead of the camera for marker localisation and also instead of the brain probe, adding the ability to precisely check the deviation of the probe TCP from the designated target point. The remaining four smaller grey markers are the IR markers used in this case for coarse localisation of the phantom, so the robot does not have to be guided manually to the four marker spheres.

After the phantom is located using the IR markers and NDI Vicra stereovision camera, the robot performs precise localisation of the four marker spheres using the calibration tool. The localisation data is then used for registration, entry and target points are calculated, defining the trajectories, as well as location and tool orientation for each of the target spheres. Phantom registration is performed using four markers whose relative positions in CT coordinates are fitted to the ones localised by the robot using Singular Value Decomposition (SVD). The robot is then moved into these positions using the calibrated parameters and the TCP deviation is recorded from the digital indicators. The phantom is then moved and/or rotated, and the procedure is performed for several positions to evaluate the effect calibration would have on the accuracy of the positioning during the procedure. This way application error can be evaluated, since both localisation and positioning are performed and measured in the robot coordinate frames, instead of using the coordinate system of an external measuring device.

Applying this type of high precision tool and reference, while also eliminating the CT as the measurement source, many errors usually present in a neurosurgical procedure can be eliminated from the setup. These can then be simulated or included separately as needed so their influence on the procedure can be examined. By using a CMM device instead of CT imaging, the error level is reduced to a micron level which is negligible for this type of use. The same goes for the tool with the digital indicators which also adds a rigid sub-micron precise tool instead of the flexible brain probe with additional clearances in the guides required for its use. Additional errors are excluded by using only one tool, instead of using a camera for localisation and probe guide for reaching the target.

III. IMPLEMENTING THE CALIBRATION

Ideally calibration should be implemented directly into the robot's control unit. However, since none of the manufacturers allows access to the kinematics definition it must be implemented separately. The most precise solution would be to calculate all the motions directly using the calibrated model, but this is also problematic due to synchronisation issues with the control unit. Even if that issue is overcome, part of the functionality of the control unit, which is certified, is now bypassed and additional certification will be required, which is even more complex when medical devices are in question.

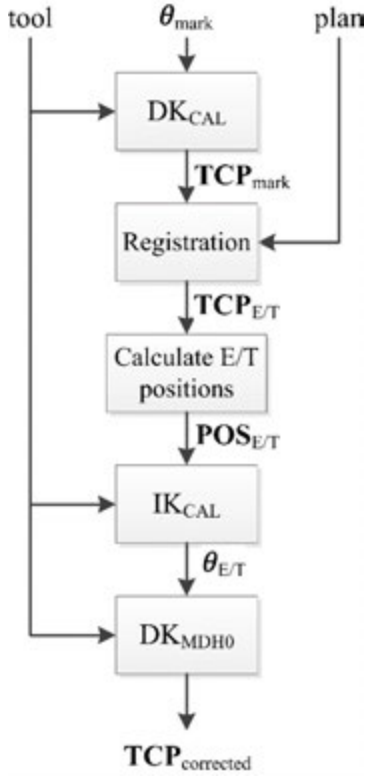


Figure 8 Diagram of the position correction process implemented for use in the RONNA workflow

To increase the precision without interfering with the control unit, while also causing a minimal impact on the RONNA workflow and software architecture, a simpler solution was selected. Instead of calculating all the positions and movements, only the critical ones will be corrected and then written in a way the control unit will understand them. As explained before, two points define a single trajectory: an entry and a target point. From those points the initial coordinate system for a trajectory is defined, the X-axis of the tool is fixed to the trajectory, and all other movements are defined from that coordinate system. This is important because once the robot enters a trajectory only linear movement along the trajectory is allowed, so calculating joint moves along the trajectory significantly complicates the issue. To assign the corrected positions to the robot, so called "corrected" points are used. The procedure is depicted in Fig.8.

When the marker spheres are localized robot's joint configurations $\boldsymbol{\theta}_{mark}$ for those spheres are recorded. These are then fed through the direct kinematic function using the calibrated model parameters (DK_{CAL}) to determine the $\boldsymbol{TCP}_{mark}$ coordinate set for all of the markers. The coordinates are then used to perform registration, i.e. a best fit between the points the robot has localized and the marker coordinates in the CT imaging (in this case CMM measured coordinates of the phantom). To perform the registration the operation plan defined by the surgeon is input into the registration function along with the marker coordinates. After registration, trajectory data

$\boldsymbol{TCP}_{E/T}$ (i.e. entry and target points) defined by the operation plan is transformed into the robot base coordinate frame. Tool positions, consisting of TCP and orientation are calculated for all entry and target points of each trajectory ($\boldsymbol{POS}_{E/T}$) and fed through the inverse kinematic function, also using the calibrated model parameters ($\mathrm{IK}_{\mathrm{CAL}}$). The result are joint configurations which can be used to position the robot precisely into entry and target positions of a trajectory ($\boldsymbol{\theta}_{E/T}$). Since programs in the robot controller are operated using tool positions, these joints are once more fed through the direct kinematic function, but his time using the default model parameters ($\mathrm{DK}_{\mathrm{MDH0}}$). Finally, the corrected TCP coordinates $\boldsymbol{TCP}_{corrected}$ are stored to be used for positioning into the trajectories by the robot controller unit. When performing position corrections this way a slight error is introduced whenever additional linear movements along the X-axis are introduced, however considering the movement length of a few millimetres and the minimal robot configuration change, the error introduced is negligible.

IV. RESULTS

The measurements were performed using the phantom described in section II. Three points were used on opposing sides of the phantom and 20 measurements for each of the points were made. Orientations around the spheres were randomly selected, however due to the position of the phantom relative to the robot and the position of the spheres relative to each other certain positions were unable to be reached for some of the spheres. The parameters were calibrated using the collected joint measurements and the sphere coordinates acquired through CMM measurements of the phantom. Deviations from the spheres using the initial parameters (MDH0) and measurements made around the three chosen spheres (60 measurements in total) resulted in a mean positional error of 0.816 mm and a maximum error of 1.843 mm. After calibration mean error was reduced to 0.101 mm and maximum to 0.225 mm, achieving eight times smaller deviations from target.

To validate whether these post-calibration results show a similar effect in an actual neurosurgical workflow, a series of phantom localisations and target positionings were performed as stated in section II, subsection D, for which three different plans were used (Fig.9). For each plan first the default parameters (MDH0) are used for positioning, and then the newly calibrated ones (CAL). Due to the dimensions of the tool and the positions of the spheres on the phantom, only the second plan could utilise all seven target spheres. The first plan (plan1) evaluates the effect of the calibration when the markers are localised with the tool set in the same orientation as the trajectories. This plan uses five spheres with parallel trajectories. The second plan (plan2) evaluates the effect when one localisation is performed for several trajectories whose orientations deviate 25°-35° from the trajectory of the tool X-axis while localising the markers. Seven trajectories are defined for the second plan. The third plan (plan3) includes four trajectories with a 65° deflection angle from the

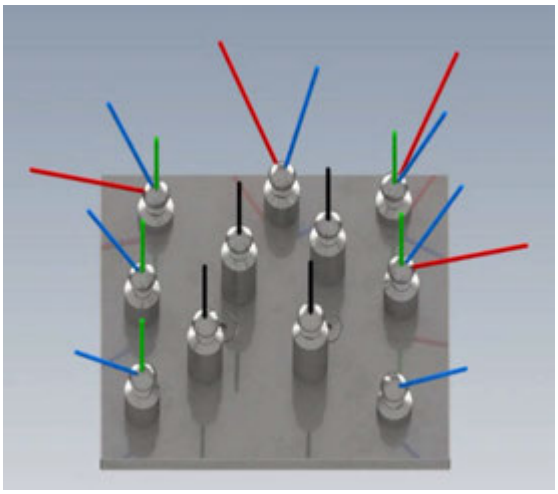


Figure 9 Phantom with localization orientations for the localization markers (four black trajectories), plan1 trajectories (five green trajectories), plan2 trajectories (seven blue trajectories) and plan3 trajectories (four red trajectories)

localisation trajectory of the tool X-axis and defines four trajectories.

For an additional validation, all the positionings performed using the MDH0 and CAL are also done using parameters calibrated through the RoboDK commercial calibration software (RDK). RoboDK uses an open loop calibration approach, measuring the TCP position using a laser tracker throughout the robot's workspace and in various joint configurations. This allows a much more thorough coverage of the influences specific parameters have on the positioning errors. The TCP position was limited to a 1000 x 1000 x 800 [mm] volume including the area in which the patient is usually located and 99 measurements in total were taken to calibrate the robot. The mean recorded error before calibration was 1.328 mm, with a maximum error of 2.378 mm, while after the calibration mean error was reduced to 0.061 mm and the maximum to 0.155 mm. Both the CAL and RDK post-calibration error values are stated in Table 1.

Ten localisations are performed for each of the trajectory plans and before each localisation the phantom position and orientation is changed. In total 160 target point measurements are performed, 50 for the first plan, 70 for the second and 40 for the third. Deviations presented in Fig.10. show that even when the first plan is used, locating the markers for each trajectory orientation [27], the mean error is cut in half when calibration is used. A significant increase in error is present for the MDH0 parameters for plan2, and the error is even three times higher when comparing plan2 and plan3. CAL and RDK plans show a

Table 1 Errors before and after calibration

Error [mm]		Mean	Std. Dev.	Max.
CAL	Before	0.816	0.407	1.843
	After	0.101	0.057	0.225
RDK	Before	1.328	0.369	2.378
	After	0.061	0.031	0.155

slight increase for plan 2 (with a couple of outliers) and a noticeable increase in error for plan3, although still at or under the error level MDH0 parameters exhibit for trajectories in plan1. Error values for individual plans are shown in Table 2. If all the deviations are plotted in a histogram (Fig.11) MDH0 parameters show a much more spread out curve, especially in the right portion showing the plan3 trajectory errors, while CAL and RDK models are rather tightly grouped around the 0.1 mm deviation for all the measurements. Although the post-calibration results are significantly better for the RDK model than for the CAL model (see Table 1), the validation difference is rather minor for the first three plans and is only really shown for the larger trajectory deviations in plan3.

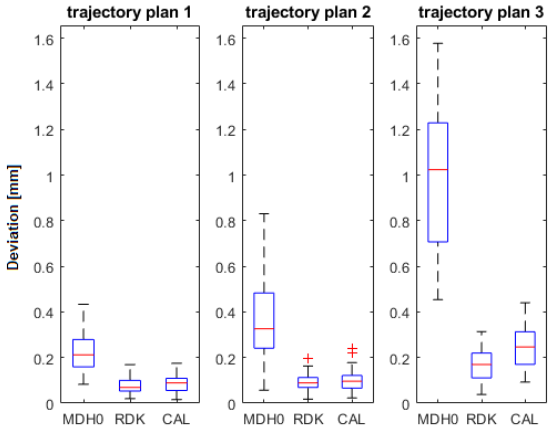


Figure 10 Deviation results for all three plans when using the MDH0, CAL and RDK model parameters

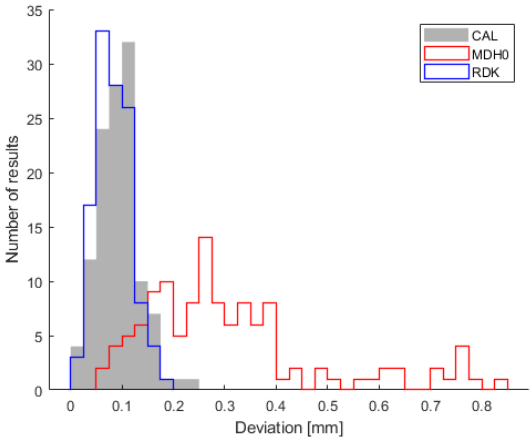


Figure 11 Histogram displaying deviation distribution for MDH0, CAL and RDK parameter sets for all three plans

Table 2 Errors for individual plans

		Errors [mm]		
Plan	Model	Mean	St.Dev.	Max.
1	MDH0	0.224	0.086	0.433
	RDK	0.077	0.034	0.169
	CAL	0.087	0.037	0.175
2	MDH0	0.371	0.200	0.831
	RDK	0.090	0.034	0.194
	CAL	0.098	0.043	0.238
3	MDH0	0.989	0.336	1.578
	RDK	0.168	0.071	0.314
	CAL	0.232	0.095	0.440

V. CONLUSION

Although the achieved error reduction is somewhat lower compared to the RoboDK calibrated parameters, the achieved results still provide a significant improvement compared to the non-calibrated parameters, especially when large trajectory deviations are present. Apart from not requiring expensive laser trackers to perform the calibration, using a simple calibrating tool as presented in the article can easily be integrated into the system to provide regular accuracy checks, as well as (semi)automated recalibration if it proves necessary.

The current calibration setup was also quite limited due to the type of reference tool used for measurement, significantly reducing positioning possibilities. Further research will include a new reference tool enabling much more flexibility. Automating the measurements will also be much easier, therefore enabling a much larger number of measurements to be performed and compare influences various positions have on the calibration. While automating the procedure observability indexes will also be added to maximise the calibration effectiveness for the given number of measurements. Also, a different modification of the MDH model is planned to be tested, using the definition presented in [28] which adds an additional MDH parameter solving the parallel axes issue present in other DH/MDH definitions.

The achieved results show great promise for implementation in the surgical process, enabling possible significant increase in accuracy and reduction in localisation time by eliminating multiple localisation orientations for multi-trajectory procedures. However, the question remains how much of the accuracy increase will be left after other error sources are reintroduced into the kinematic chain like CT imaging, camera localisation, neurosurgical probe, etc. and what final effect can be expected for the in vivo procedures. Future research will also cover these aspects implementing the rest of the error sources into the kinematic chain and examining and modelling their influence on procedure accuracy.

ACKNOWLEDGMENT

The authors would like to acknowledge the project RONNA - Robotic neuronavigation, as well as the support of the European Regional Development Fund received through the project NERO – Neurosurgical robot. Authors would also like to acknowledge the project “CRTA – Regional Centre of Excellence for Robotic Technologies”.

REFERENCES

[1] D. Chudy, ‘Chudy - 2000 - The clinical study of possibilities and usefulness of the new stereoadapter in neurosurgery’, 2000.

[2] T. Haidegger, Ed., *From Concept to Market: Surgical Robot Development*. IGI Global, 2015.

[3] Chris Wagner *et al.*, ‘Overview of Robotics in Neurosurgery’, 2018.

[4] C. Faria, W. Erlhagen, M. Rito, E. De Momi, G. Ferrigno, and E. Bicho, 'Review of Robotic Technology for Stereotactic Neurosurgery', *IEEE Reviews in Biomedical Engineering*, vol. 8, pp. 125–137, 2015, doi: 10.1109/RBME.2015.2428305.

[5] B. E. Youngerman, F. A. Khan, and G. M. McKhann, 'Stereoelectroencephalography in epilepsy, cognitive neurophysiology, and psychiatric disease: safety, efficacy, and place in therapy', *NDT*, vol. Volume 15, pp. 1701–1716, Jun. 2019, doi: 10.2147/NDT.S177804.

[6] M. Švaco, B. Šekoranja, F. Šuligoj, J. Vidaković, B. Jerbić, and D. Chudy, 'A Novel Robotic Neuronavigation System: RONNA G3', *Strojniški vestnik - Journal of Mechanical Engineering*, vol. 63, no. 12, 2017, doi: 10.5545/sv-jme.2017.4649.

[7] D. Dlaka *et al.*, 'Brain biopsy performed with the RONNA G3 system: a case study on using a novel robotic navigation device for stereotactic neurosurgery', *The International Journal of Medical Robotics and Computer Assisted Surgery*, p. e1884, Dec. 2017, doi: 10.1002/rcs.1884.

[8] B. Jerbić *et al.*, 'RONNA G4—Robotic Neuronavigation: A Novel Robotic Navigation Device for Stereotactic Neurosurgery', in *Handbook of Robotic and Image-Guided Surgery*, Elsevier, 2020, pp. 599–625.

[9] F. Suligoj, B. Jerbic, M. Svaco, and B. Sekoranja, 'Fully Automated Point-Based Robotic Neurosurgical Patient Registration Procedure', *Int. j. simul. model.*, vol. 17, no. 3, pp. 458–471, Sep. 2018, doi: 10.2507/IJSIMM17(3)442.

[10] Albert Nubiola, 'Contribution to improving the accuracy of serial robots', PhD Thesis, École de technologie supérieure, Montreal, 2014.

[11] Chen-Gang, Li-Tong, Chu-Ming, J.-Q. Xuan, and S.-H. Xu, 'Review on kinematics calibration technology of serial robots', *International Journal of Precision Engineering and Manufacturing*, vol. 15, no. 8, pp. 1759–1774, 2014, doi: 10.1007/s12541-014-0528-1.

[12] A. Filion, A. Joubair, A. S. Tahan, and I. A. Bonev, 'Robot calibration using a portable photogrammetry system', *Robot. Comput.-Integr. Manuf.*, vol. 49, pp. 77–87, Feb. 2018, doi: 10.1016/j.rcim.2017.05.004.

[13] L. Ma, P. Bazzoli, P. M. Sammons, R. G. Landers, and D. A. Bristow, 'Modeling and calibration of high-order joint-dependent kinematic errors for industrial robots', *Robotics and Computer-Integrated Manufacturing*, vol. 50, pp. 153–167, Apr. 2018, doi: 10.1016/j.rcim.2017.09.006.

[14] A. Nubiola, 'Calibration of a serial robot using a laser tracker', École de technologie supérieure, 2011.

[15] H. Tang, Z. He, Y. Ma, and X. Zhang, 'A Step Identification Method for Kinematic Calibration of a 6-DOF Serial Robot', in *Mechanism and Machine Science*, vol. 408, X. Zhang, N. Wang, and Y. Huang, Eds. Singapore: Springer-Verlag Singapore Pte Ltd, 2017, pp. 1009–1020.

[16] S. Gharaaty, T. Shu, A. Joubair, W. F. Xie, and I. A. Bonev, 'Online pose correction of an industrial robot using an optical coordinate measure machine system', *International Journal of Advanced Robotic Systems*, vol. 15, no. 4, p. 172988141878791, Jul. 2018, doi: 10.1177/1729881418787915.

[17] Y. HuangFu *et al.*, 'Research on Robot Calibration Based on Laser Tracker', in *Mechanism and Machine Science*, vol. 408, X. Zhang, N. Wang, and Y. Huang, Eds. Singapore: Springer-Verlag Singapore Pte Ltd, 2017, pp. 1475–1488.

[18] G. Xiong, Y. Ding, L. Zhu, and C.-Y. Su, 'A product-of-exponential-based robot calibration method with optimal measurement configurations', *Int. J. Adv. Robot. Syst.*, vol. 14, no. 6, p. 1729881417743555, Nov. 2017, doi: 10.1177/1729881417743555.

[19] A. Nubiola and I. A. Bonev, 'Absolute robot calibration with a single telescoping ballbar', *Precision Engineering*, vol. 38, no. 3, pp. 472–480, Jul. 2014, doi: 10.1016/j.precisioneng.2014.01.001.

[20] M. Gaudreault, A. Joubair, and I. A. Bonev, 'Local and closed-loop calibration of an industrial serial robot using a new low-cost 3D measuring device', in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, 2016, pp. 4312–4319.

[21] H. Gattringer, M. Neubauer, D. Kaserer, and A. Muller, *A Novel Method for Geometric Robot Calibration Using Laser Pointer and Cameras*, vol. 49. Cham: Springer International Publishing, 2018.

[22] A. Joubair and I. A. Bonev, 'Non-kinematic calibration of a six-axis serial robot using planar constraints', *Precision Engineering*, vol. 40, pp. 325–333, 2015, doi: 10.1016/j.precisioneng.2014.12.002.

[23] M. Švaco, B. Šekoranja, F. Šuligoj, and B. Jerbić, 'Calibration of an Industrial Robot Using a Stereo Vision System', *Procedia Engineering*, vol. 69, pp. 459–463, 2014, doi: 10.1016/j.proeng.2014.03.012.

[24] J. Denavit and R. S. Hartenberg, 'A Kinematic Notation for Lower-Pair Mechanisms Based on Matrices', *Journal of Applied Mechanics*, pp. 215–221, Jun. 1955.

[25] J. J. Craig, *Introduction to robotics: mechanics and control*, 3. ed., international ed. Upper Saddle River, NJ: Pearson, Prentice Hall, 2005.

[26] A. Joubair and I. A. Bonev, 'Kinematic calibration of a six-axis serial robot using distance and sphere constraints', *The International Journal of Advanced Manufacturing Technology*, vol. 77, no. 1–4, pp. 515–523, Mar. 2015, doi: 10.1007/s00170-014-6448-5.

[27] F. Šuligoj, B. Jerbić, B. Šekoranja, J. Vidaković, and M. Švaco, 'Influence of the Localization Strategy on the Accuracy of a Neurosurgical Robot System', *TFAMENA*, vol. 42, no. 2, pp. 27–38, Jun. 2018, doi: 10.21278/TOF.42203.

[28] S. Hayati, 'Robot arm geometric parameter estimation', 1983.

Implementation of a low-cost autonomous underwater vehicle using open source ROS components with consumer class sonar technologies

T. Aaltonen*, M. Saarivirta**, T. Kerminen*and J. Grönman**
* Satakunta University of Applied Sciences, Pori, Finland
** Tampere University/Faculty of Information Technology and Communication Sciences, Pori, Finland
toni.aaltonen@samk.fi

Abstract - **The increase in the availability of low-cost components has made it possible to design and implement new and innovative devices for the underwater robotics. Today, cost-effective and open solutions can be created to replace previous expensive and proprietary robot systems. This study presents the prototype of an autonomous underwater vehicle. The system was built using commonly available sensor components in combination with an open source code and in-house designed parts. The configuration of the autonomous underwater vehicle is introduced in this study. The results obtained from the component tests performed in the actual operating environment are presented. Based on these tests, the subsystems were found to be well-suited for underwater environment.**

Keywords - AUV; low-cost; ROS; sonar; open-source; underwater robotics

I. INTRODUCTION

Autonomous Underwater Vehicles (AUV) have been subject to scientific interest for some time, and a continuum of research projects have been developed around the theme, such as the European ARCHEOSUB project. [1] Interest in AUVs has also awakened in the business world; robots are now increasingly used in commercial sites, such as oil and gas industry, as well as fish farming. [2] The increased use of AUV equipment is supported by their low operating costs and the reduction of personnel risks in challenging conditions.

Satakunta region is a highly industrialized province which Rauma-Repola Oceanics established already in the 1980s by delivering a technically advanced Mir submarine to the Soviet Academy of Sciences to study the secrets of the deep sea. Ever since, there has been developed a cluster of companies focusing on automation and robotics in the Satakunta region. The starting point for this research was to combine those two high tech engineering sectors, maritime and robotics.

The aim of the study was to create a low cost AUV based on commercial off-the-shelf and open source components. The design of the prototype largely takes advantage of the solutions created by the open source community and the literature in the field. The AUV was built in RoboAI Research and Development Centre founded by Satakunta University of Applied Sciences (SAMK) and Tampere University. This RoboAI study was a joint project for SAMK's Automation research team and Tampere University's Data Analytics and Optimization research group.

II. BACKGROUND

The first SPURV (Self-Propelled Underwater Research Vehicle) was built in 1957 at the University of Washington, USA. It was made of aluminum and resembled a tornado. It could dive up to 3600 m and was able to travel at a speed of 4 to 5 knots. Its primary power source was the propulsion battery with two sets of 16 silver-zinc cells connected in parallel through diodes. The battery was 24 volts and 200 Ah. Secondary power supply was provided by four solid-state converters. [3] SPURV II was developed in 1973 with a greater range, speed and operating time. The SPURV II could be operated by a computer stationed on a marine vessel. Both were still operated until the 80's. [4] In 1983, ISE (International Submarine Engineering) Ltd. developed in co-operation with the Canadian Hydrographic Service (CHS) and the Department of Defense the ARCS (Autonomous and Remote-Controlled Submarine) as a test platform for autonomous submarines. It has been used since 1987 for development and demonstration of autonomous underwater technologies. Today, ARCS is used to develop new AUV technologies e.g. navigation system, algorithms and battery types. [5] In 2001, Woods Hole Oceanographic Institute Laboratory developed REMUS (Remote Environmental Monitoring Units). It was intended for marine scientific community for precision tasks in coastal areas. Its small size made it relatively unique as it was only 19 cm in diameter and weighed only 37 kg. Having four lithium-ion batteries it was capable of a speed of three knots with a range of 80 km. It was designed to be operated with a laptop. [6]

The Robot Operating System (ROS) was developed by the Stanford Artificial Intelligence Laboratory (SAIL) in 2007. ROS is becoming a very popular open-source middleware system (i. e. collection of software frameworks) and it is widely accepted by the research and industrial communities. Its goal is to make robotics system

easier to develop and share the work on other robots with minimal changes. Meaning that the code can be reused and improved by many different users and platforms. [7] ROS provides a set of tools and libraries. It supports message-passing between different processes that run on one or several computers. Processes are called "publishing", subscribing" and "messages". Both publishing and subscribing to a message are called "nodes". Data logging is also possible and a UUV (Unmanned Underwater Vehicle) Simulator package is available on GitHub containing Gazebo plugins. [8] Among others, ROS can be found in drones, service robots and AUVs. The future potential of ROS lies in the expanding applications of robotics and the developing of new, smarter applications in robotics.

RoboNation's international competition, Robosub gathers student teams from both universities and high schools around the world to challenge each other in the making and developing of AUVs.

In the USA, Riverside's student organization, RoboSub UCR from the University of California, has built an AUV called Seadragon. It is capable of completing several aquatic tasks on its own. They are using ROS, Python, C/C++ for their controls. [9] The University of Florida (UF) presents its eighth generation SubjuGator that is developed by undergraduate students in UF's Machine Intelligence Laboratory (MIL). The university has a 22-year history of autonomous vehicles. The newest SubjuGator concentrates on robust control, hardware improvements and software innovations. [10] At Michigan Technological University the Nonlinear and Autonomous Systems Lab (NASLab) has developed and combined a backseat controller for the General Dynamics Bluefin SandShark AUV. The integration of ROS is important to promote transference of code and skillset across domains. They have installed ROS on the Raspberry Pi with written drivers to initialize communication. [11]

In Spain, Underwater Robotics Research Centre (CIRS) of the University of Girona developed and tested the SPARUS II AUV. It is a lightweight classic torpedo-shaped vehicle with a hovering capability and a mission-specific payload. The software is based on ROS for multipurpose applications. [12]

Other parties have also taken interest in developing AUVs, such is a national project of Japan named "Next Generation Technology for Ocean Resources Exploration ", its nickname is "Zipangu-in-the-Ocean". HOBALIN, the hovering AUV was originally developed in 2015. However, because it was based on an old design, they decided to replace the software with a new design based on ROS. HOBALIN completed its survey missions on February 2018. The authors continue improving the AUV. [13] Hydroid (USA), owned by Kongsberg Maritime (Norway) has announced that the "Next Generation" of the REMUS 100 will be using ROS interface. Currently no open source is available. [14]

The University of Tampere has also given its own contribution to the AUV development. In 2016 the UNEXMIN project (funded by European Union's Horizon 2020) produced a high performance AUV, named UX-1 for the inspection of abandoned flooded mines. The UX-1 can dive up to 500 meters. Its spherical form with a 30 cm radius and 8 thrusters is well suited for the inspection of narrow mine channels. [15] Later on, in 2019 they have improved the propulsion system of the UX-1 by developing creative subsystems to improve the mobility of the robot. [16]

III. IMPLEMENTATION OF AUV

The SubjuGator project from the University of Florida was used as an example for the project. The University has more than 20 years of experience in AUVs and it has won the Robosub competition 3 times and been in the top 3 for 7 times. The vehicle is based on ROS. This in all was a good starting point. Although there came to be some major differences. Mainly because of the differences in the budget. The SubjuGator's sonar alone cost ca 60 000 € and the software package made 7 000 €. In addition, the Doppler Velocity Log (DVL) used to measure the vehicles speed was around 20 000€.

The second major difference is that the Robosub competition is organized in a test pool. In clear water, camera-based machine vision systems are a cost-effective and easy way to identify objects. The vehicle of this project is intended to operate in the Finnish murky waters where the possibilities of using cameras are very limited. Visibility in coastal waters are often reduced to a few dozens of centimeters and may be limited to as much as twenty centimeters. The thruster system of the project also became quite different.

A. Hull structure

The frame of the vehicle was made from aluminum profile. Two pressurized hulls were placed for the electronics. The sonar unit and the main computer were placed in the larger upper hull whereas power supplies, batteries and motor-drives were placed in the lower hull. The device was designed to be slightly buoyant so that it automatically returns to the surface in case of a failure. The choice between remote access and battery operation is made by changing the flange of the lower pressure housing. One of them is a solid flange and the other one has electrical and optical cable where the optical cable acts as a telecommunication link, and due to its tensile strength, as a traction protection for the electric cable. In addition, a pressurized air hose runs in the same bundle, the purpose of which is to make the cable slightly buoyant so that it does not stick to the shapes of the base. At the same time, buoyancy also reduces the resistance to the device.

B. Computer

The main computer was an AMD-based x86 computer. The processor was the Ryzen 2700 with 8 cores and 16 threads. The motherboard was Gigabyte B450 mATX with 2 memory slots. The main memory was 32 GB of DDR4. The graphics card was Asus Dual GTX 1070 with 8 GB of memory. The hard drive was a Samsung EVO 970 512GB NVMe SSD. The computer was powered by HDPLEX

400W DC-DC. This can be used both with current from the batteries and with the 24v DC voltage of the main power supply. The main power source which converts 240 Vac ground power to 24 Vdc was the Meanwell DPU-3200. Its efficiency was promised to be over 94% even at full 3192W power.

C. Sonars

Robotic sonar units will cost orders of magnitude more than could be procured with the budget for this project. As a result, less conventional alternatives had to be sought. The sonar used in yachting is quite inexpensive, but still quite good in image quality. In addition, they are easy to use. The problem with these is that they are not intended for such use and as a result, imaging on a computer is challenging.

This project went through various manufacturers such as Hummingbird, Lowrance, Raymarine and Garmin, looking for alternatives. It was then found that whereas the older serial fieldbuses for yachting were relatively open and used by all, such as NMEA0183 and newer NMEA2000, Ethernet-based busses are the manufacturer's own closed ones and only discuss between the manufacturer's own devices. It was not even known whether the traffic was encrypted in any way or not. The data produced by modern sonar is so accurate and bandwidth consuming that all manufacturer's sonar data travels between devices on Ethernet-based buses. There are several ways to get data from a device. The Ethernet protocol used can be determined by listening to traffic and writing a program that uses this, which appears as a manufacturer's device on the network, so that the sonar can send its data to it. The problem here is that at least two sonar units are needed to talk to each other over the Ethernet-based bus in order to listen to Ethernet traffic.

Another option is one sonar unit and some other device that uses, for example, an Ethernet-based sonar sensor on the bus. Another problem is that there is no certainty whether the traffic is encrypted or not. With all these things in mind, an alternative way of getting data on a computer was considered. Some of the better model series from Simrad and Garmin have an HDMI output. HDMI-USB3 capture cards for video processing, recording and retransmission are available in the price range of less than 100 €. For this reason, it could be assumed that the delay would be reasonable. The degree of difficulty in the implementation of this solution is considerably lower than in the determination of the bus protocol. Even if the accuracy is a bit off, it is worth the saved efforts as well as the reduced risk. For these reasons, this solution was reached. At the time of the decision there was no certainty about the functionality as there was found no prior information of such ever been done. The sonar was selected as Garmin's 8400 Xsv and the sensors were Lvs12 real-time sensors and GT54UHD-TM side-scan sonar. The solution has worked as expected, and there has been no visible delay between the image of the sonar unit and the image transferred to the computer. The effect on image quality has also proved to be insignificant.

D. Location and velocity information

The underwater depth information can be divided into two parts, the distance from the bottom and the distance from the surface. The distance to the surface is simply obtained by pressure measurement. MS580314BA01-00 was selected as the sensor. According to the sensor manufacturer's data sheet, the accuracy should be maintained within 30 cm even for longer periods of time.

The distance to the bottom is obtained by the sonar unit via the NMEA2000 bus. There has not been found any precise information of the accuracy of this, but when tested a few times with a measuring stick in the data gathering phase, the accuracy appears to be of dozens of centimetres. Of course, different bottom shapes can affect this.

The IMU, or inertia measurement unit is a commonly used sensor. The prices of these start from a couple of euros and the most expensive ones are tens of thousands of euros. In these, short-term information is usually useful, but sensor errors accumulate rapidly so that they cannot be relied upon alone, especially when calculating velocity and position from acceleration.

The weakness in calculating velocity from the sonar image is its dependence on the type of the bottom. There is no constant shape at the bottom from which the velocity could be calculated. On the other hand, there are some bottom types that the speed is available almost all the time.

At its best, the impeller only measures the flow of the water and only the velocity can be calculated in relation to the water. This can correlate well with the ground speed in certain situations, for example in small lakes, but in places where the water flows, the speed obtained in this manner is not reliable. The sensor is also easily broken down. However, this feature can be improved by installing more sensors. The impeller sensor is used to detect transverse motion. GPS location information has no cumulative error problems but does not work under water. GPS can be used as an occasional ground truth, by visiting the surface.

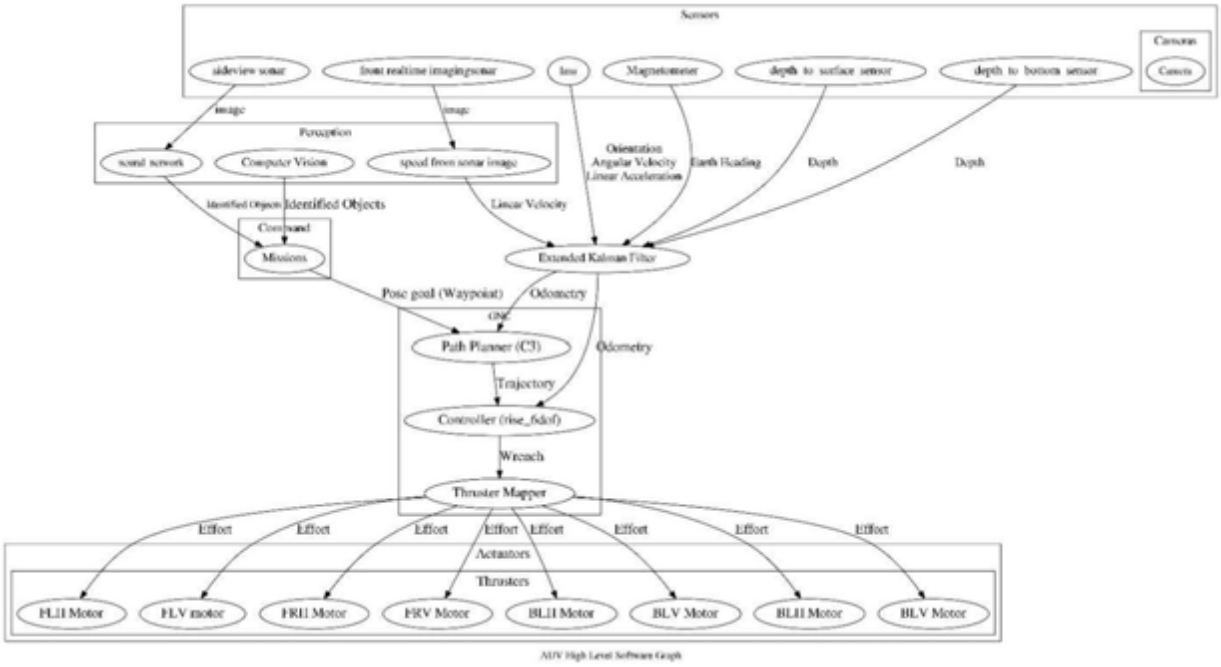


Figure 1. High level software architecture

E. *Thrusters*

There are eight thrusters and they are positioned as shown in Fig. 3. The structure decision was made because even if any of the thrusters were damaged, the task could still be completed, and the movement would still be successful in any direction. In some cases, the failure of either thrusters does not yet restrict the mobility of the device. Excluding the motors, the thruster assembly was 3D printed with PLA (polylactic acid) filament. Motors in thrusters are brushless DC motors normally used in RC helicopters. Each motor can be operated around 840W continuous power. Small zinc plates are attached to motors, for enhancing corrosion resistance. The thruster motor controllers are slightly modified VESC - open source ESC-controllers that have an RS485 bus to connect them to the computer.

F. *Software architecture*

In Fig. 1., there is a high-level architectural image of how information moves in a ROS system. The architecture is largely borrowed from the SubjuGator project that has been modified to fit the project. The mission system, path planner, controller, and thruster mapper are nearly identical to those of the SubjuGator. The Kalman filter is a robot localization package. Sonars generate data of detected objects and possible obstacles. The mission planner tells you the desired location. The sensors are collected on an extended Kalman filter, which provides the information and the desired location to create the desired route. It is then calculated how much power is desired for each motor so that the desired motion occurs. The Rise Controller monitors that the desired movement occurs.

IV. TEST PROCESS

Data gathering was carried out by attaching the sensors to various boats and the sensor data was stored on a laptop computer. The data gathering operation had several different objectives. There was the desire to learn how to calibrate and use sonars and to test sensors and bus software. The test uses ROS nodes and gathered data, as well as gathering data for teaching neural networks and calibrating algorithms. The most significant consideration for neural network-based velocity calculation is that GPS speeds which are not available below the surface of the water can be recorded while driving on the surface. In the data acquisition hardware, the ROS software ran on a laptop to which the devices were connected via USB buses. The devices are installed in the housing of the Fidobox. The sensors were attached to the boat with a bracket made of aluminum profile. Fig. 2. presents a picture of that. The rack proved to be well functional and adaptable. Data was gathered on five different boats.



Figure 2. Test equipment installed to a boat

The sealing of the device was tested before the actual dive so that any leaks would not cause equipment failure. The pressure hulls were tested by pressurizing them with water to a pressure of 5 bars. The operation of the thruster unit and engine controller were tested in a water tank to ensure that water did not affect the operation of the thrusters or cause short circuits. The motor cable connections were also tested at a water pressure of 5 bars.

V. RESULTS

In Fig. 3., the assembly of the vehicle was almost ready. The dive tests of the device have not yet begun, although many different areas have been tested. Testing has been done physically with tightness tests, testing of individual thrusters in a pool of water, software in data gathering, teaching neural networks, and running software sections in simulations. To date, no insurmountable obstacles have occurred, and it was expected that the device will also pull through the dive test. The prototype came to cost ca. 9.500€ where the sonar components are the most expensive components, ca. 4000€. Thrusters, thruster controllers and the computer cost ca. 1000€. Battery pack and BMS (Battery Management System) costs 1500€.

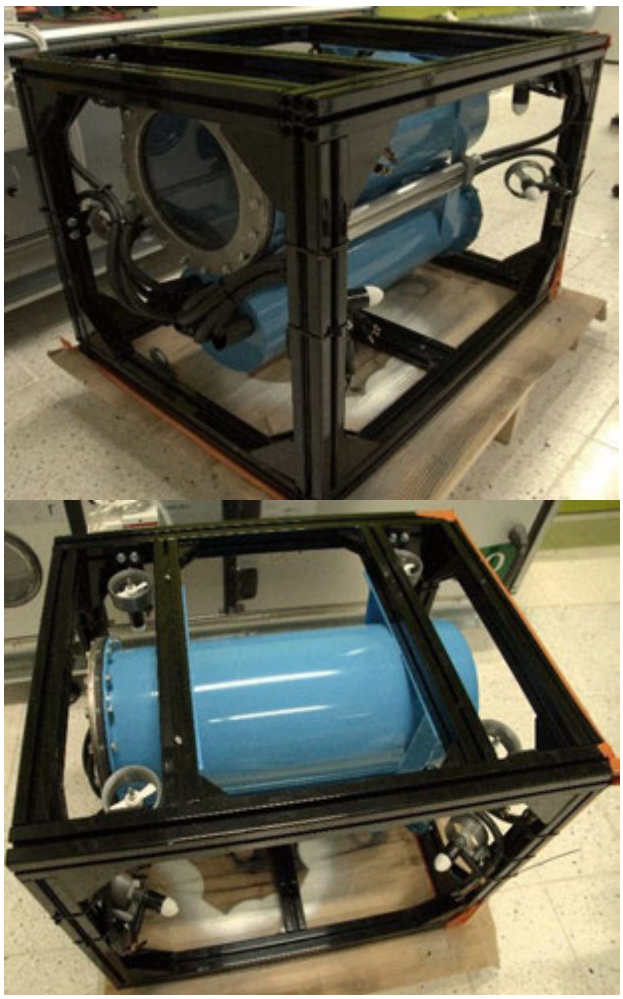


Figure 3. Pictures of the assembled vehicle

The first pressure test revealed that the seal type of the flanges was not very suitable for the purpose and was replaced with rubber. In addition, fastening bolts were added. After these modifications, the flanges were tight. The water tightness of the motor cable endings situated on the motor-side were also successfully tested at a pressure of 5 bars.

The thruster unit and motor controller perform well in the tests. More detailed thrust measurements have not yet been made, but the feeling is that there is enough thrust to move the device at the desired speed of a few kilometers per hour. The interconnection of the motor controller and the ROS software was also successfully tested.

The sonar was tested in data gathering. Data transfer with an HDMI-USB converter to a computer worked reliably, and no problems occurred after several hours of use. The associated ROS code was also tested at the same time. Optimal settings were sought for the sonar settings so that the image would be as good as possible in different ground conditions and that the targets would be identified as well as possible.

Preliminary results of the velocity calculated from the sonar image are encouraging. The velocity error compared to GPS speed is on average less than 0.2 km / h. With the increasing amount of data, the error is expected to decrease even more. It should also be noted that GPS speed is not accurate, especially at low speeds. In some situations where the measured GPS speed is incorrect, the calculated speed may be closer to the truth. How these results correlate to a practical situation where there are almost limitless new types of bottom shapes, remain to be seen.

VI. CHALLENGES

Obtaining location and velocity information below the surface is largely different from obtaining similar information in the air or on the ground, although there are some similarities. The biggest difference is the lack of GPS and similar global positioning systems. In the water radio signals do not travel deeper than the surface layers, so these cannot be used. DVL is the standard way to do velocity measurement in commercial more expensive underwater devices. The problem here is mainly the price, which limits its use in this implementation. With the selected sensors, the movement of one axis can be calculated quite accurately from the sonar image, and the pressure gives a second axis, but in the lateral direction one must rely mainly on the information of the impeller sensor and the accelerometer.

VII. Conclusion and Future Studies

This study presents the prototype of an autonomous underwater vehicle. The system was built using commonly available sensor components in combination with an open source code and in-house designed parts. The configuration of the autonomous underwater vehicle was introduced in this study. The results obtained from the component tests performed in the actual operating environment were presented. Based on these tests, the subsystems were found to be well-suited for underwater environment.

There are couple of possible actions that could be taken into consideration in the future research. Although, the Kalman filters have been a standard way to make sensor fusions for a long time, their ability to respond to changing circumstances are limited. As more data accumulates from the device tests, a neural network could be a better solution for this. It could be easier to teach the neural network, for example, situations where a certain sensor does not provide the correct information or does not provide information at all.

Another more researchable action could be developing a proper simulation model using different neural networks in Gazebo simulation environment. The simulation done by the SubjuGator project provides a good basis for the simulation model. To simulate the robot, the robot data as a URDF model is required. For the simulation, the device's water resistance information is needed, as well as the thrust produced by the motors at different power values. However, based on the data from the test run, they can be calculated. The most challenging will probably be the simulation of the sonars. The Gazebo does have LiDAR-type sensors modeled, so this could probably be a good starting point.

Acknowledgment

This work has been supported by Ulla Tuominen Foundation SR, High Technology Foundation of Satakunta, WinNova Länsirannikon Koulutus Oy Ltda and Satakunta Rescue Department Unit of Rauma, Finland. Helkama Bica

References

[1] J. Gelli, A. Meschini, N. Monni, A. Ridolfi, L. Marini, B. Allotta Development and Design of a Compact Autonomous Underwater Vehicle: Zeno AUV DOI: https://doi.org/10.1016/j.ifacol.2018.09.463, 2018

[2] H.Niu, S. Adams, K. Lee, T. Husain, N. Bose Applications of Autonomous Underwater Vehicles in Offshore Petroleum Industry Environmental Effects Monitoring DOI: https://doi.org/10.2118/09-05-12-GE, 2009

[3] H. R. Widditsch SPURV – The First Decade, University of Washington, Applied Physics Laboratory, 1973.

[4] W. E. Nodland, T. E. Ewart, W. P. Bendiner, J.B. Miller, and E. E. Aagaard SPURV II - AN UNMANNED, FREE-SWIMMING, SUBMERSIBLE DEVELOPED FOR OCEANOGRAPHIC RESEARCH, University of Washington, Applied Physics Laboratory, 1981.

[5] Autonomous Undersea Vehicle Applications Center. AUV System Spec Sheet. http://auvac.org/configurations/view/57?from_search=1, 2018.

[6] Moline. M. A., Blackwell S. M., von Alt C., Allen B., Austin T., Case J., Forrester N., Goldsborough R., Purcell M. & Stokey R. Remote Environmental Monitoring Units: An Autonomous Vehicle for Characterizing Coastal Environments. Journal of atmospheric and oceanic technology, 2005.

[7] A. Mahtani, L. Sánchez, E. Fernández, A. Martinez, "Effective Robotics Programming with ROS", 3rd ed. Birmingham, Packt Publishing Ltd, 420 s. ISBN 978-1-78646-365-4, 2016.

[8] UUV simulator, retrieved May 12, 2020, from https://uuvsimulator.github.io/

[9] Robosub UCR, "SeadragonAUV", retrieved May 12, 2020, from https://github.com/robosub-ucr/SeadragonAUV

[10] D. Volya, J. Brown, A. Puldio, N. Suhlman, E. M. Schwartz "SubjuGator 2019: Design and implementation of a Modular, High-Performance AUV, 2019

[11] J. E. Naglak, B. R. Page, N. Mahmoudian, "Backseat Control of SandShark AUV using ROS on Raspberry Pi", DOI: 10.1109/OCEANS.2018.8604630, 2018

[12] M.Carreras, C. Candela, D. Ribas, N. Palomeras, L. Magí, A. Mallios, E. Vidal, É. Vida, P. Ridao, "Testing SPARUS II AUV, an open platform for industrial, scientific and academic applications", http://arxiv.org/abs/1811.03494, 2018.

[13] T. Seta, A. Okamoto, S. Inaba, M. Sasano, "A New ROS-Base Operation Software for an Autonomous Underwater Vehicle HOBALIN and Its Test in Real Sea", DOI: 10.1109/OCEANSKOBE.2018.8559223, 2018

[14] E. Gallimore, R. Stokey, E. Terrill, "Robot Operating System (ROS) on the REMUS AUV using RECON", DOI: 10.1109/AUV.2018.8729755, 2018

[15] S. Zavari, O. Usenius, T. Salomaa, J. Villa Escusol, A. Heininen, J. Laitinen, J. Aaltonen, K. T. Koskinen,"Mechatronic Arhitecture Development of UX-1", DOI: 10.23919/ICCAS.2017.8204483, 2017.

[16] S. Zavari, O. Usenius, J. Villa, A. Heininen, T. Salomaa, J. Laitinen, J. Aaltonen, K. T. Koskinen,"Propulsion system development and power consumption in an autonomous underwater vehicle", DOI: 10.1109/ICSTCC.2019.8885503, 2019.

Improving Markerless Registration Accuracy by Mapping Facial Deformation

A. Žgaljić, M. Švaco, B. Jerbić
Faculty of Mechanical Engineering and Naval Architecture, Department of Robotics and Production System Automation, Zagreb, Croatia
adrian.zgaljic@fsb.hr

Abstract - Robotics is gaining an increasing role in stereotactic neurosurgery. Most state of the art robotic systems use marker-based patient localization and registration while markerless localization is at its infancy. One of the challenges of markerless localization and registration is face deformation. The human face can deform from the preoperative scanning phase to the operative position in the actual surgery. The face is not uniformly deformable which presents an opportunity for improving the markerless registration by using less deformable parts of the face while not registering more deformable parts. In this paper, a methodology for dividing the face surface into multiple regions and determining each region's deformability is proposed and validated on data from neurosurgical operations. To calculate the deformability of any facial region, multiple point clouds can be registered using ground truth measurements, followed by their respective deformability calculation. Ground truth registration is calculated using bone attached markers. Preliminary results are given for one intraoperative point cloud. Error is measured as the average Euclidean distance from registered cloud markers to ground truth markers. When the whole face was used for registration, the error was 4.19mm, while when using only selected parts of the face, the error was reduced to 1.78mm.

Keywords – registration; point cloud; neurosurgery; markerless

I. INTRODUCTION

In the last few years, the number of stereotactic brain surgeries is increasing. New minimally invasive surgical procedures, like intracranial thermal therapy [1], deep brain stimulation [2], genetic therapy for neurological disorders, and biopsies [3], are introduced. Although the stereotactic frame for intraoperative navigation is still considered a golden standard, the use of robots in stereotactic brain surgeries is increasing [4]. Robots are used for trajectory planning and perform faster and more accurate than surgeons [5]. The advantages of robotic systems in surgeries, compared to surgeons, are accuracy, speed, the programmability of complex 3D trajectories, repeatability, and the capability of holding still position for a long time without a tremor. The traditional method of stereotactic surgery which demands high accuracy involves the installation of a stereotactic frame on patient's head. With the frame attached, the patient undergoes preoperative radiological imaging, like CT, MRI, or both, which is time-consuming and can be a very unpleasant experience for the patient [6]. When using robots for stereotactic brain surgery, stereotactic frames are not used, but accuracy can be comparable or lower [7, 8]. Accuracy highly depends on the localization method used in robotic surgery. Localization can be achieved using markers. The higest accuracy is achieved using bone attached markers, followed by a lower accuracy when skin attached markers are used, with markerless localization being the least accurate.

In their work, Brandmeier et al. explored the accuracy of robotic system ROSA in a series of clinical studies using different registration methods. The authors concluded that bone attached markers have the same accuracy as the laser scanning of the patient's head. Although this conclusion is interesting, the fact that the accuracy error of both methods was up to a few millimetres, gives a conclusion that ROSA should be used for high accuracy procedures only with using a stereotactic frame [9]. Laser scanners for obtaining a 3D point cloud of patients heads are also used by Raabe et al. in their work [10]. Another method for obtaining point cloud of patients head for registration in neurosurgical operations, but using a stereo vision system is described in [11]. M. N. Wang and Z. J. Song in their research describe using an optical camera for markerless localization. Authors highlight the fact that all previous methods of markerless localization required that patients lie in the supine position because their face was used for scanning. This is a problem because not all trajectories could be performed in this position. The authors developed a system that uses a camera attached to a robotic arm, which can scan the patient's head from multiple angles and then create a point cloud. The downside of this method is that it requires drawing of marks on the patient's head, for which head needs to be completely shaved. The average target reegistration error (TRE) on the dummy head was 1.39 +-0.38 mm [12]. Another work that uses markerless registration is by Fan et al. In this research portable 3D scanner was used. Authors suggest that low accuracy of face scanning for registration is caused by the fact that scanners were attached to a tripod which resulted in the

limited scanning area. For testing, two phantoms were used: solid and rubber. TRE was 1.48 +-0.21 mm and 1.98 +-0.53 respectively [13]. Other interesting research showed that cheaper consumer cameras can be used for face scanning and localization. Potter and Yaniv used Microsoft Kinect v1 and 3DSystems Sense for scanning patients on multiple positions: supine, supine with a head frame, and prone. They worried that light conditions could be a problem for cameras, but it turned out that they had little effect on accuracy due to the camera's automatic exposure mode. Optimal object to camera distances was 50-60cm and 70-80cm for Sense and Kinect respectively. Measurements with patients in the supine position gave more accurate results than the patient being in a prone position. RMS error was 1.28 +- 0.01mm for Sense and 1.24 +- 0.03mm for Kinect [14]. In a recent study, scientists from the Korea Institute of Science and Technology proposed a robotic system for minimally invasive keyhole neurosurgery. A 3D surface scanner was used for markerless registration between a preoperative reference image and a patient's face in the operating room. Total system accuracy was tested with a phantom model. Calculated translation error was 0.75 +- 0.38mm and rotation error was 0.85 +- 0.16 [15]. These results indicate that markerless registration using face-scanning techniques can be accurate enough for deep brain stimulation, but these results still need to be confirmed in clinical trials, where error tends to be higher. To achieve satisfactory results in terms of accuracy, an optimal registration algorithm needs to be implemented. This topic is presented in work by Cutter et al. from the University of Birmingham. They compared two registration algorithms, iterative closest point (ICP) and coherent point drift (CPD), by registering ideal point clouds taken from MRI data with re-mashed, noisy, and smoothed versions. They found that ICP generally gives better and more consistent registration accuracy for the region of interest than CPD, with the best RMS distance of 0.88 +-0.05mm between aligned point clouds, as compared to 0.99 +-0.17mm or worse for CPD [16].

We propose a novel approach for improving the markerless registration by detecting more deformable regions of the face. Our hypothesis is that the registration accuracy can be improved by detecting those regions on a point cloud, and excluding them from the registration process. The process is divided into two steps: calculating the deformation of facial regions and the registration of preoperative and operative point clouds using less deformable regions. By calculating deformation of facial regions, facial deformation map is created which maps each facial region to it’s deformability measurement.

II. MATERIALS AND METHODS

The goal of this study is to calculate how much does each part of the patient's face deforms between preoperative scanning and the surgical procedure. The facial deformation can be calculated as the difference between multiple registered point clouds. The difference is defined as a standard deviation of Euclidean distances between correspoinding points. More deformable parts of the face will have a bigger difference while less deformable parts will have a smaller difference. We have used the clinical operative process of the RONNA system in order to acquire actual clinical data of the point clouds [17]. The RONNA clinical process consists of three main phases: the preoperative phase, the preparation phase and the operative phase. The preoperative point cloud was acquired in the preoperative phase while the operative point cloud was acquired in the preparation phase in the surgical theatre. The patient and his family member signed an informed consent detailing the entire procedure and also signed the consent to participate in this research. A written consent was also signed for using all the collected data for scientific research purposes, referring to all clinical and image data, without disclosing the patient's identity.

In the preoperative phase of the surgery, as a part of RONNA operative process, the patient is scanned on a CT scanner and a preoperative point cloud is generated from a CT scan image [17]. In the surgical preparation phase when the patient is on the operating table and after anaesthesia is induced, a 3-D scanner Ensenso N35 stereo is used to create the operative point cloud of a patient's face.

For mapping the facial deformation, a marker-based approach is used to register the preoperative and the operative point clouds. Standard retroreflective markers used in the clinical procedures with the RONNA system are used for the point cloud registration. Marker-based registration requires four reflective bone attached markers to be screwed into the patient's skull before pre-operative scanning. Coordinates of markers are automatically found on each point cloud, and all point clouds are registered using the found markers. Registered face point clouds are then divided into multiple regions by superimposing a $n \times n \times n$ 3D mesh. Deformation is measured for each region as standard deviation of Euclidean distances between correspoinding points.

A. Data acquisition

Data acquisition includes preoperative scanning of the patient with attached markers. A contrast-enhanced preoperative CT scan (Somatom Emotion®, Siemens, Erlangen, Germany) was performed the day before the surgery. A standard head protocol was used with slice width 0.75 mm, reconstruction increment 0.7 mm, image matrix 512 × 512 with a voxel size of 0.5 × 0.5 × 0.7 mm3.

The CT scan of the patient's head is transformed to a point cloud using the Osirix software. To find markers and their centres, an algorithm developed by Šuligoj et al. is used [18].

For intraoperative patient scanning an Ensenso N35 stereo 3D camera is used. The 3D camera uses infrared light to project a pattern for capturing untextured surfaces such as the human skin. The Z - accuracy of the camera is 0.18mm and pixel size is 0.32mm. The 3D scanning of the patient is performed just before the surgery, when the patient is on the operating table. Point cloud consist of approximately one milion points. Markers are detected using intensity based thresholding.

B. Facial deformation map

Once all point clouds are acquired, they are registered. This is done using markers. Firstly, one point cloud is picked and for each marker on that point cloud, corresponding markers on all other point clouds are found. When corresponding markers are found, the spatial transformation is calculated and point clouds are transformed so that markers are aligned. To divide the point cloud into regions, a 3D mesh is superimposed over the point cloud bounding box. The mesh consists of $n \times n \times n$ boxes, where n is set to 15, yielding 3375 boxes overall. Boxes that do not contain any points, such as the ones inside the patient's head, are discarded, and remaining boxes are regarded as facial regions. Overrall, 280 facial regions were detected. For each facial region, a measure of deformation is calculated as a standard deviation of all point clouds. Our hypothesis is that more deformable regions will yield a bigger difference between point clouds, while less deformable regions will be more similar and thus have a smaller standard deviation. Process of creating facial deformation map is displayed on Figure 1.

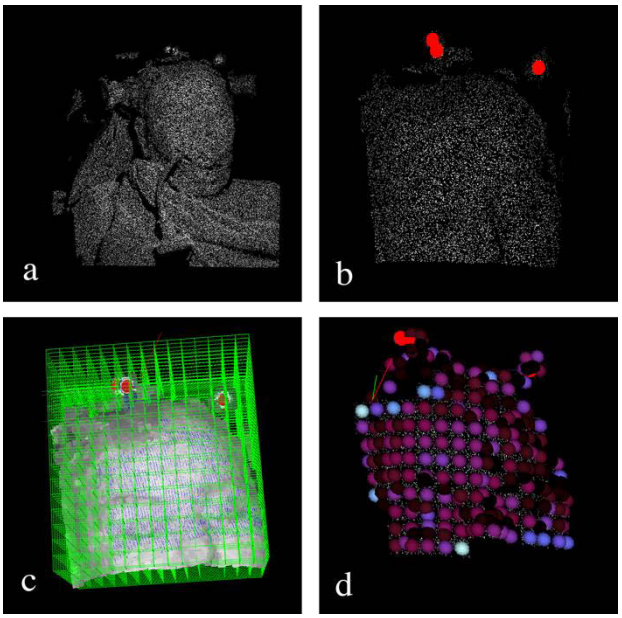


Figure 1 Creating facial deformation map:a) unfiltered point cloud b) filtered point cloud with markers found c) 3D grid superimposed over a point cloud d) deformation map where each face region is represented by a sphere, lighter color representing less deformable regions and vice versa

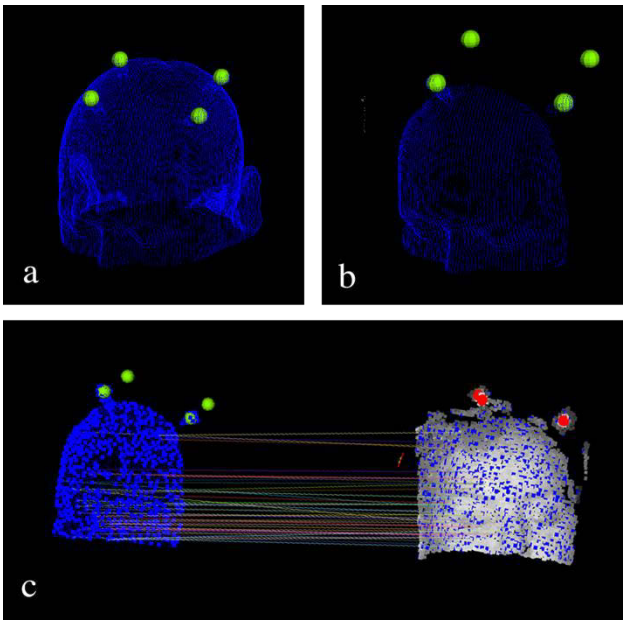


Figure 2 a) Preroperative point cloud with detected markrers b) filtered point cloud c) preoperative and operative point cloud with detected and matched keypoints

C. Patient registration

Process of markerless registration of preoperative and operative point clouds is done using keypoint matching. The process consists of five steps and is described below.

1. Downsampling

Firstly, the point cloud is downsampled using a voxel grid approach. This process reduces the size of the point cloud from approximately 100.000 points to 30.000 which significantly reduces computational time without affecting the registration accuracy.

2. Keypoint detection and descriptor calculation

Keypoints are highly-distinctive locations on a point cloud. Implementation of the algorithm described in [19] is used for the detection of Intrinsic Shape Signatures keypoints. Approximately 3.000 keypoints are detected on a 30.000 points point cloud. Descriptors are calculated for each keypoint. Implementation of the algorithm described in [20] is used to calculate the Signature of Histograms of OrienTations (SHOT) descriptors.

3. Finding correspondences

The goal of this step is to match similar keypoints by comparing their descriptors. K-nearest neighbor's search is used to match keypoints. An algorithm for matching keypoints is described below:

Suppose we have two point clouds: point cloud **A** and point cloud **B**. For each keypoint, $K_{\mathbf{A}}$ on cloud **A** find it's the nearest neighbour on the cloud **B**, $K_{\mathbf{B}}$. Store a new correspondence $K_{\mathbf{A}} \rightarrow K_{\mathbf{B}}$ and distance between keypoint's descriptors. For each keypoint $K_{\mathbf{B}}$ on the cloud **A** find it's the nearest neighbor on the cloud **B,** $K_{\mathbf{A}}$. Store a new

correspondence $K_B \rightarrow K_A$ and distance between keypoint's descriptors. Find keypoint pairs (K_A, K_B) for which both correspondences $K_A \rightarrow K_B$ and $K_B \rightarrow K_A$ exist. These keypoint pairs are the new correspondences.

4. Correspondence rejection

Not all correspondences will be correct and about half of them need to be filtered out. Random Sample Consensus is used to identify inliers. The algorithm is described in [21]. Point clouds with filtered correspondences can be seen on Figure 2.

5. Transformation calculation

Once correspondences are filtered, an Absolute Orientation algorithm is used to estimate the 6DOF transformation between the two clouds based on all remaining correspondences. Implementation of the method based on the Singular Value Decomposition (SVD) of the matrix representation of the 6DOF transformation proposed in [22, 23] is used.

After the point cloud is registered, the registration error is calculated as the average Euclidean distance between corresponding markers.



Figure 1 Flowchart describinng the whole process: data acquisition, generating facial deformation map and markerless registration

III. RESULTS

To evaluate facial deformation mapping a we have made a deformation map of the human face. The patient registration is done in three ways: using the whole point clouds, using only the face, and using only less deformable facial regions.

To achieve relatively accurate registration, point clouds need to be filtered first. This initial filtering is unrelated to the filtering using a facial deformation map. It is used to filter out all points that do not belong to the patient's face. On the operative point cloud, objects that are near the face, surgical drapes, the Mayfield clamp, and parts of the patient's head below the nose and above the forehead should be removed. On preoperative point cloud, the parts of the head below his nose and above the forehead, but also internal head structures, need to be removed so that only the face is preserved. For the initial filtering of the operative and preoperative point clouds, markers are used to determine the head orientation and all points outside of the predefined limits are removed. After the points outside of those limits are removed, outlier rejection is performed, followed by removing all points that do not belong to markers and have the intensity value grether than 250. This is done so that parts of the Mayfield clamp which are, besides markers, the brightest objects in the point cloud, are removed. Point cloud with 130.000 points is resized to 23.000 points after the initial filtering.

Next step is cloud filtering using facial deformation map, in which measurement of deformability for each facial region is. Filtering regions is done by comparing their defomability measurement with predefined threshold value. If deformability is greather than the threshold value, facial region is filtered out. The threshold for rejection of more deformable regions is set so that registration error is the lowest. If the threshold is set too high, only a few of the most deformable facial regions would be rejected and the improvement of the registration would be too small. On the other hand, setting the threshold value too low would remove most of the points and registration would be inaccurate. Final point cloud consists of approximately 10.000 points.

The registration error was calculated as the average Euclidean distance between corresponding markers on the preoperative and operative point clouds. Registration was done on four different operative point clouds of the same patient. Patient was scanned from different angles and distances by hand holding a 3D camera. When registering the preoperative and operative point clouds without initial filtering, the registration error was 1.99 mm. After the initial filtering was applied, the error was reduced to 0.45 mm. When facial regions with high deformity were filtered out and the registration was done using the filtered point cloud, the error was 0.15 mm. Registration errors for each point cloud can be seen in Table I: registration using the whole point cloud, registration using the point cloud of only the patient's face, and the registration of the face point cloud when more deformable facial regions are filtered out.

Table I. REGISTRATION ERRORS

	Whole point cloud	Face	Face filtered using deformation map
Point cloud 1	4.19	0.45	0.15
Point cloud 2	1.16	0.46	0.21
Point cloud 3	1.34	0.58	0.12
Point cloud 4	1.26	0.32	0.14
average	1.9	0.45	0.15

IV. CONCLUSION AND FUTURE WORK

Mapping facial deformation and using only less deformable regions of human face improved the registration accuracy drastically. When writing this paper, data from only one surgery was acquired and used for mapping the facial deformation due to the lack of surgical procedures during the COVID-19 period. In his paper we have presented our preliminary results and the methodology of generating a facial deformation map. In the future, more data will be gathered from numerous clinical procedures to create a more accurate deformation map of the human face. Although in this work, generating the facial deformation map and registration were done on the same patient, our plan is to generate generic facial deformation map which could be used for the registration in all future surgical procedures.

Besides creating a facial deformation map, the bone attached markers were also used for detecting facial regions as well as calculating the error. For a completely markerless approach, the detection of facial regions on a point cloud will have to be replaced with a new method. We are currently developing and testing a new method for detecting facial features using a cascade classifier on a 2D image taken with the Ensenso camera simultaneously with a 3D point cloud. Features on a 2D image will be then mapped on the 3D point cloud.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the project RONNA - Robotic neuronavigation, as well as the support of the European Regional Development Fund received through the project NERO – Neurosurgical robot. Autohors would also like to acknowledge the project CRTA - Regional Center of Excellence for Robotic Technologies funded by the European Regional Development Fund.

REFERENCES

[1] A.A. Wilfong, A. Shetty, D.J. Curry, STEREOTACTIC MRI-GUIDED LASER ABLATION OF EPILEPTOGENIC FOCI IN CHILDREN, Epilepsia 54 (2013) 280-280.

[2] B.E. Youngerman, A.K. Chan, C.B. Mikell, G.M. McKhann, S.A. Sheth, A decade of emerging indications: deep brain stimulation in the United States, Journal of Neurosurgery 125(2) (2016) 461- 471.

[3] P.A. LeWitt et al., AAV2-GAD gene therapy for advanced Parkinson's disease: a double-blind, sham-surgery controlled, randomized trial, Lancet Neurology 10(4) (2011) 309-319.

[4] G.F. Woodworth, M.J. McGirt, A. Samdani, I. Garonzik, A. Olivi, J.D. Weingart, Frameless image-guided stereotactic brain biopsy procedure: diagnostic yield, surgical morbidity, and comparison with the frame-based technique, Journal of Neurosurgery 104(2) (2006) 233-237.

[5] Jerbić B, Švaco M, Chudy D, et al. RONNA G4—Robotic Neuronavigation: A Novel Robotic Navigation Device for Stereotactic Neurosurgery. In: Handbook of Robotic and Image-Guided Surgery. Elsevier; 2020:599-625.

[6] Chudy D, Deletis V, Almahariq F, Marčinković P, Škrlin J, Paradžik V. Deep brain stimulation for the early treatment of the minimally conscious state and vegetative state: experience in 14 patients. Journal of Neurosurgery. Published online June 16, 2017:1-10.

[7] G. Minchev et al., A novel miniature robotic guidance device for stereotactic neurosurgical interventions: preliminary experience with the iSYS1 robot, Journal of Neurosurgery 126(3) (2017) 985-996.

[8] D. Langsdorff, P. Paquis, D. Fontaine, In vivo measurement of the frame-based application accuracy of the Neuromate neurosurgical robot, Journal of Neurosurgery 122(1):1-4, October 2014

[9] N.J. Brandmeir, S. Savaliya, P. Rohatgi, M.D. Sather, The comparative accuracy of the ROSA stereotactic robot across a wide range of clinical applications and registration techniques, Journal of Robotic Surgery, 2017.

[10] A. Raabe, R. Krishnan, R. Wolff, E. Hermann, M. Zimmermann, V. Seifert, Laser surface scanning for patient registration in intracranial image-guided surgery, Neurosurgery 50(4) (2002) 797-801.

[11] F. Suligoj, B. Jerbic, M. Svaco, B. Sekoranja, D. Mihalinec, J. Vidakovic, Ieee, Medical applicability of a low-cost industrial robot arm guided with an optical tracking system, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Ieee, Hamburg, GERMANY, 2015, pp. 3785- 3790.

[12] F. Meng, F. Zhai, B. Zeng, H. Ding, Guangzhi, An automatic markerless registration method for neurosurgical robotics based on an optical camera, *International Journal of Computer Assisted Radiology* and Surgery, 2017.

[13] Y.F. Fan, D.S. Jiang, M.N. Wang, Z.J. Song, A new markerless patient-to-image registration method using a portable 3D scanner, Medical Physics 41(10) (2014) 13.

[14] M. Potter, Z. Yaniv, Cost-Effective Surgical Registration Using Consumer Depth Cameras, Conference on SPIE Image-Guided Procedures, Robotic Interventions, and Modeling, San Diego, CA, 2016.

[15] S. Yoon, S. Shin, H. Cho, Y. Kim, L. Kim, D. Lee, G. Noh, Enhanced markerless surgical robotic guidance system for keyhole neurosurgery, Journal of Advanced Mechanical Design Systems and Manufacturing 11(4) (2017) 15.

[16] J.R. Cutter, I.B. Styles, A. Leonardis, H. Dehghani, Image-based registration for a neurosurgical robot: comparison using iterative closest point and coherent point drift algorithms, 20th Conference on Medical Image [Understanding and Analysis (Miua 2016) 90 (2016) 28-34.

[17] Dlaka D, Švaco M, Chudy D, et al. Brain biopsy performed with the RONNA G3 system: a case study on using a novel robotic navigation device for stereotactic neurosurgery. The International Journal of Medical Robotics and Computer Assisted Surgery. Published online December 12, 2017:1-7

[18] F. Šuligoj, M. Švaco, B. Jerbić, B. Šekoranja, J. Vidaković „Automated Marker Localization in the Planning Phase of Robotic Neurosurgery", IEEE Access, 5, June 2017.

[19] Yu Zhong, "Intrinsic shape signatures: A shape descriptor for 3D object recognition," Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on , vol., no., pp.689-696, Sept. 27 2009-Oct. 4 2009

[20] F. Tombari, S. Salti, L. Di Stefano Unique Signatures of Histograms for Local Surface Description. In Proceedings of the 11th European Conference on Computer Vision (ECCV), Heraklion, Greece, September 5-11 2010.

[21] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Commun. ACM, vol. 24, no. 6, pp. 381–395, 1981.

[22] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," Journal of the Optical Society of America A, vol. 4, no. 4, pp. 629–642, 1987.

[23] K. S. Arun, T. S. Huang, and S. D. Blostein, "Leastsquares fitting of two 3-D point sets," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 9, no. 5, pp. 698–700, 198.

Programiranje robota u učionici budućnosti

Damir Vrbanec
IV. osnovna škola Varaždin, Varaždin, Hrvatska
damir.vrbanec@skole.hr

Sažetak – Učenje programiranja trebalo bi rezultirati stvaranjem rješenja kao najvažnijim ishodom. Ono se mora temeljiti na postupku koji će učenike što više osloboditi od pisanja kôda, a što bolje naučiti stvarati apstraktne modele rješenja. Svakako je tu od jednake važnosti da se stvoreno rješenje može učenicima prikazati brzo i jednostavno. U ovom radu prikazani su primjeri takvog učenja programiranja uz pomoć edukacijskih robota.

Ključne riječi - robot, mBot, Codey Rocky, programiranje, kodiranje, učionica budućnosti.

I. UVOD

Rješavanju matematičkih problema nisu skloni svi učenici u svojem obrazovanju ili u svakodnevnom životu. Kako je programiranje usko povezano upravo sa rješavanjem matematičkih problema nije lako motivirati učenike i naučiti ih programirati. Učenje programiranja zahtijeva sposobnost dekompozicije stvarnog problema i stvaranje apstraktnog modela rješenja.

Zatim slijedi pisanje rješenja, odnosno kôda programa u nekom programskom jeziku. Pisanje kôda programa je za učenike vrlo naporan i demotivirajući posao jer ih i najmanja pogreška u pisanju kôda dovodi do neuspjeha.

Kod učenja programiranja važnije je stvoriti rješenje nego napisati sam programski kôd. Učenje programiranja bi se trebalo temeljiti na postupku koji će učenike što više osloboditi od pisanja kôda, a što bolje naučiti stvarati apstraktne modele rješenja. Svakako je tu od jednake važnosti da se stvoreno rješenje može učenicima prikazati brzo i jednostavno. U ovom radu prikazani su primjeri takvog učenja programiranja uz edukacijske robote.

Osnovni cilj ovog rada je prikazati metodologiju učenja programiranja pomoću edukacijskih robota. Ostali ciljevi bili bi:

- usmjeriti učenike na mrežne stranice za podučavanje programiranja korištenjem alata za programiranje i edukacijskih robota,
- učenicima omogućiti jednostavno učenje programiranja kroz zabavne i jednostavne primjere uz korištenje različitih digitalnih uređaja: PC računala, mobilnih telefona, tableta, Bluetooth tehnologije i sl.

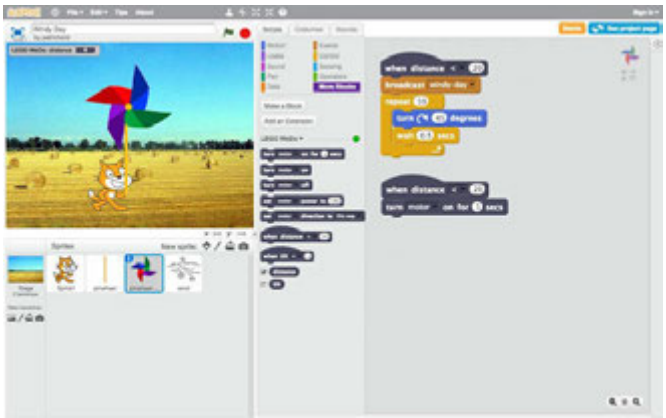
II. SCRATCH

Scratch je programski jezik koji omogućava jednostavno stvaranje interaktivnih priča, igara i animacija, te dijeljenje radova preko Interneta.

Sve što je potrebno za besplatnu instalaciju program nalazi se na poveznici Scratch.mit.edu.

Nakon jednostavne instalacijske procedure u programu Scratch kreće se sa izradom vlastitog projekta ili otvoranjem već napravljenog projekta u Scratch-u kojeg možemo izmijeniti po želji i potrebi.

Programiranje je krajnje pojednostavljeno i uspjeh je zagarantiran vrlo brzo.



Slika 1. – Sučelje programa Scratch

Početno učenje i upoznavanje sa programom Scratch:

https://www.stemlittleexplorers.com/hr/ucenje-programiranja-scratch/

Pregled svih alata Scratch programa je dostupno na Vodiču za korisnike:

https://scratch.mit.edu/projects/editor/?tutorial=all

III. POČETNO UČENJE PROGRAMIRANJA UZ ROBOT CODEY ROCKY

Programiranje robota Codey Rocky namjenjeno je učenicima mlađe dobne skupine (1. – 4. razred).



Slika 2. – Robot Codey Rocky

Robot je odličan i zabavan za početno učenje računalnog razmišljanja i kodiranja.

Robot se programira u jednostavnom grafičkom sučelju programa *mBlock*. Grafičko programiranje „povuci i ispusti“ računalno-grafičkog programa mBlock V5 razvijenog na temelju ***Scratch*** *2.0.* omogućuje djeci da brzo nauče programiranje, kontroliraju robota i realiziraju sve više funkcija s kojima robot Codey Rocky raspolaže.

Robot Codey Rocky ima više od deset naprednih elektroničkih modula i sigurno je kako spada među zanimljivije edukacijske robote uz robota mBot-a.

Pogledajmo programski kôd napisan u mBlock-u kako bi Codey Rocky izbjegavao prepreke ispred sebe:



Slika 3. – Programski kôd

Kad uredimo programski kôd moramo ga prenijeti na sam robot funkcijom *Prijenos,* a robot mora biti povezan USB kablom na računalo.



Slika 4. – Prijenos programskog kôda na robota

Programski kod sprema se u oblaku programa mBlock-i ili ga možemo spremati lokalno u spremnike računala.

Više o primjerima, zadacima i programskom kodu potražite na mrežnim stranicama:

http://www.os-cetvrta-vz.skole.hr/Informatika/codey-rocky/

IV. PROGRAMIRANJE ROBOTA MBOT

mBot je cjelokupno rješenje za učenika koji daje praktično iskustvo u programiranju, elektronici i robotici. Rad s mBot robot setom inspiriran je Scratch 2.0 programom kojeg se može povezati s mBot uređajem koristeći USB kabel, Bluetooth ili 2.4GHz wireless modul.

Ovaj edukacijski robot jednostavan je za sastavljanje. mBot set pruža široke mogućnosti da djeca lakše i uz igru nauče osnove STEAM-a (znanost, tehnologiju, inženjering i matematiku).



Slika 5. – Robotska kolica mBot

Robot sadrži 38 sastavnih dijelova koji se mogu složiti u 10 minuta, a jedinstveno obojani RJ25 priključci za žice osiguravaju više vremena za programiranje i kreativnost.



Slika 6. – mBot set za sastavljanje

Za više informacija o mBot-u, sastavljanju robota, prvom pokretanju i instalaciji mBlock softvera proučite uz PPT prezentaciju na slijedećoj poveznici:

https://drive.google.com/file/d/1TMjf2BVriKPly4v229khQMxcIGTg77Xr/view

V. GRAFIČKO PROGRAMIRANJE

„Povucite i ispustite“ računalno grafički softver mBlock V3 ili mBlock V5 razvijenog na temelju Scratch 2.0. ili mBlock V5.2.0 za Scratch3.0 omogućuje učenicima da brzo nauče programiranje, kontroliraju robota i realiziraju sve više funkcija s kojima mBot raspolaže.

mBot se može programirati kao robot koji slijedi linije, baca loptice i gura predmete, izbjegava zidove i još mnogo toga.

Programirati se ovog robota može i jezičnim kodiranjem u Arduino modu.

Jednostavan primjer upravljanja robotom mBot u sučelju mBlock-a pogledajte u ovom videu:

https://www.youtube.com/watch?v=36a6vp-xPTE#action=share

Zanimljiv projektni zadatak kroz koji učenici na jednostavan način mogu naučiti upravljati osnovnim funkcijama robota mBot kroz digitalnu i interaktivnu priču pogledajte ovdje:

https://damirvrbanec.itch.io/mbot-novi-pristup-programiranju

Više o primjerima, zadacima i programskom kodu potražite na mrežnim stranicama:

http://www.os-cetvrta-vz.skole.hr/Informatika/mbot/



Slika 7. – Sučelje programa mBlock i robot mBot

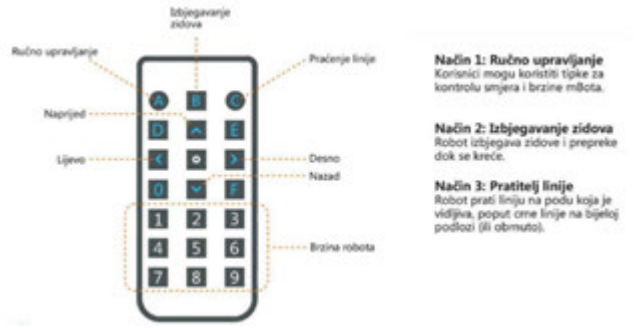
VI. PROJEKTNI ZADATAK

Uz interaktivnu priču (https://damirvrbanec.itch.io/mbot-novi-pristup-programiranju) cilj je da učenici nauče napraviti program za robota mBot-a da vozi stazom kao u ovom videu:

https://www.youtube.com/watch?v=Mg08bmCg9Xo&feature=youtu.be

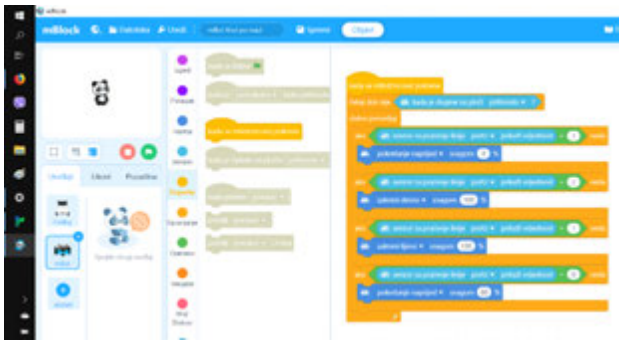
Robot mBot su zapravo robotska kolica pa prvo krenimo voziti robota naprijed, natrag i naučiti skretati. Nakon što se robot sastavi on je već spreman i programiran za rad (u tvornici su ga “naučili ponešto”). Postavite robota na ravnu podlogu (najbolje na pod, da

ne padne sa stola) i uključite ga pomoću sklopke na njegovoj upravljačkoj pločici, a nakon toga možete pomoću daljinskog upravljača upravljati robotom.



Slika 8 – Daljinski upravljač mBot-a

Sljedeći programski kod "naučiti" će robota da slijedi liniju na stazi i tako vozi u krug ili do kraja linije na stazi. Za rješenje ovog zadatka koristiti ćemo jedan od senzora mBot-a koji se zove senzor za praćenje linije (eng. *Line Follower*).
Usput ćemo naučiti robota da pričeka pritisak tipkala prije nego krene s vožnjom.



Slika 9. – Programski kod za mBot u programu mBlock

Na sljedećem videu može se vidjeti spomenuti programski kod i njegov rezultat – vožnja robota po stazi:

https://www.youtube.com/watch?v=E22yXqwj1cA&feature=youtu.be

Svi učenice vole biti nagrađeni za svoj trud stoga svatko tko dođe do kraja interaktivne piče dobiva digitalnu značku kao što je na slici.

Mnogo više od digitalne značke učenicima vrijedi novo znanje i vještine koje su stekli uz igru i zabavu. Zapravo neprimjetno oni uz igru nauče programirati robota.

Slika 10. – Digitalna značka



LITERATURA

[1] A. Budojević, A. Kanić, "Istražite Scratch 3.0 – priručnik za rano učenje programiranja", Školska knjiga 2019.
[2] P. Zenzerović, "BBC micro:bit – službeni priručnik ", VIDI 2018.
[3] S. Parker, O. Davey, "Sastav robota", Školska knjiga 2015.

ISS

International Conference on

INFORMATION SYSTEMS SECURITY

Steering Committee

Chairs: Stjepan Groš, University of Zagreb, Croatia
Tonimir Kišasondi, Oru, Varaždin, Croatia
Mario Spremić, University of Zagreb, Croatia

PAPERS

Genetic algorithm and artificial neural network for network forensic analytics

Dijana Oreški, Darko Andročec
Faculty of Organization and Informatics, University of Zagreb, Varaždin, Croatia
dijana.oreski@foi.unizg.hr, dandrocec@foi.unizg.hr

Abstract **-** Rapid development of Internet of things (IoT) technologies and their application and importance within various fields arises security issues. New threats require development of appropriate approaches to address them since information security problems could led to serious damages. This work focuses on developing methods for prediction of undesired behavior. Literature review indicated use of advanced statistical approaches such as logistic regression or multiple regression. However, in the recent years, interest among researchers for applying artificial intelligence techniques is growing. Artificial intelligence approaches shown to be powerful tool for development of efficient predictive models in various fields. Main aim of research presented here is to apply artificial intelligent techniques for intrusion analysis. Our approach is based on the neural networks and genetic algorithms. Neural networks results largely depend on the network parameters which are mostly achieved by trial-and-error. Trial-and-error approach requires a lot of time. Thus, we are applying genetic algorithm to optimize neural networks parameters. Experiments are conducted on the publicly available new dataset, Bot-IoT, consisting of legitimate and simulated IoT network traffic incorporating different types of attacks. Here, we investigate: (i) the level to which available data can be a good basis for predicting intrusion, (ii) efficiency of neural network approach supported by genetic algorithm for developing useful predictive models.

Keywords - intrusion detection, machine learning, internet of things, security, neural networks, genetic algorithm.

I. INTRODUCTION

Development of Internet of Things (IoT) and emergence of various successful examples (smart homes, smart cities...) attracted interest of cyber attackers. Various threats have emerged able to damage IoT networks, or to use IoT devices for malicious purposes. Internet of things devices and services are often not designed with security in mind, so malicious users can create botnets and other malicious software targeting things', IoT services', and IoT networks' vulnerabilities. The term malicious in network forensics refers to malicious packets and malicious traffic program (irregular traffic patterns) [1]. Malicious packets can be defined as the packets that violate the principles of network communication by exploiting vulnerabilities in security devices including gateways and gaining unauthorized access to the network hosts. IT security has always been based around standards, but the relatively recent emergence of IoT devices and applications makes it difficult to create standards and enforce them in the industry. Some examples already exist where a large number of limited IoT devices was used to launch a large scale attack (e.g., distributed denial of service (DDoS) attacks such as Remaiten and Mirai). IoT networks can both amplify and be the targets of DDoS or botnet attacks.

IoT botnets are a collection of smart devices hijacked by cybercriminals to conduct attacks like DDoS or activities like cryptocurrency mining [2]. IoT botnets often use old vulnerabilities, cryptocurrency mining, hiding their command and conquer center, and employing unique evasion tactics. Trend Micro recommends the following best practices to secure IoT devices: start security from the design phase, apply patches as soon as they become available, use secure configuration, use strong and hard to guess passwords, and apply network segmentation [2]. In IoT environments, it is also important to develop forensics methods for successful identification of inappropriate or suspicious behavior. Many tools for real-time network traffic exist, but real-time monitoring at any level requires significant human and hardware resources, so it is generally more practical to archive all traffic and analyze subsets as necessary. This process is known as reconstructive traffic analysis, or network forensics [3]. In this work, we investigate performance of hybrid methodology consisting of contrast mining techniques in feature selection phase and combination of genetic algorithm and neural networks in predictive modelling for IoT attack detection task.

The remaining sections of this article are organized as follows: In Section 2, a related work (current state-of-the-art) is given. Section 3 describes steps of the analysis by explaining dataset used in the research with a brief overview of techniques. Section 4 gives results with the focus on prediction accuracy. Finally, Section 5 concludes this paper and gives guidelines for future research.

II. RELATED WORK

Researchers and professionals in network forensics field require new tools and techniques to tackle new attack trends. Tools to assist with network forensics come in a variety of forms: packet sniffers, fingerprinting, mapping, location identification, email traffic, trace back services, and honeypots [4]. Khan et al. [1] reviewed the fundamental mechanism of network forensics techniques through an extensive review of related literature. They have proposed a thematic taxonomy for the classification of current network forensics techniques based on its

implementation as well as involved target data sets. Authors also listed the most important network forensics challenges as: intelligent network forensics tools, data extraction locations, access to IP address, data privacy, data integrity, data storage on the network devices, and high speed data transmission.

Even though IoT is slowly been assimilated in one form or another in everyday life, there is no doubt that there exist a number of issues with respect to security and privacy [5]. Zhang et al. [6] have investigated general information security background of IoT and continued on with information security related challenges that IoT will encounter. They divided the ongoing IoT research work into the following categories: object identification and locating in IoT; authentication and authorization in IoT; privacy in IoT; lightweight cryptosystems and security protocols; software vulnerability and backdoor analysis in IoT; malware in IoT; and Android platform. As the main challenges for IoT security, Zhang et al. [6] identified the heterogeneity and the large scale of smart things. Dorri et al. [7] showed how to use blockchain technology for IoT security and privacy on the case study of a smart home. Khan and Salah [8] surveyed major security issues for IoT and outlined security requirements for IoT along with the existing attacks, threats, and state-of-the-art solutions. They also discussed basic characteristics of the blockchain based security solutions and analysis of their effectiveness for securing IoT. Blockchain can be used to register and give identity to connected IoT devices. Additionally, blockchain smart contracts have the ability to provide a de-centralized authentication rules and logic to be able to provide single and multiparty authentication to an IoT Device [8]. As open challenges, Khan and Salah [8] have listed the following: IoT resource limitations; heterogeneous devices; interoperability of security protocols; single points of failure; hardware/firmware vulnerabilities; trusted updates and management; and blockchain vulnerabilities.

Major IoT security issues include IoT botnets. Kolias et al. [9] described the Mirai variants and imitators as samples of sophisticated IoT botnets. DDos attacks by Mirai and other similar botnets shows how IoT devices can pose the security risks. The main reasons malicious users choose IoT devices for botnets are: constant and unobtrusive operation of IoT devices, lack of their security, poor maintenance, considerable possible attack traffic, and no interactive or minimally interactive user interfaces. Meidan et al. [10] proposed a network-based anomaly detection method for the IoT that uses deep autoencoders to detect anomalous network traffic from compromised IoT devices. The evaluation was done by detecting IoT-based botnets Mirai and BASHLITE.

There are many existing works on IoT network forensics and using machine learning methods and techniques to tackle IoT security problems. Andročec and Vrček [11] systematically reviewed the state-of-the art to classify the research on machine learning for the IoT security, and concluded that machine learning methods were mostly used for IoT intrusion detection and authentication. The most mentioned machine learning algorithms or techniques in the primary studies of their systematic review were: Support Vector Machine, Artificial Neural Network, Naïve Bayes, Decision Tree, kNN, k-Mean, and Random forest. Koroniotis et al. [5] investigate machine learning techniques to develop IoT network forensic mechanism to track botnets. They have used the UNSW-NB15 dataset to detect botnets' attacks and their tracks. Their proposed network forensic architecture consists of traffic collection, network feature selection, machine learning techniques, and evaluation metrics. They have experimented with the following machine learning methods: Association Rule Mining (ARM), Artificial Neural Network (ANN), Naïve Bayes (NB), and Decision Tree (DT). Hossain et al. [12] proposed a forensic investigation framework called Probe-IoT using a public digital ledger to find facts in criminal incidents in IoT-based systems. It collects interactions among various IoT entities as evidence and stores them securely as transactions in public, distributed and decentralized blockchain network.

Xiao et al. [13] reviewed the IoT security solutions based on machine-learning techniques and methods including supervised learning, unsupervised learning, and reinforcement learning. The main challenges to implement the machine learning-based security techniques in IoT are partial state observations, computation and communication overhead of machine learning-based security schemes to IoT devices and networks. Supervised and unsupervised machine learning sometimes fail to detect the attacks, so backup security solutions have to be designed and incorporated. Hussain et al. [14] systematically reviewed the security requirements, attack vectors, and the current machine learning (ML) and deep learning (DL)-based security solutions for the IoT networks. They also claim that the datasets needed for ML and DL algorithms are still scarce, which makes benchmarking the efficiency of the ML- and DL-based security solutions a difficult task. Chatterjee et al. [15] presented their deep neural network-based framework that allows real-time authentication of wireless nodes. The proposed method eliminates the need for preamble-based or key-based identification of modern IoT nodes and enables low-cost secure authentication using the intrinsic properties of the radio-frequency signal.

Miettinen et al. [16] developed a system capable of automatically identifying the types of IoT devices and enabling enforcement of rules for constraining the communications of vulnerable devices. The device-type identification is based on monitoring the communication behavior of IoT devices to generate device-specific fingerprints with the help of a machine learning-based classification model. Shakeel et al. [17] introduced the deep learning based Deep-Q-Networks for reducing the malware attacks while managing the health information.

Their system examines the various malwares and the unauthorized access in IoT-based health care systems. Deep convolution neural networks were used to maintain authentication in IoT medical data transactions.

III. EMPIRICAL ANALYSIS

Primary aim of this study is to investigate performance of hybrid methodology consisting of contrast mining techniques in feature selection phase and combination of genetic algorithm and neural networks in predictive modelling for attack detection task. Feature selection is used for selecting the most relevant features from datasets [18] and removing irrelevant features. There are many benefits of feature selection: reducing the dimensionality of dataset so that the dataset is easier to handle when performing modelling and being able to reveal the relevant patterns within the dataset [19]. These benefits led to increasing use of feature selection in many domains. Feature selection is especially important in the era of big data, when certain tasks would be impossible to process without prior dimensionality reduction. One of such examples is an analysis of realistic Bot-IoT datasets presented here. Field of forensic requires big data sources for developing effective models in dealing with attacks.

In the first section of this chapter we are describing used dataset, whereas in the second section we are explaining methods used for data analysis.

A. Data description

Increasing development of public data repositories in various domains has created a database which makes data mining highly effective since data about various topics is easily and rapidly accessed. In this research, we have used dataset which is developed by Koroniotis, Moustafa, Sitnikova, and Turnbull [20]. Authors designed a new realistic Bot-IoT dataset in IoT networks. Description of designing the testbed configuration and simulated IoT sensors can be found in their paper [20]. Dataset is composed of 3 668 522 instances. It is to be noted that the original dataset was very large consisting of 72 000 000 instances and analyzing such data was very challenging. So, they have extracted 5% of the original dataset. Dataset contains, for each instance, 46 features. Such reduced dataset was also used in our research. Some of the features from original dataset have been transformed into a series of binary features so that they can be appropriately handled by the neural networks.

B. Methods overview

Koroniotis et al. [20] did statistical analysis of the explained dataset using Correlation Coeficient and Joint Entropy techniques in the feature selection. In this research we have used contrast mining approach for feature selection based on the STUCCO algorithm: SfFS (STUCCO for Feature Selection). SfFS application in feature selection gave the best results in our previous research when analyzing NSL-KDD data set for intrusion detection [21]. Idea of SfFS is first presented in [22]. SfFS tries to identify differences between two groups (in our case normal versus attack) by contrast sets which are defined as conjunctions of attributes and values that differ meaningfully in their distribution across groups. SfFs in feature selection is using threshold as cutting criterion. As an evaluation measure, relevance is used. It is defined as a measure which discriminates between features on the basis of their potential in forming rules. Contrast mining techniques are essentially defined to give the rules and measures of the quality of rules (measure that differs features with respect to their potential in defining rules) as the result. SfS in feature selection yielded with ten best features presented in Table 1.

TABLE 1. BEST 10 FEATURES EXPLANATION

Feature	***Explanation***
drate	Destination-to-source packets per second
flags number	Numerical representation of feature flags
max	Maximum duration of aggregated records
mean	Average duration of aggregated records
min	Minimum duration of aggregated records
N IN Conn P DstIP	Number of inbound connections per destination IP.
N IN Conn P SrcIP	Number of inbound connections per source IP.
seq	Argus sequence number
state number	Numerical representation of feature state
stddev	Standard deviation of aggregated records
attack	Class label: 0 for Normal traffic, 1 for Attack Traffic

Ten features are selected since Koroniotis et al. [20] also selected 10 features. As such, our results will be comparable with their work. Koroniotis et al. [20] identified following ten best features: srate, drate, rate, max, state number, mean, min, stddev, gs number, seq. Our approach for feature selection, contrast set mining, yielded features explained in Table 1.

Koroniotis et al. [20] evaluated the performance of network forensic methods by applying three machine learning algorithms. The models that were trained were: Support Vector Machine (SVM), Recurrent Neural Network (RNN) and Long-Short Term Memory Recurrent Neural Network (LSTM-RNN). Neural network are shown to give the best results. There are many linear and non-linear machine learning methods available for predictive modelling. Artificial neural network became one of the most popular one since they are considered robust learners and perform well on a wide range of applications.

Scientific contribution of the research presented here is in the methodology. Contrast set mining for feature selection is applied here, following with combination of genetic algorithm and neural networks in modelling phase. Novel combination of methods have been used in order to train and test data to generate significant features and develop highly accurate models. A novel combination of

methods in predictive modelling consist of genetic algorithms and neural networks.

Development of neural networks consists of the following steps: preparation of data and modeling, training and testing neural networks, analysis of results and selecting the best model. Before training can begin, we must define network architecture by specifying the number of units in the input layer, the number of hidden layers, the number of units in each hidden layer, and the number of units in the output layer. There are no rules as to the 'best' number of hidden layers or best number of neurons in the hidden layer. Network architecture definition is a trial-and-error process and largely affects the accuracy of the resulting trained network. Although suggestions are made for the number of neurons in the hidden layer range from one-half the number of input neurons [23], to two times the number of input neurons plus one [24] and to number of neurons in the hidden layer estimated as [25]: The number of neurons in the hidden layer = 2* $\sqrt{}$ (number of inputs + number of outputs). Since determining the size of the network is extremely important for network performances (if the network is too small it may not reach an acceptable level of accuracy, or if there are too many neurons it may result in an inability for the network to generalize as a universal approximator), we are optimizing number of hidden layers and hidden neurons by applying genetic algorithm. Oreski et al. [26] presented genetic algorithm application for neural network optimization. Their application of such methodology shown promising results in credit risk assessment. Several other authors applied this methodology in various different domains: e.g.: Sharma and Gedeon, 2013. [27], Beheshti et al., 2014. [28], Chiroma et al. 2017. [29], Kaiser et al., 2019. [30].

This approach represents automated technique that searches for quality parameters of neural network based on genetic algorithm. Genetic algorithm is applied to the problem of parameterization of the neural network.

The type of neural network used was multilayer feed-forward network with backpropagation network algorithm and activation function used was tangent hyperbolic. Number of training cycles varied from 300 till 600, and learning rate in range from 0.3 till 1.0.

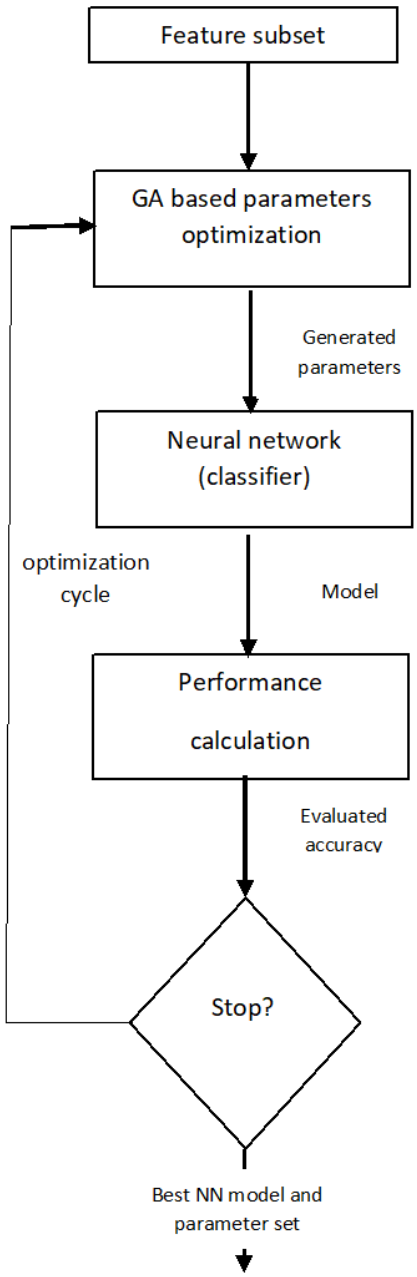


FIGURE 1. NEURAL NETWORK – GENETIC ALGORITHM ARCHITECTURE

IV. RESEARCH RESULTS

In the data analyses, we first performed feature selection and identified strong features for discriminating between the classes: attack or normal. We then assess the classification accuracy of the proposed methodology. There are several ways for testing and evaluation the accuracy of predictive models. We use the *k*-fold cross validation. This approach divides the data set into *k* subsets. Each time, one of the *k* subsets is used as the test set and the other *k-1* subsets consist the training set.

Parameters are calculated for all *k* tests. This allows good generalization of results since we get indication of how well the classifier will perform on new, unseen data. We use *k=10* and calculate confusion matrix along with four standard performance metrics: accuracy (demonstrating percentage of correctly classified records over the total number of records), precision (indicating number of class members classified correctly over the total number of instances classified as class members), recall (demonstrating the number of class members classified correctly over the total number of class members) and F1 measure as weighted average

of precision and recall. Out of 3 668 522 instances in dataset, 477 of them belongs to class *normal,* whereas 3 668 045 of them belongs to class *attack*. Table 2 represents results of analyzing dataset consisting of 10 best features after feature selection.

TABLE 2. CONFUSION MATRIX FOR TEN BEST FEATURES

True/Predict	Normal (0)	Attack (1)
Normal (0)	472	5
Attack (1)	6354	3661691

Table 3 indicates results of analyzing full dataset consisting of all features without selection.

TABLE 3. CONFUSION MATRIX FOR FULL DATASET

True/Predict	Normal (0)	Attack (1)
Normal (0)	452	25
Attack (1)	226354	3441691

Table 4 gives measures of accuracy, precision, recall and F1 for both, full dataset and dataset consisting of 10 best features.

TABLE 4. COMPARISON OF RESULTS

	10-best	Full dataset
Accuracy	0.998267	0.9382915
Precision	0.999999	0.99999274
Recall	0.998268	0.93829029
F1	0.999132	0.96815941

Our approach produced better results compared to Koroniotis et al. [20] on dataset consisting of 10 best features. Contrast set mining in feature selection identified different relevant features for given task then correlation coefficient and entropy. Contrast mining approach is focused on finding variables which make the difference between values of dependent variable. Furthermore, genetic algorithms did optimization of neural network parameters. Such approach outperforms trial and error and results with models of higher accuracy and reliability.

V. CONCLUSION

Due to heterogeneity of smart things, the security of smart things, IoT services and IoT networks is a complex practical and research problem. In this paper, we have evaluated the ability of machine learning algorithms to identify cyber-attack. We show that there is great potential in using this method. Experimental results demonstrated superiority of proposed feature selection (contrast mining) and predictive modelling approaches (genetic algorithm and neural network) with respect to previously used approach of correlation and entropy in feature selection and SVM or neural networks in classification. Nevertheless, there are some limitations of this work which should be taken into account when interpreting the results. First, contrast mining techniques in feature selection are defined with the assumption of feature independence. This approach has some advantages, but there is a limitation when features interact. Second, our approach is evaluated only on one dataset. In future work we should extend research by performing the evaluation on various data sets regarding this domain. Third, we were faced with highly imbalanced dataset, which can cause the learner to be biased towards one class: *attack*. This bias is the result of class *normal* being heavily under represented compared to the other class: *attack*. In the future research we will use methods for dealing with class imbalance: introduce weighting schemes that will give instances of the *normal* class a higher weight during training, duplicating training examples of the class *normal,* removing examples of the *attack* class. We will use each of the approaches to deal with the data imbalance to identify which one fits best for the application at hand.

REFERENCES

[1] S. Khan, A. Gani, A. W. A. Wahab, M. Shiraz, and I. Ahmad, "Network forensics: Review, taxonomy, and open challenges," *J. Netw. Comput. Appl.*, vol. 66, pp. 214–235, May 2016, doi: 10.1016/j.jnca.2016.03.005.

[2] "Into the Battlefield: A Security Guide to IoT Botnets - Security News - Trend Micro USA." [Online]. Available: https://www.trendmicro.com/vinfo/us/security/news/internet-of-things/into-the-battlefield-a-security-guide-to-iot-botnets. [Accessed: 15-Jan-2020].

[3] V. Corey, C. Peterman, S. Shearin, M. S. Greenberg, and J. Van Bokkelen, "Network forensics analysis," *IEEE Internet Comput.*, vol. 6, no. 6, pp. 60–66, Nov. 2002, doi: 10.1109/MIC.2002.1067738.

[4] R. Hunt and S. Zeadally, "Network Forensics: An Analysis of Techniques, Tools, and Trends," *Computer*, vol. 45, no. 12, pp. 36–43, Dec. 2012, doi: 10.1109/MC.2012.252.

[5] N. Koroniotis, N. Moustafa, E. Sitnikova, and J. Slay, "Towards Developing Network Forensic Mechanism for Botnet Activities in the IoT Based on Machine Learning Techniques," in *Mobile Networks and Management*, Cham, 2018, pp. 30–44, doi: 10.1007/978-3-319-90775-8_3.

[6] Z.-K. Zhang, M. C. Y. Cho, C.-W. Wang, C.-W. Hsu, C.-K. Chen, and S. Shieh, "IoT Security: Ongoing Challenges and Research Opportunities," in *2014 IEEE 7th International Conference on Service-Oriented Computing and Applications*, 2014, pp. 230–234, doi: 10.1109/SOCA.2014.58.

[7] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, "Blockchain for IoT security and privacy: The case study of a smart home," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2017, pp. 618–623, doi: 10.1109/PERCOMW.2017.7917634.

[8] M. A. Khan and K. Salah, "IoT security: Review, blockchain solutions, and open challenges," *Future Gener. Comput. Syst.*, vol. 82, pp. 395–411, May 2018, doi: 10.1016/j.future.2017.11.022.

[9] C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas, "DDoS in the IoT: Mirai and Other Botnets," *Computer*, vol. 50, no. 7, pp. 80–84, 2017, doi: 10.1109/MC.2017.201.

[10] Y. Meidan *et al.*, "N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders," *IEEE Pervasive Comput.*, vol. 17, no. 3, pp. 12–22, Jul. 2018, doi: 10.1109/MPRV.2018.03367731.

[11] D. Andročec and N. Vrček, "Machine Learning for the Internet of Things Security: A Systematic Review," in *Proceedings of the*

13th International Conference on Software Technologies, Porto, Portugal, 2018, pp. 563–570, doi: 10.5220/0006841205630570.
[12] M. Hossain, R. Hasan, and S. Zawoad, "Probe-IoT: A public digital ledger based forensic investigation framework for IoT," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2018, pp. 1–2, doi: 10.1109/INFCOMW.2018.8406875.
[13] L. Xiao, X. Wan, X. Lu, Y. Zhang, and D. Wu, "IoT Security Techniques Based on Machine Learning: How Do IoT Devices Use AI to Enhance Security?," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 41–49, Sep. 2018, doi: 10.1109/MSP.2018.2825478.
[14] F. Hussain, R. Hussain, S. A. Hassan, and E. Hossain, "Machine Learning in IoT Security: Current Solutions and Future Challenges," *ArXiv190405735 Cs Stat*, Mar. 2019.
[15] B. Chatterjee, D. Das, S. Maity, and S. Sen, "RF-PUF: Enhancing IoT Security Through Authentication of Wireless Nodes Using In-Situ Machine Learning," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 388–398, Feb. 2019, doi: 10.1109/JIOT.2018.2849324.
[16] M. Miettinen, S. Marchal, I. Hafeez, N. Asokan, A.-R. Sadeghi, and S. Tarkoma, "IoT SENTINEL: Automated Device-Type Identification for Security Enforcement in IoT," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, 2017, pp. 2177–2184, doi: 10.1109/ICDCS.2017.283.
[17] P. Mohamed Shakeel, S. Baskar, V. R. Sarma Dhulipala, S. Mishra, and M. M. Jaber, "Maintaining Security and Privacy in Health Care System Using Learning Based Deep-Q-Networks," *J. Med. Syst.*, vol. 42, no. 10, p. 186, Aug. 2018, doi: 10.1007/s10916-018-1045-z.
[18] C. Sima and E. R. Dougherty, "The peaking phenomenon in the presence of feature-selection," *Pattern Recognit. Lett.*, vol. 29, no. 11, pp. 1667–1674, Aug. 2008, doi: 10.1016/j.patrec.2008.04.010.
[19] T. Czekaj, W. Wu, and B. Walczak, "Classification of genomic data: Some aspects of feature selection," *Talanta*, vol. 76, no. 3, pp. 564–574, Jul. 2008, doi: 10.1016/j.talanta.2008.03.045.
[20] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," *Future Gener. Comput. Syst.*, vol. 100, pp. 779–796, Nov. 2019, doi: 10.1016/j.future.2019.05.041.
[21] D. Oreski and D. Androcec, "Hybrid Data Mining Approaches for Intrusion Detection in the Internet of Things," in *2018 International Conference on Smart Systems and Technologies (SST)*, Osijek, 2018, pp. 221–226, doi: 10.1109/SST.2018.8564573.
[22] D. Oreski and B. Klicek, "A novel feature selection techniques based on contrast set mining," in *14th International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED'15)*, Tenerife, Spain, 2015.
[23] J. Lawrence, *Introduction to neural networks: design, theory, and applications*, 6. ed. Nevada City, Calif: California Scientific Software, 1994.
[24] R. L. Wilson, "Business implementation issues for neural networks," *J. Comput. Inf. Syst.*, vol. 32, pp. 15–19, 1992.
[25] A. M. Flitman, "Towards analysing student failures: neural networks compared with regression analysis and multiple discriminant analysis," *Comput. Oper. Res.*, vol. 24, no. 4, pp. 367–377, Apr. 1997, doi: 10.1016/S0305-0548(96)00060-3.
[26] S. Oreski and G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 2052–2064, Mar. 2014, doi: 10.1016/j.eswa.2013.09.004.
[27] N. Sharma and T. Gedeon, "Hybrid Genetic Algorithms for Stress Recognition in Reading," in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, vol. 7833, L. Vanneschi, W. S. Bush, and M. Giacobini, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 117–128.
[28] Z. Beheshti, S. M. Hj. Shamsuddin, E. Beheshti, and S. S. Yuhaniz, "Enhancement of artificial neural network learning using centripetal accelerated particle swarm optimization for medical diseases diagnosis," *Soft Comput.*, vol. 18, no. 11, pp. 2253–2270, Nov. 2014, doi: 10.1007/s00500-013-1198-0.
[29] H. Chiroma *et al.*, "Neural Networks Optimization through Genetic Algorithm Searches: A Review," *Appl. Math. Inf. Sci.*, vol. 11, no. 6, pp. 1543–1564, Nov. 2017, doi: 10.18576/amis/110602.
[30] C. Kaiser, A. Ahuvia, P. A. Rauschnabel, and M. Wimble, "Social media monitoring: What can marketers learn from Facebook brand photos?," *J. Bus. Res.*, p. S0148296319305429, Sep. 2019, doi: 10.1016/j.jbusres.2019.09.017.

Using FireEye Endpoint Security for educational purposes

Marijo Dujmić, Damir Delija, Goran Sirovatka, Marinko Žagar
* TVZ, Zagreb, Hrvatska
Marijo.Dujmic@tvz.hr
Damir.Delija@tvz.hr
Goran.Sirovatka@tvz.hr
Marinko.Zagar@tvz.hr

Abstract **In this paper, it is presented how a commercial cybersecurity tool "FireEye Endpoint Security" can be used in the hybrid installation as a part of practical student exercises for cybersecurity training in the Network Forensics and Malware Forensics classes. The purpose of the student exercise in such scenario is to understand the capabilities of enterprise-class cybersecurity tools while keeping costs minimal for the university. The working environment is the installation of remote virtual security appliances and university laboratory equipment for independent students' work. This is a scaled-down virtual environment available for independent student work where FireEye Endpoint Security is used in demo and proof-of-concept mode, which requires minimal local installation while providing a full feature set of the tool.**

Keywords - Cybersecurity education, professional product

I. INTRODUCTION

In the cybersecurity and digital forensic training, practical exercises are of the utmost importance, especially the practical based on the current state-of-the-art commercial products. When a commercial product is used there are always issues when it comes to budget and licensing. Some vendors do have dedicated academic programs with academic licensing policies, but more vendors do not have such approach, or there is not much difference among prices available. Academic institutions with limited budget must then find acceptable solutions in cooperation with vendors. Just to be clear, the reason for lacking academic licensing programs is based on the complexity of the working solution, which often requires sophisticated networks of machines properly configured, with complex software installation and configuration. Such expensive dedicated setup is often impossible to have in academic institutions, due to limited resources and experience, while hardware sharing often fails in unpredictable ways because of interoperability issues among different (competitive) products.

As an example, preparation for the FireEye [1] demo requires complex infrastructure testing and preparation which can take few weeks to fulfill, such approach is mandatory to avoid proof of concept fail, resulting in a bad for business failure of the test. To avoid such problems but still provide students with practical experience in using a tool featured in real environments, the solution, based on the demo version of the product, was chosen. In TVZ [2] in Network forensics lectures such approach was used in cooperation with FireEye. Due to the limited budget and technical constraints, it was impossible to implement a local version of the appliance, but it was possible to use a cloud version of the appliance.

In such setup, the students' practical is done through a role of the analyst in the FireEye system, while administrative roles are dedicated to lecturers. In practice, that means the lecturer enables and configures the demo system by adding students as analysts to the demo system. The system is complex and distributed, and consists of a virtual security appliance in FireEye's private network, accessible through the internet, while the test machines, which are monitored and used in network forensic tasks, are TVZ's local machines – real, or preferably virtual inside TVZ's private network, mostly in the students laboratories. This setup allows each student (consultant in this setup) to install and use a limited number of monitored endnodes, e.g. to install FireEye agents on the local machines. These agents then connect to the virtual appliance via a secure connection providing full control over agents and monitored endnodes.

This article provides a brief insight in how a FireEye appliance could be used for educational purposes to make possible for our students to achieve learning outcomes: "Critically evaluate computer network monitoring tools",., "Build reports on the conducted digital forensic investigation of computer networks"., in a standard off-the-shelf demo configuration. Whether it would be used in a cybersecurity course at a university or just a standalone subject in college, the goal of this lab exercise would be having students getting hands-on experience on equipment found in more advanced environments. The main topic for one of the lab exercises would be based on using an advanced threat prevention platform through a FireEye product. The appropriate subject would be the one which incorporates elements of network security and network forensics. The appliance used in this lab exercise is FireEye HX, an Endpoint Detection and Response (EDR) solution.

II. LAB INTRODUCTION

FireEye is a California based company which specializes in cybersecurity. Founded in 2004, it all started with a sandboxing solution, which would use virtual machines that would download and test internet traffic before transferring it to a corporate or government network. The initial sandboxing concept translates to multiple appliance series today, and it is called the Multi-vector Virtual Execution engine – MVX. The MVX engine spins up multiple Windows and Mac OS virtual machines and analyzes potentially malicious traffic the virtual machines receive. The traffic may contain threats that come in many different formats. For example, a malicious portable executable (PE) is being downloaded onto a PC. Before reaching the PC, it passes through FireEye NX. NX's MVX engine spins up a virtual machine, detonates the malware inside its safe environment, records its behavior and decides upon a verdict. To reach a verdict, NX may use indicators of compromise (IOC) – which the appliance can obtain in three ways. It can download IOC's from FireEye's Dynamic Threat Intelligence cloud, it can generate its own IOC's by detonating malware, and finally, IOC's can be manually developed and then entered into the appliance by a Security Operations Analyst, or any other appliance operator. HX doesn't use an MVX engine but does use MVX intelligence in form of IOC's to check for any signs of compromise.

In 2013, FireEye acquired a company called Mandiant which specializes in Incident response. Together, they proceed in uncovering zero-day vulnerabilities and revealing multiple Advanced Persistent Threat (APT) group details. Some of their achievements include discovering two Microsoft Windows kernel zero-day vulnerabilities, Microsoft Word, Adobe Flash Player and Android vulnerabilities, and also uncovering details about the FIN6, APT3, FIN4, APT28 groups, etc.

Appliances developed by FireEye cover a broad range of usages – email security (EX), network security (NX), network forensics/packet capture appliances (PX and IA), endpoint detection and response (HX), fileserver protection (FX), malware analysis (AX) and the central management console (CM) for integrating all of those appliances into one collective and distributing locally generated intelligence.

The appliance used in this lab exercise is FireEye HX, the Endpoint Detection and Response appliance. HX uses threat intelligence to correlate alerts generated by FireEye network platforms, log management, and network security products with IOC's on any endpoint. Some of its main features include Exploit Guard, an engine for detecting malicious exploit processes, endpoint containment for stopping lateral movement, Malware Guard, etc. It offers very valuable detection, but also forensic capabilities as well. Using the HX console, the students are able to acquire various data (memory dumps, files, etc.) from the compromised endpoints, run a search query across all the endpoints via Enterprise Search, examine the data via FireEye's freeware forensic tool Redline, and much more.

Upon detecting any malicious activity, an alert is generated, which should be analyzed by a Security Operations Center (SOC) analyst to determine the severity of the incident, how did it occur, which machines were affected, how will the machine be remediated, etc. Depending on the appliance and the alert type, the alert contains useful details about the incident, which include malware names and hashes, malware types, protocols used, IP addresses and ports involved, MAC addresses, PCAP's, communication details, recorded sandbox activity, HTTP requests, etc.

III. PRACTICAL LAB CONFIGURATION

FireEye provided TVZ students with a demo installation as it is presented. The HX appliance was provided for 30 different users for the full demo time period, with a range of roles available to users with a goal that the learning outcome of "evaluation of network monitoring tool" and "examine the computer network for the existence of digital traces" would be achieved. Due to challenges we have in connecting to the virtual appliance, students were able to work with the appliance for less than two weeks. Main issues were the network connections to the external virtual appliance, available bandwidth, firewall rules and egress filtering, slowness of the connection, configuration of monitored endnodes – PC's in various TVZ labs under centralized management and sometimes under Deep Freeze configuration.

Since the goal of the practical is to get familiar with the tool capabilities, to get the idea how this tool is used in the real world practice, students were assigned to the analyst role, which covers practical work of malware hunting, data collecting, installing agents, etc and this was the main reason for using this virtual classroom and the student were able to "examine the computer network for the existence of digital traces". There are other roles limited to specialized tasks, like for installation, maintenance, administration etc. Such roles are to narrow for the required purpose of the students' practical work., but were useful for students to achieve "create advanced network device configurations and security settings" learning outcome.

The administrator role was assigned to the lecturer who prepared the class setup, done testing and debugging until the setup was working, installed the agents on the endpoints, and finally created students' accounts in the virtual appliance and assigned roles to the accounts. When the demo period expired, the virtual appliance was automatically deactivated and stored in the backup, so it can be reused during other future courses in TVZ. Agents installed on local machines were blocked and stopped working, not presenting any further security concerns. This is the way we created learning materials for the future students done by this generation and we learned from the student what is the most difficult task are and how to teach and train the student in the next class. This once more prove that we teacher learn from students as soon as we give them opportunity to work on they own and we monitor and facilitate their work.

Every step of the learning process was documented and analyzed, with the aim to be used in possible future exercise remakes.

IV. LAB EXERCISE

The planned duration for the lab exercise is 90 minutes. First 30 minutes is reserved for the introduction, while the exercise itself lasts for 60 minutes.

Requirements for this lab exercise are the following:

- One computer per student,
- FireEye HX,
- HX agents running on virtual machines.

The exercise was provided through TVZ's Moodle, LMS [4] (Learning Management System), presented in the figure 1, in way how it is visible to the lecturer.

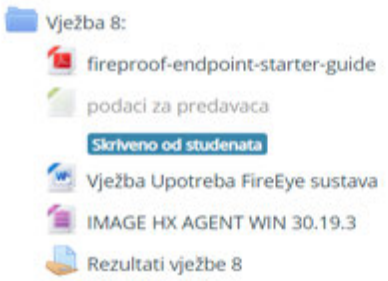


Figure 1. LMS exercise implementation

The exercise consists of 5 main tasks and one bonus assignment. The main tasks are designed to critically judge computer network monitoring tools and find and be able to perform computer network analysis on digital traces what are some of the common operations performed by security analysts. The bonus task is designed to tests students creativity – to perform computer network analysis on digital traces, it is intended for students who can do more.

A. *Student main tasks*

As it mentioned there are students tasks:

Task 1. Log into the FireEye HX appliance. Check the dashboard and see if there are any alerts that seem important.

Task 2. Generate an alert on the HX appliance. How would you do that? Explain and document.

Task 3. After you have generated an alert, analyze it and write a short report about the alert. In the report include some of the details, e.g. the hostname of the victim PC, the number of events, processes detected, malware hashes, etc.

Task 4. Analyze the malware which was used to generate the alert. What information are you able to obtain? Use the triage summary and try to enrich your report by using external sources as well, e.g. VirusTotal or Hybrid Analysis.

Task 5. Download the triage packet and use Redline to analyze it. Note any other useful information you may find, that wasn't included in the previous tasks.

B. *Bonus assignment*

Develop a simple malware that tries to communicate with a C&C center, download it to your lab computer that has an HX agent running, and execute it. The malware may incorporate some other techniques which may be characterized as malicious behavior. The purpose of this assignment is to demonstrate how FireEye reacts to threats that were never seen before – this malware didn't even exist until now.

To ease students' work, a set of sample answer sheets were provided along with FireEye documentation. The intention was to give students an image of the current setup specific for the lab, to avoid long and confusing searches through the product documentation. In Figures 2 and 3 it is possible to see the basic dashboard of the virtual appliance required in task 1. This provides students with click-to-follow visual paths to solve task 1. Same is with the other student tasks, as such approach prevents wasting very limited lab time on wandering through an unfamiliar user interface, even if one has read the documentation.

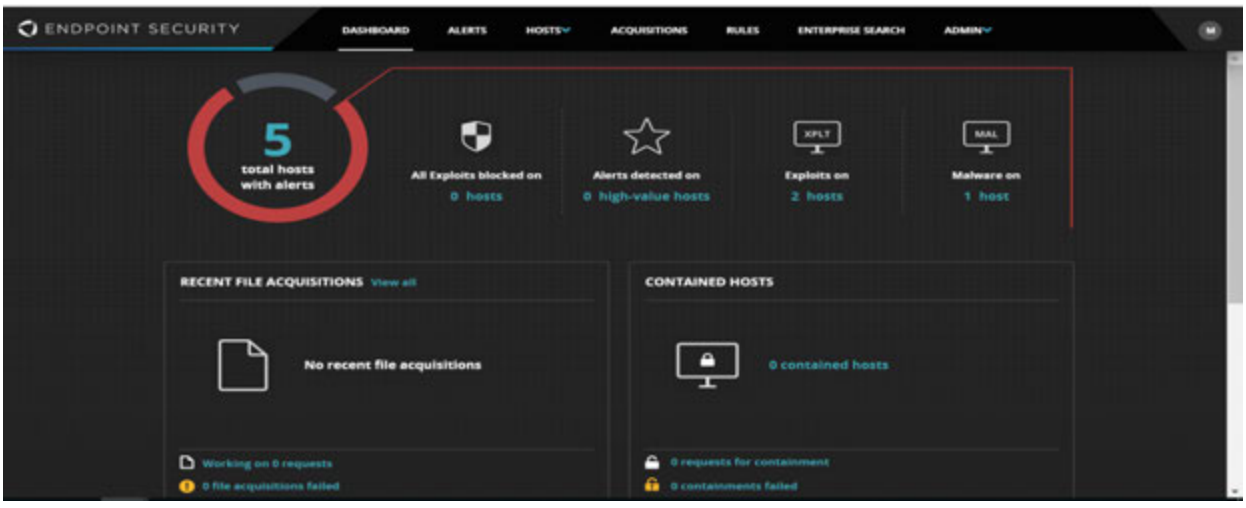


Figure 2. HX dashboard presenting the current status of the monitored system

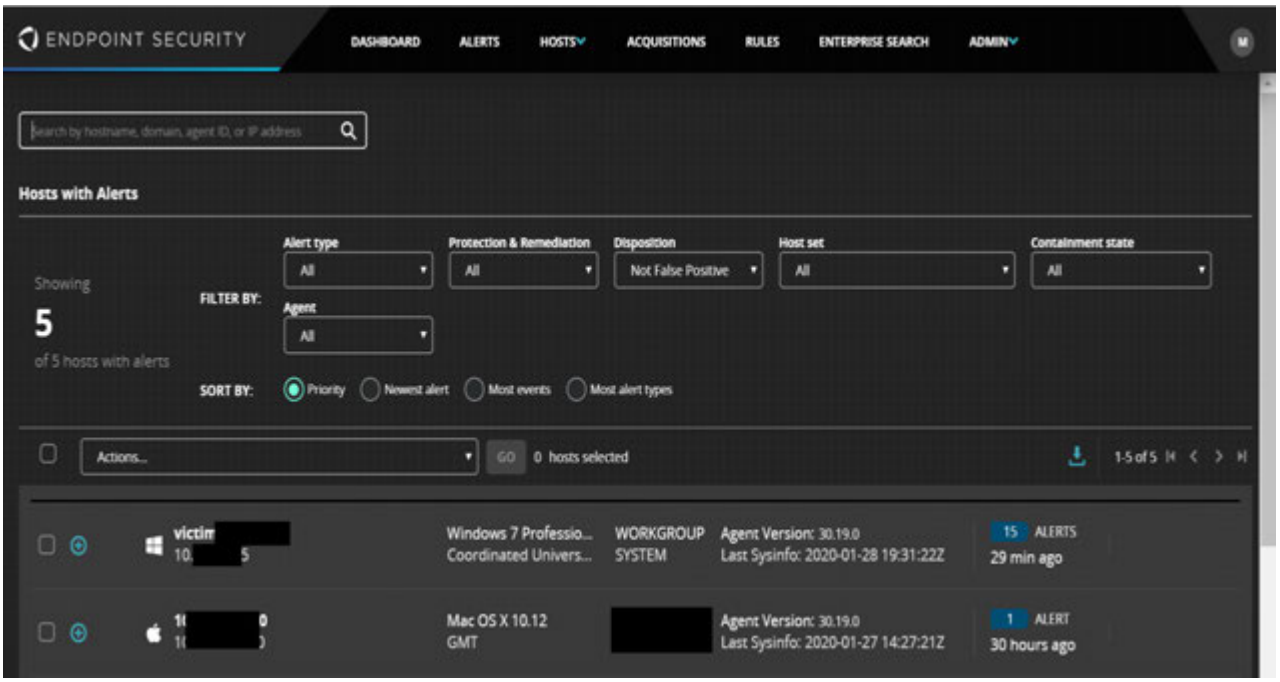


Figure 3. HX alerts list by hosts, preseting on which endpoints alerts have happened

To generate an alert, as it is required task 2, the student must download and run malware inside the virtual machine the HX agent is running on. Then, a new alert will be generated, as it is presented on the figure 3. Further analyses of the alert can be done through callback activities and alert status as it is visible on the figure 7, 6 and 5, which presents alert details.

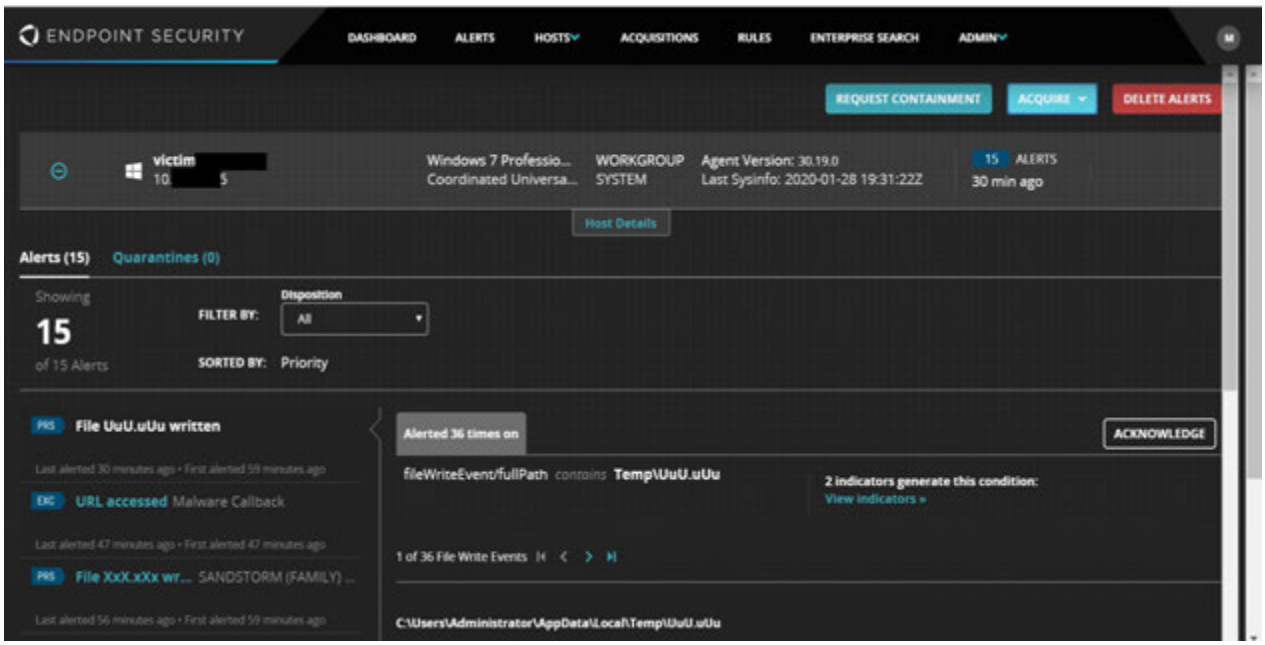


Figure 4. Alert details

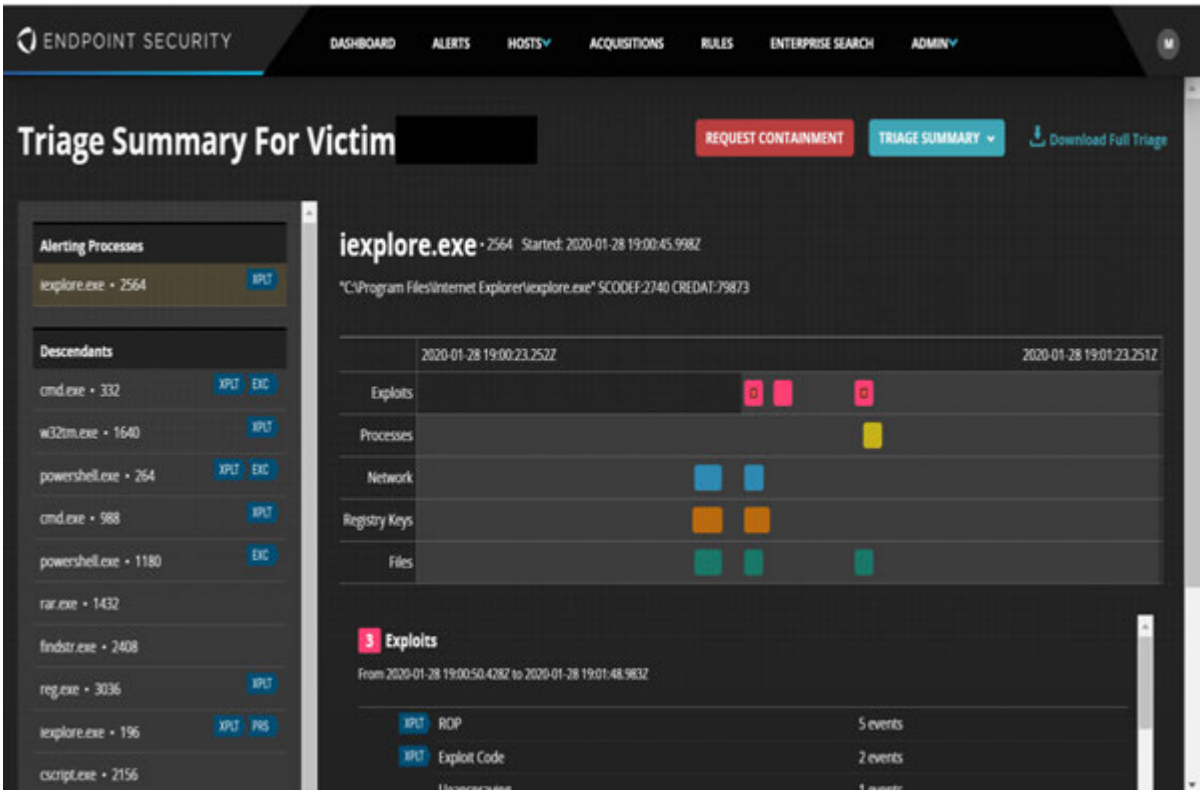


Figure 5. Endpoint Triage Summary

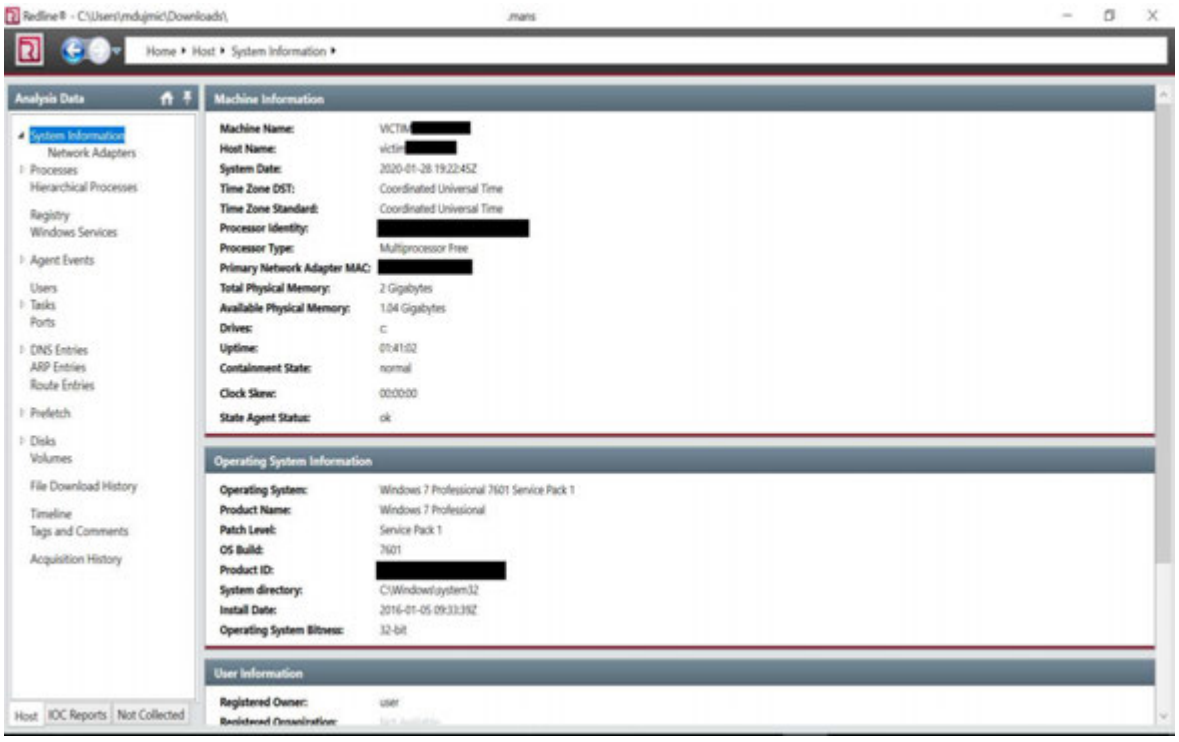


Figure 6. Basic analysis in Redline

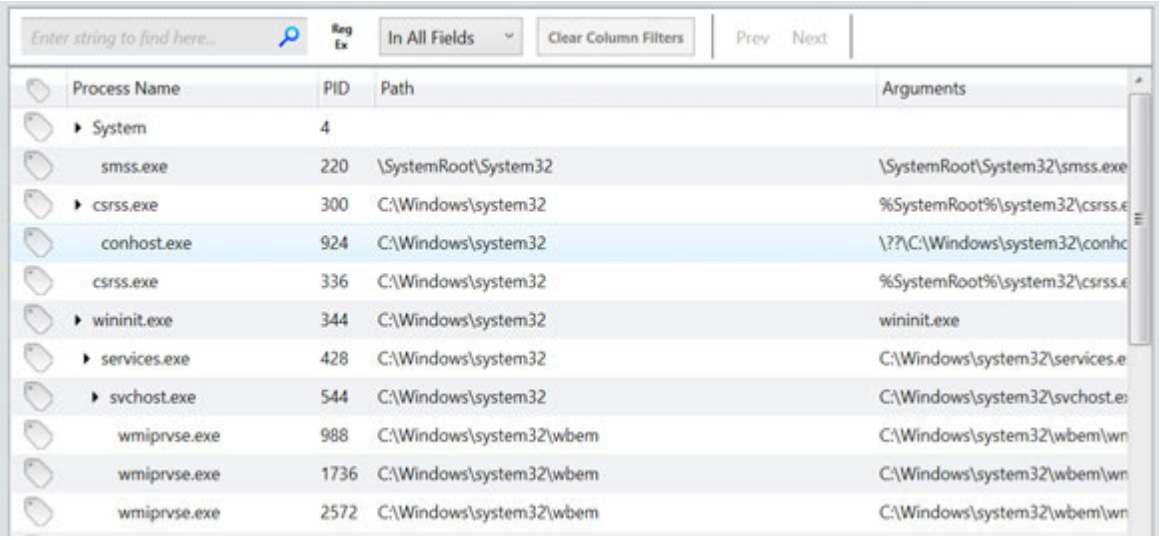


Figure 7. Process analysis in Redline

V. Student Results

The exercise results were delivered trough LMS, as students finished their assignment. Due to time limitation and network slowness, students were allowed to finish the exercise at home in case they did not manage to execute the required tasks during ordinary lab work what is again advantage of the cloud lab we describe here, as our students were able to work with professional tool in their homes as well – especially when we expect that we will work online next academic year.

All students but one in the class were able to finish the tasks successfully. The student was not able to attend lab exercises in time slot we have for this exercise (demo software limitation).

VI. Conclusion

The practical was effective, providing students with a practical glance in one of the top cybersecurity products of the day. Students' results were correct and the assigned time for the laboratory exercise was enough to fulfill tasks even with very slow response times when home access was allowed. It is important to note that the network slowness provides important insight to students why it is important to have a workable infrastructure and a SOC strategically placed in a monitored network. The problem with the student who was not able to attend this lab during the available timeframe is something characteristic for this type of hybrid lab, where timeframe depends not on the university but on third parties. To remedy such situations in future, an alternative set of practicals has to be derived with similar tools but in different timeframes.

In any way, it is necessary to show gratitude to FireEye as a company who was kind enough to provide TVZ students and lecturers a chance to work on a real system and gain valuable experience. Students were satisfied and keen even in limited conditions of this practical. Only a small amount of tool capabilities has been presented, but enough to get realistic feeling about FireEye HX's capabilities. And main goal is achieved - learning outcomes we stated for the course we did and make possible for the student to get in one practical, what give us opportunity to give our student the glimpse of real computer forensic specialist life.

Literature

[1] https://www.thz.hr, visited 11th October 2019.
[2] https://en.wikipedia.org/wiki/FireEye, visited 10th October 2019.
[3] https://www.threatprotectworks.com/datasheets/fireeye-nx-series.pdf, visited 10th October 2019.
[4] https://lms.tvz.hr/course/view.php?id=179, visited 10th October 2019.

Constructing a Set of Weak Values for Full-round MD4 Hash Function

Irina Gribanova, Alexander Semenov
Matrosov Institute for System Dynamics and Control Theory SB RAS, Irkutsk, Russia
E-mail: the42dimension@gmail.com, biclop.rambler@yandex.ru

***Abstract*—In this paper we describe the construction of a set of full-round MD4 compression function values, which are weak against the preimage attack of a special kind. The cardinality estimation for this set, obtained by using an effective probabilistic algorithm, is close to 2^{32}. According to this, the fraction of weak outputs in the set of all outputs of the MD4 compression function is about 2^{-96}. Thus, the probability to obtain an easy-invertible output by choosing a random input is significantly higher than 2^{-128}. These results demonstrate that the full-round MD4 compression function does not have the properties of a random oracle.**

***Keywords*—MD4 hash function; preimage attacks; Boolean satisfiability problem; random oracle**

I. Introduction

The MD4 hash function proposed by R. Rivest in 1994 [1] is one of the first implementations of the Merkle-Damgard construction [2], [3]. In 2011 it was officially stated that MD4 is obsolete, since there were presented efficient collision attacks for this hash function [4], [5]. However, we are not aware of any preimage attacks on hash function MD4 with realistic complexity estimations. Today the most effective preimage attack on full-round MD4 is an attack proposed by G. Leurent in 2008 [6]. This attack is based on the construction of a set of pseudo-preimages for a target hash value. The complexity of finding one pseudo-preimage is equivalent to 2^{96} calls of the MD4 compression function. The total complexity of the attack is 2^{102} calls of the MD4 compression function.

In [7] we used a technique in which SAT solvers were applied to solve cryptanalysis equations for any function of the kind f_{MD4-k}, i.e. a function specified by the first k steps of the MD4 compression function algorithm, $1 \leq k \leq 48$. Based on this technique we constructed superior in efficiency preimage attack on function f_{MD4-39}. In the present paper we develop the results obtained in [7] and show that even for the full-round version of the MD4 compression function the fraction of easy-invertible outputs in $\{0,1\}^{128}$ is relatively high. Based on this fact, we conclude that the compression function of the full-round MD4 algorithm does not satisfy the properties of a random oracle [8].

Below we present the brief outline of our paper. In the next section, we introduce the necessary concepts and preliminary results. In section 3 we describe the theoretical basis of attacks in which the preimage finding problems of functions f_{MD4-k} are reduced to the preimage finding problems of functions of a special kind. Section 4 contains the results of computational experiments. In the final section, we shortly summarize the obtained results and discuss our plans for future work.

II. Preliminaries

A. Preimage attacks on cryptographic hash functions

By $\{0,1\}^n$, $n \in \mathbb{N}_1$ we denote a set of all binary words of length n. The elements from $\{0,1\}^n$ are called Boolean vectors. A discrete function is a function of the kind:

$$f : \{0,1\}^n \to \{0,1\}^m,$$

i.e. a function that transforms binary words of length n to binary words of length m. A *hash function* is a discrete function of the kind $f : \{0,1\}^* \to \{0,1\}^c$, where $\{0,1\}^*$ is a set of all binary words of an arbitrary finite length, c is a fixed natural number called a hash length. An input of a hash function (an initial message) is processed iteratively by dividing it into parts of fixed length called blocks. To construct a hash value for one message block of length l a special function of the kind $f : \{0,1\}^l \to \{0,1\}^c$ is used. The function f is often called a *compression function*. Constants l and c can take different values depending on the hash function algorithm. For example, in hash functions from MD family $l = 512$, $c = 128$, whereas hash functions from SHA family have the following parameters: $l \in \{512, 1024\}, c \in \{160, 256, 384, 512\}$. An algorithm that finds a preimage (a message from $\{0,1\}^l$) for a given hash value from $\{0,1\}^c$ is called a *preimage attack*. The resistance to this kind of attack is one of the basic requirements for cryptographic hash functions.

It is believed that even the first standards of hash algorithms, such as MD4, have the following property: an arbitrary hash value $\gamma \in \{0,1\}^c$ has approximately 2^{l-c} preimages in $\{0,1\}^l$. Thus, the uniform distribution given on $\{0,1\}^l$ will induce a uniform distribution on $\{0,1\}^c$. This means that for a fixed hash value $\gamma \in \{0,1\}^c$, the probability to find such α that $f(\alpha) = \gamma$, by choosing a vector from $\{0,1\}^l$ in accordance with the uniform distribution, is $\approx 1/2^c$. Hence, from the properties of the geometric distribution [9], the expected number of attempts to choose a random $\alpha \in \{0,1\}^l$ such that $f(\alpha) = \gamma$ is 2^c. Thus, the complexity of the described attack can be defined as $2^c \times t$, where t is a calculating time of a value of function f on an arbitrary input. This preimage attack is essentially a brute force attack. If for a particular f there is a preimage attack which is significantly more effective than brute force attack, then f is compromised with respect to the preimage attack.

An important property of a cryptographic hash function is that such function with resistance to preimage attacks can be considered as a real example of a *random oracle*. According to [8], a random oracle is a function of the kind $O : \{0,1\}^* \to \{0,1\}^\infty$ which associates an arbitrary finite binary word with a word that is an infinite sequence of Bernoulli trials with $p = 1/2$. However, this definition is completely unconstructive. For practical applications, we need functions of the kind $\{0,1\}^* \to \{0,1\}^*$ or even $\{0,1\}^n \to \{0,1\}^m$, which have main random oracle properties in its finite variant and which can be specified by some algorithms. First of all, it is required that the algorithm defining the function is deterministic (i.e. an algorithm produces the same outputs for the same inputs). Secondly, almost all outputs (of length m) of a function that acts as a random oracle must look like Bernoulli sequences with $p = 1/2$. Let $\gamma \in \{0,1\}^m$ be an output of such a function computed for a random input. If the function has the properties of a random oracle then the complexity of finding a preimage of γ should be comparable to the complexity of finding γ in the result of m trials of tossing of an ideal coin, which is, in fact, the complexity of the brute-force attack described above. The aforesaid means that if for some hash function the amount of easy-invertible outputs in the set of all outputs is relatively high, then such a function does not satisfy the properties of a random oracle.

In the present paper we construct new preimage attacks on well-known hash function MD4. As it was mentioned above, this function was one of the first standards of cryptographic hash algorithms and has been used in practice for a long time. The lack of collision resistance for this function was convincingly demonstrated in [4]. However, the best-known preimage attack on full-round MD4 with a complexity of 2^{102} MD4 calls described in [6] cannot be implemented in practice.

Above, we introduced functions of the kind f_{MD4-k}, $1 \leq k \leq 48$, for which f_{MD4-48} is the full-round version of the MD4 compression function $f_{MD4} : \{0,1\}^{512} \to \{0,1\}^{128}$. Here we consider the compression function f_{MD4} that uses a standard value of Initial Vector (IV) and omits the finalization steps [1]. As a main result of the present paper, we show that the set $\{0,1\}^{128}$ contains $\approx 2^{32}$ vectors (values of the function f_{MD4}), for which the preimage finding problem can be solved in a few minutes on a standard PC. Thus, by choosing an arbitrary $\alpha \in \{0,1\}^{512}$ the probability to obtain an easy-invertible $\gamma = f_{MD4}(\alpha)$ is $\approx \frac{1}{2^{96}}$. The latter value is significantly greater than $\frac{1}{2^{128}}$ and thus the fraction of easy-invertible outputs of a full-round MD4 is too large to claim that this function has the properties of a random oracle. Along the way, we show that similar facts, but in a stronger form, hold for functions of the kind f_{MD4-k}, $39 \leq k \leq 47$.

B. Using SAT solvers for constructing preimage attacks on MD4-k

All currently known preimage attacks on MD4 are based on the ideas proposed by H. Dobbertin in [10]. The main conclusion made in [10] is that the cryptanalysis equations of functions of the kind f_{MD4-k} can be significantly weakened if values of some chaining variables are fixed. More precisely, H. Dobbertin specified a set of variables, which need to be assigned with an arbitrary (in the general case) constant K. This approach made it possible to find preimages for f_{MD4-32} on a personal computer. We will refer to these constraints as *Dobbertin's conditions*. Later in [11] it was proposed to construct SAT encodings for cryptanalysis equations of f_{MD4-k}, $k \in \{36, 37, 39\}$ taking into account Dobbertin's conditions and to apply SAT solver to the obtained formulas. As a result, the authors of [11] managed to find preimages for some values of f_{MD4-39} using SAT solver MINISAT [12]. In [7] we presented the record-breaking preimage attack on f_{MD4-39}, which also uses SAT solvers to solve cryptanalysis equations. The main novelty of [7] is a technique for constructing new sets of Dobbertin's conditions called *relaxation constraints*. This technique applies black-box optimization algorithms to traverse the space of possible sets of constraints and finds the best set according to several criteria. As a result, we found a set of relaxation constraints, which allows finding f_{MD4-39}-preimages for about half of any vectors randomly selected from $\{0,1\}^{128}$. The average time of finding one preimage was less than 1 minute on a PC using SAT solver MINISAT2.2.

III. PREIMAGE ATTACKS ON MD4-K VIA INVERSION PROBLEMS OF SPECIAL FUNCTIONS

In this section we present the theoretical basis of attacks described below. Recall that the preimage finding problem of an arbitrary quickly computable discrete function can be effectively reduced to the Boolean Satisfiability Problem, which is abbreviated as SAT and consists in the following. Consider a Boolean formula C (usually in the Conjunctive Normal Form, CNF) and let $X = \{x_1, \ldots, x_M\}$ be a set of variables that occur in C. The formulas of the kind x and $\neg x$, $x \in X$ are called *literals*. An arbitrary mapping $\delta : X \to \{0,1\}$ defines an assignment of variables from X, denoted as $(\delta_1, \ldots, \delta_M)$. The formula C defines a Boolean function $F_C : \{0,1\}^M \to \{0,1\}$, which associates each assignment of variables from X with a value which F_C takes on this assignment. If there is such an assignment $\delta \in \{0,1\}^M$ that $F_C(\delta) = 1$, then C is called *satisfiable*, otherwise, C is called *unsatisfiable*. Any assignment $\delta : F_C(\delta) = 1$ is called a *satisfying assignment* of C. Boolean Satisfiability Problem (SAT) for C consists in the following: to determine whether the formula is satisfiable and if so, to find a corresponding satisfying assignment of C. SAT is a well-known NP-hard problem. However, in practice, SAT can often be effectively solved for large dimensions using algorithms that have been actively developed over the past 20 years [13].

Let us consider an arbitrary function

$$f : \{0,1\}^n \to \{0,1\}^m, \tag{1}$$

specified by an efficient algorithm A_f. The *preimage finding problem* or *inversion problem* of f can be formulated as follows: for a given $\gamma \in Range\ f \subseteq \{0,1\}^m$ one need to find such $\alpha \in \{0,1\}^n$ that $f(\alpha) = \gamma$. It is commonly known that the problem of finding preimage for (1) can be

effectively (in polynomial time of A_f complexity) reduced to the problem of finding a satisfying assignment of a Boolean formula in Conjunctive Normal Form (CNF). More precisely, for the algorithm A_f we can effectively construct a CNF of a special kind, denoted by C_f. The CNF C_f is constructed from a Boolean circuit S_f representing f by using Tseitin transformation [14]. The CNF C_f is called a *template CNF* for function f [15].

Let X be a set of all variables occurred in C_f in which we select two subsets: $X^{in} = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$. These sets of variables correspond to the input and output of S_f, respectively. We will formulate in the form of a lemma an important property that will be repeatedly used below. The cited lemma uses the Boolean constraint propagation procedure that is well-known in propositional logic – Unit Propagation (UP) rule [16].

Lemma 1. *In the context of inversion of an arbitrary function* (1) *let* $\alpha = (\alpha_1, \ldots, \alpha_n)$ *be an arbitrary assignment of variables from* X^{in} *(*$\alpha \in \{0,1\}^n$*). Then the application of UP rule to CNF*

$$x_1^{\alpha_1} \wedge \ldots \wedge x_n^{\alpha_n} \wedge C_f \tag{2}$$

gives an assignment of all variables from X*. The runtime of this procedure is limited by a polynomial on the length of a binary representation of the CNF* C_f*.*

In (2) x^α is a literal x if $\alpha = 1$ and literal $\neg x$ if $\alpha = 0$. Let us give additional comments on the property formulated in Lemma 1. First, note that the application of UP to CNF (2) gives us values of variables from Y: $y_1 = \gamma_1, \ldots, y_m = \gamma_m$, such that $f(\alpha) = \gamma$, $\gamma = (\gamma_1, \ldots, \gamma_m)$. Second, due to the formulated property, the function f can be defined in the following way. Using the template CNF C_f and an arbitrary $\alpha \in \{0,1\}^n$ one can construct the formula (2), apply UP rule to it and extract from the resulted assignment of all variables from X values of variables from Y. The result would be a Boolean vector of length m which is a value of function f on input α. Facts similar to those from Lemma 1 are well-known and have been repeatedly proved earlier (see for ex. [17], [18]).

Now let us consider a function

$$f_{MD4-k} : \{0,1\}^{512} \to \{0,1\}^{128} \tag{3}$$

that corresponds to the first k steps of the MD4 compression function algorithm. Let $C_{f_{MD4-k}}$ be a template CNF for function (3). Hereinafter X denotes a set of all variables in $C_{f_{MD4-k}}$ and X^{in} – a set of input variables of (3).

To construct CNFs of the kind $C_{f_{MD4-k}}$ we use a software system TRANSALG [19], designed for automatic translation of algorithms for discrete functions into SAT. TRANSALG stores information about the connections between variables in CNF and memory cells of an abstract machine, which executes a program that specifies the function (3). This makes it possible to impose various additional constraints on data used in the process of calculating (3). For example, we can effectively add in SAT encoding various constraints on the values of chaining variables corresponding to certain steps of the MD4 algorithm. Thus, we can construct SAT encodings for the inversion problems of functions of the kind (3) that take into account various additional constraints like those suggested by H. Dobbertin.

Recall that in [10] H. Dobbertin suggested to fix some chaining variables with an arbitrary constant K, thereby deriving additional information and simplifying the cryptanalysis problem. In [11] and [7] $K = 0^{32}$ (32-bit zero constant) was used for this purpose. Different variants of chaining variables sets, whose values are fixed with a constant, can be found automatically by solving the maximization problem of a special objective function. This approach was used in [7] to find new sets of relaxation constraints which made it possible to significantly outperform the preimage attacks from [10] and [11].

In [7] an arbitrary set of relaxation constraints of the mentioned kind was presented by a Boolean vector $\lambda \in \{0,1\}^{48}$. If a vector's coordinate with number j is equal to 1 then the corresponding chaining variable at step j is fixed with a constant $K = 0^{32}$. This fact corresponds to a formula

$$\phi_j = (\neg x_{j,1}) \wedge \ldots \wedge (\neg x_{j,32}), \tag{4}$$

in which $x_{j,i}, i \in \{1, \ldots, 32\}$ denotes Boolean variables from X connected to a cell of a hash register in which a chaining variable v_j is stored, i.e. the variable at step with number $j \in \{1, \ldots, 48\}$. It is obvious that (4) has a unique satisfying assignment, namely, 0^{32}.

Now let us consider an arbitrary function of the kind (3). Let us fix some $\lambda \in \{0,1\}^{48}$ and denote by I_λ a set of coordinates of λ equal to 1. Consider the following formula:

$$C_{f_{MD4-k}}(\lambda) = C_{f_{MD4-k}} \wedge \bigwedge_{j \in I_\lambda} \phi_j \tag{5}$$

and apply UP rule to it. In this case, we are interested in such vectors λ for which the application of UP rule to $C_{f_{MD4-k}}(\lambda)$ derives values for a number of variables and does not lead to a conflict. Let us substitute all such values in CNF $C_{f_{MD4-k}}(\lambda)$ and denote the resulting CNF by $C'_{f_{MD4-k}}(\lambda)$. Hereinafter X_λ is a subset of X^{in} formed by variables values which were derived as a result of the transition from $C_{f_{MD4-k}}(\lambda)$ to $C'_{f_{MD4-k}}(\lambda)$. Also we consider a set $X^*_\lambda = X^{in} \setminus X_\lambda$ – a set of variables whose values are unknown.

For some vectors λ it is possible to highlight in set X^*_λ a subset $\hat{X}_\lambda = \{\hat{x}_1, \ldots, \hat{x}_r\}$, which has the following important property. Let $\hat{\alpha} = (\alpha_1, \ldots, \alpha_r)$ be a random vector from $\{0,1\}^r$, which is an assignment of variables from $\hat{X}_\lambda$. Consider a CNF

$$\hat{x}_1^{\alpha 1} \wedge \ldots \wedge \hat{x}_r^{\alpha_r} \wedge C'_{f_{MD4-k}}(\lambda). \tag{6}$$

For a significant part of possible $\hat{\alpha}$ the application of UP rule to (6) gives values of all other variables in CNF $C'_{f_{MD4-k}}(\lambda)$ and does not lead to a conflict.

Based on CNF (6), let us define a special function, which will be denoted below as g^λ_{MD4-k}. First, we fix an arbitrary $\hat{\alpha} = (\alpha_1, \ldots, \alpha_r)$ and apply UP rule to CNF (6). If a conflict

arises, then g^{λ}_{MD4-k} is not defined, otherwise, the values of all variables found in $C'_{f_{MD4-k}}(\lambda)$ would be derived. Among these values we choose values of chaining variables associated with steps $k-3, k-2, k-1, k$. The resulting 128-bit vector is a value of $g^{\lambda}_{MD4-k}(\hat{\alpha})$. Thus, we have the following function (partial, in general)

$$g^{\lambda}_{MD4-k} : \{0,1\}^{|\hat{X}_\lambda|} \to \{0,1\}^{128}. \quad (7)$$

In (7) and everywhere further by $\{0,1\}^{|B|}$ we denote a set of assignments of all Boolean variables from B. Below we formulate the main theoretical result of the present paper.

Theorem 1. *In the context of the inversion problem of function* (3) *consider g^{λ}_{MD4-k} which is a function of the kind* (7)*. Let $\gamma \in \{0,1\}^{128}$ be a value of g^{λ}_{MD4-k}, i.e. there is such a vector $\hat{\alpha} \in \{0,1\}^{|\hat{X}_\lambda|}$ that $g^{\lambda}_{MD4-k}(\hat{\alpha}) = \gamma$. In this case, we can effectively move from $\hat{\alpha}$ to such $\alpha \in \{0,1\}^{512}$ that $f_{MD4-k}(\alpha) = \gamma$.*

Proof. Suppose the conditions of the theorem are satisfied. Consider a template CNF $C_{f_{MD4-k}}$ and a set X^{in} of variables corresponding to the input of function f_{MD4-k}. Let us apply UP rule to CNF (5) and assume that a conflict does not arise. However, the values of some variables from X^{in} can be derived during this process. These variables form a set X_λ. Let $\alpha[X_\lambda]$ be a derived assignment of variables from X_λ. Consider the set $\hat{X}_\lambda$, which was introduced above. Thus, $\hat{X}_\lambda$ is a set of variables encoding the input of function g^{λ}_{MD4-k} and $\hat{X}_\lambda \subseteq X^*_\lambda$, where $X^*_\lambda = X^{in} \setminus X_\lambda$. Now let $\hat{\alpha}$ be a preimage of some $\gamma \in \{0,1\}^{128}$ for mapping g^{λ}_{MD4-k}, i.e. $\hat{\alpha}$ is an assignment of variables from $\hat{X}_\lambda$. In view of the above, the application of UP to CNF (6) derives values of all variables from $C'_{f_{MD4-k}}(\lambda)$, including variables from the set $\tilde{X}_\lambda = X^*_\lambda \setminus \hat{X}_\lambda$. Denote the resulting assignments of variables from $\tilde{X}_\lambda$ by $\tilde{\alpha}$. Note that $X^{in} = X_\lambda \cup \hat{X}_\lambda \cup \tilde{X}_\lambda$. Let us construct a vector $\alpha \in \{0,1\}^{512}$ from vectors $\alpha[X_\lambda], \hat{\alpha}, \tilde{\alpha}$ assuming that $\alpha[X_\lambda]$ is an assignment of variables from X_λ; $\hat{\alpha}$ is an assignment for variables from $\hat{X}_\lambda$; $\tilde{\alpha}$ is an assignment of variables from $\tilde{X}_\lambda$. Thus, $\alpha = (\alpha_1, \ldots, \alpha_{512})$ is an assignment of variables from $X^{in} = \{x_1, \ldots, x_{512}\}$. Consider a CNF

$$x_1^{\alpha_1} \wedge \ldots \wedge x_{512}^{\alpha_{512}} \wedge C_{f_{MD4-k}}. \quad (8)$$

According to Lemma 1, the application of UP rule to (8) results in the assignment of all variables that occurred in $C_{f_{MD4-k}}$ without conflicts. Indeed, a conflict would arise either when applying UP to CNF (5) or (6). Let γ' be an assignment of variables from Y as a result of UP rule applied to CNF (8). Then $\gamma' = \gamma$ because otherwise there would be a conflict. Thus, for α constructed by described way it follows that $f_{MD4-k}(\alpha) = \gamma$. The theorem is proven. □

Theorem 1 shows that the inversion problem for an arbitrary function of the kind f_{MD4-k} can be reduced to a similar problem of a special function of the kind (7). Two important comments should be made here. First, for some λ the function g^{λ}_{MD4-k} may not be defined everywhere on $\{0,1\}^{|\hat{X}_\lambda|}$. Second, not every vector $\gamma \in \{0,1\}^{128}$ has a preimage with respect to the mapping g^{λ}_{MD4-k}. However, as we will show below, the inversion problem for g^{λ}_{MD4-k} may turn out to be significantly simpler than the similar problem for f_{MD4-k}.

Let us give an example to illustrate all the above-mentioned concepts and properties. In [7] we presented a vector λ_1 which defines a new set of relaxation constraints (in [7] this vector was denoted by ρ_1). In λ_1 coordinates equal to 1 are in positions with numbers

$$14, 15, 17, 18, 19, 21, 22, 23, 25, 26, 27, 29$$

and the rest coordinates are zeros. Hence, it defines 12 relaxation constraints of the kind (4). For each $k : 39 \leq k \leq 48$ the application of UP rule to CNF $C_{f_{MD4-k}}(\lambda_1) = C_{f_{MD4-k}} \wedge \bigwedge_{j \in I_{\lambda_1}} \phi_j$ derives values of 288 variables from X^{in}, i.e. $|X_{\lambda_1}| = 288$ and $|X^*_{\lambda_1}| = 512 - 288 = 224$, respectively. In set $X^*_{\lambda_1}$ we managed to find a subset $\hat{X}_{\lambda_1} : |\hat{X}_{\lambda_1}| = 128$. Thus, in case of vector λ_1 a special function is the following

$$g^{\lambda_1}_{MD4-k} : \{0,1\}^{128} \to \{0,1\}^{128}. \quad (9)$$

It is turned out that this function is defined on the vast majority of vectors from $\{0,1\}^{128}$. We generated $\approx 10^6$ random $\alpha \in \{0,1\}^{128}$ and for each of these inputs a corresponding value of function $g^{\lambda_1}_{MD4-k}$ was found.

In fact in [7] an inversion problem of function (9) was solved for $k = 39$. More particular, $\approx 65\%$ of random $\gamma \in \{0,1\}^{128}$ have preimages with respect to $g^{\lambda_1}_{MD4-39}$. The average time of finding one such preimage is less than one minute on PC using SAT solver MINISAT2.2. For the rest $\approx 35\%$ of $\gamma \in \{0,1\}^{128}$ SAT solvers quickly proves the unsatisfiability of the CNF, constructed by substitution of vector γ in CNF $C'_{f_{MD4-k}}(\lambda_1)$ as an assignment of variables from Y.

IV. The new practical preimage attacks on functions of kind f_{MD4-k}

In this section we describe the results of computational experiments in which the inversion problems of f_{MD4-k} were reduced to the inversion problem of special functions of the kind (7). To construct different variants of function (7) various vectors λ can be used. In our experiments we studied several such vectors. The search process for each of them consisted of two stages. At the first stage, we considered a black-box optimization problem that is to maximize the number of variables from X^{in} which values were derived by UP from CNF (5). To solve this problem we applied a metaheuristic algorithm described in [7]. During the second stage, we analyzed the search history to select the so-called *promising vectors*. It should be noted that some vectors which maximize the objective function from [7] can be unsuitable for the inversion of functions of the kind f_{MD4-k}. Such a situation is, for example, when the corresponding function (7) isn't defined on the majority of its input. In this case, each such λ is considered as unpromising.

Let λ be a fixed promising vector. In accordance with the above, to define a function g^{λ}_{MD4-k} it is necessary first

to select in X^*_λ a set $\hat{X}_\lambda$ of input variables of this function. The way this step is carried out is important for the practical implementation of the corresponding attacks. We propose the following algorithm for this purpose.

Algorithm 1 Algorithm for searching the set $\hat{X}_\lambda$

```
C ← C'_{f_{MD4−k}}(λ)
for each x_i ∈ X*_λ do
    if x_i is undefined then
        λ_i ← getRandomValue();
        C ← C ∧ x_i^{λ_i}
        run UP rule for C
        if no conflict then
            add x_i in X̂_λ
        end if
    end if
end for
return X̂_λ
```

Algorithm 1 uses previously mentioned UP rule to determine "independent" variables from set X^*_λ, i.e. variables which can be assign with any values without producing conflicts. If conflict occurs, we assume that the corresponding set does not satisfy our requirements.

To accumulate vectors of the kind λ, we conducted an experiment similar to that described in [7]. The algorithm for solving the black-box optimization problem was run on the computing cluster "Academician V.M. Matrosov" of the Irkutsk supercomputer center [20]. Based on the results of this algorithm, several dozen vectors of the kind λ were found and analyzed. From these vectors, we selected 6 ones, presented in Table I. These vectors produce the most interesting attacks on functions of the kind f_{MD4-k}.

TABLE I
NEW SETS OF RELAXATION CONSTRAINTS FOR FUNCTIONS f_{MD4-k}

	Boolean vectors defining new sets of relaxation constrains	d_i
λ_1	00000000000001101110111011101000000000000000000000	128
λ_2	00000000000001101110111011101100000000000000000000	96
λ_3	00000000000011101110111011101000000000000000000000	96
λ_4	00000000001011111101110111010000000000000000000000	64
λ_5	00000000000011101110111011101100000000000000000000	64
λ_6	00000000000011101110111011101100000000000000000000	32

Let us comment on Table I. The second column contains Boolean vectors $\lambda_i \in \{0,1\}^{48}$, $i = 1,\ldots,6$ where each vector defines a set of relaxation constraints of the kind (4). If a component with a number j of the vector λ_i is equal to 1, then the corresponding chaining variable at step j is fixed with a constant $K = 0^{32}$ in the system of equations that specifies the function f_{MD4-k}. As a result, we have a set of relaxation constraints which allows us to reduce the cryptanalysis problem of function f_{MD4-k} to the inversion problem of function $g^{\lambda_i}_{MD4-k} : \{0,1\}^{d_i} \to \{0,1\}^{128}$. For example, a set of relaxation constraints corresponding to vector λ_3, which fixes values of chaining variables at steps with numbers

$$13, 14, 15, 17, 18, 19, 21, 22, 23, 25, 26, 27, 29,$$

produces a special function $g^{\lambda_3}_{MD4-k} : \{0,1\}^{96} \to \{0,1\}^{128}$.

We studied the inversion problems of functions $g^{\lambda_i}_{MD4-k} : \{0,1\}^{d_i} \to \{0,1\}^{128}$ for various $k \in \{39,\ldots,48\}$. In each case we considered a vector $\gamma \in Range\ g^{\lambda_i}_{MD4-k}$, after that the corresponding SAT instance (CNF) for the inversion problem for γ was constructed and solved using a SAT solver. For some k these problems were solved in a reasonable time (less than an hour) with a single-threaded SAT solver such as MINISAT2.2 or ROKK [21]. In our experiments we involved several nodes of the computing cluster "Academician V.M. Matrosov", where each node is equipped with two 18-core Intel Xeon E5-2695 CPUs. Thus, 36 copies of the mentioned SAT solvers can be simultaneously run on one node, wherein each solver is working on individual inversion problem. In Table II we present more detailed information about these experiments.

TABLE II
SOLVING THE INVERSION PROBLEM OF $g^{\lambda_i}_{MD4-k}$ USING ROKK

	d_i	Number of steps (k)	Average solving time in seconds (t)
λ_1	128	39	12
λ_2	96	43	4,5
λ_3	96	44	20
λ_4	64	41	5,7
λ_5	64	47	914
λ_6	32	48	509

The third column of Table II contains a maximum value of k (a number of steps used in f_{MD4}) for which the inversion problem of function $g^{\lambda_i}_{MD4-k}$ can be solved in time $\leq t$ using SAT solver ROKK. For $k = 44$ function $g^{\lambda_3}_{MD4-44}$, which is defined on $\approx 50\%$ of random inputs (from $\{0,1\}^{96}$), the corresponding inversion problem for $\gamma \in Range\ g^{\lambda_3}_{MD4-44}$ can be solved in 20 seconds (on average). For inputs on which $g^{\lambda_3}_{MD4-k}$ is defined, the absence of collisions was proven in less than an hour.

In fact, the described approach gives us a probabilistic algorithm for estimating a cardinality of a set of g^{λ}_{MD4-k} values. Since each specific output of considered functions is processed by a SAT solver in a reasonable time, the efficiency of the described algorithm is entirely determined by the cardinality of a random sample. In all computational experiments this algorithm was run on the computing cluster "Academician V.M. Matrosov".

Suppose we have a cardinality estimation for a set $Range\ g^{\lambda}_{MD4-k}$ (the range of g^{λ}_{MD4-k} values) and this estimation is relatively large. Then choosing a random input of function f_{MD4-k} we obtain its value which can be an output of g^{λ}_{MD4-k} with a relatively high probability. Moreover, according to the results of computational experiments, such outputs of these functions can be effectively inverted using single-threaded SAT solvers.

With respect to function $g^{\lambda_3}_{MD4-44}$, the above means that $\{0,1\}^{128}$ contains $\approx \frac{1}{2} \times 2^{96}$ vectors which are values of function $g^{\lambda_3}_{MD4-44}$. Thus, if we choose a random input $\alpha \in \{0,1\}^{512}$ and compute $\gamma = f_{MD4-44}(\alpha)$ then with a probability $\approx \frac{2^{95}}{2^{128}} = \frac{1}{2^{33}}$ the vector γ would be a value of function $g^{\lambda_3}_{MD4-44}$ which is expected to be easy-invertible. Thereby we show that the fraction of easy-invertible outputs of function $g^{\lambda_3}_{MD4-44}$ is $\approx \frac{1}{2^{33}}$ which is incomparably greater than $\frac{1}{2^{128}}$. Thus, we can conclude that function f_{MD4-44} does not satisfy the properties of a random oracle.

We would like to especially comment on vector λ_6, which is shown in the last row of Table I. This vector corresponds to function $g^{\lambda_6}_{MD4-k} : \{0,1\}^{32} \to \{0,1\}^{128}$, which is defined on any input from a random sample of 10^4 vectors. For $k = 48$ an output of function $g^{\lambda_6}_{MD4-48}$ (computed for an arbitrary input) can be inverted approximately in 10 minutes using SAT solver ROKK. The same SAT solver was applied to prove the absence of collisions for considered outputs. Thus, $\approx \frac{1}{2^{96}}$ of values of f_{MD4-48} are expected to be easy-invertible. Despite this value is relatively small, it allows us to conclude that even the full-round algorithm of the MD4 compression function cannot be considered as a random oracle.

V. Conclusion and Future Work

In this paper we described a new class of preimage attacks on well-known cryptographic function MD4. These attacks are based on the idea of reduction of the inversion problem of functions, which are variants of the MD4 compression function with a limited number of steps, to the inversion problems of specially constructed functions. The introduced functions have an input length which is significantly less than 512 and, in general, are substantially weaker against the preimage attacks compared with the original functions. Using this approach we show that fractions of easy-invertible outputs for all function of the kind $f_{MD4-k}, 39 \leq k \leq 48$ in $\{0,1\}^{128}$ are relatively large. For example, the fraction of such outputs for function f_{MD4-48} is $\approx \frac{1}{2^{96}}$, which is significantly greater than $\frac{1}{2^{128}}$. This fact means that even the full-round MD4 compression function does not have properties of a random oracle. In the nearest future, we are planning to construct guess-and-determine attacks, based on the concept of Inverse Backdoor Sets from [22], for special functions introduced in the present paper.

Acknowledgment

The research was funded by Russian Science Foundation (project No. 16-11-10046). Irina Gribanova is additionally supported by the Council for Grants of the President of the Russian Federation (stipend SP-3545.2019.5).

References

[1] R. L. Rivest, "The MD4 message digest algorithm," in *Advances in Cryptology - CRYPTO '90, Proceedings*, ser. Lecture Notes in Computer Science, A. Menezes and S. A. Vanstone, Eds., vol. 537. Springer, 1990, pp. 303–311.

[2] R. C. Merkle, "A certified digital signature," in *Advances in Cryptology - CRYPTO '89, Proceedings*, 1989, pp. 218–238.

[3] I. Damgård, "A design principle for hash functions," in *Advances in Cryptology - CRYPTO '89, Proceedings*, 1989, pp. 416–427.

[4] X. Wang, X. Lai, D. Feng, H. Chen, and X. Yu, "Cryptanalysis of the hash functions MD4 and RIPEMD," in *Advances in Cryptology - EUROCRYPT 2005, Proceedings*, ser. Lecture Notes in Computer Science, R. Cramer, Ed., vol. 3494. Springer, 2005, pp. 1–18.

[5] Y. Sasaki, L. Wang, K. Ohta, and N. Kunihiro, "New message difference for MD4," in *Proc. of the International Workshop on Fast Software Encryption (FSE 2007)*, vol. 4593. Springer Berlin Heidelberg, 2007, pp. 329–348.

[6] G. Leurent, "MD4 is not one-way," in *Fast Software Encryption*, ser. Lecture Notes in Computer Science, K. Nyberg, Ed., vol. 5086. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 412–428.

[7] I. Gribanova and A. Semenov, "Using automatic generation of relaxation constraints to improve the preimage attack on 39-step MD4," in *Proc. of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO'2018)*. IEEE, 2018, pp. 1174–1179.

[8] M. Bellare and P. Rogaway, "Random oracles are practical: A paradigm for designing efficient protocols," in *Proceedings of the 1st ACM Conference on Computer and Communications Security (CCS'93)*, 1993, pp. 62–73.

[9] W. Feller, *Introduction to probability theory and its applications*. Wiley, 1968, vol. 1.

[10] H. Dobbertin, "The first two rounds of MD4 are not one-way," in *Fast Software Encryption*, ser. Lecture Notes in Computer Science, S. Vaudenay, Ed. Springer Berlin Heidelberg, 1998, vol. 1372, pp. 284–292.

[11] D. De, A. Kumarasubramanian, and R. Venkatesan, "Inversion attacks on secure hash functions using SAT solvers," in *Theory and Applications of Satisfiability Testing - SAT 2007, Proceedings*, ser. Lecture Notes in Computer Science, J. Marques-Silva and K. A. Sakallah, Eds., vol. 4501. Springer, 2007, pp. 377–382.

[12] N. Eén and N. Sörensson, "An extensible SAT-solver," in *SAT*, 2003, pp. 502–518.

[13] A. Biere, M. Heule, H. van Maaren, and T. Walsh, Eds., *Handbook of Satisfiability*, vol. 185. IOS Press, 2009.

[14] G. Tseitin, "On the complexity of derivation in propositional calculus," in *Studies in mathematics and mathematical logic, Part II, Seminars in mathematics*, 1970, pp. 115–125.

[15] A. Semenov, I. Otpuschennikov, I. Gribanova, O. Zaikin, and S. Kochemazov, "Translation of Algorithmic Descriptions of Discrete Functions to SAT with Applications to Cryptanalysis Problems," *Logical Methods in Computer Science*, vol. 16, pp. 29:1–29:42, 2020.

[16] J. P. Marques-Silva, I. Lynce, and S. Malik, "Conflict-driven clause learning sat solvers," in *Handbook of Satisfiability*, A. Biere, M. Heule, H. van Maaren, and T. Walsh, Eds. IOS Press, 2009, vol. 185, ch. 4, pp. 131–153.

[17] C. Bessiere, G. Katsirelos, N. Narodytska, and T. Walsh, "Circuit complexity and decompositions of global constraints," in *Proc. of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)*, 2009, pp. 412–418.

[18] A. Semenov, "Decomposition representation of logical equations in problems of inversion of discrete functions," *Journal of Computer and Systems Sciences International*, vol. 48, no. 5, pp. 718–731, 2009.

[19] I. Otpuschennikov, A. Semenov, I. Gribanova, O. Zaikin, and S. Kochemazov, "Encoding cryptographic functions to SAT using Transalg system," in *ECAI 2016, Proceedings*, ser. Frontiers in Artificial Intelligence and Applications, vol. 285. IOS Press, 2016, pp. 1594–1595.

[20] "Irkutsk Supercomputer Center of SB RAS." [Online]. Available: http://hpc.icc.ru

[21] T. Yasumoto and T. Okuwaga, "Rokk 1.0.1," in *SAT Competition 2014*, A. Belov, D. Diepold, M. Heule, and M. Järvisalo, Eds., 2014, p. 70.

[22] A. Semenov, O. Zaikin, I. Otpuschennikov, S. Kochemazov, and A. Ignatiev, "On cryptographic attacks using backdoors for SAT," in *Proc. of the 32nd AAAI Conference on Artificial Intelligence (AAAI'2018)*, 2018, pp. 6641–6648.

Bug detection in embedded environments by fuzzing and symbolic execution

Juraj Vijtiuk
Sartura,
Zagreb, Croatia
juraj.vijtiuk@sartura.hr

Luka Perkov
Sartura,
Zagreb, Croatia
luka.perkov@sartura.hr

Antonio Krog
Sartura,
Zagreb, Croatia
antonio.krog@sartura.hr

***Abstract* — OpenWrt is an Open Source GNU/Linux distribution designed for embedded devices which, although primarily targeting home routers, can run on residential gateways, other IoT (Internet of Things) devices. RIOT is an Open Source, real-time multi-threading operating system running on numerous devices that are typically found in IoT. Fuzzing is a software testing process that uses random inputs to track unusual behaviors and crashes. Symbolic execution is a software testing technique which executes the program with symbolic instead of concrete variable values to analyze which values cause what path of the program to execute. This paper offers an overview of how we used fuzzing and symbolic execution to automatically detect crashes within OpenWrt and RIOT projects. Special focus will then be placed on detected vulnerabilities and methods for identifying such issues. Results collected by subjecting two widely used embedded distributions to fuzzing tests will be used to draw conclusions on the state of security in Open Source embedded software.**

Keywords — security, fuzzing, OpenWrt, RIOT

I. INTRODUCTION

Embedded systems continue to drive technological advancements in many domains such as electronics, healthcare, communication, home appliances and many more. The constraints of embedded devices and the incorporation of network connectivity into these devices open numerous fields of vulnerabilities that are exploited by attackers. Widely published attacks targeting commercial routers, IoT deployments and other embedded systems [1, 2, 3] reiterated the importance of embedded systems meeting high security and privacy demands and minimizing the attack surface.

The state of security in embedded systems stems from the market and the practice of how they are produced. Chipset manufacturers race against time to release newer chipset revisions and stack Open Source and proprietary features on them to stand out from the competition. These chips are purchased by original device manufacturers (ODMs) who add their own set of features on them, verify that the board is functional and ship them further on. The chipset manufacturer then focuses on shipping the next version of the chip, and the ODM on upgrading its product to work with the newer chip. Even when the hardware is brand new, the software that ships with them is usually old, and no link in the production chain has an incentive to maintain older chips and products.

The result of this state are millions of deployed devices running unpatched, unmaintained software, and the vulnerabilities they expose are being increasingly exploited by attackers. A 2019 survey by the Cyber Independent Testing Lab (CITL) [4] examined more than 6,000 firmware images spanning more than a decade and found that firmware security and security standards have not significantly improved over the last fifteen years, and even the most recent firmware images show failures to implement basic security features.

Open Source software is widely regarded as a more secure model than proprietary software, in large part because it leverages large communities that regularly test the software and provide insights and fixes for security vulnerabilities. Despite this, most Open Source software projects are still afflicted by ineffective security practices. The findings in this paper show how Open Source software can benefit by employing even the most basic fuzz testing setups to pinpoint potential security weaknesses.

The remainder of this article is organised as follows. In section II, we describe the automatic bug detection methods that were used. Section III describes the testing setup and Section IV the results of the testing. Finally, Section V concludes this work, and Section VI lists areas of improvement for future work.

II. AUTOMATIC VULNERABILITY DETECTION METHODS USED

This section introduces two automatic vulnerability detection techniques used in this paper: fuzzing and symbolic execution.

A. Fuzzing

One of the most prominent available methods of automatic discovery of software security vulnerabilities is fuzzing. Fuzzing, also known as fuzz testing, is a software testing process that uses random inputs with the goal of finding unusual behaviors and crashes. The number of academic research and projects related to fuzzing is constantly on the rise, with GitHub alone hosting around three thousand public repositories as of this writing [5].

As most modern fuzzers are mutational, the procedure of the majority of fuzzers is the following [6]:

1. The fuzzer chooses a corpus of "seed" inputs that will be used to test the target program.

2. The fuzzer repeatedly produces new inputs from at least one seed and any observation about the program.

TABLE I. OVERVIEW OF USED FUZZERS

	AFL [8]	AFLplusplus [9]	honggfuzz [10]	gramfuzz [11]	Dharma [12]	Radamsa [13]	libFuzzer [14]	Angora [15]
Collecting observations	gray-box	gray-box	gray-box	black-box	black-box	black-box	gray-box	gray-box
Open Sourced	yes	yes	yes	yes	yes	yes	yes	yes
In-memory Fuzzing	yes	yes	yes				yes	yes
Program Analysis	yes	yes	yes				yes	yes
Mutation	yes	yes	yes			yes	yes	yes
Model-based				yes	yes	yes		

3. Every new seed is evaluated on the program to produce an on observation.

4. The fuzzer records any mutated inputs that produce "interesting" behavior.

5. The fuzzer stops when it times out or reaches a pre-defined goal such as finding a particular bug.

Depending on the degree of recorded observations, fuzzers are generally classified into three groups: black-box fuzzers that observe only the input and output behavior of the tested program; white-box fuzzers that generate tests by analyzing and exploiting the program's internals (e.g. source or binary code) and the information collected during test runs; and grey-box fuzzers that obtain only some information about the program's internals to avoid the overhead created by gathering additional observations and thus enable faster test runs.

Based on how input is generated, fuzzers are traditionally divided into generation-based (model-based) or mutation-based fuzzers [7]. Generation-based fuzzers generate test cases based on provided models (e.g. network protocols, formal grammars, file formats or languages) which describe the input that the program may accept. Mutation-based fuzzers, on the other hand, rely on using and modifying seeds to provide inputs whose type is supported by the program.

B. Symbolic execution

Symbolic execution is another common technique used to automatically detect issues in software. Although widely used, symbolic execution is somewhat less popular than fuzzing. Disadvantages of symbolic execution compared to fuzzing are a lack of understanding of symbolic execution, susceptibility to path explosion in larger targets, and difficulty of setup with real world projects.

Essentially, symbolic execution is performed by a symbolic execution engine that interprets the program and - instead of concrete values - assumes symbolic values for the program's variables. The key point lies in the ability to set constraints on those symbolic variables, which are then checked during symbolic execution. An example constraint could be to check whether a variable which should only have positive integer values has taken on a negative value. A survey of symbolic execution techniques is available at [35].

As described in the original KLEE paper [36], an advantage of symbolic execution is that it can easily generate input test cases for all the paths that have been found in the target, which can then be used to extend existing unit tests.

III. TESTING SETUP

The methodology used throughout this paper aims to follow best practices for evaluating fuzzers detailed by Klees et al. [6]. This section describes the setup used for our experiment.

A. Fuzzers

Table 1. presents a summary of the fuzzers utilized in this paper and the differences in the techniques they use. The table's concept is loosely based on classification provided by Manès et al. [7]. The first row indicates the fuzzer's classification based on the degree of recorded observations. The second row denotes whether the fuzzer is publicly available. The third row shows whether the fuzzer supports in-memory fuzzing, a process where the fuzzer takes a memory snapshot of the program under test and restores it on each fuzz iteration to avoid execution overhead. The fourth row shows whether the fuzzer performs static or dynamic analysis to gather information about the program's internals. The fifth row indicates if the fuzzer mutates the input to generate new test cases, while the sixth row indicates whether the fuzzer generates test cases based on a provided model. We opted to use LibFuzzer as a starting point for most targets because it is easier to set up compared to AFL and enabled us to test a wider array of targets more quickly. For targets that were already more thoroughly tested, like json-c, we employed a variety of tools with a higher setup cost such as AFL, honggfuzz and grammar-based fuzzers, and gathered a larger starting input corpus.

B. Symbolic execution

We opted to use the KLEE symbolic execution tool [16], in large part due to its suitability for working with projects where the source code is available, as opposed to

other popular symbolic execution engines aimed at binary symbolic execution (angr [17], Manticore [18]).

of the test trials is given further in this section. The last column of the table represents the number of cores the

TABLE II. TEST RESULTS

Program	Fuzzer(s) used	Trials	Timeout	Crashes	Number of cores/threads/fuzzer instances
ubus	AFL	1	12H	0	4/8/8
	Radamsa	1	24H	0	1/1/1
UCI	AFL	1	3H	5 (1 unique)	4/8/8
		2	1H	3 (1 unique)	4/8/8
		3	12H	0	4/8/8
	AFL + gramfuzz + Dharma	1	12H	0	4/8/8
	Radamsa	1	24H	0	1/1/1
libubox (base64 encoder)	LibFuzzer	1	12H	0	1/1/1
libubox (base64 decoder)	LibFuzzer	1	12H	0	1/1/1
libubox (blobmsg_parse)	LibFuzzer (+ASAN + UBSAN)	1	12H	0	1/1/1
libubox (blobmsg_parse_array)	LibFuzzer (MSAN)	1	12H	1	1/1/1
libjson-c (json_tokener_parse_ex)	AFL	1	24H	0	4/8/8
	AFLplusplus	1	12H	0	4/8/8
	honggfuzz	1	24H	0	4/4/1
	LibFuzzer	1	12H	0	1/1/1
	Angora	1	12H	0	4/4/1
rpcd (rpc_canonicalize_path)	LibFuzzer	1	12H	0	1/1/1
odhcpd (odhcpd_valid_hostname)	LibFuzzer (ASAN + UBSAN)	1	12H	0	1/1/1
	LibFuzzer (MSAN + UBSAN)	1	12H	0	1/1/1
RIOT (base64_encode)	LibFuzzer (ASAN + UBSAN)	1	12H	0	1/1/1
RIOT (base64_decode)	LibFuzzer (ASAN + UBSAN)	1	12H	0	1/1/1
RIOT (golay2412_decode)	LibFuzzer (ASAN + UBSAN)	1	12H	0	1/1/1
RIOT (golay2412_encode)	LibFuzzer (ASAN + UBSAN)	1	12H	0	1/1/1
RIOT (sys/checksum)	LibFuzzer (ASAN + UBSAN)	1	12H[a]	0	1/1/1
RIOT (sys/bitfield)	LibFuzzer (ASAN + UBSAN)	1	12H	0	1/1/1

a.For each checksum function

C. *Programs*

The following software components were submitted to testing: UCI, ubus, libubox, libjson-c, rpcd, odhcpd and RIOT.

D. *Performance measure*

Our experiments measured the number of crashes reported by the fuzzer over a period of time, with the same starting corpus. Remarks will be made with instances where crashes are caused by the same bug.

E. *Platform and configuration*

All fuzzing experiments were conducted on a single machine equipped with a four core Intel i7-8565U CPU with sixteen gigabytes of RAM, running Arch Linux. All fuzzers were built from source using the master branch, and all the target programs were also built from source using the master branch. For KLEE, we used the official KLEE Docker image and built the programs that were tested locally in that Docker image.

IV. RESULTS

Table 2. shows an overview of the fuzzing experiments and collected results. A detailed description

fuzzing jobs were running on, the number of threads used by the jobs (the CPU used had hyperthreading which showed up as eight cores), and the number of fuzzer instances that were run. The differences are most apparent between AFL and hongfuzz: honggfuzz automatically starts a parallel workload, resulting in one honggfuzz instance running four fuzz workers, one for each physical core. On the other hand, AFL's *afl-gotcpu* tool reported that eight cores were available, so we ran eight instances of AFL with each instance using two hyperthreads on four physical cores. All programs fuzzed with AFL were fuzzed with eight instances of AFL, namely one master instance and seven secondary instances.

A. *ubusd*

Initial testing was done on *ubusd*, an integral part of the OpenWrt's micro bus architecture named *ubus* [19]. The ubusd daemon provides an interface used by other daemons to register themselves. To fuzz ubusd, we wrote a custom harness that connected to the ubus UNIX socket and passed in files that were generated by the fuzzer. The harness starts by reading the contents of a file specified by the fuzzer into a buffer, then spawns a thread which reads the contents of the buffer and sends them to the ubus UNIX socket. The main thread continues with the start up

of the ubus server unmodified. We based this on existing approaches for fuzzing network servers [20]. The component was fuzzed with AFL for twelve hours using a single starting input file: a simple text file containing the text "list". The input is then randomly mutated with code coverage increases used as a measure of success by AFL. At first glance, the minimal input used here might seem overly simple to produce significant results. However, when testing smaller projects that do not use fuzzing, we discovered that using a single simple input as a starting point can quickly produce crashes in the majority of tested projects. In cases where no crashes are detected, a larger corpus of input files can be used for further processing and test trials. This trial produced no errors.

B. UCI

Next, we fuzzed OpenWrt's main configuration interface named *Unified Configuration Interface (UCI)*. A custom harness was written which imported configuration files generated by the fuzzer. The initial corpus contained a collection of OpenWrt UCI configuration files that were minimized using AFL's tools for minimizing test cases (*afl-tmin*) and corpus (*afl-cmin*) [21, 22]. On the first trial lasting three hours, AFL reported five crashes after two hours of fuzzing, all of which were caused by the same bug [23]. The second fuzzing run lasting an hour also revealed another bug in the first fifteen minutes, where numerous crashes reported as "unique" by AFL later turned out to be caused by the same bug [24]. A third run lasting twelve hours revealed no additional bugs.

To combat the lack of discovered vulnerabilities, we generated 1 million UCI configuration files by writing a UCI configuration grammar for the gramfuzz fuzzer. Even though none of the 5 million generated files caused crashes in UCI, they were leveraged in subsequent fuzzing iterations as a starting AFL corpus. The combined corpus of 1 million grammar-generated configurations and the previously used OpenWrt configurations were additionally minimized into a corpus containing eighty-nine input configs. An additional 1 million configuration files were generated using the grammar-based fuzzer Dharma. The grammar used for Dharma was based on the grammar used with gramfuzz.

Parsing the grammars generated by Dharma did not result in any new crashes, however it did result in additional input files which revealed new paths in UCI. The total number of UCI config input files after minimization was 149. UCI and ubusd were fuzzed using only vanilla AFL. Additionally, both the UCI and UBUS harnesses were fuzzed with Radamsa for twenty-four hours which found no issues.

C. libubox.

After UCI we moved onto fuzzing the *libubox* [25] OpenWrt library, which is used extensively throughout the OpenWrt project. We started fuzzing libubox using LLVM's LibFuzzer [14], which allows feeding fuzzed inputs to the library under test via a specific entrypoint. Initial fuzzing trials were done on libubox base64 encoder and decoder, but no bugs were found after twelve hours of fuzzing both components. We combined LibFuzzer with both the *AddressSanitizer (ASAN)* [26] and *UndefinedBehaviorSanitizer (UBSAN)* [27] and fuzzed *blobmsg_parse* for twelve hours. As we expected, this approach did not discover any new crashes since the OpenWrt community started fuzzing the same functions with LibFuzzer a few weeks earlier. Additionally, we combined LibFuzzer with the *MemorySanitizer (MSAN)* [28] and fuzzed libubox. Using this combination, we uncovered and fixed a series of issues with improper usage of flexible arrays in C structures in libubox, which appeared at multiple places in *blobmsg_parse* and *blobmsg_parse_array*. We also employed KLEE to test libubox. KLEE immediately found the out-of-bound read issues that were found by fuzzing with LibFuzzer and MSAN. With contributions by OpenWrt developers, these fixes were moved to generic blob functions and macros to prevent similar issues in other use cases utilizing these structures and to minimize future issues. We continued by testing the base64 encoder and decoder, *blobmsg_parse*, *blobmsg_parse_array*, and functions in *blobmsg_json.c* with KLEE, with no issues being found. Around the same time we were fuzzing *blobmsg_json* functions, a security advisory was released regarding a potential stack buffer overflow during serialization of JSON data in *blobmsg_format_json* [29]. The advisory did not describe how the issue was found, however the example proof of concept was triggered via ubus and rpcd, so we assume that deeper fuzzing of ubus or rpcd with AFL could - given enough time - find the issue, even though LibFuzzer did not find it in our case.

The analysis of the issue was done with a combination of manual code analysis, analysis in the rr debugger [30], and valgrind output for the program. We started by writing a simple main in *blobmsg_parse_array.c* which opened a file, read it into a buffer and passed it to the *blobmsg_parse_array* function. The target was then compiled without any sanitizers and ran through valgrind. The valgrind output is shown in Figure 1. The input file that caused the initial issue is shown in Figure 2. As the file contains binary data, the output of the xxd command which outputs the file's bytes in hexadecimal is shown.

```
==10829== Invalid read of size 2
==10829== at 0x109DFC: blobmsg_namelen
(blobmsg.h:74)
==10829== by 0x109446: blobmsg_check_name
(blobmsg.c:42)
==10829== by 0x1092DD: blobmsg_check_attr_len
(blobmsg.c:79)
==10829== by 0x109A63: blobmsg_parse_array
(blobmsg.c:159)
==10829== by 0x10A7BA: main (blobmsg.c:412)
==10829== Address 0x4a2e2b4 is 0 bytes after a
block of size 4 alloc'd
==10829== at 0x483877F: malloc
(vg_replace_malloc.c:309)
==10829== by 0x10A773: main (blobmsg.c:408)
```

Figure 1. Valgrind output

```
$ xxd crash-
a3585b70f1c7ffbdec10f6dadc964336118485c4

00000000: 0300 0004
```

Figure 2. xxd output

The *blobmsg_parse_array* function tries to interpret the received data as a *blob_attr* structure, which consists of a *uint32_t* followed by a flexible array member. The existing check in libubox checked whether the received data is smaller than *sizeof(struct blob_attr)*, however as the flexible array member's size is counted as 0, the input shown above passes that check since it is exactly 4 bytes large. The issue then appears when accessing the flexible array member, even though it is empty in this case. After adding checks to prevent this issue, we continued fuzzing *blobmsg_parse_array*, which revealed additional issues of the same type. The *blob_attr* flexible array member is parsed into another structure that ends with a flexible array member *blobmsg_hdr*. The same out-of-bounds read issue appeared, so similar checks were added to prevent that issue. The same type of issues were noticed and fixed in *blobmsg_parse*. However, as the two above mentioned functions are not the only places where *blobmsg_hdr* and *blob_attr* are used, OpenWrt developers expanded the fix to account for other places where the issue might appear.

D. json-c

We fuzzed *json-c* for twenty-four hours using LibFuzzer compiled with ASAN and UBSAN and utilized three LLVM fuzz jobs, each one running on a separate core, but did not discover any issues. Next, we generated 1 million JSON files with Dharma, but this approach also did not find any issues.

Afterwards, we fuzzed json-c using AFL and leveraged a corpus containing JSON files collected from the Internet and JSON files generated by Dharma. This approach did not discover any issues in twenty-four hours. The same corpus was utilized when fuzzing json-c with hongfuzz for twenty-four hours, but this approach also did not find any issues. Fuzzing json-c with AFLplusplus for twelve hours also discovered no issues, although AFLplusplus proved to be much more successful at discovering paths than AFL: AFL discovered around 600 paths in twenty-four hours, while AFLplusplus discovered around 1300 paths in twelve hours with the same corpus. The AFLplusplus configuration consisted of eight instances: one master instance running with the exploit power schedule, three secondary instances running with a COE power schedule, two secondary instances running with the fast power schedule and two secondary instances running with the explore power schedule. Since we were unable to find bugs using fuzzing, we tested json-c with KLEE by writing and running a simple wrapper that calls *json_tokener_parse_ex* with a symbolic JSON string, but this approach also did not find any issues. Ultimately, no bugs were found in json-c, which was expected, as json-c is regularly fuzzed by Google's distributed fuzzing project, OSS-Fuzz [31].

E. RIOT

When we started looking into what part of RIOT would be best to test with fuzzers, we first consulted previous work in this field. A number of GitHub issues and pull requests already exist on the RIOT-OS github page [32]. Previous issues with coap parsing and the TCP stack were found and fixed by fuzzing. A custom black-box fuzzer for fuzzing CoAP server side implementations *fuzzcoap* was also used to test RIOT's CoAP code [33]. The RIOT 2017 Summit featured a presentation of work by Eric Sesterhenn, which resulted in 40 issues being reported in RIOT and in projects related to RIOT. The work also used fuzzing procedures and utilized honggfuzz, LibFuzzer, radamsa and AFL fuzzers.

Initially, we started to fuzz RIOT with AFL by writing a custom RIOT-OS application which would take the role of the harness and pass AFL inputs to the GNRC networking functions available in RIOT. However, due to time constraints we did not manage to successfully build such a harness, as RIOT abstracts away a lot of the build process behind Makefiles, so we ran into various issues while trying to build the custom application and instrumenting the core of RIOT. Instead, due to the ease of setup, we opted to fuzz as many separate functions with LibFuzzer. Other contributors have managed to set up fuzzing with AFL, and as of this writing a pull request was opened on RIOT-OS's GitHub with support for fuzzing the GNRC networking functions with AFL [34]. The author of the pull request has previously managed to find a dozen crashes and security issues in the fuzzed targets by using the setup available in the pull request.

Similarly to libubox, we started by fuzzing the base64 encoder and decoder for six hours with no starting inputs and instrumented with ASAN and UBSAN, but this approach did not result in any detected crashes. We continued by fuzzing the *golay2412_decode* and *encode* elliptic curve functions with the same configuration, but did not find any issues. Next, we fuzzed RIOT-OS's bitfield implementation and the checksum functions using the same configuration, but this did not reveal any bugs.

V. CONCLUSION

This paper, the experiments conducted within it and the collected results are not meant to provide a comprehensive analysis of available automatic bug detection techniques. Instead, the insights from this paper aim to emphasize the benefits that even widely used, reliable Open Source software can gain from elementary, easy-to-setup fuzzing tests. The wide list of available fuzzers can be used and combined to discover shallow bugs within the first few hours of testing. The fact that most modern software is fuzzed for weeks if not months at a time, and that all of the programs fuzzed in this paper were fuzzed for at most a day and discovered crashes in the process emphasizes the need to include fuzzing as another form of testing Open Source projects. The differences are especially noticeable when comparing json-c, a thoroughly tested library, for which we did not manage to find any issues with multiple approaches, with UCI and libubox which have not been thoroughly tested yet. Fuzzing with LibFuzzer resulted mostly in shallow, easily-reachable crashes, while a longer-lasting fuzzing run using a starting corpus of inputs with decent coverage should result in discovering more obscure bugs.

In addition to fuzzers, this paper highlighted the benefits from another underutilized form of automated bug detection: symbolic execution. Although easier to set up elementary testing, symbolic execution quickly revealed bugs that might require more thorough fuzzing

setups to detect as well as pinpointed issues that go undetected by the majority of utilized fuzzers. Symbolic execution tools, especially KLEE, provide an added benefit where even in cases where no bugs are discovered, the tools generate inputs for all unique paths discovered during exploration. These inputs can later on be used to expand existing or to create new unit tests that improve the security and stability state of tested programs. In-depth testing with KLEE, however, still requires the tester to be familiar with the target codebase.

VI. Future work

The grammars generated with gramfuzz and Dharma were mostly invalid. We will invest the effort to write better grammars which will cover a wider array of types of invalid UCI configurations, as well as generate more valid configurations.

In this paper we only managed to fuzz RIOT-OS with LibFuzzer by separately fuzzing functions of interest. Future work could focus on fuzzing RIOT-OS with AFL by using a custom RIOT-OS application as the harness. This would enable discovering deeper bugs in the code, since LibFuzzer is more suited for discovering shallower bugs. Another approach would be to reuse the test cases generated by KLEE as additions or starting points for a fuzzing corpus, which would improve fuzzer coverage.

Although KLEE has made huge advancements by providing its own symbolic versions of libc and the POSIX environment, we encountered issues when testing software that used GNU extensions to libc, or GNU-specific functions. The workaround in this case was to manually patch the software to replace the functionality with non-GNU functions. Another area of future work would consist of adding additional implementations of such functionality.

References

[1] S. Larson, “Massive cyberattack targeting 99 countries causes sweeping havoc”, http://money.cnn.com/2017/05/12/technology/ransomware-attack-nsa-microsoft/index.html, 13. 5. 2017

[2] C. Cimpanu, “Brazil is at the forefront of a new type of router attack”, https://www.zdnet.com/article/brazil-is-at-the-forefront-of-a-new-type-of-router-attack/, 12. 7. 2019

[3] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas and Y. Zhou, “Understanding the Mirai Botnet”, In USENIX Security Symposium, 2018.

[4] P. Roberts, “Huge Survey of Firmware Finds No Security Gains in 15 Year”, https://securityledger.com/2019/08/huge-survey-of-firmware-finds-no-security-gains-in-15-years/, 14. 8. 2019.

[5] GitHub, “Public Fuzzers”, https://github.com/searchq=fuzzing&type=Repositories, 25. 1. 2020.

[6] G. Klees, A. Ruef, B. Cooper, S. Wei and M. Hicks, “Evaluating Fuzz Testing”, In 2018 ACM SIGSAC Conference on Computer and Communications, 2018.

[7] V. Manès, H. Han, C. Han, S. K. Cha, M. Egele, E. J. Schwartz and M. Woo, “The Art, Science and Engineering of Fuzzing: A survey”, In IEEE Transactions on Software Engineering, 2018.

[8] M Zalewski, “American Fuzzing Lop (AFL)”, http://lcamtuf.coredump.cx/afl/, 26. 9. 2019.

[9] M. Hauser, H. Eißfeldt, A. Fioraldi, D. Maier, “American Fuzzing Lop Plus Plus (AFL++)”, https://github.com/vanhauser-thc/AFLplusplus, 31. 12. 2019.

[10] R. Swiecki and F. Gröbert, “honggfuzz”, https://github.com/google/honggfuzz, 7. 12. 2019.

[11] J. Johnson, “gramfuzz”, https://github.com/d0c-s4vage/gramfuzz, 2. 1. 2020.

[12] Mozilla Security, “Dharma”, https://github.com/MozillaSecurity/dharma, 5. 12. 2019.

[13] A. Helin, “Radamsa”, https://gitlab.com/akihe/radamsa,7. 5. 2019.

[14] The LLVM Project, “LibFuzzer”, https://llvm.org/docs/LibFuzzer.html, 13. 1. 2020.

[15] P. Chen, H. Chen, “Angora”, https://github.com/AngoraFuzzer/Angora, 17. 7. 2019.

[16] The KLEE Team, “KLEE LLVM Execution Engine”, https://klee.github.io/, 2019.

[17] Shoshitaishvili, Yan and Wang, Ruoyu and Salls, Christopher and Stephens, Nick and Polino, Mario and Dutcher, Audrey and Grosen, John and Feng, Siji and Hauser, Christophe and Kruegel, Christopher and Vigna, Giovanni, “SoK: (State of) The Art of War: Offensive Techniques in Binary Analysis”, In *IEEE Symposium on Security and Privacy*, 2016

[18] M. Mossberg, F. Manzano, E. Hennenfent, A. Groce, G. Grieco, J. Feist, T. Brunson, A. Dinaburg, “Manticore”, https://github.com/trailofbits/manticore, 12. 11. 2019.

[19] The OpenWrt project, “ubus (OpenWrt micro bus architecture)”, https://openwrt.org/docs/techref/ubus, 11. 1. 2020.

[20] javi, “Fuzzing Apache httpd server with American Fuzzy Lop + persistent mode” , https://animal0day.blogspot.com/2017/05/fuzzing-apache-httpd-server-with.html, 7. 5. 2017.

[21] D. Stender, “afl-tmin(1)”, http://www.tin.org/bin/man.cgi?section=1&topic=afl-tmin, 20. 1. 2020.

[22] D. Stender, “afl-cmin(1)”, http://www.tin.org/bin/man.cgi?section=1&topic=afl-cmin, 20. 1. 2020.

[23] Luka Kožnjak, Juraj Vijtiuk, openwrt-devel mailing list, “[OpenWrt-Devel] [PATCH] file: fix segfault in uci_parse_option”, http://lists.openwrt.org/pipermail/openwrt-devel/2019-December/026597.html, 28. 12. 2019.

[24] Luka Kožnjak, Juraj Vijtiuk, openwrt-devel mailing list, “[OpenWrt-Devel] [PATCH] file: fix segfault in uci_parse_config”, http://lists.openwrt.org/pipermail/openwrt-devel/2019-December/026600.html, 28. 12. 2019.

[25] The OpenWrt Project, “libubox”, https://git.openwrt.org/project/libubox.git, 20. 1. 2020.

[26] The LLVM Project, “Clang 11 documentation - AddressSanitizer”, http://clang.llvm.org/docs/AddressSanitizer.html, 2020.

[27] The LLVM Project, “Clang 11 documentation - UndefinedBehaviorSanitizer”, http://clang.llvm.org/docs/UndefinedBehaviorSanitizer.html, 2020.

[28] The LLVM Project, “Clang 11 documentation - MemorySanitizer”, http://clang.llvm.org/docs/MemorySanitizer.html, 2020.

[29] P. Štetiar, J. Wich, “Security Advisory 2020-01-22-2 - libubox tagged binary data JSON serialization vulnerability (CVE-2020-7248)”, https://openwrt.org/advisory/2020-01-31-2, 31. 1. 2020.

[30] Mozilla, “rr project homepage”, https://rr-project.org/, 2017.

[31] Google, “OSS-Fuzz”, https://github.com/google/oss-fuzz, 2020.

[32] “RIOT - The friendly OS for IoT”, https://github.com/RIOT-OS/RIOT, 31. 10. 2019.

[33] B. Melo, “FuzzCoAP”, https://github.com/bsmelo/fuzzcoap, 23. 6. 2018.

[34] “RIOT - The friendly OS for IoT”, “Add AFL-based fuzzing setup for network modules #13157”, https://github.com/RIOT-OS/RIOT/pull/13157, 17. 1. 2020.

[35] E. J. Schwartz, T. Avgerinos, D. Brumley, “All You Ever Wanted to Know about Dynamic Taint Analysis and Forward Symbolic Execution (but Might Have Been Afraid to Ask)”, In IEEE Symposium on Security and Privacy, 2010.

[36] C. Cadar, D. Dunbar, D. Engler, “KLEE: Unassisted and Automatic Generation of High-Coverage Tests for Complex Systems Programs”, 2008.

Forensic analysis of Windows 10 Sandbox

A. Đuranec[1], S. Gruičić[1], M. Žagar[2]
[1] INsig2 d.o.o., Zagreb, Croatia
[2] Zagreb University of Applied Sciences, Zagreb, Croatia
Antun.Duranec@insig2.com, Savina.Gruicic@insig2.com, Marinko.Zagar@tvz.hr

***Abstract* - With each Windows operating system Microsoft introduces new features to its users. Newly added features present a challenge to digital forensics examiners as they are not analyzed or tested enough. One of the latest features, introduced in Windows 10 version 1909 is Windows Sandbox; a lightweight, temporary, environment for running untrusted applications. Because of the temporary nature of the Sandbox and insufficient documentation, digital forensic examiners are facing new challenges when examining this newly added feature which can be used to hide different illegal activities. Throughout this paper, the focus will be on analyzing different Windows artifacts and event logs, with various tools, left behind as a result of the user interaction with the Sandbox feature on a clear virtual environment. Additionally, the setup of testing environment will be explained, the results of testing and interpretation of the findings will be presented, as well as open-source tools used for the analysis.**

Keywords - Windows 10; Sandbox; digital forensics

I. INTRODUCTION

With each Windows update, digital forensics examiners encounter new features that can contain evidence of illegal activities and/or can be used to hide them. To understand records generated by the Windows operating system, also known as the Windows artifacts, proper testing must be conducted. Throughout this paper, testing of newly added Windows Sandbox feature will be conducted on the clear virtual machine, while the focus is on interpreting different Windows artifacts leftover as the result of usage of the Windows Sandbox feature. Additionally, as records are analyzed, the possibility to use the Sandbox to hide data as an anti-forensic method will be addressed.

The paper is organized as follows: section 2 and 3 introduce digital forensics and setup of the testing environment, section 4 describes Windows Sandbox and the procedure for enabling the feature and section 5 combines multiple Windows artifacts, event logs and interpretation of all findings related to the usage of Windows Sandbox.

II. DIGITAL FORENSICS

Digital forensics is a branch of forensic science that includes the extraction, investigation, preservation, and presentation of digital data found in various electronic devices, often in relation to computer crime [1]. The first appearance of digital forensics was back in the late 1980s and early 1990s as a result of the appearance of personal computers and digital crime [2]. The rapid growth of digital crime forced law enforcement agencies to establish special cyber teams to carry out challenging investigations that require knowledge of information and communication technology.

As a result of the fast and rapid spread of various electronic devices, it is hard to find a crime that does not include digital evidence. Technology is used by criminals to commit a crime and to evade apprehension which presents new problems for law enforcement agencies and forensic investigators [3]. Development of many different electronic devices with specific capabilities and their usage in a crime lead to the expansion of digital forensics into branches; computer, mobile, network, and cloud forensics, that are associated with the specific capabilities of digital devices involved in digital crime. The digital forensic examination starts when a crime is committed and reported, and it includes preparation, identification of potential evidence, seizure and acquisition, analysis and report or presentation of findings.

III. TESTING ENVIRONMENT

While running, Windows operating system constantly changes data and records different information on the machine. This is done while the system is running and before shutting down. Because of many records, left after each usage of the system, testing is recommended on a clean machine or virtual environment. This approach, using a virtual environment, lowers the possibility of misinterpreting the records and provides an easy way to restore specific states of a system.

For setting up the testing environment the VMware Workstation Pro version 15 virtualization software was used. VMware Workstation Pro is a commercial tool that allows virtualization of the operating system on Windows or Linux host machines. Some of the features that VMware Workstation Pro has are high-performance 3D graphics, virtual networking, rapid duplication and snapshots [4].

Before creating a new testing environment in VMware Workstation, the Windows 10 Pro version 1909 operating system ISO image, also known as a disk image, must be downloaded. The image can be downloaded from the official Microsoft site, over the tool for the creation of installation media [5]. The next step is the creation of a virtual environment in the VMware Workstation by clicking the “File” drop-down menu in the top right corner and selecting the option “New Virtual Machine…” which

opens wizard dialog. For easier setup, the typical configuration is selected. In the section of guest operating system installation, the installer disc image file that was downloaded from official Microsoft is selected by selecting the location path to the ISO file. After that, Windows 10 Pro system version was chosen for system and all further steps are left with default values. Once the above-mentioned steps are completed, the virtual machine can be powered on and the Windows installation process will start.

IV. WINDOWS 10 SANDBOX

Windows Sandbox is one of the latest features added to the Windows 10 operating system, limited to the Pro and Enterprise edition. Microsoft defines this feature as an isolated, temporary, desktop environment for running untrusted software without the possibility to damage the user's machine. Furthermore, once when Sandbox is closed, all software, files, and state of the Sandbox environment are permanently deleted [6]. Because of the mentioned properties, a concern of using this new feature as an anti-forensic method is questioned. Anti-forensic methods are defined as a set of techniques used as countermeasures to digital forensic analysis which often include data hiding, artefacts wiping, trail obfuscation, and attacks on the forensic tools [7]. With the temporary property of Sandbox, users can hide web searches, illegal purchases, communication, program execution, etc. Furthermore, searching for related articles and terms as "digital forensics and Windows Sandbox" over Google Scholar resulted in articles related to malware analysis and usage of a different sandbox software, which is not relevant to this paper and examination of artifacts.

To conduct testing of Windows Sandbox the feature must be enabled on the system. The first step is enabling the visualization engine in VMware Workstation Pro. For this, the user needs to enable "virtualize Intel VT-x/EPT or AMD-V/RVI" option by clicking on the checkbox in the processor's section of virtual machine settings. To successfully enable this feature, the virtual machine needs to be shut down. The next step is to enable the Sandbox feature on Windows 10 system. This step is done in the Windows features dialog. One way to open the dialog is to type "Windows feature" in the Windows search bar next to the start menu. Once the dialog is open, select the Windows Sandbox feature, confirm it by clicking the "ok" button and the Windows system will automatically download the required components. Additionally, a system restart may be needed.

V. ANALYSIS OF WINDOWS SANDBOX

In this section, the focus is on the analysis of different Windows artifacts, folder structure, and event logs related to the usage of the Windows Sandbox feature. A comparison of the system's folder structure prior and after running the Sandbox is conducted. Furthermore, Windows artifacts, such as prefetch, UserAssist, and RDP (*Remote Desktop Protocol*) cache are analyzed and explained in corresponding subsections.

The analysis is conducted after shutting down the Sandbox feature to realistically represent a state of the host machine that digital forensic examiners encounter, as in most cases examiners analyze shut down system; also known as dead-box analysis.

A. Folder structure

Each Windows operating system has a similar default folder structure created through the installation of an operating system. The default folder group of interest is program files group that contains most of the applications installed on the system [8]. Often when installing additional features, in this case, Windows Sandbox, we see changes in certain parts of folder structure; often new folders are created, or new content is added to them. For comparison of folder structure, the free forensic tool FTK Imager developed by the AccessData company is used.

FTK Imager allows us to load specific logical and/or physical disk, in this case, to examine changes in the folder structure. This is done over a functionality to add new evidence item which asks examiner to choose a physical drive, logical drive, image file or content of folder as new evidence. Virtual machine for testing was created with two logical drives; drives C and D as shown in Fig. 1. The C drive contains Windows operating system while D drive represents a boot image that is required to run a virtual machine.

After enabling the Sandbox feature and restart of the testing environment, the same procedure to load the logical drive was repeated. In the drop-down menu of FTK Imager new logical drive is available for loading. Usually, the drives are represented only with a letter while this new drive is represented with full path, as shown in Fig. 2.

The drive is part of the Windows Sandbox feature and it is known as a dynamic base image that is required for the successful running of Windows Sandbox [6]. The new drive does not prove that the Sandbox feature was run on the system, but an examiner can confirm that the feature was enabled by the user. The full path to the new drive is "C:\ProgramData\Microsoft\Windows\Containers\BaseImage\<*GUID*>\BaseLayer\". As the new drive leads to a specific path, we can focus on it, compare it to the state before Sandbox, and identify changes. By comparison of the folder structure and the specific path, we can conclude that after enabling the Sandbox feature new folder "Containers" and its subfolders are created, as shown in Fig. 3.

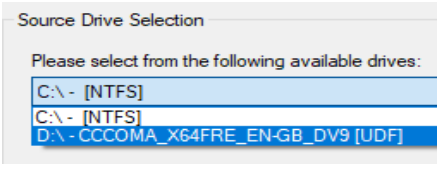


Figure 1. Logical drives on the virtual machine before the Windows Sandbox feature

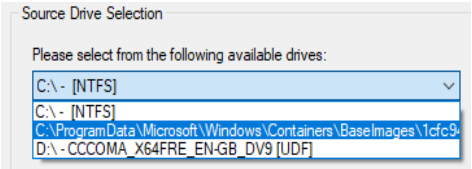


Figure 2. Logical drives on the virtual machine after enabling of the Windows Sandbox feature

The "BaseImage" folder contains the BaseLayer.vhdx file which is a virtual hard disk. It is located in a subfolder with a Globally Unique Identifier (GUID). The BaseLayer.vhdx does not contain any useful data for examiners. To check the above statement, we can load the "BaseImage" drive in FTK Imager and check the content. On Fig. 4. within the orange border, we can see the default Windows folder structure that every Windows operating system has. An additional folder that is not found in classical Windows systems is "Hives" marked with the red border. It contains base Registry files needed for the proper running of the Windows operating system virtualized through the Sandbox feature. These files contain low-level settings so that the user does not need to set up his operating system on every startup [9].

The procedure for comparing the folder structure and the files in them was repeated once again after running the Sandbox feature. This time a new sub-folder is created in the "Sandboxes" folder on the testing environment. The folder is named "c316db96f9d1-4778-a88a-bf19dfd7ec7f" and it contains "sandbox.vhdx". The modified date associated with the Virtual Hard Drive (VHDX) file matches the run time of the Sandbox feature. Additionally, while using Sandbox, the VHDX increased multiple times in size. To be sure that this VHDX represents actual drive used in Sandbox, a new file on the Sandbox environment was created. The file contains the string "This is my sandbox test! 16/12/2019". The FTK Imager was used to

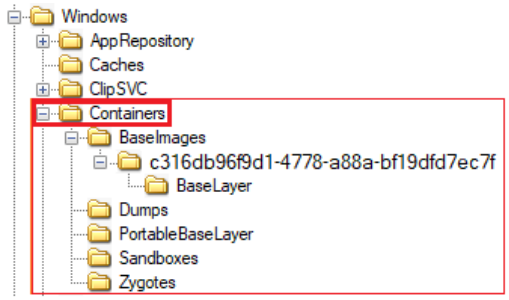


Figure 4. Folder structure after enabling of the Windows Sandbox feature

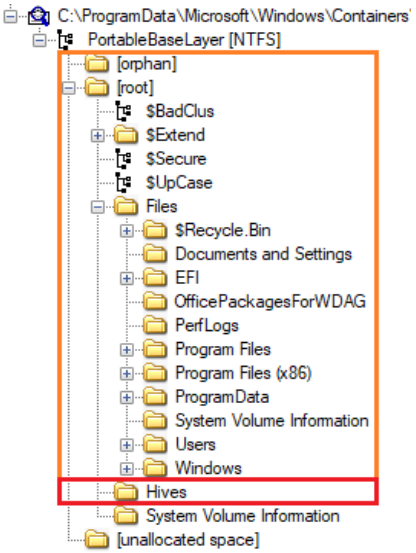


Figure 3. Folder structure of the third logical disk

open VHDX in hex view and find option was used to search for the string. Once that search was completed the matching string was found in the sandbox.vhdx file which confirms that the mentioned VHDX is used as a local drive by Sandbox feature.

After conducted tests, the Sandbox environment was shut down and the folder structure was checked once again to test Sandbox's temporary property. The "c316db96f9d1-4778-a88a-bf19dfd7ec7f" folder and sandbox.vhdx are marked as deleted immediately after shutdown. Furthermore, the sandbox.vhdx file has a size of 0 bytes which eliminates the possibility to recover this virtual hard drive, meaning that all content of drive is lost. The process was repeated and this time there was no evidence that the folder or VHDX existed in the folder structure.

With all before mentioned changes in the folder structure and additional drive that appears after enabling Sandbox, the examiner can determine that the feature is installed and enabled on the system but without specific evidence of how the Sandbox was used by the user. Furthermore, once when Sandbox environment is closed the VHDX is deleted without the possibility to recover it.

B. Prefetch

Prefetch is a well-known artifact in the digital forensics field that records the execution of programs on the Windows operating system. It was introduced with Windows XP operating system to reduce the time needed to load frequently used programs [10]. This artifact gives examiners great insights into program usage by recording the executable' name, up to nine execution date and times, run count, and path of the executable file [10]. This information is recorded inside ten seconds of the first and every next execution. Often on the system examiners can find multiple prefetch files for the identical program. Prefetch records programs by their location which means that if a user has the same program on different locations on the system each of these programs will have their prefetch file [10]. To distinguish the same programs Microsoft implemented the hashing function of executable' location path in the prefetch file name. Prefetch files have extension ".pf" and name format is "<AppName>.EXE-XXXXXXXX where the "XXXXXXXX" represents before mentioned location hash. The default location for prefetch files is on system drive "C:\Windows\Prefetch".

To examine the prefetch file related to the Sandbox feature the PECmd command-line open-source tool is used [11][12]. This simple tool allows us to analyze prefetch files and create a report of executable usage, used directories, and loaded files. To analyze prefetch records in the tool, the examiner needs to insert prefetch location, mentioned before, and output location for the report. The example of the command line to run the program is "PECmd.exe -d C:\Windows\Prefetch --html *<output directory>*". The attribute "-d" defines the location of the prefetch directory while "--html" defines the format of the report. As the report is created in Hypertext Markup Language (HTML) we can open it in any available internet browser and search for Sandbox record. The

prefetch file of interest is "WINDOWSSANDBOX.EXE-<*Hash*>.pf". In the report, an examiner can find executables created, modified and access times that are related to the Windows operating system. Furthermore, related specifically to prefetch, we can see run count with value two and associated two run times, size of executable, path's hash and operating system that generated the prefetch file. This finding is very interesting as the Sandbox was run only one time but the prefetch count has the value of two. Furthermore, the times related to the two run counts have a difference of only two seconds which is suspicious, as shown in Fig. 5. Because of that, the Sandbox was run two more times to check if every launch increases prefetch run count for two. For each additional launch of the Sandbox feature, the prefetch run count was indeed increased for two, as shown in Fig. 6. It is possible that Sandbox executable must be executed two times by computer when starting the Windows Sandbox feature which cloud be an explanation for prefetch pairs seen in the tests.

With prefetch analysis examiners can define when was Sandbox run by a user, how many times and from which location. The issue is that prefetch is system global, in other words, it is not specific to the one user.

C. *UserAssist*

UserAssist artifact was introduced in Windows XP operating system with the purpose to populate the start menu. This valuable record is stored in Windows Registry, more specifically in user-specific file NTUSER.DAT, which contains user's settings [13]. The record in NTUSER.DAT is located at "NTUSER.DAT\Software\Microsoft\CurrentVersion\Explorer\UserAssist\{*GUID*}\Count". The Global Unique Identifier (GUID) that contains a list of the executed applications is "CEBFF5CD-ACE2-4F4F-9178-9926F41749EA".

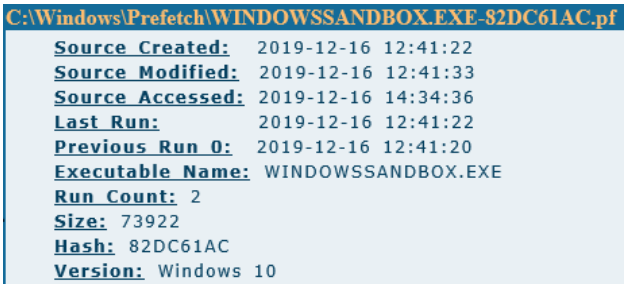


Figure 5. PFCmd report result related to the Sandbox feature after the first launch



Figure 6. PFCmd report result after two additional launchings of the Sandbox feature

To analyze this artifact the NTUSER.DAT file first needs to be extracted from the system. One of the tools that can extract it is FTK Imager with functionality to obtain system files. Note that this functionality works only on the running system. The examiner needs to specify the extraction folder and all other settings can be left default. To analyze the UserAssist the Registry Explorer open-source tool is used [12][13]. To load the before extracted file the "Show hidden files, folders, and drives" folder option needs to be enabled. After the mentioned steps the file is loaded in Registry Explorer and the examiner can search for Sandbox entry in UserAssist. The entry with the Windows Sandbox feature differs from prefetch findings. In UserAssist, Sandbox is recorded as WindowsSandbox.exe with the run count value three, which matches the exact run number of the Sandbox feature. Furthermore, UserAssist contains records of focus count and time as well as last execution time, as shown in Fig. 7. Focus count represents how many times the user selected, in this case, Sandbox as a top window on-screen while focus time represents its duration. Furthermore, the last executed time doesn't match the last run in prefetch artifact, it matches "previous run 0".

D. *Event logs associated with Sandbox*

Event logs record a wide variety of events on the Windows operating system related to the hardware and software. When enabled they can provide detailed information on what was occurring on the system at a given time. Examiners can use these records to search for Sandbox feature related logs. Logs are recorded in set of files with the extension .evtx (Event log) located at "C:\Windows\System32\winevt\logs".

The first sets of event logs are related to the Hyper-V; feature that provides hardware virtualization on Windows systems. This feature is available on Windows Pro, Enterprise and Education versions [14]. The Hyper-V event logs are associated with Windows hypervisor which is used by Sandbox for virtualization of the system. Because of that, they can contain a lot of information about Sandbox, but the event logs will not contain information on how a user interacted with the Sandbox environment. To ensure that we examine only events related to Sandbox, the clean snapshot of the testing environment was started with all event logs related to Hyper-V being deleted. After the restart of the Sandbox environment, 90 logged events were recorded in the Hyper-V-Compute log set, in one minute, with no user interaction in the Sandbox environment. Hyper-V-Compute record information about container management Application Programming Interface (API) also known as Host Compute Service (HCS) [15]. All the 90 logged events in this set are related to process ID 2072 which reference "vmcompute.exe" or more specifically HCS. In the list of logged events, after the first execution of Windows Sandbox, an examiner can find logs with event ID 2018 related to the creation of

Program Name	Run Counter	Focus Count	Focus Time	Last Executed
ABC	=	=	ABC	=
{System32}\Windows Sandbox.exe	3	4	0d, 0h, 06m, 41s	2019-12-20 14:44:21

Figure 7. UserAssist record of Sandbox execution

SnapshotSavedState.vmrs and the path to the saved file. The ".vmrs" file stores the running state of the virtual machine, in this case, the Sandbox environment, with a size of approximately 200 megabytes. The SnapshotSavedState.vmrs is saved at location "C:\ProgramData\Microsoft\Windows\ Containers\BaseImages\<*GUID*>\SnapShot" which is created after the first run of the feature. The next record with the same event ID is related to creation and saving of template file with, in this case, GUID "27c76368-c63e-44aa-8054-d92da287e19a". The event with ID 2014 record details of the creation of a virtual machine, in this case, with GUID "c316db96- f9d1-4778-a88a-bf19dfd7ec7f". The event logs with ID 2500 which contain parameter "CommandLine":"reg add" describe new registry keys and values that were added to the Sandbox environment. When the Sandbox environment is shut down the event with ID 2003 is recorded with notice of computer system termination.

The second log set to examine is Hyper-V-Worker related to the worker process responsible for running the virtual machine (VM) which will contain startup and shutdown logs of the virtual machine [15]. Hyper-V-Worker log set in table 1. can help an examiner to establish a timeline of Sandbox usage which can be compared with timestamps related to suspicious activities.

The third event log set is VHDMP which is related to the operations on a virtual hard disk (VHD). This log set can contain information about the creation of virtual hard drives that Sandbox feature uses as well as their deletion as shown in table 2.

The last test with event logs was closing the Windows Sandbox feature to determine exactly which logs are recorded in the shutdown moment. Once again, all recorded events before closing the feature are cleared. The Hyper-V-Compute recorded 11 logs from which ID 2004 recorded when the Sandbox environment was paused followed by ID 2003; terminate compute system. The Hyper-V-Worker recorded 2 events; event with ID 18516 which recorded pausing of the virtual machine followed by ID 18502 which recorded powering off the virtual machine. The VHDMP recorded an unmounting event, with ID 18, of the virtual hard disk related to Sandbox.

TABLE 1. HYPER-V-WORKER EVENT LOG SET

Event ID	Description
12148	VM started successfully
18500	VM started successfully
18601	VM successfully booted an operating system
18516	VM was paused
18502	VM was turned off
18596	VM was restored successfully
18516	VM was paused
12597	VM connected to virtual network

TABLE 2. VHDMP EVENT LOG SET

Event ID	Description
1	VHD has come online
2	VHD has been removed
12	Handle for virtual disk created successfully
15	Virtual disk object created
17	Virtual disk has been surfaced

E. Remote Desktop Protocol Cache

The Remote Desktop Protocol (RDP) developed by Microsoft provides the user with display and input capabilities when connecting to another computer over a network [16]. RDP is available in most versions of Microsoft Windows including Windows 10. It includes many different features like encryption, clipboard mapping, print redirection, remote control, etc. From Windows 2000, Microsoft introduced caching function to improve performance by saving sections of the screen that rarely change in the form of small bitmap images [16]. The bitmap images are saved in a file with .bin extension with file name format "Cache000X.bin". Each file can store up to 100 megabytes of images, with a size of 64x64 pixels. When a single file reaches 100 megabytes the new file will be generated with incremented last number in the filename. Location of the cache files is "C:\Users\<user>\AppData\Local\Microsoft\Terminal Server Client\Cache\" [16].

While testing the Sandbox feature the FileActivityWatch tool developed by Nirsoft was used to monitor read, write, and delete operations related to files on the Windows system. This tool, for every operation, displays read and write bytes, the number of the read, write and delete tasks, timestamps of first and last read and write operations as well as name and ID of the process associated with the file operation [17]. Throughout testing before the mentioned artefacts, it was noted in the FileActivityWatch tool that WindowsSandbox.exe process writes approximately 100 megabytes of data to before mentioned RDP cache file "Cache0000.bin" on the testing environment, as shown in Fig. 8. To test if the RDP cache records portions of the Sandbox' screen, the desktop background image was changed on the Sandbox environment to custom red image with text "Sandbox Screen" as shown in Fig. 9.

FTK Imager was used to extract the "Cache0000.bin" file while BMC-Tool written in Python was used to parse and extract bitmap images from the "Cache0000.bin" file. The BMC-Tools is a free tool developed by the French National Cybersecurity Agency (ANSSI) which can be



Figure 8. Windows Sandbox related to saved RDP



Figure 9. Sandbox environment background

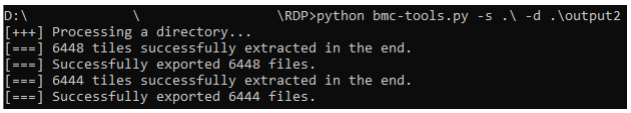


Figure 10. BMC-tools results on RDP cache

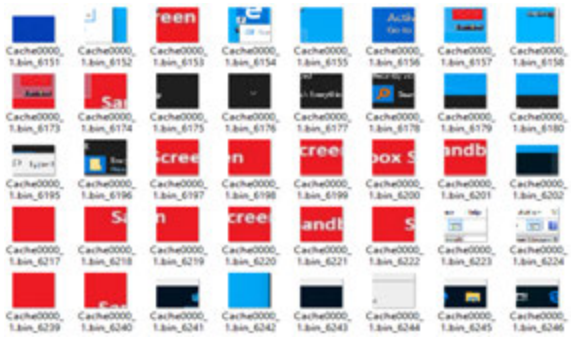


Figure 11. Sandbox environment background

downloaded from GitHub [18]. To run the tool examiner needs to specify the RDP cache file or directory to process and specify the destination directory. The example of command to run the tool is "python bmc-tools.py -s <*source file*> -d <*destination folder*>" [18].

After selecting the source and destination path the tool can be run and once when the parsing and extraction are completed the tool will report the number of extracted bitmap images also known as tiles, as shown in Fig. 10.

As a result, 6448 bitmap images were extracted from the RDP cache of the test environment. While examining the content of the images, it is noticed that the RDP cache recorded Sandbox environment red background with text, as seen in Fig. 11.

Slight issue with these bitmap images is that they are not recorded in order [19]. Some of these images can be stored partly in order or groups that examiner can easily sort manually. Interpretation of bitmap images is easy as they represent portions of the screen on the Sandbox environment, but the examiner should be aware of a risk of misinterpretation [19]. While bitmap images from RDP cache can contain program icons, dates and times, command prompt, web browsing pages, run programs, PowerShell commands, and much more other useful information, they can also contain images opened on the Sandbox environment. Because of that, the examiner must examine RDP cache images with caution as the user can open an image on the web that contains, for example, PowerShell command, which was never run.

From the above findings, it is concluded that this artifact has great value, as at this moment, it is the only record that can show the examiner how suspect interacted with the Sandbox environment.

VI. Conclusion

Windows artifacts like prefetch, UserAssist together with changes in the folder structure and event logs left behind as a result of user interaction with Windows Sandbox feature have great value when examiners want to determine if the feature was enabled and run by a specific user in specific time range, but they don't record any information of user interactions in the Sandbox environment. It is confirmed throughout testing that the Sandbox environment is indeed temporary and can be used as an anti-forensic method for hiding suspicious activities and files as the virtual hard drive and files on it were deleted without possibility to recover them with forensic tools. Unlike previously mentioned artifacts, the RDP cache contains valuable bitmap images of the screen that can give examiners information on how the user interacted with the Sandbox environment. Furthermore, examiners need to properly understand RDP cache as a misinterpretation of bitmap images can lead to the wrong conclusion which can result in evidence disposal or even wrong accusations toward the suspect. To gain better understanding of all related artifacts to the Sandbox feature, the manual analysis with validated tools is recommended and should be done personally by examiners multiple times in different environments.

References

[1] M Reith, C Carr, G Gunsch, An examination of digital forensic models, (2020, January), Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Digital_forensics

[2] Peter Sommer, The future for the policing of cybercrime, (2020, January), Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Digital_forensics

[3] E. Casey, Digital Evidence and Computer Crime, 3rd ed., Elsevier, 2011, pp. 3.

[4] VMware Workstation Pro, (2019, December), Retrieved from: https://www.vmware.com/products/workstation-pro.html

[5] Download Windows 10, (2019, December), Retrieved from: https://www.microsoft.com/en-gb/software-download/windows10

[6] H. Pulapaka, Windows Sandbox, (2019, December), Retrieved from: https://techcommunity.microsoft.com/t5/Windows-Kernel-Internals/Windows-Sandbox/ba-p/301849

[7] K. Hausknecht, S. Gruičić, Anti-computer forensics tools, (2019, December), https://ieeexplore.ieee.org/document/7973612

[8] Wikipedia, (2019, December), Program Files, Retrieved from: https://en.wikipedia.org/wiki/Program_Files

[9] Windows registry information for advanced users, (2020, January), Retrieved from: https://support.microsoft.com/en-us/help/256986/windows-registry-information-for-advanced-users

[10] O. Skulkin, (2020, January), Hunting For Attackers' Tactics And Techniques With Prefetch Files, https://articles.forensicfocus.com/2019/11/21/hunting-for-attackers-tactics-and-techniques-with-prefetch-files/

[11] B. Matt, (2020, January), Tooling Thursday: PECmd, https://medium.com/@bromiley/tooling-thursday-pecmd-83d1d6d6346b

[12] E. Zimmerman, (2020, January), Eric Zimmerman's tools https://ericzimmerman.github.io/#!index.md

[13] D. Stevens, Training, (2020, January), UserAssist, https://blog.didierstevens.com/programs/userassist/

[14] Introduction to Hyper-V on Windows 10, (2020, January), https://docs.microsoft.com/en-us/virtualization/hyper-v-on-windows/quick-start/enable-hyper-v

[15] B. Lee, Troubleshooting Hyper-V with Event Logs, (2020, January), https://www.vembu.com/blog/hyper-v-event-logs-troubleshooting/

[16] Habib, (2020, January), Digital Forensics on RDP Cache, https://cbtgeeks.com/2018/05/22/digital-forensics-on-rdp-cache/

[17] Nirsoft, (2020, January), FileActivityWatch v1.45, https://www.nirsoft.net/utils/file_activity_watch.html

[18] ANSSI, (2020, January), RDP Bitmap Cache parser, https://github.com/ANSSI-FR/bmc-tools

[19] Opentext, (2020, January), RDP Cached Bitmap Extractor, https://www.guidancesoftware.com/app/RDP-Cached-Bitmap-Extractor

Linux Forensic Triage: Overview of Process and Tools

A. Anđelković[1], K. Hausknecht[1], G. Sirovatka[2]
[1] INsig2 d.o.o., Zagreb, Croatia
[2] Zagreb University of Applied Sciences, Zagreb, Croatia
Anja.Andjelkovic@insig2.com, Kresimir.Hausknecht@insig2.com, Goran.Sirovatka@tvz.hr

Abstract - **Digital forensics dates back into the 1980s, but the importance of Linux forensics was not taken into place until recently. Linux forensics is a distinctive world compared to example Microsoft Windows forensics. Although it is commonly used as a name for the entire operating system, Linux is just the name of the kernel, a piece of software that handles interactions between the hardware and end-user applications. Its popularity has not reached the popularity of the Windows operating system, therefore, without many reliable tools on the market, it represents a bigger challenge for digital forensics investigators.**

Digital triage is the process in which an investigator collects, assembles, analyzes, and prioritizes digital evidence from a crime. Since there are not many available tools on the market for performing Linux triage, the most important part is to understand the tool and its capabilities in order to know which one to use for a certain situation.

This paper will describe how the Linux system is structured, what its architecture contains, how should one correctly approach and acquire the system, and how to understand the tools and results they provide.

Keywords – Linux forensics; digital triage; Linux architecture

I. INTRODUCTION

Digital forensic examiners use various scripts and tools that are available on the market for performing Linux forensic triage. The usage of the Linux system is constantly growing which increases the need for Linux forensic investigators, as well as the creation of proper tools. Throughout this paper, the value of manually acquiring Linux artifacts will be shown, as well as how to understand the tools and scripts that are available on the market. Before starting to use any tool, one must understand how Linux is structured in order to know where to search for useful data and to understand the output that the tool provides.

The paper is divided into 7 sections. Section 2 and 3 provide a brief overview of the Linux system and its architecture. Within section 4, the introduction to digital forensic triage is given, along with the different types and characteristics of forensic triage tools. Section 5 describes important files for Linux triage and commands for a manual search for data of forensic value, while section 6 provides detailed information on a few scripts and tools for Linux forensic triage and memory capture. Section 7 concludes the paper.

II. LINUX OVERVIEW

Linux is a family of open-source operating systems based on the Linux kernel. It was released on September 17, 1991, by the Finnish software engineer, Linus Torvalds. The idea for the creation of Linux derived from and was based on the UNIX operating system [1].

It became common to use the naming ‘Linux’ when referring to the entire operating system, but in fact, Linux is the name of the kernel. A kernel is a core of the computer system that loads first and manages interactions between the hardware and applications. The correct expression when wanting to refer to the entire operating system would be ‘Linux distribution’ or ‘Linux flavor’. This part is built on top of the Linux kernel and, depending on the distribution, it can include already installed applications or have the barest minimum [2].

Different Linux distributions offer different ways of being configured. Adjusting the system for personal needs is much more flexible than on other operating systems, but also requires higher technical knowledge. This is why Linux is often thought of as being an operating system for coders and programmers. However, over the years, Linux became more appealing to general users and therefore its usage increased [3].

As mentioned, the main difference among different Linux distributions is the kernel on which the system is built on. Almost six hundred Linux distributions exist nowadays, and the number will most likely keep increasing. Linux can be configured for different kinds of interests and needs. It can also be compiled and assembled from scratch. Some distributions include numerous pre-installed basic applications, while others have the simplest minimum. There are also distributions created for a specific purpose, for example, Digital Forensic and Forensics Toolkit (DEFT) distribution which is created for conducting digital forensic investigations with various pre-installed forensic tools. Some of the most popular widely used distributions are Debian, Fedora, openSUSE, Gentoo, Linux Mint, Ubuntu, and CentOS [3].

Its reliability evolved from being more resistant to viruses, crashes, slow operation, and expensive repairs than other operating systems. Therefore, most of the

Internet, all of the world's top supercomputers, and the world's stock exchanges are run on the Linux system. Also, some of the most popular platforms, like Android and iOS, are established on the Linux operating system [4].

III. LINUX ARCHITECTURE

Linux system architecture contains four layers, as it is shown in Figure 1. The central layer consists of all the physical devices that are connected to the system. The central layer is called hardware and is contained of a hard disk drive, Random Access Memory, Central Processing Unit, motherboard, etc. Secondly goes the Kernel, which is the core component that manages interactions with the layers above and below it. It is responsible for all important activities. The last two layers are Shell and Applications. Shell is the bridge between the user and the Kernel. It takes inputs from users and forwards them to the upper layer. Likewise, it takes outputs from Kernel and sends results back to the Shell, for users to be seen. The last layer, Applications, is the closest one to users. Applications utilize programs that run on Shell and provide the user with the functionalities of the system [5].

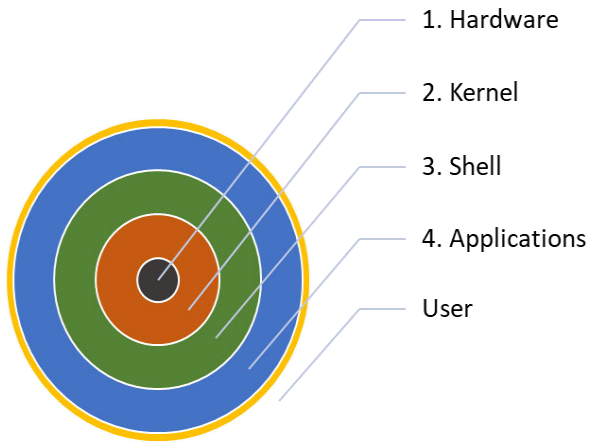


Figure 1 Linux architecture overview

A. *Linux structure*

The Linux File Hierarchy Structure or the Filesystem Hierarchy Standard (FHS) defines the structure and content of the directories in Unix-like operating systems. The FHS project started in 1993, and the purpose was to agree on how directories should be systematized, and which files should be stored where so that distributions could have a single reference point from which to work [6].

In Linux, everything starts from the *root* directory, which can be seen in Figure 2. Root directory would be equivalent to the Microsoft Windows C:\ folder. Every file and directory start from there and only root users can write under this directory. Root users are the ones with the administrative privileges, such as changing the ownership over the files. In other words, root users have the highest level of access rights on the system [7].

Some of the most common top-level directories, which can contain forensically sound evidence are listed and described in Table 1.

Table 1 Most common top-level directories

Directory	Description
/bin	Essential system binaries
/etc	System configuration files
/home	User directories
/opt	Optional or third-party application
/tmp	Temporary files (cleared on reboot)
/usr	User related software
/var	Variable data, log files

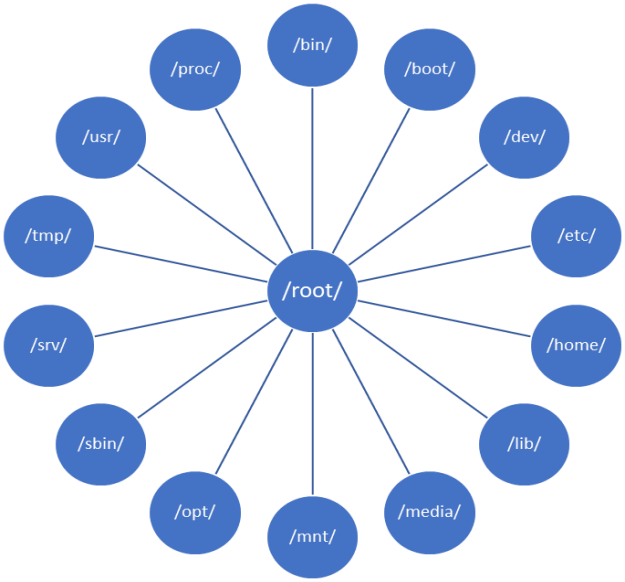


Figure 2 Linux directory structure

B. *Journaling file system*

To investigate the Linux system, a forensic investigator should know the system's file structure. Being able to understand how a certain file system handles files can be helpful when recovering evidence such as deleted files and file activity.

Most of the Linux systems use a journaling file system that keeps track of system changes which are still not committed to the file system. Recorded changes are written down in a data structure, known as "journal". In case a power failure occurs and the system unexpectedly shuts down, such file systems can be brought back more quickly with less chance of becoming corrupted [8].

The most commonly used journaling file system for Linux is the third extended filesystem (ext3 or ext3fs), which was added in 2020 to the kernel from version 2.4.15. The newest journaling file system for Linux is ext4 (fourth extended filesystem). It is the default file system for many Linux distributions and can support volumes with sizes up to one exbibyte [9].

IV. DIGITAL FORENSIC TRIAGE

The term "triage" comes from the field of medicine, in which medical personnel determine the priority of patients due to their condition. Triage in digital forensics can be the crucial step as its purpose is to quickly identify data that are likely to contain the evidential value. During the

process of digital forensic triage, investigators collect, extract, and analyze the acquired digital evidence. By using triage, investigators can decide not to perform a full disk image but only capture what is necessary to prove or disprove investigation goals [10]. Though the purpose of triage is to quickly identify data of evidential value, these tools have certain limitations. Issues that could occur when working with triage tools are that they will not bypass the encryption or passwords and most of them need administrator privileges [11].

The digital triage comes in two forms, based on the type of data that is being investigated and methods used – live triage and post-mortem triage. Traditional forensic investigations were focused on analyzing data from a powered-off computer. Investigators would take the disk from a powered-off computer, create its duplicate and start the analysis. Since certain data gets lost when the device gets powered-off, investigators started to analyze the live system. When dealing with the live computer, running processes, notifications, recent email messages, recently visited web sites, passwords, etc. can be captured [11].

A. Forensic triage tools characteristics

Not every tool is appropriate for every situation, nor will every tool produce all of the needed results. The most important part is to understand the tool and its capabilities in order to understand the results they produce. Triage tools and tools' updates must be validated before being used. Tool validation is performed in order to preserve the integrity of evidence and to state the correctness of the results. Tools should not modify data in any way. Investigators must not only know what the tools output, but also how they do it. That way they will also be able to describe and demonstrate how these tools are relevant to their investigation and how they changed the inspected system without setting their investigation in danger. The precondition that every triage tool should have is the capability to collect data in a reasonably short time without changing the data [12].

Tools can be divided into three categories – open-source, free, and commercial tools. Open-source tools have a freely available program code and support is based on the community. Commercial tools offer plenty of options and customizations, and have excellent support, but are usually extremely expensive. Free tools are a subcategory of commercial tools. They are made free to use by a company that is providing support for them. In comparison to standard commercial tools, free tools do not have full functionality or they are free to use for a certain period of time As Linux is a family of open-source operating systems, so are most of the tools used for them [12].

All tools and scripts that will be presented throughout this paper are open-source and free for download.

V. LINUX TRIAGE COMMANDS AND FILES

Tools used for Linux triage are usually built out of Linux commands. Investigators should know these commands to understand how these tools perform the work and collect information. It is not always enough just to run the tool and wait for it to produce gathered information. In the following chapters, some of these commands will be mentioned, along with their importance for the investigation.

A. Gathering user account information

When analyzing the system, volatile data should be captured first. Volatile data is the type of data that would be lost if the system were powered off. This type of data is stored in a computer's temporary memory while it is running. Computer's memory is investigated since it can lead to identifying attacks and malicious behaviors, gather information on recently performed actions (visited web sites, received/sent email messages, etc.), bypass the encryption, collect password information, and many more.

Collection of the user account information is of great importance while performing forensic investigation. Best way to gather information from the Linux system is by using the terminal in which the commands are inputted. The command "w", when typed into the terminal, will display users that are logged on the system. This command will also display what is the user doing on the system and when the user logged into the system [13].

Besides these commands, there are important files under the /etc directory that should be investigated. All this information could be helpful when building an event timeline. The /etc directory contains application's configuration files, startup, shutdown, start and stop script for every individual program. A configuration file is a local file used to manage the operation of a program – it must be static and cannot be an executable file [14]. Inside this directory are many files of interest for digital forensic investigators when performing Linux triage. System configuration, like hostname, time zone, operating system information, etc., could be found. For reading each of these files, the command "cat", along with the file name, is used. Knowing the hostname can be useful in the further investigation to connect other information with the specific action or user. Hostname file can be opened from the terminal with the command "cat /etc/hostname" [15]. To adjust the event timeline, useful information is to know in which time zone is the system set. This information can be found with the command "cat /etc/timezone" [15].

Another two important files are /etc/passwd and /etc/shadow. These files contain information on users and their passwords. The /etc/passwd includes basic information about each user account on the system, including the root user who has full administrative rights, system service accounts, and actual users. [16].

The /etc/shadow file stores the actual password in an encrypted format for the user's account with the additional properties related to the user's password. In other words, it stores secure user account information. [17].

B. Network information

Part of digital triage is to check the network information. Network information includes the local network information, a list of ports opened to the outside that could be used to compromise the network, related programs to those ports, and many more.

In order to check the IP address that is assigned to the system, "ifconfig -a" command is used. This command lists out the number of received and transmitted packets and assigned IP address [18].

During this process, it is also important to review each of the network connections and open sockets. This information can be gathered with the "netstat" command. To use this command, the user must be in root mode in order to receive the information. This command will list out all the active internet connections which can be very useful for the investigators. There are additional options with this command. For example, with the "netstat -tulpn" command, only servers will be listed, but when using the "netstat -natp" command, established connections will also be listed. It will give out the protocol that is used, the number of received and sent packets, local and foreign address, the state, PID (Process Identifier) number and program name (e.g. Firefox) [18].

C. Running programs and processes

Running programs and processes can point to a certain activity that can prove or dispute certain actions. The command "top" will show all running processes, while command "ps -aux" will display all processes [19]. Listing out the processes can point to certain activities performed by the user or lead to the suspicious activity running in the background. To check for opened files on a Linux system, the "lsof" command will provide information on it. In its output, this command returns information about files that are opened by processes. Usage of this command can be very useful in malware investigations since it can discover what was used by the user. This command can be modified with different options to obtain a different output. To display IPv4 or IPv6 files, either "lsof -i -4" or "lsof -i -6" commands are used. Also, it can be specified to display processes that are belonging to a certain user. For that, the username must be included in the command, in a way that "lsof -u [USERNAME]" command is used [20].

Most of Linux triage tools are based on these commands but can also be used to create own scripts and personalize them according to what needs to be acquired.

VI. LINUX TRIAGE SCRIPTS AND TOOLS

The main purpose of digital triage is to identify the most relevant artifacts. Two important characteristics that triage tools must have is that they are reliable and do the work fast. These requirements are imperative for building triage tools.

As mentioned before, useful information can often be found in the computer's Random Access Memory (RAM). Digital investigators can search for information simply by reviewing the readable text and performing keyword searches. Since physical memory in modern computers has increased and will continue to increase, such an approach is not efficient. That is why it is important to have specialized knowledge of data structure in memory and associated tools.

When talking about Linux triage, not many tools are available on the market, but those that are, showed justifying results. A great number of tools and scripts are open-source and available for download.

A. Acquire Volatile Memory for Linux (AVML)

When coming across a live computer, capturing memory is imperative. When talking about Linux, the traditional approach was to use "dd" and point to a device file such as /dev/mem or /dev/kmem. Modern Linux system restricted access to these virtual devices so the new approach was required. The current method and the most popular tool for capturing memory is Linux Memory Extractor (LiME), which is, in fact, Loadable Kernel Module (LKM). The problem with this method is it requires knowledge of the target distribution or kernel, which can sometimes present an issue. Also, it requires to be compiled for every kernel version separately. This is where Microsoft introduced its memory acquisition tool called AVML (Acquire Volatile Memory for Linux).

AVML tool is a volatile memory acquisition tool written in Rust programming language. Its advantage among other acquisition tools is its possibility to acquire memory without knowing the target operating system or even its kernel and does not require any additional libraries on the target machine. This is the issue with most other acquisition tools since they require a loadable kernel module to be built on the exact same kernel version as the target [21]. This tool also has certain limitations as it does not support every target distribution or kernel.

It is enough to build the tool once and place it on any storage device for further use. When wanting to capture a memory from a target machine, it can be done without any on-target compilation needed.

AVML acquires memory from three memory sources - /dev/crash, /proc/kcore, and /dev/mem. It uses the LiME output format, when not using optional compression.

B. IR Triage Toolkit script

IR Triage Toolkit is an open-source collection of scripts that can be used to create a toolkit for incident response and volatile data collection. It can be used for both Linux and Windows operating systems and downloaded for free from a GitHub webpage [22].

When the script is run, it dumps the content of the system's memory and outputs several other system's information in a text file. Additionally, it creates detailed timestamped logs of the given command and creates a SHA-256 checksum hash of the files saved to the storage directory [22].

After the toolkit is downloaded, it needs to be created in the desired directory. In order to do so, the bash script "create-toolkit" is started with the command "sudo bash create-toolkit", as it is shown in Figure 3. Following, a bash script named "run" will be generated inside the chosen directory. When running the "run" script, volatile data will be saved and ready for further analysis.

When looking at the background of the toolkit by opening its bash file, it is visible what the toolkit does. The toolkit first acquires a full memory dump by using the integrated LiME tool. Afterward, it collects various system information and outputs it into separate text files.

The output files contain information on the network, opened files, running processes, users, system, and mounted devices. At the end of the process, the toolkit creates SHA-256 checksum for all files.

Gathering information with IR Triage Toolkit is very fast with the well-structured output. It collects the main basic information on the system and users, and the bash script can be easily modified and adjusted for specific needs. The script is based on numerous Linux commands, some of them which were mentioned in the chapter before. The tool is small in size when downloaded but requires larger storage capacity for dumping the content – depending on the memory size of the target device.

```
@ubuntu:~/ir-toolkit$ sudo bash create-toolkit
Directory to save toolkit: ./create-toolkit /tool
```

Figure 3 Command for creating IR Triage Toolkit

C. *FastIR Collector Linux*

The tool FastIR Collector Linux is a script based on Linux commands which collects various artifacts on the live Linux system and can be downloaded for free from a GitHub webpage. There is also a separate script created for collecting artifacts from Windows operating systems. Some of the artifacts it collects are system information, last logins, connections, handles, user's data, and many more. Collected results are outputted in .csv and .txt files [23].

The tool is written in Python programming language and must be run as a root user with a "python" command, as it is shown in Figure 4. It will output the mentioned information in a separate directory.

Information collection is performed and outputted within seconds. This tool also collects information from /etc/passwd and /etc/shadow files. Its disadvantage, in comparison to IR Triage Toolkit, is that it does not dump the system's memory, and is harder to adjust the script's code for personal needs.

```
@ubuntu:~/Fastir$ sudo python fastIR.py
```

Figure 4 Running FastIR Collector python script

D. *IR Rescue*

IR Rescue consists of two scripts that collect numerous forensic data from both 32-bit and 64-bit Windows systems and Unix systems. The scripts are intended for incident response and forensic investigators for use at various stages in the analysis and investigation process. [24].

The script also comes with an executable for dumping memory, called "linpmem-2.1.post4" [24].

The file is composed of binary directives – true or false. By choosing between true and false options, data types to collect and advanced tools for running will or will not be performed. The configuration file can be modified in the way it is needed for a certain purpose or situation. For example, a memory dump is set to "false" by default, and therefore will not be performed unless otherwise is not defined. This may make a procedure a bit more complicated but also gives the examiners different possibilities to adjust the script for specific needs. Examples of setting options to "true" or "false" are shown in Figures 5 and 6. For example, if the web browser's data is needed to be collected, it will be set to state "true", as shown in Figure 5. On the other hand, if a complete memory dump is not needed, it can be set to false, as seen in Figure 6. After the configuration file is modified and the script is run, outputted data will be structured in directories, in either a textual (.txt) format or as a database (.sqlite). Further on, .sqlite files, which are created for web browsers, can be opened with the command "sqlitebrowser [FILENAME]".

```
# web
web-all=true
        # collect browser information
        # "mozilla", "firefox"
web-browers=true
```

Figure 5 Collecting web browser information with IR Rescue

```
# memory
memory-all=false
        # dump the RAM to AFF4 format
        # Pmem "linpmem-2.1.post4" (64-bit ELF)
memory-dump=false
```

Figure 6 Turning off memory dump creation in IR Rescue script

E. *Live Response Tool Collection*

Live Response Tool Collection is a set of tools for collecting data from a live system. It covers Windows, OS X, and Unix/Linux operating systems. It is created in a way that a non-technical person can easily use it. The built-in triage option collects volatile data, using a variety of scripts. The script saves data to a directory with the computer name and timestamp under the directory from which the script was run. It also generates text files containing the MD5 and SHA-256 hashes of every collected file and the full path to that file. There is also a log text file that contains each command that was run by the script and, if present, any error message from running a certain command.

The script should be run under administrative privileges as it allows larger data collection. When running the Linux bash script, the tool collects information and stores it under five subdirectories. Information is collected very fast and outputted in a well-

structured way, where each file is automatically named according to the content it contains.

"BasicInfo" subdirectory contains alternate data streams, last activity view, running processes, list of hidden files, etc. Under "NetworkInfo" subdirectory information about ARP tables, allowed/denied hosts, routing tables, Internet settings, and socket statistics are stored. "PersistanceMechanisms" subdirectory contains information related to persistence mechanisms on the system, including autoruns, loaded drives, and scheduled tasks. "UserInfo" subdirectory includes user's information, like logins, a listing of the users, and information on the current user [25].

The bash script can be adjusted for specific needs and purposes, and it is well-documented, giving the information on what each section does and where it is outputted.

VII. CONCLUSION

The consumption of the Linux operating system is expanding, especially among coders and programmers, but, unfortunately, it is also widely used for performing criminal activities. In order to investigate activities on a Linux system, forensic investigators must understand the system's architecture and file structure.

When approaching a live computer, it is imperative to capture volatile data. Once the system is powered off, volatile data becomes irretrievable. Linux distributions may differ in the kernel version on which they are built, making the memory capture harder since the tool must be built on the same kernel version. Microsoft company presented a tool for capturing the memory of a Linux system, called AVML. This tool distinct from others as it can acquire memory without knowing the kernel version, making the whole process easier.

Throughout this paper, various commands were described. Tools and scripts that are being used for performing Linux triage are based on these commands. Since there are various tools and scripts available on the market, it is not enough to just run them but to understand how they work and the results they produce. Few free-for-download scripts were described throughout this paper. Their support is mainly based on the community, making the tool validation process even more important. The best practice is to use multiple different tools for the same action, as not every tool will produce the same amount of results, nor will one tool be appropriate for every situation. Sometimes the used tool can only be an investigator's matter of preference. As can be seen in Section VI., all of these scripts are easy for use. This is one more reason for forensic investigators to be familiar with the Linux commands and file structure so that they can adjust the existing scripts as they need them for a particular situation. Knowing the file structure leads to knowing how to navigate yourself through files that could have forensic value.

REFERENCES

[1] "Linux", (2019, December), Retrieved from: https://en.wikipedia.org/wiki/Linux

[2] "What is Linux", (2019, December), Retrieved from: https://opensource.com/resources/linux

[3] A. Williams and B. Turner, "Best Linux distros of 2019: for beginners and advanced users", (2019, December), Retrieved from: https://www.techradar.com/best/best-linux-distros

[4] "What is Linux", (2019, December), Retrieved from: https://www.linux.com/what-is-linux/

[5] "Linux Architecture", (2019, December), Retrieved from: https://tecadmin.net/tutorial/linux/linux-architecture

[6] "Filesystem Hierarchy Standard", (2019, December), Retrieved from: https://en.wikipedia.org/wiki/Filesystem_Hierarchy_Standard

[7] "Root User", (2020, January), Retrieved from: https://www.ssh.com/iam/user/root

[8] Sayak Boral, "What is a Journaling File System?", (2019, December), Retrieved from: https://www.maketecheasier.com/journaling-in-file-system

[9] Ramesh Natarajan, "Linux File Systems: Ext2 vs Ext3 vs Ext4", (2019, December), Retrieved from: https://www.thegeekstuff.com/2011/05/ext2-ext3-ext4/

[10] V. Jusas, E. Gahramanov, and D. Birvinsksas, "Methods and Tools of Digital Triage in Forensic Context: Survey and Future Directions", (2020, January), Retrieved from: https://www.researchgate.net/publication/315928408_Methods_and_Tools_of_Digital_Triage_in_Forensic_Context_Survey_and_Future_Directions

[11] INsig2, "Basic digital forensics for IT professionals", Retrieved from: https://insig2-and-zyberglobal.learnworlds.com/author/course?courseid=dfir-it

[12] INsig2, "Fundamentals of digital forensics for lawyers and judges", Retrieved from: https://insig2-and-zyberglobal.learnworlds.com/author/course?courseid=fundamentals-of-dfir

[13] Aaron Kili, "11 ways to find user account info and login details in Linux", (2020, January), Retrieved from: https://www.tecmint.com/find-user-account-info-and-login-details-in-linux/

[14] Linux Foundation, "Chapter 3. The Root Filesystem", (2019, December), Retrieved from: https://refspecs.linuxfoundation.org/FHS_3.0/fhs/ch03s07.html

[15] Tho Le, "Linux Forensics – Some useful artifacts", (2020, January), Retrieved from: https://medium.com/@tho.le/linux-forensics-some-useful-artifacts-74497dca1ab2

[16] V. Gite, "Understanding /etc/passwd File Format", (2019, December) Retrieved from: https://www.cyberciti.biz/faq/understanding-etcpasswd-file-format/

[17] V. Gite, "Understanding /etc/shadow file", (2019, December), Retrieved from: https://www.cyberciti.biz/faq/understanding-etcshadow-file/

[18] Chandan Kumar, "10 Useful Linux Networking Commands", (2020, January), Retrieved from: https://geekflare.com/linux-networking-commands/

[19] Chris Hoffman, "How to manage processes from the Linux terminal", (2020, January), Retrieved from: https://www.howtogeek.com/107217/how-to-manage-processes-from-the-linux-terminal-10-commands-you-need-to-know/

[20] Narad Shrestha, "10 losf command examples in Linux", (2020, January), Retrieved from: https://www.tecmint.com/10-lsof-command-examples-in-linux/

[21] Microsoft, "AVML (Acquire Volatile Memory for Linux)", (2020, January), Retrieved from: https://github.com/microsoft/avml

[22] R. Shipp, "ir-triage-toolkit", (2020, January), Retrieved from: https://github.com/rshipp/ir-triage-toolkit

[23] SekoiaLab, "FastIR Collector Linux", (2020, January), Retrieved from: https://github.com/SekoiaLab/Fastir_Collector_Linux

[24] D. Fernandes, "ir-rescue", (2020, January), Retrieved from: https://github.com/diogo-fernan/ir-rescue

[25] B. Moran, "Live Response Collection Overview", (2020, January), Retrieved from: https://www.slideshare.net/BriMorLabs/live-response-collection-overview

Overview of Mac system security and its impact on digital forensics process

D. Sladović [1], D. Topolčić [1] and D. Delija [2]
[1] INsig2 d.o.o., Zagreb, Croatia
[2] Zagreb university of applied science, Zagreb, Croatia
Danijel.Sladovic@insig2.com, Davorka.Topolcic@insig2.com, damir.delija@tvz.hr

Abstract **-** Nowadays there are 3 main operating systems used, and Mac OS is one of them. Until now Apple published many iterations of their operating system and with that introduced many new features that are related to system security. Even though security-related changes go unnoticed, in the world of digital forensics this presents a challenge. Today encryption can be implemented on both, hardware and software level, which can make imaging Mac OS difficult. Besides, security which is meant to protect, user data is also used by criminals to restrict access to their computers. **This paper will focus on the differences and problems that occur while creating a forensic image and extracting data from Mac OS. On top of that this paper will depict the impact devices equipped with "T1" or "T2" security chip have on digital forensic process and remediation methods.**

Keywords - Mac OS forensics; forensic image; data extraction; T1; T2; encryption

I. INTRODUCTION

Through this paper, the reader will be introduced to the field of Mac OS forensics and the problems that a forensic investigator can encounter. Another focal point will be the comparison of Apple's "T1" and "T2" security chips, their features and benefits for the users, but also what impediments can an investigator encounter while analyzing devices having these chips. The next part will show what the investigator must do if he encounters an unlocked "Apple" device to preserve as much digital evidence as possible. Few software solutions will be tested to see what are their limitations regarding the bypassing of Apple's security features and what can an investigator do to recover or image a disk if an Apple device (i.e. Macintosh notebook) is protected or locked in different ways. Finally this paper will be summed up with a conclusion according to research and testing of available tools.

II. MAC OS FORENSICS

Digital forensics investigators traditionally deal with Windows machines, but, since the increase of Apple's popularity, their devices can be seen everywhere, from regular users, enthusiasts to corporate use for music production, photography, video editing, web development, and many more. These are the reasons why every investigator today must have the core skills to analyze Apple devices, or at least be knowledgeable on how to do the triage. A forensic investigator will have to follow a protocol during the examination and acquisition process because every operating system is different, artifacts are located in different locations on the hard drive and all these various artefacts can be useful during an investigation [1]. Important groups of forensic artifacts found on Mac OS are:

- System artifacts – consist of records related to the system configuration, OS version, Time zone, Language, MAC address, Start-Up folders, etc.
- User Profile – records related to user settings, Keychain, Recent folders, DOCK (persistent apps), Safari browsing history, Apple Mail, USB devices
- Logs – System logs, Network logs, User logs, etc.

Some important steps before seizing or acquiring data from MAC OS devices are:

- get the administrators/users or firmware password, File Vault password, Recovery key, iCloud credentials, Apple Id and password (if possible) in case the device is locked - in some cases that will be the only possible way to create a disk image;
- know how to analyze and acquire data from Hierarchical File System (HFS+) and Apple File System (APFS);
- it is important to find out what the allocation block size is because if the block sizes are not identical data won't be copied properly which is the main reason for errors appearing during the acquisition (destination drives for the acquisition must be block size aware);
- it is important to create a logical image of the disk while the machine is unlocked and powered on, if FileVault is enabled;
- know how to recognize a Fusion disk is important, because if both disks are not retrieved and imaged data will not be recovered;
- if all disks from a RAID are not found, data can't be recovered this is why it is important to know how to recognize the use of RAID;
- recover the original charger of the device;
- if the device is unlocked or turned on, it is important to remember to turn off "Secure Boot" and enable booting from external devices on devices that have T2 chips [1,2,3].

III. MAC OS SECURITY

From a forensic standpoint, Apple has built their reputation around their encryption standard which can obstruct a forensic investigation. Apple devices are very powerful when it comes to destroying data or making them unrecoverable. This is because of their encryption on the newer machines which is from the start turned on because of the T2 chip. Another point regarding the Mac OS encryption (FileVault) is that in case of remote

wiping only the decryption keys are deleted which leaves all the data scrambled and unrecoverable.

A. T1 security chip

Apple T1 chip, model number "APL1023" is an ARMv7 System on a Chip (SoC) that was introduced in 2016 and its main purpose was to handle the processing for the "Touch ID" sensor in the first MacBook Pro that was equipped with the Touch Bar. As previously mentioned, the T1 chip is based on 32-bit ARMv7 SoC from Apple which is running "bridgeOS" variant of "watchOS". The T1 chip was also used to lock sensitive components like the built-in microphones and cameras. It was also used for the "System Management Controller" (SMC) which is responsible for heat and power management, battery charging and the sleep and wake up functionality of the Mac OS. T1 is also used to check if MacOS is running on genuine Apple hardware.

T1 has a built-in "Secure Enclave" that stores the user's fingerprint or "Touch ID". It even protects security keys against malware attacks because it is designed to prevent brute force attacks on the chip and its secure enclave [4].

B. T2 security chip

Apple T2 chip, model number "APL1027" was introduced in October 2018. The T2 chip is made to act as a co-processor, it is based on the Apple A10 processor which is a 64-bit ARMv8 processor that can be found in the iPhone 7 and 7 Plus. The difference in the A10 and T2 processors is that the T2 has only one T801x core. This processor, as the T1 has, a separate Secure Enclave Processor (SEP). The T2 processor is used to implement the" Secure Boot" feature. Since the T2 chip has better performance than the T1 chip, it also handles more tasks such as early boot tasks. It is securing data storage at-rest, it encrypts data on the SSD using dedicated AES hardware that has a 256-bit key tied to a unique identifier in the chip itself, which does not affect the SSD's performance. To make the data on the Mac OS more secure in cases when the attacker has physical access to the machine Apple placed the T2 security chip between the CPU and the storage, CPU no longer has direct access to the data through the PCIe/NVMe path which ultimately makes imaging the disks harder for forensic investigators. Flash storage is placed on an isolated bus, and the user data can only be accessed by the "Direct Memory Access" (DMA) crypto engine [4]. Because of the T2 chip encryption and physical placement, when the "FileVault" is enabled, the encryption is done fast and by the T2 chip enabling the user to normally use the mac OS during the encryption process. The decryption key is a combination of the Macintosh's unique hardware ID and user-provided passphrase, which adds another layer of encryption on top of the regular T2 chips encryption.
Additional features of the T2 chip are:

- encryption is on by default;
- it prevents booting from external device unless the security options are not changed (changing security options requires administrator password);
- the live-boot drive cannot be imaged;
- a Macintosh with T2 chip can be booted into "Target disk mode" and connected to another Macintosh computer that is booted to acquisition software (MacQuisition, Recon Imager);
- Macintosh with T2 chip can be imaged by selecting the physical drive or the APFS container;
- if File Vault is off, the T2 Macintosh logical data can be browsed through.

In case of repairs, the Mac OS goes through a hardware check to ensure genuine parts have been used for replacement [5].

C. Firmware password

The firmware password is a type of hardware-level security that prevents people from resetting the user's password or reinstalling OS X without authenticating themselves first. Ultimately this means that digital forensic investigators will not be able to access any features that will allow them to boot from external bootable media and create a disk image of the Mac OS.

Features that are protected by the firmware password:

- blocked function to start up from optical disk (use "C" key);
- blocked function to start up from the diagnostic volume of the install DVD (use "D" key);
- blocked the function to start up from a NetBoot server (use "N" key);
- blocked function to start up from Target Disk Mode (only for machines that offer this feature), so-called "T" key;
- blocked function to start in Verbose mode (activated by pressing the "command + V" keys during startup);
- blocked function to start in single-user mode (Command+S);
- block the reset of Parmeter RAM (PRAM) which is accessed using the Command+option+P+R key combination;
- a password is required to execute commands in Open Firmware mode (this mode is entered with the Command+option+O+F key combination);
- blocks the Safe Boot mode function by prompting for a password (the mode is accessed by pressing the "Option" key during startup) [6].

The firmware password makes the process of disk imaging even harder for the examiner. In case when the firmware password cannot be obtained from the suspect, it can be impossible for even the most skilled forensic investigator to image the suspect's Macintosh computer. This is because Mac OS will not allow any boot from any other device, except the one containing the OS, even with "Secure boot" and "External boot" protection turned off.

The BIOS chip that containes the firmware password can in some cases be replaced with another chip that is specifically programmed to work on the machine. It is also possible to remove the password using a method called "SPEG programming". It allows the programmer to remove the password from the BIOS chip without desoldering it. But these two methods are not forensically

sound because, in both cases, the Mac OS is reset to its factory settings and all data is lost [8].

Another method of possibly removing or bypassing the firmware password is by using the "Matt Card" to bypass the built-in EFI ROM permanently. The "Matt Card" is an alternative to removing and exchanging the EFI ROM chip with a preprogrammed one. This "card" can be plugged into a connector on the motherboard and, as long as it is connected, the firmware password will not be needed. To use the "Matt Card" the investigator must turn off the Macintosh computer, plug in the correct "Matt Card" for the motherboard in the locked Macintosh computer and power on the computer. This hardware is used by Macintosh repair businesses to gain access to Macintosh devices but can be in some cases useful while imaging the devices [16].

D. Secure Enclave

As mentioned earlier, the secure enclave is a hardware-based key manager that is isolated from the main CPU which creates an additional security layer. The "Secure Enclave" contains the keychain decryption key and can be accessed only by authorized applications. The benefit of using "Secure Enclave" is that, after the user stores a private key to the "Secure Enclave", the handling of the key is done only by the "Secure Enclave" which ultimately makes it hard for a key to become compromised. When applications are handled by the "Secure Enclave", the user receives only the operation outputs such as cryptographic signature verification outcome or encrypted data. The only downside of this whole process is that by using "Secure Enclave" to handle these operations the stored password have to be decrypted and saved in plain text to the system memory.

The benefits of the "Secure Enclave" are balanced because of few restrictions. The "Secure Enclave":

- is a hardware feature of the A7 series processor. It is only supported by iOS devices that have the A-series processor and Macintosh computers equipped with the Touch bar required for the Touch ID feature;
- it stores only 256-bit elliptic curve private keys which can be used to create and verify cryptographic signatures or for elliptic curve Diffie-Hellman key exchange;
- preexisting keys cannot be imported. The keys used by the "Secure Enclave" must be created inside it. This is the fundamental principle that makes the "Secure Enclave" secure from attacks.

This is another example of how the Apple security features make a locked Macintosh computer hard to investigate or to retrieve evidence from, especially when the subject of the investigation requires authentication or decryption keys [7].

E. FileVault

FileVault is a disk encryption program available for Mac OS X 10.3. and later. It performs on-the-fly encryption on volumes of a Macintosh computer. When FileVault was first introduced, it only encrypted the user's "Home" directory, but it had a lot of implementation problems and also had a lot of bugs. It was fully redesigned and published in 2011 with macOS X 10.7 ("Lion"). FileVault 2 was introduced with the functionality of full-disk encryption (FDE) and when it is enabled the entire contents of a drive becomes encrypted. In case of shutdown, all data becomes unrecoverable until the password is entered again. FileVault 2 made it possible to remotely wipe the disk in case of laptop loss. That is possible because FileVault 2 relies on the encryption key and, when a user wipes the drive secured with FileVault 2, it erases the key and makes all the data unobtainable for anyone. Another interesting fact about FileVault 2 is that even if the Macintosh machine is run in "Guest mode" it can remotely wipe the machine in case of theft or loss [14].

Apple Macintosh devices equipped with T2 chips have an additional security feature inherited from the iOS devices - the delay after a specific amount of incorrectly input passwords. Similar to iOS in which the device gets disabled after 9 attempts of entering the password, the Macintosh machines have a maximum of 30 attempts of entering the user password with delays specified in Figure 1. And if a user or investigator uses up 30 attempts to log in without success than rebooting the Macintosh computer into Mac OS Recovery grants 10 more attempts to log in. If that is not sufficient enough, then the user or investigator has 30 attempts per each additional method of decryption which are: iCloud recovery, FileVault recovery key, institutional key. When all the attempts are exhausted, the "Secure Enclave" will no longer process any decryption attempts of the volume and the volume remains useless or unrecoverable and can only be erased to allow a clean system to be installed on it. Using this method Apple is making brute force attacks useless [15].

Attempts	Delay Enforced
1–14	none
15–17	1 minute
18–20	5 minutes
21–26	15 minutes
27–30	1 hour

Figure 1. Macintosh delays between password attempts[15]

IV. IMAGING MAC COMPUTERS

Imaging a Macintosh computer is, as in all branches of digital forensics, a very important step. Unlike the Windows machines that everybody is used to image, Mac OS can present a greater challenge to the forensics investigator. That is because of all the security features that these machines are equipped with, such as secure boot, external boot, user and firmware password. Because of all these security measures the best-case scenario for a forensic investigator is to find a powered on and unlocked Macintosh computer. The investigator will then have access to all, or most of the data, from the machine. For example, the newest Macintosh computers give the investigator a chance to image the hard drives on it, but without the user's password the "Physical memory" or RAM can't be imaged using even specialized tools such as "MacQuisition". Today all Mac OS users are required by the OS to secure their computers with a password. On top of that systems newer than Mac OS version 10.7 do

not allow auto logging by default which makes it harder to acquire a forensic image of a machine. Since the release of macOS 10.14 Mojave, Apple implemented Fusion Drives for the APFS file system. This functionality can combine two different drives such as SSDs and HDDs that have a different capacity (SSD 128GB, 1,2,3 GB HDD) combining their capacity. The important fact about "Fusion drive" is that it creates a separate virtual disk or, so-called an "APFS Container" (shown in Figure 2), with the combined capacity of both the disks married. The important fact to remember regarding the imaging of these drives is that the forensic investigator must choose to image the container as the source disk and the AFF4 image format [10].

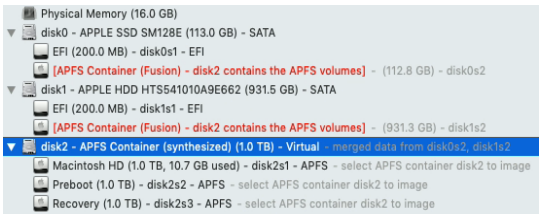


Figure 2. Recognizing and imaging a "Fusion Disk"[9]

A. *MacQuisition disk imaging*

MacQuisition [10] is a forensic acquisition and imaging software that supports many Mac OS, it also has the functionality to image "Physical memory" or RAM from a live machine. It is capable of running within the OS X boot environment. MacQuisition is also the first forensic tool capable of creating a physical image of Macintosh equipped with the Apple T2 chip. MacQuisition supports booting in a forensically safe environment and can acquire data from over 185 different Macintosh computer models in the native environment and can even image "Fusion Drives" [11].

During testing, MacQuisition proved to be an excellent tool to image a powered off or live Mac machine. During a live acquisition (when the Mac OS is unlocked) MacQuisition first prompts for the user password. After entering the password MacQuisition instantly detects that full disk encryption is enabled and lists the encrypted drives (like in Figure 3) which is a good reminder for the investigator to create a logical image of drives while they are unlocked. A RAM dump can also be created but only in case the investigator knows the user password and MacQuisition can only do a logical image of the machine, even though BlackBag states that they can create a physical image of a T2 equipped Macintosh, it is only possible to do so while the machine is shut down. That is because the T2 chip is used for real-time decryption by the operating system.

If performing a postmortem acquisition, the investigator must know at least the user password to be able to image the drives. The drives on T2 equipped Macintosh computers are encrypted by default. Also, if File Vault 2 is enabled it can be imaged, but the investigator has to provide the File Vault password, recovery key, or the keychain file of the Mac OS to decrypt and image the disk. The encrypted disk can then be imaged. Using MacQuisition the investigator can make a logical or physical copy of the suspect's disk. For another scenario, which includes the Macintosh to have "Secure Boot" enabled, the investigator will not be able to boot any forensic software due to "Secure Boot" and will be prompted to restart to the macOS X or to access the recovery mode using the key combination "Command+R" to disable the "Secure Boot" and "External Boot" option. To successfully boot the machine using a forensic software, the investigator will have to know the user password that protects the "Startup Security Utility" and disable all the security features in it.



Figure 3. MacQuisition Full Disk Encryption detection

B. *Recon Imager data acquisition*

Recon Imager [19] is another commercial forensic tool made to image Mac OS. It is a bootable imaging utility that supports all modern Intel-based Apple computers, including the newest Macintosh computers. Recon Imager is based on a Mac OS environment, modified to be forensically sound and ensures write protection for internal and external media. Like MacQuisition, Recon Imager also has the functionality to identify Apple File System (APFS) containers and disk volumes, FileVault, Fusion drives and Core storage volumes. Recon Imager has 3 different versions or modes that provide support for Macintosh computers with different hardware. Another functionality of Recon Imager is the possibility of imaging RAM without the need to type in the user or administrator password. The image of RAM will only contain residual data from previous sessions because the RAM imaging functionality is accessed from the boot environment. And the amount of data left in the RAM can differ, depending on the size of available memory and software that was run on the machine [12, 13].

V. RECOVERING DATA FROM FILEVAULT

As mentioned earlier, to recover user's data from an encrypted FileVault disk the best-case scenario is to have the user's password. The password can be recovered during the interview with the user or suspect, or by finding it written down somewhere. The investigator has everything needed to gain access to users' data. If password could not be acquired, the next step is to search for the recovery key which enables the investigator to decrypt the disk using recovery mode accessed by pressing and holding "Command+R" keys. Another option is to recover the key from the iCloud backup and decrypt the drive. The next scenario would be to use the keychain from the suspect's Mac OS and decrypt the drive with it. If the locked Macintosh is a part of a company network, which has a group policy set to enable FileVault by default, it is very likely that the system administrator has set the policy to save all the passwords centrally. This means that the administrator potentially

can provide the password to unlock the Mac, with the assumption he is not involved in the crime. The last solution would be to look for another user with an "Administrator" account that has a known password, or an easier way would be to guess the password using the users hint. All the passwords can be changed from the "User & Groups" settings and using this method the suspect's account can be unlocked. This method can be used for cases that take place in a company and a less likely scenario would include cases where private people are involved. If the investigator cannot use any of the previously mentioned methods, he will be forced to recover the password using different recovery methods - brute force attack, dictionary attack, etc. The biggest problem with recovering the key in all brute force attacks is that it takes too long, even with a high-end computer.

A. *FileVault Cracker*

The first tool to try recovering the FileVault password with would be "FileVault Cracker" which is an open-source tool made to recover the encrypted drive password using a dictionary attack on it. This tool was developed as a private project and it is not finished which means that some bugs or errors can occur during use. It can be used to recover passwords from "CoreStorage" drives using HFS+ file system. The software is available on GitHub [17], written by the user "mac made". It is an XCode project that must be compiled or published to work. The software has a simple and intuitive user interface (Figure 4) and the only thing that is required from the user is to enter the encrypted drives UUID. FileVault Cracker already has a dictionary file in it, which the user can change easily using the user interface. Additional settings are to generate case variants for words up to 20 characters, to generate common derivations for words up to 20 characters and to choose the number of threads the software will use while recovering the password.

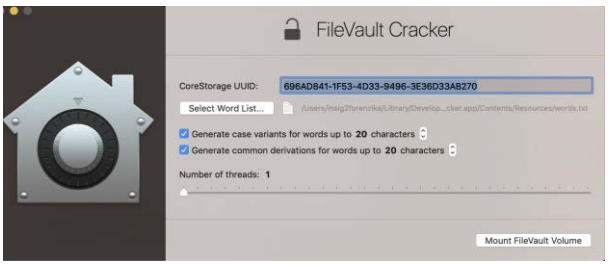


Figure 4. FileVault Cracker

B. *John the Ripper and Hashcat*

The tools John the ripper and Hashcat can also be used to recover FileVault2 passwords. This chapter will in short explain how to recover a FIleVault2 password on "HFS+" and "APFS" file systems.

To start the recovery of FileVault 2 password on "HFS+" file system some requirements have to be fulfilled:

- this will work for target systems on a MacBook Air running macOS v10.12.6;
- attacker machine has to be an iMac running Mac OS 10.14.2;
- the investigator must download and compile "fvde2john" and "Hashcat" on iMac.

The next step in this process is to run the MackBook Air in "Target Disk Mode" (this means that the Macintosh device is working as an external drive), the target device must be connected to the iMac machine via thunderbolt, FireWire, or USB. When the two machines are connected, the investigator can run the command "`diskutil list`" on the iMac to see all available disks. After the investigator has found the target disk (which is marked as external), the next step is to identify the "Apple_Boot Recovery HD" drive. Then the "Recovery HD" drive must be mounted using "`diskutil mount /dev/disk_s_`". The name of the disk can be found in the "IDENTIFIER" column. The next command to run is "`find /Volumes/Recovery\ HD -name Encry*`" to find the file containing the string "Encry" in itself. To gain the hash of the encrypted volume the next command must be executed "`sudo./fvdetools/fvdeinfo -e/Volumes/Recovery\HD/com.apple.boot.S/System/Library/Caches/com.apple.corestorage/EncryptedRoot.plist.wipekey -p don't-know /dev/disk2s2`". The MacBook can then be unmounted and "hashcat" run using the next command "`./hashcat/hashcat -a 0 -m 16700 -o found.txt hash.txt wordlist.txt`" which will recover the decryption password. The recovery process duration can vary depending on the simplicity and length of the password and encryption algorithm [18].

In the next few lines the recovering the FileVault 2 password from an APFS file system will be explained. This process can be used on a MacBook Pro running MacOS v10.14.2 which will also be the target machine. For this method of password recovery, the attacking machine is running Ubuntu 19.04. For this process to be successful "apfs-fuse" must be installed, which is a driver for read-only interpreting data and recognizing disks using the APFS file system. After the "apfs-fuse" driver is successfully installed, which will be indicated with the message "[100%] Built target apfs-fuse" in the Ubuntu "terminal", the next step is to put the source or the suspect's machine in to "Target Disk Mode" and connect it to the Ubuntu machine. Using Ubuntu "terminal" the investigator must determine which disk name belongs to the encrypted drive. The correct drive is going to be determined using the command "`cat /proc/partitions`" through the "terminal" window. The output of the terminal will look like Figure 5. When the name of the encrypted drive is determined, the investigator is ready to acquire the encrypted disc's hash value. The hash value will be generated using the command "`sudo ./apfs-dump-quicq /dev/sdb log.txt`" in which "sdb" represents the name of the encrypted disk. The values used to create hash can be found next to the identifiers "Salt", "Iterat's", "KEK Wrpd". These values will look similar to Figure 6. To arrange the hash to be recognized by "hashcat", the values next to identifiers will be combined respectively with the prefix "$fvde$2$16$". Between each value (Salt,

Iterat's, KEK Wrpd) will be a "$" sign connecting them into a hash value. When the hash was created the recovering of the password using "hashcat" can be started. If "hashcat" is not installed on the Ubuntu machine, the first step is to install it. The command "`hashcat -a 0 -m 18300 -o found.txt hash.txt wordlist.txt`" can be used to start the password recovery. This particular command is using hashcat's dictionary attack to recover the FileVault 2 password from an APFS drive [18].

```
$ cat /proc/partitions
major minor  #blocks  name

   8        0  125034840 sda
...
   8       16  118489088 sdb
```

Figure 5. Identifying the target drive name

```
[KEK]
Unk 80   : 0
UUID     : 48BCAEEB-4E7A-C5D2-B7EB-C21DCD1366F9
Unk 82   : 00000000 0002 15 174
KEK Wrpd: 2FDEAFAA0F6A971F674B487270A5AE59578B29FB377F76E35CF
Iterat's: 100000
Salt     : 692B540738291E8B5248A74444E5B1EF
```

Figure 6. Identifying the target drive name

VI. CONCLUSION

Through this paper, most of the issues and impediments that the investigator has to overcome have been covered. And the whole process of bypassing the security system will only allow the investigator to acquire or image a drive of Mac OS. The T1 and T2 security chips and their functionality have been covered. Because of these security upgrades which made it harder for investigators to deal with Apple or specifically Macintosh devices, all this information lead to the conclusion that the best and easiest way of acquiring data is when the Apple device is found turned on and unlocked. As shown in one of the examples, even when the Mac OS is unlocked, the physical memory cannot be imaged without entering the user's password which is in the world of Windows forensic almost a trivial task. The recovery of FileVault2 passwords is a difficult task to do by itself, even on the older Macintosh machines that do not have the T2 security chip. It is well known in the forensic community that a password longer than 10 characters is impossible to recover it using brute force attacks. And the use of uppercase, lowercase and special signs in passwords make the recovery exponentially more difficult. The T2 security chip plays a big role in the security of Apple devices and digital forensics, mostly because of the security features that prevent the acquisition of disks on Apple machines. The biggest problems for forensic investigators is the placement of the T2 chip. It is placed between the CPU and the disk on a secure bus, which makes it impossible to acquire and decrypt data without using the T2 chip for decryption. All in all, Apple's security and encryption is getting stronger with every new iteration. And even now there are only few solutions that can recover data from Apple devices that are encrypted, which makes the forensic investigation of such devices impossible for an investigator who don't know how to use these solutions.

VII. REFERNCES

[1] Niranjan Reddy, Mac OS Forensics, Practical Cyber Forensics, Apress, Berkeley, CA, pp 101-132, July 16, 2019

[2] Sarah Edwards, Sans: Mac and iOS Forensic Analysis and Incident Response (2020). Retrieved from: https://www.sans.org/course/mac-and-ios-forensic-analysis-and-incident-response

[3] Kevin J. Ripa, Computer Evidence Recovery, Forensic Acquisition of Mac Computers (Mar 8, 2016). Retrieved from : http://www.computerpi.com/forensic-acquisition-of-mac-computers/

[4] Pepijn Bruienne, Duo Secutiry, Apple iMac Pro and Secure Storage (may 2,2018). Retrieved from: https://duo.com/blog/apple-imac-pro-and-secure-storage

[5] John Martellaro, The Mac Observer, How Apple's T2 Security Chip Affects Your Disk Storage (Jan 10, 2019). Retrieved from: https://www.macobserver.com/tips/deep-dive/apple-t2-security-chip-disk-storage/

[6] Forensic Focus, MackBook Air firmware password (Jun 23, 2014). Retrieved from: https://www.forensicfocus.com/Forums/viewtopic/t=11911/

[7] Apple Develoer, Storing Keys in the Secure Enclave (2020). Retrieved from: https://developer.apple.com/documentation/security/certificate_key_and_trust_services/keys/storing_keys_in_the_secure_enclave

[8] ALLservice, Service Forum, Unlocking firmware, PIN & icloud. Apple iMac, Macbook (Apr 09, 2013). Retrieved from: https://www.allservice.ro/forum/viewtopic.php?t=2724&sid=36bd813863cfa6bad9c56a69e4df4ea7

[9] Ashley Hernandez, BlackBag, Insights Blog, Apple T2 Chip Systems: Create Decrypted Physical Images With MacQuisition (May 08,2019). Retrieved from: https://www.blackbagtech.com/blog/apple-t2-chip-decrypted-image/

[10] BlackBag Technoligies, MacQuisition 2019 R1.2 (May 30,2019). Retrieved from: https://www.blackbagtech.com/software-downloads/releaseNotes/mq2019r1_2.pdf

[11] BlackBag Technoligies, MacQuisition Quick Start Guide (2020). Retrieved from: https://www.blackbagtech.com/macquisition-quick-start-guide/

[12] Sumuri, Recom Imager (2020). Retrieved from: https://sumuri.com/recon-imager-manual/

[13] Sumuri,YouTube ,Recom Imager – Booting up and Interface Overview (Apr 21, 2017). Retrieved from: https://www.youtube.com/watch?v=9H-V4226Gb0

[14] Glenn Fleishman, Macworld, How to encrypt your Mac with FileVault 2, and why you absolutely should (Feb 5, 2015). Retrieved from: https://www.macworld.com/article/2880039/how-to-encrypt-your-mac-with-filevault-2-and-why-you-absolutely-should.html

[15] Der Flouder, T2, FileVault and brute force attack protection (Nov 01, 2018). Retrieved from: https://derflounder.wordpress.com/2018/11/01/t2-filevault-and-brute-force-attack-protection/

[16] CMIzapper, Technology for your Mac repair business, Matt Cards (2020). Retrieved from: http://www.cmizapper.com/products/mattcard.html

[17] GitHub, FileVaultCracker (Oct 17, 2018). Retrieved from: https://github.com/macmade/FileVaultCracker

[18] tinnyapps.org, Cracking FileVault 2 (HFS+ or APFS) (May 27, 2019). Retrieved from: https://tinyapps.org/docs/cracking-filevault.html

[19] Sumuri, Recom Imager (2020). Retrieved from: https://sumuri.com/software/recon-imager/

[20] Openwall, John the ripper password cracker (2020), Retrieved from : https://www.openwall.com/john/

[21] GitHub, hashcat , (2020). Retrieved from : https://github.com/hashcat/hashcat

Red Teams - Pentesters, APTs, or Neither

Ivan Kovačević, Stjepan Groš
University of Zagreb, Faculty of Electrical Engineering and Computing,
Laboratory for Information Security and Privacy,
Unska 3, 10000 Zagreb, Croatia
{ivan.kovacevic, stjepan.gros}@fer.hr

***Abstract*—In cybersecurity, red teams emulate real threats and launch attack operations to assess an organization's security posture and provide experience to it's defenders. Although red team engagements share similarities with penetration testing and vulnerability scanning, they differ significantly in terms of goals and purpose. Moreover, just mimicking real threats does not make red teams APTs. This paper summarizes various academic and industrial views on the topic of cyber red teaming and aims at clarifying the common properties that differentiate it from other types of security assessments and real attackers. Based on evidence from case studies and various expert opinions, we have concluded that red teaming is a distinct type of security assessment, and that penetration tests are often mislabeled as red team engagements.**

***Index Terms*—cybersecurity, red teaming, penetration testing, security assessment, advanced persistent threat**

I. Introduction

The term *red team* allegedly stems from the cold war, when red used to represent the Soviet Union in western countries [1]. Today, red teams are still used in the military, extending their operations to the cyberspace [2] [3], but have seen a wide adoption in the industry as well. Many companies in the industry provide various red team services, that often resemble penetration testing or even vulnerability scanning, with nomenclature being guided in large part by marketing [4]. This paper provides a survey of red team related topics in cybersecurity from various sources, and proposes classification and definitions for red teams and red teaming, and their relationship to penetration testing, APTs, vulnerability assessment, and cybersecurity exercises.

The structure of this paper is as follows. Section II describes how we performed the literature search. Then, Section III analyses the usage of the term *red team*, first giving the military views, followed by the views from academia and the industry. In Section IV, we describe the cases of three recent targeted attacks performed by *Advanced Persistent Threats* (APTs) to draw a comparison between them and red teams. The paper finishes with conclusion in Section VI, followed by a list of references.

II. Methods

We surveyed a diverse set of academic articles and industry blogs related to red teams, penetration testing, security exercises, and threats. To find articles and reports, several literature searches were performed using Google Scholar and Google Search between December 2019 and February 2020, with the search terms "red teams in the military", "red teams in cyber security", "red team framework", "red teaming vs penetration testing", "red team exercise framework", "penetration testing framework", "ukraine cyber attack power grid", and "swift cyber attacks". After finding articles of interest, additional articles were found among their references and citing articles.

To find relevant blogs and insights from the industry, we searched for prominent Twitter accounts belonging to experts from the industry, which often have associated blogs. They have tens of thousands of followers from the cybersecurity community, meaning that a wide professional audience considers their views, experience, and advice as relevant. The search on Twitter was performed simultaneously with the searches on Google Scholar. Instead of searching Twitter using search terms, we relied on accounts related to cybersecurity that one of the authors already followed, and the accounts that those accounts followed. Out of those, Twitter accounts with resources most relevant for this paper are @DanielMiessler, with 103 000 followers, @PwC_UK, with 96 000 followers, and @DarkReading, with 223 000 followers as of February 2020.

Out of the results, we singled out a collection of representative papers and blogs related to questions discussed in this paper.

III. Survey of Usage of Terms

This section provides views regarding red teams and red teaming from the military, academia, and industry.

A. Red Teams and Red Teaming in the Military

Before cybersecurity became what it is today, red teams and red teaming were used in the military, government organizations, and management [2]. Gold and Hermann [2], in their report on military use of red teams, define red teams and related processes in very broad terms, using their properties and goals. Besides targeting personnel and physical systems, they argue that red teams in the military can also target information and computer systems, bringing them firmly into the scope of this paper.

First of all, a red team is established to challenge aspects of an organization's plans, programs, assumptions, etc. [2], while *red teaming* is a process that includes the red team, the blue team, and management, and defines reporting and product use [2]. As noted by [2], strategic level red teaming aims to challenge the organization's assumptions and visions, operational level tests postures, war plans, and portfolios, while tactical level trains military units and gives feedback on programs in development. The authors differentiate and define three types of red teams: (i) Surrogate adversaries and competitors, (ii) devil's advocates, and (iii) sources of judgement outside normal processes [2].

Surrogate adversaries and competitors serve the purpose of training blue team members, so they would be better prepared to face real adversaries. These types of teams can be restricted to imitate the behaviour of known potential adversaries, but can also display original tactics to counter those known to be employed by the blue team. Devil's advocates analyse the existing assumptions, strategies, plans, programs, and processes, anticipate problems and possible surprises, and provide critique and improved alternatives. Finally, the idea of sources of judgement outside normal processes is that red team members can be consulted when performing decisions [2]. While the first type of red teams has similarities to the corresponding use cases in cyber security, devil's advocates are effectively reviewers, and sources of judgement are consultants. To be successful, a red team must [2]:

(i) take its assignments seriously,
(ii) be independent and not bound by bureaucracy,
(iii) be involved in decision-making,
(iv) have adequate interaction with the blue team, and
(v) respect data confidentiality

Similarly to [2], the UK Ministry of Defence [5] defines red teaming as an independent application of structured, creative and critical thinking techniques to inform the end user for making a better decision or producing a better product.

Red teams often appear in the context of cybersecurity exercises. Ablon et al. [3] define a framework for NATO cybersecurity exercises, and analyse related long term strategies. The goal of such exercises is to promote cooperation of members, and to assess and train the cybersecurity capabilities of military and industrial personnel [3]. Exercises are usually organized on a smaller range of levels, e.g. strategic and operational level tabletop exercises for higher military officers, missing real IT infrastructure and personnel, or tactical and technical exercises with red teams attacking a cyber range defended by the blue teams [3]. Future exercises would aim to combine all levels, integrating IT exercises with exercises focused on higher officers, and even kinetic exercises, to promote communication and control among the trained organizations and personnel [3].

An example of a military cyber security exercise is the exercise Locked Shields, described in [6]. This exercise was performed on a cyber range simulating a NATO mission network, with classified and unclassified parts, that was partitioned among a large number of blue teams from various countries [6]. Each blue team had to defend their partition from attacks, and was penalized if those attacks succeeded [6]. The red team was tasked with performing attacks from a predefined scenario, such as *Denial-of-Service* (DoS), website defacement, data wiping, etc [6]. This provided a training opportunity for the blue team, and gave performance information that identified the areas the teams could improve on [6].

Other use cases for red teams in the military are *Responsive Cyber Defense* (RCD) and *Cyber Offense*. In RCD, Communication and Information Systems (CISs) under attack are defended by attacking the CIS of the attackers or other involved parties [7]. Cyber offense is concerned with striking the opponent's military, government, and sometimes even civilian cyber infrastructure supporting the war effort [8]. It can be performed as tactical operations to support combat forces [8], and in strategic operations to disrupt services in the enemies' rear and lower their morale [8]. Lewis [8] sees a red team as focused on testing defenses, but also capable to provide a basis for offensive capabilities [8]. To provide offensive capabilities, the red team needs to be expanded with planning capabilities, operational context, and intelligence [8].

B. Red Teams and Red Teaming in Academia

Most academic work regarding red teams in cybersecurity that we surveyed belongs either in the category of

formal models for various red team emulation systems or penetration testing. This section contains several related viewpoints we found during research.

Hoffmann made a survey of simulated penetration testing [9] in 2015. He described several systems that attempted to replace humans in penetration testing, by preforming attack planning over a model of the target system [9]. In essence, attack actions are defined using preconditions and effects, and an attack plan that leads to the specified goal is created [9]. Actions can range from technical attacks, such as service vulnerability exploitation, to *socio-technical* attacks, such as spear-phishing and looking over someone's shoulder to obtain information [9]. In addition, to realistically simulate attackers, the simulation includes their acquisition of information regarding the targeted systems, as well as their usage of common sense knowledge and guesses, including some levels of uncertainty [9]. The goal of the approaches he surveyed is to realistically simulate human hackers and determine the vulnerabilities of the IT system, or basically to perform network penetration tests.

Applebaum et al. [10], [11] and Miller et al. [12] describe red teams as emulators of real adversaries. They argue that adversary emulation enables defenders to find weaknesses and vulnerabilities in their systems, and see their systems from an adversaries perspective. Moreover, they describe the differences between red team assessments and penetration tests. While network penetration tests aim at compromising the network and finding vulnerabilities, red teams also perform other tasks, such as reconnaissance, tool and infrastructure development [12], and extensive post-compromise work [11] to achieve the emulated adversaries' goals [12]. The authors also mention adversary emulation exercises as a distinct category from both penetration testing and red teams, but do not elaborate it's differences in regard to red teams. To sum up, red teams are viewed primarily from a technical perspective, focusing on the emulation of adversaries' TTPs and their detection by the defenders.

C. Red Teams and Red Teaming in the Industry

The authors in [13] made a series of interviews with influential professionals from the industry regarding their opinions on various red teaming topics. Their results have shown that there is a plethora of different views on red teaming and that many of those stem from organizations naming vulnerability assessments inconsistently, usually related to marketing [13].

The association of banks in Singapore provides guidelines for red team exercises for its member organizations [14]. They focus on red teams exercises that emulate real threats using TTPs, which they refer to as *Adversarial Attack Simulation Exercises* (AASE). The guidelines, in addition to defining specific roles involved, and a methodology for organizing such exercises, discuss the key differences between penetration tests, red team exercises, and real adversaries. According to them, some key characteristics of penetration tests are as follows [14]:

(i) they are limited to technical assets,
(ii) they follow industrial technical methodologies,
(iii) they are known to all parties involved, and
(iv) their objectives are to find and assess vulnerabilities and weaknesses of the targeted systems, so they can be patched.

On the other hand, some key properties of AASEs are [14]:

(i) they can involve both technical assets and physical domain,
(ii) they can target people and processes,
(iii) they are kept secret from everyone not involved in their organization for their duration, and
(iv) they aim at testing the defense and detection capabilities of the organization.

Finally, real adversaries have the following key properties [14]:

(i) are not bound by laws, ethics, and the well being of the company, its employees, and partners, and
(ii) unlike penetration tests, and red teaming assignments, they are not under the company's control.

In his book on red teaming [1], Oakley describes penetration testing and red teaming as two offensive security techniques in cyberspace, and uses the term *red team* as an umbrella term for the methodologies of both. According to him, red teaming serves several purposes [1]:

(i) it evaluates how the organization and it's employees are prepared to attacks, and identifies whether technologies are useful for preparedness, or if resources are wasted,
(ii) it evaluates the defenses of the organization,
(iii) it evaluates how well the organization monitors for malicious activity, be it physical, cybernetic, or related to policies and procedures, and finally,
(iv) it evaluates how well the organization responds to those activities

It can be noted that this definitions show similarities to the properties of red teams described in Section III-A.

Although he mentions penetration testing and red teams interchangeably, it is interesting to note that when he mentions penetration testing, it is mostly related to individual pentesters or outside contractors who perform it, and when he mentions red teams, they are mostly related to groups of people and internal organizatinal units [1]. Finally, he distinguishes automated vulnerability scans from penetration tests, because they are non-invasive and can uncover only a small set of vulnerabilities, leading to a false sense of security [1].

Daniel Miessler, an experienced cybersecurity professional, writes a series of blogs regarding various cybersecurity topics. In [4], he discusses the terminology and differences between various types of cyber security assessments.

He defines a vulnerability assessment as a technical assessment useful for companies with low to medium maturity in regards to security, that aims to yield as many vulnerabilities as possible in the environment, coupled with severity and priority information [4]. Vulnerability assessments are undertaken when it is assumed or known that security problems exist in the environment [4].

A penetration test, on the other hand, is a technical assessment set out to achieve a specific goal, such as e.g. stealing critical data, or obtaining access to a resource [4]. They make sense in organizations that already have a mature security posture, and that need to validate their configurations for security problems [4]. These organizations should already have cybersecurity policies and procedures in place, such as e.g. regular patching and deleting inactive user accounts [4]. The properties of an penetration test differentiating it from red teams are the following [4]:

(i) it has a definition and scope,
(ii) it is a point-in-time test, and
(iii) it has specific goals for success or failure

Finally, a red team is a continuous service that emulates real threats to improve the blue team's and overall company's performance in regards to cyber security [4]. Although red teams can share TTPs with penetration tests, red teams are focused on improving the blue team, while penetration testing, as mentioned earlier, validates the configuration of the environment [4]. Miessler considers the following elements that all red teams should have [4]:

(i) organizational independence,
(ii) defensive coordination,
(iii) continuous operation,
(iv) adversary emulation, and
(v) efficacy measurement.

Red teams are appropriate in organizations with a mature security posture that actively detect and respond to threats using blue teams, and are often corporate, operated from within the organization [4].

IV. Examples for comparison of red teams and APTs

This section will provide examples of three targeted attacks that have been attributed to well known *Advanced Persistant Threats* (APTs). Those examples will service as a basis for later comparison of APTs with red teams in Section V. The first attack aims to disrupt critical infrastructure, the second one is basically a financial crime, and the third one is concerned with industrial cyber espionage.

In December 2015, several Ukrainian power distributors were struck with a sophisticated cyberattack, causing a blackout for approximately 225 000 customers [15]. The attackers used the *BlackEnergy* malware, that was disseminated through several email spear phishing campaigns in early 2015, using malicious Office documents [16]. Users from the attacked power distributors opened the attachment, providing attackers with an initial foothold within their networks, from which they collected information and moved laterally towards the *Industrial Control System* (ICS) networks [15]. Once they reached the ICS networks, they likely performed detailed reconnaissance and reconstructed the ICS configuration within their own organization, finally using expert knowledge of those ICS systems to craft malicious firmware [15]. When the target time of attack came, the final phase involved installing the malicious firmware, locking out the operators in the affected companies, wiping workstations and servers, and cutting off power for the customers. Based on malware samples and intelligence, this attack was attributed to a Russian, possibly government controlled, threat actor [17].

The cyber attack on the national bank in Bangladesh happened in February 2016, resulting in one of the largest bank heists in history [18]. Attackers obtained access to the bank's systems in early 2015, and moved laterally towards the SWIFT access point, used to order transactions from the Bangladesh bank's accounts in other banks [18]. Logistic preparations for the attack are known to have started in May 2015, when two Chinese citizens opened fraudulent accounts in the RCBC bank in the Philippines [19], together with a local bank manager [20]. Finally, the attackers modified the SWIFT software in a highly sophisticated manner [18], and ordered fraudulent transactions in a carefully planned time-frame [21]. The transactions were performed on Thursday, 4th

February, the last day before the weekend in Bangladesh, and at the same time directly preceding a national holiday on the Philippines on Monday, 8th February [22], giving the attackers enough maneuver space to withdraw and launder the money in casinos in Manilla [21]. Based on malware samples and various information, this attack was attributed to the APT Lazarus, controlled by the People's Republic of Korea [23].

A series of worldwide espionage attacks, nicknamed *Operation Cloud Hopper*, started in 2016, and continued since, targeting industries that China deemed as strategic [24]. This attacks were specific, because instead of attacking organizations directly, they focused on compromising their *Manage Service Providers* (MSPs), making them a variant of supply chain attacks [24]. MSPs posses a lot of data regarding their clients, and usually have some level of remote access to their IT infrastructure, providing many opportunities to the attackers [24]. Once the clients systems were compromised, sensitive data was exfiltrated trough the MSPs network [24]. Based on malware samples, compile times, and other information, the operation was attributed to a well known Chinese threat actor, often involved in espionage [24].

V. Results

Based on [2] and [5], we can conclude that in the military domain, a red team, according to it's member specialties, can perform various types of assignments, such as performing adversary simulation and competitor exercises, to train blue teams, and perform reviews and consulting. Also, the team can perform on tactical, operational, and strategic levels. Red teaming, on the other hand, is the process, that includes other actors, infrastructure, time, etc., in which activities of the red team are performed.

Red teaming can include the tactical and technical levels, as exemplified by the blue and red team interaction in the military cybersecurity exercise Locked Shields [6], but according to definitions from [2] and [5], can also be focused on high-level exercises entirely devoid of IT infrastructure, such as those for higher officers mentioned in [3].

From [8], it can be taken that a typical NATO red team will be mostly technically skilled, but have deficiencies when it comes to operation planning and intelligence gathering. Nevertheless, it can provide a basis to develop offensive [8] and RCD [7] capabilities.

Since the tabletop exercises for higher-level personnel challenge, among others, security policies and procedures, and that they indirectly influence the IT infrastructure, vulnerabilities, and the overall security posture, they can also, by definitions in Section III-A, be considered as red teaming.

From the sources in Subsection III-B, not only can it be concluded that red teams can perform penetration tests, but also that the properties of red teaming as described in [11], are essentially the same as the descriptions of penetration tests in [9]. Moreover, when we compare the academia articles with the reasoning from the military in Section III-A, it is clear that the academic cases focus on individual technical subtasks of what is considered cybersecurity red teaming in the military.

Industry, as seen in Section III-C, lacks a common vision regarding red teams and penetration tests, although efforts such as [4] show that there are attempts at providing a common terminology. Miessler [4] argued that red teams are unique because it primarily aims at training the organization's defenders, and lasts much longer than penetration testing, preferably continuously.

Finally, from the APT targeted attacks, three of which were described in Section IV, it can be seen that they can involve various third parties and suppliers, and, like described in [14], do not care for the welfare of people and organizations. Unlike red teams, APTs do not attempt to train blue teams, and are not expected to organize meetings with them and file reports on activities and conclusions. APTs also involve various support operations, among others, money laundering, and managing botnets, such as those surveyed in [25], that are not directly related to red teams. Finally, when comparing cyber offensive operations described by [8], and the attacks from Section IV, it can be seen that there are many similarities, especially with the case from Ukraine. The attacks lasted for prolonged periods of time, included logistic support and used sophisticated domain knowledge, with concrete operational and strategic goals in mind. From the enemies perspective, an offensive military cyberspace operation could resemble an APT attack.

A. Proposed definition of red teaming

Based on insights from Section III, summarized on Table I, and Figures 1 and 2, we propose the key properties of red teams and red teaming as follows.

Red team. A red team is a group of people with various skills, some having technical skills and knowledge of TTPs, some more focused on tactical planing, and some offering support in various tasks, such as e.g. gathering intelligence about the targets. Some members can have multiple skills at once, and the combination of the skills within the team is mainly determined by the types of

	Red Team	Red Teaming	Penetration testing	Cyber Security Exercises	Real targeted attack
Group of experts	yes	N/A	N/A	N/A	N/A
Adversary simulation	sometimes	yes	sometimes	sometimes	N/A
Duration of operations	N/A	long	short	short	long
Tactical level planning	yes	yes	yes	sometimes	yes
Operational level planning	consulting	no	no	sometimes	sometimes
Strategic level planning	consulting	no	no	sometimes	sometimes
Tabletop exercises	sometimes	no	no	sometimes	?
Finds technical vulnerabilities	N/A	few	many	few	N/A
Finds organizational vulnerabilities	N/A	yes	sometimes	sometimes	N/A
Goal	N/A	training	security validation	training	varies
Ethical and authorized by target	N/A	yes	yes	yes	no

TABLE I: Comparison of various terms

exercises and assessments that the organization needs. It is important to note that the team does not perform operational and strategic planning, that are rather provided by other parts of the organization, or the military in case of military red teams. However, they can provide consulting for operational and strategic planning. The team can either be established for internal use in an organization as a continuous service for training the defenders, or can be provided as a service to other companies. In both cases, it needs to have enough independence to provide its services with adequate quality. This also means that a red team can also be a part of military offensive operations, or APTs, and that these types of application are determined by the circumstances, and operational and strategic goals the red team serves.

Red teaming. Red teaming is a methodology trough which a red team, with authorization, in accordance to the organization's policies and defined scope, conducts planned attack exercises with the goal of training the organization's defenders and employees. While organizers of the exercise within the organization are informed about at least its high-level details, the defenders are usually not informed in advance, and do not know that an exercise is in progress. We denote the exercises in which red and blue teams perform short-lasting exercises on a cyber range, with blue teams knowing that they are performing an exercise, with the term *cybersecurity exercise*, and do not consider them red teaming in this sense.

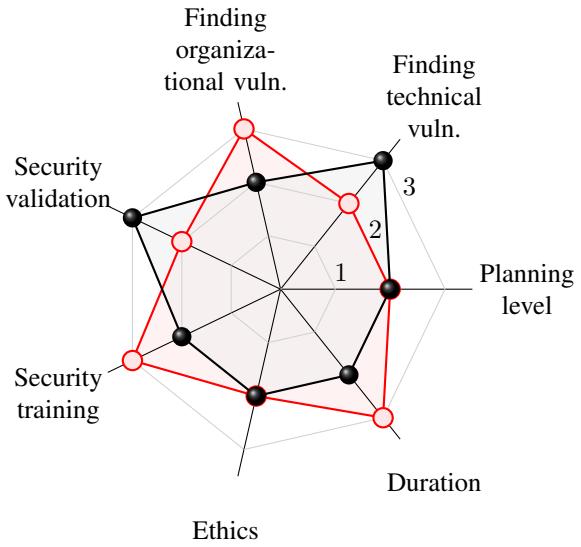


Fig. 1: Comparable dimensions of penetration testing and red teaming. A higher value for a dimension means that the type of assessment provides a larger value for that dimension.

VI. Conclusion

This paper gave an overview of red teaming and related topics in cybersecurity. It surveyed various views on red teaming from the military, academia, and industry, combined with examples of real targeted attacks performed by APTs, and proposed definitions for red

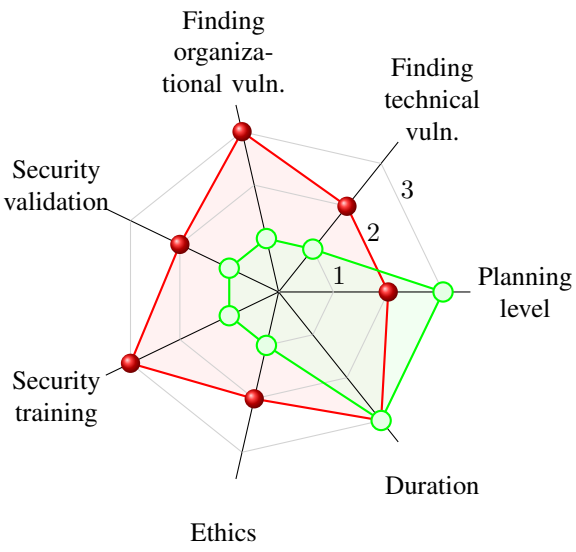


Fig. 2: Comparable dimensions of targeted attacks and red teaming. A higher value for a dimension means that the type of team has a larger value for that dimension.

teams and red teaming, that clearly differentiate them from penetration testing and APTs.

The main difference between red teaming and penetration testing is that the first one concerns with continuously training the organization's defenders, either by simulating real attackers, acting as devil's advocates, or through consultation, while the second one lasts a short period of time and primarily focuses on validating the organization's security posture. Companies often mix the terms, to a great part due to marketing, as noted by Miessler [4].

Red teams and APTs are conceptually different, because a red team lacks it's own strategic and operational planning, and adheres to the organization's policies. Therefore, if the organization has a policy that allows the red team to attack other organizations with a given malicious operational goal, that organization could be viewed as a threat, and in case it has large enough amounts of resources allocated for those attacks, as an APT.

REFERENCES

[1] J. G. Oakley, *Professional Red Teaming: Conducting Successful Cybersecurity Engagements*. Apress, 2019.

[2] T. Gold and B. Hermann, "The Role and Status of DoD Red Teaming Activities," OFFICE OF THE UNDER SECRETARY OF DEFENSE WASHINGTON DC, Tech. Rep., 2003.

[3] L. ABLON, A. BINNENDIJK, Q. E. HODGSON, B. LILLY, S. ROMANOSKY, D. SENTY, and J. A. THOMPSON, "Operationalizing cyberspace as a military domain," 2019.

[4] D. Miessler, "Information security assessment types," *Available: https://danielmiessler.com/study/security-assessment-types/*, 2019.

[5] C. The Development and S. W. S. R. Doctrine Centre, Shrivenham, "Red teaming guide," 2013.

[6] B. Kulich, "Lessons learned from military cyber defence exercises," *Science & Military Journal*, vol. 9, no. 1, p. 47, 2014.

[7] P. Brangetto, T. Minárik, and J. Stinissen, "From active cyber defence to responsive cyber defence: A way for states to defend themselves legal implications," *NATO Legal Gazette*, no. 35, pp. 16–28, 2014.

[8] J. A. Lewis, "The role of offensive cyber operations in nato's collective defence," *Tallinn Paper*, vol. 9, 2015.

[9] J. Hoffmann, "Simulated penetration testing: From" dijkstra" to" turing test++"," in *Twenty-Fifth International Conference on Automated Planning and Scheduling*, 2015.

[10] A. Applebaum, D. Miller, B. Strom, C. Korban, and R. Wolf, "Intelligent, automated red team emulation," in *Proceedings of the 32nd Annual Conference on Computer Security Applications*, 2016, pp. 363–373.

[11] A. Applebaum, D. Miller, B. Strom, H. Foster, and C. Thomas, "Analysis of automated adversary emulation techniques," in *Proceedings of the Summer Simulation Multi-Conference*, 2017, pp. 1–12.

[12] D. Miller, R. Alford, A. Applebaum, H. Foster, C. Little, and B. Strom, "Automated adversary emulation: A case for planning and acting with unknowns," 2018.

[13] M. J. Carey and J. Jin, "Tribe of hackers red team: Tribal knowledge from the best in offensive cybersecurity," 2019.

[14] ABS, "Red team: Adversarial attack simulation exercise guidelines for the financial industry in singapore," 2018.

[15] D. U. Case, "Analysis of the cyber attack on the ukrainian power grid," *Electricity Information Sharing and Analysis Center (E-ISAC). Available: https://ics.sans.org/media/E-ISAC_SANS_Ukraine_DUC_5.pdf*, vol. 388, 2016.

[16] GReAT, "Blackenergy apt attacks in ukraine employ spearphishing with word documents," *Kaspersky APT Reports. Available: https://securelist.com/blackenergy-apt-attacks-in-ukraine-employ-spearphishing-with-word-documents/73440/*, 2016.

[17] J. Finkle, "Us firm blames russian 'sandworm'hackers for ukraine outage. reuters," *Available: https://www.reuters.com/article/us-ukraine-cybersecurity-sandworm-idUSKBN0UM00N20160108*, 2016.

[18] S. Shevchenko, "Two bytes to $951 m," *BAE Systems Threat Research Blog. Available: https://baesystemsai.blogspot.com/2016/04/two-bytes-to-951m.html*, 2016.

[19] A. Katz and W. Fan, "A baccarat binge helped launder the world's biggest cyberheist," *Bloomberg, Available: https://www.bloomberg.com/news/features/2017-08-03/a-baccarat-binge-helped-launder-the-world-s-biggest-cyberheist*, 2017.

[20] K. Lema, "Philippine court orders jail for former bank manager over bangladesh central bank heist," *Available: https://www.reuters.com/article/us-cyber-heist-philippines/philippine-court-orders-jail-for-former-bank-manager-over-bangladesh-central-bank-heist-idUSKCN1P40AG*, 2019.

[21] S. Quadir, "Spelling mistake stops hackers stealing $1 billion in bangladesh bank heist," *Independent; Available: https://www.independent.co.uk/news/world/asia/spelling-mistake-stops-hackers-stealing-1-billion-in-bangladesh-bank-heist-a6924971.html*, 2016.

[22] R. K. Byron and M. F. Rahman, "Hackers bugged bangladesh bank system in jan," *Available: https://web.archive.org/web/20160312145208/ http://www.asianews.network/content/hackers-bugged-bangladesh-bank-system-jan-11271*, 2016.
[23] K. J. Higgins, "Inside the north korean hacking operation behind swift bank attacks," *Available: https://www.darkreading.com/perimeter/inside-the-north-korean-hacking-operation-behind-swift-bank-attacks–/d/d-id/1332969*, 2018.
[24] P. Coopers, "Operation cloud hopper," *PwC UK Cyber security and data privacy. https://www.pwc.co.uk/cyber-security/pdf/cloud-hopper-report-final-v4.pdf*, 2017.
[25] S. S. Silva, R. M. Silva, R. C. Pinto, and R. M. Salles, "Botnets: A survey," *Computer Networks*, vol. 57, no. 2, pp. 378–403, 2013.

File Fragment Classification with Focus on OLE and OOXML classes

K. Skračić*, F. Rukavina**, K. Miličić**, J. Petrović**, P. Pale**
* Ericsson Nikola Tesla d.d.
** Department of Electronic Systems and Information Processing
** University of Zagreb, Faculty of Electrical Engineering and Computing
Kristian.Skracic@ericsson.com, Fil.Rukavina@gmail.com, {Karlo.Milicic, Juraj.Petrovic, Predrag.Pale }@FER.hr

Abstract - Classification of file fragments is a crucial step in digital forensics and determining file types based on available data fragments. Currently explored file fragment classification methods other than forensic hand-examination rely on machine learning techniques. Those methods most commonly use features based on byte frequency distribution as inputs in artificial neural networks. In this paper, some new approaches to file fragment classification are explored. Older MS Office file format files (doc, ppt, and xls), and the new MS Office format (docx, pptx, and xlsx), which were previously shown to be difficult to differentiate between, were joined into two separate higher-level classes due to similarities in the included files' structure. Different approaches to specifically differentiating between subtypes in each of those two higher-level classes are further explored in the paper. The results suggest small increases in classification accuracy can be achieved using the proposed approach.

Keywords - file fragment classification; file type detection; OLE; OOXML; artificial neural network; feedforward neural network

I. INTRODUCTION

Classification of file fragments is an important step in digital forensics, particularly for file recovery. Efficient file classification is necessary for the reduction of the computational cost of recovery, while also increasing the accuracy of information extraction from both their contents and their metadata.

Early research heavily depended on the availability of the file extension and magic numbers to classify file fragments. This is only possible when the file fragments can be connected to the extracted metadata. In practice, it was shown that such information can easily be tampered with and is often not available in real-world scenarios.

In recent years, content-based file fragment classification algorithms have been developed to alleviate that issue, all of them trying to determine the file type based on just a part of the file bitstream. The usual approaches consist of taking hand-engineered features that are further utilized by a machine learning model such as k-NN [1], decision trees [1], support vector machines (SVM) [1]–[3], and neural networks [4]–[6].

Among those, feedforward neural networks have usually provided the best classification results, in some cases, with significant overlapping between certain file types. For example, the work by Vulinović et al. [4] shows that a significant overlapping of classification results was present in case of samples belonging to one of the three older Microsoft Office (pre-2003) files (doc, ppt, xls), as well as between samples belonging to one of the three newer Microsoft Office (post-2003) files (docx, pptx, xlsx). This should not come as a surprise, since doc, ppt, and xls files share the same inner structure referred to as Object Linking & Embedding (OLE), a proprietary Microsoft technology, while Microsoft's newer Office file types docx, pptx, and xlsx, also use the same inner structure consisting of xml files and other embedded multimedia files packed in a zip archive. In this paper, the authors attempt to exploit this fact to improve classification results first by joining the three OLE file types in a single higher-level class and the three OOXML file types in another higher-level class, and afterwards by distinguishing between them using features that were not useful in their differentiation from other classes, but were shown to be useful in differentiating between those specific file types.

The rest of this paper is organized as follows. Section II covers a brief overview and analysis of related research. Section III explains the newly explored approach, used features and how the dataset was constructed. It also gives a detailed description of the used artificial neural network (ANN) architectures. Section IV gives a detailed presentation of the experimental setup, followed by an analysis of the results. Section V describes conclusions and future work.

II. OVERVIEW

Previous research on file fragment classification has shown that machine learning methods can be used to determine the type of file the fragment belonged to [1]. One of the main challenges of file fragment classification is to identify the file type of a file fragment based on its content when no header or other information is available. Since file fragments belonging to a file header can contain magic numbers and other unique patterns that can identify the file fragment, they are excluded from the dataset as they introduce bias. Consequently, most previous research has excluded the fragments that were derived from the header or the footer of a file [7].

To foster reproducibility and future comparison of the results, most research on file fragment classification has been based on publicly available datasets. As stated by numerous researchers [8], comparing results that were obtained using different datasets could be highly misleading. At the time of writing the most widely used dataset for file fragment classification is the GovDocs1 dataset [9] which is freely available and used by many researchers [10].

A file fragment can be represented using N-grams as feature vectors. In case when N = 1, the unigram (or histogram) simply counts the number of occurrences of each byte in a file fragment. This representation is also referred to as Byte Frequency Distribution (BFD). Similarly, bigrams count the occurrence of a

sequence of two bytes. As there are 256 unique values in a single byte, bigrams have 65.536 possible patterns. Most research is focused on using unigrams and bigrams since the feature space is of manageable size. Using higher values of N may result in an extremely large number of sparsely populated features which makes it computationally intensive to train machine learning models.

File fragment classification methods that utilize N-gram vectors as features rely on recognizing specific patterns in file fragments. Previous research has shown that using BFD as feature vectors was successful at discovering patterns in some file types. The work by Fang et al. [3] shows that a Support Vector Machine (SVM) classifier was successfully trained and achieved an average prediction accuracy of 53.31% across 18 different file types.

A drawback of using N-Grams as feature vectors was shown on file types that do not have a set of patterns common across all or most of their fragments. For example, compressed and encrypted files do not have common patterns because their very nature is to suppress patterns. In compressed files, most patterns are removed for more efficient storage, while encrypted files are created to seem as random as possible to make cryptanalysis more difficult. Fragments coming from such files are called high entropy file fragments [11], [12]. Because these fragments are hard to classify, they are rarely used in existing research as their prediction accuracy is usually quite low. There are exceptions such as the work by Cox et al. [13], in which the authors chose to include file header information, thus introducing bias in the results.

For such file fragments to be recognized from other file fragments, previous research has utilized various statistical measures. The work by Penrose et al. [11] considers the use of various NIST randomness tests as features for classifying high entropy file fragments. Since high entropy file fragments contain no patterns, the premise is that their entropy or randomness measures could be used to improve classification accuracy. The study trains an artificial neural network with only randomness measures as features, no N-Grams were used. The study reported an accuracy score of 76% and 70% for encrypted a compressed fragment respectively. As the study only used encrypted and compressed, it is unclear if the features can discriminate between high entropy file fragments and fragments from other file types.

Previous research [4] has shown that N-grams can be used in conjunction with statistical entropy measures to successfully classify file fragments coming from compressed file fragments, as well as other file fragments. This paper implements and explores the idea of merging doc, ppt, and xls file types into a single OLE class, since all those files are based on the same OLE structure, and merging docx, pptx, and xlsx file types into the OOXML class. Those merges are motivated by the fact that those types are often mistaken for one other due to the similarities in their structure. If in the first step they would be merged in a general OLE and the general OOXML class, which should improve the prediction accuracy, while in the next step different features or machine learning models can be used to differentiate specifically between the three OLE subtypes and the three OOXML subtypes. Features or machine learning models that have not been successful enough in the classification of many different file types could, however, be useful in differentiating specifically between those subtypes of OLE and OOXML and are therefore worth exploring.

III. METHODOLOGY

A. Data representation and preprocessing

Data used for this research was generated from the GovDocs1 dataset, by randomly sampling 20000 samples of 512-byte blocks from file types containing at least 1000 samples in GovDocs1. The first and last block of every file was omitted not to include header and footer information. OLE and OOXML classes were created using equal percentages of samples from doc, ppt, and xls files in the first case – and labelling them as OLE class (in figures labeled *ole*), and equal percentages of docx, pptx, and xlsx classes, and labelling them as OOXML class (in figures labeled as *ooxml*). In total, 18 prediction classes or file types were used in the dataset: swf, pdf, html, txt, ps, jpg, rtf, gif, xml, png, gz, csv, doc, ppt, xlsx, docx, pptx and xlsx. As described earlier, OOXML class and OLE class were used to substitute for older and newer Microsoft Office file types. The training and validation data were split with a factor of 0.2. The unigrams were preprocessed to fit between the values [-1,1], which proved to be optimal for the approach described in section IV. Experimental setup and results.

B. Used models and aproaches

Several different machine learning models were evaluated and compared. Feedforward Neural Networks (FNNs) were expected to provide the best results, as it has been reported in the related literature, abut other models including Support Vector Machines (SVMs), Random Forests (RFs), Vector Logic Regression (VLR) and an ensemble of the three aforementioned methods. Some of those approaches have been used in other works as well [1]–[4].

IV. EXPERIMENTAL SETUP AND RESULTS

Using a FNN consisting of four layers of 256x1500x750x18 neurons with *relu* activation function [14], the average overall accuracy of 69.39% was achieved for the original 18 classes (not including general OLE and OOXML classes, but rather individual classes doc, ppt, xls, docx, pptx, xlsx, and other 12 file types). The corresponding confusion matrix averaged over 10 runs is presented in Table 1.

Table 1. Confusion matrix for 18 different classes before joining

	docx	pptx	xlsx	swf	doc	ppt	pdf	html	xls	txt	ps	jpg	rtf	gif	xml	png	gz	csv
docx	**48.4**	6.5	6.5	11.2	0.2	1.5	4.7	0.0	0.0	0.0	0.0	5.3	0.0	2.5	0.0	7.7	5.7	0.0
pptx	8.5	**37.4**	6.1	11.8	0.4	1.7	5.0	0.0	0.0	0.0	0.0	11.8	0.0	2.6	0.0	9.7	4.9	0.0
xlsx	1.9	1.4	**83.3**	5.5	0.0	0.2	2.8	0.0	0.1	0.0	0.0	0.7	0.0	0.4	0.0	1.9	1.7	0.0
swf	4.1	6.6	7.2	**46.6**	0.4	0.7	5.9	0.0	0.2	0.0	0.0	10.2	0.0	3.3	0.0	8.5	6.2	0.0
doc	1.8	1.8	0.8	2.5	**78.2**	3.1	1.3	0.0	2.0	0.2	0.0	2.7	0.2	0.6	1.0	2.1	1.8	0.0
ppt	5.8	6.8	4.0	9.0	5.0	**36.9**	3.9	0.0	1.1	0.0	0.1	10.6	0.0	2.0	0.0	9.4	5.4	0.0
pdf	2.9	3.9	8.9	13.5	0.3	0.8	**45.2**	0.1	0.1	0.5	0.8	6.2	0.0	4.0	0.1	6.4	6.3	0.0
html	0.0	0.0	0.0	0.0	1.3	0.0	0.0	**86.8**	0.0	5.6	0.1	0.0	0.1	0.0	6.1	0.0	0.0	0.0
xls	0.2	0.0	0.2	0.3	14.8	1.2	0.2	0.0	**81.9**	0.1	0.0	0.3	0.0	0.1	0.4	0.2	0.1	0.0
txt	0.0	0.0	0.0	0.0	0.6	0.0	0.0	1.1	0.0	**96.5**	0.1	0.0	0.1	0.0	0.9	0.0	0.0	0.8
ps	0.0	0.0	0.1	0.1	0.1	0.3	1.7	0.1	0.0	1.5	**95.5**	0.1	0.4	0.0	0.0	0.2	0.1	0.0
jpg	2.3	11.0	2.1	8.5	0.5	1.7	7.2	0.0	0.1	0.0	0.0	**56.3**	0.0	1.8	0.0	5.1	3.3	0.0
rtf	0.0	0.0	0.0	0.0	0.6	0.2	0.1	0.0	0.0	1.2	0.6	0.0	**96.9**	0.0	0.4	0.0	0.0	0.0
gif	1.2	1.8	2.5	3.6	0.3	0.1	2.7	0.0	0.0	0.0	0.0	2.8	0.0	**79.3**	0.0	3.3	2.5	0.0
xml	0.0	0.0	0.0	0.0	0.7	0.0	0.0	2.1	0.0	1.7	0.0	0.0	0.1	0.0	**95.4**	0.0	0.0	0.0
png	5.1	4.5	6.4	12.2	0.1	1.4	5.2	0.0	0.0	0.0	0.0	5.5	0.0	3.6	0.0	**47.2**	8.7	0.0
gz	5.8	4.5	8.2	14.9	0.1	1.1	6.4	0.0	0.0	0.0	0.0	4.8	0.0	4.1	0.0	12.2	**37.9**	0.0
csv	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.5	0.0	0.0	0.0	0.0	0.1	0.0	0.0	**99.2**

After substituting the three older MS Office file types with a general OLE class and the three newer MS Office file types with the general OOXML class using the same network and activation function, the overall accuracy of 73.06% was achieved for the 14 classes. The corresponding confusion matrix averaged over 10 runs is presented in Table 2.

Table 2. Confusion matrix after joining OLE and OOXML

	swf	pdf	html	txt	ps	jpg	rtf	gif	xml	png	gz	csv	OLE	OOXML
swf	**40.9**	9.5	0.0	0.0	0.0	15.0	0.0	2.7	0.0	11.0	11.7	0.0	1.9	7.4
pdf	8.7	**47.2**	0.1	0.2	1.2	11.2	0.1	2.8	0.0	10.1	10.1	0.0	1.6	6.7
html	0.0	0.1	**91.0**	3.8	0.2	0.1	0.1	0.0	4.3	0.0	0.0	0.2	0.3	0.0
txt	0.0	0.0	2.0	**95.4**	0.5	0.1	0.4	0.0	0.6	0.0	0.0	0.7	0.4	0.0
ps	0.1	0.9	0.0	0.3	**97.8**	0.1	0.3	0.1	0.1	0.1	0.1	0.0	0.2	0.1
jpg	5.7	6.8	0.0	0.0	0.1	**68.9**	0.0	1.0	0.0	5.8	4.5	0.0	1.6	5.7
rtf	0.0	0.1	0.4	0.5	0.6	0.0	**97.5**	0.0	0.2	0.0	0.0	0.0	0.8	0.0
gif	3.0	5.0	0.0	0.0	0.0	3.5	0.0	**77.7**	0.0	3.1	3.3	0.0	0.9	3.6
xml	0.0	0.1	4.0	2.3	0.1	0.0	0.3	0.0	**92.9**	0.0	0.0	0.0	0.3	0.0
png	7.5	7.5	0.0	0.0	0.0	8.7	0.0	2.7	0.0	**52.9**	11.7	0.0	1.5	7.6
gz	8.9	9.6	0.0	0.0	0.1	8.7	0.0	3.5	0.0	14.8	**43.8**	0.0	1.7	9.0
csv	0.0	0.0	0.1	0.5	0.0	0.0	0.0	0.0	0.1	0.0	0.0	**99.2**	0.1	0.0
OLE	2.9	3.0	0.4	0.4	0.3	6.3	0.0	0.7	0.3	4.9	3.2	0.1	**73.9**	3.7
OOXML	8.1	8.5	0.0	0.0	0.0	12.5	0.0	2.7	0.0	11.8	10.1	0.0	2.4	**44.0**

The higher accuracy when using the merged OLE and OOXML classes is meaningless unless the merged subclasses, the three OLE and three OOXML classes, can be differentiated between more efficiently than before. To validate that assumption, separate machine learning models were trained to differentiate between doc, ppt, and xls file types within the OLE class, and between docx, pptx, and xlsx file types in the OOXML class. Several approaches have been evaluated for this, including FNN, SVM, RF, and VLR. Using soft voting together in an ensemble the classifiers achieve an accuracy of 87% for OLE subclasses. which is greater than the FNN on the same problem (accuracy of 83%). OOXML classes show a smaller improvement of 75% using an ensemble and 73% by using a FNN. SVMs showed the best results with the regularization parameter C set to 5, and the gamma is set to the inverse of the number of features. The SVM method being used is one vs all. The RF classifier showed the best results at 50 estimators, and the LR classifier shows little improvement with the modification of parameters. The results of the ensemble classifier are presented in Table 3.

Table 3. Differentiating subclasses of OLE and OOXML

OLE	doc	ppt	xls
doc	**80.**	16.	4.
ppt	5.	**94.**	1.
xls	12.	12.	**86.**

OOXML	docx	pptx	xlsx
docx	**71.**	23.	6.
pptx	27.	**67.**	6.
xlsx	3.	4.	**93.**

Finally, a cascade of classifiers using OLE and OOXML classes in the first step using the best-evaluated architecture, followed by the model differentiating between OLE subclasses, and OOXML subclasses achieved an average accuracy of 72.66%.

Table 4. Confusion matrix for 18 different classes after joining OLE and OOXMLS

	docx	pptx	xlsx	swf	doc	ppt	pdf	html	xls	txt	ps	jpg	rtf	gif	xml	png	gz	csv
docx	**61.0**	6.0	2.0	8.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	5.0	0.0	4.0	0.0	9.0	9.0	0.0
pptx	4.0	**57.0**	2.0	6.0	0.0	2.0	3.0	0.0	0.0	0.0	0.0	14.0	0.0	5.0	0.0	8.0	8.0	0.0
xlsx	1.0	1.0	**78.0**	5.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	3.0	7.0	0.0
swf	4.1	6.6	7.2	**46.6**	0.4	0.7	5.9	0.0	0.2	0.0	0.0	10.2	0.0	3.3	0.0	8.5	6.2	0.0
doc	1.0	2.0	0.0	1.0	**82.0**	2.0	1.0	0.0	2.0	0.0	0.0	3.0	0.0	1.0	0.0	2.0	3.0	0.0
ppt	3.0	6.0	2.0	5.0	5.0	**62.0**	3.0	0.0	2.0	0.0	0.0	12.0	0.0	3.0	0.0	9.0	10.0	0.0
pdf	2.9	3.9	8.9	13.5	0.3	0.8	**45.2**	0.1	0.1	0.5	0.8	6.2	0.0	4.0	0.1	6.4	6.3	0.0
html	0.0	0.0	0.0	0.0	1.3	0.0	0.0	**86.8**	0.0	5.6	0.1	0.0	0.1	0.0	6.1	0.0	0.0	0.0
xls	0.0	0.0	0.0	0.0	13.0	1.8	0.0	0.0	**85.0**	0.0	0.0	0.0	0.0	0.1	0.0	0.2	0.0	0.0
txt	0.0	0.0	0.0	0.0	0.6	0.0	0.0	1.1	0.0	**96.5**	0.1	0.0	0.1	0.0	0.9	0.0	0.0	0.8
ps	0.0	0.0	0.1	0.1	0.1	0.3	1.7	0.1	0.0	1.5	**95.5**	0.1	0.4	0.0	0.0	0.2	0.1	0.0
jpg	2.3	11.0	2.1	8.5	0.5	1.7	7.2	0.0	0.1	0.0	0.0	**56.3**	0.0	1.8	0.0	5.1	3.3	0.0
rtf	0.0	0.0	0.0	0.0	0.6	0.2	0.1	0.0	0.0	1.2	0.6	0.0	**96.9**	0.0	0.4	0.0	0.0	0.0
gif	1.2	1.8	2.5	3.6	0.3	0.1	2.7	0.0	0.0	0.0	0.0	2.8	0.0	**79.3**	0.0	3.3	2.5	0.0
xml	0.0	0.0	0.0	0.0	0.7	0.0	0.0	2.1	0.0	1.7	0.0	0.0	0.1	0.0	**95.4**	0.0	0.0	0.0
png	5.1	4.5	6.4	12.2	0.1	1.4	5.2	0.0	0.0	0.0	0.0	5.5	0.0	3.6	0.0	**47.2**	8.7	0.0
gz	5.8	4.5	8.2	14.9	0.1	1.1	6.4	0.0	0.0	0.0	0.0	4.8	0.0	4.1	0.0	12.2	**37.9**	0.0
csv	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.5	0.0	0.0	0.0	0.0	0.1	0.0	0.0	**99.2**

V. CONCLUSIONS

In this paper, the approach of merging OLE and OOXML file types into unique classes has improved validation accuracy to 74% compared to 70% (without the proposed approach). The increase in prediction accuracy is relatively small, yet meaningful, and it does not increase classification time.

Several machine learning techniques have been explored for differentiating specifically between doc, ppt, and xls file types, and between docx, pptx, and xlsx file types. The results regarding encoded and archived files suggest this approach is feasible in their classification as well, and that classifier ensembles could be used to slightly increase prediction outcomes.

REFERENCES

[1] I. Ahmed, K.-S. Lhee, H.-J. Shin, and M.-P. Hong, "Fast Content-Based File Type Identification," in *Advances in Digital Forensics VII*, 2011, pp. 65–75.

[2] S. Fitzgerald, G. Mathews, C. Morris, and O. Zhulyn, "Using NLP techniques for file fragment classification," *Digit. Investig.*, vol. 9, pp. S44–S49, Aug. 2012, doi: 10.1016/j.diin.2012.05.008.

[3] F. Wang, T.-T. Quach, J. Wheeler, J. B. Aimone, and C. D. James, "Sparse Coding for N-Gram Feature Extraction and Training for File Fragment Classification," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 10, pp. 2553–2562, Oct. 2018, doi: 10.1109/TIFS.2018.2823697.

[4] K. Vulinović, L. Ivković, J. Petrović, K. Skračić, and P. Pale, "Neural Networks for File Fragment Classification," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2019, pp. 1194–1198, doi: 10.23919/MIPRO.2019.8756878.

[5] Y. Wang, Z. Su, and D. Song, "File Fragment Type Identification with Convolutional Neural Networks," in *Proceedings of the 2018 International Conference on Machine Learning Technologies*, New York, NY, USA, 2018, pp. 41–47, doi: 10.1145/3231884.3231889.

[6] Q. Chen *et al.*, "File Fragment Classification Using Grayscale Image Conversion and Deep Learning in Digital Forensics," in *2018 IEEE Security and Privacy Workshops (SPW)*, May 2018, pp. 140–147, doi: 10.1109/SPW.2018.00029.

[7] I. Ahmed, K.-S. Lhee, H. Shin, and M. Hong, “Fast file-type identification,” Mar. 2010, doi: 10.1145/1774088.1774431.
[8] S. Axelsson, “Using Normalized Compression Distance for Classifying File Fragments,” in *2010 International Conference on Availability, Reliability and Security*, Feb. 2010, pp. 641–646, doi: 10.1109/ARES.2010.100.
[9] S. Garfinkel, P. Farrell, V. Roussev, and G. Dinolt, “Bringing science to digital forensics with standardized forensic corpora,” *Digit. Investig.*, vol. 6, pp. S2–S11, Sep. 2009, doi: 10.1016/j.diin.2009.06.016.
[10] C. Grajeda, F. Breitinger, and I. Baggili, “Availability of datasets for digital forensics – And what is missing,” *Digit. Investig.*, vol. 22, pp. S94–S105, Aug. 2017, doi: 10.1016/j.diin.2017.06.004.
[11] P. Penrose, R. Macfarlane, and W. J. Buchanan, “Approaches to the classification of high entropy file fragments,” *Digit. Investig.*, vol. 10, no. 4, pp. 372–384, Dec. 2013, doi: 10.1016/j.diin.2013.08.004.
[12] V. Roussev and S. L. Garfinkel, “File Fragment Classification-The Case for Specialized Approaches,” in *2009 Fourth International IEEE Workshop on Systematic Approaches to Digital Forensic Engineering*, May 2009, pp. 3–14, doi: 10.1109/SADFE.2009.21.
[13] J. A. Cox, C. D. James, and J. B. Aimone, “A Signal Processing Approach for Cyber Data Classification with Deep Neural Networks,” *Procedia Comput. Sci.*, vol. 61, pp. 349–354, Jan. 2015, doi: 10.1016/j.procs.2015.09.156.
[14] X. Glorot, A. Bordes, and Y. Bengio, “Deep Sparse Rectifier Neural Networks,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Jun. 2011, pp. 315–323, Accessed: Jun. 25, 2020. [Online]. Available: http://proceedings.mlr.press/v15/glorot11a.html.

AI safety: state of the field through quantitative lens

Mislav Juric*, Agneza Sandic ** and Mario Brcic**
* Student at University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
** University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
mislav.juric2@fer.hr, agneza.sandic@fer.hr, mario.brcic@fer.hr

Abstract - Last decade has seen major improvements in the performance of artificial intelligence which has driven wide-spread applications. Unforeseen effects of such mass-adoption has put the notion of AI safety into the public eye. AI safety is a relatively new field of research focused on techniques for building AI beneficial for humans. While there exist survey papers for the field of AI safety, there is a lack of a quantitative look at the research being conducted. The quantitative aspect gives a data-driven insight about the emerging trends, knowledge gaps and potential areas for future research. In this paper, bibliometric analysis of the literature finds significant increase in research activity since 2015. Also, the field is so new that most of the technical issues are open, including: explainability and its long-term utility, and value alignment which we have identified as the most important long-term research topic. Equally, there is a severe lack of research into concrete policies regarding AI. As we expect AI to be the one of the main driving forces of changes, AI safety is the field under which we need to decide the direction of humanity's future.

Keywords - AI safety; technical AI safety; research; surveys; bibliometrics

I. INTRODUCTION

The field of Artificial Intelligence (AI) safety is concerned with answering a very important question (along with its variations) - „How can we make AI safe for humans?". As the field of AI safety evolved over the years, many developments have been published on the topics related to AI safety. Although there are a lot of survey papers giving an overview of different aspects of the field of AI safety, there is a lack of quantitative insight into the state of the field of AI safety. Namely, there is a lack of data which tells a story about current and past trends and gives us empirical evidence to where research contributions would be most valuable.

In this paper we shall divide AI safety in the following hierarchy of sub-fields:

1. technical AI safety - deals with the technical issues of achieving safety and utility. It is subdivided by SRA classification [1]
   a) specification (S) – defines the purpose of the system
   b) robustness (R) – designing systems withstanding perturbations
   c) assurance (A) – monitoring, understanding, and controlling system during operation
2. AI ethics – mainly deals with the questions of moral responsibility and utility for the humans.
3. AI policy – deals with the questions how legal and governance systems need to be setup with respect to the new AI technologies (not only on national, but also on supranational level)

This paper is structured as follows. We shall explain our research methodology in section II, present our results and elaborate upon them in section III. In section IV we give our view on the current and possible future developments in the field of AI safety. Finally, in section V we shall give our conclusion and state the limitations to our research.

II. RESEARCH METHODOLOGY AND SCOPE

In this paper, we use bibliometrics as the basis of our research. In AI safety there are plenty of surveys (see Table 1), but most of them are specific to their certain sub-fields. There is a general survey [2], but it is of enumerative and descriptive nature, while we aim to supplement it with quantitative overview.

We have used online databases to identify indexed work sampled across different subfields of AI safety. We used the following databases: SCOPUS, Web of Science (WoS), and Google Scholar (GS). Since the area of AI safety is rather new, number of published work is smaller than for the well-established areas. This is especially the case for WoS. GS offers high-volume of work, including important pre-print sources, but the quality varies substantially and the result exploration seems to be limited. SCOPUS offers good trade-off between the volume and quality, paired with flexible searching and result exploration. So, we have manually fine-tuned our search-queries to SCOPUS and then we have used them across all three databases. We have fine-tuned queries until we got purity of 90% over returned results – purity established empirically by sampling across the results. We did our best to look into as many as subfields as possible, but we leave improvements, with respect to covered sub-fields, to the future work.

We have searched the following areas: AI ethics, AI policy, robustness (R), explainability and intepretability (A), fairness (S), value alignment (S), reward hacking (S), interruptibility (A), safe exploration (R), distributional shift (R), and AI privacy (A). Other areas such as verifiability have proven to be elusive for search queries in the terms of results' purity and relevance. The respective search queries for each selected area are given in the appendix VI.

Table 1 AI safety surveys

Subject	References
General AI safety	[2]–[9]
AI ethics	[10]–[18]
AI policy	[19]–[24]
Interpretability/XAI	[25]–[32]
Adversarial robustness	[33]–[37]
Fairness/bias	[38]–[40]
Value alignment	[41]–[44]
Safe exploration	[45]
interruptibility	[46]

Identified documents have the following distribution: conference papers (47.71%), journal articles (39.89%), reviews (3.89%), books (3.57%), book chapters (3.11%), and other (2.83%). We chose to focus on papers from 1985 till 2019 with the relevant data retrieved on January 26, 2020.

III. AI SAFETY ARTICLES

In Figure 1 we can see the trends of growth in AI ethics and AI policy. AI ethics has seen steady growth since 2003, with visible explosion in interest since 2010. On the other hand, AI policy has had no significant amount of work until 2018 since when it is experiencing strong growth last two years.

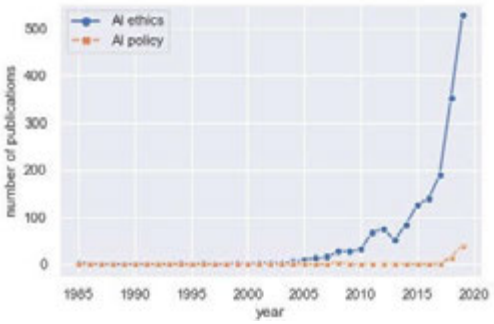


Figure 1 Number of papers in AI ethics and policy papers published each year

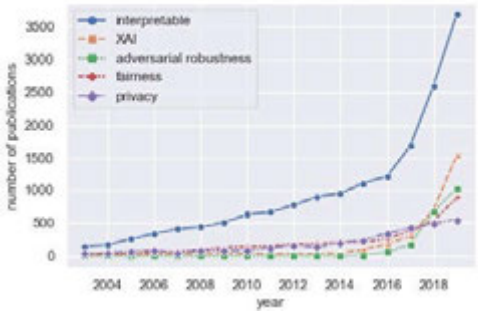


Figure 2 Number of papers in technical AI safety, high volume topics

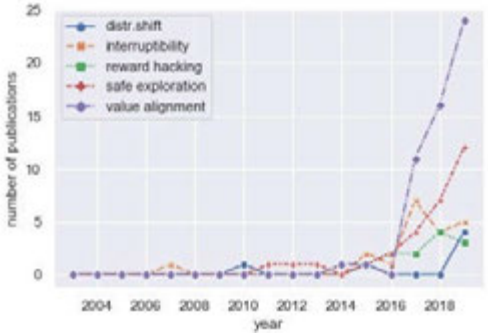


Figure 3 Number of papers in technical AI safety, low volume topics

Figure 2 and Figure 3 show the change of popularity for different subfields of technical AI safety over the last 20 years, grouped by their popularity level. The whole field of AI safety is seeing strong growth which is mainly driven by strong growing sub-fields. **Interpretability (with explainable AI (XAI))** is the *single most* important growth generator for the field. *Strong growth* is shown by the fields of AI ethics, adversarial robustness and by the smaller volume topics of value alignment and safe exploration. These fields are driven mainly by the near-term applications in diversity of areas, such as transportation, medicine, biology, robotics, etc. Slight growth includes topics of fairness and privacy. Finally, the topics of safe exploration, distribution shift, interruptibility, and reward hacking seem to be emerging and are likely to become more intense venues in the future. Interruptibility and reward hacking are more focused on long-term research goals and do not get as much attention as other topics. These are areas that could be covered by non-publication decisions by some of the research organizations.

rank	Journal	#	Journal	#
1	Expert systems with applications	249	AI and Society	32
2	IEEE Access	237	Ethics and Information tech.	32
3	Neurocomputing	200	Futures	17
4	Information sciences	199	Minds and machines	16
5	Applied Soft Computing journal	154	Phil.and Tech./ Science and Engineering Ethics	14

Table 2 Published articles per journal

In Table 2 we have identified top 5 journals for each technical AI safety and AI ethics respectively. We must emphasize that a considerable amount of interesting research and ideas in the area is not published in peer reviewed outlets. At best, they are published at pre-print services and otherwise they are published in blogs and their respective comments. This makes idea sharing harder and reinventing more likely. Some organizations, such as MIRI, have decided not to publish most of their work due to security reasons.

AI governance had such a small volume of work that journal ranking makes no sense. Some of the journals in AI policy are: „Computer Law And Security Review", „Contemporary Security Policy", and „Ethics And International Affairs".

Tool/Technique	#
Machine Learning	3052
Classification	2089
Deep Learning	1910
Forecasting	1289
Feature Extraction	1112
Decision Trees	1017
Fuzzy Systems	929
Optimization	899
Clustering Algorithms	763
Genetic Algorithms	738
Knowledge Based Systems	636

Table 3 Popular AI safety research tools and techniques

Table 3 summarizes most used tools and techniques in research based on their frequency as an article keyword. Machine learning tops the list, due to the latest surge of applications and proliferation. Classification seems to be the most studied problem setting, followed by forecasting. Deep learning is at the top of algorithmic approaches, followed by the decision trees which are more transparent model. Optimization is a tool used in learning systems and it can aim at different criteria. Genetic algorithms are an approach to optimization of hard problems. Fuzzy systems and knowledge based systems seem to be emerging in popularity, starting to be proposed as supplementary techniques to deep learning which bring complementary benefits where deep learning seems lacking.

IV. Recommendations for Future Research

Generally, we welcome more concrete empirical work which would bring much needed information to fuel the future development. This calls for necessarily *multidisciplinary* work with both computational experimentation (such as [47]), and real-world testing (such as [48], [49]). Creation of many rich and diverse simulation environments and benchmarks would improve both system building and safety testing, since learning algorithms are data-hungry and real-world has too small experience bandwidth.

A. Explainability

One of the most important open problems in explainability is that there is no agreement on what an explanation is. Some works define decision tree, set of rules or an image as good explanation [31], with most of the work appealing to the mere intuition. Evaluation of comprehensibility of explanations to humans is underexplored [26]. There is still no algorithm that provides both high accuracy and explainability. In that regard, new hybrid techniques hold potential to achieve more effective explanations [32]. Looking further, the utility of explainability to the safety in the long-term is unclear since such approaches lack guarantees for adversarial schemes. Namely, possibly harmfully incorrect, but plausible explanations can be generated. Moreover, the size and dynamic of gap between the true and the most incorrect plausible explanation are interesting questions. Also, the nature of limits to the explainability of advanced concepts (for example in future science) to humans is not yet understood. Overstepping such limits makes explanations too far removed from the reality and their utility fades out.

B. Value alignment

Work on reinforcement learning (RL) and learning a reward function from actions and preferences will be very important for advancing the field of value alignment [2]. More advanced than value learning, value discovery through aligned algorithms could enable finding better reward functions that unlock new opportunities. It is more advanced prospect than value learning. Researching into reward corruption and side effects of optimizing for a goal which doesn't capture human values fully is just at the beginning. Currently, the most interesting work are approaches combining recursive factorization and bootstrapping for scalable reward learning and value alignment both in cooperative [50], [51] and adversarial [52] settings. There are many open questions about feasibility of such ideas with respect to error bounding, automated factorization process, validity of assumptions, etc. Mesa-optimizer [53] is an important concept, especially for the long-term research. It introduces multiple levels of value alignment within learned optimization. However, their relevance to the current practice is yet to be demonstrated.

C. AI policy

Transparency, accountability, reliability, security, corrigibility, interpretability, value specification, ability to limit capability, and performance and safety guarantees for particular AI systems are all important for future wide-spread applications and should be important parts and aims of future policies [21], [24]. Regulating research related to AI seems to be a possible long-term approach to regulating AI developments, but not at the cost of restricting scientific progress [5]. AI regulation could work globally only if there is concensus between the major AI research organizations. Although there are many

papers related to AI governance, there is a severe lack of concrete AI policy suggestions.

D. Corrigibility

Corrigibility, reasoning that reflects that an agent is incomplete and potentially flawed in a dangerous ways, is something that needs to be worked on [54]. Safety measures which are easily integrated within the AI development environment seems like something which would be tremendously valuable for improving AI safety. One promising method of improving AI safety in the short-term, while developing, is containment through virtualization [55].

E. Safe exploration and distr. shift

Preventing catastrophic mistakes from occurring while training a reinforcement learning model is a non-negotiable need when the system interacts with the real world and not within a simulated environment [8]. Work in detecting and overriding an agent's action when it seems too dangerous seems like a good approach to reduce the number of catastrophic mistakes [2]. Proposing new conceptual solutions to the problem of safe exploration and distributional shift seems to be something that could bring a lot of value, but testing and/or extending existing solutions may be valuable as well. Hybridization of symbolic and sub-symbolic approaches might be a valid approach here.

F. Adversarial robustness

Deep neural network models are highly vulnerable to adversarial attacks, which curtails their applications. There are special methods which reduce success rates of different types of adversarial attacks, but there are no general defensive methods successful against all attacks [34]. Open questions include why do the adversarial examples exist in the first place and why are they transferable [36]. Security verification of models to adversarial attacks is an open research challenge [35].

G. AI ethics

Further work into human trust is necessary, especially from the aspect of human-AI interaction [13]. In applications, privacy concerns need to be taken seriously if we are to construct AI which is ethical [18], [56]. Ethics vary across the globe and evolve over time [12], [48], [49]. More research into human moral preferences is important to guide our development and policies.

V. Conclusions and Limitations of the Research

In this paper, we have surveyed the field of AI safety through quantiative lens. We have found various trends in the field of AI safety. We identified interpretability and explainability as strongest research topic in the near-term, but we have raised the question of utility in the long-term since such approaches lack guarantees for adversarial schemes [25]. Also, there are limits to explaining increasingly advanced concepts (for example, in future science) in a comprehensible way to humans. After overstepping such limits, the utility of explanations would fade out. Robustness-based topics are growing in importance with the incoming technical applications. AI policy is seeing its first greater contributions, but there is a severe lack of concrete policies. However, we would like to point out value alignment as the most important subfield in the long-term as it continues its sudden growth from relative obscurity. It is the most promising research direction for achieving the coexistence and cooperation with the computationally more capable agents than ourselves. If their goals are aligned with ours, they have no incentive to harm. In the case of misalignment, it is hard to have general guarantees through other approaches against incentivized, computationally superior agents.

This research has limitations, some of them are: underestimation/overestimation due to the search queries, non-publication bias of some research fields in peer-reviewed outlets, bias due to the increase of number of publishing outlets, reliance on the descriptive statistics to derive facts and insights about AI safety as we did not read all of the covered papers.

Literature

[1] "Building safe artificial intelligence: specification, robustness, and assurance." [Online]. Available: https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1. [Accessed: 27-Jan-2020].

[2] T. Everitt, G. Lea, and M. Hutter, "AGI Safety Literature Review," *ArXiv180501109 Cs*, May 2018.

[3] R. Yampolskiy and J. Fox, "Safety Engineering for Artificial General Intelligence," *Topoi*, vol. 32, no. 2, pp. 217–226, Oct. 2013, doi: 10.1007/s11245-012-9128-9.

[4] P. J. Scott and R. V. Yampolskiy, "Classification Schemas for Artificial Intelligence Failures," *ArXiv190707771 Cs*, Jul. 2019.

[5] K. Sotala and R. V. Yampolskiy, "Responses to catastrophic AGI risk: a survey," *Phys. Scr.*, vol. 90, no. 1, p. 018001, Dec. 2014, doi: 10.1088/0031-8949/90/1/018001.

[6] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine Learning Testing: Survey, Landscapes and Horizons," *ArXiv190610742 Cs Stat*, Dec. 2019.

[7] Y. K. Dwivedi *et al.*, "Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *Int. J. Inf. Manag.*, p. 101994, Aug. 2019, doi: 10.1016/j.ijinfomgt.2019.08.002.

[8] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," *ArXiv160606565 Cs*, Jul. 2016.

[9] S. Russell, D. Dewey, and M. Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence," *AI Mag.*, vol. 36, no. 4, pp. 105–114, Dec. 2015, doi: 10.1609/aimag.v36i4.2577.

[10] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data Soc.*, vol. 3, no. 2, p. 2053951716679679, Dec. 2016, doi: 10.1177/2053951716679679.

[11] J.-A. Cervantes, S. López, L.-F. Rodríguez, S. Cervantes, F. Cervantes, and F. Ramos, "Artificial Moral Agents: A Survey of the Current Status," *Sci. Eng. Ethics*, Nov. 2019, doi: 10.1007/s11948-019-00151-x.

[12] A. Hagerty and I. Rubinov, "Global AI Ethics: A Review of the Social Impacts and Ethical Implications of Artificial Intelligence," *ArXiv190707892 Cs*, Jul. 2019.

[13] B. W. Israelsen and N. R. Ahmed, "'Dave...I can assure you ...that it's going to be all right ...' A Definition, Case for, and Survey of Algorithmic Assurances in Human-Autonomy Trust Relationships," *ACM Comput. Surv. CSUR*, vol. 51, no. 6, pp. 113:1–113:37, Jan. 2019, doi: 10.1145/3267338.

[14] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nat. Mach. Intell.*, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.

[15] J. Morley, C. Machado, C. Burr, J. Cowls, M. Taddeo, and L. Floridi, "The Debate on the Ethics of AI in Health Care: A Reconstruction and Critical Review," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3486518, Nov. 2019.

[16] B. S. Barn, "Mapping the public debate on ethical concerns: algorithms in mainstream media," *J. Inf. Commun. Ethics Soc.*, vol. ahead-of-print, no. ahead-of-print, Jan. 2019, doi: 10.1108/JICES-04-2019-0039.

[17] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Sci. Eng. Ethics*, pp. 1–28, Dec. 2019, doi: 10.1007/s11948-019-00165-5.

[18] S. Spiekermann, J. Korunovska, and M. Langheinrich, "Inside the Organization: Why Privacy and Security Engineering Is a Challenge for Engineers," *Proc. IEEE*, vol. 107, no. 3, pp. 600–615, Mar. 2019, doi: 10.1109/JPROC.2018.2866769.

[19] A. Ramamoorthy and R. Yampolskiy, "Beyond MAD?: The race for artificial general intelligence," *ITU J. ICT Discov.*, vol. 2018, no. 1, pp. 77–84, Mar. 2018, doi: http://handle.itu.int/11.1002/pub/812a0228-en.

[20] C. Coglianese and D. Lehr, "Transparency and Algorithmic Governance," *Adm. Law Rev.*, vol. 71, p. 1, 2019.

[21] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi, "Artificial Intelligence and the 'Good Society': the US, EU, and UK approach," *Sci. Eng. Ethics*, vol. 24, no. 2, pp. 505–528, Apr. 2018, doi: 10.1007/s11948-017-9901-7.

[22] M. Brundage and J. Bryson, "Smart Policies for Artificial Intelligence," *ArXiv160808196 Cs*, Aug. 2016.

[23] C. Cath, "Governing artificial intelligence: ethical, legal and technical opportunities and challenges," *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 376, no. 2133, p. 20180080, Nov. 2018, doi: 10.1098/rsta.2018.0080.

[24] A. Dafoe, "AI Governance: A Research Agenda." Governance of AI Program, Future of HumanityInstitute, University of Oxford, 2018.

[25] R. V. Yampolskiy, "Unexplainability and Incomprehensibility of Artificial Intelligence," *ArXiv190703869 Cs*, Jun. 2019.

[26] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 0210–0215, doi: 10.23919/MIPRO.2018.8400040.

[27] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[28] J. Townsend, T. Chaton, and J. M. Monteiro, "Extracting Relational Explanations From Deep Neural Networks: A Survey From a Neural-Symbolic Perspective," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15, 2019, doi: 10.1109/TNNLS.2019.2944672.

[29] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[30] A. Preece, "Asking 'Why' in AI: Explainability of intelligent systems – perspectives and challenges," *Intell. Syst. Account. Finance Manag.*, vol. 25, no. 2, pp. 63–72, 2018, doi: 10.1002/isaf.1422.

[31] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv. CSUR*, vol. 51, no. 5, pp. 93:1–93:42, Aug. 2018, doi: 10.1145/3236009.

[32] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," *Auton. Agents Multi-Agent Syst.*, vol. 33, no. 6, pp. 673–705, Nov. 2019, doi: 10.1007/s10458-019-09408-y.

[33] N. Akhtar and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018, doi: 10.1109/ACCESS.2018.2807385.

[34] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial Attacks and Defences: A Survey," *ArXiv181000069 Cs Stat*, Sep. 2018.

[35] X. Wang, J. Li, X. Kuang, Y. Tan, and J. Li, "The security of machine learning in an adversarial setting: A survey," *J. Parallel Distrib. Comput.*, vol. 130, pp. 12–23, Aug. 2019, doi: 10.1016/j.jpdc.2019.03.003.

[36] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019, doi: 10.1109/TNNLS.2018.2886017.

[37] J. Zhang and C. Li, "Adversarial Examples: Opportunities and Challenges," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–16, 2019, doi: 10.1109/TNNLS.2019.2933524.

[38] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ArXiv190809635 Cs*, Sep. 2019.

[39] X. Zhang and M. Liu, "Fairness in Learning-Based Sequential Decision Algorithms: A Survey," *ArXiv200104861 Cs*, Jan. 2020.

[40] D. L. Coates and A. Martin, "An instrument to evaluate the maturity of bias governance capability in artificial intelligence projects," *IBM J. Res. Dev.*, vol. 63, no. 4/5, pp. 7:1-7:15, Jul. 2019, doi: 10.1147/JRD.2019.2915062.

[41] D. Abel, J. MacGlashan, and M. L. Littman, "Reinforcement Learning as a Framework for Ethical Decision Making," in *AAAI Workshop: AI, Ethics, and Society*, 2016.

[42] J. Taylor, E. Yudkowsky, P. LaVictoire, and A. Critch, "Alignment for Advanced Machine Learning Systems," 2016.

[43] T. Arnold, D. Kasenberg, and M. Scheutz, "Value Alignment or Misalignment - What Will Keep Systems Accountable?," in *AAAI Workshops*, 2017.

[44] T. LaCroix and Y. Bengio, “Learning from Learning Machines: Optimisation, Rules, and Social Norms,” *ArXiv200100006 Cs Stat*, Dec. 2019.

[45] J. García, Fern, and O. Fernández, “A Comprehensive Survey on Safe Reinforcement Learning,” *J. Mach. Learn. Res.*, vol. 16, no. 42, pp. 1437–1480, 2015.

[46] N. Soares, B. Fallenstein, S. Armstrong, and E. Yudkowsky, “Corrigibility,” in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[47] J. Leike *et al.*, “AI Safety Gridworlds,” *ArXiv171109883 Cs*, Nov. 2017.

[48] E. Awad *et al.*, “The Moral Machine experiment,” *Nature*, vol. 563, no. 7729, pp. 59–64, Nov. 2018, doi: 10.1038/s41586-018-0637-6.

[49] E. Awad, S. Dsouza, A. Shariff, I. Rahwan, and J.-F. Bonnefon, “Universals and variations in moral decisions made in 42 countries by 70,000 participants,” *Proc. Natl. Acad. Sci.*, Jan. 2020, doi: 10.1073/pnas.1911517117.

[50] P. Christiano, B. Shlegeris, and D. Amodei, “Supervising strong learners by amplifying weak experts,” *ArXiv181008575 Cs Stat*, Oct. 2018.

[51] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, “Scalable agent alignment via reward modeling: a research direction,” *ArXiv181107871 Cs Stat*, Nov. 2018.

[52] G. Irving, P. Christiano, and D. Amodei, “AI safety via debate,” *ArXiv180500899 Cs Stat*, Oct. 2018.

[53] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant, “Risks from Learned Optimization in Advanced Machine Learning Systems,” *ArXiv190601820 Cs*, Jun. 2019.

[54] N. Soares and B. Fallenstein, “Agent foundations for aligning machine intelligence with human interests: a technical research agenda,” in *The Technological Singularity*, Springer, 2017, pp. 103–125.

[55] J. Babcock, J. Kramar, and R. V. Yampolskiy, “Guidelines for Artificial Intelligence Containment,” *ArXiv170708476 Cs*, Jul. 2017.

[56] L. Rothenberger, B. Fabian, and E. Arunov, “RELEVANCE OF ETHICAL GUIDELINES FOR ARTIFICIAL INTELLIGENCE – A SURVEY AND EVALUATION,” *Res.--Prog. Pap.*, May 2019.

VI. APPENDIX

The queries from the Table 4 were used for searching through all the fields. The only exception is the query for AI privacy which was used only on the fields: title, abstract, and keywords.

Table 4 Search queries used for bibliometric research

Topic	**Search query**
explainability	("artificial intelligence" OR "machine learning" OR "AI") AND "explainable"
interpretability	("interpretable" OR "interpretability") AND "artificial intelligence"
<general>	"AI safety" OR "safe AI"
Adversarial robustness	"adversarial examples"
AI ethics	"AI ethics" OR "machine ethics" OR "friendly AI" OR "good AI" OR ("superintelligence" AND "risk") OR ("existential risk" AND "AI")
AI policy	"AI policy" OR "AI governance"
fairness	("AI" OR "algorithmic") AND ("bias" OR "fairness" OR "discrimination") AND "ethics"
Interruptibility&corrigibility	"AI" AND ("interruptibility" OR "corrigibility") AND "risk"
Safe exploration	"AI" AND "safe exploration"
Distributional shift	"Artificial intelligence" AND "distributional shift" AND ("generalization" OR "safety")
Reward hacking	("reward hacking" OR "wireheading" OR "reward tampering" OR "reward gaming") AND "AI"
AI privacy*	"privacy" AND ("AI" OR "artificial intelligence")
Value alignment	("value alignment" OR "alignment problem") AND "artificial" AND "ethics"

miproBIS

International Conference on

BUSINESS INTELLIGENCE SYSTEMS

Steering Committee

Chairs: Boris Vrdoljak, University of Zagreb, Croatia
Mirta Baranović, University of Zagreb, Croatia
Matteo Golfarelli, University of Bologna, Cesena, Italy
Mihaela Vranić, University of Zagreb, Croatia

Program Committee

Members: Marko Banek, University of Zagreb, Croatia
Mirta Baranović, University of Zagreb, Croatia
Bartosz Bebel, Poznan University of Technology, Poland
Ladjel Bellatreche, University of Poitiers, France
Ljiljana Brkić, University of Zagreb, Croatia
Alfredo Cuzzocrea, University of Calabria, Italy
Todd Eavis, Concordia University, Montreal, Canada
Enrico Gallinucci, University of Bologna, Cesena, Italy
Dragan Gamberger, Ruđer Bošković Institute, Zagreb, Croatia
Paolo Garza, Polytechnic University of Turin, Italy
Matteo Golfarelli, University of Bologna, Cesena, Italy
Igor Mekterović, University of Zagreb, Croatia
Damir Pintar, University of Zagreb, Croatia
Sandro Skansi, University of Zagreb, Croatia
Zoran Skočir, University of Zagreb, Croatia
A Min Tjoa, Vienna University of Technology, Austria
Mladen Varga, University of Zagreb, Croatia
Mihaela Vranić, University of Zagreb, Croatia
Boris Vrdoljak, University of Zagreb, Croatia

INVITED PAPER

The Linked Data Enterprise as Enabler for both Intra- and Inter-organizational Business Data Integration and Usage

A M. Tjoa
TU Wien, Vienna, Austria
amin@ifs.tuwien.ac.at

Abstract - **Information silos and conceptual heterogeneity within organizations remain to be main problems towards an effective utilization of available information resources.**

Since decades, proper data and knowledge-integration and extraction remain to be giant tasks in the information systems arena. This paper briefly introduces the Linked-Data Concept as a possible approach for easing this problem.

Linked Data can be utilized on the one hand within organizations ("intra-enterprise") for the integration of possible existing heterogeneous systems and on the other hand as enabler of cross-organizational data integration ("inter-enterprise") in a networked economy.

The "Linked Data Enterprise" unifies both concepts of intra- and inter-enterprise networks as a very flexible and extendible instrument of data integration.

Special emphasis is laid on the applicability of the FAIR Guiding Principles for data management and stewardship that was originally introduced for scientific data.

Keywords – Linked Open Data, Ontology Based DataAccess, Business Value, Linked Data, Linked Enterprise DataNetworks, Enterprise Content Management.

I. INTRODUCTION

Since its emergence, the concept of Linked Open Data (LOD) attracted a lot of attention from the scientific community. From a business perspective, the use of Linked Open Data and especially the combination of Linked Open Data with internal company records (e.g. past sales figures, shipping fees, capacity utilization, customer data, transportation routes) provides the potential of an incomparable information source which can be used by small, medium and large-size companies and across various industry segments, to gain a competitive advantage.

The widespread adoption of electronic information systems by enterprises has as its ultimate goal the complete, integrated access to all data. With linked data, that promise is now at hand. The main enterprise benefits to linked data reach from the elimination of internal data silos to the easy interlinkage of enterprise, industry-standard, open public and public subscription data. Other benefits can be obtained in the complete data modeling transparency of any future legacy schema, the ease of updates and changes of existing schemas, leading to the end of the tremendous efforts so far needed for harmonizing and even "re-architecting" legacy schemas due to organizational changes like merger and acquisitions.

Ultimately, the revised organizational culture of a Linked Data Enterprise will value both the act of sharing and producing of data as equally important. [1]

In 2016 the G20 state leaders endorsed the „FAIR Guiding Principles for scientific data management and stewardship" to enforce the Findbility (F), Accessibility (A), Interoperability (I), and Reuse (R) of digital assets. These principles should ultimately enable a worldwide machine-actionability with none or minimal human intervention. [2]

The FAIR-principles are considered as a guarantee for the promotion of open science and for the facilitation of access to research results that can be jointly shared by the scientific community.

This paper is aiming at propagating the need of an equivalent guiding principles for Intra- and Inter-organizational Linked Data Enterprises. [3]

II. LINKED DATA ENTERPRISE (LDE)

The original idea for the use of Linked Open Data in enterprises can be sought in the Linked Enterprise Data (LED) approach in order to open up a data architecture, in which data sources (from different enterprises) are linked based on their meaning and content. This further allows the use of semantic technologies to interpret, link, further process and display heterogeneous data based on standards amongst enterprises as a driver for interoperable cross-organizational data integration [4].

In [5] the notion of Enterprise Linked Data (ELD) was introduced for the internal use of the Linked Data WITHIN organizations for purposes of a solid integration of data in heterogeneous systems within an enterprise.

In this paper we want to emphasize on the concept of Linked Data Enterprise (LDE) [6] aiming at BOTH a meaningful semantic integration of data sources WITHIN organizations ("intra-enterprise" or ELD-based data integration) AND the cross-organizational data integration across several independent business and governmental entities ("inter-enterprise" or LED-based data integration).

For an inter-organizational content management enhanced by LDE we urgently need the necessary guidelines for a fair-collaboration amongst the various players, which could be governments, research institutions, (competing) enterprises, and individuals delivering relevant data as the new "oil of the 21st century". This problem has become even more apparent due to the extremely successful results of deep-learning applications based on huge amount of available data. In the intra-organizational use of open data (ELD) we also have to consider that ethical principles for the protection of client and citizen data stored within enterprises.

III. FAIR GUIDING PRINCIPLES

For scientists and IT-professionals guidelines for the design, development and applications of LDE-solutions as essential as the use of open data for research data or AI-applications.

Of particular interest is the introduction of the FAIR-principles [2] which has gained relevance due to the recommendation of the G20 state leaders in 2017 where the support of "appropriate efforts to promote open science and facilitate appropriate access to publicly funded research results on findable, accessible, interoperable and reusable (FAIR) principles" was unanimously agreed. [7]

The FAIR principles are very much suited to the Linked Open Data basic elements, which are predominantly built based on a "Semantic Web Technology Stack" with RDF, XML and an ontological foundation. All FAIR-principles are introduced via strict recommendations on meta-data. [2].

In particular the compliance with the following FAIR-principles are formulated as follows:

F1. (Meta)data are assigned a globally unique and persistent identifier
F2. Data are described with rich metadata (defined by R1 below)
F3. Metadata clearly and explicitly include the identifier of the data they describe
F4. (Meta)data are registered or indexed in a searchable resource

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
A2. Metadata are accessible, even when the data are no longer available

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (Meta)data use vocabularies that follow FAIR principles
I3. (Meta)data include qualified references to other (meta)data

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

The analogy to the ethical principles of LDE-work and the use of "big data" for deep-learning and other machine learning techniques in various applications in the area of Industry 4.0, Social Networking, and Intelligent Recommender Systems etc. is apparent. Therefore, similar principles are generally relevant for the use of (open) data.

In 2019 the principles of FAIR are investigated due to their ethical compliance according to the 2030 United Nations goals of the Sustainable Development Goals (SDG) where "No one should be left behind". Thereafter the Global Indigenous Data Alliance (GIDA) released the CARE Principles for Indigenous Data Governance as a complementary guide extending FAIR to include Collective benefit, Authority to control, Responsibility, and Ethics (CARE).

The University of Vienna has adopted both FAIR and CARE principles for its scientific repository. [9]

Although the original concept of CARE was the protection of the heritage and data of indigenous populations, it turns out that these principles are also relevant for the protection of people in general. The decision taken for scientific repositories to follow both FAIR and CARE can be regarded as a role model for future successful implementation of LDE in enterprises in inter- and intra-organizational applications.

IV. CONCLUSION

LDE causes new ethical, technological and governance challenges. These include risks of (un)intended misuse with a huge potential for outcomes which could be extremely harmful in a world where "No one should be left behind".

We have still to ask the question if Code of Ethics or Guiding Principles like FAIR and CARE, which are not legally binding need Laws and Regulations to complement the gaps left to guarantee trust and appropriate use of LDE. Especially due to the potential misuse of Big Data further investigations on an LDE-based commercial world should be further investigated.

We urgently need further work on how principles described in this paper can be operationalized the design and implementation in order to minimize risks and negative outcomes.

V. REFERENCES

[1] Allemang, D., Semantic Web and the Linked Data Enterprise. In: Wood, D. (ed) Linking Enterprise Data. Springer Science+Business Media, LLC, 2010. DOI: 10.1007/978-1-4419-7665-9.

[2] Wilkinson, Mark D et al., "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* vol. 3 160018. 15 Mar. 2016, doi:10.1038/sdata.2016.18

[3] EU High-Level Expert Group on AI (2019), Ethics guidelines for trustworthy AI, URL: *https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai* (accessed on July 15th, 2020)[4] A.Blumauer, Linked Data in Unternehmen. Methodische Grundlagen und Einsatzszenarien . in: T. Pelligrini, H. Sack and S. Auer (ed.), „Linked *Enterprise Data – Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologie"*, pp. 3-20. X. media.press, Springer, Berlin, 2014.

[5] A.Polleres, S. Kirrane and J. D. Fernández, Enterprise_Linked_Data_Whats_missing, *https://aic.ai.wu.ac.at/~polleres/presentations/20160921.pdf,* 2016 (accessed on July 15th, 2020)

[6] N. Novak and A M. Tjoa, Towards a Business Value Framework for Linked Enterprise Data, in *Proc. IEEE-RIVF International Conference on Computing and Communication Technologies,* 2019.

[7] G20 Leaders' Communique Hangzhou Summit, *https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967* (accessed on July 15th, 2020)

[8] GO FAIR, "FAIR Principles", *https://www.go-fair.org/fair-principles/ ,2016* (accessed on July 15th, 2020)

[9] FAIR and CARE Data Principles, https://datamanagement.univie.ac.at/forschungsdatenmanagement/fair-und-care/ (accessed on July 4th, 2020h

PAPERS

Time series model for sales predictions in the wholesale industry

Tomislav Hlupić, Dražen Oreščanin, Ana-Marija Petric
Poslovna inteligencija d. o. o.
tomislav.hlupic@inteligencija.com, drazen.orescanin@inteligencija.com,
ana-marija.petric@inteligencija.com

***Abstract* - The prediction process in sales is a basis for a successful ongoing planning process for any organization. Wholesale companies, being B2B oriented, have to plan their organisational environment carefully to optimize the costs and maximize revenue. As the sales process is intersected with logistics, having precise sales predictions optimizes both sales and logistics processes. In order to track the sales towards a customer, we propose a data mart built on the top of the data warehouse to be used with daily loads of outgoing invoices and uninvoiced shipments data.**

Predictions are based on ARIMA model, one of the most popular forecasting models for the time series. The data is aggregated on a weekly level, as it was proven to be the most useful in this process. For the prediction purposes, we are focusing only on the outgoing invoices. From the business perspective, each product is tracked with data about the sales market, customer, quantity, and the date. In the article, the process of data preparation will also be included as it is the crucial step for successful prediction.

Keywords – forecasting models; sales predictions; time series; wholesale industry

I. INTRODUCTION

In the wholesale industry, planning processes have a significant impact on various aspects of the business. Originally, specialized planning software based on the data warehouse was being used for the purposes of creating forecasts, but more novel approaches are being implemented to augment the existing processes. One of those is implementing an automatic prediction model based on the data which is already obtained as a part of daily used analytical solution. Novel approaches are not meant to replace the existing systems, but rather to be an addition to the already established business processes support.

The data warehouse, being time-variant, is an excellent basis for creating time series based on the historical data stored in the fact tables. Time series are on the most common approaches for data forecasting though applying various algorithms. The ARIMA (Auto-Regressive Integrated Moving Average), being one of the widely used models in the time series forecasting and a form of the Box-Jenkins method, is considered as a basis for the computer-based forecasted data. The ARIMA model was chosen as a basis for developing a forecasting model as the analyzed time series were showing the possibility of achieving valuable insights after creating the stationary time series from the original ones.

Forecasting methods have been a research area in the field of economics for a significant time, with a significant focus on the sales forecasting [1][2][3]. In a broader picture, the forecasting spans throughout the whole business scope from obtaining the goods up to their delivery to the customer [4]. Traditionally, forecasting has been done either semi-automatically through user input or though predefined models.

A more automated process consists of implementing models, such as ARIMA [5] [6], into forecasting process where the data is fetched either from the operational environment or analytical systems, such as data warehouses. With the higher computing power, neural networks were introduced both as a hybrid (including ARIMA models [7] [8] [9]) and independent systems used for more precise forecasting.

The article covers the business aspect of the proposed solution, describing the planning process and predictions process in Chapter 2. The data model used as the basis for the prediction solution is described in Chapter 3. The description of the data preparation process and its importance in the solution are covered in Chapter 4. Finally, the description of the prediction model is given in Chapter 5.

II. PLANNING PROCESS AND PREDICTIONS

In modern business, planning process is one of the critical processes in the company. Company can measure success of its business operations only if actuals can be compared to business plan. Traditionally, business plan is created on yearly base and usually this process is addressed as annual budgeting process. Main outputs of that process are planned financial statements that are describing business results of the company at the end of next fiscal year – Profit & Loss statement, Balance Sheet and Cash Flow. To be able to create those three main statements, business plans with revenues and operational and capital expenses must be made throughout the company, covering all major business functions – sales, distribution, procurement,

manufacturing, customer support and also all supporting business functions.

Sales planning is one of the most important parts of the planning process, because total revenue of the company and revenue growth is highly dependent on sales. Traditionally, sales planning is done based on sales results for first nine months of current year with estimated sales till the end of the year as a base. Sales plan for next year is then based on corporate targets for next year, expected seasonality, planned launches of new products or expansions to new markets. Also, influence of changes in the market such as economy growth in important markets or entrants of new competitors are taken into account.

Once the annual budget is in place, companies are tracking actual execution of the plan and making forecasts for the next periods, mostly till the end of the year. In such forecasts, first several months are based on actuals and the rest may be based on annual budget or forecasted values that are taking into account results from finished periods and information about market changes that were not available at the time when annual budget was created.

For wholesale company's main issue in planning process is that they are selling and delivering their products to retailers as their customers and they actually in most cases don't know when their products are sold to end customers and who those customers are. Also, their products can reach end-customer in any time between delivery to the retailer and end of validity period of the product, and that period may vary from one or few days for fast moving consumer goods to many months or years for goods with longer validity periods.

Approach described in this paper that use ARIMA method is useful in two parts of planning process:

- Creating of estimation of remaining sales for current year as a base for next year annual budget
- Creating forecasts based on sales results in first several months of the year, including also longer time series from previous year or years

III. DATA MODEL

The source of the data for the prediction algorithms can vary from the flat files (e. g. Comma Separated Value – CSV) to the complex solutions based on various types of the databases. The most common approach for storing the data for the analytical purpose is a data warehouse, a database specifically modeled for analytical queries. The data warehouses main components are dimensional and fact tables, where the dimensional tables hold the descriptive data, while the fact tables contain the measures of the business' subject areas.

The time variant feature of the data warehouse enables the usage of the time-dependent predictive analytics, such as the chosen ARIMA method. All the records in the fact tables have a timestamp, which is usually a date with the business meaning. The timestamps allow the analytic system to create summarizations over a certain period of time, which might stretch from a day to a year.

Another important aspect of the analytical systems are data marts, functional areas of the data warehouse that have a focus on a specific business area [10]. Data marts are either independent smaller data warehouses or a part of the enterprise-wide data warehouse [11]. The basis of the proposed data model is a sales data mart which will be described in detail. The model itself is loosely based on the Retail model developed by Poslovna inteligencija [12].

The source data which the model is based on are the invoices sent to the partners stored in the operational environment and is integrated to the data warehouse through the daily ETL process. Each invoice contains the invoice date, partner data (name, market), invoice items which are the products being sold and the quantity for each sold invoice item. The invoice details are stored to the corresponding dimensions, while the invoice sales data together with the invoice item granularity are stored to the sales fact. The data in the fact table are denormalized, so each record, besides the foreign key to the products dimension, also contains the foreign keys to the invoice related dimensions together with the relevant invoice data (invoice number, invoice date…).

The data from the fact table are exposed to the external systems, including the prediction engine, through the view whose structure is created using keys and aggregations, as explained in the next chapter. Besides the view, a table was created to which the pivoted prediction results are stored and connected to the dimensional data through the codes, instead of the surrogate keys. The codes were preferred over the surrogate keys in order to have a human-readable data available for the data preparation process.

The data model, shown on the **Figure 1**, in the data mart follows the star schema and includes the sales fact (f_sales), time dimension (d_date), partner dimension (d_partner), market dimension (d_market) and the product dimension (d_product). In the data model, the product price is stored only in the fact table, thus enabling the historization and allowing pricing flexibility depending on several factors.

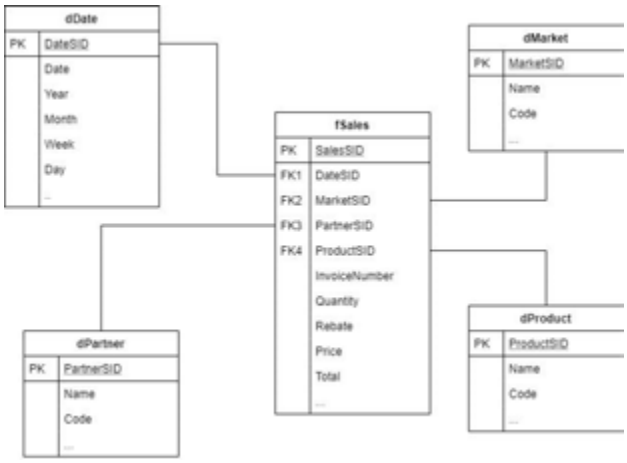


Figure 1. The proposed data model

IV. DATA PREPARATION

The data preparation process is as crucial for the successful prediction process as the built prediction model. Through the data preparation process, a source data is once again structured to the human-readable model which, in the case of the relation databases, consists of the measures being bind to the codes through the foreign key relations. The main reason for the codes being used instead of raw database keys are the manual operations being done throughout the process.

In case of sales forecasting, data preparation process includes removing records which make the dataset unnecessary larger and the prediction process slower (such as obsolete entries with no impact or even negative impact on the prediction process), adding prediction-related attributes, grouping and aggregating the data and classification.

Clearing the dataset for the implemented prediction model consists mostly of removal of the products' sales data when the products are discontinued from the wholesale portfolio. Usually those products are either marked with a flag in the dimension or have a derived flag based on the threshold (e. g. a period of time in which the product didn't appear in any invoice). Removing those products, besides the performance, has impact on the prediction model's correctness percentage.

Also, several other data matters, from the business perspective are tackled during the data preparation process, one of which is the business impact of product returns. Typically, returns are shown as negative sales in the data and, depending on the business perspective, may be included in the prediction dataset, completely removed or another dataset might be created based on the data. The returns' data contain valuable information about the 3rd partner sales and another model could be created for those predictions, as they impact the sales predictions, the products purchase needs and, finally, impact the company's profit. Another aspect is the cancellation of the purchases that first need to be analyzed prior to the decision how to include them in the dataset. Typically, they are included, but the check whether all the cancellations amounts have been met by the original invoice amounts should be applied and the outliers detected in order to interpret the final prediction results.

Aggregations performed are only summarization of the records by the prediction keys, where the time dimension allows the records to be aggregated on an arbitrarily granularity level which fits best the data depending on the model. Most common aggregations are daily, weekly and monthly and the choice variously depends on several factors, like the sales frequency, volume and finally the built model.

As the data preparation process lays on the business intelligence layer, is it a common practice to use the views on the data warehouse as the source of the data. The view reflects the dataset structure and can be either used to ingest the data directly into the prediction engine or to store the data into a table with the same structure and exposed again through the business intelligence layer. The approach depends on the volume of the sales data, the frequency of the data refresh and the frequency of the data ingestion to the prediction engine.

V. PREDICTION MODEL

The main reason for choosing ARIMA as a forecasting model is the ease of implementation and fairly easier understanding from the user's perspective when analyzing the data. Generally, ARIMA model can be simplified as a model which takes previous points in the time, puts weight on them (so the most recent data has more impact on the prediction than the previous data), reduces the seasonality and removes the random movements from the time series. The model is adjusted for the data by changing the properties values until the output of the model forecasting predictions reaches the optimal result (the level of prediction correctness).

Mathematically, ARIMA model is described as a discrete time linear equation with noise in the following form

$$\left(1-\sum_{k=1}^{p}\alpha_k L^k\right)(1-L)^d X_t = (1+\sum_{k=1}^{q}\beta_k L^k)\varepsilon_t$$

and consists of 3 parameters:

p: *Lag order* - the number of lag observations included in the model, also called the lag order
d: *Degree of differencing* - the number of times the raw observations are differenced
q: *Order of moving average* – the size of the moving average window

The autoregressive component (AR(p)) stands for the fact that the past values in the series' regression equation are used, where *p* is the previously specified lag order. The integrated component (I(d)) specified the differentiation

by subtracting current and previous values of the series *d* times. The moving average component (MA(q)) represents the error of the model calculated as a combination of the previous error terms.

- Autoregressive model of the order *p* can be defined as:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

where ε_t is the white noise (assumed mean of errors). The definition indicates that the autoregressive model consists of the multiple regressions with the *p* lagged values of y_t as predictors. To create a different time series pattern we change the parameters $\phi_1 \ldots \phi_p$.

- The degree of differencing (d) is defined as:

  $y'_t = y_t - y_{t-1}$

  which computes the difference between two consecutive observations.
- Moving average of the order *q* is defined as follows:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

where the value of of y_t is the weighted moving average of the past forecasting errors. Typically, we will limit the forecast errors to a recent period, as the more historic errors do not impact the model significantly.

On a high level, we can abstract the prediction algorithm using ARIMA as following:

1) Define the model ARIMA (p, d, q)
2) Fit the defined model on the training dataset (smaller subset of the original time series)
3) Run the predictions on the latter data using the fitted model
4) If the outcome doesn't fall into the defined confidence level (the correctness of the prediction compared to the existing data), adapt the properties p, d and q of the model

The main benefit of introducing the ARIMA model into the sales predictions for the wholesale industry is the possibility of running the model automatically as the significant amount of data arrives into the data warehouse. Usually, the process of traditional planning and forecasting includes a vast amount of manual work using specialized tolls, while our implementation of the model on the top of the existing data warehouse can automatically adjust the parameter models through the *auto.arima*() function even without any user intervention. The models can be fitted on a daily basis, providing the end users a possibility of faster determination of the strategic and tactical decisions without the delay which typical human intervention includes.

One of the prerequisites for using ARIMA is to make the series stationary. This means that we have to select optimal parameters to deal with trend, seasonality and periodicity of a series. When we deal with a single series we can explore and manually select the optimal parameters for the series by creating ACF and PACF plots and extracting the correct values for series' parameters from the plot. However, our data set consists of 1000+ series, each of which has a different seasonality and trend so manually setting the parameters would be time consuming and ineffective. Due to this reason, we take advantage of the generic *auto.arima()* function in R to obtain the optimal parameters for each of the analyzed time series.

Forecast is run for each of the products in a loop, ensuring that the forecast is made correctly for each of the products. We have opted out to use the *forecast()* function from the {forecast} package in R [14]. This function is generic but the advantage is that we can forecast either on a time series itself or the time series model. Since we are fitting each product individually with *auto.arima(),* we are forecasting from the fit rather than the raw series. Forecast is suitable for univariate time series. Since we are running forecast inside of a loop, we are treating each of the parallel series as univariate, ensuring that the predictions are based on the correct parameters. In this case, the required object of the function forecast is of the class Arima, rather than time series. Additional parameters used in our model are **h** which designates the number of periods for forecasting (in our case 120 days) and **level** which indicates the confidence level for prediction intervals (in our case 95, signifying 95% confidence interval). This confidence interval is rather high, as the normal confidence interval is set as a vector of values ranging between 85 and 95.

The *forecast* function returns an object of class "forecast" which is a list containing at least these elements:

- model - a list containing the information on the fitted model
- method - the name of the forecasting method (character string)
- mean - point forecasts as a time series
- lower - lower limits for prediction intervals
- upper - upper limits for prediction intervals
- level - the confidence values associated with the prediction intervals
- x - the original time series
- residuals - residuals from the fitted model. For models with additive errors, the residuals are x minus fitted values
- fitted - fitted values (one- step forecasts) [13]

We also include the plots for each of the series inside the loop which compare forecasted and actual values to visually examine the predictions. Lastly, inside of the loop we also compute and store individual accuracy percentages for each of the series. When we work with multiple parallel series we expect that some of the series will have a lower accuracy then others, and computing

individual accuracy (rather than overall model accuracy) is useful to pinpoint those series which have low accuracy so that we can extract them for further examination and analysis.

It is possible, especially when we deal with large amounts of distinctive series, that the series with low accuracy will have specific features rendering them unsuitable for ARIMA analysis (i.e. we may have seasonal products or gift packages, which are made in specific quantities and sold on the basis of availability rather than demand). This means that they would normally not be forecasted in the same way as the other series. If the information about seasonal and/or special products is not available in the source data, detailing the specific features of these kinds of series, our model will treat them the same as the other ones, which will then lower the overall model accuracy. For business purposes, it is advisable to always run both the overall model accuracy as well as the individual accuracy when we are dealing with multiple univariate parallel series.

VI. CONCLUSION

The prediction models are a novel approach and an addition to the existing planning processes in enterprises. A lot of enterprises have incorporated a data warehouse as a basis of their analytical systems, which stores the integrated data from the operational environment. As the data warehouses store the data with the temporal component, it is possible to create time series which server as a basis for the prediction models.

The data mart, stand-alone or as a part of the enterprise data warehouse, with the purpose of storing and analyzing the sales data uses a single fact table and several dimensions which all together track the sales on the granularity of a sold product per invoice. Through a view in the business intelligence layer, the prediction engine ingests the exposed data that went through the data preparation process which results in the prepared dataset. The data preparation includes removing business-wise irrelevant data, separating the unsuitable data into a new dataset and finally creating aggregations.

The forecasting process is based on running a loop on the univariate time series containing the product sales data that produces forecast data, the correctness percentage (compared to the actual sales data) and the plots that visualize actual and predicted sales. Finally, both model overall accuracy and individual product accuracy are taken into account, as a number of products incorporate various features that make them stand out and behave divergently.

Based on the developed model and forecasting process, a future research will be focused on enhancement of the forecasting, mainly on introduction of the neural network-based forecast accompanying the existing ARIMA models.

REFERENCES

[1] J. T. Mentzer and J. E. Cox, "Familiarity, application, and performance of sales forecasting techniques," J. Forecast., vol. 3, no. 1, pp. 27–36, Jan. 1984.

[2] C.-W. Chu and G. P. Zhang, "A comparative study of linear and nonlinear models for aggregate retail sales forecasting," Int. J. Prod. Econ., vol. 86, no. 3, pp. 217–231, Dec. 2003.

[3] A. P. Ansuj, M. E. Camargo, R. Radharamanan, and D. G. Petry, "Sales forecasting using time series and neural networks," Comput. Ind. Eng., vol. 31, no. 1–2, pp. 421–424, Oct. 1996.

[4] E. V. Gijo, "Demand forecasting of tea by seasonal ARIMA model," Int. J. Bus. Excell., vol. 4, no. 1, p. 111, 2011.

[5] M. Shukla and S. Jharkharia, "Applicability of ARIMA Models in Wholesale Vegetable Market," Int. J. Inf. Syst. Supply Chain Manag., vol. 6, no. 3, pp. 105–119, Jul. 2013.

[6] J. Martin and T. Setzer, "On the Robustness of ARIMA-Based Benchmarks for Corporate Financial Planning Quality," in 2014 47th Hawaii International Conference on System Sciences, 2014, pp. 1221–1229.

[7] J. V. Hansen and R. D. Nelson, "Time-series analysis with neural networks and ARIMA-neural network hybrids," J. Exp. Theor. Artif. Intell., vol. 15, no. 3, pp. 315–330, Jan. 2003.

[8] F.-M. Tseng, H.-C. Yu, and G.-H. Tzeng, "Combining neural network model with seasonal time series ARIMA model," Technol. Forecast. Soc. Change, vol. 69, no. 1, pp. 71–87, Jan. 2002.

[9] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," Neurocomputing, vol. 50, pp. 159–175, Jan. 2003.

[10] P. Pipe, "The Data Mart: A New Approach to Data Warehousing," Int. Rev. Law, Comput. Technol., vol. 11, no. 2, pp. 251–262, Oct. 1997.

[11] A. Bonifati, F. Cattaneo, S. Ceri, A. Fuggetta, and S. Paraboschi, "Designing data marts for data warehouses," ACM Trans. Softw. Eng. Methodol., vol. 10, no. 4, pp. 452–483, Oct. 2001.

[12] "Poslovna Inteligencija Data Warehouse Data Model for Retail and Distribution." [Online]. Available: https://dwh-models.com/retail/. [Accessed: 23.01.2020].

[13] M. S. Peiris and B. J. C. Perera, "On Prediction With Fractionally Differenced Arima Models," J. Time Ser. Anal., vol. 9, no. 3, pp. 215–220, May 1988.

[14] "Package Forecast" [Online] Available: https://cran.r-project.org/web/packages/forecast/forecast.pdf [Accessed: 7. 7. 2020]

Business Intelligence Approach to Support Decision Making in Publishing Sector

D. Borissova[*,**], N. Keremedchieva[*] and D. Keremedchiev[*,***]
[*] Institute of Information and Communication Technologies at Bulgarian Academy of Sciences / Department of Information Processes and Decision Support Systems, Sofia, Bulgaria
[**] University of Library Studies and Information Technologies, Sofia, Bulgaria
[***] New Bulgarian University, Sofia, Bulgaria
dborissova@iit.bas.bg, ninakeremed@gmail.com, delyan.keremedchiev@gmail.com

***Abstract* – The article deals with decision making approach when generation of different offers for printing of books and particularly for printing of monographs. In accordance to the law for academic staff development in Republic of Bulgaria, the monograph should be at least 100 pages with 1800 signs within page. Depending on the monograph subject, different pictures, diagrams, formulas and tables may be included. All of these considerations are taken into account when generate different offers depending on the quality of printing pictures, type of paper and number of pages. To cope with such problems a framework of web-based decision support system is proposed. The three-layer architecture described web-application is capable to process the input information about the monograph contents and to generate different offers in accordance to different users' preferences.**

Keywords – Business intelligence; decisin making; publishing; web-application.

I. INTRODUCTION

The contemporary business environment is constantly changing and along with actual challenges the proper business model should be provided. To cope with such type of problems the decision support system can be used. The main advantage of decision support systems is the ability to integrate different business models in respect of various activities [1]. This could be realized by using of business intelligence that mixes data mining, algorithms, visualization and other approaches to help businesses make better decisions [2, 3].

The digital revolution reflects also on the publishing book industry that moves from analogue to digital formats as other creative industries [4]. This change need to cope with different business models that firms are used to propose their product and services. The lack of dominant publishing design and different goals of publishers corresponds to various business models are to be redesigned to meet the new digital changes. The contemporary trends of digitalization lead to the appearance of new forms of organization and management that are related with virtual enterprises corresponded to the real ones. In this regards, an approach to develop of a new e-media based on the business model related with technology transfer office is proposed in [5]. This model is based on virtual agents considered as fundamental element of developed e-media to describe some information processing functions. The identification of the processes and related good practices in similar service-based management should be considered including also cyber resilience [6, 7].

Business itself is changed over the time and the actual information should be properly processed to meet the current business activities. To support different business activities there exist different decision support system realized as desktop application or as web-based tools [8, 9]. The advantages of online analytical processing can be used to enables users to extract and query data in order to analyze it from different points of view. Such business intelligence queries could be used to generate different budgeting offer to be able to response of different users' points of view. In similar situations, decision-making can be done by different individuals or by group decision making. The difference between experts' experience within group can be taken into account by using different importance for their opinions when aggregating the final group decision [10, 11]. Thus, by using of set of processes, architectures, and technologies it is possible to convert raw data into useful information that contribute profitable business actions.

Recently, there exists a renewed interest in publishing sector along with its integration into multinational entertainment and information conglomerates [12]. It has been observed a special attentive to the results of concentration, the importance of publishing as a global phenomenon, and the conjunction between books and other entertainment and information media. Commercialization being the main theme intertwined with decentralization discussed in [13]. Significant China's publishing industry changes are result of combining of commercialization, decentralization, privatization and internationalization. The latest trends in print on-demand for academic publishing (2018–2025) identified regarding the use of academic books and associated publishing strategies are related with three trends [14]. These trends impact on the print-on-demand book and are related with 1) more 'digital first' academic books with optional print-on-demand editions, 2) more integrated book publishing ecosystems enabled by digital technologies and 3) more offer author-centric platforms publishing services with new technologies concerning the physical book, more

multimedia and more custom books. It is shown that Chinese academic publishers should improve the quality of academic e-books and to transform to the academic content service providers, and along with that to improve the cooperation with stakeholders [15]. To be successful, the e-book market should work together including authors, publishers, online book stores, book distributors, reading device manufactures, etc. [16].

Different approaches of the problems related with setting up a virtual company in publishing sector are discussed in [17]. The advantage of such virtual publishing house is the ability to carry out orders faster and cheaper than traditional and could be easily adapted to market changes. To make this possible, the workflow and process management in printing and publishing company is to be known. The efficiency of the production process requires digitization at all stages except the start and end stages [18]. All of these aspects offer competitive advantages to printing and publishing companies to adapt their business processes according to the technological and organizational framework of digital printing. In addition, a three-stage model is presented for direct publishing and the bargaining between the author and the publisher [19].

All of these motivate the authors to investigate the publishing sector through the prism of business intelligence and to propose a web-based tool to support authors in decision making when printing a monograph. This paper seeks to address the last modification of the law for academic staff development in Republic of Bulgaria related with monograph printing. The rest of the paper is organized as described. Section 2 provides the definition of monograph corresponding to the law for academic staff development in Republic of Bulgaria. Section 3 describes the proposed framework of decision support system for calculation of offers for books/monographs printing. An algorithm for estimation whether the manuscript could be considered as monograph is also presented. In Section 4 numerical testing of the developed prototype and discussion on obtained results are given. Section 5 contains conclusions for the using of proposed approach and future investigations.

II. Problem Description

The latest corrections made in 2019 about the law for academic staff development in Republic of Bulgaria concern the definition of monograph. The monograph should consist of at least 100 pages and 1800 signs within page. Due the monograph subject, the issue could involve additional components except the plain text as pictures, diagrams, formulas. That means the overall number of pages of monograph has to be reduced by some coefficient considering number of rows taken from the formulas and another coefficient to take into account the pictures size. In addition, there are two more components that reflect on the correct offer for the price, namely type of printing paper and color printing. At least but not last, the number of circulation is to be taken in consideration too.

III. Framework of Designing a Decision Support System for Calculation of Offers for Monographs Printing

To tackle with described above problem a framework of decision support system for calculation of offers for monograph printing is developed. The proposed decision support system is realized as web-application with business intelligence that is able to process the input data, taking into account the given restrictions and requirements is such way to meet the actual definition of monograph accordingly the law for academic staff development in Republic of Bulgaria. The three-layer architecture of proposed web-application is shown in Fig. 1.

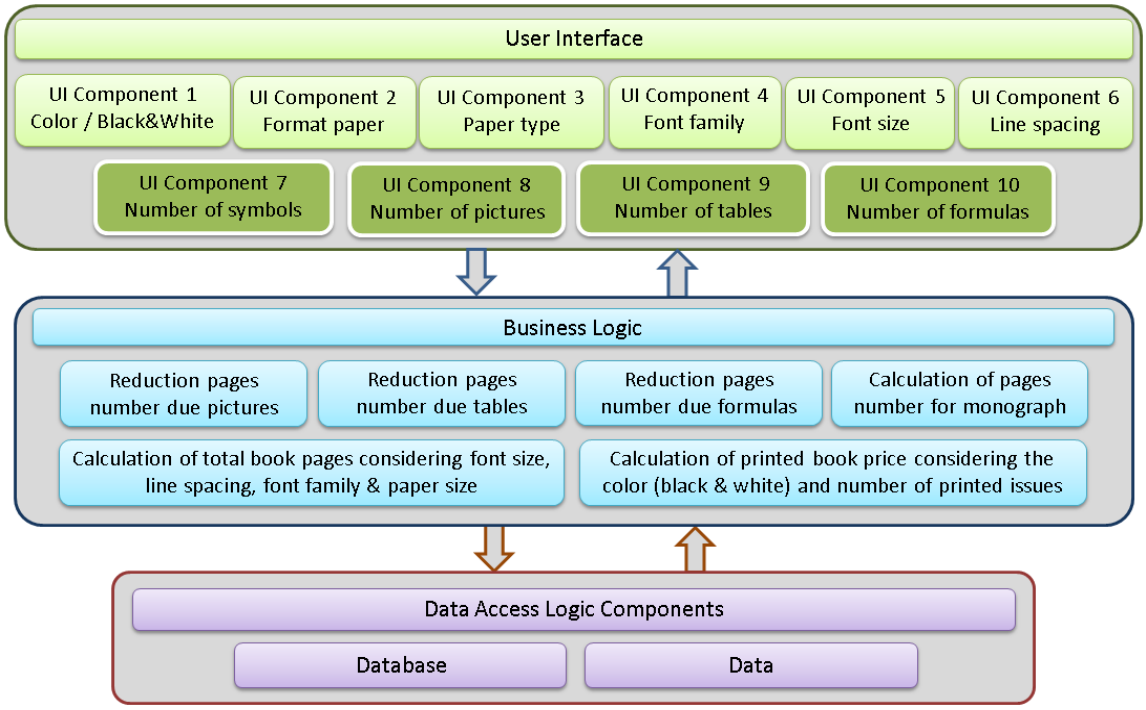


Figure 1. Architecture of decision support system for calculation of offers for books/monographs printing

The first layer is the presentation layer that is realized by proper user interface including different items to get the input information. The main role of this layer is to provide the interaction between the user and sending requests to the data layer through business logic layer. The second layer is the business logic layer that play essential role as it provides the needed calculation between input data from user interface and the data layer. The third layer is the data layer that communicates with the business logic and serves to store data and their accessing.

To design a web-based tool to be able to answer the question if the manuscript corresponds to the definition for monograph corresponding to the law for academic staff development in Republic of Bulgaria a workflow algorithm is proposed. This algorithm along with its main two stages is illustrated in Fig. 2.

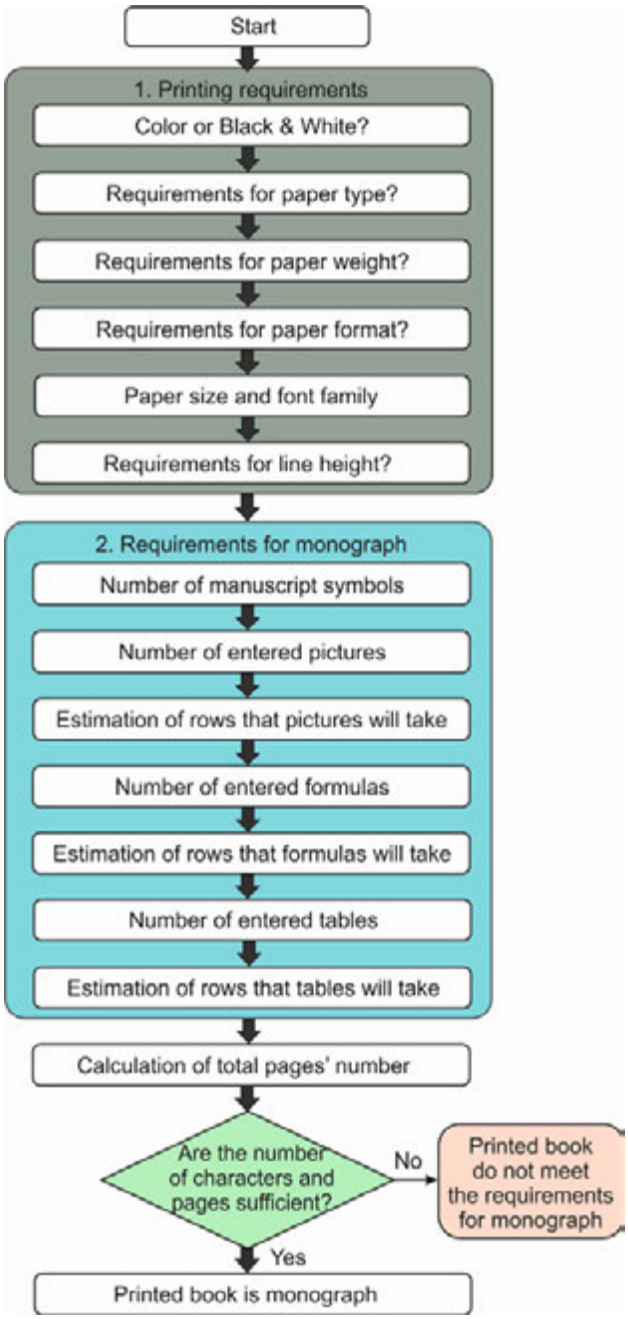


Figure 2. Algorithm for estimation whether the manuscript could be considered as monograph

The first stage concerns the needed parameters of which the printed manuscript should meet, while the second stage is focusing on the manuscript contents – text symbols, pictures, formulas, tables, etc. The main elements of the printed book include information about the color or black and white printing, type of paper (bond paper, gloss coated paper, matt coated paper, recycled paper, silk coated paper, uncoated paper or watermarked paper). The next components of the printed book are the paper weight and dimensions of the sheets. Once the printed format is known, the font size, font family and line height are to be determined. Depending on these three parameters the number of overall sheets of book could be vary significantly. When the number of symbols of entire manuscript is entered, it is possible to calculate the total number of pages without including pictures and formulas if any. In case of present of pictures and formulas it is needed to estimate the number of rows that formulas will take along with some estimates about the required rows or sheets taken by the pictures.

Depending on the available web server, some parameters influencing on its efficiency usage have to be considered [20]. This analysis will reflects in better organization of users' requests. In some cases, it could be better to use optimization in determination of parameters by using of maximal/minimal values [21].

All of this information together with additional information for the required number of printed books can be used to calculate the total price of the printing issue. Based on described above components for each given manuscript, the corresponding offers with different parameters could be generated and presented to the authors.

IV. NUMERICAL TESTING

To demonstrate the applicability of proposed framework of decision support system for calculation of offers for monograph printing a case-study for particular manuscript is considered. Before generating the offer, the proposed algorithm in Fig. 2 should be perform to answer the question if the manuscript can be considered as monograph accordingly the definition for monograph from the latest corrections of the law for academic staff development in Republic of Bulgaria.

A case-study of given two manuscripts are used to test the applicability of proposed framework shown in Fig. 1 and developed algorithm in Fig. 2. The parameters of these two manuscripts used as input data are shown in Table I. These input data of the given manuscripts are used to prepare two printed books with equal parameters as follows: color printing, paper format A5, paper weight 90 gr, single line height, Times New Roman font with size of 10 pt.

TABLE I. MANUSCRIPTS' PAREMATERS

No	Manuscripts' parameters			
	Number of symbols	*Number of pictures*	*Number of tables*	*Number of formulas*
1	222099	4	8	6
2	130646	16	18	17

Based on the proposed three-layer architecture of decision support system framework for calculation of offers for monograph printing, a prototype of web-based application is developed. The first part of the designed graphic user interface contains the required parameters concerning the printed book including color or black and white printing, format paper, type of paper, paper weight, font family and line height. The second part of developed application contains the parameters of manuscript that should be printed as a book. This section requires entering the specific manuscript' parameters that are used in the business layer to perform needed calculations. The third part of graphical user interface contains the button, which invoke JavaScript calculations to answer the question whether the manuscript could be considered as monograph in accordance to definition of the law for academic staff development in Republic of Bulgaria.

All of these considerations are taken into account in developing graphical user interface of the prototype of decision support application for calculation of different offers with different parameters for printing of books/monographs. The input data of manuscript-1 are shown in Fig. 3.

The applicability of the developed prototype for generation of offers for books/ monographs printing is numerically tested by using the input data of two particular manuscripts with parameters as shown in Table I. Taking into account the number of total symbols, pictures, tables and formulas for each manuscript, the obtained data as a result of performing the algorithm for estimation whether the manuscript could be considered as monograph are summarized in Table II.

TABLE II. PRINTED BOOKS' PAREMATERS

No	Offer for color printed books			
	Number of pages	*Number of issues*	*Price per unit*	*Manuscript is monograph?*
1	205	30	4.37	**yes**
2	128	30	2.73	**no**

To determine the number of pages of the printed book some expert estimation are taken into account. For example, the estimation of rows taken by each formula is considered as equal to three rows in determination of total pages. Each picture and table from the manuscript are considered with equivalent height that corresponds to a half page of printed book.

Using the estimation about total page number together with number of printed issues allow to determine the expected price of the printed manuscript (Fig. 3). The next important feature of the proposed framework of decision support system for calculation of offers for books printing is the possibility to evaluate whether the manuscript could be considered as monograph accordingly the latest corrections of the law for academic staff development in Republic of Bulgaria. This is possible to realize by using the given definition of monograph, where the monograph should contain at least 100 pages and 1800 symbols per single page. This calculation is made perform within business layer, where proper reduction of rows is to be considered due the pictures, tables and formulas are available in each manuscript. Except this, the final information about the printed book is also available including number of total pages of the book, number of issues and price.

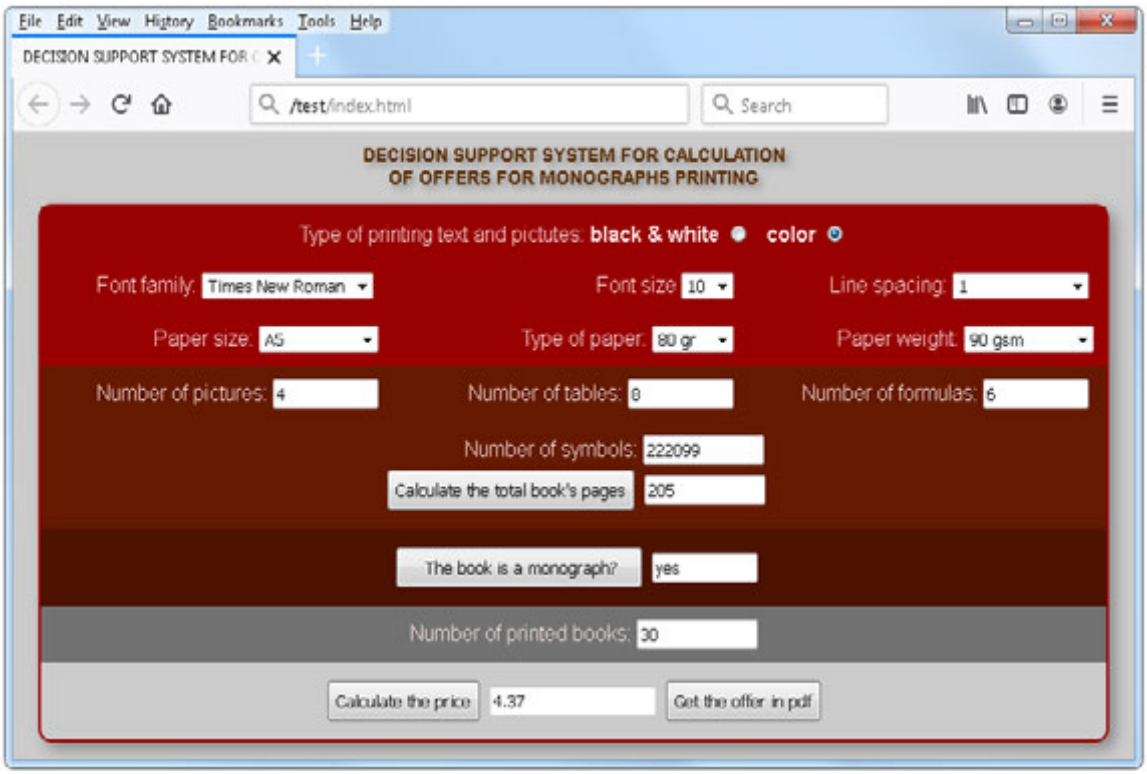


Figure 3. GUI of developed prototype for calculation of offers for book/monograph printing

The obtained results for manuscript-1 show that it could be considered as monograph in accordance to the definition given in the law for academic staff development in Republic of Bulgaria. The parameters for printing manuscript-1 (from Table II) shows total pages number equal to 205 that is over 100 with entire number of symbols are 222099 and is over the needed 180000. Unlike to manuscript-1, the second option concerning the parameters for printing manuscript-2 shown that it could not be considered as monograph. The total pages number equal to 128 is over 100, but overall symbols are less then needed 180000 symbols as it is written in by definition in the law for academic staff development in Republic of Bulgaria.

After entering the input data about each manuscript and checking whether manuscript could be considered as monograph, the final offer can be generated in pdf format. Two additional offers are generated for the same manuscripts considering the requirement for 100 printed issues instead 30. The obtained expected price per unit is visualized in Fig. 4.

The comparison of generated two different offers for both manuscripts for 30 and 100 printed issues show that order of more printed issues led to decrease the price per unit. For example, the single price for manuscript-1 could be reduced from 2.73 to 2.59 if 100 printed issues are ordered. The same situation is observed for the manuscript-2 – decreasing the price from 4.37 to 4.15 per single issue (Fig. 4). The prices are without VAT (20 %) and flat tax (10 %) and do not include some extra costs related with type of cover and design, ISBN and proofreading. Other printed book' parameters can be changed to get other possible offers like line spacing, font type, font size, etc.

The proposed approach allows combining the business analytics together with data mining and data visualization by using the developed prototype of web-based application to make data-driven decisions. Thus, the authors will have more than one option to make his/her preferable decision in offer selection. All of these confirm the applicability of the proposed framework and its implementation as decision support system for calculation of different offers for books/monographs printing.

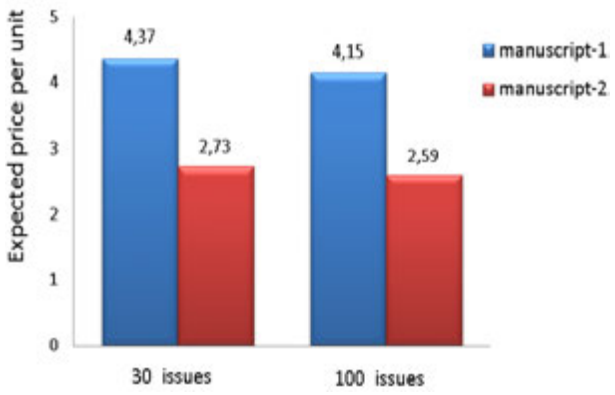


Figure 4. Dependence between expected price per unit and number of ordered printed issues

The described prototype of decision support system for generation of offers for books/monographs printing is realized as web-based application. Technologies like HTML5, CSS and JavaScript are used to demonstrate the applicability of the proposed framework in Fig. 1 and developed algorithm for estimation whether the manuscript could be considered as monograph (Fig. 2).

The future developments are focused on implementation of proper database to store the parameters of manuscripts instead entering the data about manuscript each time. Availability of such database could contribute to automate the process of offer generation by loading of already saved data for the manuscripts. In such way, it will be easier to load the parameters of manuscripts and by setting of different printing options different offers will be generated and authors will have variety of possible offers to choose from. Thus, the authors of manuscripts will have variety of possible offers to choose from.

V. CONLUSION

The article deals with problems related with determination of offers for printing of manuscripts. The basic parameters of manuscripts related with pictures, formulas, tables and overall symbols are taken into account in generating the offer. Except the parameters of manuscripts, the parameters of printed book are also considered. This includes color or black and white printing, paper format, paper type, paper weight, line height, and used font. These two main features are used to answer the question whether the manuscript could be considered as monograph in accordance to definition of the law for academic staff development in Republic of Bulgaria. For the goal, a two-stage algorithm for determination if the manuscript could be considered as monograph is proposed.

The numerical testing is done by using the input data for two manuscripts with different parameters. The results show that one of them can be considered as monograph, while the second one do not meet the definition given by the law for academic staff development in Republic of Bulgaria. This proves the functionality of proposed algorithm together with described architecture of decision support system for calculation of offers for books/monograph printing realized as web-application.

The future developments are related with integration of proper database to store the parameters of manuscripts. In such way, it will be easier to load the parameters of manuscripts and by setting of different printing options different offers will be generated and authors will have variety of possible offers to choose from.

REFERENCES

[1] L. Kirilov, V. Guliashki, and B. Staykov, "Web-based decision support system for solving multiple-objective decision-making problems", Technological Innovations in Knowledge Management and Decision Support, 2019, pp. 150–175. DOI: 10.4018/978-1-5225-6164-4.ch007.

[2] D. Borissova, I. Mustakerov, and D. Korsemov, "Business intelligence system via group decision making", Cybernetics and Information Technologies, vol. 16(3), 2016, pp. 219–229.

[3] D. Borissova, and I. Mustakerov, "An integrated framework of designing a decision support system for engineering predictive

maintenance", Int. Journal Information Technologies & Knowledge, vol. 6(4), 2012, pp. 366–376.

[4] S. Øiestad, and M. M. Bugge, "Digitisation of publishing: Exploration based on existing business models", Technological Forecasting and Social Change, vol. 83, 2014, pp. 54–65. https://doi.org/10.1016/j.techfore.2013.01.010.

[5] R. Andreev, D. Borissova, A. Shikalanov, and T. Yorgova, "Model-driven design of eMedia: Virtual technology transfer office". In: Lugmayr A., Stojmenova E., Stanoevska K., Wellington R. (eds) Information Systems and Management in Media and Entertainment Industries. International Series on Computer Entertainment and Media Technology. Springer, Cham, 2016, pp 279-298. https://doi.org/10.1007/978-3-319-49407-4_14.

[6] V. Shalamanov, "Institution building for IT governance and management", Information & Security: An International Journal, vol. 38, 2017, pp. 13–34.

[7] V. Shalamanov, "Towards effective and efficient IT organizations with enhanced cyber resilience", Information & Security: An International Journal, vol. 38, 2017, pp. 5–10.

[8] D. Borissova, and I. Mustakerov, "A framework for designing of optimization software tools by commercial API implementation", Information Journal of Advanced Engineering, Management and Science, vol. 2(10), 2016, pp. 1790–1795.

[9] D. Borissova, E. Tabakova, and V. Grigorova, "An architecture of web-based application for thyroid disease identifying", Trends Journal of Sciences Research, vol. 2(1), 2015, pp. 46–49.

[10] D. Borissova, D. Korsemov, and I. Mustakerov, "Multi-criteria decision making problem for doing business: Comparison between approaches of individual and group decision making", In: Saeed K., Chaki R., Janev V. (eds) Computer Information Systems and Industrial Management. Lecture Notes in Computer Science, vol. 11703, 2019, pp. 385–396.

[11] I. Mustakerov, and D. Borissova, "A Web application for group decision-making based on combinatorial optimization", In: Proc. of 4th International Conference on Information Systems and Technologies, 2014, pp. 46–56.

[12] B. Luey, "Publishing as medium", International Encyclopedia of the Social & Behavioral Sciences (2nd Ed.), 2015, pp. 620–623. https://doi.org/10.1016/B978-0-08-097086-8.95029-3.

[13] Q. Yun, "Digital Publishing: Challenges and Opportunities", China's Publishing Industry. From Mao to the Market. 2019, pp. 197–219. https://doi.org/10.1016/B978-0-08-100919-2.00007-3

[14] S. Wilson-Higgins, "Trends in academic book publishing that impact on-demand", The Impact of Print-On-Demand on Academic Books. Chandos Information Professional Series, 2018, pp. 109–118. https://doi.org/10.1016/B978-0-08-102011-1.00008-0.

[15] Fu Wen-Qia, Zhang Meib, and Yan Ling-Yanc, "Academic e-book publishing in China: An investigation of current status and publishers' attitudes", The Journal of Academic Librarianship, vol. 44, 2018, pp. 15–24.

[16] Chin-Chao Lin, Wen-Chih Chiou, and Shih-Sin Huang, "The challenges facing e-book publishing industry in Taiwan", Information Technology and Quantitative Management (ITQM2013). Procedia Computer Science, vol. 17, 2013, pp. 282–289.

[17] E. Pesikov, O. Zaikin, and E. Kushtina, "Strategic decision risk management in a virtual publishing house on a basis of computer modelling", IFAC Proceedings, vol. 42(4), 2009, pp. 87–90. https://doi.org/10.3182/20090603-3-RU-2001.0405.

[18] M. Glykas, "Workflow and process management in printing and publishing firms", International Journal of Information Management, vol. 24, pp. 523–538, 2004.

[19] M. Mustonen, "Direct publishing and the bargaining between the author and the publisher", Information Economics and Policy, vol. 48, 2019, pp. 40–47. https://doi.org/10.1016/j.infoecopol.2017.10.003.

[20] S. Dimitrov and T. Stoilov. "Loading test of Apache HTTP server by video file and usage measurements of the hardware components", In Proc. of 14th Int. Conference on Computer Systems and Technologies, CompSysTech'13, 2013 pp. 59–66. https://doi.org/10.1145/2516775.2516799.

[21] K. Stoilova, T. Stoilov, and V. Ivanov, "Bi-level optimization as a tool for implementation of intelligent transportation systems", Cybernetics and Information Technologies, vol. 17(2), 2017, pp. 128–133.

ERP Solutions in Cloud Technologies as a Driver for Digital Transformation of Businesses

S. Križanić, T. Šestanj-Perić and A. Kutnjak
Faculty of Organization and Informatics, Varaždin,
University of Zagreb, Croatia
skrizanic@foi.unizg.hr; tperic@foi.unizg hr; akutnjak@foi.unizg.hr

Abstract - **Introduction of digital platforms has become an important part of digital transformation of businesses. Typically, these platforms include, among other things, cloud technology-based solutions. The growth of data streams caused by digital technologies results with large amounts of data that need to be processed and stored. Enterprise Resource Planning systems (ERPs) as a common type of information systems which process and handle business data aim to do this with the lowest costs. The role of cloud technologies in this context is the promise to contribute to this aim. This article examines the existing literature on the role of ERP solutions in cloud technologies as a driver for digital transformation of businesses. The research questions addressed in this paper are following: How does the digital transformation paradigm affect the application of ERPs in enterprises by using cloud technology? How do cloud based ERPs impact enterprise profitability? Which issues have been reported on this subject? The aim of this article is to explore the role of cloud technologies in digital transformation of businesses from the aspect of shifting from owning ERP solutions physically on the user-side to hosted ERPs on a remote service.**

Keywords - digital transformation; cloud technologies; ERP; information systems

I. INTRODUCTION

Today, cloud technologies have found their place in various business segments of many companies. Transition to cloud based enterprise solutions is no longer something unusual, as such solutions are known to contribute to lower initial deployment costs, faster implementation, and no need for internal software and hardware implementation. This article examines the existing literature on the impact of the digital transformation paradigm on the application of Enterprise Resource Planning systems (ERPs) using cloud technology. Digital transformation for an enterprise means adapting to the impact of digital technologies and integrating them into the organization's external interface and ongoing internal processes [1]. The most commonly used information systems that support business processes are related to ERPs, Customer Relationship Management (CRM), Supply Chain Management (SCM), and similar systems [2]. According to [3] "ERPs strongly optimize data and processes, whereas soft factors like user experience and user engagement have not been developed with the same intensity". ERPs have traditionally been implemented as on-premise software, but today the cloud computing delivery model is increasingly attractive, because of its cost advantage as one of the reasons [4]. However, the advanced cloud based information models that enable cloud ERP providers to separate business data of a user company from its internal hardware and servers put some risk on user companies, such as: transparency of IT operations of cloud ERP provider and limited control [5].

The idea of this research is to determine which literature deals with cloud computing, ERPs and digital transformation in general. The research questions we are trying to answer through a literature review are:

1. How does digital transformation (DT) affect the application of ERPs in enterprises using cloud technology?
2. How do cloud based ERPs affect enterprise profitability?
3. Which issues have been noticed in the implementation of cloud based ERPs?

The motivation for writing this article stems from a curiosity to primarily identify the advantages, but also the disadvantages that cloud computing brings to the development of ERPs.

In addition, by in-depth analysis of selected articles, it is analyzed if there are recognized the determinants of digital maturity. These determinants by which we analyzed the gathered literature are: Strategy Orientation, Customer centricity, ICT and process infrastructure, Talent, capability and capacity strengthening, and Innovation culture and organizational commitment.

The paper is conceptualized as follows: the methodology section describes how the literature research was performed and cites the selected articles. Two scientific databases were used to search the literature: Scopus and Web of Science (WoS). The subsequent three sections process the literature related to research questions, and then there is a section that considers digital maturity in the context of cloud ERP application. At the end, there is short discussion section and the article is concluded with the final considerations.

II. METHODOLOGY

This chapter describes the research strategies applied in this article. The scientific literature databases used for the research were Web of Science (WoS) and Scopus. The same keywords were used to search both databases for literature published between 2014 and 2019 (Table I).

This research has been conducted as part of the wider research in the project Development of innovative platform for digital transformation of enterprises (RDI) which is funded by European Union through the European Regional Development Fund (ERDF).

TABLE I STRATEGIES DATABASE RESEARCH

WoS strategy of research	N^0 of results
ALL FIELDS: ("ERP" AND "digital transformation" AND ("cloud" OR "cloud computing")) Timespan: 2014-2019. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC.	3
Scopus strategy of research	**N^0 of results**
ALL ("ERP" AND "digital transformation" AND ("cloud" OR "cloud computing")) AND PUBYEAR > 2013 + ALL FIELDS	146

Searching only titles and abstracts for the keywords DT, cloud computing and ERPs, returned only 2 articles. Thus, we widened the keywords search to bodies of the articles (results in Table I). Scopus results were sorted by relevance, based on the number of appearances of keywords in all the fields. There were in total 16 articles included in the in-depth analysis, which showed that for the most part of them, only two of the three topics (keywords) are substantially covered (Table II, sign "-" indicates that the term is only briefly mentioned). Only four articles examine thoroughly all three keywords of our interest. This was a problem for comprehensive analysis of existing literature about the application of ERP solutions in cloud technologies, as a driver for advanced digital transformation of businesses. For this reason, we additionally analyzed this articles through the determinants of digital maturity.

TABLE II A LIST OF SELECTED ARTICLES THAT DEAL WITH CLOUD COMPUTING, ERPS AND DIGITAL TRANSFORMATION

Authors, Year of publication	Cloud computing	ERPs	DT
A. M. AlBar, and M. R. Hoque, 2017 [14]	+	+	-
K. Andriushchenko, et al., 2019 [21]	-	+	+
P. M. Asprion, B. Schneider, and F. Grimberg, 2018 [3]	-	+	+
C. Gebayew, I. R. Hardini, G. H. A. Panjaitan, N. B. Kurniawan, and N. B. Suhardi, 2018 [8]	-	-	+
J. Grabis, and J. Kampars, 2018 [22]	-	+	+
L. Kappelman, V. Johnson, R. Torres, C. Maurer, and E. McLean, 2018. [17]	+	+	+
V. Koscheyev, V. Rapgof, and V. Vinogradova, 2019 [1]	-	-	+
M. Makhlouf, and O. Allal-Chérif, 2019 [7]	+	-	+
Y. Masuda, S. Shirasak, S. Yamamoto, and T. Hardjono, 2018 [18]	+	+	+
Z. H. Pathan, Z. Jianqiu, U. Akram, Z. Latif, M. K. Khan, and M. Z. Tunio, 2017 [10]	+	+	+
D. Sokač, and R. Picek, 2019 [9]	+	+	+
H. Srimathi, and A. Krishnamoorthy, 2019 [19]	-	+	+
U. M. Z. Usman, M. N. Ahmad, and N. H. Zakaria, 2019 [23]	+	+	-
T. P. Ventulett, and L. M. Villegas, 2018 [24]	+	-	+
T. Walther, 2018 [20]	+	-	+
S.-K. Yoo, and B.-Y. Kim, 2018 [6]	+	-	+

Because of the lack of appropriate literature about cloud ERP and DT, we examined additional literature that covers the topic of using cloud ERPs in enterprises. Search of Scopus database using "cloud ERP" keyword, resulted with 88 articles. Those with more than 10 citations were selected for the review (table III).

TABLE III A LIST OF ADDITIONAL ARTICLES SELECTED FOR THE REVIEW

Authors, Year of publication
B. R. Chang, H.-F. Tsai, and Y.-C. Tsai, 2014 [15]
Y.-W. Fan, C.-D. Chen, C.-C. Wu, and Y.-H. Fang, 2015 [13]
S. Gupta, S. Kumar, S. Kumar Singh, C. Foropon, and C. Chandra, 2018 [11]
B. Johansson, A. Alajbegovic, V. Alexpoulos, and A. Desalermons, 2015 [4]
C. López Vargas, and A. Ishizaka, 2017 [12]
G. C. A. Peng, and C. Gala, 2014 [5]

III. IMPACT OF DIGITAL TRANSFORMATION ON THE APPLICATION OF ERPS, USING CLOUD TECHNOLOGY

The aim of this chapter is to explore what the existing literature says about the impact of digital transformation through the usage of cloud computing (technologies) on the application of enterprise resource planning systems.

Digital transformation strategies help companies to achieve competitiveness and sustainable growth [6]. They want to transform their information systems for faster integration of technological and sociological changes such as cloud computing, big data and mobility [7]. The determinants of cloud ERP adoption, which can help in building awareness regarding it, for example, are: complexity, compatibility, ICT skill, observability, ICT infrastructure, top management support, regulatory environment and organizational culture [14].

Some of the benefits of digital transformation for business are: increased customer satisfaction, improved customer experience, improved profitability, increased revenues from products and services, reduced costs [8].

Enterprise resource planning systems are affected by digital transformation and, therefore, new trends in the development of ERPs have emerged, in the form of technologies like cloud ERP and Internet of Things ERP [9]. Cloud computing is a platform using remote grid servers on the Internet to store, manage and process data, rather than on personal computers or local servers. Cloud computing has brought new applications or services to enterprises, like cloud Enterprise Resource Planning, cloud Customer Relationship Management and cloud Supply Chain Management [10]. There was also a survey with the aim to investigate essential factors of cloud computing adoption, in which CEOs, managers and IT experts from SMEs participated to exhibit the knowledge about technology adoption. The results showed that cloud computing adoption in SMEs "is influenced by 7 factors: relative advantage, cloud computing compatibility, dynamic complexity, regulatory support, competitive pressures, managerial support and firm size" [10].

There are two types of ERP solutions being employed so that organizations can use it, namely, an on premise ERP and cloud based ERP. In an on-premise ERP, it is company's responsibility to maintain its own IT infrastructure [11]. Cloud applications, also known as Software-as-a-Service (SaaS) applications, can take many

forms, ranging from simple personal document storage tools (e.g., Google Docs) to email systems (e.g., Google Apps Email) and business-wide applications (e.g., cloud ERPs) [5]. The cloud computing application makes the deployment and upgrade of IT-related resources easier, allowing a continuous and agile alignment of rapidly changing business needs with the enterprise ERPs [12]. Organizations that are satisfied with the existing ERPs usually retain their *status quo* and find the transition to a cloud based ERP system less valuable and high-risk [13].

The emergence of cloud computing is transforming the way ERP solutions are designed, developed, deployed, implemented, updated, maintained and even paid for [12]. In the cloud, ERP packages and related data are hosted by the third party vendors who manage and control the infrastructure. Businesses can use their ERPs and data as on-demand services using a Web browser, without physical system installations on local computers and storage the data on local servers [5]. Cloud ERPs can provide flexibility, cost-efficiency, scalability, adaptability, availability and configurability of data in any enterprise [14].

Survey was conducted in [23] on cloud ERP adoption in Nigerian manufacturing SMEs. They hypothesise the following relationship based on literature: security and privacy issues, complexity – negative relation to cloud ERP; cost savings, technological capability, compatibility with existing values, top management support, firm size (larger firm, more positive effect), pressure from competitors, government support – positive relation to cloud ERP. The respondents find only cost savings, compatibility and privacy issues as important in predicting cloud ERP adoption, which means that there are issues that managers currently do not, and should additionally consider when assessing the adoption of cloud computing technologies.

IV. THE INFLUENCE OF CLOUD BASED ERPS ON ENTERPRISE PROFITABILITY

Typical cost issues associated with on-premise ERPs can be significantly improved by deploying cloud solutions [5]. Small and mid-sized enterprises (SMEs) are well-suited to implement a cloud ERP as it positively reflects on profits. The long-term costs of implementing a cloud ERP are lower for SMEs, allowing them to reduce their overall IT costs [4]. Cloud solution providers can offer powerful IT infrastructure (e.g., hundreds of servers and huge data storage capacity) for ERP users and their growing amount of data [5]. Cloud based ERP solutions offer access to the state of the art infrastructure, IT expertise and mobility of services through a sustainable payment model for SMEs, which typically do not have the resources to absorb the costs for in-house maintenance, as opposed to large enterprises. In addition, the advantage for small and medium-sized enterprises is also in a better focus of all available resources on the essential areas of business and in increased competitiveness [4]. The maintainability of cloud ERP solutions is considered to be the most important criterion in the assessment of ERP vendors [12].

Under the complicated cloud infrastructure, moving ERP data from one cloud provider to another can be expensive and time-consuming [5]. Besides that, cloud ERP does not necessarily make a great financial sense to large enterprises, as large number of ERP users can results in expensive subscriptions fees, which could in the long term cause higher costs than maintaining an on-premises ERP system [4].

V. ISSUES NOTICED IN THE IMPLEMENTATION OF CLOUD BASED ERPS

It was thought that business managers and users may not have sufficient knowledge to offer in-depth insights into cloud ERP issues. In contrast, IT consultants and ERP professionals, who work in the frontier areas of the IT industry, were expected to have a deeper knowledge of the issues that occur in cloud ERPs [5]. Today, it is considered that the implementation of cloud ERP solutions and its success depend on people both within the organization (employees and top management) as well as those outside (ERP solution vendor). Involving employees in the implementation process facilitates the transition and allows employees to become familiar with the technology [11]. Many IT consultants are "effectively trained to have a good mixture of knowledge and experience on both ERP and cloud solution implementation" [5]. Lack of knowledge about how to work within a cloud environment can lead to a biased evaluation of cloud ERP adoption [13].

The service-oriented hosts in enterprises such as ERPs, are often faced with the crucial problem of unexpected downtime or system failure that can potentially cause data loss and system shutdowns. It is difficult for a host to transfer all data to another host in a timely and instant manner and continue its task as usual [15].

Choosing a cloud ERP system can be considered as more challenging than choosing an on-premises ERP [12]. Lacking successful adoption cases of cloud ERPs, issues and anxiety of adoption of such systems cannot be eliminated with attractive functionality [13]. When data from the ERPs are hosted by a third-party cloud provider, companies have less control than the cloud provider to access and download their most important enterprise and customer data. Such a lack of control in a cloud environment introduce further data security threats of enterprise customer data [5].

VI. DIGITAL MATURITY IN THE CONTEXT OF CLOUD ERP APPLICATION

Enterprises should be digitally mature in order to accomplish digital transformation. Digital maturity assumes that the enterprise already has established information technology and human capacity who are ready to raise the existing information system to a higher level i.e. to the digital transformation. Key determinants of digital maturity are [16]:

1. Strategy orientation, which includes vision, leadership and management;
2. Customer centricity, which consists of customers' experiences and prediction of customers' needs;
3. ICT and process infrastructure, which include information and communication resources and management of business processes;
4. Talent, capability and capacity strengthening, which include permanent investment in new skills, capacities and knowledge;
5. Innovation culture and organizational commitment, which consist of commitment to organizational and innovation culture, also as commitment to organizational factors.

According to [3] a recommendation to improve the innovation culture and organizational commitment would be in developing a cloud based ERP platform where companies could build their own ERP system by assembling appropriate Web services, which could be offered by different software vendors.

In Table IV, there are findings from the literature analysis that are partly or fully related to the determinants of digital maturity. In other words, there are features noticed in the articles that could be related to the determinants of digital maturity in the context of the cloud ERP system implementation as a step towards digital transformation of businesses.

TABLE IV RESEARCH METHODS AND FINDINGS FROM LITERATURE ACCORDING TO DETERMINANTS OF DIGITAL MATURITY

Determinants of digital maturity	Research methods and findings	References
Strategy orientation	Online survey (2017, EU IT executives, 276 respondents): DT ranks 2nd in importance, regardless of company size by revenues. Within the top ten IT investments, cloud computing (CC) ranked 3rd, and ERP 5th, larger companies by revenues invested more in ERPs. CC is also ranked high (3rd) within technologies that should receive more investment, spending on ERP is mostly considered sufficient. In general, for 2018 it was projected that spending on cloud services would make a much higher percentage of total budgets. The mostly used measure of IT performance was cost control/reduction, while revenue growth, productivity improvements or profitability measures are much less used.	[17]
	Case study (pharmaceutical company): Two critical success factors that can support an IT strategy by promoting cloud, mobile and digital computing in an enterprise are commitment from CIO and top management, and collaboration between the architecture and communities on digital platforms.	[18]
	Online survey (CEO, Business Operation Manager and IS managers, 136 respondents): In cloud ERP adoption, a high level of management support is considered to be the most important critical success factor.	[14]
	Questionnaire (CEO's, managers and IT experts of SMEs, 103 respondents): The timely adoption of cloud computing enables small and medium enterprises (SMEs) better business communications with their stakeholders, customers, suppliers and partners.	[10]
	Case study (oil and gas industry): Oil and gas industry is lagging behind in the use of IT due to safety concerns (problem of using equipment in hazardous areas). Testing three catastrophic scenarios proves that, besides the need to have devices that operate in specific dangerous environments, their interoperability on a shared cloud is required to improve efficiency, productivity and security.	[24]
Customer centricity	Theoretical paper: In the education system, the admission process, counseling, course registration, fee payment, student support services, attendance maintenance, exam results announcement and etc. are made online using ERP-supported systems.	[19]
	Theoretical paper (cement industry): Clients, who are using clouds to store their data, have the advantage of not having to invest in expensive hardware (such as servers) or in software support and updates.	[20]
	Theoretical paper (model development): Based on their original purpose, ERPs strongly optimize processes and data, while factors like user experience and user engagement have not been developed with the same intensity.	[3]
	Theoretical paper (model for managing the information infrastructure): Generation of a new design of digital ERP system, where the goal of this restructuring is faster response to the demands of customers and more flexible production due to individual user requirements. This can be achieved by using social responsibility, mobility, analytics and cloud technologies in the development of ERP platforms.	[21]
ICT and process infrastructure	Online survey (2017, EU IT executives, 276 respondents): Among the largest IT investments are investments in cloud computing (e.g., SaaS, PaaS, IaaS, etc.), ERP, analytics and related technologies.	[17]
	Online survey (CEO, Business Operation Manager and IS managers, 136 respondents): The technological context for implementing cloud ERP solutions includes ICT infrastructure, technological and communicational skills of employees that facilitate the flow of information within an organization.	[14]
	Survey and structural model (executives and staff members of IT corporations; 198 respondents): Task characteristics and technology characteristics were found to have a positive influence on the performance of cloud computing adoption.	[6]
	Theoretical paper: The Learning Information Service (LIS) standard is established between Learning Management System and ERP.	[19]
	Theoretical paper (construction companies): Operational model change needed to comply with digital transformation; this change includes focus on cooperation with client, use of variety of digital tools and new management system.	[1]
	Theoretical paper and framework: The cloud services for the Internet of Things can be used to ensure a smooth development process and proper functioning in the early stage of development.	[9]
	Questionnaire (CEO's, managers and IT experts of SMEs, 103 respondents): Cloud computing is very dynamic and offers business-oriented and cost-effective cloud services like SaaS, ERP, SCM, PaaS, IaaS.	[10]
	Theoretical paper: The digital transformation approach consists of three components - business process architecture, transformation process, and technical architecture. The transformation process defines the steps to be taken to redesign a traditional business process into a data business process.	[22]

Determinants of digital maturity	Research methods and findings	References
Talent, capability and capacity strengthening	Case study (pharmaceutical company): Specialized architecture framework performing architecture reviews and enabling alignment between IT strategy and each architectural solution in IS including digital IT solutions - this proposed framework tested on ERP issues; a specialized ERP board, digital board and CRM board in aligning IT strategy promoting cloud, mobile, and digital IT with target architectures in these areas.	[18]
	Survey questionnaire (manufacturing SMEs, 436 respondents): Cost savings, compatibility and lack of data privacy are some of the most important factors considered in predicting the technological context and adopting a cloud ERP solutions.	[23]
	Exploratory study (173 cases of companies): Cloud computing provides the ability to implement ambitious transformation programs, which is otherwise hampered by huge investments in information systems.	[7]
	Case study (oil and gas industry): Aggregating all information into the cloud enables numerous parties to gather and securely distribute information in real time.	[24]
	Questionnaire (CEO's, managers and IT experts of SMEs, 103 respondents): The lack of IT professionals in SMEs causes an increase in the complexity of cloud computing; example - adopting of SaaS (ERP), IaaS (Oracle's cloud) and PaaS solutions requires IT oriented professionals.	[10]
	Literature review: A skill is an individual ability to use technology like cloud computing to implement digital transformation on the organization.	[8]
Innovation culture and organizational commitment	Survey questionnaire (manufacturing SMEs, 436 respondents): Organizations with a higher level of technical capability are highly ranked for adopting cloud computing solutions.	[23]
	Theoretical paper (model development): Enterprises should be able to define the future role of their ERPs. More specifically, enterprises should define whether they wish to realize ERP implementation or digital transformation.	[3]
	Theoretical paper (model for managing the information infrastructure): Social responsibility, mobility, analytics and cloud technologies are the basis on which a digital enterprise is formed, but using the right tools does not make the enterprise digital. Enterprises have to be able to rebuild their information infrastructure according to the demands of the digital world.	[21]
	Survey and structural model (executives and staff members of IT corporations; 198 respondents): Technological readiness, organizational factors and environmental factors have a positive impact on the adoption of cloud computing, whereby environmental factors have a relatively greater impact which can provide a competitive advantage when starting a new business.	[6]
	Case study (pharmaceutical company): They identified 36 action items with architectural risks in ERP projects, and proposed solutions to mitigate those risks. They propose the creation of specialized Architecture Boards such as ERP Board, Digital Board and CRM Board, to improve the quality of the architecture, and use of enterprise portals to enhance communication with the employees who resist change.	[18]

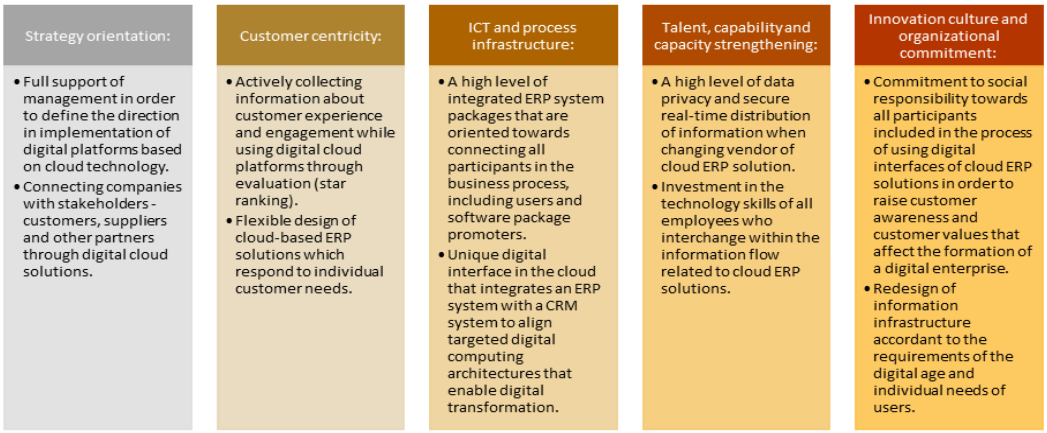


Figure 1 Guidelines how to reach digital maturity related to ERP platforms based on cloud computing.

Figure 1 shows an overview of our guidelines that outline the expected direction of the application of cloud ERP solutions in order to reach digital maturity levels of SMEs which use ready-made "all in one" ERP solutions in their businesses. Reached maturity level represents the basis for digital transformation.

VII. DISCUSSION

Since there is a number of different cloud ERP solution vendors on the market, the implemented IT solution should be carefully selected, considering the size of the enterprise and the business nature. The implementation of a cloud ERP solution in the form of a digital platform is a path toward digital transformation of an enterprise, through which attention should be paid to the determinants of digital maturity that can define the success of digital transformation. According to the definition of each determinant of digital maturity, we offer some guidelines that should be considered while implementing a cloud ERP solution. The design of a cloud ERP solution is one of the elements that makes the difference between the various solutions of cloud ERPs

on the market. However, design is not the most important requirement when choosing an ERP solution. Elements such as the cost of a cloud ERP solution implementation, the cost of monthly subscriptions and the security of information stored in the cloud technology are some of the most important elements when choosing an ERP solution.

VIII. CONCLUSION

Today, ERPs are increasingly popular in the form of cloud solutions. This paper highlights the advantages and disadvantages of using a cloud ERP system. Although the cloud ERP system has its advantages, it does not necessarily mean that these benefits will have an equal impact for all businesses. We have learned that enterprise size significantly influences the adoption of cloud ERP solutions. SMEs are more inclined to embrace the cloud based ERP solution, while large businesses need to weigh on their own whether it is worthwhile to maintain their existing ERP system or to implement a cloud based solution.

With the high level of information flows that are quickly exchanged inside and outside the company, businesses today are better informed about ready-made solutions in the market, making business decisions about resources, employees, customers, partners and suppliers more easily. However, what is the difference between the final solutions on the market today? Innovation is the element that distinguishes one cloud ERP solution provider from another and contributes to the competition on the market.

The main reasons for moving to a cloud ERP system are lower costs in the long run and the speed of implementing the new solution. The main problem identified is the lack of control on the user side. With the increase of the level of security and stability of the cloud system, as well as the increase of freedom of users who manage the ERP system, an increase in the number of enterprises that choose to implement cloud solutions is expected.

REFERENCES

[1] V. Koscheyev, V. Rapgof, and V. Vinogradova, "Digital transformation of construction organizations," IOP Conference Series: Materials Science and Engineering 497(1), 2019.

[2] S. Križanić, T. Šestanj Perić, and K. Tomičić-Pupek, "The changing role of ERP and CRM in Digital Transformation," Economic and Social Development, 41st International Scientific Conference on Economic and Social Development, Book of Proceedings, pp. 248-256, 2019.

[3] P. M. Asprion, B. Schneider, and F. Grimberg, "ERPs towards digital transformation," Studies in Systems, Decision and Control 141, pp. 15-29, 2018.

[4] B. Johansson, A. Alajbegovic, V. Alexpoulos, and A. Desalermons, "Cloud ERP Adoption Opportunities and Concerns: The Role of Organizational Size," in System Sciences (HICSS), 48th Hawaii International Conferece, pp. 4211-4219, 2015.

[5] G. C. A. Peng, and C. Gala, "Cloud ERP: a New Dilemma to Modern Organisations?," Journal of Computer Information Systems, vol. 54 (4), pp. 22-30, 2014.

[6] S.-K. Yoo, and B.-Y. Kim, "The effective factors of cloud computing adoption success in organization," in Journal of Asian Finance, Economics and Business, vol. 6 (1), pp. 217-229, 2018.

[7] M. Makhlouf, and O. Allal-Chérif, "Strategic values of cloud computing transformation: A multi-case study of 173 adopters," in Journal of Global Information Management, vol. 27 (1), pp. 128-143, 2019.

[8] C. Gebayew, I. R. Hardini, G. H. A. Panjaitan, N. B. Kurniawan, and N. B. Suhardi, "A Systematic Literature Review on Digital Transformation," 2018 International Conference on Information Technology Systems and Innovation, ICITSI 2018 – Proceedings, pp. 260-265, 2018.

[9] D. Sokač, and R. Picek, "Methodological approach of integrating the Internet of Things with enterprise resource planning systems," ICEIS 2019 – Proceedings of the 21st International Conference on Enterprise Information Systems 2, pp. 312-319, 2019.

[10] Z. H. Pathan, Z. Jianqiu, U. Akram, Z. Latif, M. K. Khan, and M. Z. Tunio, "Essential factors in cloud-computing adoption by SMEs," in Human Systems Management, vol. 36 (4), pp. 261-275, 2017.

[11] S. Gupta, S. Kumar, S. Kumar Singh, C. Foropon, and C. Chandra, "Role of cloud ERP on the performance of an organization, Contingent resource-based view perspective," in The International Journal of Logistics Management," vol. 29 (2), pp. 659-675, 2018.

[12] C. López Vargas, and A. Ishizaka, "GAHPSort: A new group multi-criteria decision method for sorting a large number of the cloud-based ERP solutions," in Computers in Industry 92, pp. 12-24, November 2017.

[13] Y.-W. Fan, C.-D. Chen, C.-C. Wu, and Y.-H. Fang, "The Effect of Status Quo Bias on Cloud System Adoption," in Journal of Computer Information Systems, vol. 55 (3), pp. 55-64, 2015.

[14] A. M. AlBar, and M. R. Hoque, "Factors affecting cloud ERP adoption in Saudi Arabia: An empirical study," in Information Development, vol. 35 (1), pp. 150-164, 2017.

[15] B. R. Chang, H.-F. Tsai, and Y.-C. Tsai, "High-Performed Virtualization Services for In-Cloud Enterprise Resource Planning System," in Journal of Information Hiding and Multimedia Signal Processing, vol. 5 (4), pp. 614-624, October 2014.

[16] I. Pihir, K. Tomičić-Pupek, and M. Tomičić Furjan, "Digital Transformation Playground – Literature Review and Framework of Concepts," in Journal of Information and Organizational Sciences, vol. 43, pp. 33-48, 2019.

[17] L. Kappelman, V. Johnson, R. Torres, C. Maurer, and E. McLean, "A study of information systems issues, practices, and leadership in Europe," in European Journal of Information Systems, vol. 28 (1) pp. 26-42, 2018.

[18] Y. Masuda, S. Shirasak, S. Yamamoto, and T. Hardjono, "Architecture Board Practices in Adaptive Enterprise Architecture with Digital Platform: A Case of Global Healthcare Enterprise," in International Journal of Enterprise Information Systems, vol. 14, pp. 1-20, January-March 2018.

[19] H. Srimathi, and A. Krishnamoorthy, "Integration of student system using iPaaS," in International Journal of Scientific and Technology Research, vol. 8 (9), pp. 602-606, 2019.

[20] T. Walther, "Digital transformation of the global cement industry," 2018 IEEE-IAS/PCA Cement Industry Conference, IAS/PCA 2018, pp. 1-8, 2018.

[21] K. Andriushchenko, et al., "Processes of managing information infrastructure of a digital enterprise in the framework of the ‹‹Industry 4.0›› concept," Eastern-European Journal of Enterprise Technologies, pp. 60-72, 2019.

[22] J. Grabis, and J. Kampars, "Application of microservices for digital transformation of data-intensive business processes," ICEIS 2018 – Proceedings of the 20th International Conference on Enterprise Information Systems, pp. 736-742, 2018.

[23] U. M. Z. Usman, M. N. Ahmad, and N. H. Zakaria, "The determinants of adoption of cloud-based ERP of Nigerian's SMEs manufacturing sector using TOE framework and DOI theory," in International Journal of Enterprise Information Systems, vol. 15 (3), pp. 27-43, 2019.

[24] T. P. Ventulett, and L. M. Villegas, "Enabling the best by preparing for the worst: Lessons from disaster response for industrial IoT in oil and gas," Society of Petroleum Engineers – Abu Dhabi International Petroleum Exhibition and Conference 2018, 2018.

Towards an Agile Framework for Business Intelligence Projects

M. Prouza*, S. Brodinová, A.M. Tjoa*
*TU Wien, Vienna, Austria
martin@prouza.at, brodinova@gmail.com, amin@ifs.tuwien.ac.at

***Abstract* - Business Intelligence (BI) solutions provide decision makers with information about their business and beyond as a utility to help them to make better decisions for their business purpose. Establishing such a BI system in a company can often be very complex and may end up in a financial disaster due to cancelling the project because of no valuable output after a specific time. Complex technologies, different stakeholders from various domains in a company, and vague requirements are often the initial situation for a BI project and reasons for delayed project delivery. Thus, being flexible and adaptable in the project are key factors for success. An agile approach can be the answer for a successful BI project. This paper deals with a conceptional agile framework for such a project. The focus is put on identifying yet unspecific or unknown aspects during a starting phase of a BI project.**

Keywords – Busines Intelligence; Agile; Agile setup; Agile Business-Intelligence; Scaled agile;

I. INTRODUCTION

Information is a valuable resource, especially for decision makers within an organization (e.g. a company, government institution) [1]. Business Intelligence (BI) solution targets to generate information from data and provide the information in a proper way to decision makers (i.e. end users) as a utility for their daily business and strategic decisions. A proper definition of BI for this article can be formulated as "*...an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance.*"[2].

Incorporating a proper BI system within an organization is nowadays highly needed to get the required information for decision support. The value of specific information increases when it is related to additional information beyond the own domain. For example, marketing promotions with finance and product sales. Moreover, the time aspect must also be considered in some cases in order to have the right information at the right time is necessary to increase the quality of decisions [3].

Establishing such a BI system in an organization can be challenging due to several reasons [4]. Firstly, stakeholders from different departments have individual interests. Secondly, heterogeneous source systems need to be consolidated and handled in various ways. Setting data from different domains and source systems in relation is usually not trivial. Moreover, the development team must deal with unclear and vague requirements especially at the beginning of the BI project. Finally, various skillset of the project members is needed to build up a BI system. For handling such a complexity within a project, a conventional plan-driven waterfall method may not be the best approach. An agile approach seems to be a more appropriate way to establish a BI system [3, 5]. Therefore, agile methodologies are often used in a complex and dynamic environment to facilitate on one hand the handling of uncertainties in defined processes, on the other hand to get fast feedback from end users. Fast feedback helps to facilitate that end users get what they really need, which is often hidden behind the requirements [6].

In recent years, agile approaches have been distributed very fast and a lot of frameworks such as Scrum [7], Large-scaled Scrum (LeSS) [8], disciplined agile delivery (DaD) [9], Kanban or Scaled Agile Framework (sAFE) [10] were introduced providing a set of processes, tools, and artifacts for an agile approach for a small project or the whole organization. Nevertheless, BI development cannot be considered as a conventional software project [11] (e.g. transactional system or online shopping platform) due to a complex architecture, the number of integrated tools, involved departments, and people.

The rest of the paper is organized as follows. Section II briefly reviews state-of-the-art of the BI architecture and highlights several important key facts that should be followed in order to achieve an agile project. Section III deals with agile approaches in a BI context. Section IV provides a conceptional framework with the focus on the initial implementation phase of a BI project with an agile approach. Finally, Section V concludes the paper.

II. THE NATURE OF BUSINESS INTELLIGENCE

A valid description of a BI architecture is provided by K. Collier [12] and consists of four broad layers: staging, integration, presentation and visualization. Each layer (i.e. stage) has its own purpose, targets and specific characteristics. The borders between these layers may shift a bit, depending on the concrete developing approach and tools. However, we strongly recommend distinguishing between these layers as it is essential for the agile framework (see Section V).

While the staging layer should be implemented very simply to get data fast, the second layer requires building advanced automated processes and being robust against changes at the same time. Indeed, the integration layer represents the core of the BI architecture where data with business value is kept over time. Traditionally, this layer is usually modelled in third norm technique. However, using the third norm technique does not follow the agile principles as the changes and adaptions of such a model during the project lifetime are very complex and expensive [5]. Therefore, an adaptable modelling technique is recommended for an agile approach. D. Linstedt and M. Olschimke [13] published the DataVault 2.0 modelling approach in 2013. Importantly, they introduced a new modelling approach of the integration layer inspired by the network of neurons. The approach could be described as a combination of the third norm and star schema modelling [14]. The flexibility is also an important characteristic for the information layer as the rules for transforming data into information can frequently change during the project. Finally, heterogeneous analysis and visualization tools, depending on the specific business purpose are typical for the last layer.

DataVault 2.0 generally encourages scalability of a data warehouse and the ability to change, adapt or extend the model during a project with no or reasonable costs. The approach is used for the presented framework where each layer is described in some detail.

A. Staging Layer – Layer 1

The first layer consists of the operational data sources and the staging area where source data are extracted to the data warehouse environment. Data from source systems arrives and technical metadata such as loading date, or the source system are added. If real time analysis is necessary for specific data (e.g. from Internet-of-Things sensors), this layer may also be omitted. The data in this layer are usually archived or deleted after they are loaded into the next layer.

B. Integration Layer – Layer 2

The integration layer (also called raw vault) is the basis of a BI system. The data are loaded from the previous layer using automated Extract-Transform-Loading (ETL) processes and persisted in prepared data structures [13]. The data from different business domains and source systems are integrated into this layer which means they are set in relation and combined. A business-oriented approach is used for integration with business keys describing each business object (e.g. a customer) [13]. The data structures in this layer are based on three basis entity types: hubs, links and satellites. Hubs usually contain a set of business keys an organization has like customer number or product number. Links represent the relations between business keys which are usually transactional data like sales. And entities from type satellite stores contextual data like product name, sales price or customer zip code. Figure 1. represents a simple model with all three entities and their relations. Data in this layer are kept in a raw manner, meaning that only hard rules are applied (e.g. standardize date formats). It is not recommended applying any business rules which

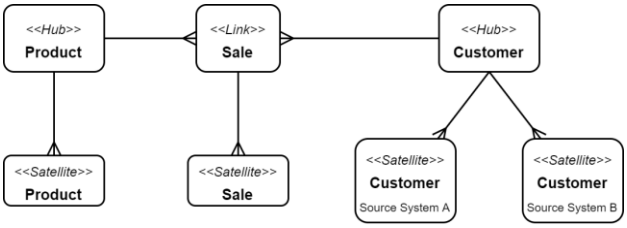


Figure 1. Hub, link, and satellite entities set in relation

results in changing the meaning of the data [13] which is part of the next layer. In some cases, it may be possible to perform analysis on the raw data but usually end users do not have direct access to the data [13]. Auditability of the reporting system is given due to the concept of technical historization.

C. Information Layer – Layer 3

The goal of the information layer is to conduct data preparations and transformations in order to achieve a comfortable platform of the information for the decision makers and their business purposes. The term information is used by [13] because data are now provided in an interpretable structure which delivers information to end user. Turning data into information is performed by applying business rules on the integration layer. In this layer, a lot of refactoring work comes through a development team and, therefore, this layer needs to be flexibly concepted. It should be emphasized that business rules or other incorporated logics may change over time. Indeed, these rules may even change during implementation due to better understanding of the business, the data and more concrete specification of the requirements.

In order to react on changes fast enough, D. Linstedt and M. Olschimke [13] recommended starting with non-physical views on the integration layer, which is usually fast and cheap to develop, instead of building complex expensive ETL processes. If performance issues require improvement steps on this layer like adding database indices, partitioning or the views are the necessary steps to be performed.

Conventionally information for business analyses is provided in a data mart or information mart, designed with a dimensional modelling technique [14], where specific business performance indicators such as sales and revenue can be analyzed in greater detail. The level of the granularity is defined by dimensions (e.g. states with corresponding regions or departments with sub-units).

In some cases, it might be appropriate to use raw data marts, directly applied on the staging layer. This kind of data marts is used for fast rollout of simple but usable reports which can be used for review, adaptions and further requirements [13]. Raw data marts might be not of a high acceptance for the end users such as decision makers as the quality of the data can be considerably low due to a lack of data cleaning or applying business rules or standards. This is very crucial, for example, in the financial domain due to demand for accurate financial indicators. In contrast to decision makers, data scientists see the advantages of raw marts as they often need access

to the data as soon as possible, even if data are raw and unclean.

D. Visualisation Layer – Layer 4

The last layer provides end-users with the prepared information from the previous layers (i.e. data mart or raw mart). The provided information is used for specific analysis (e.g. data mining) and reporting tools like dashboards or OLAP cubes (see more [15]).

III. Toward Agile Approach for BI Project

The essential idea of agile is to be flexible and fast enough to react on changes during the BI project. Therefore, agile follows an iterative approach aiming at providing potentially shippable functionalities after each iteration, even if such functionalities are not released to production at the moment. In another word, functionality is added continuously after each iteration till project end. A single iteration follows a time-boxed approach and usually lasts from one up to four weeks [7].

The concrete duration may depend on the specific organization and culture, used tools, involved people or the number of teams, and knowledge or project phase. One week is commonly used for the initial development phases to react fast on the frequent changes which are common at the beginning of the BI projects. In contrast, two or three weeks should be considered in later phases. Two weeks provides faster deployment cycles which conclude in faster access to new data for end users [13] and also have benefits when working with multiple individual teams. A three-week iteration approach can have benefits of focusing on design and development in the first week, followed by managing and testing in the second week, and finishing the iteration with the deployment and the acceptance of the delivery in the third week [13]. A four-week iteration approach should be used, when working in multiple individual teams. In such situation, each team has the opportunity to define their individual duration (i.e. one, two, or four weeks), while keeping an overall team iteration length of four weeks.

Figure 2. illustrates a simplified project approach using agile components starting with an overall initial project workshop at the beginning of implementation. Each iteration starts with a planning workshop, where the Product Owner - an agile role who is responsible for the requirements - shows the development team the desired working items (e.g. user stories or developer stories) for the next iteration. These items are already refined, estimated and prioritized in prior workshops with the end users and development team. After all necessary details

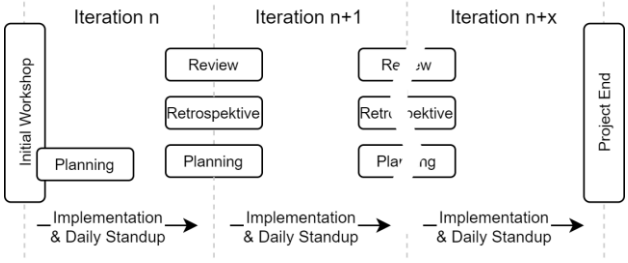


Figure 2. Simplified presentation of an agile approach

are clarified, the development team commits the work for the next iteration and starts with the implementation of the work items followed by short daily stand-up meetings every day between the team members. At the end of an iteration new functionality is presented to the end user by the developers in a review meeting and feedback is taken for further adaptions. The retrospective meeting gives the team the opportunity to inspect themselves, adapt, evolve increasing efficiency, quality, and motivation, during the project lifetime, The outcomes of the review meeting and the retrospective is used as input for the planning meeting for the following iteration [7, 8, 10].

Such iterative approach strongly requires a closer collaboration between all members of a BI project. These principles assure that accurate information is delivered to the right person at the right time, being one of the main goals of a BI system targets. Unsurprisingly, an agile approach seems to be far more suitable building up a BI platform than the traditional waterfall approach [5, 12].

A. Agile Mindset for BI Solution

The main characteristics of the agile mindset are flexibility, collaboration, periodical demonstrations and feedback. For example, K. Collier [16] stated that a very important factor is to involve management as well as end users in the collaboration. Showing and providing useable functionality in regular cycles has a very positive impact on a BI project and increases the trust in the BI project itself and the team members. Main advantages of an agile approach are, periodical demonstration of useable and practical software to end users, and the fast feedback to the development team if the solution follows the expectations and ideas of the end users. If expectations are not met, even in some fine details, an agile approach considers this in form of changes, which can be prioritized and implemented in one of the following iterations.

In traditional waterfall approaches change requests are often not welcome and may imply in a lack of requirements [17]. This is completely contrary compared to an agile mindset [18]. Detail description of the agile approach (e.g. events, team-roles, artefacts) are not part of this paper. Scrum guide [7] provides and good insight into the agile workflow.

B. Agile Requirements

Agile requirements are generally short and precise descriptions of system behavior and user needs, written in the form of so-called user stories [19, 20]. The user stories are the essentials in the agile methodology [20]. Instead of writing long and time-consuming documents, the user stories allow to react fast on the changes of the business rules and other requirements over time. Besides the user stories, regular agile refinement meetings are necessary to detail the requirements, which helps the development team with understanding the business needs in order to design and create a proper solution.

In the context of BI projects, the requirements for all the layers (i.e. staging, integration, and presentations and visualization) are commonly combined in one user story. Usually, the main focus of such user story is put on the description of the visualization layer (see Section II), more

specifically on the front end functionalities. Unfortunately, the excessive focus on the visualization layer only can considerably lower the quality of the final solution. A proper method is provided by [5] with introducing developer stories as a special type of requirements for back end (i.e. other three layers) behavior.

C. Developer Stories

Developer stories are not much different from user stories. Their focus is rather on the implementation of at least the staging and integration layer due to the specific characteristics, the used technology, and the needed skills for each layer. Therefore, the implementation of each layer can be a very powerful concept to increase the quality of the final solution [5]. However, the importance of involving the developer stories must be clearly communicated with members of the BI project, especially management. Some member might see this kind of stories and narratives as unnecessary additional work and not being aware of the advantages of such an approach. There are discussions about a direct business value of developer stories for the end users and, therefore, some experts might not consider as agile.

IV. Towards an Agile Framework

An efficient agile framework, especially at the beginning of a BI project, can be considered as a key success factor. An ill-designed project framework may lead to a lot of coding with no valuable output or mediocre quality. This section presents a framework, which forces fast integration and delivery of components with the flexibility to deal with uncertainties or changes during implementation. For simplicity, we use the notation layer 1,2,3,4 for staging, integration, information or visualization layer.

A. Prepare Development Phase

We strongly recommend including a preparation phase as an initial phase, also known as iteration zero where technical preparations and nontechnical prework is conducted. One or more workshops with the team members and business users (e.g. end users) should be held to get a backlog of user stories which must be prioritized by their business value. A unified vision of the expected outcome and scope must be clear to all team members and stakeholders as it is essential for a successful agile approach. Besides, the technical aspect must also be prepared like creating the infrastructures and configuration of the client software. The end of this phase starts with the first iteration planning meeting. In practice, the quality of requirements is of low quality at the beginning and the necessary information is often incomplete for the development. All these aspects should be communicated to team members as well as stakeholders.

B. Initial Development Phase

The initial development phase of a BI project is challenging – a new development team has to deal with a bunch of agile requirements where the quality will increase after certain iterations. The team usually has no specific knowledge about the domains and the data they will work with. Moreover, stakeholders tend to ask for results as soon as possible which raises the pressure on the team. Therefore, a proper setting of the agile approach can help the team work efficiently and deliver potentially shippable BI functionality early.

Developing BI functionality requires the implementation on all four layers where each layer depends on the previous one, we highly recommend using the concept of developer stories and splitting the requirements in manageable sizes. This concept helps the developers focusing on the specific characteristics of each layer.

Figure 3. visualizes an illustrative development cycle where new BI functionality is added within two iterations, and each layer must be implemented. Interestingly, Figure 3. shows that *Layer 1* with *Layer 2* and *Layer 3* with *Layer 4* must be implemented and tested together within one iteration which is optional but there is a simple and useful logic behind it. *Layer 1* and *Layer 2* have the same characteristics, following simple patterns which can be automated. In contrast, *Layer 3* and *Layer 4* are aiming at turning data into information by adding business rules and providing it to the end users with proper analysis tools. The described approach also works well under specific circumstances. Assuming, data are already integrated at the very beginning of a project, *Layer 1* and *Layer 2* can easily be omitted. If real-time integration for real-time analysis is needed, *Layer 1* can be omitted.

The main idea of this approach is to deliver fully operational and independent components as soon as possible and deploy it on a test system where the processes and integration can run in a regular manner. This also is important to gain experience and identify inconsistency early before the next iteration.

C. Dealing with Uncertainty

Uncertainties are usually not welcome in a software project and commonly in any kind of project. Dealing with uncertainty means dealing with the unknown, leading to failures and errors during implementation or testing. Consequently, the satisfaction of not only stakeholder but also team members may decrease during a project. Typical examples of uncertainty in the BI project are a lack of requirements, insufficient data profiling, unknown changing behavior of the data, and hidden business rules.

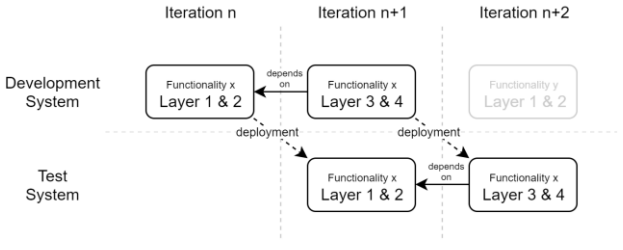


Figure 3. Development cycles for new BI functionality in an agile project

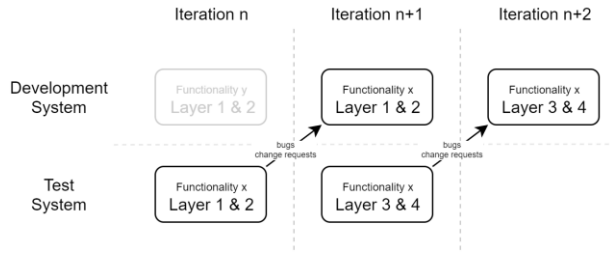


Figure 4. Consider uncertainties in an agile approach

Figure 4 demonstrates a suggested solution on how to deal with uncertainties in the BI project. A test system is used to test the functionality implemented in a previous iteration. ETL processes may show unexpected behavior, or the quality of a report is not as good as expected and adaptions are desired for the end user. These issues can be considered in one of the following iterations. Such a solution will work under several assumptions. Firstly, requirements must be split into achievable user and developer stories in order to recognize unpredictable behavior early. Secondly, the length of the development cycle needs to be setup appropriately. Finally, high priority bugs should be fixed as soon as possible, i.e. within the current iteration.

D. End User Collaboration

End user collaboration is an essential part of the agile principles [18] and a significant success factor establishing a BI system. One of the main targets of a BI system is to help end users, such as decision makers to make better decisions or data analyst to improve the quality of data they analyze. Generally, the BI system is used as a utility to make business more efficient due to faster access to actual information about a specific domain. Therefore, not only fast user access but also an active involvement of end users to the development cycle is strongly recommended.

Figure 5 demonstrates the concept of such a collaboration. At the end of each iteration, a review meeting with all stakeholder intents to present the finished work of an iteration. Giving end users the possibility to have access to the components as early as possible. For example, decision makers should have access to the visualization layer (i.e. *Layer 4*) while data analysts need to check the data in the integration or presentation layer (i.e. *Layer 2* or *Layer 3*) depending on the character of the analysis. This is a proper way to get fast feedback and accordingly correct requirements or change requests due to newly provided information. Such a concept has a great advantage over waiting too long till all layers are fully finished.

E. Scaling Teams

Agile teams are usually small teams between three up to nine members [7]. Some protagonists work with teams up to 15 members [21]. The optimal size depends on the individuals and the environment in a project. We recommend splitting the team into smaller teams of a proper size due to higher communication effort and complexity of coordination. The splitting strategy can vary from a project to a project and should evolve over

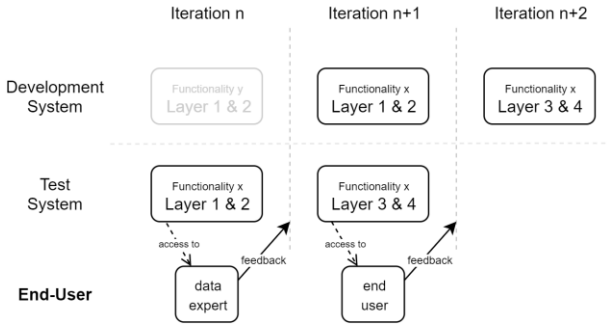


Figure 5. The role of end user in the development cycles in the agile BI project

time. For example, a layer-based splitting strategy can have two teams, one team working on *Layer 1* and Layer *2* and the second team developing *Layer 3* and *Layer 4.* Another common approach is a domain-based strategy where each smaller team works on all layers but with one domain only, e.g. financial domain.

V. CONCLUSION

In this paper, we address common challenges when establishing a BI project. The paper reviews technical and practical aspects of incorporating agile framework in order to manage the difficulties which arise due to a complex BI architecture. The discussed agile fundaments are, for example, predefined processes, events, a clear architecture with four separated, and flexible but robust data structure. We also focus on how to deal with the difficulties which arise in the initial phase of a BI project and which are often carried along the whole project. The recommended solution is, for example, the end user collaboration during project time. This helps to create a single perspective on the expected results for the team members and gives the possibility for direct feedback to the developers about the solution. Several aspects, such as the size of a team or the length of the development cycles, are discussed in order to choose an appropriate setup, depending on the specific characteristics of the BI project.

VI. REFERENCES

[1] E. Turban, *Decision support and business intelligence systems*, 9. ed., internat. ed. ed. Boston, Mass. [u.a.]: Boston, Mass. [u.a.] : Pearson, 2011.

[2] G. Inc., "BI Glossary," Glossary. [Online]. Available: https://www.gartner.com/en/information-technology/glossary/business-intelligence-bi.

[3] M. Muntean, "Agile BI – The Future of BI," *Informatica Economică,* vol. volume 3, pp. 114-124, 09/30 2013, doi: 10.12948/ISSN.

[4] D. Larson and V. Chang, "A review and future direction of agile, business intelligence, analytics and data science," *International Journal of Information Management,* vol. 36, no. 5, pp. 700-710, 2016, doi: 10.1016/j.ijinfomgt.2016.04.013.

[5] R. Hughes, *Agile Data Warehousing Project Management: Business Intelligence Systems Using Scrum*. Elsevier Science, 2012.

[6] P. Serrador and J. K. Pinto, "Does Agile work? — A quantitative analysis of agile project success," *International Journal of Project Management,* vol. 33, no. 5, pp. 1040-1051, 2015, doi: 10.1016/j.ijproman.2015.01.006.

[7] K. Schwaber, *Agile Project Management with Scrum*. Sebastopol: Sebastopol : : Microsoft Press, 2009.

[8] C. Larman and B. Vodde, *Large-scale scrum : more with less* (Large-scale scrum). Boston: Boston : : Addison-Wesley, 2017.

[9] S. W. Ambler and M. Lines, *Disciplined agile delivery : a practitioner's guide to agile software delivery in the enterprise* (Disciplined agile delivery). Upper Saddle River, N.J.: Upper Saddle River, N.J. : : IBM Press, 2012.

[10] R. Knaster and D. Leffingwell, *SAFe 4.5 distilled : applying the Scaled Agile Framework for Lean enterprises* (SAFe 4.5 distilled). Boston: Boston : : Addison-Wesley, 2019.

[11] W. Yeoh and A. Koronios, "Critical Success Factors for Business Intelligence Systems," *Journal of Computer Information Systems,* vol. 50, no. 3, pp. 23-32, 2010, doi: 10.1080/08874417.2010.11645404.

[12] K. Collier, *Agile analytics : a value-driven approach to business intelligence and data warehousing* (Agile analytics). [Place of publication not identified]: [Place of publication not identified] : Addison Wesley, 2012, pp. 1 online resource (xxxv, 329 p.) :, ill.

[13] D. Linstedt and M. Olschimke, *Building a Scalable Data Warehouse with Data Vault 2.0*. Morgan Kaufmann Publishers Inc., 2015.

[14] R. Kimball and M. Ross, *The data warehouse toolkit : the complete guide to dimensional modeling*, 2nd ed. ed. New York: New York : : Wiley, 2002.

[15] I. Bogdan-Andrei and P. Sorina, "Business Intelligence. A Presentation of the Current Lead Solutions and a Comparative Analysis of the Main Providers," *Database Systems Journal,* vol. V, no. 2, pp. 60-69, 2014.

[16] R. A. Mungree D., Morien D., "A Framework for Understanding the Critical Success Factors of Enterprise Business Intelligence Implementation," presented at the AMCIS 2013, 2013.

[17] M. Stoica, M. Mircea, and B. Ghilic-Micu, "Software Development: Agile vs. Traditional," *Informatica Economica,* vol. 17, no. 4, pp. 64-76, 2013, doi: 10.12948/issn14531305/17.4.2013.06.

[18] K. Beck *et al.*, "Manifesto for Agile Software Development," in *Manifesto for Agile Software Development*, ed, 2001.

[19] M. Cohn, *User stories applied : for agile software development* (Addison-Wesley signature series User stories applied). Boston: Boston : : Addison-Wesley, 2004.

[20] I. Inayat, S. S. Salim, S. Marczak, M. Daneva, and S. Shamshirband, "A systematic literature review on agile requirements engineering practices and challenges," *Computers in Human Behavior,* vol. 51, pp. 915-929, 2015, doi: 10.1016/j.chb.2014.10.046.

[21] S. Ambler, "Adapting Agile Methods for Complex The Agile Scaling Model (ASM): Adapting Agile Methods for Complex Environments," 01/11 2009.

A Different Approach for Clique and Household Analysis in Synthetic Telecom Data Using Propositional Logic

Sandro Skansi*, Kristina Šekrst ** and Marko Kardum***
* Faculty of Croatian Studies, Zagreb, Croatia
** Faculty of Humanities and Social Sciences, Zagreb, Croatia
*** Faculty of Croatian Studies, Zagreb, Croatia
sskansi@hrstud.hr

Summary – In this paper we propose an non-machine learning artificial intelligence (AI) based approach for telecom data analysis, with a special focus on clique detection. Clique detection can be used to identify households, which is a major challenge in telecom data analysis and predictive analytics. Our approach does not use any form of machine learning, but another type of algorithm: satisfiability for propositional logic. This is a neglected approach in modern AI, and we aim to demonstrate that for certain tasks, it may be a good alternative to machine learning-based approaches. We have used a simple DPLL satisfiability solver over an artificially generated telecom dataset (due to GDPR regulations), but our approach can be implemented on any telecom data by following the SAT encoding we have developed, and the DPLL solver can be substituted by a more advanced alternative such as CDCL. This paper extends the method presented in [1] for banking logs to data containing caller information, and proposes a more efficient encoding.

Keywords – Clique Detection; SAT Solving; DPLL; Household Identification; SAT encodings; Telecom Data Analysis

I. INTRODUCTION

The problem of clique detection (or informally, in the case of telecom data, reciprocal callers) is considered a rare event, unsuitable for machine-learning algorithms. Rare-event problems usually deal with unbalanced datasets, while most of machine-learning algorithms presuppose balanced classes in predictive classification. An example of using a non-machine-learning algorithm for rare events is [2], which relies on genetic algorithms, based on population-based optimization. Even though modern deep-learning methods hardly ever consider such occurrences, rare events are often interesting events pointing to unusual or peculiar behavior, or are able to give us insight into previously unknown relations. For example, detecting reciprocal payers in [1] may lead to fraud detection or uncovering money-laundering schemes.

The clique problem is one of well-known computational and optimization problems of finding complete subgraphs or cliques: subsets of graph vertices such that every two distinct vertices in it are adjacent. Clique detection – the task of finding such a clique – is NP-complete (NP stands for nondeterministic polynomial time), i.e. its validity can be quickly tested in polynomial time, but there is no known way to find a solution quickly (see [3] for more insight). If posed as a non-decision problem, clique detection is NP-hard, and due to it, exact solutions also exhibit exponential complexity. In order to solve NP-complete problems, we are left with brute-force search methods that check if every possible candidate in the list satisfies the given statement. Since such methods are unusable in large datasets with an exponential rise of computational complexity, various heuristic and approximation methods are often used for coarse-grained results. Because finding a solution to one NP-complete problem guarantees a solution to another of the same class, we aim to use a non-machine-learning AI algorithm based on Boolean satisfiability that performs better in changeable dynamic datasets.

In real-world settings, social networks are commonly represented as graphs, where vertices stand for people, the edges represent their acquaintances, and a clique is a subset of mutual friends, i.e. people that know each other. Such networks may be used to analyze business relationships or family relations, in order to improve sales and marketing techniques. In this paper we are focusing on cliques as household members, and our analysis can show that a detected clique may be really connected or just formally connected. We differentiate between *functional* and *official households*. Functional households comprise household members that actually live together, while official households are seen as households only on paper, for a variety of purposes, such as studying or working abroad, or even tax evasion.

Recently, the concept of a household has been of great importance in business and finance. For example, Guiso and Sodini [4] consider household economy as a new field since financial services and products used by households constitute a substantial part of financial industries in first-world countries. Households are today even more directly involved in financial decisions, such as pension funds, streaming accounts and services, or loan markets, which may be of interest to businesses in sales, marketing or banking. In order to detect households, we presuppose that household members communicate with each other more

often than with others. Even though it may seem counterintuitive that family members communicate more using phones since they live together and therefore their communication over the phone would seemingly decrease in quantity, recent research [5], [6] has addressed this issue, stating that mobile phones are regularly used, especially between parents and young people. Hence, we equate cliques to functional households using telecom data and presupposing their mutual communication.

By grouping people into households, such analysis may be used for marketing and sales, especially for telecom providers offering different plans and packages. For example, the services may be offered to one household leader to call rather than spend time on contacting three, four, five or even more household members. In order to do that, we require households to be functional, and not official, and a clique is a distinguishing feature of such households. Cliques outside of a household indicated a closed social network, and various standard social network analyses can be used to identify a household leader [7]. By identifying a household leader, a probability of successful product placement is thus increased.

II. SATISFIABILITY

SAT or the Boolean satisfiability problem is the issue of determining a satisfying assignment of a given Boolean formula. Usually, the formula needs to be in the so-called conjunctive normal form (CNF), in which we deal with a conjunction of one more clauses, where a clause is defined as a disjunction of possibly negated literals. That is, we are dealing with a conjunction of different disjunctions, often used in automated theorem proving. We need to note that the conversion into CNF may lead to exponentially large formulas, thus slowing the automation process. A clause $A_1 \vee A_2 \vee \ ... \ \vee A_n$ is valid iff there exist i, j such that $A_i = \neg A_j$. A CNF formula $c_1 \wedge c_2 \wedge ... \ c_n$ is valid if each of its clauses is valid.

The Boolean satifiability problem (SAT) was the first known NP-complete problem, proved independently by Cook [8] and Levin [9] in the early 1970's, when the notion of NP-complete problems still had not been established. Cook and Levin had shown that NP problems can be reduced to SAT problems for CNF formulas, which also includes clique-detection reductions [10]. The first SAT solvers were propositional truth tables developed by Charles Sanders Peirce [11] and Ludwig Wittgenstein [12]. The first real SAT algorithm was the Davis-Putnam algorithm [13], and a complete backtracking-based search extension was the Davis–Putnam–Logemann–Loveland (DPLL) algorithm from 1962 [14]. After more than 50 years, DPLL forms the basis of most modern SAT solvers today. Since decision problems inside the NP class have worst-case exponential complexity, efficient SAT solvers are constantly being improved, especially from the early 2000s [15] up to recent years [16].

DPLL algorithm chooses a literal, assigns a truth value to it, and thus by simplifying the formula, it recursively checks if it is satisfiable. If that turns out to be the case, the original formula is satisfiable, otherwise the same procedure is performed with the opposite truth value. If there is a unit clause, we have a unary constraint, and one promising assignment to the corresponding variable. While trying different assignments, the algorithm is building partial solutions, which might prove successful or not. The basic process of DPLL is:

DPLL Algorithm:

1. **Guess a variable.**
2. **Find all unit clauses from the last assignment and assign the needed value.**
3. **Loop step 2 until there is no change.**
4. **If the current assignment does not yield true for every clause, break from the recursion and try a different assignment.**
5. **If it is possible, guess another variable. The algorithm terminates if there is a solution, i.e. all clauses are satisfied, or step 5 is unable to be performed since there is no solution.**

Standard DPLL implementations use CNF files in the widely accepted DIMACS (named after the Center for Discrete Mathematics and Theoretical Computer Science) file format, which defines Boolean expressions stated in conjunctive normal forms. Such input file is a textual file containing comments, problem type and clauses of the parsed formula:

1. The file may begin with comment lines denoted with a lowercase *c*, which can occur throughout the file, but are usually placed in the beginning.
2. The problem line denoted with lowercase *p* consists of the problem type (*cnf* for CNF files) and the number of variables and clauses.
3. The list of clauses is given line by line, where each line is a clause, and clauses are separated by spaces, tabs or newline/carriage-return characters. A variable is defined as *i*, while its negated version is represented as *-i*. The variables are assumed to be numbered from 1 to *n*. The definition of a clause may extend to multiple lines; however, the end of a clause is determined with a final value of 0.

III. IMPLEMENTATION

In order to efficiently analyze reciprocal callers, we want to use hypergraphs in order to better streamline the analysis of functional and official households. A hypergraph is a generalization of a graph, where an edge can connect any number of vertices, that is a hypergraph H is a pair (V, E), where V is a set of vertices/nodes, and E is a family of subsets of V, known as hyperedges. Hypergraphs are usually represented using multi-dimensional arrays, and are often used in improving SAT solvers, with various optimization-methods such as decomposition [17] or [18].

As noted in the introduction, our approach is based on converting a dummy Telco dataset with standard parameters to a hypergraph, and then encoding it to a CNF formula written in the DIMACS file format, and then finally using a satisfiability-solving algorithm to find a

satisfiable assignment of the formula which corresponds to a clique of size *k*, where *k* is a parametrization argument. This in turn means that all *j* up to *k* will be checked in the final algorithm, resulting in *O(nk)*, where *n* is the complexity of the satisfiability algorithm, i.e. polynomially increasing the overall complexity.

The first question is about the dataset format. We will be needing a simple comma-separated values file (CSV) of the form:

numberFrom, IDcallerFrom, numberTo, IDcallerTo

The synthetic dataset we have used was made by randomly generating 50000 entries consisting of 8-digit strings in the numberFrom column, one number at the time with a 0.03 probability of resampling a number which has occured previously. The *numberTo* column entries where generated by randomly sampling from the *numberFrom* entries. The *IDcallerFrom* column was made by assigning random IDs to the numbers, with a 0.08 probability of repeating the same ID for different numbers. The *IDcallerTo* was inferred from the *IDcallerFrom* column.

Additional parameters such as time or date can be included, but we will not be using them in our approach. Whether filtering by additional parameters can improve the overall effectiveness of finding households we leave open for future research.

The conversion of this dataset to a hypergraph is quite simple. A first step is to notice that a conversion to a graph based on *IDcallerFrom* and *IDcallerTo* is trivial. A hypergraph is then obtained by adding all numbers (*numberFrom*) belonging to a *IDcallerFrom* and making a hyperedge between all of them and the *IDcallerTo*. In the same manner the *IDcallerTo* is expanded. The final hyperedge exists encompassing all *numberFrom* and *numberTo* iff there is a call between any pair of them.

We use a hypergraph since we want to be able to equate e.g. 12340001 calls 00001111 and 56782222 calls 00001111 when 12340001 and 56782222 belong to the same *IDcallerFrom*. We are not interested in the direction of the call, but we keep this information since it is the usual way Telco data is presented.

The next step is to encode this hypergraph in a CNF formula. We will use an encoding of the ID-graph (a graph using only *IDcallerFrom* and *IDcallerTo* where two nodes are connected iff there is call between the respective IDs) and then a simple expansion of clauses will create the needed hypergraph encoding. We follow a slight modification of the approach taken in [19] to encode the ID-graph to SAT, while the subsequent expansion to a hypergraph is a simple extension of the original encoding which is, to the best of our knowledge, our small contribution. The encoding of hypergraphs that we formulate is then just a simple generalization of the graph encoding. The idea behind the encoding is that we want a propositional variable $x_{j,m}$ to be true if and only if *m* is the *j*-th node in the clique. As we said earlier, the transformation is parameterized by clique size, and we denote it by *k*. If the final CNF is satisfiable, there is a clique of size *k*, and its nodes are exactly the components that are true. Get such an output, a proper encoding is needed, and two types of constraints are needed:

1. For each *j*, there is a *j*-th node in the clique, i.e. $x_{j,n1} \vee x_{j,n2} \vee ... \vee x_{j,nM}$, where nL, $1 =< L =< N$ are all nodes in the ID-graph

2. It is impossible that two nodes *n* and *m* not connected in the ID-graph both belong to the clique: $\neg x_{i,n} \vee \neg x_{j,m}$, for every node *n*, *m*, and every *j*, *n*.

The expansion to the hypergraph is then a simple matter of adding to every clause a copy of these constraints for every *numberFrom* and *numberTo* present in nodes that are connected in the hypergraph.

IV. ANALYSIS

To our knowledge, this is the first application of a SAT solver to telecom data in order to inspect and evaluate reciprocal callers. One possible objection might be that a simple lookup using a dictionary or a database query may seem faster on smaller datasets. However, such telecom data is not suitable for different methods since 1) datasets are big enough so that brute-force method complexity may rise exponentially 2) datasets are extremely dynamic and large, and may not be suitable for relational models. For example, relations per database limits are easily reached, along with maximal relation sizes and columns/rows per table. In case of simple scripting lookups, memory and disk space can be a huge issue, along with building possible dictionaries, while machine-learning methods are generally problematic for "needles in the haystack" data classification. Advances have been made in the cases of credit-card fraud detection, where fradulent activities are rare events, but in such cases there is a strong need of labeled datasets, and contextual information is needed to build accurate profile-based systems, which is often against privacy regulations [20]. Since different datasets and different providers may have different data quality, unbalanced datasets remain an issue in data mining and machine-learning procedures. Therefore, we used a propositional-logic algorithm to efficiently solve NP-complete consequences, in order to acquire a solution suitable for everyday practices and business purposes, hence eliminating unnecessary cost and resources.

V. CONCLUSION

Our goal was to show that artificial-intelligence approach not based on machine-learning may be used for telecom data analysis. An application of a Boolean satisfiability solver to telecom data may yield interesting relations for detecting different kinds of households and their properties: 1) do members of each household actually communicate with each other 2) if so, one can use this finding for business advantage, especially for marketing and sales departments by identifying functional households and their leaders. Our basis was the DPLL algorithm, which can be substituted by more advanced

counterparts, such as CDCL (conflict-driven clause learning) [21], in which backtracking to appropriate decision level is non-chronological. As a final note, this method and analysis may be used for other datasets and clique analysis of different social connections and statuses, in order to gain more insight for business and research.

BIBLIOGRAPHY

[1] S. Skansi, and B. Dropuljić. Identifying Reciprocal Payers in Banking Logs using SAT Solvers. IEEE Conference Publications: 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1500–1503. 2016.

[2] G. Weiss, and H. Hirsh. Learning to Predict Extremely Rare Events. AAAI Technical Report WS-00-05. 2000.

[3] A. Douik, H. Dahrouj, T. Y. Al-Naffouri, and M.-S. Alouini. A Tutorial on Clique Problems in Communications and Signal Processing. arXiv:1808.07102 [cs.IT]. 2018.

[4] L. Guiso, and P. Sodini. Household Finance: An Emerging Field. Handbook of the Economics of Finance. Elsevier B. V., pp. 1397–1532. 2013.

[5] K. Devitt, and D. Roker. The Role of Mobile Phones in Family Communication. Children & Society, vol. 23, pp. 189–202. 2009.

[6] M. Castells, M. Fernández-Ardèvol, J. Linchuan Quiu, and A. Sey. Mobile Communication and Society: A Global Perspective. The MIT Press. 2007.

[7] M. Zubair Shafiq, M. U. Ilyas, A. X. Liu, and H. Radha. Identifying Leaders and Followers in Online Social Networks. IEEE Journal on Selected Areas in Communications 31(9), pp. 618–628. 2013.

[8] S. A. Cook. The Complexity of Theorem-Proving Procedures. Proceedings of the 3rd Annual ACM Symposium on Theory of Computing, pp. 151–158. 1971.

[9] L Levin. Универсальные задачи перебора. Проблемы передачи информации. 9 (3): pp. 115–116. 1973. [Universal'nye perebornye zadachi/Problemy peredachi informatsii]

[10] N. Dahale, N. S. Chaudhari, and Maya Ingle. CLISAT – Clique Encodings in Implementation of SAT. Mathematical Sciences International Research Journal, vol. 3 (2), pp. 565. 2014.

[11] A. Biere, M. Heule, H. van Maaren, and T. Walsh (eds.). Handbook of Satisfiability. IOS Press. 2009.

[12] Wittgenstein, L. Tractatus Logico-Philosophicus. London: Routledge. 1974 [Original edition 1921].

[13] M. Davis, and H. Putnam. A Computing Procedure for Quantification Theory. Journal of the ACM, vol. 7, no. 3, pp. 201–215. 1960.

[14] M. Davis, G. Logemann, and D. Loveland. A Machine Program for Theorem Proving. Communications of the ACM, vol. 5, no. 7, pp. 394–397. 1962.

[15] M. Moskewicz, C. Madigan, Y. Zhao, L. Zhang, and S. Malik, Chaff: Engineering an Efficient SAT Solver. 39th Design Automation Conference. 2001.

[16] T. Toda, and T. Soh. Implementing Efficient All Solutions SAT Solvers. Journal of Experimental Algorithmics 1.12. 2016.

[17] We Li, and P. van Beek, Guiding Real-world SAT Solving with Dynamic Hypergraph Separator Decomposition. 16th IEEE International Conference on Tools with Artificial Intelligence, Boca Raton, FL, USA, 2004, pp. 542–548. 2004.

[18] F. Capelli, A. Durand, and S. Mengel. Hypergraph Acyclicity and Propositional Model Counting. International Conference on Theory and Applications of Satisfiability Testing 2014, pp. 399–414. 2014.

[19] M. Sipser. Introduction to the Theory of Computation. Boston: Thompson Course Technology. 2006.

[20] N. Yousefi, M. Alaghband and I. Garibay. A Comprehensive Survey on Machine Learning Techniques and User Authentication Approaches for Credit Card Fraud Detection. arXiv:1912.02629

[21] J. P. Marques-Silva; K. A. Sakallah. GRASP-A New Search Algorithm for Satisfiability. Digest of IEEE International Conference on Computer-Aided Design (ICCAD). pp. 220–227. 1996.

Sentiment analysis of tweets about COVID-19 disease during pandemic

Goran Matošević*, Vanja Bevanda**
* Fakultet ekonomije i turizma Dr. Mijo Mirković, Pula, Hrvatska, goran.matosevic@unipu.hr
** Fakultet ekonomije i turizma Dr. Mijo Mirković, Pula, Hrvatska, vanja.bevanda@unipu.hr

Abstract – **In this paper authors are presenting results of sentiment analysis of tweets – textual posts generated by users of Twitter in April 2020 in 6 selected countries: USA, UK, Spain, Italy, Sweden and Germany. There are two datasets formed in this research: first containing tweets about COVID-19 in selected countries, and second, containing tweets from top 10 politicians, users of Twitter in selected countries. Results show that most tweets from both data sources share similar positive sentiments, with minor differences between countries. Top emotions include "trust", "fear" and "anticipation".**

Keywords – sentiment analysis; twitter analysis; covid-19; corona virus

I. INTRODUCTION

Twitter is social network website where users can post short textual messages in which they express their opinion about specific topic. Sentiment analysis is process of extracting emotions and sentiments from text. Performing sentiment analysis on tweets is common research practice to gain insight into twitter users opinion towards selected topics – although it is not very reliable method, and results need to be taken with a grain of salt [10].

Corona virus spread begun in January 2020 and by April 2020 had affected almost all countries over the world [4]. The pandemic peaked in March and April 2020, and by the time this paper was written, future trend is unknown. Most affected countries in April 2020 were USA, Italy, Spain and UK. Germany and Sweden was included in this research because of their specific response to this pandemic – Germany had one of lowest death rate in Europe, and Sweden is the only country which did not take high restriction measures, allowing the virus spread freely among population with aim to develop herd immunity [7].

In this paper tweets during highest pandemic of COVID-19 are extracted from Twitter and explored to gain insight into how twitter users (and exposed politicians) are communicating and which feelings they are expressing during these difficult times.

II. SENTIMENT ANALYSIS

Sentiment analysis is text analysis technique for extracting emotions from text [20]. It can be used to gain insight on customers opinions about products, brands or other topics by analysing their feedback, comments on social media, surveys or any other textual sources. Sentiment analysis focus on polarity of each text segment (comment, feedback, paragraph, etc.) that is analysed. Polarity can be negative, neutral or positive, or even have more detailed scales (like "very positive", "positive" and "slightly positive", etc.). Also, sentiment analysis can be used to extract emotions from textual data, like anger and happiness. For better understanding, when referring to polarity, term "sentiment" will be used in this paper, and when referring to emotions, term "emotion classification" will be used.

There are two main approaches on how sentiment analysis is done: using lexicons and using machine learning [11]. Lexicon based sentiment analysis uses predefined lists of words that are categorized by sentiment and emotions. For example words "good", "best" and "beautiful" express positive sentiment. In text analysis these words are then detected and counted to calculate the sentiment of text. The second approach, using machine learning, is more complex – it involves using textual data that are marked with sentiments and emotions which are then used in process of training with selected machine learning algorithm [14, 15]. Machine learning algorithm extracts important features from text and builds a model that can be used to classify new (unseen) text, based on rules that model learned in training phase. In this paper we are using lexicon based sentiment analysis, more precisely the NRC emotion lexicon [17].

Sentiment analysis results can be used in marketing, for social media monitoring, customer relationships management, brand monitoring, etc. It can help business to identify customers satisfaction, bad or good parts of products or services, customer reactions on various marketing actions, etc. [18] Business can use sentiment analysis to gain actionable insights not only by analysing social media content generated from their customers, but also their competitors. By listening customer opinions on competitor products or services, business can take advantage of these insights to find actionable areas in which they need to improve or apply additional marketing campaigns [19].

III. PREVIOUS RESEARCH

Using data from social networks to track viruses, gather more information about their spread and public opinions and sentiments has been topic of interest for many research. In this section previous work is presented in this order: work regarding sentiment and emotion analysis over text from tweets about different pandemics,

following by most important work that uses tweets for other applications (predicting the spread of pandemics), and lastly work related to tweets about COVID-19.

Authors of [2] in their paper investigated tweets about H1N1 during 2009-2010 in UK and its influence on vaccinations trends and usage of antivirus medicaments in UK. They found that most tweets are descriptive and share official information, but also some contained public health system critics. They analysis show scepticism against economic side of vaccinations and lower trust in antivirus medicaments. They also researched the impact of important public persons opinions on these topics.

Authors of [3] used data from Twitter to assess public awareness of WHO recommendations about Zika virus. Their dataset contained more than 1 million tweets gathered during 65 days period. They analysed words and hashtags used in tweets about zika virus and crated word clouds for visual analysis. Also, sentiment analysis showed that tweets with negative sentiment had majority (61%) in their dataset.

Another interesting approach of using data from twitter is done in [18], who built a model for prediction of number of people infected by zika virus for 1 week ahead. Their data set contained tweets from USA during epidemic. Results showed that number of infections can be predicted 1 week ahead with reasonable accuracy.

Authors of [9] used machine learning approach to extract sentiments from tweets about vaccinations against HPV virus. They collected tweets from November 2015 until March 2016 targeting hashtags related with HPV virus. They targeted only tweets on English language (while we target more languages and tweets from geographic locations). Sentiment analysis by tweet sentiment in this research is refined by further categorizing negative tweets in 5 categories: "safety", "efficacy", "resistant", "cost" and "other". They identified a breaking point caused by an article published in New York Times about HPV which raised the positive sentiment of tweets from 35.8% to 66.21%. Authors believe that sentiment analysis is not reliable indicator of public sentiment, but still important one which can help in recognizing important trends and help solving problems.

Another relevant article that investigates sentiments on twitter is [16]. They used machine learning to classify tweets about Alzheimer's Disease into 6 predefined categories that reflect how twitter users stigmatize people with Alzheimer's disease. Their model has accuracy between 86% and 95% and shows how sentiment analysis can be done on specific use cases and gain valuable results for interdisciplinary research.

In [9] authors analysed twitter data about flue spread in 2015. During one month (November 2015) they collected tweets from Twitter stream API about flue based on keywords „sick", „cough" and „flue". For sentiment analysis they used machine learning approach trying to predict flue spread in certain cities in Canada and USA. Their research showed that twitter can contain valuable data about flue spread and recommend to include social media data in FluWatch application. They showed the importance of analysing social media and compared results from twitter with results from traditional medical records concluding that twitter data have good potential in determining flue spread.

Similar article [8] also focuses on predicting pandemic based on interaction on social networks and posts by health organizations. They built a system with which they can collect information about particular pandemic from different sources in short amount of time.

In [5] authors used tweets to investigate sentiments on twitter about Ebola virus. Beside tweets, their data source was built using newspaper articles too. For grouping articles by topic machine learning clustering algorithm was used. Also, sentiment analysis was done by machine learning approach and results segmented by country showed that newspaper articles contained more emotions then tweets – which was, according to authors, a surprising result.

Authors of [1] conducted similar research on COVID-19 tweets. Their dataset ranges from 2nd February to March 15 and included tweets from users around the world. They segmented tweets in 4 basic topics: virus origin, virus source, impacts on people, countries and economy, and tweets about virus prevention and infection risk control. Their sentiment analysis shown that most tweets on these topics are positive, except tweets about death rates. Only tweets on English are extracted (using Twitter API). Our research differs on time frame and segmentation by country, and share similar results on positive sentiment in most cases.

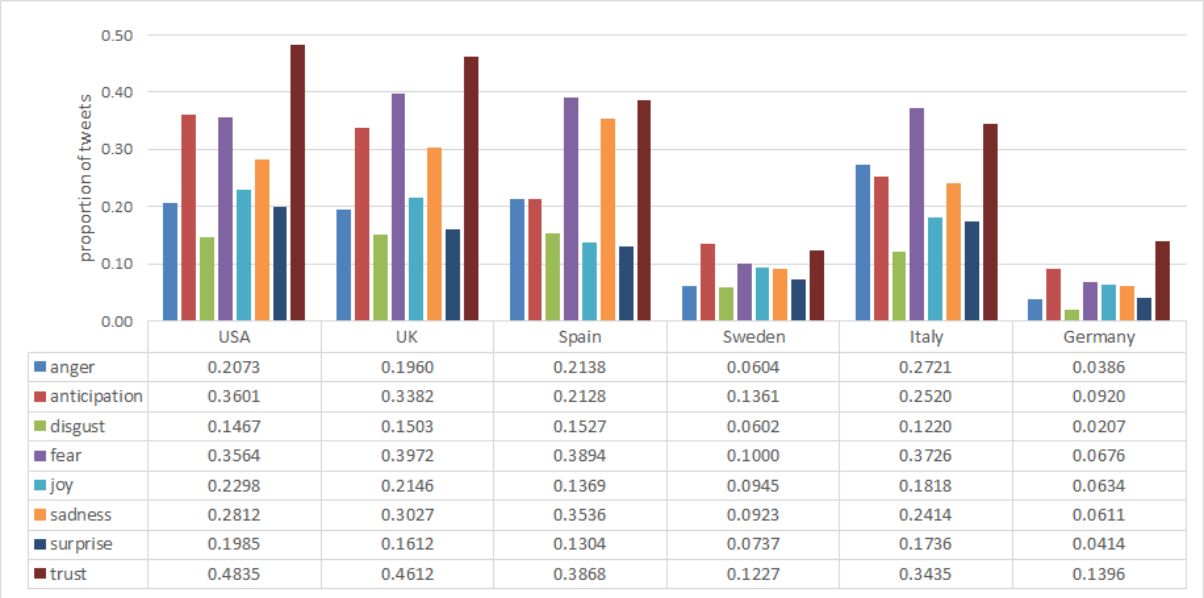
IV. RESEARCH METHODS

Two datasets are formed in this research: "Covid19" and "Politicians". The "Covid19" dataset was formed by random extract of 1,000 tweets per each day in April 2020 and per each 6 countries. Only tweets with "#covid19" are extracted. This results in 30,000 tweets per country, except for Sweden where this number is 6,783 – probably because of low number of users tweeting about Covid-19 in that country.

TABLE 1. NUMBER OF TWEETS IN "POLITICIANS" DATASET

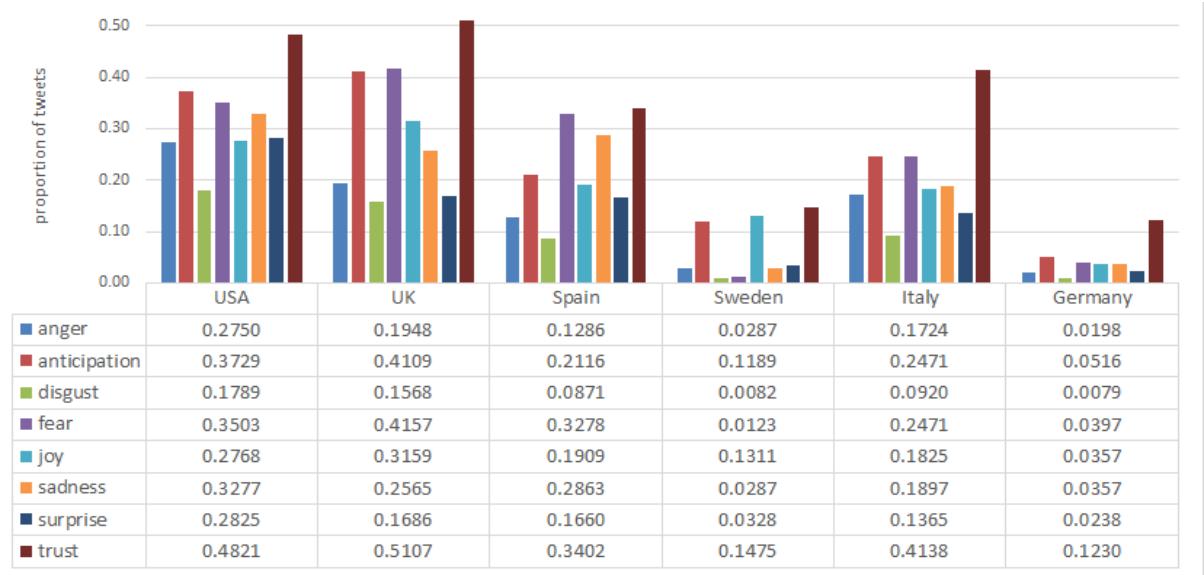
Country	Number of tweets	Percentage of tweets containing words related to pandemic
USA	531	19.96%
UK	421	8,31%
Italy	696	3,74%
Spain	241	10,79%
Germany	252	10,72%
Sweden	244	6,67%

The "Politicians" dataset was formed by extracting tweets of politicians from selected 6 countries in April 2020. List of politicians that actively use Twitter in



	USA	UK	Spain	Sweden	Italy	Germany
anger	0.2073	0.1960	0.2138	0.0604	0.2721	0.0386
anticipation	0.3601	0.3382	0.2128	0.1361	0.2520	0.0920
disgust	0.1467	0.1503	0.1527	0.0602	0.1220	0.0207
fear	0.3564	0.3972	0.3894	0.1000	0.3726	0.0676
joy	0.2298	0.2146	0.1369	0.0945	0.1818	0.0634
sadness	0.2812	0.3027	0.3536	0.0923	0.2414	0.0611
surprise	0.1985	0.1612	0.1304	0.0737	0.1736	0.0414
trust	0.4835	0.4612	0.3868	0.1227	0.3435	0.1396

Figure 1. Proportion of tweets in each emotion class per country from "Covid-19" dataset



	USA	UK	Spain	Sweden	Italy	Germany
anger	0.2750	0.1948	0.1286	0.0287	0.1724	0.0198
anticipation	0.3729	0.4109	0.2116	0.1189	0.2471	0.0516
disgust	0.1789	0.1568	0.0871	0.0082	0.0920	0.0079
fear	0.3503	0.4157	0.3278	0.0123	0.2471	0.0397
joy	0.2768	0.3159	0.1909	0.1311	0.1825	0.0357
sadness	0.3277	0.2565	0.2863	0.0287	0.1897	0.0357
surprise	0.2825	0.1686	0.1660	0.0328	0.1365	0.0238
trust	0.4821	0.5107	0.3402	0.1475	0.4138	0.1230

Figure 2. Proportion of tweets in each emotion class per country from "Politicians" dataset

selected countries is extracted from SocialBakers website[1] which monitors these statistics. For each country 10 profiles are selected. The number of tweets extracted is shown in Table 1.

Datasets were analysed with Syuzhet R package [2] which uses NRC method for sentiment analysis [17]. NRC method uses word lexicon which categorizes words in 8 basic emotions: anger, fear, anticipation, trust, surprise, sadness, joy, disgust, and two basic sentiments (polarity): positive and negative. It's a binary word-emotion table built by manually annotating words in selected language. Emotional Lexicon Sentiment scores, which represent word count, are binarized and sentiment is calculated by this rule: if tweet contained more positive than negative words, sentiment is positive, otherwise it is negative. If tweet contained the same amount of positive and negative words, sentiment is set to neutral. Also, neutral tweets are those who didn't have any positive or negative words, which may mean that sentiment was unable to identify with NRC method.

Tweets were pre-processed to exclude special characters, URL-s, hashtags and user tags. To get insights into word frequencies used in both dataset, word clouds [6] are generated (Figure 4). Appropriate text pre-processing was done before word cloud creation (stop words removal).

[1] https://www.socialbakers.com/statistics/twitter
[2] https://cran.r-project.org/web/packages/syuzhet/index.html

V. RESULTS AND DISCUSSION

Results of emotion classification of "Covid-19" and "Politicians" dataset are graphically presented in Figure 1 and 2.

From Figure 1 one can conclude that in most countries "trust" and "fear" are top emotions expressed in tweets about COVID-19. Tweets from Italy and Spain, where pandemic was high, express "fear" as top emotion. On the other side, tweets from Sweden express "anticipation" as top emotion, following by "trust". Similar situation is in tweets from Germany. "Sadness" is also emotion with high peaks in most countries. Authors believe that low proportions of emotions in tweets from Germany and Sweden are due low quality of NRC lexicons for German and Swedish language, so these results must be taken with caution. By comparing data from Figure 1 and 2 one can conclude that public and politicians shared similar emotions when tweeting about Covid-19 disease.

The same issues is visible on Figure 3, which presents sentiment polarity of tweets. High neutral sentiment for Sweden and Germany needs to be interpreted as limitation of NRC sentiment analysis method for German and Swedish language. Nevertheless, one can see that in all countries expect Spain positive sentiment of tweets is in majority.

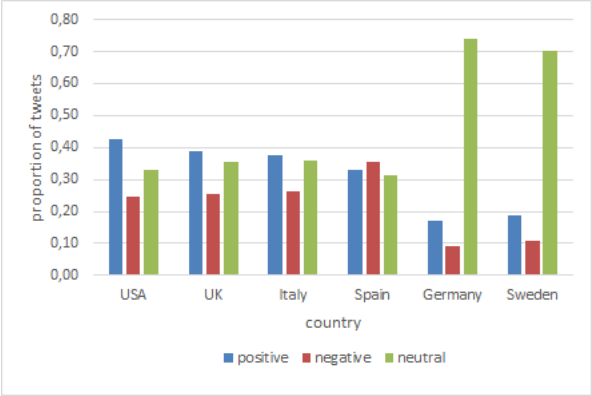


Figure 3. Polarity of tweets from "Covid-19" dataset

Since tweets for "Covid-19" dataset was collected daily in April 2020 (1,000 tweets per day), a time analysis of sentiment did not show any interesting trends change or patterns, except for USA, which is showed in Figure 4.

From Figure 4 one can see that on April 17 there was significant drop of positive tweets from USA and rise of negative ones. By exploring statistics about Corona virus cases, new cases, and deaths in USA in April, on April 17 there was no peaks in either. Maybe exploring news headlines, politicians actions or events in USA at that time could reveal potential causes – which was out of the scope of this paper, but it's a suggestion for further work.

Word frequencies analysis for "COVID-19" dataset are presented with word clouds on Figure 5. They reveal that twitter users from Spain and Sweden used their country name most frequently in tweets about COVID-19. Similar patterns are present in Germany and Italy, while USA and UK have the word "people" in focus. USA users of twitter mentioned their president Trump most frequently. Italian president Conte was also in focus by twitter users from Italy. Other words frequent in all countries include: "health", "today", "deaths" and "home".



Figure 4. Sentiment polarity change of tweets from USA over time

By examining tweets from politicians in these countries (the "Politicians" dataset), this research revealed that they expressed positive emotions with these frequent words: "government", "health", "today", "workers", and of course "trump" by USA politicians. This may indicate that politicians communicated the most about government decisions that affected health of their nation and economic situations. Table 1 shows the proportion of tweets about pandemic in this dataset based on words presence related to pandemic: covid, covid-19, corona, virus, coronavirus, sars-cov-2, stayathome (with similar translations in other languages), pandemic and epidemic. These proportions are smaller then author expectations indicating that politicians did not use these words often when posting on Twitter (except for USA dataset where this percentage is higher then other countries).

VI. CONCLUSION

With sentiment analysis of words used in tweets during COVID-19 pandemic in April 2020 from selected 6 countries presented in this paper authors showed how emotions expressed in tweets differ from each country. In 5 of 6 analysed countries positive sentiments are expressed through "trust" and "anticipation" emotions. Spain was the only country where "fear" was the top emotion, following by "sadness". Polarity analysis revealed that positive sentiment is in majority in all countries except Spain, where negative sentiment is in slight majority. By analysing sentiment polarity change over time (daily) in selected countries, positive sentiment was in majority, following by neutral and negative. Research revealed that there was an interesting polarity change in USA dataset on April 17 that could be good topic for further research.

In countries where pandemic is handled in different way, like Germany (where health system controlled the

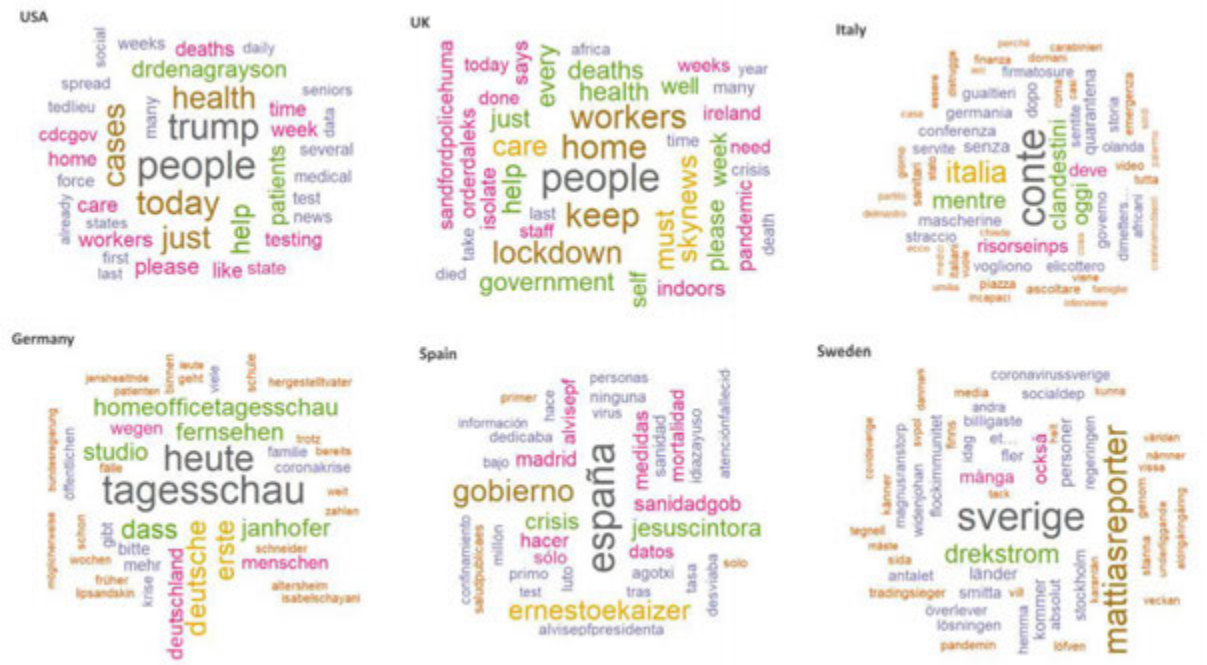


Figure 5. Most frequent words used in tweets about COVID-19 in selected countries

virus) and Sweden (which relies on population immunity), beside "trust" top emotion is "anticipation". Word frequencies analysis revealed that tweets about COVID-19 were used in combination with country names and national president names (Trump and Conte), while tweets from politicians contained words related to "government".

Results of these research can be used to better understand twitter users opinions about COVID-19 pandemic. Datasets can be further explored in different ways, to gain insight on how and what twitter users communicated during pandemic.

Limitations of this analysis is in dataset size and time frame when it was collected. It would be interesting to have data from wider time frame, to include before and after pandemic tweets and see how sentiments change over longer time.

LITERATURE

[1] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi and Z. Shah, "Top concerns of Tweeters during the COVID-19 pandemic: infoveillance study.", Journal of Medical Internet Research, vol. 22.4, 2020.

[2] A. McNeill, P.R. Harris and P. Briggs, "Twitter influence on UK vaccination and antiviral uptake during the 2009 H1N1 pandemic," Frontiers in public health, vol. 4, February 2016.

[3] C. J. Su and Y. Li, "Sentiment analysis and information diffusion on social media: the case of the Zika virus," in BIBE 2018, International Conference on Biological Information and Biomedical Engineering, VDE, 2018.

[4] Countries where COVID-19 has spread, https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/, accessed on May 1, 2020.

[5] E. H. J. Kim, Y. K. Jeong, Y. Kim, K. Y. Kang and M. Song, "Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news," Journal of Information Science, vol. 42.6, 2016, pp. 763-781.

[6] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds, " in 47th Hawaii International Conference on System Sciences, IEEE, 2014.

[7] I. Bremmer, C. Kupchan and S. Rosenstein, "Coronavirus and the Sweden Myth," The New York Times, https://www.nytimes.com/2020/05/04/opinion/coronavirus-sweden-herd-immunity.html, accessed on May 4, 2020.

[8] I. P. Cvijikj and M. Florian, "Understanding social media marketing: a case study on topics, categories and sentiment on a Facebook brand page," in Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, 2011.

[9] J. Du, J. Xu, H.Y. Song, and C. Tao, "Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data," BMC medical informatics and decision making, no. 17.2, July 2017.

[10] J. Jussila, V. Vuori, J. Okkonen, and N. Helander, "Reliability and perceived value of sentiment analysis for Twitter data," Strategic Innovative Marketing, Springer, Cham, 2017, pp. 43-48.

[11] K. Ahmed, N. El Tazi, and A. H. Hossny, "Sentiment analysis over social networks: an overview," in IEEE international conference on systems, man, and cybernetics, 2015.

[12] K. Byrd, A. Mansurov, and O. Baysal, "Mining twitter data for influenza detection and surveillance," in Proceedings of the International Workshop on Software Engineering in Healthcare Systems, 2016.

[13] M. A. Al-garadi, M. S. Khan, K. D. Varathan, G. Mujtaba, and A. M. Al-Kabsi, "Using online social networks to track a pandemic: A systematic review," Journal of biomedical informatics, vol. 62, 2016, pp. 1-11.

[14] M. Anjaria, M. and R. M. R. Guddeti, "A novel sentiment analysis of social networks using supervised learning," Social Network Analysis and Mining, vol. 4.1, 2014.

[15] M. Hammad, and M. Al-awadi, "Sentiment analysis for arabic reviews in social networks using machine learning," Information Technology: New Generations, Springer, Cham, 2016, pp. 131-139.

[16] N. Oscar, P.A. Fox, R. Croucher, R. Wernick, J. Keune, and K. Hooker, "Machine learning, sentiment analysis, and tweets: an examination of Alzheimer's disease stigma on Twitter," Journals

of Gerontology Series B: Psychological Sciences and Social Sciences, vol. 72.5, March 2017, pp. 742-751.

[17] S. M. Mohammad, and P. D. Turney, "Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon," in Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text, Association for Computational Linguistics, 2010.

[18] S. Masri, et al, "Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic," BMC public health, vol. 19.1, 2019.

[19] W. He, et al. "A novel social media competitive analytics framework with sentiment benchmarks," Information & Management, vol. 52.7, 2015, pp. 801-812.

[20] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams engineering journal, vol. 5.4, 2014, pp. 1093-1113.

Role of Business Intelligence Systems in Croatian Higher Education Quality Assurance

M. Cvitanušić Brečić*
* Agency for Science and Higher Education
marina.cvitanusic-brecic@azvo.hr

Abstract - **Quality assurance in higher education in Croatia is conducted by and in line with the European standards and guidelines developed by the European umbrella organization for quality assurance in higher education, European Association for Quality Assurance in Higher Education (ENQA). One of the standards for internal quality assurance directly relates to the information management stating that "Institutions should ensure that they collect, analyse and use relevant information for the effective management of their programmes and other activities." Internal quality assurance is, or should be, at the core of every higher education institution, influencing its performance and quality perceived by its users, but also external evaluators (e.g. government and professional bodies that regulate or influence the work of the institution). Higher education institutions can ensure the quality of their services only through a well-developed internal quality assurance system that can provide information for timely decisions. This paper will analyse the requirements and obligations that higher education institutions have regarding the collection, processing and analysis of data within the scope of their work. In addition, it will provide examples of existing and possible use of information systems that lead to improvement of internal quality assurance system of the higher education providers.**

I. INTRODUCTION

A. Business Intelligence Concept

Regardless of whether we define Business Intelligence (BI) in a broader sense, as an IT-based managerial philosophy [1], or in a narrower sense, as information technology that helps organisations manage business information, the ultimate role of the BI is the same – to provide reliable and up-to-date information for timely decision-making.

Although the term Business Intelligence is mostly used in relation to private/commercial businesses, the role it fulfils should also be important in decision-making at institutions operating in public interest, which is a fundamental role of higher education institutions (HEIs). It could be argued that the responsibility towards the taxpaying public is more important for publicly funded HEIs than for private ones, which are primarily responsible to their owners.

Whether we generally consider Business Intelligence as a process, or IT software solutions that support decision making, data is behind everything. Based on a large amount of data (sometimes referred to as Big Data) on all business processes, institutions should be able to assess their current position and decide on how to reach the intended goal, to the mutual satisfaction of employees and users of services. This goal is very similar to those of quality assurance.

B. Quality Assurance Concept

The concept of quality assurance (QA) has become an inescapable descriptor of business performance in the last two decades, both in private and public sector. Higher education systems, and HEIs in particular, have not been spared the new work paradigm. The notion of quality assurance is being introduced into the higher education systems throughout Europe, mostly within official procedures of assessing quality of HEIs. In the Croatian higher education system, the concept of quality assurance was introduced first as a part of prescribed procedure of institutional re-accreditation, and subsequently in all external evaluation procedures in Croatian higher education and science. An independent national public body, Agency for Science and Higher Education (ASHE), was established and modelled after the best European practices in quality assurance in science and higher education. ASHE is tasked with carrying out different external quality assurance evaluations, with institutional re-accreditation being the most (data) demanding, by evaluating HEIs' provision of education services against legal requirements, as well as institutional development of quality assurance and, ultimately, quality culture[1].

Although quality is not (all) about data, part of every external evaluation procedure is submission of considerable amount of data regarding core activities of the evaluated institution. Within the framework of each external evaluation procedure, the assumption is that the institution has relevant and reliable data about its activities[2]. The collection and managing of data on HEIs' core activities (study programmes, students, employees, etc.) and sharing that data with other stakeholders (ASHE, Ministry of Education and Science, National Statistics Bureau) are partly regulated by different national regulations. Assessing the quality of data on HEIs' activities, the way in which they are collected and analysed, i.e. used in decision-making, is part of the internal and external quality assurance system. Therefore we could argue that information system supporting the quality assurance system and management processes, helps enhancing the overall

[1] In its Glossary of Basic Terms and Definitions in Higher Education Quality Assurance, UNESCO defines *quality assurance* as "An all-embracing term referring to an ongoing, continuous process of evaluating the quality of a higher education system, institutions, or programmes." and for the *quality culture* states that "it refers to a set of shared, accepted, and integrated patterns/principles of quality to be found in the organizational cultures and the management systems of institutions."

[2] ESG, Standard 1.7. of the Standards and guidelines for internal quality assurance (*Institutions should ensure that they collect, analyse and use relevant information for the effective management of their programmes and other activities*).

success, and to provide assessable information about quality assurance system [2].

C. Business Intelligence in quality assurance

If we presuppose that all HEIs maintain data on their (business) activity, regardless of whether this is driven by need for improvement, or just complying with regulations, it can be asserted that HEIs have some kind of Management (Business Intelligence) information system. Some institutional systems are more advanced, but all of them maintain data on their respective business processes. How HEIs use their data in (strategic) decision-making procedures that contribute to quality assurance is subject to evaluation within external evaluation procedures, especially institutional re-accreditation and audit. This paper will outline all the requirements placed on HEIs regarding the collection and management of data that are part of their business system. Whether there is a possibility of developing better and improved Business Intelligence system for (strategic) decision-making in higher education, is a question that requires future research.

II. QUALITY ASSURANCE IN HIGHER EDUCATION

A. Quality assurance in higher education

Defining what quality assurance in higher education is, and which procedures and methods should be used to assure and improve it, is not an easy task, neither for national bodies tasked with monitoring quality assurance, nor for HEIs themselves. Quality does not refer to the same for students, their parents, society, labour market or HEIs. Each stakeholders group has its own perspective towards quality of higher education.

In order to evaluate perceived quality from the perspective of each of the stakeholder group, European Association for Quality Assurance in Higher Education (ENQA)[3] has, in cooperation with the European Students' Union (ESU), the European Association of Institutions in Higher Education (EURASHE) and the European University Association (EUA), developed The Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG) [3]. All (national) quality assurance agencies that aim at membership of ENQA have to undergo external evaluation by foreign experts consortium formed by ENQA in order to prove their ESG compliance in all their evaluation procedures. One of the biggest benefits of being an ENQA member is ensuring better assessment and (automatic) recognition of HE qualifications on the European level, and worldwide.

The ESG have 3 integral parts that provide guidelines for establishing quality assurance of a higher education system in a country:

Part 1: Standards for internal quality assurance,
Part 2: Standards for external quality assurance,
Part 3: Standards for quality assurance agencies.

Main differences between guidelines for evaluation criteria for internal and external quality assurance are provided in Figure 1.

European Standards and Guidelines for Internal Quality Assurance	European Standards and Guidelines for External Quality Assurance
1.1. Policy for quality assurance	**2.1. Consideration of internal quality assurance**
1.2. Design and approval of programmes	2.2. Designing methodologies fit for purpose
1.3. Student-centred learning, teaching and assessment	2.3. Implementing processes
1.4. Student admission, progression, recognition and certification	2.4. Peer-review experts
1.5. Teaching staff	2.5. Criteria for outcomes
1.6. Learning resources and student support	2.6. Reporting
1.7. Information management	2.7. Complaints and appeals
1.8. Public information	
1.9. On-going monitoring and periodic review of programmes	
1.10. Cyclical external quality assurance	

Figure 1. Main differences between guidelines for evaluation of internal and external quality assurance

Source: ESG (2015)

Basic idea behind the ESG and quality assurance system is that institutions have to build and enhance their internal QA system, which should support decision-making process on the basis of analyses of institutional procedures and data. External QA agencies, such is ASHE, are expected to evaluate the quality of internal QA system of HEI.

B. Quality assurance in Croatian higher education

Systematic quality assurance in higher education and science in Croatia began in 2010 with the re-accreditation of all Croatian HEIs, conducted by the Agency for Science and Higher Education. Becoming a full member of ENQA and being listed in EQAR (European Quality Assurance Register for Higher Education), both in 2011, the Agency proved its reliability as a quality assurance agency working in the European Higher Education Area. In the period 2010-2016, ASHE re-accredited all public and private HEIs in Croatia, in accordance with the national regulations, but also in line with the ESG.

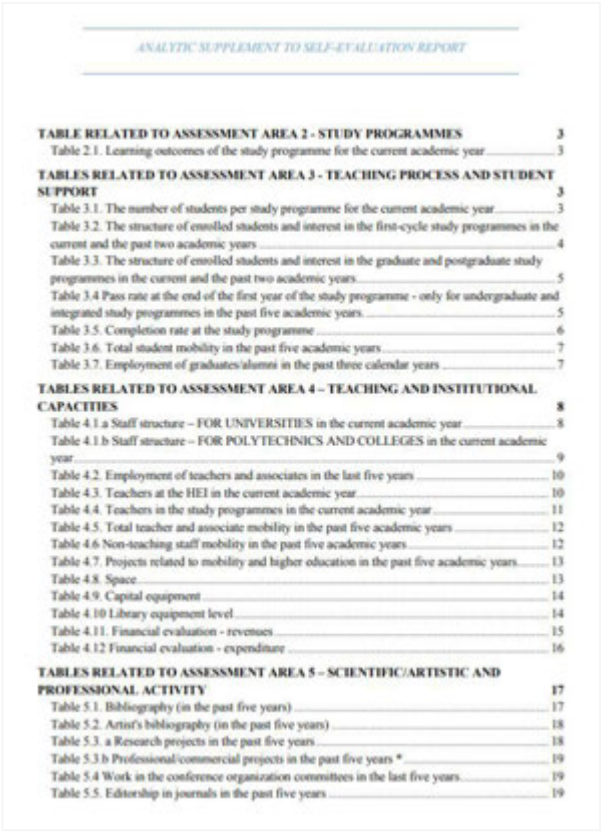
In order to provide relevant data and evidence within the procedure of institutional re-accreditation, but also other external evaluation procedures conducted by ASHE, HEIs have to submit a considerable amount of data about their (business) activities into ASHE's information system MOZVAG[4]. Collecting and analysing data on HEIs activities was found the most demanding and time-consuming part of the evaluation procedure(s), both for HEIs and ASHE.

Within the legally prescribed procedure of re-accreditation, which is the most demanding external evaluation procedure, HEIs have to submit a significant amount of data concerning all aspects of institutional activities that are being evaluated through national Standards for evaluation of the quality of HEIs, in line with the ESG. Areas being evaluated are: I) Internal

[3] ENQA is an umbrella organisation which represents its members at the European level and internationally, especially in political decision making processes and in co-operations with stakeholder organisations (https://enqa.eu/).

[4] ASHE information system for data collection in external evaluations, system description available here: *https://wiki.srce.hr/display/TUT2/Moduli+-+nova+verzija+Mozvaga*

quality assurance and the social role of the higher education institution, II) Study programmes, III) Teaching process and student support, IV) Teaching and institutional capacities and V) Scientific/artistic activity. To better understand the amount of data involved, Fig. 2 shows the content of the Analytical Supplement, which is a mandatory document in re-accreditation, and is automatically created from the MOZVAG system based on HEI's data entered manually or transferred from other systems. All data within the second assessment area - study programmes - are provided for each study programme that HEI delivers, which for some HEIs can be more than 30 programmes.



ANALYTIC SUPPLEMENT TO SELF-EVALUATION REPORT

TABLE RELATED TO ASSESSMENT AREA 2 - STUDY PROGRAMMES 3
Table 2.1. Learning outcomes of the study programme for the current academic year 3
TABLES RELATED TO ASSESSMENT AREA 3 - TEACHING PROCESS AND STUDENT SUPPORT 3
Table 3.1. The number of students per study programme for the current academic year 3
Table 3.2. The structure of enrolled students and interest in the first-cycle study programmes in the current and the past two academic years 4
Table 3.3. The structure of enrolled students and interest in the graduate and postgraduate study programmes in the current and the past two academic years 5
Table 3.4 Pass rate at the end of the first year of the study programme - only for undergraduate and integrated study programmes in the past five academic years 5
Table 3.5. Completion rate at the study programme 6
Table 3.6. Total student mobility in the past five academic years 7
Table 3.7. Employment of graduates/alumni in the past three calendar years 7
TABLES RELATED TO ASSESSMENT AREA 4 – TEACHING AND INSTITUTIONAL CAPACITIES 8
Table 4.1.a Staff structure – FOR UNIVERSITIES in the current academic year 8
Table 4.1.b Staff structure – FOR POLYTECHNICS AND COLLEGES in the current academic year 9
Table 4.2. Employment of teachers and associates in the last five years 10
Table 4.3. Teachers at the HEI in the current academic year 10
Table 4.4. Teachers in the study programmes in the current academic year 11
Table 4.5. Total teacher and associate mobility in the past five academic years 12
Table 4.6 Non-teaching staff mobility in the past five academic years 12
Table 4.7. Projects related to mobility and higher education in the past five academic years 13
Table 4.8. Space 13
Table 4.9. Capital equipment 14
Table 4.10 Library equipment level 14
Table 4.11. Financial evaluation - revenues 15
Table 4.12 Financial evaluation - expenditure 16
TABLES RELATED TO ASSESSMENT AREA 5 – SCIENTIFIC/ARTISTIC AND PROFESSIONAL ACTIVITY 17
Table 5.1. Bibliography (in the past five years) 17
Table 5.2. Artist's bibliography (in the past five years) 18
Table 5.3. a Research projects in the past five years 18
Table 5.3.b Professional/commercial projects in the past five years * 19
Table 5.4 Work in the conference organization committees in the last five years 19
Table 5.5. Editorship in journals in the past five years 19

Figure 2. Analytic supplement from MOZVAG (version from 2019.)

In order to optimize the evaluation process and facilitate the delivery of HEIs' data, the existing ASHE information system MOZVAG was upgraded in 2017. The main improvements relate to optimization of data collection, where the required data can be retrieved from the existing information systems that HEIs use to store data and keep records of their business activities (e.g. Information System of Higher Education Institutions - ISVU[5], the Croatian Scientific Bibliography database – CROSBI[6] and Database of Project Activities in Science and Higher Education in Croatia[7]).

[5] *https://www.isvu.hr/javno/hr/index.shtml*
[6] *https://www.bib.irb.hr/*
[7] *https://pdb.irb.hr/*

The data in the mentioned systems can provide a good basis for the development of institutions' Business Intelligence System.

III. BUSINESS INTELLIGENCE SYSTEMS AND QUALITY ASSURANCE

Some studies have shown that the use of information technology has a positive impact on the quality assurance procedures of HEIs, both for HEIs themselves and for agencies conducting external evaluations [4].

Using different software and applications and/or developing information systems for collecting, processing and analysing data on core activities of HEIs is nowadays more rule than exception. Therefore, it is not surprising that assessing the quality of these tools is also a part of overall quality assurance system, whether internal or external.

Investing in development of a system that will support everyday processes and decision-making at HEIs is not only driven by desire for a better analytical basis for decision-making, but is often, especially for public HEIs, the result of existing obligations arising from different legal regulations and requirements of different external evaluations procedures. Private HEIs are forced to invest more in the quality of decision-making systems to ensure sustainability, since their funding is not secured by public money. Nevertheless, requirements concerning quality assurance of Croatian HEIs do not differ between private and public HEIs.

Within re-accreditation of HEIs conducted in the first cycle (2010-2016) it has been noted that HEIs do not adequately use their existing data and information for improving their business processes. For example, most of HEIs keep updated information on student grades and progression, but only a small number of them use this data to identify students who are at risk of dropping out.

A. Data requirements in national regulations (legal and procedural requirements)

The Act on Scientific Activity and Higher Education [5], stipulates in Article 90. that "Higher education institutions electronically keep records and data collections, process the collected personal and other data for the purpose of performing tasks within their jurisdiction."

The mandatory data that HEIs should collect and process include:

- register on students' personal data, their exams and other study requirements,
- register of issued certificates of completion of studies and acquired academic and professional titles and academic degrees;
- records on the personal data of employees, teaching staff and teaching load;

- other records prescribed by law, implementing acts and general acts of the higher education institution;

In addition, the same Article also stipulates that the process of keeping and maintaining mandatory records shall be an integral part of the internal quality assurance and improvement system, and shall be subject to the external quality assurance and quality improvement procedures.

Ordinance regulating the requirements for performing higher education activity and re-accreditation of higher education institutions [6] in Articles 17 and 18 stipulates that HEIs have to establish information system for collecting, managing and processing data on the organisation of higher education institution and the organisation and delivery of study programmes, as well as information on quality assurance. Moreover, within re-accreditation of higher education institution, the required information system has to be evaluated.

The audit, as an external evaluation procedure, assesses the efficiency and degree of development of the internal quality assurance system of a HEI in accordance with national [7], European (ESG, 2015) and international standards, as well as its continuous improvement of quality culture. One of the audit criteria dealing with implementation and monitoring of the QA processes is assessing in what way HEI collects, manages and uses information in a decision-making process.

B. Data requirements in line with the ESG

The ESG are the basis for quality assurance in the European Higher Education Area (EHEA). They are used by HEIs and QA agencies as a reference document for building and developing internal and external quality assurance. It is important to emphasize that the responsibility for quality assurance, according to ESG, lies primarily with the higher education institution. That is the reason why internal quality assurance system of HEI is an integral part of quality of HEI's business activities in general.

Some of the standards for assessment of the internal quality assurance system directly cite the need for an information system as a basis for data collection. The analysis of this data should lead to better process management and quality assurance in core activities. Those standards are:

- 1.4 Student admission, progression, recognition and certification

…Institutions need to put in place both processes and tools to collect, monitor and act on information on student progression.

- 1.7 Information management

Institutions should ensure that they collect, analyse and use relevant information for the effective management of their programmes and other activities.

Reliable data is crucial for informed decision-making and for knowing what is working well and what needs attention. Effective processes to collect and analyse information about study programmes and other activities feed data into the internal quality assurance system. The following data is of interest:

-- Key performance indicators;
-- Profile of the student population;
-- Student progression, success and drop-out rates;
-- Students' satisfaction with their programmes;
-- Learning resources and student support available;
-- Career paths of graduates.

- 1.8 Public information

Institutions provide information about their activities, including the programmes they offer and the selection criteria for them, the intended learning outcomes of these programmes, the qualifications they award, the teaching, learning and assessment procedures used, the pass rates and the learning opportunities available to their students as well as graduate employment information.

- 1.9 On-going monitoring and periodic review of programmes

Institutions should monitor and periodically review their programmes to ensure that they achieve the objectives set for them and respond to the needs of students and society

C. Existing IT solutions in higher education and science with a potential for Business intelligence system

As mentioned, Business Intelligence system can use different information sources that combined provide quality data and analyses for decision-making.

In Croatia there are several possible information systems and platforms that provide possibilities to maintain data on the part of activity of HEIs, and, more important, they are free of charge for all HEIs. Those solutions are ISVU, CROSBI and Database of project activities, which combined provide most of the data necessary for quality data analysis of core activities of HEIs.

Information System of Higher Education Institutions (ISVU) is a project funded by the Croatian Ministry of Science, Education and Sports started in the 2001. ISVU is comprehensive information system that provides support for education related processes taking place within a higher education institution. As a part of the project, a data warehouse was developed to provide reporting and analytical feature. Unfortunately, the Ministry failed to ensure that all HEIs submit data to ISVU, regardless of whether they use ISVU or other software to maintain their data and legally prescribed datasets. Today, around 80% of HEIs use ISVU as their main information system for data management on students and teaching process.

Data on HEIs bibliography, including all publications of their teachers, can be stored in CROSBI, and data on project activities of a HEI can be entered in

recently upgraded Database of Project Activities in Science and Higher Education. Both databases are maintained by the Centre for Scientific Information of the Ruđer Bošković Institute, and offer connectivity to other systems, including MOZVAG.

In addition to these existing information systems, Information system on Croatian Scientific Activity (CroRIS), which is currently being developed within EU funded project "Scientific and Technological Prediction", anticipates to store and analyse all relevant data on scientific activity in Croatia. Since it is still under construction, its contribution to development of Business Intelligence in the higher education system remains to be seen.

IV. Conclusion

In today's digital world, the use of information systems and IT tools to store and analyse data is implied. Higher education institutions in Croatia strive to live up to these expectations, but due to the lack of a unique solution for everyone to apply, the quality of data collection and processing varies among the institutions. In quality assurance procedures, the ASHE evaluates, among other things, the quality of data collection at HEIs, as well as the use of data and their analysis in decision-making and achievement of strategic goals.

Although HEIs use several different information systems to record information about their activities, they do not recognise the opportunities offered by linking all data and creating information system that would be used for improvement of their organisational and decision-making processes. The external evaluations conducted by ASHE so far indicate the need to improve the use of existing information sources, primarily through linking data, especially in terms of creating a business intelligence system.

References

[1] P. Rausch, A. F. Sheta, and A. Ayesh, Eds., *Business Intelligence and Performance Management: Theory, Systems and Industrial Applications*. London: Springer London, 2013.

[2] T. C. Kahveci, Ö. Uygun, U. Yurtsever, and S. İlyas, 'Quality Assurance in Higher Education Institutions Using Strategic Information Systems', *Procedia - Social and Behavioral Sciences*, vol. 55, pp. 161–167, Oct. 2012, doi: 10.1016/j.sbspro.2012.09.490.

[3] ENQA, EUA, ESU, and EURASHE, *Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG)*. European Students' Union. Bruxelles, Belgium, 2005.

[4] 'Utilization of ICTs in quality assurance and accreditation of higher education: Systematic literature review - PDF Free Download', *zapdf.com*. https://zapdf.com/utilization-of-icts-in-quality-assurance-and-accreditation-o.html (accessed Aug. 03, 2018).

[5] 'The Act on Scientific Activity and Higher Education (Official Gazette 123/03, 198/03, 105/04, 174/04, 02/07, 46/07, 45/09, 63/11, 94/13, 139/13, 101/14, 60/15)'. .

[6] 'Ordinance on the content of licence and conditions for issuing licence for performing higher education activity, carrying out a study programme and re-accreditation of higher education institutions (OG 24/2010)'. .

[7] ASHE, 'Ordinance on audit of quality assurance systems at higher education institutions in the Republic of Croatia with Criteria for Assessment of Development and Efficiency of Internal Quality'. 2017.

LIST OF FREQUENTLY USED ACRONYMS

ASHE	**A**gency for **S**cience and **H**igher **E**ducation
CROSBI	**Cro**atian **S**cientific **Bi**bliography
ENQA	**E**uropean **N**etwork for **Q**uality **A**ssurance (since 2004: European Association for Quality Assurance in Higher Education)
ESG	**E**uropean **S**tandard and **G**uidelines for Quality Assurance in Higher Education (full name: Standards and Guidelines for Quality Assurance in the European Higher Education Area, 2015)
HEI	**H**igher **E**ducation **I**nstitution
ISVU	Information System of Higher Education Institutions (in Croatian: **I**nformacijski **S**ustav **V**isokih **U**čilišta)
MOZVAG	ASHE information system for data collection in external evaluations
QA	**Q**uality **A**ssurance

Study of customer behavior in online B2B shopping

E. Exenberger* and J. Bucko**
*Technical University of Košice/Department of Applied Mathematics and Business Informatics, Košice, Slovak Republic, emil.exenberger@tuke.sk
**Technical University of Košice/Department of Applied Mathematics and Business Informatics, Košice, Slovak Republic, jozef.bucko@tuke.sk

***Abstract* – The main goal of our study is to analyze behavior of the online shopping customers in a particular company, which is distributor of electronic components for industrial production. To reach the main goal we perform a data analysis of the data obtained from the company´s online sales process by using data mining technique – mining association rules. The purpose of the analysis is to identify and describe the significant association rules in order to provide information to support management decision making.**

***Keywords* – classification, association rules, data analysis, customer behavior, online shopping**

I. Introducion

Neoclassical economics assumes that each customer's goal is to maximize value with every purchase. However, customers are often influenced by subjective assumptions or momentary emotions. These aspects are not taken into account by neoclassical economics and considered absent and inadmissible in the decision-making process. Over time, more and more attention has been paid to these factors. Methods have been discovered that have been able to identify and analyze these factors, which prompted a new approach - behavioral economics.

Behavioral economics as a scientific area is focused on the analysis of such behavior in the economic environment, which contradicts the basic principles of neoclassical economics.

Companies often carry out various analyzes to improve their market position. These analyzes also include consumer behavior analysis, which can create a picture of what the customer thinks and what purchasing strategies he uses consciously or unconsciously in shopping process [1]. They are based on data they have collected over a period of time. Today, business processes use modern technology that enables real-time consumer data collection. Real-time collection ensures that data is up to date and ready for immediate analysis. These benefits can also be used to explore online consumer behavior. Many authors realize the benefits of real-time customer analysis and use it in their work.

II. Literature review

Some authors use social networks to sentiment analysis of customers experiences based on big amount of their posts [2] [3] [4] [5] [6]. Real-time counting system is used to analyzing customer movement [7]. It is possible to detect and analyze emotions of customer by using depth sensor and emotion recognition application programming interface integrated in tool named Intel RealSense [8]. Many other methods allow you to monitor and analyze online consumer behavior. In our article, we focus on real-time analysis of data obtained from online sales.

When making purchases, whether online or in stores, there are goods that are closely related. If there are sets of goods, the occurrence of which is related to the existence of another set of goods, these relationships can be expressed by the **Association Rules (AR)**. This concept was introduced by R. Agrawal, T. Imielinski and A. Swami in 1993 [9]. The relationship between the examined goods can be found by observing and analyzing the shopping carts of individual customers.

Association rules mining (ARM) is often used at present in various business sectors like medicine [10] [11] [12], language studies [13], E-transaction security [14], but also customer analysis [15] [16]. For example Jooa et al. using ARM to recommend places to people that they might want to visit based on their movement history and current location [17]. M. Y. Avcilar and E. Yakut mine AR in an environment of clothing and accessories. They compare parameters between generated ARs. The result is 25 470 ARs generated for the selected company and a proposal for the use of sales results [18].

The most used method for mining AR is the method named **Apriori** designed by R. Agrawal and R. Srikant in 1994 [19]. Since then, several authors have modified it to improve algorithm. H. Wu et al. propose a modified Apriori method called "an improved algorithm", which reduces the number of operations performed in association rules mining [20]. Modified Apriori method by X. Yuan use improves the operating efficiency of this method by e. g. decreasing number of needed scans of database [21].

Modifications to the Apriori method have been also made to adapt it to the analysis of data in specific areas. Y. Guo, M. Wang and X. Li suggest a modified Apriori method and recommend it for mining AR to increase the efficiency of data acquisition in order to provide relevant recommendations to online customers in real time [15]. M. Kaur and S. Kang introduce a new method for analyzing consumer behavior to support decision-making and increase sales of companies based on the modified Apriori method, so that it can analyze not only static data, but to enable AR analysis and generation in real time [16].

III. Research question and research method

Based on the literature review, we know that ARM has been tested and continues to be used in various fields of science and practice. On the other hand, while studying literature, we did not meet with ARM testing in the area of B2B electronic components distribution what is aim of our research. Our research question is to find out if there are association rules $A => B$ in B2B distribution of electronic components, which are valid with 100% probability. The association rule $A => B$ with validity 100% means that in every purchase where product A is located, product B is in this purchase too. Based on the research question, we have set the following hypothesis.

H1: There are no association rules in the B2B online distribution environment of electronic components, which are valid with 100% probability.

In our research, we observed how often a customer who bought the electronic component of interest also bought another in the same purchase. When selecting a research site, it was a precondition that the company had the possibility of selling online, with the emphasis being on data being as robust as possible.

The dataset represents data from one year of online purchases of the company that distributes electronic components named SOS Electronic. After communicating with the webteam leader and the project manager, we found that the company does not use any BI methods to analyze consumer behavior in more detail in online purchases. Based on additional information, we conclude that the implementation of unused BI methods should be focused on analyzing consumer behavior in real-time online purchases and the results of this analysis should support business process decision making such as consumer behavior, customer structure and analysis often purchased product groups, which is convenient for using the proposed methodology.

Each record in the dataset contains the following attributes:

- customer id;
- state of customer;
- order date;
- item id;
- brand of good;
- kind of good;
- order priority;
- item category id;
- quantity ordered;
- price per unit;
- payment method;
- order form;
- sales officer;
- sales representative;
- customer type.

We used the attributes *customer id*, *order date* and *item id* for our analysis. To keep data anonymous, customer and item names have been replaced by identification numbers. We carried out the whole process of work in the R software environment for its free availability and a number of such packages that allow to perform all the steps of our analysis. Table I shows an exemplary portion of the data that represented input to the analysis performed. Source data contains *customer id*, *order date* and *item id* in each line.

TABLE I
DATA SAMPLE

customer id	date	item id
309	01/31/2018	94 721
309	01/31/2018	482 229
309	02/02/2018	15 327
309	02/02/2018	26 409
311	01/04/2018	134 909
311	01/04/2018	61 639
311	01/04/2018	61 733

Source: prepared by authors

To corroborate the hypothesis, we mined association rules through the Apriori method [19], which is the most commonly used method in ARM. We used the $arulez$ and $arulezviz$ packages to mine ARs using the Apriori method and command `apriori()`. To program Apriori function into the R environment, M. Hahsler and B. Gruen [22] used original [9] and also more modern designs of this method [23] [24] [25].

First, list of transactions (shopping carts) will be created and within them will be recorded a set of goods that custom customer has purchased in one transaction. The purchase of one customer will be considered as one transaction in one day. To find association rules we use `apriori()` command, that uses metrics as $itemset$, $frequency$, $support$, $confidence$ and $Lift$.

$Itemset$ represents a set of all items from which individual purchases are made, so one purchase constitutes a subset of itemset. $Frequency(A)$ is the absolute number of occurrences of product A in the observed quantity of purchases N. $Support(A)$ is a relative representation of the number of items A in total number of purchases N (see equation no. 1).

$$Support(A) = \frac{frequency(A)}{N} \tag{1}$$

$Confidence(A \implies B)$ determines the likelihood of existing product B in purchases containing product A (see equation no. 2).

$$Confidence(A => B) = \frac{frequency(A,B)}{frequency(A)} \tag{2}$$

$Lift(A => B)$ represents the correlation between product A and product B - as in a particular purchase, the occurrence of product A affects the occurrence of product B in the same purchase (see equation no.3).

$$Lift(A => B) = \frac{Support(A,B)}{Support(A) * Support(B)} \tag{3}$$

Based on the calculated value, one of three situations occurs:

- $Lift(A => B) < 1$ - the presence of product A in a purchase has a negative impact on the presence likelihood of the product B. The lower the calculated value $Lift$ is, the less likely it is for product B to be purchased with product A. The lowest possible value of the $Lift$ value can be equals to 0;

TABLE II
SAMPLE OF TRANSACTIONS

#	Transactions
1	94721, 482229
2	15327, 26409
3	83897;83917;83915;83899;25769
4	200655
5	253415
6	134909;61639;61733;26485;61867;137561;100273;105799
7	79397;11843;59271;58811;144089
8	75865;75893;75837;75877;75889;75933;75857;73819;73641;75829;209165;162865
9	81399;38845;66041;120665;61427;61411;61733;61867;49973;35289;37665;39091;159407
10	80837;24683;81731;80665;31483;115141;109321;141767;31519;98299;116049

Source: prepared by authors

- $Lift(A => B) = 1$ - the presence of the product A in the purchase does not affect the likelihood of product B in the same purchase;
- $Lift(A => B) > 1$ - the presence of product A in the purchase has a positive effect on the likelihood of product B in the same purchase. The higher the calculated value is, the greater the likelihood of product B in the purchase with product A. Lift value can be maximal equal to the number of tracked purchases N.

When analyzing ARs, we look for those ARs that meet the minimum values of the $confidence$ and $support$ parameters that we reffered as $min_confidence$ and $min_support$. For each AR found must apply:

- $Confidence(X \implies Y) \geq min_confidence$;
- $Support(X \implies Y) \geq min_support$.

The $min_support$ parameter can also be expressed by the condition parameter $min_frequency$, wherein the relationship between them is represented by equation no. 4 [26] [27].

$$min_support * N = min_frequency \tag{4}$$

The result of ARM will be a table that expresses the association rules and the associated $support$, $confidence$, $Lift$ and $count$ values. The established hypothesis will be considered true if mined ARs do not meet the condition $confidence = 1$. During testing, the $min_support$ value will be set to 0.01 by default and gradually decreased until Apriori results have at least one association rule meeting the condition $min_confidence = 1$. If hypothesis will be refused, additional ARs will be mined to provide them business together with suggestions of their use, which will be described in the discussion.

IV. RESULTS

First, we generated transactions by using command using $arulez$ package [22]. Table II shows sample of 10 transactions generated from dataset (see Table I). One transaction represents set of items purchased by custom customer in one day. The total number of transactions generated from 185 706 rows in source data was 48 998. The maximum number of items purchased per purchase was 183.

After that we started searching ARs using Apriori method [9], where parameters were determined at $min_confidence = 1$ and $min_support = 0,01$. At $min_support = 0.000734709$, the program found the first AR that met the specified conditions. On the basis of this result, we reject the established hypothesis. The program also generated additional ARs.

Table III shows absolute abundance of ARs found for individual frequency values. The value Frequency represents how many times AR was actually applied in the analyzed transactions and row Abundance represents how many ARs were found for each Frequency value.

TABLE III
ABSOLUTE ABUNDANCE OF ARS FOR INDIVIDUAL FREQUENCY VALUES

Frequency	15	16	17	18	19	20	21	22	25	26	29	36
Abundance	104	32	12	18	6	7	5	3	1	2	1	1

Source: prepared by authors

Table IV shows the first 20 ARs with the highest support value. Column lhs (left horizontal side) represents products that are input of implication of association rule and column rhs (right horizontal side) represents product that is output of implication of association rule. We terminated the ARM for computational complexity at $min_support = 0.0003061349$, which means condition $min_frequency = 15$. Sum of ARs generated under these conditions was 192, which we consider sufficient for the analysis suggested in the discussion of this article.

Interestingly, ARs with a higher $support$ value sometimes have a lower lift value than ARs with a lower $support$ value. We will explain this phenomenon using the example AR $\{A \implies B\}$. If product B were in 100% purchases and A in 1% of purchases, $confidence$ would be equal to 1, but lift would be 1, indicating that product A in the purchase does not affect the occurrence of B at all. On the other hand, if product B were in only one purchase and product A also

TABLE IV
FIRST 20 ARS WITH THE HIGHEST SUPPORT VALUE

#	lhs		rhs	support	confidence	lift	count
1	{83897,83917}	=>	{83915}	7,35E-04	1	690,13	36
2	{103727,130543}	=>	{103725}	5,92E-04	1	352,51	29
3	{73135}	=>	{73133}	5,31E-04	1	1 689,62	26
4	{101317,102971}	=>	{101315}	5,31E-04	1	263,44	26
5	{101317,102971,102973}	=>	{101315}	5,10E-04	1	263,44	25
6	{83947,83949}	=>	{83945}	4,49E-04	1	544,43	22
7	{170669,341857}	=>	{170667}	4,49E-04	1	426,08	22
8	{101797,103727,130543}	=>	{103725}	4,49E-04	1	352,51	22
9	{525529}	=>	{472251}	4,29E-04	1	2 227,23	21
10	{102851,103055}	=>	{103051}	4,29E-04	1	777,76	21
11	{58949,61639,61791}	=>	{171731}	4,29E-04	1	157,55	21
12	{171731,58949,61791}	=>	{61639}	4,29E-04	1	230,04	21
13	{171731,58949,61639}	=>	{61791}	4,29E-04	1	418,79	21
14	{103053,149793}	=>	{103055}	4,08E-04	1	844,81	20
15	{84025,84031,84033}	=>	{84027}	4,08E-04	1	1 224,98	20
16	{171731,58949,61867}	=>	{61791}	4,08E-04	1	418,79	20
17	{171731,58949,61867}	=>	{61639}	4,08E-04	1	230,04	20
18	{58949,61639,61791,61867}	=>	{171731}	4,08E-04	1	157,55	20
19	{171731,58949,61791,61867}	=>	{61639}	4,08E-04	1	230,04	20
20	{171731,58949,61639,61867}	=>	{61791}	4,08E-04	1	418,79	20

Source: prepared by authors

in only one purchase with product B, $Lift$ value would be at maximal value.

V. DISCUSSION

The results show that there are ARs in B2B online distribution of electronic components that are 100% probable. Due to results set hypothesis is false. This can be used as an incentive for further scientific research to analyze consumer behavior in more depth. Rejection of the hypothesis as one of the results of this research could thus serve as a premise for further analyzes. In practice, mined ARs can be used to analyze consumer baskets or predict consumer behavior.

In the **analysis of consumer baskets**, association rules will not be followed in the form of implications ($X_1 = \{A \implies B\}$) but as sets of goods in the form of consumer baskets ($X_1 = \{A, B\}$). These consumer baskets will be frequently-frequented sets with sufficient credibility as they have been created from association rules that meet the conditions of $min_support$ and $min_confidence$. Tracking these baskets allows business management to identify the links between products within the baskets. If identified correctly, baskets can be used as an incentive to change and manage business processes that could lead to increased efficiency.

Another benefit of the analysis may be the **prediction of consumer behavior** directly in the customer's online purchasing process through the use of established association rules. This should lead to reaching a customer who is likely to be interested in buying the goods or services that the seller is offering. The analysis should therefore help the seller answer the questions: What is the assumption that the customer will be interested in buying the goods or services we offer? What goods will they most likely want to buy and when will they most likely want to buy them? The answers to these questions will enable the executives to reach the customer through different marketing channels effectively.

When selecting ARs, an enterprise should focus on $support$ and $Lift$ values that speak of the presence and relationship of products in the same purchase. When analyzing consumer baskets, management should focus on the ARs with the highest $Lift$ value. Indeed, the $Lift$ value speaks of how much is the product from rhs connected with the product(s) in lhs. In the case of predicting consumer behavior, the $Lift$ value is not so significant because at that time we do not observe the exclusivity of the link between the two products. What is important in this case is the value of $confidence$, which tells about the current validity of the ARs.

The results of the analysis performed in the article can also be used as a stimulus for further scientific analysis of consumer behavior in the B2B distribution of electronic components. In our next research, we plan to focus on other business sectors, comparing research results.

VI. CONCLUSIONS

Behavioral economics makes it possible to examine those aspects and factors of customer behavior in the purchasing process that cannot be explained through knowledge of neoclassical economics. In the introduction, we pointed to the phenomenon of real-time data analysis. In the literature review,

we summarized the current state of knowledge and the use of real-time consumer behavior analysis. We focused on the use of association rules mining in current scientific studies. Among the many areas where the ARM method is used, we have mentioned recent studies that use the method in the online sales environment.

After reviewing the literature, we established a hypothesis based on the research question. The task of the research was to verify the truth of the hypothesis - to see if there are association rules in the area of online distribution of electronic components with 100% probability. We defined the terms used in mining association rules and set the conditions that conditioned the analysis of the established research hypothesis. Subsequently, we described the research site and character of the data that formed the input to the analytical part.

This was followed by the analysis and mining of association rules. After finding the first association rules with $confidence = 1$, we rejected the hypothesis. We continued mining to $min_frequency = 15$, with a total of 192 association rules.

In the discussion we outlined the use of results in the form of analysis of consumer baskets and prediction of consumer behavior. The paper can serve as a process model for mining association rules in data or as a stimulus for further analysis in the field of Business Intelligence.

ACKNOWLEDGEMENT

The research was realized within the national project "Decision Support Systems and Business Intelligence within Network Economy" (Contract No. 1/0201/19) funded by Grant Agency for Science; Ministry of Education, Science, Research and Sport of the Slovak Republic.

REFERENCES

[1] C. Giri, S. Thomassey, and X. Zeng, "Customer analytics in fashion retail industry," in *Functional Textiles and Clothing*. Springer, 2019, pp. 349–361.

[2] S. Ranjan, S. Sood, and V. Verma, "Twitter sentiment analysis of real-time customer experience feedback for predicting growth of Indian telecom companies," in *2018 4th International Conference on Computing Sciences (ICCS)*. IEEE, 2018, pp. 166–174.

[3] M. Geetha, P. Singha, and S. Sinha, "Relationship between customer sentiment and online customer ratings for hotels-An empirical analysis," *Tourism Management*, vol. 61, pp. 43–54, 2017.

[4] G. Vinodhini and R. M. Chandrasekaran, "Sentiment analysis and opinion mining: a survey," *International Journal*, vol. 2, no. 6, pp. 282–292, 2012.

[5] W. Duan, Q. Cao, Y. Yu, and S. Levy, "Mining online user-generated content: using sentiment analysis technique to study hotel service quality," in *2013 46th Hawaii International Conference on System Sciences*. IEEE, 2013, pp. 3119–3128.

[6] S. Khatoon, "Real-time twitter data analysis of Saudi telecom companies for enhanced customer relationship management," *International Journal of Computer Science and Network Security*, vol. 17, no. 2, pp. 141–147, 2017.

[7] S. I. Cho and S.-J. Kang, "Real-time people counting system for customer movement analysis," *IEEE Access*, vol. 6, pp. 55 264–55 272, 2018.

[8] J. Purnama, J. Yapri, T. Winarta, S. Oliver, and M. Galinium, "The Intel Realsense Depth-Camera Performance for Real-Time Customer Satisfaction Analysis using Facial Expression Detection," in *Journal of Physics: Conference Series*, vol. 1175, no. 1. IOP Publishing, 2019, p. 12076.

[9] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Acm sigmod record*, vol. 22, no. 2. ACM, 1993, pp. 207–216.

[10] D. Luo, C. Xiao, G. Zheng, S. Sun, M. Wang, X. He, and A. Lu, "Searching association rules of traditional Chinese medicine on Ligusticum wallichii by text mining," in *2013 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, 2013, pp. 162–167.

[11] A. L. Buczak, B. Baugher, E. Guven, L. C. Ramac-Thomas, Y. Elbert, S. M. Babin, and S. H. Lewis, "Fuzzy association rule mining and classification for the prediction of malaria in South Korea," *BMC medical informatics and decision making*, vol. 15, no. 1, p. 47, 2015.

[12] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.

[13] A. Z. Adamov, "Mining Term Association Rules from Unstructured Text in Azerbaijani Language," in *2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*. IEEE, 2018, pp. 1–4.

[14] S. Askari, S. Md, and A. Hussain, "E-Transactional Fraud Detection Using Fuzzy Association Rule Mining," *E-Transactional Fraud Detection Using Fuzzy Association Rule Mining (January 1, 2020)*, 2020.

[15] Y. Guo, M. Wang, and X. Li, "Application of an improved Apriori algorithm in a mobile e-commerce recommendation system," *Industrial Management & Data Systems*, vol. 117, no. 2, pp. 287–303, 2017.

[16] M. Kaur and S. Kang, "Market Basket Analysis: Identify the changing trends of market data using association rule mining," *Procedia computer science*, vol. 85, pp. 78–85, 2016.

[17] J. Jooa, S. Bangb, and G. Parka, "Implementation of a recommendation system using association rules and collaborative filtering," *Procedia Computer Science*, vol. 91, pp. 944–952, 2016.

[18] M. Y. Avcilar and E. Yakut, "Association Rules in Data Mining: An Application on a Clothing and Accessory Specialty Store," *Canadian Social Science*, vol. 10, no. 3, pp. 75–83, 2014.

[19] R. Agrawal, R. Srikant, and Others, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, 1994, pp. 487–499.

[20] H. Wu, Z. Lu, L. Pan, R. Xu, and W. Jiang, "An improved apriori-based algorithm for association rules mining," in *2009 sixth international conference on fuzzy systems and knowledge discovery*, vol. 2. IEEE, 2009, pp. 51–55.

[21] X. Yuan, "An improved Apriori algorithm for mining association rules," in *AIP conference proceedings*, vol. 1820, no. 1. AIP Publishing LLC, 2017, p. 80005.

[22] Datacamp.com, "apriori function — R Documentation." [Online]. Available: https://www.rdocumentation.org/packages/arules/versions/1.6-4/topics/apriori

[23] J. Lepping, "Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery," 2018.

[24] C. Borgelt and R. Kruse, "Induction of association rules: Apriori implementation," in *Compstat*. Springer, 2002, pp. 395–400.

[25] C. Borgelt, "Efficient implementations of apriori and eclat," in *FIMI'03: Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations*, 2003.

[26] H. Jabeen, "Market Basket Analysis using R." [Online]. Available: https://www.datacamp.com/community/tutorials/market-basket-analysis-r

[27] J. Paralič, *Objavovanie znalostí v databázach*. Elfa, 2003.

Fruit firmness prediction using multiple linear regression

1st Tomislav Ivanovski
Department of Informatics
University of Rijeka
Rijeka, Croatia
tomislav.ivanovski@student.uniri.hr

2nd Guoxiang Zhang
China Agricultural University
Beijing, China
zhangguoxiang@cau.edu.cn

3nd Tomislav Jemrić
Department of Pomology
University of Zagreb
Faculty of Agriculture
Unit of Horticulture and Landscape Architecture
Zagreb, Croatia
tjemric@agr.hr

4nd Marko Gulić
Faculty of Maritime Studies
University of Rijeka
Rijeka, Croatia
marko.gulic@pfri.hr

5nd Maja Matetić
Department of Informatics
University of Rijeka
Rijeka, Croatia
majam@uniri.hr

***Abstract*—Smart agriculture is a term used to describe the utilization of digital technologies used in optimizing agricultural food production systems. In order to increase the efficiency of manufacturing process, modern tools for collecting, storing and analyzing electronic data are used.**

The focus of this paper is creation and comparison of peach firmness prediction models using various machine learning algorithms. The size of the data set, which is used to construct machine learning models described in this paper, is small. Because size of the data set has a large impact on the performance of the machine learning algorithm, models of different complexities were developed to tackle this problem.

Simple linear regression is used as fundamental techniques for predicting numerical outcome variable, the peach firmness using a single predictor variable. By extending simple linear regression model so that is can accommodate multiple predictors, multiple linear regression model is obtained, which is the top performing model when applied to the dataset described in this paper. The backpropagation neural network model is developed and its performance is compared to the performance of regression models.

***Index Terms*—smart agriculture, BP neural networks, machine learning, linear regression, prediction models, fruit firmness**

I. Introduction

With the arrival of the digital era, the transmission of large amounts of data at high speeds became possible. Increase in computer memory and processing power yield the possibility of applying various machine learning methods for data analysis and visualization. Using these methods useful information can be obtained based on which predictive models can be created for future predictions. With increased availability of quality sensor data used in agriculture and increasing computation power for analyzing it, new innovations in this field arise [13]. Subjective methods for determining proper harvest time, such as classification based on fruit color, are commonly used, but they do not yield precise results since are subjected to personal interpretation.

Modern measurement techniques such as colorimetry, visible imaging, spectroscopy, and spectroscopy imaging [2,7] produce high-quality data using precise sensor measurements. Similarly, electrical impedance spectroscopy is the impedance measurement method. Due to the high correlation between impedance and peach firmness, the data of measured impedance can be used to fit prediction models. Various machine learning algorithms can be used to analyze collected data and to create predictive models. Simple linear regression is the most simple example of a supervised learning algorithm [17] and, as such, is used as a reference for performance comparison with other machine learning algorithms in this paper. Using a simple linear regression model, the outcome variable can be predicted based on the value of one predictor variable, assuming the linear relationship between the two. Simple linear regression is used for firmness prediction based on measured impedance.

Multiple linear regression algorithm also assumes a linear relationship between the response variable and multiple predictor variables. Regression modeling means not only estimating the coefficients but also choosing which predictors to include and in what form [17]. For firmness prediction, which is described in more detail in section IV, three predictor variables are selected. Finally, backpropagation neural network model is fitted. Its structure supports capturing very complex relationships between predictors and outcome variable [17], and its performance is compared to other models.

This paper compares the prediction performance of different

machine learning models and gives closer insight into the correlation between all the measured variables.

II. Related works

Different machine learning algorithms are used for prediction purpose in a vast amount of applications. Nanyang Zhu et al. give summary of major deep learning algorithms, including concepts, limitations, implementation, and training processes used in smart agriculture in [5]. Similar work is done in [14,15], where Wang Xiang et al. present wireless multi-gas-sensor system for real-time cold chain monitoring. Shelf life prediction model based on the theory of zero or first-order reaction kinetics model and the Arrhenius equation is examined. Study in [16] presents data mining methods which analyze temperature, relative humidity, and CO_2 characteristics in the fresh peaches and nectarine cold chain logistics. Based on the knowledge extracted by data mining methods it is possible to identify changes which are used as an early warning to take effective measures in advance and resolve any problems that may cause unexpected product quality loss. Use of of backpropagation neural networks for estimating the temperature in areas where sensor placement is not feasible is described in [6]. Here R. Badia-Melis et al. proved that neural networks showed the most precise temperature approximations in comparison to other methods, given the non-linear relationship between temperature source and temperature target [8]. Research on peach firmness prediction, based on peach dielectric properties by using a multi layer neural network model, is presented in [4]. G. Zhang et al. successfully constructed highly accurate multi layer neural network model for firmness prediction. They implement k-means clustering and principal component analysis in order to achieve dimension reduction. Research on correlation between dielectric properties and maturity of the banana fruit is done in [21]. D. Jamaludin et al. showed that using impedance measurement it is possible to detect different stages of banana maturity. In addition, simple linear regression model has been constructed for predicting maturity based on measured impedance. Similar research is presented in [9] by U. Ahmad et al., where linear discriminant analysis machine learning method is implemented for melon maturity prediction based on image data. In [20] S. Sabzi at al. present an ANN model developed for the estimation of apple maturity using areal images.

III. Data set description and methods

A. Data set description

The data set described in this paper consists of measured variables of two hundred peaches. Measuring was done for two consecutive days. Measurements on one hundred peaches were done on the first day, while the measurements on the other one hundred peaches were obtained the day after. These two sub sets are referred to as the first and second sub set. Initial data consisted of twenty-one measured variables. Dimension of the data set had to be reduced using the expert's domain knowledge in order for machine learning algorithms to operate efficiently [17]. Impedance is measured using probes with alternate voltage of 500 mV and 10 kHz frequency that were inserted into a peach. Based on measurements for height, width and mass, value for density if obtained, while using penetrometer peach firmness is measured. Soluble solids content is obtained using a refractometer and juice extracted from a peach, while titratable acidity is measured by combining peach juice and alkaline solution.

B. Variable selection

Variables included in the final data set used throughout the experiment are firmness, impedance, density, and ratio of soluble solids content and titratable acidity. Dielectric properties give insight into structural characteristics of solid materials and biological tissues. They can be obtained using electrical impedance measurement, because of passive electrical properties, plant tissue impedance is related to cellular ionic content, membrane structures and viscosity [1]. Changes in membrane structure reflect changes in fruit ripening [10,11,12]. Because the firmness is a good maturity indicator [3], it is used as target variable in developed models. Correlation between all the measured variables is also examined in the experiment. In Fig. 1, the correlation matrix between variables is shown in which correlation coefficients define linear dependence between two variables. It is possible to see that the largest correlation is between impedance and firmness. Fig. 2 depicts the scatter

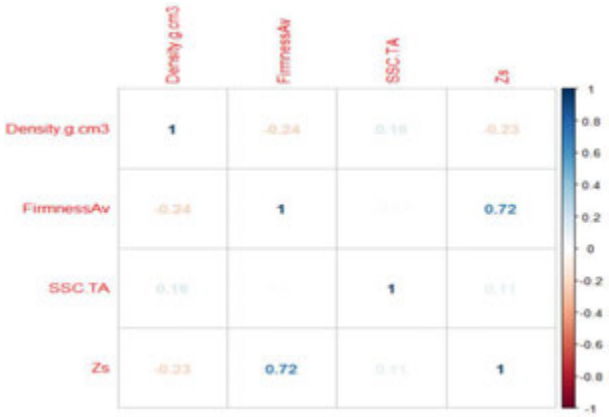


Fig. 1. Data set correlation matrix

plot matrix of the data set. High linear dependency between firmness and impedance variables is noticeable.

Next, the scatter plot of firmness versus impedance is shown in Fig. 3. Circles highlighted blue represent data points of the second sub set, while data points representing the first sub set are highlighted red. It is possible to conclude that the variance of the second sub set is greater than the variance of the first one. Based on expert's domain knowledge increased variance of the second sub set whose measurements were obtained one day after the first sub set is a result of high maturation dynamic. Fig. 4 depicts box plots of the same variable in two of the sub sets. Twenty-fifth and seventy-fifth percentiles are similar, but the rest of the data range is wider for the second sub set [17]. Hence, the variance of the second sub set is

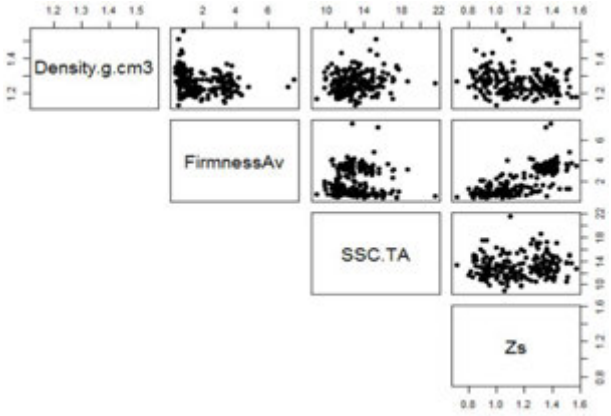


Fig. 2. Data set scatter plot

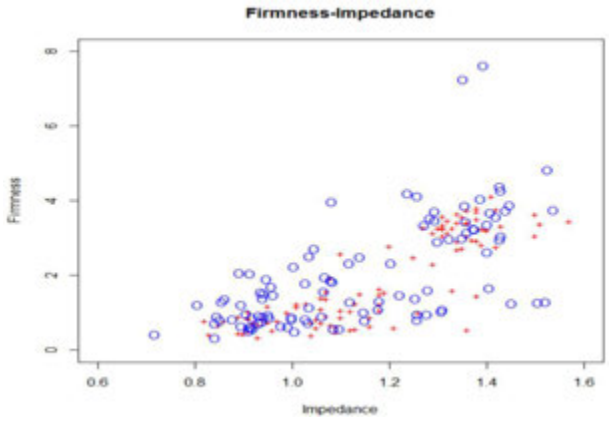


Fig. 3. Scatter plot of firmness and impedance

larger than the first one. Finally, min-max data normalization

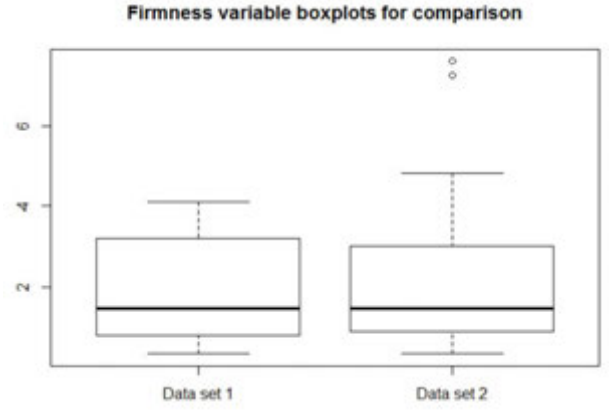


Fig. 4. Firmness box plots for two sub sets

is done to give equal importance for all measurements in terms of variability when creating machine learning models. Data set is divided in two categories; the training set which is eighty percent of the original data randomly sampled, and the test set which is twenty percent. Each of the following chapters introduces a machine learning model used in firmness prediction.

C. Simple linear regression

Simple linear regression is a popular approach for supervised learning. It is used to fit a relationship between a numerical outcome variable and a predictor variable [17,18]. Simple linear regression assumes linear relationship between predictor and response.

$$Y = \beta_0 + \beta_1 \cdot X + e \tag{1}$$

- β_0: Intercept coefficient
- β_1: Slope coefficient
- X: Predictor variable
- e: Mean-zero random error term

Least squares approach is used to fit the linear regression model. It adjusts the values of β_0 and β_1 coefficients, and minimizes the sum of squared deviations between the actual outcome and outcome predicted by the model.

D. Multiple linear regression

The only difference between simple and multiple linear regression is in the number of predictor variables. Multiple linear regression, just like simple linear regression uses least square approach for adjusting the coefficients. A generic expression for multiple linear regression model that approximates relationship between the predictors and outcome variable is an extension of the simple linear regression model and is the following [18]:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + ... + \beta_n \cdot X_n + e \tag{2}$$

- $\beta_{0..n}$: Multiple linear regression coefficients
- $X_1..n$: Variable values
- e: Mean-zero random error term

E. Artificial neural network

Artificial neural networks are models used for classification and prediction tasks in supervised learning. They are inspired by the model of the biological brain in which neurons are interconnected. The neural network mimics the way the human brain learns and is considered a ”black box” in terms of interpretability [17]. A weakness of the neural network is that it can easily overfit the data, causing significantly better performance on training comparing the performance on test data. This can be seen in section IV where prediction accuracy for trained models is compared. An artificial neural network consists of the input layer, output layer, and arbitrarily number of hidden layers, a shown on Fig. 5.

IV. Results

As an initial step, simple and multiple linear regression models are created and tested. The experiment is concluded in a way that ten simple linear regression models are created with impedance as a predictor variable and firmness as a response. Impedance is chosen as a predictor based on a large correlation value as mentioned in chapter III-B. Each model is created on randomly sampled eighty percent of data and tested on the

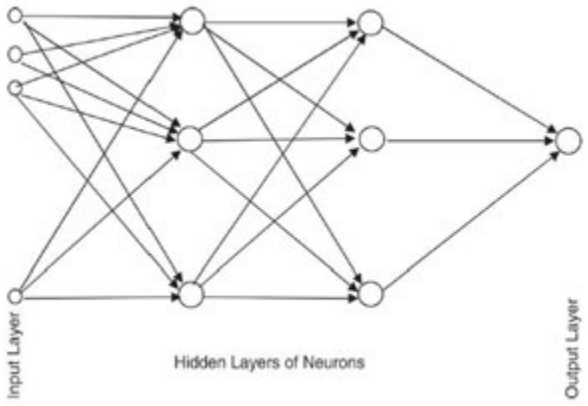


Fig. 5. Multi layer artificial neural network [17]

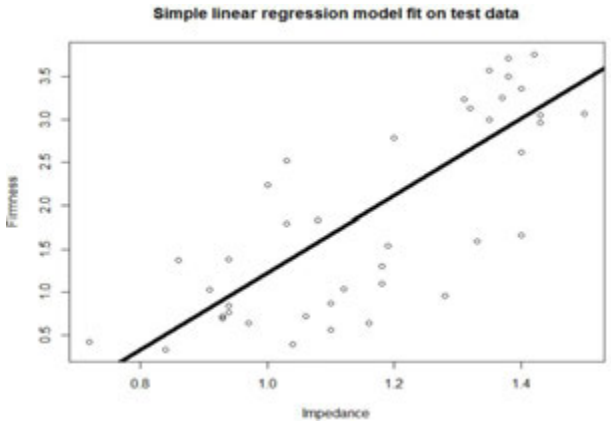


Fig. 6. Simple linear regression model fit on test data

remaining twenty percent. Residual standard error, multiple r-squared, and F-statistics statistical measurements give the insight to model fit. Large F-statistic indicates a dependency between predictor and response variable, r-squared represents a proportion of variance in response that is explained in the predictor variable and residual standard error gives insight to the overall model fit. A lower value of residual standard error indicated a better model fit [18]. Residual standard error expression used for model fit measurement follows:

$$RSE = \sqrt{\frac{RSS}{n-p-1}} \tag{3}$$

Where:

- RSS: Residual standard error
- n: Number of observations
- p: Number of predictor variables

It is important to differentiate the goodness-of-fit and predictive accuracy of the model. The large F-statistic value indicates good predictive performance if sampling has been done correctly. However this does not mean that model with the highest F-statistic will yield best performance on newly introduced data. For assessing model's prediction performance, the difference between actual outcome value and predicted outcome value has to be calculated. Mean squared error estimation is used for prediction performance evaluation.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i')^2 \tag{4}$$

Where:

- n: Number of observations
- Y_i: Actual outcome value
- Y_i': Predicted outcome value

Test MSE is calculated for ten created models. In Fig. 6 model with the lowest test mean square error value is shown since it has the best prediction accuracy of all created simple linear regression models. Test MSE is used as a measurement for model prediction accuracy. The best simple linear regression model has test MSE value of 0.6695544. This value will be used throughout the experiment for comparison of prediction accuracy with other models. In order to test the influence of the other two predictors on the response variable a multiple linear regression model is introduced. Again, ten models are fitted using randomly sampled train data sets. Predictor variables are impedance, SSC/TA, and density, while the response variable is firmness. When applied to the case of firmness prediction using three predictors previously mentioned, the expression takes the following form:

$$firmness = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot SSC.TA + \beta_3 \cdot impedance + e \tag{5}$$

For every trained model, test MSE is calculated based on which most accurate model is selected. The most accurate multiple linear regression model yields test MSE value of 0.6315375, which is 5.7 percent more accurate than the best simple linear regression model. Coefficient values for this model are the following:

- β_0 = -1.80234
- β_1 = -0.59464
- β_2 = -0.06343
- β_3 = 4.58600

Using multiple linear regression model with all of the predictors variables it is possible to acquire the most accurate model.

The most prominent characteristic of artificial neural network is its ability to capture highly complicated relationships between the predictors and outcome variable [17]. Because of its good predictive performance, an artificial neural network model is created. Besides examining if accurate firmness prediction model using impedance, density, and ratio of soluble solids content and titratable acidity measurements can be obtained, performance of artificial neural network on a small and highly variable dataset is examined and compared with other, simpler statistical learning methods.

Data set for training and testing are scaled using min-max normalization. Effect of this normalization gives variables equal importance in terms of variability [17] and as a result gradient descent algorithm that is responsible for artificial neural network weight tuning will converge more quickly. ANN models contain three neurons in the hidden layer, and the data set is divided in 80:20 ratio for training and test set respectively. Model predictor variables are impedance, SSC/TA, and density, while the response variable is firmness, just like in other models. ANN model is shown in Fig. 7. As in previous chapters, test MSE is calculated for every

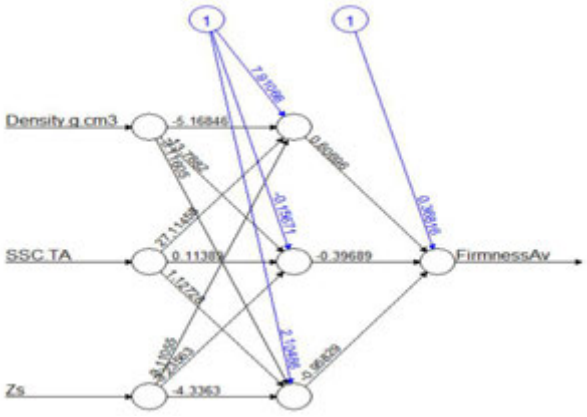


Fig. 7. Artificial neural network used for firmness prediction

trained ANN model, and the most accurate model is selected. Obtained test MSE has the value of 0.6543438, which is a better result then simple linear regression, but slightly worse than multiple linear regression model.

V. Conclusion

The result of the experiment shows the possibility of using machine learning models for firmness prediction. Based on large correlation between impedance and firmness, it is possible to conclude that impedance can be used as predictor variable for predicting firmness in machine learning models. The more mature the fruit is, the lower the impedance is, since the proportion of liquid and ions in fruit is increased. Based on the experiment's results, it is possible to conclude that multiple linear regression model gives most accurate predictions. This is expected since the dataset used is very small and the majority of variables in it have a large variance. In the case of small data sets like this one, simple machine learning models give more accurate predictions [19] since they apply global structure to the data. In contrast, artificial neural networks learn directly from the data based on each observation given in the data set. In the case where data is scarce, advanced machine learning models like ANN's tend to over fit to train data set and have degraded performance when applied to the test set. The next research step is to obtain a larger data set on which the same models will be trained. In addition, classification and regression trees will be used as dimension reduction techniques to identify key predictor variables. It is then to expect the best prediction performance from the artificial neural network.

Acknowledgment

This work has been fully supported by the University of Rijeka under the project number uniri-drustv-18-122.

References

[1] I. Jocsak, G. Vegvari, E. Vozary. (2019) Electrical impedance measurement on plants: a review with some insights to other fields. Theor. Exp. Plant Physiol. (2019) 31:359–375

[2] Bo Li, J. Lecourt, G. Bishop. (2018) Advances in Non-Destructive Early Assessment of Fruit Ripeness towards Defining Optimal Time of Harvest and Yield Prediction—A Review. Plants 2018, 7, 3

[3] F. Roger Harker, John H. Maindonald. (1994) Ripening of Nectarine Fruit Changes in the Cell Wall, Vacuole, and Membranes Detected Using Electrical Impedance Measurements. Plant Physiol. (1994) 106: 165-171

[4] Guoxiang Zhang, Qiqi Fu, Zetian Fu, Xinxing Li, Maja Matetić , Marija Brkić Bakaric, Tomislav Jemrić (2020). A Comprehensive Peach Fruit Quality Evaluation Method for Grading and Consumption. Appl. Sci. 2020, 10, 1348;

[5] Nanyang Zhu, Xu Liu, Ziqian Liu, Kai Hu, Yingkuan Wang, Jinglu Tan, Min Huang, Qibing Zhu, Xunsheng Ji, Yongnian Jiang, Ya Guo. (2018) Deep learning for smart agriculture: Concepts, tools, applications, and opportunities. Int J Agric & Biol Eng Vol. 11 No.4

[6] R. Badia-Melis, U. Mc Carthy, L. Ruiz-Garcia, J. Garcia-Hierro, J.I. Robla Villalba. (2018) New trends in cold chain monitoring applications - A review. Food Control 86 (2018) 170-182

[7] Satyam Srivastava, Shashikant Sadistap. (2017) Non-destructive sensing methods for quality assessment of on-tree fruits: a review. Springer Science+Business Media, LLC 2017

[8] Ganjar Alfian, Jongtae Rhee, Hyejung Ahn, Jaeho Lee, Umar Farooq, Muhammad Fazal Ijaz, M. Alex Syaekhoni. (2017) Integration of RFID, wireless sensor networks, and data mining in an e-pedigree food traceability system. Journal of Food Engineering

[9] Usman Ahmad, Dwi Pamungkas Bermani, Mardison. (2018) Color Distribution Analysis for Ripeness Prediction of Golden Apollo Melon. Telkomnika, Vol.16, No.4, August 2018, pp. 1659 1666

[10] Pietro Ibba, Aniello Falco, Biresaw Demelash Abera, Giuseppe Cantarella, Luisa Petti, Paolo Lugli. Bio-impedance and circuit parameters: An analysis for tracking fruit ripening. Postharvest Biology and Technology 159 (2020) 110978

[11] Marco Grossi, Bruno Riccò. (2017) Electrical impedance spectroscopy (EIS) for biological analysis and food characterization: a review. J. Sens. Sens. Syst., 6, 303–325, 2017

[12] G. M. Weaver, H.O. Jackson. (1966) Electric impedance, in objective index of maturity in peach. Canadian journal of plant science Vol. 46

[13] Sonka, Steve. (2016) Big data: fueling the next evolution of agricultural innovation. Journal of Innovation Management 4.1: 114.

[14] Xiao, Xinqing; Li, Zhigang; Matetić, Maja; Brkić Bakarić, Marija; Xiaoshuan Zhang, (2017) Energy-efficient sensing method for table grapes cold chain management // Journal of cleaner production, 152.

[15] Wang, Xiang; He, Qile; Matetić, Maja; Jemrić, Tomislav; Zhang, Xiaoshuan, (2017) Development and evaluation on a Wireless Multi-Gas-Sensors System for improving traceability and transparency of table grape cold chain // Computers and electronics in agriculture, 135

[16] Wang, X., Matetić, M., Zhou, H., Zhang, X., & Jemrić, T. (2017). Postharvest quality monitoring and variance analysis of peach and nectarine cold chain with multi-sensors technology. Applied Sciences, 7(2), 133.

[17] Galit Shmueli, Peter C. Bruce, Inbal Yahov, Nitin R. Patel, Kenneth C. Lichtendahl Jr. Data mining for business analytics. John Wiley & Sons, Inc. (2018)

[18] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer Science+Business Media New York 2013 (Corrected at th printing 2017)

[19] Andrew Ng. (2019) Machine Learning Yearning Technical Strategy for AI Engineers, In the Era of Deep Learning. 2018 Andrew Ng. All Rights Reserved.

[20] Sajad Sabzi, Yousef Abbaspour-Gilandeh , Ginés García-Mateos, Antonio Ruiz-Canales, José Miguel Molina-Martínez, Juan Ignacio Arribas. (2019) An Automatic Non-Destructive Method for the Classification of the Ripeness Stage of Red Delicious Apples in Orchards Using Aerial Video. Agronomy 2019, 9, 84

[21] D. Jamaludin, S. Abd Aziz, N. U. A. Ibrahim. (2014) Dielectric Based Sensing System for Banana Ripeness Assessment. International Journal of Environmental Science and Development, Vol. 5, No. 3, June 2014

Applying the Decision Tree Method in Identifying Key Indicators of the Digital Economy and Society Index (DESI)

A. Kutnjak, L. Hrustek and S. Križanić
University of Zagreb / Faculty of Organization and Informatics, Varaždin, Croatia
akutnjak@foi.unizg.hr; lhrustek@foi.unizg.hr; skrizanic@foi.unizg.hr

***Abstract* – The fourth industrial revolution is affecting the way people live, work and communicate. Driven by rapidly changing technology, it reflects on a new paradigm - digital transformation (DT). DT seeks to strike a balance between improving business operations and new opportunities that the revolution brings. Essentially, DT brings changes in the business model making it fully end-user oriented. Well prepared are those organizations that accept transformation at all levels and incorporate DT into strategic plans. Because organizations approach the transformation process in different ways, so is the impact of DT on their competitiveness level also different. There are various methods for measuring the success of the transformation process, and Digital Economy and Society Index (DESI) is one of them. It measures the digital competitiveness of EU Member States through defined dimensions and sub-dimensions, and ranks the observed countries accordingly. This paper is aimed to identify key dimensions and sub-dimension that affect the rank, using the method of the decision tree for this research.**

Keywords - digital transformation; digital economy and society index; decision tree; data mining

I. INTRODUCTION

The digital revolution is present in all segments of society, so it grabs and influences changes in the lives of individuals as well as in the sphere of business. Everyday happenings are imbued with concepts from the digital world, and their impact reflects on all economic spheres and market conditions. Organizations that follow current changes and developments in the market are certainly aware of the latest trend called digital transformation.

Through digital transformation, organizations seek to implement new business initiatives - new business models, new products and services, based on a fully personalized customer experience. In doing so, they go beyond traditional ways of doing business, foster networking with the community and the entire ecosystem, and emphasize the importance of constant innovation for the purpose of restructuring strategy and how to market [1]. Digital transformation changes business models of organizations, meaning that with the use of digital technologies, it impacts the way they work and create value. To all the above, the potential of new digital technologies is being exploited which directly affects on business process improvement, governance and organizational structure [2]. Given the complexity of the transformation process, organizations need to define an adequate digital strategy that, as such, will be the guiding principle for the integration of the entire business and represent a systematic elaboration of priorities and ways of implementing them. In addition, the strategy must have clearly measurable objectives, whereby focus is not on only technology, but technology will represent a strategic tool for achieving the goals set [3].

Success in achieving digital strategies is not just related with individuals or organizations - namely, it also defines their respective goals and the way how they are measured. One of the measurable indicators of a country's digital performance is the Digital Economy and Society Index (DESI). It investigates the digital performance of EU Member States, their digital competitiveness and efficiency across predefined dimensions.

In line with the aforementioned index, the following research question arises: to what extent is each individual component of the dimension of the DESI index (sub-dimension) a predictor of the digital development class? The answer to this research question will be explored using decision tree, one of the data mining method, using the RapidMiner tool.

The paper is structured as follows. The research methodology is defined at the very beginning. Thereafter, a review of the literature is given to explain the concept of data mining and the decision tree method. The application of the decision tree method, based on the analysis of the DESI index, is presented in the next section. The conclusion at the end of the paper gives a systematic overview of the results and the lessons learned.

II. RESEARCH MODEL

The processing of the research topic began with a review of published papers available in scientific, bibliographic databases - Web of Science (WoS) and Scopus. The search in the mentioned databases was carried out on the basis of the complex query "*data mining*" and "*decision tree**", in the title of papers. Search was limited to the type of papers and it included only articles and papers published at conferences.

This research has been conducted as part of the wider research in the project Development of innovative platform for digital transformation of enterprises (RDI) which is funded by European Union through the European Regional Development Fund (ERDF).

The search resulted with 196 articles in Scopus and 90 in WoS. An additional limit was placed on the citation of papers and accordingly, papers with a citation of more than 40 citations were taken into account. Consequently, the number of papers obtained in Scopus was 11, while in WoS there were 6. A basic comparison of the datasets showed that some results were found in both platforms, which ultimately resulted with 11 relevant papers directly related to the research topic.

The term "*digital transformation*" and "*decision tree*" were further searched through Google Scholar. This search included 5 new papers in to the research. Finally, a set of 16 papers was analyzed, with the aim of gaining insight into the research topic and answering a research question previously asked.

III. LITERATURE REVIEW

Data mining is a way of automated detection of non-trivial, previously unknown patterns and rules [4], therefore, many pattern recognition algorithms and techniques have been developed to detect complex and complicated but potentially useful patterns, between input variables and output targets [5].

Data mining techniques involve the exploration and analysis of large amounts of data to discover relevant knowledge [6], and are particularly significant today when the digital transformation of all aspects of life has led to the storage of vast amounts of data [7]. Data mining, as the basis of knowledge discovery in data, has been successfully applied in many fields, including finance, health and medicine [8], manufacturing, e-commerce and more [9]. Identified and recognized potential patterns bring benefits to researchers, which cannot be identified by mathematical and statistical methods. The fundamental objective of data mining is to discover hidden information in the data that will be used to support decision making, to make predictions, and to evaluate or provide a basis for developing new strategies and knowledge [4]. In literature, decision tree and neural networks are mentioned as two basic methods of data mining [5]. This methods are considered popular personalized learning methods and for optimizing personal learning [10].

The decision tree is the most significant method that can be used to classify data according to rules derived from input variables and presents tree-shaped data, to investigate the importance of influencing factors [9]. Different decision tree algorithms correspond to different types of data. Of the many available algorithms, the most appropriate method for processing continuous data, such as anthropometric data, is the classification and regression tree (CART). CARTs developed in the areas of pattern recognition, statistics and machine learning, are of particular interest in data mining because they use symbolic and interpretative presentation [11], and represent a model that is both predictive and descriptive [4]. CART puts pre-processed data at the root of a tree, and then, through a series of sorting processes, determines the best separation points it uses to shape the tree-shaped structure [9]. The classification tree evolves when the prediction target is discretely evaluated, while the regression tree develops when the target is numerical and continuous [11].

The decision tree is a type of tool based on intuitive rules and from this arises the attractiveness of the model [5]. It has a flow chart structure consisting of inner nodes, leaf and branch nodes [7], and is drawn with the root at the top and the leaves at the bottom [5]. Each inner node represents a decision [7] and involves testing a specific data attribute [6]. Each outgoing branch corresponds to the possible outcome of the test [7], respectively represents all possible test results [6]. Each leaf node represents a class or category [7] that indicates the majority class within the end partition [6]. To classify an unlabeled data sample, the classifier tests the sample attribute values against the decision tree [7]. The classification phase [6] is the path followed from the root to the leaf node that holds the class prediction for that sample [7]. Starting from the root, the instance value of a particular attribute decides which branch will go at each inner node. Whenever a leaf node is reached, the associated class tag is assigned to the instance [6]. Different leaves may make the same choice, but each leaf makes that choice for a specific reason. The tests are chosen so that they can best be distinguished from the targeted choices. Each path from root to leaf represents a decision rule [5].

The flexible structure of the decision tree model enables it to adapt to training data, and the model has the ability to create understandable rules and recognize the significance of variables [5]. Decision tree models can be easily converted to a set of IF - THEN rules using a set of reasoned attributes [5] and are used for decision making [7]. A set of rules can be expressed in the following form:

IF <condition> THEN <prediction (class)> [12].

The advantage of this knowledge presentation is that it is intuitive to users [12] and can be easily explained. Some other advantages of the decision tree are that (a) it is not necessary to specify the structure of the model in advance, thus reducing the incompatibility between the model structures and the explanations; (b) they have the ability to model nonlinear systems, represented by the more complex relationships involved in human behavior; and (c) induced patterns, in most cases, can be drawn from a subset of observations with homogeneity [5].

The decision tree cannot capture correlations between attributes or rule sets, and has a problem processing continuous data, which is why the data must first be grouped into ranges, manually or automatically by a software tool. Selected ranges may inadvertently hide useful patterns. One of the drawbacks is related to assessment when a tree model makes the decision about the variable to which it will divide the node, the decision cannot be changed or improved due to the absence of a backtracking technique [5]. Decision trees can sometimes be quite difficult to understand when the size of a tree is too large [13].

It can be concluded that this favorite method among analysts, the decision tree method, despite its minor drawbacks, provides transparent calculations in the modeling process [14], greater accuracy of calculations [15], and provides relevant forecasting models [16].

IV. APPLYING THE DECISION TREE METHOD IN IDENTIFYING KEY INDICATORS OF DESI

A. Description of the used data

For the analysis based on the decision tree, in this article data from the Digital Economy and Society Index (DESI) for the European Union (EU) Member States was used. DESI is a joint index that includes a summary of 30 relevant indicators of European digital performance. Indicators are measured in five areas called dimensions, namely *Connectivity*, *Human Capital*, *Use of Internet*, *Integration of Digital Technology* and *Digital Public Services*. Each of the dimension within itself measures sub-dimensions, while they include detailed indicators. The data used for the analysis has been published on the European Commission's website [17] and the research is conducted in the area of Digital Single Market, Digital Economy and Society.

This article uses sub-dimension data expressed in an index for all 28 European Union Member States for 2018, and its structure is presented in Table 1.

TABLE I. DESI STRUCTURE

Digital Economy and Society Index (DESI)	
Dimensions / (weight)	**Sub-dimensions / (weight)**
1 Connectivity (25%)	1a Fixed Broadband (18.5%)
	1b Mobile Broadband (35%)
	1c Fast Broadband (18.5%)
	1d Ultrafast Broadband (18.5%)
	1e Broadband Price Index (9.5%)
2 Human Capital (25%)	2a Internet User Skills (50%)
	2b Advanced Skills and Development (50%)
3 Use of Internet (15%)	3a Internet Use (25%)
	3b Activities online (50%)
	3c Transaction (25%)
4 Integration of Digital Technology (20%)	4a Business digitisation (60%)
	4b e- Commerce (40%)
5 Digital Public Services (15%)	5a e-Government (80%)
	5b e-Health (20%)

B. Results of applying the decision tree method

The process of analyzing the data began by clustering, using k-Means algorithm, with the values of each of the dimensions of individual countries being included in the analysis (Figure 1). The clustering process of the 28 EU Member States was based on the realized value of the DESI dimensions. After several attempts, the number of clusters k = 4 was defined, to best fit the amount and size of the data. The benefit of adding or reducing the number of clusters was minimal, with the view that any change in the cluster had visibly impaired the consistency of country affiliation to a particular cluster.

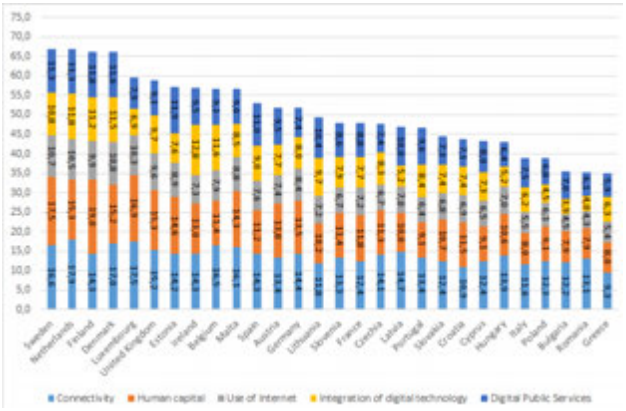


Figure 1. Data used for clustering

Clustering of countries into four groups or classes is visible, according to the analyzed characteristics, named as follows: (3) **digital leaders**; (1) **followers**; (0) **class III**; (2) **class IV**.

The data selected and the ranking of EU countries show that Sweden, Netherlands, Finland and Denmark are the most competitive digital countries (class (3)). By contrast, the least digitally competitive countries are Bulgaria, Romania and Greece. Poor competitiveness is also evident in Croatia, which is in (0) class III according to clustering.

A detailed overview of the countries, based on their assigned classes, is shown in Table 2 below.

TABLE II. COUNTRY CLASSES ACCORDING TO CLUSTERING

Class/Class name			
(3) Digital leaders	**(1) Followers**	**(0) Class III**	**(2) Class IV**
Sweden	Luxembourg	Spain	Hungary
Netherlands	United Kingdom	Austria	Italy
Finland	Estonia	Lithuania	Poland
Denmark	Ireland	Slovenia	Bulgaria
	Belgium	France	Romania
	Malta	Czech	Greece
	Germany	Latvia	
		Portugal	
		Slovakia	
		Croatia	
		Cyprus	

The process of clustering countries into four clusters allowed the continuation of the analysis based on the decision tree method, which reveals the extent to which a particular dimension and sub-dimension is a predictor of digital development. For the model itself, input variables (independent) were defined, sub-dimension values, while

the class represents the output variable, that is, the target variable.

Figure 2 shows the decision tree for the first DESI dimension - *Connectivity*. It indicates that the sub-dimension "Fast Broadband" is the most important for classifying countries according to the mentioned DESI dimension. The following rule for the classification of countries has been deduced from tree branching: IF (("Fast Broadband" > 0.562) and ("Broadband Price Index" > 0.144) and ("Fixed Broadband" ⩽ 1.138)) THEN classify country = digital leaders. IF (("Fast Broadband" > 0.562) and ("Broadband Price Index" ⩽ 0.144)) THEN classify country = followers. IF (("Fast Broadband" ⩽ 0.562) and ("Broadband Price Index" ⩽ 1.170) and ("Mobile Broadband" > -0.960) and ("Fast Broadband" > -0.733) and ("Ultrafast Broadband" > -0.062)) THEN classify country = class III. IF (("Fast Broadband" ⩽ 0.562) and ("Broadband Price Index" ⩽ 1.170) and ("Mobile Broadband" ⩽ -0.960)) THEN classify country = class IV.

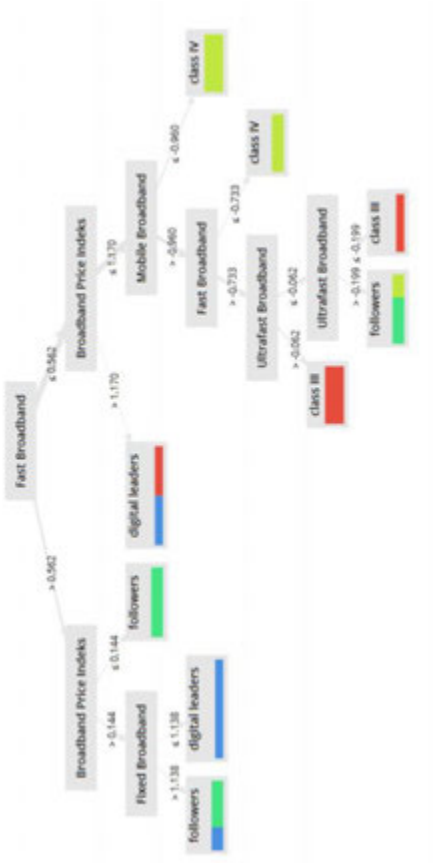


Figure 2. Decision tree - dimension *Connectivity*

The second dimension of the DESI index relates to *Human Capital* and its associated decision tree is shown in Figure 3. The tree indicates that the main sub-dimension for classifying countries into specific classes is "Internet User Skills". A simple classification rule for digital leaders can be deduced from the generated tree: IF ("Internet User Skills" > 1.131) THEN classify country = digital leaders. IF (("Internet User Skills" ⩽ 1.131) and ("Advanced Skills and Development" ⩽ 0.541) and ("Advanced Skills and Development" ⩽ -0.181) and ("Internet User Skills" ⩽ 0.047) and ("Internet User Skills" ⩽ -0.709)) THEN classify country = class IV.



Figure 3. Decision tree – dimension *Human Capital*

According to decision tree, the most significant sub-dimension of the dimension *Use of Internet* is "Transaction" (see Figure 4). The following country classification rules may be defined: IF ("Transaction" > 1.007) then classify country = digital leaders. IF (("Transaction" ⩽ 1.007) and ("Transaction" ⩽ 0.229) and ("Internet Use" ⩽ -0.012) and ("Internet Use" ⩽ -0.993)) THEN classify country = class IV.

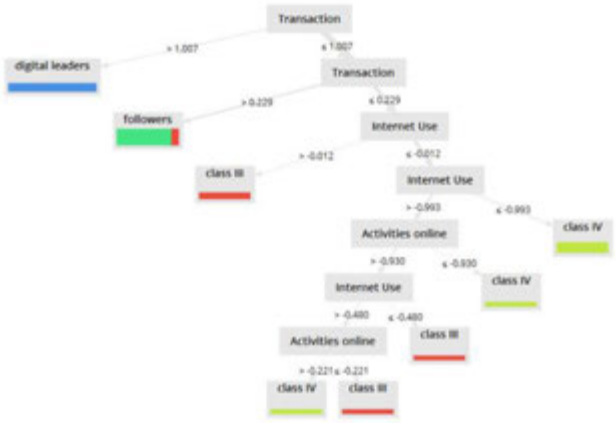


Figure 4. Decision tree – dimension *Use of Internet*

Decision tree related to dimension *Integration of Digital Technology* is shown in Figure 5. The most significant sub-dimension in this segment, and relevant to the classification, stands out "Business digitisation". Accordingly, the following classification rules may be mentioned. IF (("Business digitization" > 0.633) and ("Business digitization" > 1.617)) THEN classify country = digital leaders. IF ("Business digitization" ⩽ 0.633) and ("e-Commerce" ⩽ 0.028) and ("Business digitization" ⩽ 0.041) and ("Business digitization" ⩽ - 0.310)) THEN classify country = class IV.



Figure 5. Decision tree – dimension *Integration of Digital Technology*

The last decision tree (shown in Figure 6.) refers to the fifth dimension of the DESI *Digital Public Services*, where "e-Health" stands out as the most significant sub-dimension. The classification rules are as follows: IF ("e-Health" > 1.445) THEN classify country = digital leaders. IF ("e-Health" ⩽ 1.445) and ("e-Government" ⩽ -0.645)) THEN classify country = class IV.

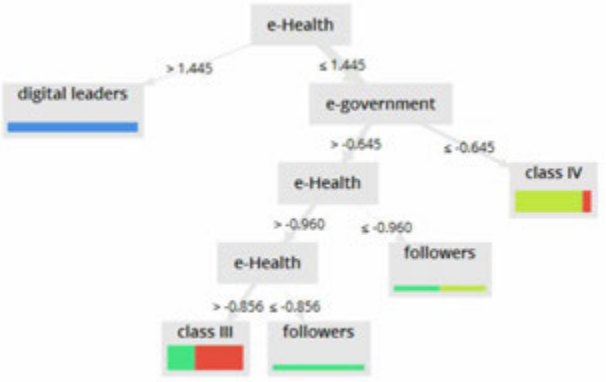


Figure 6. Decision tree – dimension *Digital Pubic Services*

The application of the decision tree method and the generation of its algorithms shows that the relevant variables, which enable position prediction in determining the digital competitiveness of European Union countries, can be deduced. The most relevant sub-dimensions are those that appear as splitting criteria in the decision tree, that is, the most important are those sub-dimensions that are at the top of the tree. According to the results, the most relevant dimension for *Connectivity* is "Fast Broadband", for *Human Capital* variable "Internet User Skills", for *Use of Internet* it is "Transaction", dimension *Integration of Digital Technology* identifies sub-dimension "Business digitisation" as most relevant for classification, while "e-Health" is a prominent variable of the dimension *Digital Public Services*. Thus, the answer to the initial posed research question was established.

V. CONLUSION

This paper analyzes the DESI index, which describes the digital competitiveness of EU countries across the five major dimensions and their sub-dimensions. DESI was analyzed using clustering and decision tree methods with the RapidMiner tool. The first step in conducting this research was to determine the digital development classes of countries for which clustering was used. Furthermore, the decision tree method has generated models for each of the dimensions of the DESI index, which show us key, predictive variables for determining the digital competitiveness and development of EU countries. Based on IF-THEN rules, the conditions for classifying countries into one of the previously defined classes were obtained: digital leaders, followers, class III and class IV.

The scientific contribution of this paper is the analyzed structure of DESI index, which identifies key indicators and their dimensions. These indicators can be some guidelines for countries investments at their vision of improving competitiveness in the market and achieving successful results in the coming period. The analysis of DESI variables has identified those that are the most relevant to the digital competitiveness of countries – fast broadband, Internet user skills, transaction, business digitisation and e-Health. Therefore, a detailed decision tree analysis can become a guide for developing and adjusting country's strategy to improve digital competitiveness, and an indicator of what areas of effort should be directed to make the country more digitally competitive in the market as well.

REFERENCES

[1] S. Nambisana, M.Wright, and M. Feldmanc, "The digital transformation of innovation and entrepreneurship: Progress, challenges and key themes," Research Policy, pp. 1-9, 2019.

[2] C. Matt, T. Hess and A. Benlian, "Digital Transformation Strategies," Business & Information Systems Engineering, 57(5), pp. 339–343, 2015.

[3] G. C. Kane, D. Palmer, A. N. Phillips, D. Kiron and N. Buckley, "Strategy, not Technology, Drives Digital Transformation," MITSloan Management Review, 2015.

[4] W. J. Kuo, R. F. Chang, D. R. Chen, and C. C. Lee, "Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images," in Breast Cancer Research and Treatment, vol. 66, Issue 1, 2001, pp. 51-57.

[5] C. Xie, J. Lu, and E. Parkany, "Work Travel Mode Choice Modeling with Data Mining: Decision Trees and Neural Networks," Transportation Research Record: Journal of the Transportation Research Bord, vol. 1854, Issue 1, 2003, pp. 50-61.

[6] S. C. Chen, M. L. Shyu, M. Chen, and C. Zhang, "A Decision Tree-based Multimodal Data Mining Framework for Soccer Goal Detection," in IEEE International Conference on Multimedia and Expo, 2004, pp. 265-268.

[7] M. Kamber, L. Winstone, W. Gong, S. Cheng, and J. Han, "Generalization and Decision Tree Induction: Efficient Classificationin Data Mining," in IEEE Proceedings Seventh International Workshop on Research Issues in Data Engineering. High Performance Database Management for Large-Scale Applications, 1997, pp. 111-120.

[8] M.A. Karaolis, J.A. Moutiris, and C.S. Pattichis, "Assessment of the Risk Factors of Coronary Heart Events Based on Data

Mining With Decision Trees," IEEE Transactions on Information Technology in Biomedicine, vol. 14, Issue 3, 2010, pp. 559-566.

[9] C. H. Hsu and M. J. J. Wang, "Using decision tree-based data mining to establish a sizing system for the manufacture of garments," The International Journal of Advanced Manufacturing Technology, vol. 26, Issue 5-6, 2004, pp. 669-674.

[10] C.F. Lin, Y. Yeh, Y.H. Hung, and R.I. Chang, "Data mining for providing a personalized learning path in creativity: An application of decision trees," in Computers & Education, vol. 68, 2013, pp. 199-210.

[11] C. Apte and S. Weiss, "Data mining with decision trees and decision rules," in Future Generation Computer Systems, vol. 13, Issues 2-3, 1997, pp. 197-210.

[12] D. R. Carvalho and A. A. Freitas, "A hybrid decision tree/genetic algorithm method for data mining," in Information Sciences, vol. 163, Issues 1-3, 2004, pp. 13-35.

[13] N. D. Anuar, H. Sallehudin, A. Gani, and O. Zakari, "Identifying folse alarm for network intrusion detection system using hybrid dana mining and decision tree," in Malaysian Journal of Computer Science, vol. 21, Issue 2, 2008, pp. 101-115.

[14] W. Chen, S. Zhang, R. Li, and H. Shahabi, "Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling," Science of The Total Environment, vol. 644, 2018, pp. 1006-1018.

[15] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction," in Ecosystem, vol. 9, Issue 2, 2006, pp. 181-199.

[16] L. R. Iverson, A. M. Prasad, S. N. Matthews, and M. Peteres, "Estimating potential habitat for 134 eastern US tree species under six climate scenarios," Forest Echology and Management, vol. 254, Issue 3, pp. 390-406, 2008.

[17] European Commission, "The Digital Economy and Society Index (DESI)," https://ec.europa.eu/digital-single-market/en/desi.

Determining the location of postal centers in B&H using machine learning clustering method and GIS

Amel Kosovac*, Ermin Muharemović*, Muhamed Begović*, Edvin Šimić*
* Faculty of Traffic and Communications, University of Sarajevo, Bosnia and Herzegovina
muhamed.begovic@fsk.unsa.ba

Abstract -**The rapid development of technology is directly affecting the growth and development of e-commerce shipments, especially in the Business to Customer segment. An increase in e-commerce shipments has a strong impact on the express delivery industry. In these conditions, a very significant challenge is how to organize a postal network. The problem that arises is how many postal centers, and at what locations, should be implemented in a specific geographical area in order to optimize the level of service for the users. Solving this challenge has latterly received increased attention in both industry and academia. The aim of this paper is to firstly provide a concise overview of current approaches in the process of determining the optimal location of postal centers. The second part of the paper will focus on proposing an approach that will rely on machine learning methods for clustering in defined conditions and specific geographical environment using appropriate geographic information tools for spatial data analysis and visualization.**

Keywords - postal centers; B2C model; shipments pickup and delivery; machine learning; geographic information systems

I. INTRODUCTION

The development of technology has introduced new ways of communication between people and enabled a whole new approach for products sale. Today there is almost no geographical barrier in the market. E-commerce has completely eliminated these barriers and created new challenges in the postal industry [1]. This type of market allows people in rural areas to have direct access to buying a different type of products. The customer wants the ordered product to be delivered to his address as soon as possible. These new conditions have put a grave task for postal companies in terms of network organization, delivery quality and cost management.

In order to find a balance between customer expectations and costs, the postal companies have to pay more attention to problems of defining the optimal number of postal centers in some geographic areas. Based on the number of shipments in pickup and delivery for a specific geographic area, new approaches enable the simpler ways for the determination of the optimal number of postal centers. To reduce costs, postal companies adopt a hub-and-spoke network structure [2]. The location of the central warehouse or hub is determined in the geographical area in which is the highest percentage of picked up and delivered shipments, as well as on the development of transport infrastructure.

In this paper, the determination of clusters or postal centers is made according to the number of shipments in the pickup and delivery using the clustering *K-means* machine learning method. The data used in this research are for a period of one month, with over a 30.000 number of shipments in pickup and delivery. The research presented in this paper integrates the use of machine learning clustering and GIS software for postal centers number and position determination in Bosnia and Herzegovina.

II. RELATED RESEARCH

The pioneer of researching the facility location in relation to customers was Weber in 1909 with the main focus of reducing the total distance and overall cost. The one-hub siting problem introduces by O'Kelly [3] and it is equivalent to a Weber least-cost location model. This paper has introduced the problem of locating hub facilities so as to serve a set of interacting cities. The problem of hub determination can be viewed in two ways, from network characteristics and location problems.

Article [4] describes the simple plant location problem (SPLP) considering an important family of discrete, deterministic, single-criterion and widely applicable optimization problems. In addition to the methods described above, clustering methods are also used to determine the optimal center location.

Fuente and Lozano used cluster analysis and cost evaluation for determining warehouse number and location in Spain [5]. One of the basic algorithms for clustering is *k-means* clustering. Because of its simplicity and easy application, it is often used today even though it was described more than 60 years ago. Choosing a postal center location is a problem that has been around for a long time and it is researched mainly by supply chain researchers. The Warehouse Location Problem (WLP) problem has often been described as a clustering problem, and many papers have developed methods to solve this problem. In addition to this problem, we need to look at the postal center in the network perspective, particularly as a hub-and-spoke system.

Hub location problem is the problem that frequently arises in the design of transportation and distribution systems, postal delivery networks and airline passenger flow [6]. A hub-and-spoke network typically involves three simultaneous decisions: deciding the optimal number of hub nodes, their locations and the appropriate allocation of the non-hub nodes to the hubs [7]. When using the *K-means* method, we can also select the number of clusters, and this is one of the very important factors when using this method.

Authors in [8] illustrate some of the key issues: measuring similarity, forming clusters, and deciding on the number of clusters that best represent a structure that will show in our paper later on. Paper [9] finds the optimal allocation of warehouses and hospitals to demand cluster centers. In the paper [10] authors transform a special location-allocation problem into a clustering problem. Clustering methods used for logistics networks optimization in GIS allow us to visualize problems and make it easier to decide on the number of clusters relative to many practical issues that software cannot take into account. However, recent studies integrate machine learning into GIS software, as will be shown in this paper. Several machine learning models were used with GIS integration to determine the potential site for the proposed hotel based on business success indicators [11].

Previously, GIS was used only as a tool for better visualization of existing data and previously conducted research. GIS however, provides far greater possibilities than mere visualization. Nowadays, GIS is used for analyzing large amounts of data generated by a complex system such as the traffic. It performs demanding spatial analyzes while at the same time helping in making decisions important to many segments of traffic engineering [12]. In [13] authors used the clustering method in GIS to develop a tool for planning and managing the reuse of agricultural drainage water for irrigation in the Nile Delta. There is also an approach [14] to conduct the geochemical analysis by using the machine learning approach. For their analysis, Simple *K-means* is used. The decision support system in logistics inbound and transportation system with the application of Geographic Information System (GIS) was presented in [15]. The clustering method is often used for solving the problem of location and network routing in order to reduce costs [16]. In recent years, there has been great progress in solving the problem of facility location based on clustering. Continuous facility location problem and its application to a clustering problem are also described in [17]. The authors of this paper propose good approximation algorithms for Euclidean and squared Euclidean distance functions.

III. ANALYSIS AND INTERPRETATION OF RESULTS

The application of machine learning methods to optimization processes in various fields enables the detection of relations and patterns among large amounts of data.

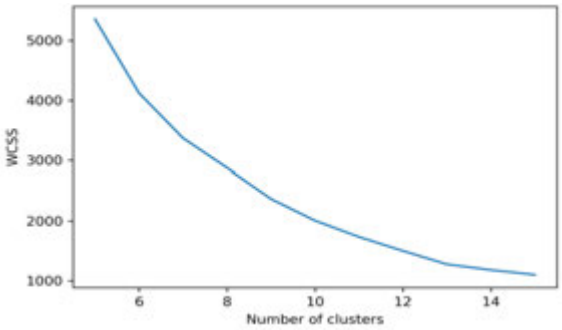
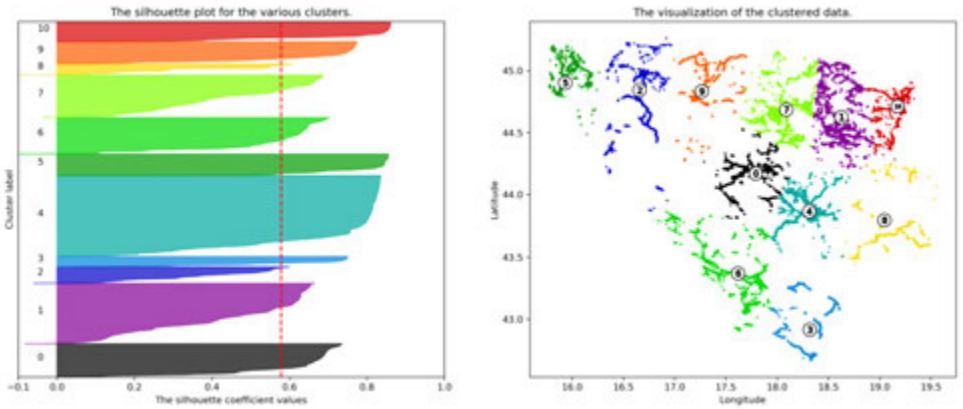


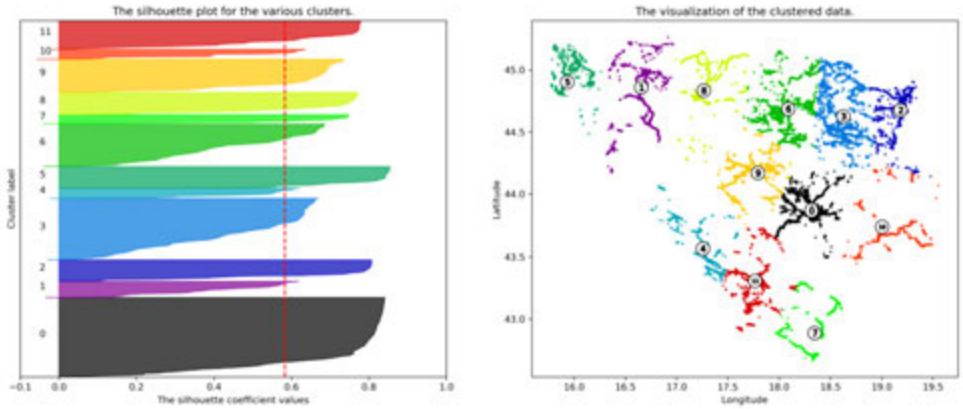
Figure 1. Results of the Elbow method

In the practical implementation, real-world data were used on the locations of pickup and delivery of shipments. These data were generated by one of the operators of commercial courier services that covers the territory of Bosnia and Herzegovina.

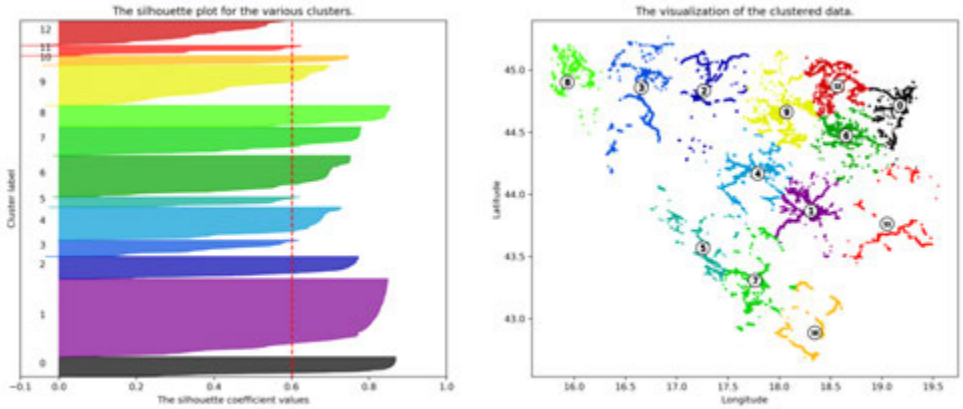
Data exclusively considered the B2C (Business-to-consumer) packet pickup and delivery segment. The authors of this have ignored the existing layout of the postal centers and have tried to determine the optimal number and location of the postal centers. Given that the data were unlabeled, it is clear that the described problem has to be solved by unsupervised machine learning methods. Additionally, as it is necessary to determine the affiliation of each point to a particular center from which a limited geographical area would be served, the problem is clearly characterized by clustering methods. Clustering is one of the most used data exploration techniques for grasping knowledge about data structure. It allows identifying subgroups in the data so that the data in the same subgroup (cluster) is very similar, while the data points in the different subgroups are very different. To implement clustering, the *K-means* method was used, primarily because of its ease of implementation and the fact that it fits well with large datasets. Considering the fact that the research includes over 30.000 records of individual pick-ups and deliveries, it is clear why K-*means* was suitable. Random initialization trap (choosing initial cluster center points so that the algorithm gives a false positive model) was avoided using *K++ means* advanced method. The algorithm was implemented using *Python* programming language. Program code includes *pandas* and *NumPy* libraries for importing and preprocessing the data, *sklearn* for implementing the *K-means* method, and finally *matplotlib* for graphical displaying of the results. The *K-means* method requires manual determination of the number of clusters before implementing the algorithm. Due to the geographical specificity of Bosnia and Herzegovina, the minimum number of centers should be 5. Existing companies that deal with the commercial services of pickup and delivery have an average of about 10 centers. Therefore, the research was conducted by implementing the *K-means* method with a set of different values of the number k, where $k \in [5,15]$ and compared the numerical and graphical results for each selected number of clusters. In the first attempt to determine the optimal number of clusters *Elbow method* was used.



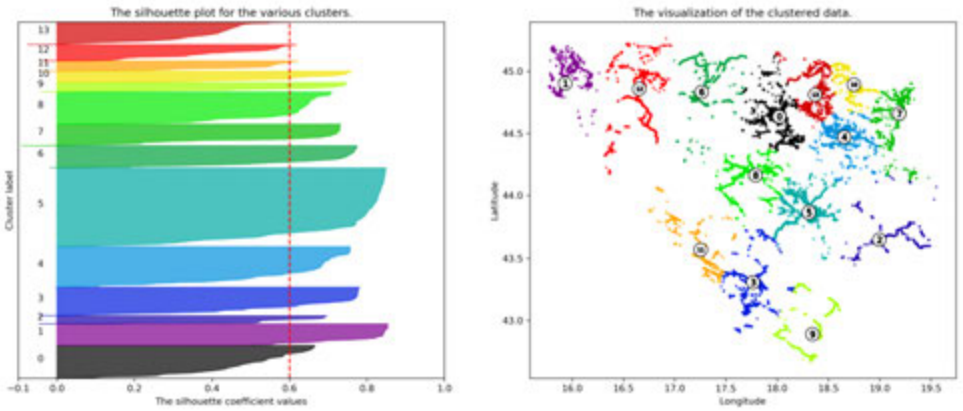
(a) Silhouette analysis for K-means clustering: k=11



(b) Silhouette analysis for K-means clustering: k=12



(c) Silhouette analysis for K-means clustering: k=13



(d) Silhouette analysis for K-means clustering: k=14

Figure 2. Determining the optimal number of clusters using silhouette analysis

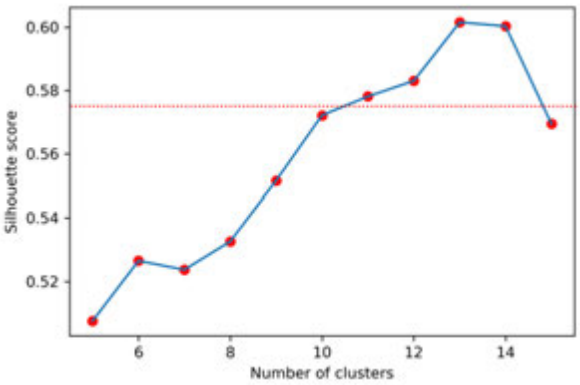


Figure 3. Results of the Silhouette score method

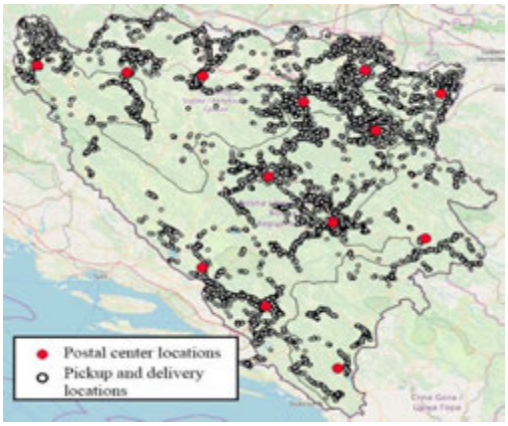


Figure 4. Part of QGIS visualization for insight in real-world of the determined geographical location of centers

The *Elbow method* is a heuristic method for evaluating consistency within a cluster. The number of clusters is determined by increasing clusters as long as it results in significantly better data modeling. However, the results obtained were vague and difficult to interpret (see Fig. 1). It was very difficult to determine where the "*elbow*" was located, that is, for each increase in the number of clusters, the value of the *WCSS parameter* (within-cluster sum of squares) decreased with relatively the same intensity.

If C denotes a cluster, S_i denotes all shipments that belong to one cluster, d is the distance between each shipment S_i and its corresponding cluster C, then WCSS can be defined as (n is the observed number of clusters):

$$WCSS = \sum_{S_i \in C_1} d(S_i, C_1) + \sum_{S_i \in C_2} d(S_i, C_2) + \cdots \sum_{S_i \in C_n} d(S_i, C_n) \quad (1)$$

As the *Elbow method* did not produce clear results for interpretation, a *Silhouette analysis* was introduced. Silhouette value is a way of assessing the similarity of a feature and cluster to which it belongs. It can take a value between -1 and +1. A higher value means that a particular feature is more evident to belong in its own cluster, that is, it is more distant from neighboring clusters. If too many points have a small or even negative value, then it is highly likely that the number of clusters is not ideal for particular data.

If $d_a(i)$ denotes the mean distance between S_i and all other data points in the same cluster, and $d_b(i)$ denotes the smallest mean distance of S_i to all points in any other cluster, of which is S_i not a member, then silhouette value can be defined as:

$$s(i) = \frac{d_b(i) - d_a(i)}{\max\{d_a(i), d_b(i)\}} \quad (2)$$

In Fig. 2, one can see in detail the *Silhouette score* for each value of k, where $k \in [11,14]$, as well as for each individual cluster and location within that cluster. In addition, the two-dimensional Cartesian coordinate system shows the locations of each individual pickup and delivery, as well as the estimated ideal positions of the centers (clusters). Affiliation to a cluster is interpreted by common color for all locations within its area of coverage.

Using this method has yielded much clearer results. It showed that for a set threshold value of 0.577, the number of centers between 11 and 14 is acceptable, and the best results were achieved with k = 13 (see Fig. 3.).

Unfortunately, the upbuilding of a postal center in an optimal location determined by the application of these algorithms is often not practically possible. The point of the cluster center may be in an unfavorable location. These reasons may include a location in the city center where such facilities are not feasible, the absence or high cost of buying or renting suitable land for construction, protected areas where construction is prohibited, inadequate access traffic infrastructure, etc. Therefore, certain optimal locations of centers determined by the algorithm must be adjusted in space according to the characteristics of the environment.

The output of results, that is, the position of clusters (postal centers) in the coordinate system of the *matplotlib* library does not give insight into the real geographical position of the points. It is much easier to draw conclusions using geoinformation systems that provide rich visualization and additional analysis of spatial data. For this reason, the data *.csv* output from *Python* has been imported into *QGIS (Quantum Geographical Information System)* software and interpreted the actual geographical position of certain clusters using the *Openstreet* basemap. From the visualization in GIS (see Fig.4.), it was concluded that certain positions of the centers are closest to the following larger settlements: Sarajevo, Zenica, Bihać, Mostar, Trebinje, Posušje, Goražde, Bijeljina, Tuzla, Brčko, Banja Luka, Prijedor, and Doboj.

IV. CONCLUSION

In optimizing the costs of organizing a postal network, great attention is paid to the quality of service delivery and customer satisfaction. Only through the good organization of pickup and delivery, and an optimal number of centers

is it possible to achieve a high level of customer satisfaction in the B2C segment. The analysis of the *K-means* method results demonstrates that there is potential for the application of machine learning and geoinformation systems for determining the number and spatial dispersion of postal centers.

The data used in this research were related to the distance between the delivery/pickup locations and respective postal centers. The focus was not on other factors such as type of shipment, the weight of shipments, development, and condition of transport infrastructure, postal center fleet capacity, number of stops in pickup and delivery, etc. In future research, it is desirable to test the accuracy of the method by considering multiple factors, that is, multiple dimensions that influence the selection of the optimal number of centers and their location. Also, it would be useful to compare and verify the results using some other clustering methods such as hierarchical clustering.

REFERENCES

[1] S. Čaušević, E. Muharemović, B. Memić, and M. Begović, "Integration of logistics information systems with electronic sales channels," in *International Scientific Conference "Science and Traffic Development" - ZIRP 2018*, 2018, pp. 53–62.

[2] M. Masaeli, S. A. Alumur, and J. H. Bookbinder, "Shipment scheduling in hub location problems," *Transportation Research Part B: Methodological*, vol. 115, pp. 126–142, 2018.

[3] M. E. O'Kelly, "Location of Interacting Hub Facilities.," *Transportation Science*, vol. 20, no. 2, pp. 92–106, 1986.

[4] J. Krarup and P. M. Pruzan, "The simple plant location problem: Survey and synthesis," *European Journal of Operational Research*, vol. 12, no. 1, 1983.

[5] D. de la Fuente and J. Lozano, "Determining warehouse number and location in Spain by cluster analysis," *International Journal of Physical Distribution & Logistics Management*, vol. 28, no. 1, pp. 68–79, 1998.

[6] M. Naeem and B. Ombuki-Berman, "An efficient genetic algorithm for the uncapacitated single allocation hub location problem," in *2010 IEEE World Congr. Comput. Intell. WCCI 2010 - 2010 IEEE Congr. Evol. Comput. CEC 2010*, 2010.

[7] M. E. O'Kelly, "Hub facility location with fixed costs," *Papers in Regional Science*, vol. 71, no. 3, pp. 293–306, 1992.

[8] C. Beckett, L. Eriksson, E. Johansson, and C. Wikström, *Multivariate Data Analysis (MVDA)*. 2017.

[9] L. Özdamar and O. Demir, "A hierarchical clustering and routing procedure for large scale disaster relief logistics planning," *Transportation Research Part E: Logistics and Transportation Review*, vol. 48, no. 3, pp. 591–602, 2012.

[10] K. Liao and D. Guo, "A Clustering-based approach to the capacitated facility location problem," *Transactions in GIS*, vol. 12, no. 3, pp. 323–339, 2008.

[11] Y. Yang, J. Tang, H. Luo, and R. Law, "Hotel location evaluation: A combination of machine learning tools and web GIS," *International Journal of Hospitality Management*, vol. 47, pp. 14–24, 2015.

[12] S. Čaušević, A. Deljanin, M. Begović, and E. Deljanin, "Potentials and advantages of applying geographic information systems in various fields of traffic engineering," in *Road and Rail Infrastructure V*, 2018, vol. 5, pp. 1285–1290.

[13] M. Shaban, B. Urban, A. El Saadi, and M. Faisal, "Detection and mapping of water pollution variation in the Nile Delta using multivariate clustering and GIS techniques," *Journal of Environmental Management*, vol. 91, no. 8, pp. 1785–1793, 2010.

[14] G. H. Alférez, J. Rodríguez, B. Clausen, and L. Pompe, "Interpreting the geochemistry of southern California granitic rocks using machine learning," in *Proc. 2015 Int. Conf. Artif. Intell. ICAI 2015 - WORLDCOMP 2015*, 2019, no. Lil, pp. 592–598.

[15] R. Tangkitjaroenmongkol, S. Kaittisin, and S. Ongwattanakul, "Inbound Logistics Cassava Starch Planning," in *Eighth Int. Jt. Conf. Comput. Sci. Softw. Eng. Inbound*, 2011, pp. 204–209.

[16] S. Barreto, C. Ferreira, J. Paixão, and B. S. Santos, "Using clustering analysis in a capacitated location-routing problem," *European Journal of Operational Research*, vol. 179, no. 3, pp. 968–977, 2007.

[17] L. A. A. Meira and F. K. Miyazawa, "A continuous facility location problem and its application to a clustering problem," in *Proc. ACM Symp. Appl. Comput.*, 2008, pp. 1826–1831.

Defining ERP system selection methodology – research carried out in small and medium-sized production companies in Croatia

Alan Mahmutović *, Milorad Nikitović **
* Elektroprojekt, Zagreb, Croatia
** College for Information Technologies, VSITE, Zagreb, Hrvatska
alan.mahmutovic@elektroprojekt.hr; milorad.nikitovic@vsite.hr

Abstract - Each company has to independently decide on the manner of selection of an optimal ERP solution. In order to achieve realistic results, it is necessary to, before every interview or presentation with suppliers, and especially before the announcement of a tender, define a selection model. From the existing literature, it is evident that there are a number of factors that could be deemed important in defining a selection model. This paper recommends 4 factor groups with 24 factors within each group. The questionnaires were sent to the project managers in companies who have implemented an ERP system and to ERP implementers in companies which supply ERP solutions. The importance of factor groups and individual factors represents a basis for potential buyers of ERP systems with regarding to their implementation in procurement contracts in which they should be used to a logical extent.

Key words - ERP system solution, selection methodology, SME.

I. INTRODUCTION

The acronym ERP (Enterprise Resource Planning) was created by the Gartner Group in the 1990s to refer to a system that integrates business processes [1]. In the days that have passed since this basic definition was established, the Gartner Group itself changed this definition several times – and more. In the early 2000s, the Gartner Group also defined ERP II, followed by ERP III ten or so years later. The reason for the first migration was to integrate SCM_(Supplier Chain Management), sales CRM (Customer Relationship Management) and PLM (Product Lifecycle Management) into the original version of ERP. The following migration towards ERP III was impelled by new technological breakthroughs towards Cloud Computing, SOA (Service Oriented Architecture) and the application of hybrid ERP solutions through server and web offers [2].

In defining the term and acronym of ERP, one of the essential definitions that must be taken into consideration is that which was coined by APICS (American Production and Inventory Control Society), and which defines ERP as a framework for organizing, defining, and standardizing the business processes necessary to effectively plan and control an organization so the organization can use its internal knowledge to seek external advantage [3].

The term SME refers to micro, small and medium-sized enterprises.

The importance of this research is reflected in the fact that, in 2017, micro, small and medium-sized enterprises accounted for 99.7% of Croatia's economy. That same year, micro, small and medium-sized enterprises employed 73.2% of all employees in private enterprises [4].

The definition of micro, small and medium-sized enterprises in Croatia is aligned with the definition of the European Commission [5]:

- Micro enterprises:
  - less than or 10 persons employed and
  - annual turnover and balance sheet up to or equalling 2 million Euro,
- Small enterprises:
  - 11 to 50 persons employed and
  - annual turnover and balance sheet up to or equaling 10 million Euro,
- Medium-sized enterprises:
  - 51 to 250 persons employed and
  - annual turnover up to or equalling 50 million Euro, and annual balance sheet up to or equalling 43 million Euro.

II. SELECTING THE OPTIMAL ERP SOLUTION-REVIEW OF LITERATURE

According to the literature, implementation methods and strategy have a significant effect on the critical success factors of the entire implementation procedure. According to Kumar et al. [6], this method implies pre-defined steps as the ″recipe″ for implementing ERP solutions. According to them, the method is 80% the work of consultants, 40% the work of ERP suppliers and 20% the work of hardware suppliers (some pointed out in the research the importance of more offered).

How can we describe the ERP implementation life cycle? There exist a number of definitions, depending on how we view this issue. In the context of this paper, we can say that this term implies the time period that has

elapsed between the moment that the formal decision to build a new system (or significantly improve the existing one) was made, and the moment that the formal decision on termination of use or implementation was made.

Esteves and Pastor [7] analysed the ERP implementation life cycle according to the general context depicted in Figure 1.

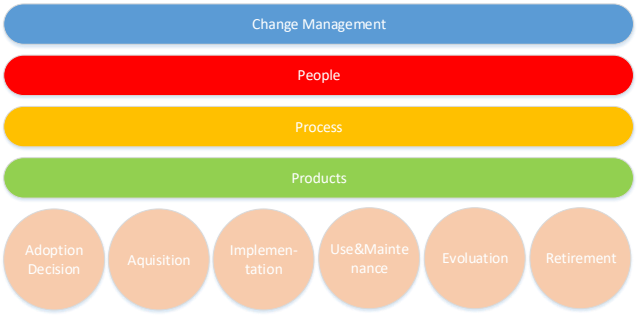


Figure 1. General context of the ERP life cycle (according to [7])

The general context consists of phases and dimensions. Phases represent various parts of the life cycle of an ERP organisation system, while dimensions refer to different viewpoints from which the phases can be analysed. The dimensional view presents a set of interconnected problems. Thus, change management includes culture and business culture, organisational structure, functions and skills, management of strategical changes and the reorganisation of business processes.

Pursuant to the work of Razi and Hossain [8], Filbert [9], Gupta and Kumar [10], [11], [12], Maheshwari et al. [13], Chen et al. [14] and Markus and Tanis [15], the following life cycle can be applied to our conditions:

- initiation and analysis of needs,
- procurement and planning,
- execution and control,
- testing and launch and
- support.

From the stated life cycle, it is evident that the process of selecting an optimum ERP solution stretches through the first two phases of the life cycle – initiation and analysis of needs, and procurement and planning.

A variety of very different approaches to the selection of an ERP solution can be found in the literature. These approaches can be divided into two basic groups. One of them implies defining 3-4 key factors, up to a maximum of ten or so factors. This approach is mostly connected with companies that either sell solutions or manage implementation. The other approach involves an attempt at grouping factors with corresponding definitions, both when it comes to factors and groups.

This paper [16] proposes 5 factors that need to be taken into consideration in the selection of appropriate ERP software: return on investment, translation of needs into requirements, introduction of functionalities that were not even considered, let alone applied in the previous phase, selection of an appropriate cost model, and the ability of the new software to remain in step with the changes in the core business of the enterprise and market changes. The paper [17] also proposes processes for the evaluation and selection of ERP packages. In order for these processes to be effective, the success of the following activities needs to be taken into account: formation of a committee for the selection of a solution, analysis of the needs of the enterprise, selection of criteria and the corresponding selection procedure (shortlist of suppliers, presentation, visits to companies from the list of users). A somewhat different view of the key factors and processes is presented in paper [18]. They include: the realistic change of needing an ERP solution (with an emphasis on new), assessing the key stakeholders and creating a list of requirements from the ERP solution, a shortlist of potential suppliers, sending a request for an offer and inviting suppliers to make their presentations and reaching a final decision. In the paper [19], the methodology for selection is divided into two groups: analysis of needs and analysis of suppliers. Analysis of needs comprises the following factors: understanding the critical factors for the success of implementation, measuring success, understanding existing business processes and the improvements that need to be made, precision in the defining of needs, the potential involvement of employees, risk management knowledge, finding the best internal user and ensuring the involvement of the top management. The analysis of suppliers comprises the following factors: identification of potential suppliers, finding a good consultant, drafting a request for an offer, the presence of all important persons at the presentation, reference check, checking if the system is suitable for the company, negotiating the price, conclusion of the contract and acknowledgment of total expenses. The company ERP Focus [20] proposes 10 steps for the safe selection of a suitable ERP solution. All of these steps have been covered in this part of the chapter.

Molnar et al. [21] have defined the selection process with 17 criteria, without any grouping of the same. In essence, these criteria can be divided into two groups. The first group is the position of the ERP solution and supplier on the global ERP market, and the second is the ability to resolve some of the segments of business functions, though not all of them. The latter segment of criteria is, to an extent, related to the critical factors for the success of ERP implementation.

Pitic et al. [22] propose an approach for evaluating and selecting an ERP solution that is based on four criteria groups. These are the preliminary phase (the PM and team, with the help of consultants, define the elements of the ERP project), the analysis phase (definition of the key business processes and identification of potential ERP suppliers), evaluation phase (defining the elements of the presentation, defining the technical criteria for selection and preparing the elements of the technical report) and the contract phase (identifying contract elements, defining a negotiation strategy, negotiation and the decision itself).

Haddara [23] proposes the selection of an ERP solution through 3 phases. The first phase implies a vision

of business from the aspect of the ERP solution, the second phase involves needs, the desire for change and limitations, while the third phase refers to the evaluation of the ERP solutions that made it to the shortlist.

Haddara [24] points towards the evident discrepancy in research of the reasons for the critical nature of the selection process. In selection, the harmonisation of the ERP system with organisational demands and needs is placed at the forefront. Therefore, according to him, the selection phase is pivotal, as if a company selects an inappropriate ERP solution, the entire project could be destined to fail.

III. RESEARCH METHODOLOGY

Pursuant to research of the available literature, we have compiled a list of the factors/criteria that may have a significant effect on the selection of a suitable ERP solution. 24 factors have been defined. In the attempt to rank them according to any manner, it was not simple to define which part of the key persons in the company should be in charge of each of them. Due to this reason, the factors were divided into the following four groups:

- factors related to the total expenses of implementation (6):
    - latent implementation expenses,
    - hardware and network expenses,
    - internal resource expenses,
    - licence expenses,
    - software maintenance expenses, and
    - external consultant expenses.
- analysis of the internal factors of the company (4):
    - vision and strategy (reason for selecting an ERP solution),
    - knowledge of the key factors for the successful implementation of ERP,
    - involvement of key persons in the ERP implementation project, and
    - potential budget (readiness to adapt).
- systematic software analysis factors (9):
    - appropriate functionalities according to production type,
    - flexibility (simplicity of adapting) the system,
    - reliability of the system,
    - simplicity of use,
    - suitability for the business processes of the company,
    - possibility of being upgraded by the user,
    - compatibility with other systems,
    - technological platform (cloud or server), and
    - option of supplementing the system with solutions that will coexist with the ERP solution.
- factors related to suppliers (5):
    - reputation of the supplier on the local ERP market,
    - potential of the supplier in the ERP implementation process,
    - possibility of developing and executing additional components,
    - terms and conditions related to the customer support plan, and
    - references according to the production type.

In the questionnaire itself, questions focused on the importance of each of the listed factors individually, not as part of their respective groups. The intent was for the importance of the group to be obtained as the summation of individual responses.

Therefore, the questionnaire consisted of 24 factors listed in alphabetical order, each of which had to be ranked according to their importance for the selection of an ERP solution:

- has no influence (1),
- has an influence (2) and
- has a strong influence (3).

Realistically, such a distribution of meaning on 3 levels could be questioned. However, involving 5 levels could have been potentially more difficult for the respondents. Moreover, this study wished to, in a certain manner, connect with research on the latent expenses of ERP implementation and the critical factors of ERP implementation, which were evaluated according to the same scale.

The questions were sent by e-mail to 73 production companies in Croatia in which, during 2012, research on the critical success factors (CSFs) in ERP implementation was performed. In 2018, research on the latent expenses of ERP implementation was performed in the same companies. At the same time, in addition to the questions sent by e-mail, a web programme that users were able to fill in was created, with an appearance identical to that of the word document. The choice of using the word document and returning the same by e-mail was a consequence of experiences from previous research. Pursuant to the aforementioned we asked two research questions:

- what is the perception of the importance of the selection of an ERP solution through the proposed factors in Croatian enterprises, and how would you rank the list of factors for the selection of an ERP solution according to their importance?

The questionnaires were sent in early January 2019, and they could be filled in by the end of March. As we did not receive the expected number of responses, we also included interviews in the research methodology. These interviews were performed in May and June 2019. One of the reasons that prompted us to decide on performing interviews was that additional explanations of each of the factors in the selection of an ERP solution might be necessary. Some of the names of the factors were ambiguous and could have been interpreted in different ways. Dividing certain factors into simpler segments would have increased the number of factors, which could have had the effect of respondents refusing to participate in the research.

IV. RESULT ANALYSIS

It needs to be stressed that, prior to sending the e-mails, we performed interviews with one of the 73 contact persons that we had selected, as we have had a consultant relationship with a better part of these companies for more than 20 years, whereas we have been performing our third study for the past 7 years. After first contacts were established, it became evident that some of the companies were not likely to participate in the research. A total of 46 companies completed the questionnaire in any form. From this number, 19 questionnaires were completed by CEOs, technical directors or production directors, 22 by ERP implementation project leaders and 5 by advanced users. For 23 companies, the current ERP solution is the second implementation, but not of the same solution. 27 companies did not fill in the questionnaire due to the following reasons:

- part of them did not participate in the selection of the ERP solution (significant fluctuations of employees), and therefore they found it difficult to contemplate many of the factors,
- those that were witness to the selection and introduction of the system were not in the selection process, as the solution was obtained through the logic of parent-subsidiary,
- a part of them considered that contemplating such a great number of factors did not make any sense, i.e. that it was not necessary, and
- one smaller part of the respondents simply did not fill in the questionnaire or give any feedback.

Therefore, pursuant to these 4 assertions and taking into consideration the fact that the questionnaire was filled in by 46 companies, i.e. 63% of them, it can be concluded that Croatian production companies have a perception of the importance of the factors that influence the selection of a suitable ERP solution. This is one aspect of the response to the first research question.

At the level of all 24 factors that influence the selection of an optimum ERP solution, in 21.38% of cases it was stated that there was no influence. 41.39% of factors were assessed to have an influence, while 37.23% of them were assessed to have a strong influence (Figure 2). To sum up, the responses has an influence (2) and has a strong influence (3) add up to 78.62 %, which indicates that the respondents have an awareness of the great importance of the proposed factors, which is the second aspect of the response to the first research question.

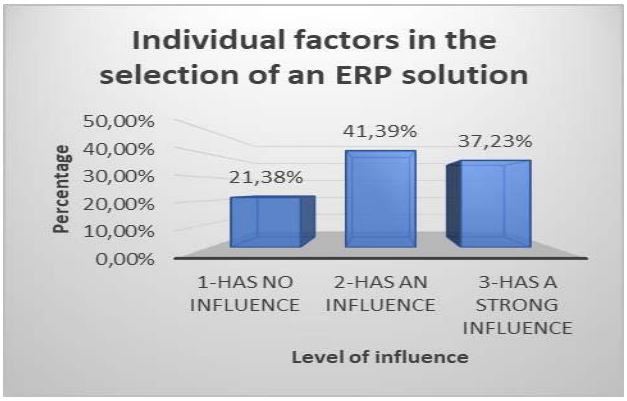


Figure 2. Individual factors in the selection of an ERP solution (1-has no influence, 2-has an influence, 3- has a strong influence)

The respondents assessed the following as the least important factors (1- has no influence) in the selection of an ERP solution:

- latent ERP implementation expenses,
- external resource expenses, and
- hardware and network expenses.

The factors with the strongest influence out of all 24 factors (3 – has a strong influence) were:

- appropriate functionalities according to production type,
- reliability of the system, and
- technological platform (cloud or server).

If the second (2- has an influence) and third (3- has a strong influence) option are added together, the stated three factors have also been assessed as the most influential.

This provides a response to the second research question.

The stated results, in and of themselves, do not provide sufficient arguments for the selection of an ERP solution. First of all, all who are part of the decision-making chain for selecting an ERP solution are not likely to have sufficient knowledge of every single factor. Due to this reason, selection according to the stated groups is important, as it is indisputable that certain people will be able to give a more confident and relevant assessment of a group and the factors within it, than of every factor separately.

As for factors related to implementation expenses, the following had the most powerful influence:

- licence expenses,
- external consultant expenses, and
- software maintenance expenses.

The fact that expense-related factors are of great importance is evident even to a person with average familiarity with the implementation of an ERP process. The distribution of influences concerning factors related to implementation expenses is shown in Figure 3.

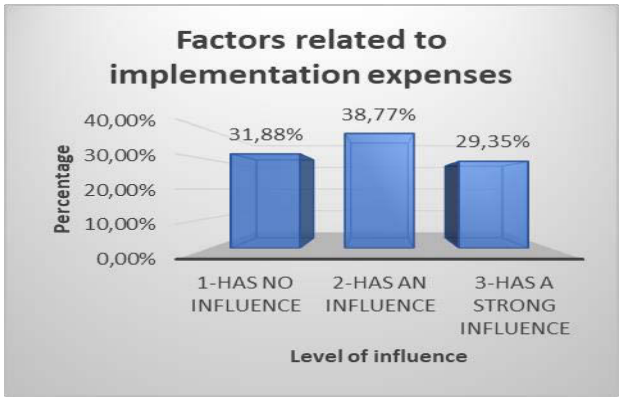


Figure 3. Factors related to implementation expenses (1-has no influence, 2-has an influence, 3- has a strong influence)

Within the group Analysis of internal company factors, the following factors had the strongest influence:

- involvement of key persons in the ERP implementation project, and
- potential budget (readiness to adapt).

The distribution of influences at the level of the group Analysis of internal company factors is shown in Figure 4.

Within the group Systematic software analysis factors, the following factors had the strongest influence:

- appropriate functionalities according to production type,
- reliability of the system, and
- technological platform (cloud or server).

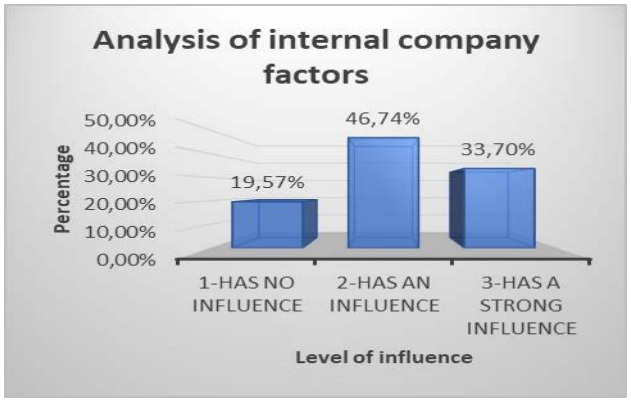


Figure 4. Analysis of internal company factors (1-has no influence, 2-has an influence, 3- has a strong influence)

The factor related to the possibility of being upgraded by the user was given the least importance in this group.

The distribution of influences, in a summary and percentage sense, in the group Systematic software analysis factors, is shown in Figure 5.

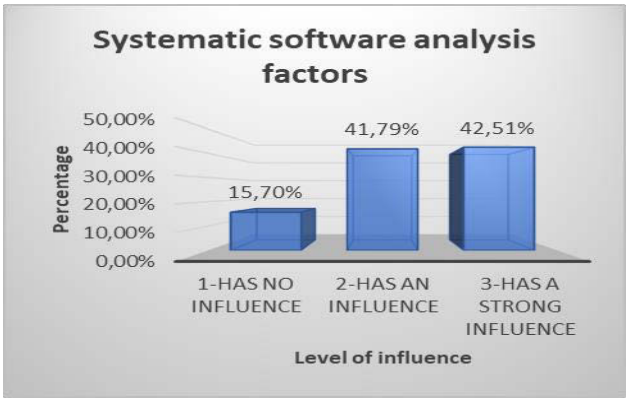


Figure 5. Systematic software analysis factors (1-has no influence, 2-has an influence, 3- has a strong influence)

Within the group Factors related to suppliers, the following factors had the strongest influence:

- potential of the supplier in the ERP implementation process,
- reputation of the supplier on the local ERP market, and
- possibility of developing and executing additional components.

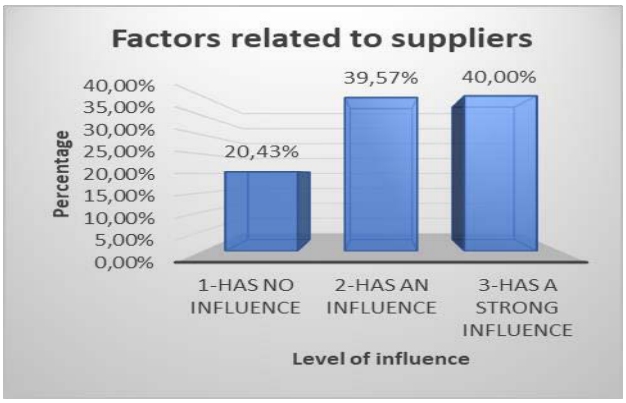


Figure 6. Factors related to suppliers (1-has no influence, 2-has an influence, 3- has a strong influence)

The factor related to references according to production type was given the least importance in this group.

The distribution of influences, in a summary and percentage sense, in the group Factors related to suppliers, is shown in Figure 6.

On the basis of the processing of results and the data contained in Figures 3-6, several things are immediately evident.

First of all, for the factor of Systematic software analysis factors, 2- has an influence and 3- has a strong influence add up to 84.3%. 15.70% of respondents opted for answer 1 – has no influence. This factor group has the strongest influence of all groups.

Secondly, at 31.88 %, the group of Factors related to implementation expenses received the highest percentage of answer 1- has no influence.

Finally, the factors in the groups Analysis of internal company factors and Factors related to suppliers obtained similar results when it comes to answer 1 – has no influence. The difference is only in answer 2- has an influence and 3- has a strong influence, in which number 3- has a stronger influence is higher in the group Factors related to suppliers, therefore, suppliers are considered to be more critical than internal company factors.

V. CONCLUSION

Selecting an unsuitable ERP solution paves the path towards unsuccessful implementation, excessive expenses related to the entire implementation process, missed deadlines and unsatisfied users. The situation becomes even more grave when we take into consideration the fact that, today, a new ERP solution is created every 7-10 years.

The idea behind this research was to establish the importance of each of the factors in the selection of an ERP solution, as well as the importance of groups of factors, as it is indisputable that, in a systematic approach to the selection of an ERP system supported by artificial intelligence, the importance of each of these groups needs to be taken into consideration.

Any future in-depth analysis of the results obtained with this research should take two things into consideration.

First of all, it would be good to focus the selection of a solution towards artificial intelligence. Support for a selection through artificial intelligence should, without any doubt, also be based on a much larger sample of respondents.

Secondly, the results according to groups that have been presented in this work indicate that future research should attribute a certain coefficient for correcting the importance of these groups. Only in this manner can a software-based selection process approach reality.

LITERATURE

[1] https://www.gartner.com/en/information-technology/glossary/enterprise-resource-planning-erp (accessed on 19/08/2019)

[2] L. Herbean, D. Fotache, ERP III: Promise of a new generation, Proceedings of the 2014 International Conference, on Informatics and Economy, Education, Research & Business Technologies, ASA Bucharest (ISSN: 2247-1480)

[3] APICS Dictionary, 16th Edition 2018.

[4] http://www.cepor.hr/wp-content/uploads/2015/03/SME-report-2018-HR.pdf (accessed on 20/12/2018)

[5] https://ec.europa.eu/commission/index_en (accessed on 20.12.2018)

[6] V. Kumar, B. Maheshwari, U. Kumar, ERP systems implementation: Best practices in Canadian government organizations, Government Information Quarterly, 19, pp. 147-172, 2002.

[7] M. J. Esteves, A. J. Pastor, An ERP Life-cycle-based Research Agenda, First International Workshop in Enterprise Management and Resource Planning: methods, Tools and Architectures – EMRPS'99, Venice, Italy, 1999.

[8] A. M. Razi, M. M. Hossain, ERP Implementation: Examining Interdependencies among Pre-Implementation, Implementation and Post Implementation Phases, Journal of Integrated Enterprise Systems, Vol. 1, No 1, pp. 61-66, 2012.

[9] L. Filbert, Understanding the Phases of a Manufacturing ERP Implementation, 2012, (downloaded form http://info.iqms.com/IQMS-Manufacturing-ERPexpertise/bid/83949/Understanding-the-Phases-of-a-Manufacturing-ERPImplementation, accessed on 23/10/2013)

[10] P. C. Gupta, A. Kumar, Evaluation of Enterprise Resource Planning Life Cycle on the scale of ERP Implementation failure, International Journal of Research in IT, Management and Engineering, Vol. 2, Iss. 8,pp. 61-68, 2012.

[11] http://ww35.indiastudychanel.com/resources/145958-ERPimplementation-phases-enterprise (accessed on 23/10/2013)

[12] https://www.indiastudychannel.com (accessed on 23/10/2013)

[13] B. Maheshwari, B, V. Kumar, U. Kumar, Delineating the ERP institutionalization process: go live to effectiveness, Business Process Management Journal, Vol. 16, No. 4, pp. 744-771, 2010.

[14] C. C. C. Chen, H. C. C. Law, C. S. Yang, Managing ERP Implementation Failure: A Project Management Perspective, IEEE Transactions on Engineering Management, Vol. 56, No. 1, February 2009., pp. 157-170, 2009.

[15] M. L. Markus, C. Tanis, The Enterprise system experience – From adoption to success, Framing the Domains of IT Research: Glimpsing the Future Through the Past, Chapter X, Pinnaflex Education Resource, pp. 173-207, 2000.

[16] https://noblue.co.uk/news-updates/b2b/5-tips-for-selecting-the-right-erp-software-whats-the-best-fit-for-your-business/ (accessed on 22/12/2017)

[17] https://www.managementstudyguide.com/evaluation-selection-erp-packages.htm (accessed on 20/12/2017)

[18] https://www.erpfocus.com/erp-selection-checklist.html (accessed on 19/12/2017)

[19] https://www.180systems.com/article/system-selection-methodology/ (accessed on 10/12/2017)

[20] https://www.erpfocus.com/ten-essential-erp-selection-criteria-2640.html

[21] B. Molnar, G. Szabo, A. Benczur, Selection Process of ERP Systems, Business System Research Journal, Vol. 4, No 1, pp. 36-48, Zagreb, March,2013.

[22] L. Pitic, S. Popescu, D. Pitic, Roadmap for ERP evaluation and selection, Procedia Economics and Finance 15, pp. 1374-1382, 2014.

[23] M. Haddara, ERP Selection: The SMART Way, Procedia Technology 16, pp. 394-403, 2014.

[24] M. Haddara, ERP system selection in multinational enterprises: a practical guide, International Journal of Information Systems and Project Management, Vol. 6, No. 1, pp. 43-57, 2018.

Implementacija ERP sustava iz perspektive konzultanata

Izidora Kostanjevec i Ruben Picek
Fakultet organizacije i informatike, Sveučilište u Zagrebu, Varaždin, Hrvatska
{izidora.kostanjevec, ruben.picek}@foi.hr

***Sažetak* - Suvremeno poslovanje danas, nezamislivo je bez informacijsko komunikacijske tehnologije, a ERP sustavi predstavljaju samu srž uspjeha svakog poslovanja. Uspješna implementacija ERP sustava ključ je transformacije poslovanja poduzeća. Postoje mnoga istraživanja koja govore o različitim aspektima uvođenja ERP sustava, ali vrlo malo se istražuje kako sami konzultanti ERP sustava vide aktivnosti koje provode u mnogobrojnim implementacijama. Stoga je cilj ovog rada analizirati uvođenje ERP sustava u organizaciju, ali iz perspektive konzultanata koja obuhvaća pristupe i tijek implementacije, probleme i rizike s kojim se susreću te post implementacijske aktivnosti. Kao ishodišna točka analizirana su istraživanja te tematike pronađena u znanstvenoj literaturi, a odnose se na stanje u Hrvatskoj, ali i šire. Rad prikazuje rezultate istraživanja provedenog u hrvatskim poduzećima koja je bave uvođenjem ERP sustava, a iz kojih se može zaključiti da su poduzeća koja obavljaju djelatnosti konzultantskih usluga i uvođenja ERP rješenja u Hrvatskoj uglavnom mala ili mikro poduzeća za čijim uslugama postoji velika potražnja iz različitih grana industrije. Doprinos ovog istraživanja usmjeren je na razumijevanje kritičnih čimbenika uspjeha uvođenja ERP sustava iz perspektive konzultanata ERP sustava.**

Ključne riječi - uvođenje ERP sustava, implementacija, konzultanti ERP sustava, transformacija poslovanja, poslovni procesi

I. UVOD

ERP sustavi omogućuju pojednostavljenje i promjenu načina poslovanja, a ujedno su jedan od najučinkovitijih alata za reinženjering poslovnih procesa. Osim radikalnih promjena koje istodobno mogu predstavljati prednost i nedostatak ERP sustava, najistaknutije prednosti uvođenja ERP sustava su poboljšanje izvođenja poslovnih procesa, smanjenje troškova, povećanje produktivnost zaposlenika, standardizacija i rast poslovanja. Globalno gledano, integrirani ERP sustavi povećavaju učinkovitost i djelotvornost svih operacija [1]. S obzirom na niz prednosti koje pružaju ERP sustavi nije iznenađujuće što je ERP tržište i dalje jedno od najbrže rastućih segmenata u IT industriji [2]. Nedavnim svjetskim istraživanjem [3] utvrđeno je da su najčešći razlozi uvođenja ERP rješenja usmjereni na poboljšanje uspješnosti poslovanja, pozicioniranje organizacije za rast te smanjenje obrtnog kapitala i pružanje boljih usluga kupcima. Stanje u Hrvatskoj zasigurno je drugačije, stoga je cilj ovog rada istražiti pogled konzultantskih kuća na uvođenje ERP sustava te definirati i analizirati aspekte njihovog poimanja na uvođenje ERP sustava.

Ovaj rad sastoji se od 4 cjeline. U prvom odjeljku predstavljena su uvodna razmatranja koja upućuju na svrhu i ciljeve ovog rada, zatim slijedi povezivanje rada sa srodnim istraživanjima i predstavljanje korištenog postupka istraživanja. U odjeljku tri prikazani su rezultati provedenog istraživanja, a na samom kraju izveden je zaključak.

II. PREGLED ZNANSTVENE LITERATURE

ERP sustavi javljaju se kao odgovor na sve učestalije zahtjeve i promjene na tržištu, a koje su rezultat jačanja konkurencije i globalizacije općenito. Organizacije su prisiljene prilagoditi svoje sustave i uvesti ili nadograditi ERP sustave kako bi se prilagodili okruženju [4]. D. Žabjek i suradnici smatraju da su ERP sustavi postali imperativ za poduzeća koja nastoje steći konkurentsku prednost [5].

U Hrvatskoj je provedeno istraživanje M. Nikitović i V. Strahonja [6] čiji je predmet bila implementacija ERP rješenja u 120 hrvatskih tvrtki koristeći pomoć deset iskusnih konzultanata. Istraživanje se temeljilo na otkrivanju kritičnosti pojedinog čimbenika, a početni broj kritičnih čimbenika bio je 81 koji se kasnije u procesu filtriranja smanjio na 32. U istraživanju sudjelovali su direktori tvrtki, napredni korisnici te voditelji projekata ispitanih tvrtki. Prema mišljenju direktora, najvažniji kritični čimbenici su analiza i motiviranost za potrebom ERP sustava, uključenost i participacija korisnika, potpuno razumijevanje ključnih problema vezanih uz ERP implementaciju, organizacijska spremnost, pažljiv izbor ERP rješenja, podrška najvišeg vodstva i drugo. Voditelji projekta smatraju da su najvažniji kritični čimbenici podrška najvišeg vodstva, analiza i motiviranost za potrebom ERP sustava, pristup i metodike implementacije,

očekivanja poslovodstva, pažljiv izbor ERP rješenja i partnerski odnos s dobavljačem. Posljednju skupnu čine napredni korisnici, a prema njihovom mišljenju kritični čimbenici su analiza i motiviranost za potrebom ERP sustava, jasna vizija i poslovni ciljevi, pažljiv izbor ERP rješenja, podrška najvišeg vodstva, ERP se tretira kao program, a ne projekt i komunikacija unutar organizacije. Na temelju prosjeka odgovora ispitanika ključan faktor je analiza i motiviranost za potrebom ERP sustava, a zaključak kojeg su izveli autori iz istraživanja je postojanje nesrazmjera između angažiranih novčanih sredstava i stručno-znanstvenih istraživanja.

K. Šimunović i suradnici [7] proveli su istraživanje o ulozi ERP sustava u poslovnom procesu. Istraživanje pomoću ankete provedeno je u 30 malih do srednjih poduzeća u Hrvatskoj koje su uvele ERP rješenja ili ga nastoje uvesti. Od 30 malih do srednjih ispitanih tvrtki, 90% ispitanih koristi ERP sustav. Većina poduzeća točnije 59% koriste ERP sustave razvijene u Hrvatskoj, a 41% poduzeća koristi ERP sustav stranih dobavljača od čega je 37% sustava prevedeno na hrvatski jezik. Od ukupnog broja poduzeća koja su uvela ERP sustav, 96% poduzeća se izjasnilo da su prošli kroz edukaciju za rad u ERP sustavu. 54% ispitanih tvrde da je edukacija izvedena savršeno dok 46% korisnika nije zadovoljno načinom provedene edukacije. Svega 24% poduzeća bila su nezadovoljna uvođenjem ERP sustava zbog čega su pružali otpor. Bitne prednosti koje su ispitanici uočili prilikom korištenja ERP sustava su olakšano praćenje troškova, ušteda vremena i generiranje različitih analiza i izvještaja. M. Heredia-Calzado i A. Durendez [8] svoje istraživanje usmjerili su također na mala i srednja poduzeća koja koriste ERP sustave. Autori su postavili hipoteze koje analiziraju na podacima prikupljenim od 137 španjolskih malih i srednjih poduzeća primjenom metode strukturalnih jednadžbi. Glavni cilj bio je objasniti važnost upravljanja znanjem i profesionalizaciju kao način za postizanje konkurentske prednosti u poslovanju malih i srednjih poduzeća. Empirijski rezultati potvrđuju tri ključna doprinosa u istraživačkoj literaturu. Prvi od njih upućuje na poboljšanje procesa upravljanja kroz promicanje potrebe za korištenjem ERP sustava u malim i srednjim poduzećima. Drugi doprinos ovog istraživanja usmjeren je na predlaganje ERP sustava kao važnog posrednika između konkurentskih prednosti i upravljanja znanjem i profesionalizacije. Autori smatraju da kroz uspostavu strategija obogaćivanja znanja i vještina mogu postići veću motivaciju zaposlenika i minimalizirati otpornost na promjene. Posljednji doprinos istraživanja ukazuje na nedostatke u korištenju ERP sustava i predlaže načine njihova savladavanja, a to su prilagodba modula koji bi olakšali komunikaciju korisnika sa sustavom, veći pristup informacija i kvaliteta podataka, poboljšanje koordinacije unutar odjeljenja i profesionalizacija procesa. Za razliku od prethodno spomenutih istraživanja, S. Malinić i M. Todorović [9] proveli su istraživanje na devet poduzeća pri čemu su isključena mala i srednja poduzeća. Temeljem provedenog istraživanja, dobiveni su rezultati koji upućuju da je u početnim fazama procesa implementacije ERP sustava veća pažnja usmjerena na uspostavljanje osnovnih modula, a složenost rješenja može ometati razvoj računovodstva i ostalih odjela. Također, za potpunu implementaciju ERP sustava potrebno je dugo vremena što može utjecati na uspješnost upravljanja.

Implementaciju ERP sustava istraživali su i C.K.M. Lee i suradnici [10] s ciljem dobivanja odgovora na pitanje: Kako uspješna implementacija ERP sustava može pomoći transformiranju procesa u poduzećima? Došli su do zaključka da ERP sustav poboljšava vidljivost podataka i pravovremenu isporuku, ali je manje učinkovit u poboljšanju produktivnosti zaposlenika. Ističu kako je ključna faza implementacije ERP sustava upravo faza analize i planiranja u kojoj se razmatraju aspekti upravljanja implementacijom. Ova faza predstavlja ključnu fazu koja je zbog loše taktike i strategije često rezultat neuspjeha implementacije. Neuspjeh implementacije prouzrokuju i nekompetentni konzultanti i korisnici koji pružaju otpor na promjene, stoga je važno razvijati funkcionalne, tehničke i međuljudske vještine [11].

Za analizu primjene ERP sustava korišteno je i dvanaestomjesečno istraživanje Panorama Consulting Solutions za razdoblje od ožujka 2017. do veljače 2018. godine u kojem je sudjelovalo 237 ispitanika [12]. Najčešća industrija koja je uvodila ERP sustave bila je proizvodnja. U odnosu na prethodnu godinu proizvodnja preuzima vodeću poziciju (43%) u uvođenju ERP rješenja. Ostale grane industrije koje uvode ERP sustave su maloprodaja i distribucija te financije i osiguranje s udjelom od 11%, a zatim slijede informacijska tehnologija, profesionalne usluge, neprofitne organizacije, obrazovanje, zdravstvo i telekomunikacije. Najpoznatiji izabrani ERP dobavljači prema provedenom istraživanju su Oracle, SAP, Microsoft, Sage, Epicor, Infor, Netsuite, IFS, IQMS i brojni drugi. Na tržištu ne postoji apsolutna dominacija pojedinog dobavljača, a dva istaknutija ERP dobavljača su Oracle kojeg koristi 16% ispitanika te SAP s 15% ispitanika. Prema istraživanju Hasan N. i suradnika [13] najzastupljeniji ERP dobavljači također su Oracle (62%) i SAP (30%). Ispitanici istraživanja Panorama Consulting Solutions smatraju da uvođenje

ERP sustava omogućuje rast tvrtke, smanjuje obrtni kapital, omogućuje pružanje bolje usluge klijetima, olakšava posao, poboljšava integraciju sustava na više lokacija. B. Erkayman [14] proučavao je uspostavljanje ERP sustava u poduzeće, a ključni čimbenici koje je izdvojio kao presudne za uspješnost projekta su tip, veličina i opseg sustava. Autor je istaknuo važnost provedbe projekta putem sljedećih faza: „inicijacija, planiranje, analiza, realizacija, tranzicija i operacija. Također, smatra da ERP sustav nije moguće implementirati odjednom te da je potrebno svima objašnjavati što se događa i zbog čega se događa čime se smanjuje otpor na promjene. Gazaleh M., Abdallah S. i Khan M. [15] istraživali su važnost unutarnjih snaga nakon implementacije ERP sustava. Ova studiji empirijski potvrđuje da organizacijska spremnost, interaktivni faktori komunikacije te upravljanje znanjem i nakon same implementacije ERP sustava pozitivnu utječu na održivost ERP sustava.

III. ISTRAŽIVANJE

A. *Predmet istraživanja*

U svrhu prikupljanja informacija i podataka od konzultantskih kuća u Hrvatskoj izrađena je anketa čiji predmet istraživanja predstavlja pogled konzultanata na uvođenje ERP sustava.

B. *Opis istraživanja*

Cilj ovog istraživanja je utvrditi poziciju domaćih u odnosu na strana ERP rješenja, objasniti pristup i tijek implementacije, predstaviti probleme i rizike s kojima se susreću konzultanti i iznijeti stavove konzultanta o uspješnosti provedene implementacije.

Anketna pitanja odnosila su se na sljedeće podatke:

- Znanje, vještine, zanimanje i zvanje konzultanata koji rade na implementaciji
- Vlastito razvijena ERP rješenja ili korištenje rješenja nekog poznatog dobavljača
- Broj provedenih implementacija, veličina organizacija i grane industrije koje implementiraju ERP rješenja
- Vrijeme trajanja implementacije te potrebni resursi
- Rizici i problem prilikom implementacije
- Metodike koje koriste konzultantske kuće prilikom implementacije i koliko slijede faze korištenih metodika.
- Zahtjevi klijenata nakon implementacije
- Mogućnost nadogradnje ERP sustava.

C. *Analiza rezultata*

Istraživanje je provedeno u 15 poduzeća u Hrvatskoj koja se razlikuju po veličini s obzirom na kategoriju broja zaposlenih od mikro do velikih poduzeća s rasponom od 7 do 600 zaposlenih temeljem čijih je tvrdnji moguće analizirati sliku hrvatskog tržišta ERP rješenja.

Profil tvrtke i zaposlenika

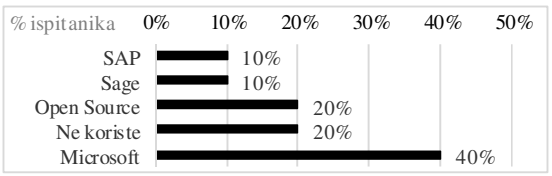
Od ukupno 15 anketiranih tvrtki, najčešći odgovori prikupljeni su od *vlasnika* što u odnosu na ukupan broj ispitanika iznosi 53,4%. Iz prikupljenih informacija o profilu zaposlenika saznalo se da 60% ispitanika ima zanimanje *informatičara* dok je 26,7% ispitanika *ekonomskog* zanimanja. Znanja i vještine kojima raspolažu ispitanici u većini slučajeva odnose se na znanja i vještine stečena u radnom odnosu, a posebno ističu:

- Programiranje, upravljanje, analiza, vođenje poslovanja, rad na IT projektima, objektni sustavi
- Poznavanje humanističkih znanosti, stranih jezika te informatike
- Rad u SQL-u (eng. Structured Query Language), Microsoft Office-u, poslovna integracija
- Računarstvo i arhitektura informacijskih sustava
- Poznavanje poslovnih procesa i financijskih izvještaja, logično rasuđivanje i odlučivanje
- Usvajanje korisničkih zahtjeva te pretakanje istih u projektnu dokumentaciju potrebnu da se projekt što kvalitetnije odradi

Na temelju prikupljenih podataka, poduzeća u Hrvatskoj koja obavljaju djelatnost konzultantskih usluga i nude ERP rješenja po kategorizaciji prema broju zaposlenih spadaju u mikro i mala poduzeća. Od ukupnog broja anketiranih, 4 poduzeća broje do 10 zaposlenika i predstavljaju mikro poduzeća, a 7 poduzeća broje od 18 do 32 zaposlenika zbog čega su kategorizirana kao mala poduzeća. Dva poduzeća zapošljavaju više od 50 zaposlenika, a manje od 100 dok svega 2 poduzeća imaju zaposleno više od 400 zaposlenika, točnije 440 i 600 zaposlenih u poduzeću.

Ponuda ERP rješenja

Ispitane konzultantske kuće razvijaju vlastita rješenja, ali koriste i gotova rješenja dobavljača. Omjer razvijenih vlastitih ERP rješenja i korištenja rješenja stranih dobavljača je 60% - 40% što bi značilo da je 9 tvrtki razvilo vlastito rješenje, a 6 koristi ERP rješenje SAP, Microsoft, Sage dobavljača. Utvrđeno je da postoji veći broj Hrvatskih konzultantskih kuća koje su razvile vlastito ERP rješenje i zbog čega dovode domaću ponudu ERP rješenja u povoljniji položaj u odnosu na strana ERP rješenja na Hrvatskom tržištu.



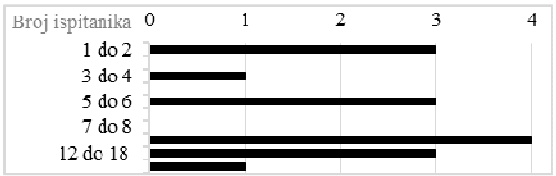
Slika 1. Korištenje ERP rješenja stranog dobavljača

Od 10 ispitanika koji su odgovorili na anketno pitanje vezano uz korištenje ERP rješenja nekog od stranih dobavljača, 40% ih se izjasnilo da koristi ERP rješenje Microsoft-a, 10% SAP i Sage, 20% koristi Open source rješenja, a 20% ne koristi rješenja stranih dobavljača. Najčešće industrije koje koriste rješenja ispitanika su Proizvodnja, Maloprodaja i distribucija, Građevinarstvo, Hotelijerstvo i ugostiteljstvo, Veleprodaja i logistika. U većini slučajeva ERP rješenja pokrivaju potrebe svih industrija. Ispitanici su na pitanje o najčešćoj veličine tvrtke koje koriste vaše rješenje mogli odabrati između više mogućnosti stoga su se male tvrtke 7 puta pojavljivale kao najčešća veličina tvrtke, srednje tvrtke 11, a velike 6 puta. Prema prikupljenim podacima, najčešći korisnici ERP rješenja ispitanika su srednje tvrtke.

Implementacija ERP sustava

Provedenim istraživanjem dobiven je uvid u broj provedenih implementacija na godišnjoj razini. Osam ispitanika odnosno 53,3% izjasnilo se kako godišnje provedu do 5 implementacija, 26,7% godišnje provedu do 10 implementacija, a jedna tvrtka zajedno s partnerskom mrežom godišnje provede više od 50 implementacija.

Prosječno vrijeme potrebno za implementaciju razlikuje se od tvrtke do tvrtke. Ispitanici su mogli odabrati između vremenskih intervala izraženih u mjernoj jedinici mjesec. Zadani intervali prikazani su na sljedećem grafikonu, a ispitanici su se formirali u približno jednake grupe s obzirom na potrebno prosječno vrijeme za implementaciju.



Slika 2. Vrijeme trajanja implementacije ERP rješenja

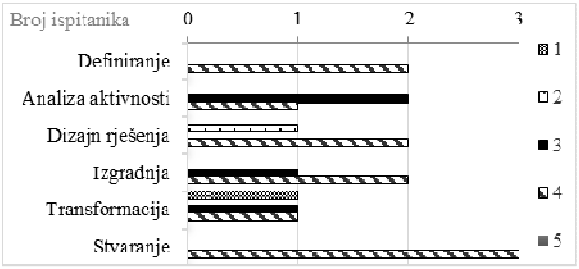
Od ukupnog broja ispitanika, 26,7% izjasnilo se kako je potrebno vrijeme za implementaciju od 8 do 12 mjeseci, po 20% ispitanika točnije po 3 ispitanika navode 3 različita vremenska intervala, a to su intervali od 1 do 2 mjeseca, 5 do 6 mjeseci i 13 do 18 mjeseci. Jedan od ispitanika naveo je vlastiti interval s prosječnim vremenom potrebnim za implementaciju u trajanju od 2 do 6 mjeseci.

Metodike implementacije koje se koriste su ASAP, Sure Step, vlastito razvijena metodika i kombinacija metodika. 40% ispitanika koristi kombinaciju metodika, 33,3% razvilo je vlastitu metodiku, 20% koristi ASAP metodiku, a 6,7% koristi Sure Step metodiku. Analizirajući odgovore 10 ispitanika koji su naveli faze vlastite metodike ili kombinacije metodika moguće je zaključiti da sva poduzeća započinju s fazom analizom stanja odnosno snimkom stanja kod klijenta. Neke od faza koje su naveli ispitanici su sljedeće:

- Analiza stanja, dokumentacija, dizajn, razvoj, implementacija, edukacija, UAT, Go-live i postGo-live support)
- Snimka stanja (organizacija, procesi, dokumenti, izvještavanje), Postavljanje BPM strukture (priprema sustava, izrada prilagodbi), Import podataka iz prethodnog sustava, obuka krajnjih korisnika, Kick off produkcija
- Analiza postojećeg rješenja, kritični poslovni procesi u postojećem sustavu, fit/gap analiza zahtjeva, razvoj dorada, setup, treninzi i testiranje, go-live, postprodukcijska podrška
- Snimka stanja, razrada projekta, usuglašavanje rješenja, razvoj, interni test, eksterni test, implementacija.

Ispitanici koji koriste metodike AIM, Sure Step ili ASAP ocijenili su koliko slijede faze implementacije pojedine metodike. Koliko slijede faze pojedine metodike ocijenili su ocjenom od 1 do 5 odnosno od odgovora slijedimo minimalno, dovoljno, dobro, vrlo dobro i potpuno.

Na sljedećem grafikonu prikazane su faze implementacije AIM metodike.

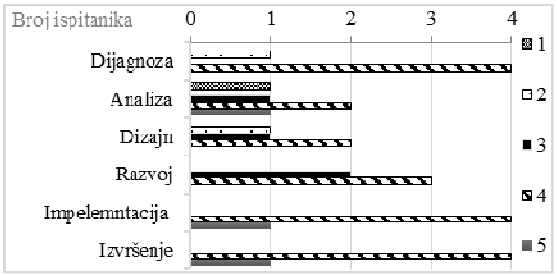


Slika 3. Faze implementacije AIM metodike

Prvu fazu *Definiranje* ocijenila su dva ispitanika ocjenom 4 to jest, dvije konzultantske kuće vrlo dobro slijede prvu fazu AIM metodike. Drugu fazu predstavlja *Analiza aktivnosti*, a slijede ju tri ispitanika,

od toga jedan ispitanik ocjenom vrlo dobro, a dva ispitanika ocjenom dobro. Minimalnom ocjenom jedan ispitanik ocijenio je fazu *Transformacije*. Prosječna ocjena praćenja faza implementacije AIM metodike je vrlo dobar. Tri poduzeća koja koriste faze AIM metodike gotovo svaku fazu ocijenili su ocjenom 4 na temelju čega se može zaključiti da je poduzeća koriste AIM metodiku, ali da ju ne prate u potpunosti u svim pogledima.

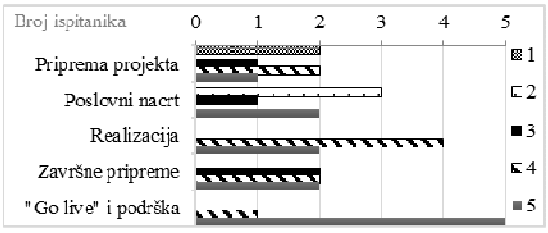
Na sljedećem grafikonu prikazane su faze Sure Step metodike.



Slika 4. Faze implementacije Sure Step metodike

U odnosu na faze AIM metodike, faze Sure Step metodike primjenjivanije su i njihovo praćenje ocijenjeno je većim ocjenama. Faze metodike Sure Step su općenite zbog čega je koristi veći broj konzultanta u odnosu na AIM metodiku. Kod Sure Step metodike postoji po jedan ispitanik u tri faze koji praćenje faza metodike Sure Step ocjenjuje ocjenom potpuno, to su faze *Analize*, *Implementacije rješenja* i *Izvršenja*.

Posljednja metodika, ujedno i najzastupljenija među ispitanicima je ASAP. To je metodika čije se faze prema rezultatima ankete najviše slijede.



Slika 5. Faze implementacije ASAP metodike

Posljednju fazu ASAP metodike *„Go live“ i podrška* gotovo u potpunosti slijedi 6 tvrtki. Od ukupno 6 ispitanika, 5 ih je odgovorilo da u potpunosti slijede fazu *„Go live“ i podrška,* a jedan se ispitanik izjasnio da tu fazu slijede vrlo dobro. U svakoj fazi postoji najmanje jedan ispitanik koji fazu slijedi u potpunosti zbog čega su faze metodike najviše slijeđene.

Rizici s kojima se susreću konzultantske tvrtke prilikom implementacije ERP sustava klijentima su: otpor korisnika, nelogični procesi, nepodudarnost između ERP sustava i organizacijske kulture, koncentracija odgovornosti, ljudi i njihovo viđenje što bi trebalo, 13,4 % ispitanika navodi da se susreće sa svim navedenim rizicima, a najčešći rizik je otpor korisnika. Organizacijska nespremnost klijenta, nerealna očekivanja i nerealni zahtjevi korisnika, nejasna vizija i poslovni ciljevi klijenta, nemotiviranost zaposlenika za promjene, loša komunikacija unutar organizacije i prilagodba korisnika na novi sustav neki su od problema s kojima se susreću konzultantske tvrtke na hrvatskom tržištu. Ispitanici ističu kako problem predstavlja sve navedeno, a kao najčešći problem navode organizacijsku nespremnost klijenta za promjene.

Post implementacijske aktivnosti

Posljednji dio ankete predstavljale su post implementacijske aktivnosti koje poduzeća provode. Svih 15 ispitanika potvrdilo je da njihove tvrtke nakon završenog projekta implementacije ERP sustava pružaju konzultantske usluge korisnicima i da provode post implementacijske prilagodbe. Neki od zahtjeva nakon završenog projekta implementacije su nadogradnje sustava u svrhu automatizacije poslovnih procesa, poslovna inteligencija, dodatne vertikale, optimizacija procesa i radnji, smanjenje broja operacija i dokumentacije, prilagođeni izvještaji, isporuke specifičnosti kojih se klijent ne želi odreći, dorade i usklađivanje procesa s postojećim sustavom, integracije na nove sustave, promjena zakonskih okvira i podrška, rekonfiguracija sustava, dodatni izvještaji i edukacija novozaposlenih.

Sva poduzeća pružaju post implementacijske prilagodbe, ali ne omogućuju sva neovisnu nadogradnju prema potrebama i zahtjevima poslovanja. 93,3% ispitanika izjasnilo se kako njihova konzultantska tvrtka daje mogućnost neovisne nadogradnje prema potrebama i zahtjevima klijenta, a 6,7% ispitanika točnije 1 ispitanik naveo je da njihovo poduzeće ne omogućuje nadogradnje jer je sustav zatvoren.

Mogućnosti neovisne nadogradnje koje nude konzultantske kuće su:

- Razvoj i implementacija dodatnih funkcionalnosti, BI, maloprodaja, Hospitality i Forecorut
- Modularni pristup, parametarsko postavljanje Inteface-a između core sustava i nadogradnji
- Set Web servisa kojima se postiže integracija s raznim informacijskim rješenjima
- Specijalizirana rješenja za diskretnu proizvodnju, upravljanje životom i automatskom konfiguracijom proizvoda

Za kraj, ispitanici su morali ocijeniti koliko je uvedeni ERP sustav u skladu s poslovnim procesima poduzeća. Ponuđen raspon ocjena bio je od 1 do 10. Najveći broj ispitanika ocijenio je sklad implementiranog sustava s poslovnim procesima ocjenom 8. Tri osobe usklađenost implementiranog sustava ocijenile su ocjenom 9, dvije osobe ocjenom 7. Jedan ispitanik smatra da je njihov implementiran sustav u 100%-tnom skladu s poslovnim procesima klijenta stoga je usklađenost između procesa ocijenio ocjenom 10.

IV. ZAKLJUČAK

Implementacija ERP sustava iz perspektive korisnika/klijenta tema je o kojoj se mnogo piše i govori, ali istovremeno nedostaju podaci u kojima se promatra proces implementacije iz perspektive konzultanata. Provedeno istraživanje daje uvid upravo u te podatke, a temeljem istog, može se zaključiti da je implementacija ERP rješenja složen proces, u kojem svaka faza iziskuje kvalitetnu provedbu plana, a najvažniju ulogu imaju ljudi koji svojim kompetencijama omogućuju jednostavnije i brže odvijanje faza implementacije. ERP sustavi sve su primjenjivaniji u poslovanju, a razmjerno tome raste i broj konzultantskih kuća koje nude ERP pakete. U Hrvatskoj su to mikro i mala poduzeća koja na godišnjoj razini provedu do 5 implementacija, a u većini slučajeva razvijaju vlastita ERP rješenja dok nekolicina koristi ERP rješenja stranih dobavljača. Najzastupljeniji strani dobavljač je Microsoft, a slijede ga SAP i Sage. Prilikom same implementacije, konzultanti se susreću s brojnim rizicima kao što su nepodudarnost između ERP sustava i organizacijske kulture, otpor korisnika, nelogični procesi, ali i problemima poput nespremnosti klijenta, nerealnih očekivanja i zahtjeva korisnika, nejasne vizije i poslovnih ciljeva klijenta. Usprkos brojnim rizicima i problemima, konzultanti ERP sustave implementiraju u tvrtke iz različitih grana industrije kao što su Proizvodnja, Maloprodaja i distribucija, Građevinarstvo, Hotelijerstvo i ugostiteljstvo, Veleprodaja i logistika. Uspoređujući rezultate provedenog istraživanja konzultantskih kuća u Hrvatskoj i istraživanja u svijetu [12], vidljivo je kako je glavna grana industrije koja u najvećem postotku uvodi ERP rješenja u oba slučaja upravo *proizvodnja*. Hrvatske konzultantske tvrtke prate trendove te uočavaju važnost neovisne nadogradnje i integracije ERP-a s dodatnim funkcionalnostima.

LITERATURA

[1] I. Peric, P. Groselj, A. Sujova, M. Kalem, K. Greger, i J. Kropivsek, „Analysis of Implementation of Integrated Information Systems in Croatian Wood Processing Industry“, *Drv. Ind.*, sv. 70, izd. 2, str. 129–139, lip. 2019.

[2] G. Assenza *i ostali*, „Oscillatory Activities in Neurological Disorders of Elderly: Biomarkers to Target for Neuromodulation“, *Front. Aging Neurosci.*, sv. 9, 2017.

[3] „2017-ERP-Report.pdf“. [Na internetu]. Dostupno na: https://www.dropbox.com/home/_ERP_2017%20(Izidora)/Panorama%20materijali?preview=2017-ERP-Report.pdf. [Pristupljeno: 10-lis-2019].

[4] C. Barth i S. Koch, „Critical success factors in ERP upgrade projects“, *Ind. Manage. Data Syst.*, sv. 119, izd. 3, str. 656–675, tra. 2019.

[5] D. Žabjek, A. Kovačić, i M. Indihar Štemberger, „Business Process Management As An Important Factor For A Successful Erp System Implementation“, *Economic research - Ekonomska istraživanja*, sv. 21, izd. 4, str. 1–18, pros. 2008.

[6] M. Nikitović i V. Strahonja, „Critical Success Factors Aspects of the Enterprise Resource Planning Implementation“, *Journal of Information and Organizational Sciences*, sv. 36, izd. 2, str. 135–146, pros. 2012.

[7] K. Šimunović, G. Šimunović, S. Havrlišan, D. Pezer, i I. Svalina, „Uloga ERP sustava u poslovnom procesu i edukaciji“, *Tehnički vjesnik*, sv. 20, izd. 4, str. 711–719, kol. 2013.

[8] M. Heredia-Calzado i A. Durendez, „The influence of knowledge management and professionalization on the use of ERP systems and its effect on the competitive advantages of SMEs“, *Enterp. Inf. Syst.*, sv. 13, izd. 9, str. 1245–1274, lis. 2019.

[9] S. Malinić i M. Todorović, „How Does Management Accounting Change Under The Influence Of Erp?“, *Economic research - Ekonomska istraživanja*, sv. 25, izd. 3, str. 722–751, ruj. 2012.

[10] C. K. M. Lee, L. Zhang, P. X. Lee, i K. O. Au, „Using ERP Systems to Transform Business Processes: A Case Study at a Precession Engineering Company“, *International Journal of Engineering Business Management*, sv. 1, izd. 1, str. 19–24, ožu. 2009.

[11] P. Bingi, M. K. Sharma, i J. K. Godla, „Critical Issues Affecting an ERP Implementation“, *Information Systems Management*, sv. 16, izd. 3, str. 7–14, lip. 1999.

[12] „2018 ERP Report.pdf“. [Na internetu]. Dostupno na: https://www.dropbox.com/home/_ERP_2017%20(Izidora)/Panorama%20materijali?preview=2018+ERP+Report.pdf. [Pristupljeno: 10-lis-2019].

[13] „Factors affecting post-implementation success of enterprise resource planning systems: a perspective of business process performance: Enterprise Information Systems: Vol 13, No 9“. [Na internetu]. Dostupno na: https://www.tandfonline.com/doi/full/10.1080/17517575.2019.1612099?casa_token=rtLEHKDiTP4AAAAA%3ACtSGfY58Edyu8YsSpHJFrfdAJ3OtRPH3jCsbUYgP1CMFgEJPsfxKJ7PnBxpYpaUb9xRKc2wEquvX. [Pristupljeno: 20-stu-2019].

[14] B. Erkayman, „Transition to a JIT production system through ERP implementation: a case from the automotive industry“, *International Journal of Production Research*, sv. 57, izd. 17, str. 5467–5477, ruj. 2019.

[15] M. Ghazleh, S. Abdallah, i M. Khan, „Critical internal organization's forces influencing sustainability of post ERP in UAE service industry: A confirmatory factor analysis approach“, *International Journal of Organizational Analysis*, pros. 2018.

DE-DS

International Conference on

DIGITAL ECONOMY AND DIGITAL SOCIETY

Steering Committee

Chairs: Mladen Mauher, Zagreb University of Applied Sciences, Croatia
Edvard Tijan, University of Rijeka, Croatia

Members: Dragan Čišić, University of Rijeka, Croatia
Alen Jugović, University of Rijeka, Croatia
Đani Juričić, Jožef Stefan Institute, Ljubljana, Slovenia
Anđelko Milardović, Institute for Migration and Ethnic Studies, Zagreb, Croatia
Mladen Mauher, Zagreb University of Applied Sciences, Croatia
Ana Perić Hadžić, University of Rijeka, Croatia
Davor Salamon, Zagreb University of Applied Sciences, Croatia
Vjeran Strahonja, University of Zagreb, Varaždin, Croatia
Edvard Tijan, University of Rijeka, Croatia
Marijana Vidas-Bubanja, Belgrade Business School, Serbia

Program Committee

Members: Saša Aksentijević, Aksentijević Forensics and Consulting, Ltd., Viškovo, Croatia
Mladen Mauher, Zagreb University of Applied Sciences, Croatia
Edvard Tijan, University of Rijeka, Croatia

PAPERS

E-Democracy Tools Adoption: Experience of Austria, Croatia, Italy, and Slovenia

V. Roblek*, I. Strugar ***, M. Meško**, M. Pejić Bach*** and B. Jaković***
* Faculty of Organisational Studies, Novo mesto, Slovenia
** University of Primorska, Faculty of Management, Koper, Slovenia
***University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia
e-mail address: vasja.roblek@gmx.com, istrugar@efzg.hr, maja.mesko@gmail.com, mpejic@efzg.hr

***Abstract* - The paper focuses on the importance of e-democracy and the introduction of ICT in democratic processes, which in the age of computerization has a significant impact on the development of e-participation in developed societies. Based on comparisons of e-participatory tools in Austria, Croatia, Italy, and Slovenia, we want to show the current situation at both the local and national levels. It should be emphasized that with the development of e-participation tools, Croatia and Italy are ahead, while Austria and Slovenia are lagging behind the current trends in the development of e-participation. The reasons for this are mainly due to a lack of confidence in online security and a decline in interest for e-democracy, especially in Slovenia.**

Keywords – e-democracy; e-participation; decision-making; European Union; e-voting

I. INTRODUCTION

The concept of informatization is an integral part of the 4th industrial revolution, where it appears in the context of the development and implementation of various smart systems [1]. These technologies play a role in de-bureaucratization, citizen deliberation, participation and decision-making in an urban environment [2], [3]. The management of the urban environment is part of the "intellectual framework of second-order cybernetics and presents urban subjectivity in terms of data flows. Planning as a political practice is superseded by environmental-behavioral control, in which subjectivity is supra-individual (permeating the city with sensing nodes) and infra-individual (making citizens into sensing nodes)" [4]. During the first decade of the 21st century, social innovation e-participation projects based on the digital platforms and cyber-physical systems were introduced (e.g. open government, platforms for participative budgeting, platforms for citizen participation in public policy decision making). These platforms and systems have become a key generator of change in political administration, bureaucratic administration and customer service delivery [5].

Public administration, digitalization, and informatization are created at the national micro-level (bottom-top approach). Technologically and socially advanced urban communities are taking advantage of smart governance systems and platforms, known as e-government (e-gov) and e-democracy (e-dem). This not only provides all the precursor functions but is established through a cyberspace and digital platform (deliberative, participative and decision-making system) of a new, more liberal and participatory approach that enables the introduction of direct democracies in technological advanced urban communities (e.g. smart region, smart city, smart village) [6]. It is going for a democratic process of self-organization within the scope of (cyber) community and government interactions [7].

e-Democracy has become an important factor in citizens' decision-making. World-leading countries based on online citizen participation rates in 2018 are Denmark, Finland, the Republic of Korea, Netherlands, and Australia. Japan, New Zealand, Spain, United Kingdom and the United States [8]. The importance of the topic of socio-political and socio-technological development led us to the research question: What is the situation about the development and implementation of e-Democracy tools in Austria, Croatia, Italy and Slovenia? The goal of the paper is mainly to present tools that are available in the analyzed countries. The paper gives a realistic picture of the situation and serves as a basis for both the civilian public and politicians to further discussions.

II. THE EMERGENCE OF THE E-DEMOCRACY IN EUROPEAN UNION

A. Theoretical frameworks of e-democracy

Information and communication technology, which was introduced in the early 1980s, removes time and space barriers for interactions between people and institutions, which means that it allows for a new approach to direct democracy [9]. The development of ICT has since the 1990s onward shown that it can be used as a tool for participatory and deliberative models of democracy.

The concept of e-Government, as it is known in the 21st century is presented in Figure 1. E-Gov is divided

according to purpose to the public e-Service and e-Democracy model. Authors are using different definitions of e-Dem. So today the term e-Dem is torn between understanding e-democracy as part of e-government and in the connection with the term cybernetic policy and e-participation [10]. Nguyen and Alexander [11] proposed a plebiscitary model, in order to develop a conventional ideal democracy as a political system, in which citizens actively participate in public decision-making by electronic voting. In contrast to Nguyen and Alexander, Barber [12] suggested a so-called model of deliberative e-Dem. Barber thus believes that "the e-Dem consultation model should focus on citizen participation in debates on public affairs rather than e-voting". Kim [13] considers the e-Dem consultative model as a medium in cyberspace for implementing processes of the public sphere described by Habermas [14].

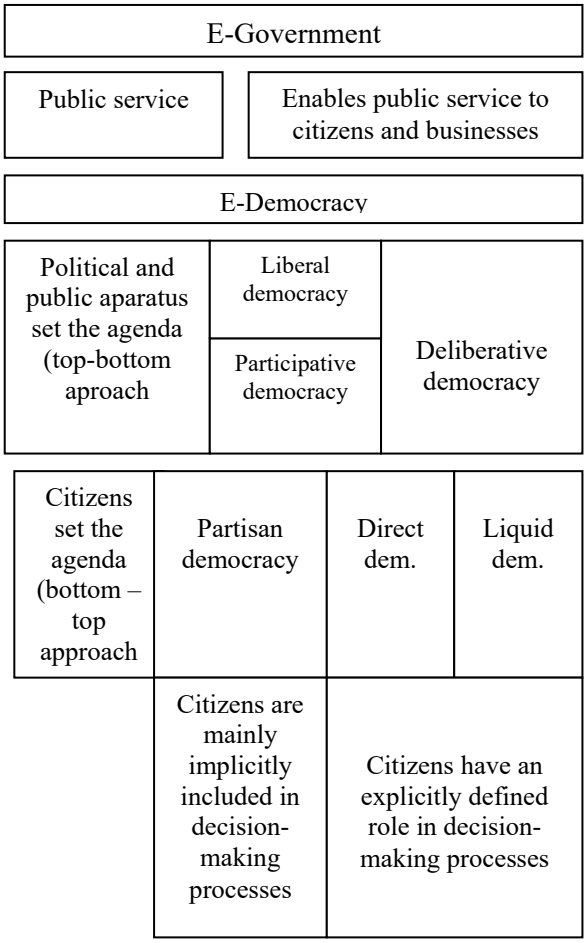


Figure 1. E-Government concept and e-Democracy

Deliberative democracy theory focuses on the deliberative quality of the policy-making process. "The theory holds that the legitimacy of democratic decisions can be enhanced if such decisions are preceded by a delusion that is free from distortion and unequal power between deliberators" [15]. Deliberative democrats are divided between "those who focus on deliberation between political elites and those who have a more populist orientation and engage in the debates of lay citizens" [16]. Despite their substantive differences, both the deliberative and the participatory models present a significant impact of information and communication technology on future democratic development. With the introduction of the citizen participation by policy decision making in the community that runs through web portals and mobile applications, authorities should be paying increasing attention to the problem of exponentially rising amounts of information that have the potential to cause information overload and lead citizens into frustration and apathy [17].

Van Dijk [18] estimated that the contribution of the development of e-Democracy since the 1990s had been limited primarily to facilitating access and exchange of political information. It should be noted that in the last few years, we are witnessing increased censorship of online (e.g. via Facebook, Twitter, Instagram, YouTube, etc.) political posts in Europe. Opponents of the new EU copyright directive (Directive (EU) 2019/790) believe that the directive is introducing web content filtering, which will be resulted in the removal of content from the web [19]. On the other side, non-mediated governance aims to minimize the role of political administration and its bureaucracy, with technology shifting the decision-making power to citizens, by allowing administrative procedures to be led by technology itself [1]. Informatization is enabling direct or fundamental democracy, which has emerged as a radical contrast to traditional forms of representative democracy. This means that political decisions, or at least essential decisions, are made by the public. Unlike participatory democracy, which limits the extent of public participation and leads to decisions that are not legally binding on government or parliament, direct democracy is understood as a form of democracy, where citizens take control of both the design of key processes and the decision-making process. All decisions made in this way become legally binding, where a precondition for establishing such a model is that the Internet becomes a key factor in establishing democracy and thus replacing traditional forms of communication [21]. One such concept is liquid democracy (LD) which is connected to direct democracy. In the LD, the citizen decides for himself or herself whether he wants to participate in the decision on a particular issue or to pass his vote on to someone who may be more aware of the issue or issue and has the necessary knowledge to make a decision [22].

The role of e-Democracy at the end of the twenties of the 21st century is thus mainly focused on the online involvement of citizens in the formulation of opinions and political decision-making. E-Dem thus enables citizens who have the right to vote to participate equally in the proposal, development, and formulation of government decisions. It is going for a process that contains "social, economic and cultural conditions that allow for free and equal practice of political self-determination". E-Dem is presented as a democratical tool that enables the involvement of individual citizens in

political decision-making through e-participation techniques (for consultation purposes) and e-votes [1].

B. The differences between the e-democracy theory in EU and USA

The crucial difference in the development of e-Democracy concepts in the EU and the US is that the EU has focused on more democratic strategies, while the US is focused more on the technological development itself [24]. It should be noted that European policies focused on technological projects should choose the projects that are most beneficial to EU citizens. Since the Tindeman Report in 1976, which stated that a European project must be understood as being beyond just an integrated market [23]. Thus, we can conclude that the European model of e-Democracy as an EU project must be citizen-centered.

The EU is considering establishing its own form of digital sovereignty, by balancing many elements of cyberspace and establishing its borders in cyberspace and ensuring its security [24]. The development of cyberspace has thus led to many challenges regarding the regulation of the Internet, and the setting of minimum regulatory standards regarding its interoperability. These rules are inherently controversial, given the specific nature of the network's operation and the need to balance all fundamental human rights related to the exercise of freedom of expression and association, which are often contrary to other rights and public interests. The North American and European approaches on development and control of cyberspace are different, and this shows how complex and delicate they are [25]. European citizenship is expected to gain in importance in this case, because the European political space and its identity are not limited by nationalities. Allowing the sharing of common values and characteristics of the user in cyberspace will mean greater integration in the European cyberspace, which is crucial for the level of the single digital market and e-democracy [26].

C. The emergence of smart governance

Smart governance can be defined as a political democratic governance framework, which has for its goal to introduce the combination of digital technologies (e.g. urban data platforms as interfaces for smart governance) and innovative practices for the purpose of improving the government (public) service delivery and citizen inclusion in a political process, which is categorized as the first phase of participation. In the second phase responsive, transparent and accountable decision – making can be seen, in the context of developing and implementing public policies and concrete public projects [27, 28]. Different authors (Bolivar [29]; Castelnovo, Misuraca, & Savoldelli [30] concluded that smart governance, under the condition of efficient use of the power and authority of politicians.

This has presented a framework of a political and public administration tool that elicits trust and public confidence in government institutions at all levels. Smart governance plays a vital role in the smart city initiative, creating a complex dialogue between governments (supply), citizens (demand) and other stakeholders [30, 31]. It can be said that the informatization of the public apparatus contributes to a political decision – making and other attributes that contribute to ethics [32].

As can be seen in figure 2 the implementation of smart governance leads to a change in traditional representation in political and public administration, which can use for communication with the citizens the top-bottom approach or bottom-top approach. It depends on social consensus on the level of implementation of smart governance that brings different distributions of decision-making power in the community.

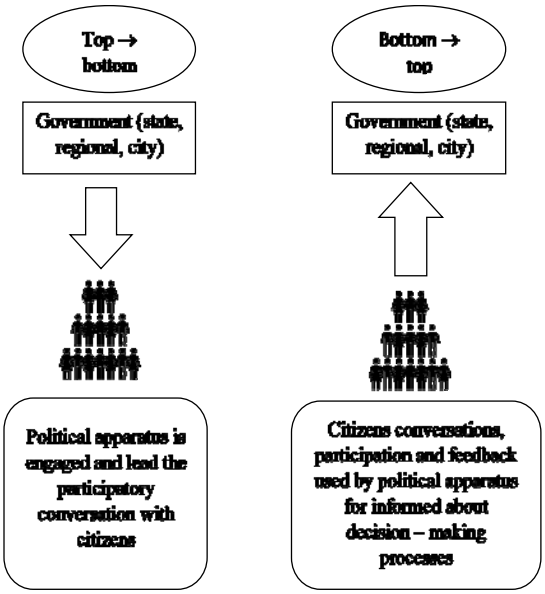


Figure 2. E-Democracy approaches

The responsibility of politicians includes governance of domains, according to the constitution and legislation. The political systems (democratic or not) have to be viable, and for this, they need value and stability. This process includes open government data, transparency, accountability and evidence of governing efficacy. Corruption has an impact on efficacy, together they are responsible for the loss of trust and public confidence. According to Schildkraut [33], efficiency and citizen trust represent the political capacity to govern, where the loss of trust can lead to political instability.

III. E-DEMOCRACY APPROACHES IN AUSTRIA, CROATIA, ITALIA AND SLOVENIA

In this section, we focus on the progress of e-democracy development in Austria, Croatia, Italy and Slovenia. Based on the analysis of websites and expert texts, we have selected some of the most important examples for each country. The analysis focused on the functions of e-information, e-counseling, e-initiative and e-voting.

A. Austria

E-government is in Austria focused primarily on the provision of e-services for legal entities and citizens. It is surprising that Austria is considered as one of the leading countries in the EU, in regards to the provision and use of e- Gov tools (e.g. portal Help.gv.at), but in the context of the emergence of e-participation it is far behind. The cause of the underdevelopment of e-democracy and e-participation mechanisms in Austria is found in citizens' mistrust of internet security (data confidentiality) and lack of education in the operation of ICT technologies [34]. This situation has resulted in citizens not yet being able to participate and decide directly on public affairs, through digital media. In addition, in Austria, it is not possible to vote online. The Austrian government has since 2012 enabled access to public data (data.gv.at) by creating a public information catalog, which represents the Austrian open government [35]. In Austria development of e-Democracy and e-participation depends on the organization of civil society, where the leading developer such a project is the "Network Civil Society" (Netzwerk Zivilgesellschaft). This project is affiliated with the "Academy of Civil Society" (Akademie der Zivilgesellschaft), where the network brings together various initiatives to share experiences and best practices.

The Academy offers several courses in e-democracy and e-participation and encourages voluntary activity in this field [36]. Their activities caused that in Austria come to the informal forms of e-dem and e-participation, which are mostly organized through websites such as Avaaz. Thus, in Austria, a website has been developed, dedicated to social activism in the form of online petitions. Carrying out informal online petitions connects different stakeholder groups with different target groups, such as public authorities, private companies and other organizations. There are no specific procedures in place for the implementation of these petitions, such as whether, if and how those addressed by the petition must respond to it [37]. An example of such a petition is the request of Graz residents. The resident requests participation in the preparation of municipal urban planning and wants to be informed about future projects at an early stage. The citizen also expressed the view that the central concern of the city should be to strengthen the participation of citizens in the preparation and decision-making of urban policies [38]. At the level of development of e-participation in cities, it is necessary to mention the participatory portal of the Vienna city administration (Partizipationsplattform der Stadt Wien: Ideen entwickeln. Diskutieren. Umsetzen.). As the name of the web site itself indicates, it is intended for consultation with citizens, who can also submit concrete proposals to the administration through it. For example, the portal discusses the preparation of design proposals and the implementation of the participatory budget for 2017 and 2019 in Margareten and Simmering counties, in addition to the portal, direct discussions in the town quarters and video conferences, discussions on the Urban Digital Agenda 2020, digital health, Internet of Things, etc. [39].

B. Croatia

At the state level (top-level), it is necessary to mention two web portals that fulfill the functions of e-participation: Imamopravoznati.org and E-consulting. In the case of Imamopravoznati.org, it enables sending inquiries to public authorities in Croatia. Following the Right of access to information act (OG 85/15) [40], the portal allows all citizens to search for information held by a state authority in Croatia. The seeker will get a response if it is a valid piece of information within the meaning of the law [41]. The e-Savjetovanje is a web site intended to consult the interested public about a large number of draft laws and regulations adopted in Croatia every year. In public consultations, they usually prepare draft laws proposed by the ministries, as well as various regulations proposed by state agencies and offices. The purpose of the e-counseling platform is to enable interested publics to become acquainted with the content of the regulations in preparation, before they enter into force, with the opportunity to read and participate in the drafting of their provisions in advance. Any consultation also entails the obligation of the authority which has prescribed the regulation to advise that it must be informed after the consultation has taken place. Besides, all comments received and all statements by public authorities are published on the e-counseling page [41].

Croatian cities have only Rijeka, Zagreb, Pula, and Osijek marked as open data on their web pages. Access to GIS (Geographic Information System) web pages of Croatian municipalities are widely available, but municipal web sites vary depending on the data access options. The City of Zagreb is considered dominant in the quality of GIS services. Rijeka offers the best range of publicly available open data for individuals and legal entities [42]. In the area of citizen participation in decision-making processes, only a few cities in Croatia have so far developed an interactive digital implementation of public consultation, which has also developed the possibility for citizens to decide on the formulation of a city budget online. In the area of e-participation and open budget planning, Croatia's leading cities are Pula and Rijeka, and the Croatian public expects that the cities will develop standards that will be followed by other Croatian cities. Both sites have developed an online e-counseling format for advising citizens. In Croatia, they consider it necessary to develop e-counseling systems in all cities, but also to introduce them at a national level, which enables the interactive participation of interested publics in legal procedures, legislation and other regulations and acts. Open Budget Planning includes Rijeka, Zagreb (both in the big cities category) and Sisak (in the smaller cities category) cities that have developed an interactive interface for budget planning and control, while all other cities have a spreadsheet of budgets with visualizations. In the availability of budget documents, Split, Rijeka, Pula and

Karlovac are the most successful cities in terms of the longest time period of the accessible documents. The cities Rijeka and Pula also provide e-participative budgeting to their citizens [42].

C. Italy

In Italy, the development of e-democracy began in 2004, when the State Ministry of Innovation and Technology co-financed projects for the development of digital citizenship and citizen participation in territorial governance. This experience has led to the further development of institutional websites with an aim to enable access to public information and the creation of forums to facilitate public debate. As part of the Open Data project, Italy has developed a website at www.dati.gov.it at the national level, which enables access to approximately 150 publicly accessible databases [43]. Access to geo-data has also been developed throughout the country [44]. In Italy, along with institutional initiatives, private (often amateur) projects have begun to emerge, aimed at creating programmatic websites for the development of direct and participatory democracy. Thus, in 2013, individual members of the Democratic Party, civic choice and left-wing ecology and freedom joined a liquid feedback-based e-democracy platform [45]. As of 2015, the platform is no longer active. In 2013, a consultative referendum was held in Salento, for the first time in Italy, with a unique digital ballot [46]. At the local level, two provincial projects need to be mentioned. Emilia Romani set up the "io Partecipo +" website. The portal allows citizens to get involved in the process of collaborative design by engaging in "virtual room" discussions as part of the preparation for some regional policy decisions [47]. In Tuscany, the first region in Italy to set up a participatory website ("Partecipa Toscana"), which allows public participation and deliberation within the regional political system (e.g. integration of thematic forums, dialogues or other forms of debate between citizens). An important public discussion via the website took place in 2016 on the development of the port of Livorno.

As part of the Regional Participation Project, Tuscany has also set up its own "Open Toscana" website for open access to public information. The Tuscan project has also encouraged Emilio-Romagna, Sardinia and Umbria to set up their own e-participatory projects [48]. In 2015, Milan launched a two-year test project on participatory budgeting, which was carried out in four phases: The first phase was open consultation with citizens, where particular attention was given to social groups, such as young people and immigrants. The second phase of joint planning involved one group of 30 citizens per district. The third phase involved voting online or in person at selected dates. Those that were 14-18 years old could only vote in person with a ballot paper. The fourth phase involves the implementation of selected projects. The primary participants are municipal officials, technicians and experts, although the public monitors their progress on the board's website. The implementation of the projects took three years and ended in 2018 [49]. The implementation of the participatory budget also took place in 4 phases in 2018 in Bologna [50].

D. Slovenia

The key web portal of the Slovenian public administration is eGovernment. The e-democracy site is located within this portal. The site allows citizens to: Participate in the form of commentary on the drafting of regulations with their opinions, proposals, comments and initiatives to the proponents and decision-makers; provides information on e-access and contacts of public officials and individual ministries; provides basic information on democracy and its implementation, on fundamental human rights and freedoms and on the regulation of power in Slovenia and the EU; it also includes a guide to the procedures of drafting regulations. In this guide any citizen can find the basic process of drafting a regulation before it is submitted to the National Assembly for discussion [51]. Open Access to Public Information in Slovenia is provided by the Slovenian Open Data Portal (OPSI). OPSI thus represents a single national online point for the publication of open data for the entire public sector [52]. Civil society is also involved in the development of the latest approaches to e-democracy and e-transparency in Slovenia, following the latest trends in open data and open government. Thus, NGO Today has developed a new parameter today. It is open-source software that, in addition to visually displaying the results of the vote, allows users to see which Members have participated in committee and parliamentary meetings. The site also makes it easy to share results on social media and embed results in online news [53]. At the local level, e-participation in Slovenia is extremely poorly developed, even though all municipalities have websites. However, very few municipal websites include e-survey or e-forum. They are mainly intended for information, access to public information and e-services [54].

IV. Conclusion

All four countries have access to public information at both national and local levels. Only in the field of e-participation, in Italy and Croatia, at the local level, is much more developed than in Austria and Slovenia. In Austria, only the city of Vienna stands out with its website. Unfortunately, at the local level, except for necessary information and e-service, Austria and Slovenia are not offer its citizens e-participation solutions. At the national level, all countries provide only basic options for informing and commenting on proposed draft legislation. Civil society plays a much higher role in raising awareness and developing e-participatory tools in Austria, Slovenia, and Italy. The next steps of the development and introduction of the e-participatory tool will require consensus in all four countries, between the public institutions and citizens. Public institutions will have to invest in innovative e-democracy technology, while citizens will have to be aware of the importance of these tools and to use them carefully.

Further research on e-Democracy in selected countries will be focused on the professional public opinions about the next steps of implementing e-Democracy tools and their importance for the emergence of new models of democracy, as well as socio-technical perspective [55]. It will also be necessary to carry out a survey among the citizens and identify their pro and contra opinions and suggestions about the e-democracy and its further emergence.

REFERENCES

[1] A. Paulin, Smart city governance. Amsterdam: Elsevier, 2019.

[2] I. Calzada, "The techno-politics of data and smart devolution in city-regions: Comparing Glasgow, Bristol, Barcelona, and Bilbao," Sys., vol 5, no. 1, 18, 2017.

[3] B.Wilson, and A. Chakraborty, "Planning Smart (er) Cities: The Promise of Civic Technology", J. of Urb, Tech., vol 26, no. 4, pp. 1-23, 2019.

[4] M. Krivý, "Towards a critique of cybernetic urbanism: The smart city and the society of control", Plann, Th. vol. 17, no 1, pp. 8-30, 2018.

[5] M. Barrett, E. Davidson, J. Prabhu, and S.L.Vargo, "Service innovation in the digital age: key contributions and future directions," MIS q. vol 39, no. 1, pp. 135-154, 2015.

[6] D. Helbing, S.B. Frey, G. Gigerenzer, E. Hafen, M. Hagner, Y. Hofstetter, ... and A. Zwitter, "Will democracy survive big data and artificial intelligence?" in Towards digital enlightenment, D. Helbing, Ed. Cham: Springer, 2019, pp- 73-89.

[7] J. Edelenbos, I. van Meerkerk, and T.Schenk, "The evolution of community self-organization in interaction with government institutions: Cross-case insights from three countries," The Am, Rev. of Publ. Ad., vol 48, no. 1, pp. 52-66, 2018.

[8] United Nations E-government survey 2018, retrieved from https://publicadministration.un.org/egovkb/Portals/egovkb/Documents/un/2018-Survey/E-Government%20Survey%202018_FINAL%20for%20web.pdf

[9] A. Toffler, Future shock, New York: Random House, 1970.

[10] M. Kneuer, "E-democracy: A new challenge for measuring democracy", Int. Pol. Sci. Rev., 37(5), pp. 666-678, 2016.

[11] T.D. Nguyen, and J. Alexander, "The coming of cyberspacetime and the end of the polity," in Cultures of internet: Virtual spaces. Real histories, living bodies, R. Shields, Ed. London: Sage Publications, 1996, pp. 99-124.

[12] B.R. Barber, Strong democracy: Participatory politics for a new age, 20th ed. Berkeley: Univ of California Press, 2003.

[13] K.C. Kim, "A Cross-national Analysis of Global E-government," Pub. Org. Rev., vol 7, no 4, pp. 317-329, 2007.

[14] J. Habermas, The structural transformation of the public sphere, trans. T. Burger. Cambridge: MIT Press, pp. 85-92, 1989.

[15] D. Held, Models of democracy. Stanford: Stanford University Press, 2006.

[16] L.D. Ponet, and E. J. Leib, "Fiduciary's law's lessons for deliberative democracy," BUL rev. 91, 1249, 2011.

[17] A. Davis, "New media and fat democracy: the paradox of online participation, " New med. & soc. vol. 12, no. 5, pp. 745-761, 2010.

[18] M.G.A.J van Dijk, "Digital democracy: Vision and reality," In Public administration in the information age: Revisited, I. Snellen, M. Thaens, and W. deDonk, W., Eds. Amsterdam: IOS-Press, pp. 49-61

[19] C. Nichols, "The state of internet censorship in Europe," For. Pol. Jour. Feb 21, 2019.

[20] Z. Asatryan, T. Baskaran, T. Grigoriadis, and F. Heinemann, "Direct democracy and local public finances under cooperative federalism," The Scan. Jour. of Eco., vol 119, no, 3, pp. 801-820, 2017.

[21] R., Lindner, L. Hennen, and G. Aichholzer, Electronic democracy in Europe. Berlin: Springer, 2016

[22] C.R. Vieira, "The cyberpolitical space of the European Union: an overview of e-Government, e-Democracy and interoperability in the European space and its citizenship". UNIO–EU Law Jour., vol 4, no.2, pp. 117-127, 2018.

[23] Hamburg Institute of International Economics, Ed., "Tindemans-report: Europe à la carte, Intereconomics", Hamburg: Verlag Weltarchiv, vol. 11, no. 2, pp. 36.

[24] A. Chadwick, and C. May, "Interaction between states and citizens in the age of the internet: "e - Government" in the United States, Britain, and the European Union," Gov. vol. 16, no. 2, pp. 271-300, 2003.

[25] European parliament, Prospects for e-democracy in Europe, part I: Literature review, 2018.

[26] B.P. Anand, and J. Navío-Marco, "Governance and economics of smart cities: Opportunities and challenges," Tel. Pol., vol.42, no. 10, pp. 795-799, 2018.

[27] S. Barns, "Smart cities and urban data platforms: Designing interfaces for smart governance, culture and society," vol.12, pp. 5-12, 2018.

[28] R.P.M. Bolívar, "Governance models and outcomes to foster public value creation in smart cities," Sc. Reg., vol. 17, no. 1, pp. 57-80, 2018

[29] W. Castelnovo, G. Misuraca and A. Savoldelli, "Smart cities governance: The need for a holistic approach to assessing urban participatory policy making," Soc. Sci. Com. Rev., vol. 34, no. 6, pp. 724-739, 2016.

[30] V.G. Pereira, P. Parycek, E. Falco and R. Kleinhans, "Smart governance in the context of smart cities: A literature review," Inf. Pol., vol. 23, no. 2, pp. 143-162, 2018.

[31] V. Roblek, "Smart city Vienna," In Smart city emergence: Cases from around the world, L. Anthopoulos, Ed., Amsterdam: Elsiever, pp. 105-127

[32] M. Yolles, "Governance through political bureaucracy: an agency approach," Kybernetes, vol. 48, no. 1, pp. 7-34, 2019.

[33] J.D. Schildkraut, "National identity in the United States, In Handbook of identity: Theory and research, J.S. Schwartz, K. Luyckx, and L.V. Vignoles, Eds. New York: Springer, 2011, pp.846-965.

[34] R. Pohoryles, "Austria: Many attempts, but much still has to be done," In eDemocracy and eParticipation, R. Pohoryles, and R. Decheva, Eds. European Liberal Forum, pp. 60-63. Retrieved from www.liberalforum.eu

[35] Open Data Österreich, retrieved from https://www.data.gv.at, 2019.

[36] Activism - Petition Online – Österreich, retrieved from https://petition.activism.com/de_AT/#, 2019.

[37] Nachhaltig in Graz (Sustainable in Graz), "Demo gegen Bauprojekt: Partizipation in der Stadtplanung Graz(Demo against construction project: Participation in urban planning Graz)," 2019. Retrieved from https://nachhaltig-in-graz.at/Veranstaltung/demo-gegen-bauprojekt-partizipation-in-der-stadtplanung-graz/

[38] Partizipationsplattform der stadt Wien (Vienna participationplaform), Diskussionen (Discussions), 2019. Retrieved from https://www.partizipation.wien.at/en.

[39] Zakon o pravu na pristup informacijama (Act on the Right of Access to Information), 2015. Retrieved from https://www.zakon.hr/z/126/Zakon-o-pravu-na-pristup-informacijama

[40] Imamopravoznati (We have rights to know). 2019. Retrieved from http://imamopravoznati.org

[41] e-Svetovanje (e-Counseling), 2019. Retrieved from https://esavjetovanja.gov.hr/ECon/Dashboard

[42] Apsolon, Analiza digitalne spremnosti hrvatskih gradova 2019 (Digital readiness analysis of Croatian cities, 2019), 2019 Retrieved from www.apsolon.com

[43] I dati aperti della pubblica amministrazione (public administration open data), 2019. Retrieved from www.dati.gov.it

[44] Repertorio nazionale dei dati terriotoriali (National directory of spatial data), 2019. Retrieved from www.geodati.gov.it., 2019

[45] R., De Rosa, Cittadini digitali: L'agire politico al tempo dei social media (Digital citizens: Political behavior at the time of social media). Milano: Fabio Gialain, 2014.

[46] M. Pennisi, "Nei comuni di Melpignano e Martignano si sono raccolte le preferenze dei cittadini con una speciale urna digitale (In the municipalities of Melpignano and Martignano the preferences of citizens were gathered with a special digital urn)," Ecco com'è andata Retrieved from https://daily.wired.it/news/tech/2013/05/08/voto-elettronico-salento-2578295.html

[47] Regione Emilia-Romagna, "Partecipazione (Participaton), 2019. Retrieved from https://partecipazione.regione.emilia-romagna.it/iopartecipo

[48] S. Fletcher, The Tuscany regional participation policy, Italy," 2019. Retrieved from https://participedia.net/method/5594

[49] P. Spada, "Participatory budgeting in Milan - cycle 2015-2016," 2018. Retrieved from https://participedia.net/case/4838

[50] Commune di Bologna, "Bilancio partecipativo (participative budget)," 2019. Retrieved from http://partecipa.comune.bologna.it/bilancio-partecipativo/info

[51] eUprava RS, "Država in družba," 2019. Retrieved from https://e-uprava.gov.si/drzava-in-druzba/e-demokracija.html

[52] Odprti podatki Slovenije., 2019. Retrieved from https://podatki.gov.si

[53] Parlameter. from https://parlameter.si

[54] M. Haček, and S. Kukovič, "Razširjenost orodij e-demokracije in e-participacije v slovenskih občinah (Prevalence of e-democracy and e-participation tools in Slovenian municipalities)," Teorija in praksa, vol. 50, no. 5/6, pp. 736-752, 2013.

[55] P. Bakunzibake, G.O. Klein, G.O., S.M. Islam, "E-Government Implementation Process in Rwanda: Exploring Changes in a Socio-technical Perspective", Business Systems Research, vol. 10 no. 1, pp. 53-73, 2019

The importance of digitizing the management processes of sports associations in the city of Zadar

J. Miočić

Department of Tourism and Communication Studies University of Zadar, Zadar, Croatia

josip.miocic@szgz.hr

Abstract: **Digital transformation is an opportunity and a challenge for associations, where a modern society is expected to adapt to the market and use modern infrastructure and tools through which associations offer their services and benefits to users. Communication in those processes represents the transfer of information, ideas, knowledge, results, mutual agreement, etc., or more precisely, it is one of the preconditions without which there is no work within the associations. The digitization process itself is a step in the process of delivering digitized content to users. The term digitalization also implies the use of digital technology to create new values in the work processes of an association and to change business models, thereby representing a process that leads to the overall digitization of the association's work. In this paper, the research was conducted through a survey about the use of digital technologies in the administrative and management processes of associations, based on the example of the members of the Sports Association of the City of Zadar (SACZ). The questionnaire concluded that the associations aspire to use digital tools, that is, it was found that in the segment of administrative and management processes, members associations of the SACZ clearly lag behind modern standards, and only partly use modern tools in the training process. In response to the hypothesis, it was found that the associations do not sufficiently use information technology for the purposes of managing the administrative and management processes of the associations. Thus, there can be concluded that 71% of respondents believe it is necessary to enable the use of a modern e-application that integrates all management processes.**

Keywords: **digitalization, information, associations, transformation.**

INTRODUCTION

With the development of IT technology that now becomes an unavoidable segment of daily life, a need arises to digitize the management processes of associations, both regarding administration and training. Until now, various associations have generally used tools such as Word and Excel in order to make their administrative obligations more practical. With the arrival of Google's various popular solutions (Drive, Calendar, Docts, etc.) a simpler way for them to function on a daily basis became possible (14). However, this method also lasted briefly due to the fact that documents were "scattered" on all sides, and the constant transcription of those present for the activities in electronic documents, reviewing the statements to determine who is present and in what form they use the tools, prompted the need to digitize the processes. By using modern tools in the processes of digitization, a need was identified to create an e-application that will digitize the above-mentioned activities of an association, and that would in turn slowly contribute to the solving of administrative problems. The e-application is fully in compliance with the applicable laws defining the frameworks of the business of associations, which in turn corresponds to all profiles of existing associations, including the ones focusing on sports. The aim of the paper is to examine the attitudes and opinions of the respondents regarding the methods and usage of digital technologies in everyday work processes of associations and to determine in which work segment the associations are most frequently using digital technologies. The respondents in the paper are 124 sports associations, which represent 89% of the total number of active sports associations that are members of the Zadar City Sports Community.

II.DIGITIZATION

In recent years, modern IT technology has been a factor that significantly affects the development of society (Society 5.0) and the economies within societies (Industry 4.0). By interacting with the environment and by processing enormously large amounts of data through digital tools, complex algorithms have introduced modern technology to life (8). The digitization process itself is just one step in the process of delivering digitized content to users. Attempts to digitize the management process in associations are conducted on a daily basis, based on trial and error. In societies, a digital transformation brings about a number of changes, and for nonprofit employees it involves acquiring new skills, upgrading knowledge, methods of doing things, etc. (4). Within that context, the leading man of the world's top medical research institute *TGen* - J. Lowey, states that during the process of digitization, it is crucial to have a good team that is always ready to embrace new technologies and keep pace with technological changes that go beyond a formal education" (3). In today's society, the transformation of business organizations, which is based on digital technology, represents a basic and now indispensable segment of business work-cycle. As such, digitalization represents the transformation of an organization's business model using digital technologies. The beginning of digitalization can be

observed as the application of various technologies "such as *Cloud* computing, *Big Data*, *Internet of Things* (*IoT*), while today, in the context of digital transformation, we can talk about the use of robots, artificial intelligence (AI), 3D printing, drones, or more precisely, the so-called "Industry 4.0" (2). Digitization and digital transformation have a major impact on business processes carried out by organizations and/or associations, as they not only change the way these entities work, but also the content of the work, focusing on the end user (15). Users in contemporary societies have drastically changed both their expectations as well as behavior patterns. Therefore, in order for non-profit organizations and/or associations to be attractive and competitive on the market and to attract customers, they must adapt to the requirements and change their business. Namely, users do not expect associations to respond to their demands, but rather they expect them to anticipate and create their future needs, even before the actual needs were defined by the users themselves. Therefore, IT and their applications in the business function are one of the tools that help to make management processes in organizations more successful. The application of the digital documentation management tools, during the life cycle of a particular process is of particular importance (13). The business function of the associations contains numerous and varied documentation, for which it is of the utmost importance that it be updated and archived so that the productivity of the management process can be increased (9). The digital tools for managing the business of non-profit organizations, which include associations, are the following: tools for monitoring administrative and training needs, financial tools and tools for improving the management process. The business activities of associations are currently conceptualized to consist of members, employees and external users and the processes they carry out - financial statements, administrative procedures, sports results and the training process (10).

III.PRINCIPLES OF INTERNAL COMMUNICATION IN THE CONTEX OF THE DIGITAL TRANSFORMATION

Communication and internal communication involve the transfer of information, therefore digital communication is the physical transfer of information converted into a digital form through a digitization process, whereby a computer is used to process and store data. The internal communication strategy of a digital transformation will, of course, depend on the size of the organization and the type of technology being implemented. It is crucial to inform the users about relevant issues in a timely manner and to adapt their digital transformation process by getting adequate feedback (17). Often, this feedback can be surprising by exposing, for example, a decrease in efficiency due to the lack of relevant information. As part of the digital transformation process, it is necessary to ensure a comprehensive education process for users and to properly distribute them within the different levels of education, depending on their experience and workplace to date (11). The study "The Impact of Internal Communication on Employee Satisfaction at Work" indicated the need for well-organized communication channels in the organization (1). This will reflect the awareness of the leaders of an organization regarding the fact that the learning processes of employees / users for the purpose of digitization do not stop when the organization is digitized, but instead that the processes continue.

IV. INDUSTRY 4.0 AND SPORTS

In today's society, Industry 4.0 is a strategic approach to connect systems based on "Internet technology aimed at establishing communication between machines, people, products", business systems, or markets. We were introduced to Industry 4.0 by German experts in 2011 at the World Exhibition of Industrial Technology "Hannover Messe as part of the German industrial strategy for the period up to 2020" (5). Due to its universal principles, Industry 4.0 is also being applied outside of actual production, in contexts such as healthcare, administration, sports, education, etc. This reduces the cost of providing services, while increasing the efficiency of the system. The authors see the concept of Industry 4.0 as CIM 2.0 rather than the fourth industrial revolution. In continuation, it is evident that Industry 4.0. has deeply penetrated into all spheres including sports, management, manifestations, various sports organizations and their processes (16). The most important technology trends are *lot - Internet of things*, *cloud computing* - a way to deploy and store data, additive manufacturing – such as 3D printing, smart sensors - different ways of processing data, automation, AI, machine learning and deep learning, cybersecurity, *blockhain* technology - increased connectivity, the need for security. These tools and technologies offer the opportunity to make progress in the field of sports, especially in the training process and, what is important for this work, in the field of increasing the efficiency of business management. The benefits for organizations are visible through lower operating costs, improved communications, productivity, broad customer base, outside collaboration, collaboration, improved documentation and process security, use of fewer infrastructures to achieve better sports results. Through their association, IT and the digital transformation have become the perfect tool to bring associations and services closer to their users and to make them accessible and visible at all times. For this reason, organizations must be sensitive to changes in networking, digitalization and industry 4.0 in order to respond to them and correlate to changes in society 5.0. However, a question is posed whether associations use all of the assistance provided by digitalization in the business and process management of sports organizations? In the case of the Sports Community of the City of Zadar, as well as the work of its members, an attempt will be made to answer the posed question.

V. E-APPLICATION

The E-application is a tool that represents the simplest and most complete solution for the work of an association and that is the management process. The application is a relatively easy to use tool with, of course, the required level of IT literacy by users and accessibility to the Internet. The application offers easy enrollment, membership registry tracking, group creation, and group assignment through organizational groups, etc. This also allows the tracking of finances, attendance records, absences, medical examinations, member sports scores, the financial records of an organization, sports scores, training process, etc. It is especially important to emphasize that modern applications allow the creation and printing of all necessary documents with information and the logo of the organization, registration forms, payment slips on HUB3 with 2D or QR code, various certificates, reports and the like. In addition to the above, the e-application contains tools that allow non-profit organizations to keep their simple financial books (cash register, cash and receipts book, entry and exit accounts), as well as all the necessary tools for conducting the training process. From everything previously mentioned, it can be concluded that the e-application reduces the inefficiency and unproductiveness of the administrative and training staff of an organization, thereby taking advantage of the digitization of the process (2).

VI. THE SUBJECT AND PURPOSE OF THE RESEARCH ON THE EXAMPLE OF THE SPORTS COMMUNITY OF THE CITY OF ZADAR

The typical topics in the field of sports do not usually address the use of digital technology in organizations, especially the use of modern technologies in the administrative and management processes of associations. An example of this is the program implemented by the Sports Community of the City of Zadar (SCCZ) - the Program of Public Requirements for Sports in the City of Zadar. This program is of importance for the society as a whole at all levels and needs to be adapted to contemporary social needs. That is, it is necessary to accept achievements, the trends in the field, which opens the possibility of contributing to the quality work of associations and achieving better or more adaptive results. Similar research conducted by Kesić and Vlasic (7), focused however more on the related customer satisfaction and use of digital technology, supports the conducted research. The research was encouraged by the project *The Development of an Innovative Platform for the Digital Transformation of Enterprises*, as well as the work of Pihir, Križanić and Kutnjak from 2018, which dealt with similar issues in another field. The aim of this paper is to examine the attitudes and opinions of the respondents regarding the way and use of digital technologies in everyday work processes of associations, and to determine in which work segment of a sports associations digital technologies are most frequently being used.

VII. RESEARCH METHODS

For the purpose of this research, and the examination of respondents and their use of digital technologies in the daily work of the association, the method of interviewing by means of a questionnaire was used. The used questionnaire consists of 12 questions, divided into two groups - general and research questions. The general part deals with the general data related to the respondent, while the second part is focused on questions intended to answer the proposed hypothesis. "The survey is a quantitative form of research that aims at capturing the different views and opinions of respondents" (Miočić et al., 2019), and to identify the levels and areas of use of digital technologies. The questionnaire was designed and then initially tested on a group of 6 subjects, and later on completed by all other research participants - 124 of them.

VII.I. INTERVIEWEE SAMPLE AND COURS OF TESTING

The sample of respondents consists of 124 associations, members of the Sports Community, that is, the associations that joined the examination. They make up a total of 89% of active membership. The proposed paper analyzes the results of the survey questionnaire, that was filled out at the headquarters of the Sports Community. After completing the questionnaire by the respondents, the questionnaires were introduced in printed form into the *LimeSurvey* program, and then exported to the *SPSS* statistical program for further statistical processing.

VIII. SURVEY RESULTS AND DISCUSSIONS

A group of questions related to the satisfaction with the way the data was collected formalized the answer that the way of collecting data was partially satisfactory, which is evident from the example of the question (Table 1) "How satisfactory is the way of collecting data for the purpose of evaluating the work of sports associations?". More than half of the respondents (57%) believe that the method of collection is partially satisfactory. 18% of these considered the data collection method to be neither satisfactory nor unsatisfactory, while 17% stated that the data collection method was completely satisfactory. A possible reason for such results might be the data collection using paper-based methods. However, it should also be emphasized that the analyzed data collection is considered partially unsatisfactory by 6%, and completely unsatisfactory by 2% of respondents.

Table 1 - Results for the question - To what extent is the current way of collecting data for the purposes of evaluating the performance of sports associations satisfactory or not?

Completely unsatisfactory	Partially unsatisfactory	Neither satisfactory nor unsatisfactory	Partially satisfactory	Completely satisfactory
2%	6%	18%	57%	17%

The following group of questions was related to the use of digital technologies (Table 2) - "How much do you use digital technologies in your functioning and business processes? (administrative processes, training processes and financial processes)". The largest number of respondents stated that they do not use or use digital technologies (37%), a slightly smaller percentage of respondents (30%) stated that they do not use digital technologies, a 30% of respondents stated that they only partly use digital technologies, while 3% stated that they use digital technologies.

Table 2 – Results for the question - How much do you use digital technologies in your functioning and business processes?

Do not use	Do not use/use	Partially use	Use
30 %	37 %	30 %	3 %

The question "Within which segment of the business procedures the associations are using digital technology? (Table 3) (training processes, financial processes, administrative-management processes), resulted in 53% of the respondents stating that they use digital technology for the purpose of training processes, a smaller percentage of respondents (31%) answered that they use digital technology for the purpose of financial processes, while a significantly smaller number of respondents (16%) stated that they use digital technologies for the purpose of administrative-management processes.

Table 3 - Results for the question - Within which segment of the business procedures the associations are using digital technology?

Training processes	Financial processes	Administrative-management processes
53 %	31 %	16 %

Table 4 – Results for the question - To what extent can the use of an e-application (process digitization) for the day-to-day operations possibly enhance the work of sports associations?

Greatly	*Partially*	*Insignificantly*
71 %	*22%*	*7%*

By observing the presented data, it becomes clear that associations are not satisfied with the "paper-based" approach to collecting data and that they tend to use digital processes. However, it is clear from the respondents' answers that most of them are currently not sufficiently using digital technologies in their everyday work, and that only a small number of associations does so. The questions related to the segments in which digital technologies are being used have clearly defined that in the segment of administrative and management processes, associations are clearly lagging behind modern standards, or more precisely, that they using primarily modern tools in the training process. Through the question concerning the use of both digital technologies and the improvement of the work of an association, the respondents significantly determined that the use of non-introduced technologies can improve the work of associations (71%), or more precisely, they recognized the importance of modern technologies for the improvement of the work process of the associations. Based on the conducted research, a number of topics were raised through the questions used in the questionnaire that should be addressed in the future (e.g., educating the association employees about digital processes, sport communication processes within associations). The survey questionnaire examined the different attitudes and opinions of the respondents and determined the levels of use of digital technologies, as well as the areas in which they were used. In accordance with the set goal and the results of the questionnaire, we can confirm the hypothesis that the sports associations in the City of Zadar do not sufficiently use information technology for the purposes of managing the administrative and management processes of the various associations.

IX. CONCLUSION

In conclusion, in order for the organizations, even those within the field of sports, to be successful globally, they need to embrace technological changes and the benefits of a digital transformation. To achieve this the organizations and their leaders must use processes that they do not control, but instead use the benefits of the digital technology tools and the acquired knowledge of the employees of an association. It is clear from the previously presented arguments that a comprehensive and global view of technology and its capabilities must be adopted. The area presented in the paper represents a golden niche for the use of appropriate modern digital systems and tools. In the City of Zadar, 124 representatives of associations, which make up 89% of active sports associations which are part of the Sports Community of the City of Zadar, clearly confirmed this with the presented research results. This is confirmed in this paper through the use of an e-application as a tool that represents the simplest and most complete solution for the work of a sports association, that is, the management process at the administrative level. Continuous training, acquisition of new skills and knowledge is equally important for a successful sports organization, as well as for the digital system user himself/herself, or in generally, for the use of modern information technology in the field of sports. Therefore, it can be emphasized that the digitalization within the context of sports associations is being lived without actually being aware of it.

REFERENCES

[1] Bolfek, B.; Milković,V; Lukavac, M. (2017). Utjecaj interne komunikacije na zadovoljstvo zaposlenika radnim mjestom. Oeconomica Jadertina 1/2017, Zadar: Sveučilište u Zadru.

[2] Domitrović, I. (2016). Digitalna transformacija – Slijede velike promjene, https://dnevnik.hr/vijesti/hrvatska/digitalna-transformacija. -444413.html, accessed: [August 24th 2018]

[3] Dobrinić, D.; Gregurec, I., (2016). Integrirani marketing. Fakultet organizacije i informatike, Varaždin. [3] Dörner, K.; Edelman, D.

[4] Dugalić, S. (2018) Sport, mediji i digitalizacija, SPORT - Nauka i Praksa, Vol. 8, № 1, 2018, pp. 56-69 Beograd: Fakultet za fizičku kulturu i menadžment u sportu.
[5] Oztemel, E.; Gursev, S. (2018). Literature review of Industry 4.0 and related technologies. Journal of Intelligent Manufacturing. https://doi.org/10.1007/s10845-018-1433-8 accessed: July 2nd 2019.
[6] https://www.e-udruge.eu
[7] Kesić, T.; Vlašić, G., (2018). Marketinška istraživanja i predviđanje potražnje (Chapter 4); Stvaranje zadovoljstva, lojalnosti i vrijednosti za kupce (Chapter 5) http://www.efzg.unizg.hr/UserDocsImages/MAR/gvlasic/Predavanje%205-6.pdf,
[8] Krämer A.; Tachilzik T.; Bongaerts R. (2017). Technology and Disruption: How the New Customer Relationship Influences the Corporate Strategy. In: Khare A., Stewart B., Schatz R. (eds) Phantom Ex Machina. Springer, Cham
[9] Milanović, Dragan. *Strategija razvoja hrvatskog športa.* // Olimp 4(2000). URL: http://www.hoo.hr/downloads/olimp%204%20-%20cijeli.pdf (2017-01-17)
[10] Milanović, Ljubica. Upravljanje poslovnim procesima i znanjem primjenom informacijske tehnologije u hrvatskim poduzećima: specijalistički poslijediplomski rad. Zagreb: Ekonomski fakultet, 2009. URL: https://bib.irb.hr/datoteka/448612.final.pdf (2017-01-22)
[11] Novak, Mijo; Sikavica, Pere. *Poslovna organizacija*. Zagreb: Informator, 1999.
[12] Pihir, I.;Križanić, S.; Kutnjak, A. (2018) Digitalna tranformacija marketinga u malim i srednjim poduzećima – pregled postojećih istraživanja, Varaždin: CRODMA.
[13] Stančić, H. (2009.) *Digitalizacija,* Zagreb: Zavod za informacijske studije.
[14] Stolterman, E., Croon Fors, A. (2004). Information Technology and the Good Life. Information Systems Research, Relevant Theory and Informed Practice. Springer, Boston, MA. Umeå University, pp 686-692, 2004.
[15] Schereiber, M. (2017). Companies in digital transformation. Elektrotechnik und Informationstechnik, 134(7), pp 197-2013,2017.
[16] Watson, H., & Rodrigues, R. (2018). Bringing privacy into the fold: Considerations for the use of social media in crisis management. Journal of Contingencies and Crisis Management, 26(1), 89-98.
[17] White, C., Vanc, A., Stafford, G. (2012) Internal Comunication, Information Satisfaction and Sense of Community, The Effect of Personal Influence, University of Tennessee.

Youth Perception on 5G Networks in Osijek

Davorin Turkalj, Ph.D*, Ivan Kelić, Ph.D**, Juraj Rašić, MA***
* Associate Professor, J. J. Strossmayer University of Osijek, Faculty of Economics in Osijek, Osijek, Croatia
** Assistant Professor, J. J. Strossmayer University of Osijek, Faculty of Economics in Osijek, Osijek, Croatia
*** PhD Student, J. J. Strossmayer University of Osijek, Faculty of Economics in Osijek, Osijek, Croatia
davorin@efos.hr, ivan.kelic@efos.hr, juraj.rasic@efos.hr

ABSTRACT

For cities to function correctly, many systems and infrastructure, such as water, electricity, education, healthcare, information systems, communication technology, etc., are needed to increase the quality of life in the community. In today's world of advanced technologies, fast and complex changes in the environment, larger and developed cities, and even a few smaller ones, aspire to well-designed systems and infrastructure to get the epithet of a smart city.
The development of technology and globalization has enabled the development of this type of city. Smart cities began to develop as society realized the importance of new technological, organizational, and logistical solutions aimed at improving the quality of life in them. Smart cities are fundamental to the development of society and the economy in the coming period, as all activities related to digital technology will work incomparably faster thanks to the development of the new mobile network of the 5th generation.
The introduction of the new 5G network is considered to be one of the essential prerequisites for creating a smart city and an opportunity for infrastructure development and construction of the first 5G network in the Republic of Croatia, founded by the European Action Plan "5G for Europe", was given to Osijek. Considering the fact that Osijek will become the first 5G city in the Republic of Croatia, this paper demonstrates the research performed among the members of the "Z generation" of Osijek with the main goal of examining their views and familiarity with the new 5G network.

Key words: ***Smart city, 5G network, development***

I. INTRODUCTION

The general acceptance of technology at all levels of business activities is accompanied by the need for faster information flow.
The Internet, as a 21st century phenomenon relying on modern technology, has enabled users to search and retrieve information in real time, send multimedia recordings, and embrace technology as a complement to their daily activities. The result is the rapid development of devices that require a more technologically advanced infrastructure that enables the implementation of these activities. Technological advancements are visible in all spheres, mostly in the daily use of applications, platforms, and software that allows users to perform everything from the simplest of tasks to the complex tasks associated with the job they do. Since technological application has become an everyday activity, the trend is to develop network infrastructure in urban cities - cities that allow their residents, with the help of technology, to stay more comfortable and live in the city by enhancing certain activities like power/water/waste/transport management by giving the aforementioned activities their independent services and applications. In order to develop the so-called Smart cities, it is essential to create the prerequisites for infrastructure development - wireless internet technology, or the 5G network. Based on the "5G for Europe" Action Plan, which contains guidelines and recommendations for promoting the development of fifth-generation mobile telecommunications infrastructure and services by 2020 [1], the City of Osijek has been selected as the first city in the Republic of Croatia to build a 5G network and commission commercial 5G services. Since the application of 5G technology in urban areas is in its infancy, and there has been no research done on the effects of the 5G network, the authors aim to understand the potentials of the application of 5G technology in economic terms and what is the perception of this topic for the "Z generation" as the largest consumers of online content about adopting said technology.

II. LITERATURE REVIEW

The evolution of telecommunication technologies in the last four decades has advanced at a rapid rate.
The fifth-generation (5G) wireless network, as announced, is set to roll out in 2020. As the next generation of wireless internet technology, 5G will

provide a better user experience and improve business utility with faster data transfer and more reliable connectivity.
According to the European Commission, the new 5G network will be noticeably faster than the existing 4G network, making the devices more efficient, and the network itself will provide greater reliability and higher availability.



Figure 1 - Smart city with 5G

As it can be seen in Figure 1, [2] using the 5G network, there will be a development in artificial intelligence, and widespread use can be expected in industry, agriculture, health, education, as well as in building new smart city infrastructure, thus improving the way public services function [2].

The introduction of the 5G network is one of the essential prerequisites for creating smart cities. Smart cities can be defined as a high-tech intensive, and advanced city that connects people, information, and city elements using new technologies to create a sustainable, greener city, competitive and innovative commerce, and increased quality of life [3]. In order for the city to develop, it is necessary to adapt the ICT infrastructure that is visible through the development of key areas: Smart mobility, Smart living, Smart economy, Smart environment, Smart governance and Smart citizens. It is about integrating solutions, providing interoperability and cohesion among systems within a city in full collaboration with all stakeholders involved [4]. Some authors consider that one of the key elements that stands out in the smart (intelligent) city is the utilization of networked infrastructures to improve economic and political efficiency and enable social, cultural, and urban development [5].

Cities around the world are pushing for a transformation of technology to improve the environmental, financial, and social aspects of urban life. It is necessary to point out that smart cities must integrate technologies, systems, services, and capabilities into a cellular network that is sufficiently multi-sectoral, flexible for future developments, and open-access. [6] Information and Communication Technologies (ICT) are crucial for the successful implementation of Smart Cities. The scope of application of 5G technology is diverse. The areas where there is significant progress and benefit is undoubtedly the automotive sector, which is thought to use 5G technology for cooperative automatic driving. The way it will work is that thousands of vehicles will exchange real-time traffic data.

With 5G technology, cars will be connected and use a safety application that will be able to see road conditions well ahead of vehicles, which will undoubtedly reduce the number of accidents and increase driver safety. Apart from the automotive sector, 5G technology will prove useful for managing assets and interventions in hospitals, as well as enhancing the quality of experience of surgeons using operating robots. Due to the super-fast data transfer using the 5G network, surgeons will be able to control the robot and perform movements of the surgeon who performs the virtual operation remotely.5G could also be useful for time-critical process control, factory automation, remote control and enterprise communication, and related goods, thus creating smart factories, which would be of great importance to all robotics-based manufacturing processes. Energy companies would use 5G to balance demand and supply in the future, at a time when there is an increase in the amount of energy being generated through solar panels and using wind, as a balancing act for unpredictable demand for electric vehicles. But apart from the great benefits for the various industries, significant changes and advances will be experienced by the media and entertainment sectors. 5G will be used for Ultra High Fidelity media, integrated media, cooperative media production and shared gaming. You could play virtual reality-based games on all continents or watch football games in 3D from the living room, but with the experience of being in a stadium.

According to a European Commission report, the 5th Generation of Communication Infrastructure (5G) is expected to support the digitalization of economic and social activities worldwide. The European Commission strongly supports the international establishment of a new communications structure that is seen as a future of telecommunications and computer infrastructure that will connect billions of users. International collaboration is a critical instrument for reaching a global consensus on 5G vision, standards, and spectrum requirements. In this regard, the European Commission has signed joint statements on 5G with Brazil, China, Japan, and South Korea. Cooperation is also well established with India and the United States, and informal targeted cooperation is ongoing with Taiwan.



Figure 2 - 5G International Cooperation

As shown in Figure 2, [7] the Joint Declaration on Strategic Cooperation in Information and Communication Technology (ICT) and 5G was first signed with South Korea in June 2014, then with Japan in May 2015 and China in September 2015, ending with Brazil in February 2016.

III. YOUTH PERCEPTION ON 5G NETWORKS – SITUATION AND PERSPECTIVES

By the request of the European Union to have at least one city covered by super-fast internet by 2020, the Government of the Republic of Croatia has decided to declare Osijek the first Croatian 5G city by December 31st 2020. This initiative was launched in synergy with the Slavonija, Baranja and Srijem project, which is intended to contribute to the development of the eastern part of Croatia.

The Slavonija, Baranja and Srijem project presents an initiative by the government of the Republic of Croatia, the goal of which is to provide support to the economic growth of the five counties in Slavonia using the funds from European structure funds and the budget. The Ministry of Regional Development and European Union Funds is in charge of the coordination of the project, and over ten ministries are taking part in it. The role of the World Bank Group in the project is to ensure technical aid as support to the enforcement of the main project initiatives with a special highlight on the following: the diagnostic of opportunities for new investments, export and new work places; a base of planed infrastructure projects and determining the priority main projects and the strategy, as well as an integrated program of projects and reforms.

The latter is also an introduction into the announced realization of the Ericsson Nikola Tesla research and development center in Osijek in which over a hundred young experts will be employed for the development of the latest 5G networks.

Given the many controversies that have emerged with the news about Osijek being Croatia's first 5G city, the authors conducted a preliminary survey on young citizens' perceptions of the benefits and importance of introducing a 5G network.

The authors questioned the perceptions of young people as one of the groups in society that should be more of an initiator of the use of new 5G technology through the online software Google Forms. The survey was conducted from mid-January until the beginning of February 2020. 150 respondents from Osijek-Baranja County participated in the study since the county was selected to test 5G technology. The authors wanted to find out more about the perception of the youth regarding their readiness to accept 5G technology and also the extent to which they are familiarized with the aforementioned technology. The success of any business subject primarily depends on the consumer and what they are ready to accept and how much they are willing to pay for it [8]. The structuring and conceptualizing of the development and promoting of the product begins with segmentation. One of the variables might be the year of birth i.e. the generation which we can define as a group of people determined by years. In order to be able to do that, one needs to know the age structure of the consumer. The sample for review can be the current and future age structure of the consumers: the WWII/swing generation was born in 1932 or earlier. The baby boomer generation encompasses those born between 1946 and 1964. The Generation X relates to those born between 1965 and 1976. The Millenials generation are those born between 1977 and 1995. The Generation Z encompasses those born between 1995 and 2010. The emerging generation is created after 2010 when there was a rapid decline in newly born children due to unfavorable economic opportunities (adjusted according to Jurin et. al. [9]

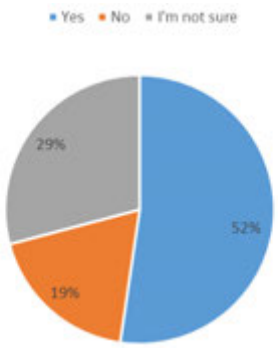
The authors selected a suitable sample - members of the "Z generation." Although different ranking systems define the range of the “Z generation” in different ways, the authors used the following classification for their work, which states that they were born between 1995 and 2000 [10].

This generation group is now entering their key consumer era, consuming online content to the fullest, and embracing technology as a boon.

In technological terms, the mentioned generation is more interconnected and has grown up with technology and is unaware of a world without information technology, unlike previous generations, so they are much more connected. Therefore, it is assumed that precisely this generation will generally be ready to accept contemporary technology. Among the respondents, there is an uneven distribution of male respondents represented by (30.5%) and women represented by (69.5%).

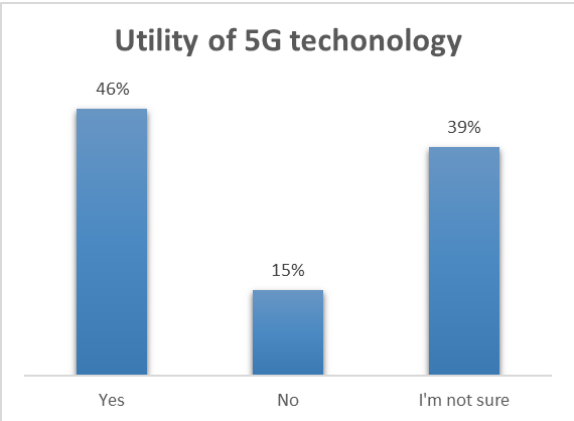
The largest share of respondents (83.4%) reported having completed secondary education, but it should be emphasized that these respondents are mostly students and are in the process of obtaining a university degree.

Consequently, the most represented age groups of the respondents are those aged 18-20 years (75.5%) and 21-24 years (13.2%). Almost all respondents (98.7%) own a smartphone with which they connect to the Internet. The fact that internet access plays a significant role in their daily lives is confirmed by the fact that 40.1% spend 3-5 hours online daily, and 32.7% spend even more than 5 hours online daily.
Although the benefits of using the 5G network have been extensively explored, there is a certain resistance from citizens who are pushing to suspend 5G technology until its long-term biological effects and potential adverse health and environmental consequences are investigated.
The purpose of this preliminary research is to determine the familiarity of the respondents - members of the "Z generation" - with its meaning. Familiarity with the term 5G network is evident from the questionnaire, using the question: "Are you familiar with the term 5G network?"



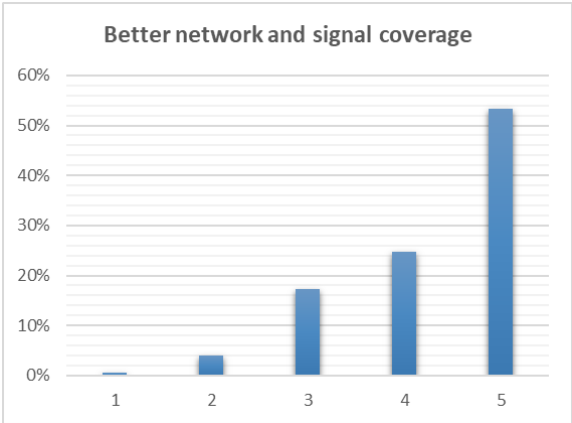
Graph 1 - Familiarity with the term 5G network?

According to the answers in Graph 1, it is evident that 52% of the respondents are familiar with the term 5G network, but also that the other 48% are not familiar with the term 5G network, or they are not sure what it means and represents.



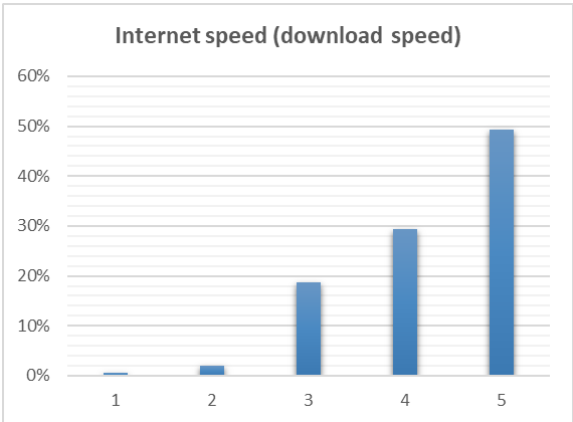
Graph 2 - Utility of 5G technology

When it comes to the usefulness of 5G technology itself, 46% of respondents think 5G technology is useful, 15% believe that the introduction of 5G technology will not be of use to us, and 39% of respondents are not sure about the benefits associated with the introduction of 5G technology, as seen in Graph 2.



Graph 3 - Better network and signal coverage

Respondents were asked to rate the importance of the elements offered when using the Internet in general. Graph 3 shows that among offered elements, respondents chose "network and signal coverage" as the element with the highest priority. A total of 118 respondents rated "network and signal coverage" as an important and extremely important element when using the Internet. The introduction of 5G technology does not mean automatically switching off the 4G network. Moreover, in places where there will be no coverage with the 5G network, it will be replaced by 4G, and in some cases, 3G networks. The synergistic effect of 5G and 4G network availability should, to the greatest extent, result in reliable signal and excellent coverage.



Graph 4 - Internet speed (download speed)

Furthermore, in terms of importance, as seen in Graph 4, "internet speed" is also highly rated. Specifically, 81 respondents rated internet speed as extremely important when using the Internet. The most common association with the introduction of

the next generation of wireless networks is the upload/download speed.
According to the results of the research, we can conclude that internet speed is important to the respondents since internet speed is a prerequisite for using certain activities on the media - such as downloading multimedia, video streaming, faster data loading, etc.
Part of the results of a study conducted by PwC shows factors that would affect respondents' willingness to pay more to their telecom provider to switch to a 5G network. Poor readiness in the home internet category can be attributed to the insufficient awareness by the respondents of the fact that the 5G network covers the home internet as well.

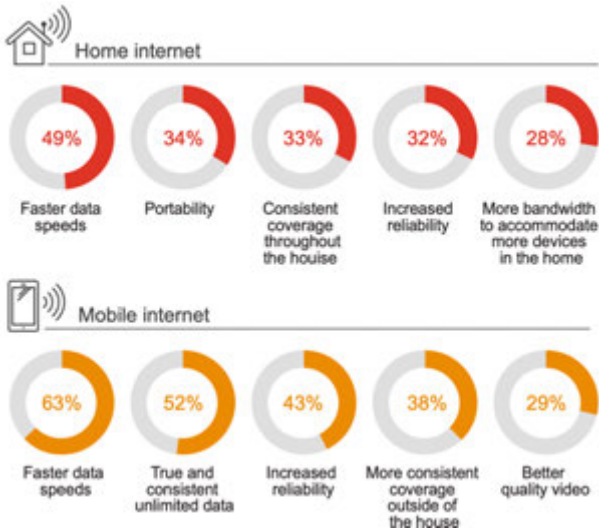


Figure 3 - Factors that make respondents willing to pay a higher price for a service based on 5G technology

Comparative analysis with research conducted in September 2018 is interesting. On a sample of 1,000 U.S. citizens aged 18-64 years, familiarity with 5G was found in 46% of respondents. [11]
The segment from the abovementioned research is shown in Figure 3 [11] where similarities are observed in the data collected and in the sample from Croatia. Speed, reliability, and coverage are the most important factors in adopting new technology, and as a result, one can conclude a willingness to pay a higher cost of service if these elements are delivered.

IV. CONCLUSION

According to the guidelines of the European Union Action Plan "5G for Europe", the goal is to fully implement the infrastructure expected from 2020 onwards, which means developing infrastructure and associated applications and platforms for the development of Smart Cities. Global revenues for 5G technology in 2025 are estimated at € 225 billion, and the development of this technology is crucial to Europe's competitiveness in the global market. We can conclude that 5G technology has great potential that is still not recognized in the minds of potential users. The project to introduce 5G technology will certainly raise the quality of life and enable the development of Smart Cities, but it is essential to educate users about the potential of implementing the technology.
The research shows that the Z generation members, who are mostly pioneers in the use of technology, are not sufficiently familiar with the technology. Most respondents are not sure what potentials this technology brings, but they are ready to accept the technology if they gain faster and easier access to information. The question remains whether the end-users and large systems (e.g. smart cities) will be willing to invest in devices that support 5G technology. It is up to the business entities to create content that will motivate users to embrace the technology or models of information distribution.
The fact that cities must take into consideration is different age groups. A limitation of the paper is that the research was conducted on only one age group (generation) who are adopting much more modern technology. It is the author's wish to extend the research in the future to other generational groups to gain an insight into what items are essential to them when using technology, and whether they are willing to start using 5G technology if their use of the abovementioned entities improves. It is important to note that younger users are willing to put in the effort and resources to embrace new technology. Since the "Z generation" is considered the future of the global economy and is the target group of most businesses, the development of applications and platforms for smart cities will certainly turn to the needs of that generation. The bottom line is that 5G technology has been one of the main topics in the digital industry and other industries over the past two years, and the whole world is striving to develop infrastructure and media that use the infrastructure in a way that enhances the benefits and corrects the disadvantages to make the systems work better.

REFERENCES

[1] European Commision (2020.) 5G europe action plan [Online] Available at: https://ec.europa.eu/digital-single-market/en/5g-europe-action-plan

[2] European Commision (2020.) What can we expect from 5G technology [Online] available at: https://ec.europa.eu/croatia/content/what_can_we_expect_from_new_5G_technology_hr, (accessed : 5.2.2020.)

[3] T. Bakıcı, E. Almirall, and J. Wareham, “A Smart City Initiative: The Case of Barcelona,” Journal of the Knowledge Economy 2: 1 (2012) 1–14.

[4] Rao, Sriganesh K., and Ramjee Prasad. "Impact of 5G technologies on smart city implementation." Wireless Personal Communications 100.1 (2018): 161-176.

[5] EGER, JM. "CYBERPLACE: Building the smart communities of tomorrow." I Way 21.1 (1998): 33-38.

[6] Albino, Vito, Umberto Berardi, and Rosa Maria Dangelico. "Smart cities: Definitions, dimensions, performance, and initiatives." Journal of urban technology 22.1 (2015): 3-21.

[7] European Commision (2019.) 5G international cooperation [Online] Available at: https://ec.europa.eu/digital-single-market/en/5G-international-cooperation

[8] Brown, S., Foxall, G., Goldsmith, R. (2007). Psihologija potrošnje u marketingu. Jastrebarsko: Slap.

[9] Jurin, E., Gorup, T., Karuza, I., & Marić, P. (2016). Priručnik o doživljajima. *Zagreb, Hrvatska turistička zajednica*.

[10] Maas, Judith. "Generations at Work: Managing the Clash of Veterans, Boomers, Xers, and Nexters in Your Workplace." MIT Sloan Management Review 41.2 (2000): 98.

[11] PwC Consumer Intelligence Series (2018) 5G survey. [Online],Available at: :https://www.pwc.com/us/en/advisory-services/publications/consumer-intelligence-series/promise-5g.pdf , (accesed 5.2.2020.)

Digitalization Impacts on Innovation Networks: Policy and Estimations Issues

O. Kudrina[*], V. Omelyanenko[*, **], H. Shevtsova[**], V. Samoday[*], Yu. Mashyna[***], and V. Bilyk[****]
[*] Sumy State Pedagogical A. S. Makarenko University, Sumy, Ukraine
[**] Institute of Industrial Economics of NAS of Ukraine, Kyiv, Ukraine
[***] Kyiv National Economic Vadym Hetman University, Kyiv, Ukraine
[****] Cherkassy National B. Khmelnitsky University, Cherkassy, Ukraine
omvitaliy@mail.com

***Abstract* – The research deals with the digitalization impact on the innovation networks development. Different issues of U2B-B2U cooperation digitalization within the innovation networks development were considered. The main practical recommendations for digitalization of innovation networks were proposed. Suggestions for estimations of digitalization impacts on innovation network were made.**

Keywords – digitalization, innovation network, policy, technology transfer

I. INTRODUCTION

All sectors of the economy are undergoing a digital transformation, which will become a driver for changing existing educational institutions and creating new ones.

All new trends inexorably capture the educational market and change it. Trends in the implementation of digital technologies in educational and research activities are set by commercial organizations (private universities, business schools, corporate universities etc.). But public universities and institutions are starting to think more and more about digital transformation. It is important for them to prepare, compete with the wave of educational startups that emerged with the digital revolution. The formation of an online environment is an area that still requires serious efforts from universities. But in the modern conditions the areas of information and communications technologies ICT application are expanded within the universities networked strategies. Modern trends do not even give clear idea of what the university will look like in 20 years, how the campuses or laboratories of universities will look like, how much education and research activities will move into the virtual reality.

In our study we will discuss the issues of an economic and organizational nature in order to understand what new university models look like now within the U2B – B2U (Universities to Business – Business to Universities) cooperation, what are the prospects and future steps needed to reform a modern university and create fundamentally new models of higher education based on ICT. This objective requires considering the limitations and possibilities of new university models as part of network strategies.

II. LITERATURE REVIEW

Modern innovation transformations require focus on universities, which are open to contestation concerning their role in public sphere and the politics of knowledge production. Where universities are discussed, higher education is focused on, because this is undergoing an intense audit-driven neoliberal re-engineering [1].

It's common knowledge that higher education, science and technology policy discourse is a key point of innovation system development [2].

Research activities have expanded research-intensive academic departments [3]. It relates to existing research on organization in academic leadership in many ways [4].

The paper [5] develops an analytical framework for analyzing the factors of global flows in higher education. Here we can note that attention would be primarily given to issues if Industry 4.0 impacts on higher education, including the problems of rapid learning and innovations development. Therefore one of the most striking features of this problem is re-assembling knowledge production in university [6]. Industry 4.0 revolution causes the transformation of social and economic systems [7] and higher education institutions (HEIs) can be considered as a source of innovations.

Research [8] deals with the problem of education space that supports the learning needs of students. In the same in study [9] particular attention is given to distance education. It is clear from these observations that higher education should be based on innovation management and cooperative marketing. A number of key issues arise from the statement. For instance, in research [10] diverse aspects of universities influence on the development of the society is considered. E.g. it includes countless smart city and community initiatives. These cases bring together city planning, industry, universities and citizens to improve the urban life of individuals and organizations with the integrated use of versatile information, digital and communication technologies. Within these issues the quality management system in universities, which comprehensively covers education, research and social collaboration activities, and which can ensure the universities' successful partnership in smart city projects, was studied.

In [11] an analysis of three challenges to established understandings of higher education: digitalization, commodification and precarity were proposed. Universities remain the most important organizations involved in developing knowledge and providing means

of social mobility.

The study [12] aims to identify and analyze factors that determine the implementation of Information and Digital Technologies of smart manufacturing. Digitalization enables administrative solutions, data security systems, research data storage and learning resources, as well as opportunities for better collaboration [13].

From these facts, we may conclude that digitalization provides opportunities for the exchange of accumulated experience and knowledge, which allows different stakeholders to learn more and make more informed decisions. ICT support of higher education institutions participation in innovation networks [14; 15], including international networks [16] is important issue. Examples of such ICT support can be found in nanoinformatics [17] application for R&D and industrial analysis.

The given points confirm the idea that strategies and approaches for ICT infrastructure management substantiation is a complex task, proposed by authors in previous studies [18]. Therefore the objective of this study is to consider the issues of ICT support of cooperative marketing of HEIs in innovation networks by indicating some strategies and approaches for ICT management.

III. Universities Networked Strategies Smart Environment Levels

Based on experience of leading universities and business schools [19], it is possible to form a conceptual model of smart environment for universities networked strategies, which consists of five levels.

The first level is the most important and is represented by scientific and pedagogical workers, students, industry and academic partners of the university, graduates and applicants. The first level is, in fact, the internal and external stakeholders of the university.

The second level is represented by basic information services. Their task is to create a single information space for digital interaction within the university using flexible tools. Examples of such services are video screens for lectures and seminars, wireless communications throughout the university (including dormitories), cloud storage for storing and exchanging data, professional printing, etc.

The third level includes services that greatly facilitate the lives of students and teachers in a modern university. For foreign teachers and students, they are already an obligatory element of the university; a number of services in most universities are still at the initial stage.

The fourth level is the most resource-intensive in terms of implementation, but at the same time it allows the university to get the highest added value. It consists of services such as digital marketing, research project management, procurement management, interaction with applicants and students.

The fifth level consists of digital technologies, which are highly likely to be widely used in the university environment (virtual reality, network labs etc.).

Advanced levels of smart environment for their development require appropriate infrastructure, and smart environment.

IV. Universities Networked Strategies Smart Environment Policy Recommendations

According to modern innovation policy approach, to ensure the educational services competitiveness, HEIs should implement appropriate quality assurance procedures, develop innovation management policies, standards of practice-oriented educational programs, form and implement a strategy for continually improving the quality of education through R&D interaction with business and society. At the same time, the innovation management in modern HEI should include the digitalization the main directions of innovation activity.

Management of formation and practical application of intellectual resources becomes a key task for HEIs in the framework of new role in development. Therefore, ICT tools that are involved in the creation of intellectual resources are largely decisive. Experience of number of leading world universities and expert's reviews confirm that critical thinking and communicative skills are the two core competencies of Industry 4.0.

As a consequence, the role of the communication component of the HEIs interaction with its environment is growing, which forms relevant requests for educational strategies and can facilitate the application of various academic resources for the purposes of socio-economic development. In the context of the foregoing it is expedient to consider the peculiarities of strategic role of academic institutions, in particular in the context of educational and research missions. The analysis shows the practical absence of system processes in research and social missions, which leads to the loss of development potential and impairs the implementation of an educational component that is not related to practice.

For the purposes of ICT tools role analysis we consider the innovation networks. Today, innovative networks are actively involved in modern business processes, offering effective ways to implement complex projects. The growing number of multifunctional networks offers high availability, security and service quality for business processes. But at the same time business and public sector face with a number of challenges related to development of network infrastructure and introduction of ICTs, because digital transformation leads to a number of problems (especially security problems) and opportunities (formation of interaction synergy).

From an innovative point of view, we propose to consider the role of ICT within a systematic approach. At Stanford, the innovation system is defined as inter-organizational, political, economic, technological and non-military systems, through which a favorable environment for business development is formed, supported and developed. From this definition we can define four main components of an innovation system:

1. Purposeful process: the innovation system is built around the process of commercialization of innovations, which has a clear result in the form of innovative products in the market;

2. Subjects interacting with the process of commercialization of innovations;

3. Relationships between them;

4. An environment that provides resources for subjects of commercialization of innovations.

Therefore the strategic development of the HEIs networked-based innovation system is a factor that equally contributes to both increasing the scientific reputation and its financial sustainability, attracting students, new employees and innovative resources. To do this, we need to orient the system to specific commercial outcomes.

Specialized ICT offer new platforms (applications) for interaction and integration the educational & research processes. Domestic and international R&D cooperation, which becomes effective through modern ICT tools, is crucial for solving modern research problems. In general, collaboration tools and services create new opportunities for future research challenges overcoming.

It can be accurately noted that individual ICT solutions, network infrastructure and scalable solutions will play major role in HEIs R&D and technology transfer. Such ICT can be considered as a key for educational processes reengineering. Thus, the higher education system digital transformation is important to maintain competitiveness in the era of Industry 4.0.

The role of innovation networks is growing significantly as future education reduces the distance between employer and employee, and the educational services market will increase its degree of flexibility. As a result, ICTs in innovation and education networks will participate in the following education and R&D areas:

- large corporations will be able to order certain competencies (projects) without mediation of specialized government or private firms, including project approach (learning through real R&D projects);
- organizations that are in particular in need of collective intelligence and team building can enter the educational market themselves and offer specialized courses or services;
- carriers of new interactive educational technologies will be able to combine into the virtual professional communities and to form the community colleges with their own standards and teaching methods. Subsequently, these structures can compete with traditional universities;
- to implement a model of future education, a high-tech infrastructure is needed that will be developed based on up-to-date databases based on Big Data or cloud technologies;
- there are prototypes of stock exchanges for educational opportunities for joint and individual learning, platforms for educational content stores;
- mass attestation tests to determine the overall educational level will gradually lose relevance. They will be replaced with online evaluation and certification systems that allow for an external, verified assessment of the availability of individual competencies and skills.

In these conditions, the educational sector needs to achieve significant progress in digitalization and the use of leading innovations in ICTs in the implementation of existing and new innovation tasks (Fig. 1).

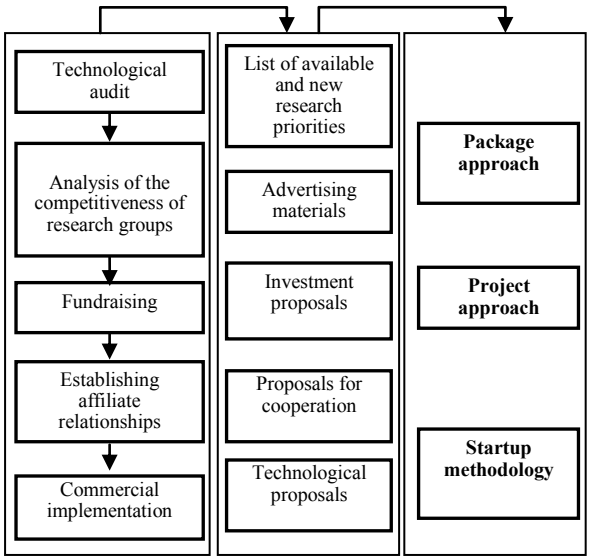


Figure 1. ICT application framework in HEIs innovation system

Thus, the role of ICT for the development of innovation and educational processes of modern universities can be defined in the following areas:

- creation of educational programs based on unified scientific-educational and innovative process with application of interdisciplinary problem-oriented and project-oriented educational technologies;
- organizational structure and project management methods, which are appropriate to the new task, have been developed;
- diversified sources of HEI funding and effective system of crowdsourcing;
- increase in the number of cooperation projects and practical oriented scientific and educational services;
- developed infrastructure of interaction of HEI with the external environment, which includes agents from the regional, national and international dimensions.

In each of these areas the role of ICT through the special indicators can be considered. So we came to two aspects of information aspect indicators: (1) availability and quality of the information and (2) readiness and stimulus for using the information.

For development of a smart environment appropriate for networked strategies purposes we propose to consider the elements of system design methodology [20; 21]. Among other things it includes a comprehensive analysis and evaluation of the customer's task, goals and objectives of complex technical systems, the formation of a source

data system for system design, the development of an interconnected system of criteria and indicators for innovation and its components. Rational Unified Process (RUP) model supports the basic principles of system design and Model-Driven Systems Development (MDSD), which is used both for education and R&D.

Virtual reality technology, e.g. IC.IDO, also has significant potential, which, in combination with engineering systems for product design, is used both by large international corporations and medium and small companies to increase the competitiveness of products on the market. The IC.IDO system is a model of a real product with all its specific characteristics and behavior in real conditions, thereby providing the most accurate teaching aid for conducting trainings and advanced training courses for academic staff.

Such methodological foundations can solve the problem of ICT support of universities R&D strategies.

V. CONLUSION

In research it was concluded that universities should perceive the use of new educational technologies as a tool to increase their own competitiveness within the innovation networks.

A general conclusion is made concerning that processes of testing and implementing such technologies develops the university's digital educational system but at the same it requires individual strategy that should corresponds with R&D strategy.

As a result of models of ICT support of universities R&D strategies considering the main issues and cases of ICT application in universities for technology transfer management thought the creation of smart environment of HEIs researches programs for business were outlined. It is shown that ICT application framework in HEIs innovation system should be based on package and project approaches and startup methodology.

Examination of some cases has produced a more complete understanding of ICT-based university networked strategies. Discussion of the present findings has confirmed the smart environment for networked strategies of HEIs not only helps to form unique innovation competencies, but at the same time expands the creative opportunities for teachers and researchers.

ACKNOWLEDGMENT

The publication was publicly funded by Ministry of Education and Science of Ukraine for developing of research project № 0117U003855 «Institutional and technological design of innovation networks for Ukraine national security systemic providing», research project № 0119U100179 «Development of scientific and methodological foundations and practical tools for evaluating of the product innovations commercial (market) prospects».

REFERENCES

[1] J. Cruickshank, and R. Abbinnett, "Neoliberalism, Technocracy and Higher Education: Editors' Introduction," Social Epistemology, vol. 33, iss. 4, pp. 273–279, 2019.

[2] T. Nokkala, "Making the case for policy – persuasiveness in higher education, science and technology policy discourse," European Journal of Higher Education, vol. 6, iss. 2, pp. 144–158, 2016. DOI: 10.1080/21568235.2015.1086664

[3] G. Gibbs, C. Knapper, and S. Piccinin, "Disciplinary and contextually appropriate approaches to leadership of teaching in research-intensive academic departments in higher education," Higher Education Quarterly, vol. 62 (4), pp. 416–436, 2008.

[4] C. E. Tømte, T. Fossland, P. O. Aamodt, and L. Degn, "Digitalisation in higher education: mapping institutional approaches for teaching and learning," Quality in Higher Education, vol. 25, iss. 1, pp. 98–114, 2019.

[5] S. Marginson, and E. Sawir, "Interrogating global flows in higher education," Globalisation, Societies and Education, vol. 3:3, pp. 281–309, 2005. DOI: 10.1080/14767720500166878

[6] C. Muellerleile, and N. Lewis, "Re-assembling knowledge production with(out) the university," Globalisation, Societies and Education, vol. 17, iss. 1, pp. 1–10, 2019.

[7] L. Melnyk, O. Kubatko, I. Dehtyarova, O. Matsenko, and O. Rozhko, "The effect of industrial revolutions on the transformation of social and economic systems," Problems and Perspectives in Management, vol. 17 (4), pp. 381–391, 2019.

[8] B. Taggart, "Modern higher education students within a non-traditional higher education space: not fitting in, often falling out," Research in Post-Compulsory Education, vol. 21:1–2, pp. 86–97, 2016. DOI: 10.1080/13596748.2015.1126420

[9] O. Zawachi-Richter, and S. Naidu, "Mapping research trends from 35 years of publications in distance education," Distance Education, vol. 37 (3), pp. 245–269, 2016.

[10] J. Anttila, and K. Jussila, "Universities and smart cities: the challenges to high quality," Total Quality Management & Business Excellence, vol. 29:9-10, pp. 1058–1073, 2018.

[11] J. Holmwood, and Ch. Marcuello Servós, "Challenges to Public Universities: Digitalisation, Commodification and Precarity," Social Epistemology, vol. 33:4, pp. 309–320, 2019. DOI: 10.1080/02691728.2019.1638986

[12] M. Ghobakhloo, "Determinants of information and digital technology implementation for smart manufacturing," International Journal of Production Research, 2019. DOI: 10.1080/00207543.2019.1630775

[13] J. Khalid, B. R. Ram, M. Soliman, M. Khaleel, and M. S. Islam, "Promising digital university: a Ali pivotal need for higher education transformation," International Journal of Management in Education, vol. 12, iss. 3, pp. 264–275, 2018.

[14] T. Tirto, V. Omelyanenko, and Y. Ossik, "ICT Support for Industry 4.0 Innovation Networks: Education and Technology Transfer Issues," Lecture Notes in Mechanical Engineering. Springer, Cham, 2020. DOI: 10.1007/978-3-030-22365-6_36

[15] O. Prokopenko, V. Omelyanenko, T. Ponomarenko, and O. Olshanska, "Innovation networks effects simulation models," Periodicals of Engineering and Natural Sciences, vol. 7, no. 2., pp. 752–762, 2019. DOI: 10.21533/pen.v7i2.574

[16] O. Prokopenko, Yu. Eremenko, and V. Omelyanenko, "Role of international factor in innovation ecosystem formation", Economic Annals-XXI, vol. 3-4 (2), pp. 4–7, 2014.

[17] V. Omelyanenko, and D. Volodin, "Nanoinformatics application framework for R&D and industrial analisys," 2017. DOI: 10.1109/NAP.2017.8190183

[18] O. Prokopenko, O. Kudrina, and V. Omelyanenko, "ICT Support of Higher Education Institutions Participation in Innovation Networks," CEUR Workshop Proceedings, vol. 2387, pp. 466-471, 2019.

[19] Sidorov G., "Digital University: the use of digital technology in modern educational institutions," ITWEEK, 2017. https://www.itweek.ru/idea/article/detail.php?ID=192831

[20] New high technology in technology: Encyclopedia. Part 21. Cosmonautics and solving the problems of the development of civilization in the 21st century. Moscow: ENTSITEH, 2002.

[21] M. Cantor, Rational Unified Process for Systems Engineering, IBM Rational Software, 2003.

A Literature Review of Digital Transformation in Healthcare

L. Ivančić*, LJ. Milanović Glavan * and V. Bosilj Vukšić*
* University of Zagreb, Faculty of Economics and Business, Department of Informatics, Zagreb, Croatia
livancic@efzg.hr; ljmilanovic@efzg.hr; vbosilj@efzg.hr

Abstract - **Digital transformation has reached out to every aspect of human lives, including personal and institutional provided healthcare. In order to inspect changes occurring in the healthcare domain due to digital transformation, a literature review is conducted. A systematic literature search in the Clarivate Analytics' Web of Science database was carried out, along with the text mining in order to analyze the papers. The following attributes are considered for the classification and analysis of the papers: time component; methodological approach; research areas, and; topic identification. Results revealed that the digital transformation of healthcare has just recently begun to induce more interest in the research community. Furthermore, a mixture of scholars' profiles investigates the topic, including engineering and information scientists, as well as medical and healthcare service researchers. Additionally, by using text mining, we have identified four discourses of topics that are addressed in the literature when considering the digital transformation of healthcare. The interdisciplinary nature of the topic calls for the inclusion of diverse experts from the medical, informatics and public administration areas in order to advise proper government actions and stimulate the research field.**

Keywords – digital transformation; digitization; healthcare; literature review; text mining

I. INTRODUCTION

The proliferation of numerous digital technologies such as the Internet of Things (IoT), RFID, artificial intelligence and robotics, brings new business possibilities and revenue streams in numerous sectors if implemented properly and strategically. Digital transformation is a process of change occurring due to the application of these digital technologies [1]. It can be observed at the organizational or the industry level. From the organizational or company point of view, the emphasis is made on how to employ "digital and other technologies along with organizational practices to create a digital culture" [1, p. 36]. When observing the digital transformation trend on the industry level, the appearance of disruptive business models that are changing the established ways of work and government regulations are stressed [2]. The healthcare sector is not an exception in this trend.

Healthcare Information Technology (HIT) is usually associated with electronic health records (EHR). Integration of patients' EHR alongside the entire healthcare system stems from the general practice doctors to hospital treatments and access to pharmaceuticals. This integration is considered to be essential to a better and cost-efficient healthcare system [2]. However, some reports suggest that traditional EHR-based IT systems are barriers to the inclusion of new digital technologies and patients data they provide into the HIT [3]. According to some authors [4], [5], the healthcare system is shifting from the institutional-provided care to the more personal and human-centric approach due to digital technologies. In order to shed some light on this phenomenon, a systematic literature review of the digital transformation of healthcare is conducted.

Hence, the objective of this paper is to investigate the state and progress of research on digital transformation in healthcare. In order to do so, a systematic literature search has been carried out and identified relevant papers have been analyzed according to the following attributes: time component (*Year*); types of papers that deal with the topic (*Methodological Approach*); research areas that cover the topic (*Research Area*), and; topics' identification and categorization (*Clusters of Terms*).

The remainder of this paper is organized as follows. After the introduction, a methodology section explains three steps that have been undertaken in order to address the research objective of this paper. The methodology section is followed by the presentation of literature review results. Next, a text mining analysis reveals the digital transformation in healthcare topics' landscape. In the end, we conclude the paper with a short summary and recommendations for future research directions.

II. METHODOLOGY

The adopted methodological approach for this paper has incorporated three steps. The first step has concerned the identification of relevant literature. After the literature collection, the analysis step has consisted of two parts. The first part of the analysis has been focused on the investigation of the papers according to the three attributes: time component; methodological approach, and; research areas. In the last step, we have employed a text mining. With the aid of the text mining, we have discovered distinct discourses of topics investigated under the umbrella of digital transformation in healthcare. Identification of topic discourses, along with the other observed parameters, has enabled us to shape the recommendations for the future research agenda in the field.

A. *Systematic Literature Search*

A systematic literature search in the Clarivate Analytics' Web of Science (WoS) database was carried out in order to meet the objectives of this paper. Web of Science is one of the most trusted research databases, covering reputational indexes, such as SSCI, SCI-E, and Current Contents Connect database. Hence, it was chosen as a source of outlets for finding the relevant literature.

We have used English for the inclusion criterion since it is a language of mutual understanding in the scientific community nowadays and the majority of papers in respectful databases are written in English. Hence, other authors from the IS field, including healthcare related research, also explicitly state this inclusion criterion in their papers (see for instance [7], [8]). Giving the interdisciplinary nature of the field, the search was not limited to a certain index or a research field. On top of that, we have also included conference papers and book chapters in addition to the journal articles. This search strategy is strongly recommended for literature reviews in the IS domain [6]. Moreover, we haven't limited our search to a certain time period. The adopted approach has enabled us to collect a comprehensive base of papers for the next step, i.e. conducting our analysis. The search strategy for the literature collection was therefore formed as presented in Table I.

The search was conducted in April of 2019. It has resulted in thirty-one papers in total. The authors have inspected the collected papers and came to the agreement that the results of the search corresponded adequately to the researched topic. Therefore, the search results have not been altered, i.e. no refinement of the search results was employed after this content assessment.

After the content assessment, papers have been coded, analyzed and classified in accordance with the chosen attributes, which is the subject of the next section.

B. *Text Mining*

The third step of our research employed text mining in order to investigate the terms and topics discussed in the digital transformation of healthcare literature. In order to do so, identified papers have been analyzed with the assistance of VoSviewer software.

Text mining is a computational technique for extracting insightful information from unstructured and semi-structured data [9]. As such, text mining has gained more attention in the last several years in diverse research areas, including the IS field, where it has been used as a bibliometric analysis technique [10]. Text mining as a bibliometric technique can reveal hidden structure and patterns in text and papers. For instance, it can reveal the research landscape in terms of topics that are studied; the most influential papers and researchers; or the evolution of the field. Hence, bibliometric analysis by text mining is based on the clustering and topic modeling techniques [9].

When performing bibliometric analysis using text mining, researchers usually turn to dedicated software for bibliometric networks visualization [11]. These networks visualize clusters of information according to the chosen parameters and inputs. Hence, we can distinguish several visualization networks. The most commonly used ones are "citation, co-citation, bibliographic coupling, keyword co-occurrence, and co-authorship networks" [11, p. 285].

VoSviewer is one of the software that has the ability to visualize all of the mentioned networks. For the purposes of this paper, a keyword co-occurrence network is created in VoSviewer using abstracts and titles of papers, also referred to as the "term map" [12]. The terms will be joined into a cluster if the distance between the terms in the cluster is smaller than the distance from the terms that belong to the other clusters. The distance of the terms is based on their co-occurrences in the text [12].

TABLE I. STRATEGY FOR THE SYSTEMATIC LITERATURE SEARCH

Search String Component	***Employed Strategy***
Database	Clarivate Analytics' Web of Science
Search terms	TOPIC: ("digital transformation") AND TOPIC: (healthcare)
Paper types	ARTICLE OR PROCEEDINGS PAPER OR BOOK CHAPTER OR REVIEW
Language	ENGLISH
Timespan	All years
Indexes	SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI

III. LITERATURE REVIEW RESULTS

The distribution of the papers according to the *Year* attribute is presented in Fig. 1. A steadily growing trend of research concerning the digital transformation of healthcare is visible since 2015. Only one paper per year can be found before 2016, with zero papers researching the topic prior to 2010. These results are in perfect agreement with the previous research on digital transformation [13] and can be explained by utilizing the "digital transformation" term in the search strategy. Digital transformation is a relatively new research domain. Accordingly, transformations of healthcare occurring due to digitization have only recently started to be more investigated.

In order to inspect what types of papers deal with digital transformation in healthcare, papers have been classified according to the *Methodological Approach*. The scheme

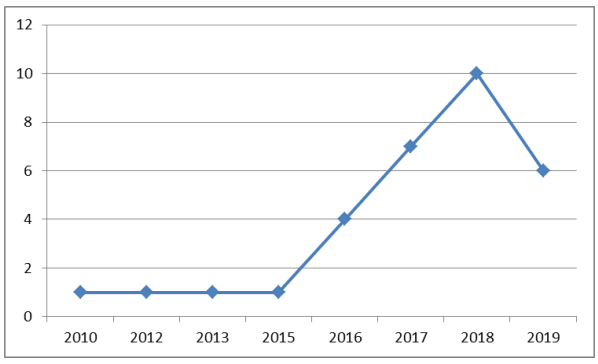


Figure 1. Number of papers over the years

proposed in a systematic literature review paper from the IS domain [14] has been used for the classification of papers. Following this scheme, papers have been divided into three categories, namely: empirical papers (quantitative or qualitative approach); theoretical applied papers, and; literature review papers [14].

Distribution of the papers with respect to the *Methodological Approach* attribute is visible in Fig. 2. Most discovered papers are theoretical papers, in total 17 of them, including opinion papers and proposed engineering frameworks, which count for 55% of the researches. 29% of the researches are empirical papers, applying quantitative or qualitative approach. Out of those, only 2 papers employ quantitative methodology, whereas 7 papers employ some sort of methodology that can be regarded as qualitative. 16% of the researches, i.e. 5 papers in total, are literature review papers. Only one of the review papers employs strict research protocol, including a clear description of the process of relevant literature identification, and thus can be labeled as a systematic literature study [15].

When it comes to research areas that cover the topic, the interdisciplinary nature of the field is manifested. Research results with respect to the *Research Area* attribute are presented in Fig. 3. Three principal types of researchers' profiles can be identified. Firstly, academics from the computer science and engineering field are eminently interested in the topic, followed by medical experts. These medical experts are interdisciplinary scholars under the categories of healthcare service science and medical informatics, fields closely related to information science. Lastly, five papers can be placed in the business economics field. Other research fields are represented to a minor extent. If we observe only the affiliations and roles of the article authors, a strong preponderance of engineering scientists and medical experts is evident.

IV. TEXT MINING ANALYSIS FOR TOPICS' LANDSCAPE

In order to inspect the topic landscape of digital transformation in healthcare, the identification and categorization of terms was made using VoSviewer software.

Classification of papers with respect to the *Clusters of Terms* attribute is presented as a term map and visible in Fig. 4. The term map is created using titles and abstracts of all of the papers identified as relevant in the systematic literature search step. Hence, the links in Fig. 4 are based on the co-occurrence of the items, where the items are relevant keywords extracted from the text [16]. Initially, 74 terms (keywords) have been recognized by the software. After the manual removal of general words, such as "research", "way", and similar, 37 terms remained and were used for the creation of the term map. In this process, four clusters of terms have been identified, and are marked in different colors in Fig. 4. Due to graphical limitations of the software, not all of the terms in clusters are visible in the Fig. 4, i.e. some of the terms that are represented by circles do not have their corresponding written term. Nevertheless, the list of exact terms per clusters can be obtained from the software by zooming-in

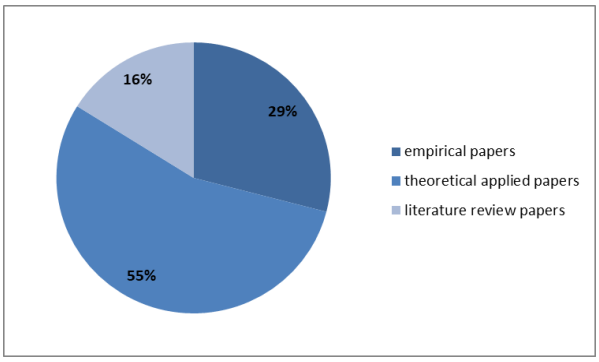


Figure 2. Distribution of papers across the Methodological approach attribute



Figure 3. Research areas of the papers according to the WoS database classification made with the assistance of WoS summary tool

to the particular cluster, or on the Action panel. Accordingly, in the descriptions of the clusters in the following text of this chapter, we have itemized the terms per clusters by putting them into quotation marks.

The first cluster (Cluster 1) is marked by red color and is placed in the center of the term map. It can be described as *Technology and aspects of proper implementation*. "Healthcare data", "personal health record", "IoT", "cloud", and "integration", are some of the terms that are placed in Cluster 1. Other terms in this cluster are "advantage", "key enabler", but also the "privacy concern". This means that although the advantages of technology in healthcare are recognized and discussed, there are still numerous issues mainly concerning the privacy of personal data, and after that proper implementation, enhancement, and integration of new technology and existing healthcare systems.

The second cluster, Cluster 2, is placed in the lower-left corner of the term map and marked in green color. Cluster 2 deals with *Strategic and government issues*. The topic of this cluster is "ongoing digital transformation" and "digitalization" that provides great "potential", as well as "challenges" for the healthcare industry. These potentials and challenges of the "digital hospital" and digitized health system are observed from the strategic level and aspect of government actions.

Cluster 3 is depicted in blue color, on the left upper corner of the term map. This cluster contains keywords such as "digital technology", "platform", "electronic health records", and "patient". The topic of this cluster can be labeled as *Digital technologies for enhancing service offerings*. It is recognized that outcomes for patients can be beneficial with the digital transformation of healthcare, mainly in terms of new "service offerings". Consequently,

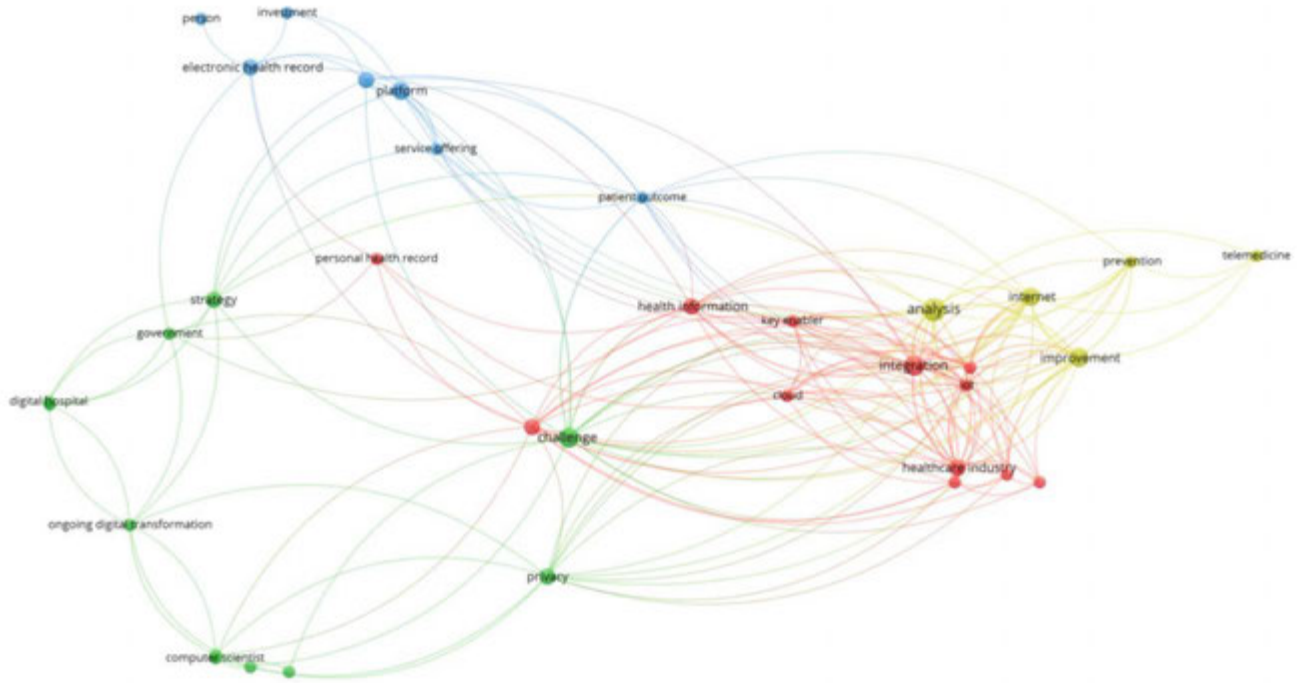


Figure 4. Clusters of terms in terms map network visualization

new service offerings come with its costs. Hence, needed "investments" for these services are also being discussed.

The last cluster, Cluster 4, is marked by yellow color and positioned on the right side of the term map. *Telemedicine and prevention in healthcare* is the topic of Cluster 4. This cluster includes several terms, namely "telemedicine", "internet", "prevention", "improvement", and "analysis". The potential of new technologies joined with information that can be found on the internet and transferred over the internet network offers great possibilities for the transformation of healthcare. This is the case both with the personal healthcare where the patients can have a more active role in their wellbeing and prevention, as well as with the institutional provided healthcare in terms of telemedicine.

V. CONCLUSION

In this paper, we have enveloped the state and progress of the research on digital transformation in healthcare, as well as its topics' landscape.

We have found out that the digital transformation of the healthcare sector is not being investigated as much as one would expect. Only thirty-one articles have matched our search inquiry, and most of them have been published in the last three years. Further, there is a preponderance of theoretical papers that propose a framework or give an opinion but lack empirical validation. When it comes to researchers that investigate the topic, frameworks are developed by the engineering scientists, while the opinion papers and some others are mainly written by medical professionals. Hence, researchers from the IS field and business economics that have the most experience in new technologies and digital transformation are surprisingly underrepresented in researching transformations of the healthcare sector. In the end, by the means of text mining, we discovered four discourses of topics that are being investigated in the literature:

- *Technology and aspects of proper implementation;*
- *Strategic and government issues;*
- *Digital technologies for enhancing service offerings, and;*
- *Telemedicine and prevention in healthcare*.

Based on our findings, we have shaped the recommendations for the future research agenda in the field. First of all, we strongly encourage empirical research that would stimulate the field. This mainly refers to case studies or technical papers of good implementation of technologies in the existing healthcare system, and diverse best practice examples. The interdisciplinary nature of the topic calls for the inclusion of diverse experts from the medical, informatics and public administration areas in order to advise proper government actions and bring valuable research results. Accordingly, we advise IS scholars to cooperate with other experts and become more involved in the propelling transformations that are already bringing beneficial possibilities into the healthcare system. This cooperation could especially strive towards advancing the discourses of *Strategic and government issues*, and *Telemedicine and prevention in healthcare*.

ACKNOWLEDGMENTS

This research has been fully funded by the University of Zagreb under the project "Digital transformation in Croatian Companies". We would also like to acknowledge Associate Professor Tomislav Hernaus, Ph.D., from the Faculty of Economics and Business Zagreb, for introducing us to the VoSwiever software.

REFERENCES

[1] L. Ivančić, V. Bosilj Vukšić, and M. Spremić, "Mastering the Digital Transformation Process: Business Practices and Lessons Learned," Technol. Innov. Manag. Rev., vol. 9, no. 2, pp. 36–50, Feb. 2019.

[2] H. C. J. Lucas, R. Agarwal, E. K. Clemons, O. A. El Sawy, and B. Weber, "Impactful Research on Transformational Information Technology: An Opportunity to Inform New Audiences," MIS Q., vol. 37, no. 2, pp. 371–382, Feb. 2013.

[3] M. N. Walsh and J. S. Rumsfeld, "Leading the Digital Transformation of Healthcare," J. Am. Coll. Cardiol., vol. 70, no. 21, pp. 2719–2722, Nov. 2017.

[4] R. Haux, "Health information systems – past, present, future," Int. J. Med. Inform., vol. 75, no. 3–4, pp. 268–281, Mar. 2006.

[5] D. Sharma, G. Singh Aujla, and R. Bajaj, "Evolution from ancient medication to human-centered Healthcare 4.0: A review on health care recommender systems," Int. J. Commun. Syst., no. May, p. e4058, Sep. 2019.

[6] J. Webster and R. Watson, "Analyzing the Past to Prepare for the Future: Writing a Literature Review," MIS Q., vol. 26, no. 2, p. 3, 2002.

[7] S. T. M. Peek, E. J. M. Wouters, J. van Hoof, K. G. Luijkx, H. R. Boeije, and H. J. M. Vrijhoef, "Factors influencing acceptance of technology for aging in place: A systematic review," Int. J. Med. Inform., vol. 83, no. 4, pp. 235–248, Apr. 2014.

[8] E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, "Process mining in healthcare: A literature review," J. Biomed. Inform., vol. 61, pp. 224–236, Jun. 2016.

[9] M. Allahyari et al., "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques," in Proceedings of KDD Bigdas, 2017, p. 13.

[10] D. Yu, Z. Xu, W. Pedrycz, and W. Wang, "Information sciences 1968–2016: A retrospective analysis with text mining and bibliometric," Inf. Sci. (Ny)., vol. 418–419, pp. 619–634, Dec. 2017.

[11] N. J. van Eck and L. Waltman, "Visualizing Bibliometric Networks," in Measuring Scholarly Impact, Y. Ding, R. Rousseau, and D. Wolfram, Eds. Cham: Springer International Publishing, 2014, pp. 285–320.

[12] N. J. van Eck and L. Waltman, "Text Mining and Visualization Using VOSviewer," ISSI Newsl., vol. 7, no. 3, pp. 50–54, 2011.

[13] V. Bosilj Vukšić, L. Ivančić, and D. Suša Vugec, "A Preliminary Literature Review of Digital Transformation Case Studies," in ICMIT 2018: 20th International Conference on Managing Information Technology, 2018, pp. 737–742.

[14] L. Ivančić, D. Suša Vugec, and V. Bosilj Vukšić, "Robotic Process Automation: Systematic Literature Review," in Business Process Management: Blockchain and Central and Eastern Europe Forum. Lecture Notes in Business Information Processing, vol. 361, C. ; Di Ciccio, R. ; Gabryelczyk, L. García-Bañuelos, T. Hernaus, R. Hull, M. Indihar Štemberger, A. Kő, and M. Staples, Eds. Cham: Springer, 2019, pp. 280–295.

[15] S. K. Boell and D. Cecez-Kecmanovic, "On being 'Systematic' in Literature Reviews in IS," J. Inf. Technol., vol. 30, no. 2, pp. 161–173, Jun. 2015.

[16] N. J. van Eck and L. Waltman, "VOSviewer Manual." 2019.

Creating of Digital Life in Art Museums

Dijana Ilišević*, Nataša Banović-Ćurguz* and Sarita Vujković**
* Mtel /Division for Technics, Banja Luka, Bosnia and Herzegovina
dijana.ilisevic@mtel.ba, natasa.banovic-curguz@mtel.ba
** Museum of Contemporary Art of Republic of Srpska, Banja Luka, Bosnia and Herzegovina
saravujkovic@gmail.com

Abstract - **Digital transformation affects all segments of human life. Cultural institutions are no exception. Information and communications technologies and technological innovations are reshaping the role of museums and art galleries which are becoming an integral part of the digital society. Visitors' expectations are changing and art museums are moving from collection-oriented to user/visitor-oriented institutions with strong interaction between individuals and museums around the world. This paper presents the use of new technologies in everyday digital life in museums. We also present the Museum of Contemporary Art of Republic of Srpska, and highlight the new challenges and opportunities that this institution expects in the digital age of the 21st century.**

Keywords - ***digital society, art, museums, technology, cultural heritage***

I. INTRODUCTION

As technology and application changing so dramatically, what will be communication model in future when physical and virtual worlds grow more connected, erasing the lines between online and offline. When transitioning to a digital society innovative model must be designed to ensure everyone benefit from opportunities in 5G environment (digital inclusion- gender, social etc.). With 5G in initial phase, industry and academia are started to look what 6G will be. In addition, to improvements in speed, capacity and latency, 6G offers shift from connected things to connected intelligence with extreme demand for new performance metrics [1].

Digital representation of reality will become important as reality itself and let people exist and communicate in parallel realities. Today, digital realities can matter as much as physical reality ever did, opening new mindset, and redefining the existing reality with new digital model. Defining arising issues in digital environment can help to rethinking technology and its role in digital society as a major public interest with equal opportunities for all [2].

These changes will have a broader impact on society, including aspects of regulation, policy, social impacts and new business model. Detailed following of technological changes and trends, with understanding consequences it has on society as a whole, is the most important step towards creation of strategy that can answer to modern trends. Digital transformation is a journey every country has to accomplish. Building healthy ecosystem based on digital mindset will bring opportunities and challenges and transform the way we communicate. Without appropriate strategies, opportunities can be lost, perhaps for good.

The role and mission of cultural institutions is reshaping by technological innovation. They become the places that reflect the actual values of the society and give an identity to the city. Including informative and educational features, museums have transformed into institutions with interactive users of their collections. Visitors' expectations are changing and art museums are moving from collection-oriented to user/visitor-oriented institutions with strong interaction between individuals and museums around the world. The digitization of cultural heritage, with advancement in communication models (internet, application, virtual tours etc.) is causing a novel model in traditional process of accessing information related to artworks and cultural institutions.

In this paper we shed light on museums, use of new technologies in everyday digital life in museums, institution concept which has known from Ancient Greek. We have presented the Museum of Contemporary Art of Republic of Srpska (MSURS), Bosnia and Herzegovina which has influence as a city meeting point in Banja Luka and Bosnia and Herzegovina and contributes significantly to the identity of the region. We have tried to identify the new challenges and opportunities that MSURS expects in the 21st century digital age.

II. DIGITAL SOCIETY

There is no doubt digital technology has changed society as a whole. Analysis must be started by looking at users of Future Communication (FC) model and the services they are likely going to be used in digital society. This is approach from service point of view or anything as a service [3]. Once we understand demands and driving factors, we have possibility to identify the requirements for future communication model. Furthermore, we need to appoint what will become new reality in digital world? Every day, physical and virtual worlds grow more connected, blurring the lines between online and offline. Our daily lives are being sent, transmitted, received, and saved in binary form using anywhere-anytime communication model, and ensure human existence in parallel realities. Dynamic rules that support the full potential of future ICT (Information and Communication Technology) networks will require redesigning or building new institutions, transformation of regulatory practice and governance, new skills and competencies, and significant changes to mindset and culture (legacy systems, institutions and processes).

The biggest change in FC model is that information is expected to be free but entertainment is surprisingly merchandise that people are willing to pay for. New FC model ensure globalization, in way which art is made, presented and unerstood. Innovative models of cultural colobration and interantional nature of art exchange are evolved connecting museums, artists and users of museum´s content. Social networks have become extremely popular all over the world often as a main source of news, information and it is important to notice that can be influenced. In recent scandal (Cambridge Analytics) it can go for even more scary anty-democracy scenario. Social Bots (SB) represents powerful model to influence relevant content and manipulate public opinion. As given in [4], five challenges of SB are identified. These results (Bastion framework) propose solutions to detect human behavior of those autonomous entity.

The modern lifestyle is mobile, online with screens and in mixed environment (physical and digital realities grow more and more connected). Future communication model and advanced services/application will shape digital society. How many realities are today? Virtual Reality (VR) refers to the creation of a completely artificial environment with no connection to physical world. Another VR refers to digital reality that brings real world closer – remote reality. Augmented reality (AR) is a virtual overlay upon the real world (smartphones, tablets, etc.) and often is taken as best of both worlds. If we understand demands and driving factors, we have opportunity to build society with equal opportunity for all. New realities are reflected on user experience of cultural content, giving museums different models to expand access to cultural heritage. In the future museums will ensure real-time sharing of cultural content/heritage in different realities, with technology as a foundation to this goal. In recent past, mobile devices have been changing target of online advertising. The internet of Things (IoT) will open up a new environment for online advertising. With prediction of numerous number of devices connected in IoT ecosystem, opportunities for expanding advertising space is promising direction (building new application and services). The IoT will open a new digital landscape, as an advertising market place as given in [5]. Furthermore, digital society and communication models are still under construction, as an open space for new world of innovation and connection with users of museum´s content.

A. Museums in Digital Society

Technological innovation is reshaping role of museums in digital age, building digital identity, and museums are still seeking for effective steps when it comes to diversity and inclusion nature that could make a real difference in future. Museums as an institution that connect past and future could be in center of a digital society, as a good example of institution that create innovative model for understanding technology to create digital society.

This model put museums in interdisciplinary environment and build future digital identity that will be inclusive, educational and increasingly socially oriented, as given in Fig. 1.



Figure 1. An illustrative example of museum in digital age.

The basic roles of museums such as education, collection and preservation are influenced by technology. With modern lifestyle (social networks, digital archive, multimedia display systems) museums role must be spread to building relationships with society and promote current values to achieve truly digital identity. New model recognizes museums as a place for entertainment (International Museum Day and the European Night of Museums), as well as education and cultural authority. Since communication model has been changed, there is reduction in permanent exhibition and museums mindset shift from visitors to user. With this model, museums offer social opportunities to users to accomplish a universal nature of culture.

Museums must embrace the opportunities technology will ensure to transform their position in digital society. The most important task will be to recognize efficient communication model and path to support successful use of technology innovation and user demands. Digital technologies will be used in two ways, to promote the events, exhibitions, workshops, performance etc. and to improve visitors' experience.

B. Digital Technologies in Contemporary Museums

Contemporary museums became integral part of their communities, local and global and sharing their ideas and collections through social media, blogs, websites ensuring living in digital environment where access to content and communication using digital applications and platforms is 24/7.

With digital technology as a foundation, visitors' experience can be improving, given as follows:

- Events (exhibitions, workshops, performance etc.) must include digital technology to create and support art. New foundation force museums to go beyond physical world and to connect with visitors through digital applications and platforms.
- Internet will continue to be a powerful resource for museums to go beyond physical and geographical boundaries. In addition, digital technologies ensure analytic tool for trends and evaluation of access and use of museums digitized collection and content.

- Physical and digital interaction with visitors will important tool to be considered by museums. Strategy in digital society for museums include diversity, inclusion and ensure their integration in society as a whole.
- Digital technology has a huge impact on art, and definition of art in future. Museums are forced to integrate with digital artist, mixed media, augmented reality and robotics in the creation of contemporary art. The use of video, photography and a variety of digital media has made it possible to endlessly multiply products of art, granting original artworks rarity status.

III. ENHANCING MUSEUM EXPERIENCE WITH DIGITAL TECHNOLOGIES

A. Musem of Contemporary art of Republic of Srpska

Public institution Museum of Contemporary Art of Republic of Srpska, Bosnia and Herzegovina (MSURS) is a central museum institution for performance of museum activities in the Republic of Srpska (founded in 1971) [6]. The home of the MSURS is one of Banja Luka's finest edifices, originally an Austro-Hungarian railway station built in 1891, as given in Fig. 2. For it to be used as a museum, it was adapted in full compliance with museological standards in 1981.

The museum has four departments: Collections department; Information and documentation centre and library; Education programmes department; and Department of exhibitions and programme activities. Museum of contemporary art of the Republic of Srpska is the first museum in Bosnia and Herzegovina whose name contains the word 'contemporary' and which narrowly focuses on contemporary art.

MSURS is dedicated to promoting the development of contemporary art in its sphere of influence, functioning as a hub that collects materials and studies and presents contemporary art practices. Also, the MSURS relies on its cooperation with relevant international institutions for its programme activities, which allows it to keep up with and present the latest developments and tendencies in art [7].



Figure 2. Museum of Contemporary Art of Republic of Srpska, Bosnia and Herzegovina [6].

The activities undertaken by MSURS have certainly helped the rise and consolidation of a vivid and dynamic local contemporary art scene, whose achievements have been recognized internationally. The Museum's most ambitious projects to date have been the organization, production and presentation of the pavilions of Bosnia and Herzegovina at the 55th International Art Exhibition of la Biennale di Venezia in 2013 ('The Garden of Delight' by Mladen Miljanović) and at the 57th Venice Biennale in 2017 ('University of Disaster' by Radenko Milak).

B. Digital Technologies in MSURS

The MSURS, which has become an urban focus, tries to adapt different digital experiences to the museum space. MSURS is an example of how digital technologies are used to enhance the museum experience. It incorporates some of the digital technology seen in major contemporary art museums. Although they may change temporarily depending on the type and content of the exhibited works, they can also be permanently housed in museum spaces.

In recent years, MSURS has organized several exhibitions where the impact of digital technology has been significant. The following is an overview of several exhibitions that give insight into the transition to digital thinking to achieve a social and public mission.

The exhibition 'Exceptionalities: Contemporary Aspects of Cultural Heritage/The New Trends of the Future Permanent Exhibition' was staged in honour of the tenth anniversary of MSURS. It included a selection of works from the collection of the Museum purchased or received as gifts to suggest possible guidelines for the kind of dialogue that would underscore the need for museums to purchase works of art, as well as some ways in which contemporary art and the special media it uses, such as videos, installations, performances and objects combining a range of materials, can be preserved as permanent cultural heritage. Figure 3 shows part of works in exhibition 'Exceptionalities', as an illustrative example of usage different media. The exhibition presented works by 32 artists belonging to different generations and created using various media, when intimacy and privacy prevailed over the context of public engagement and signification.

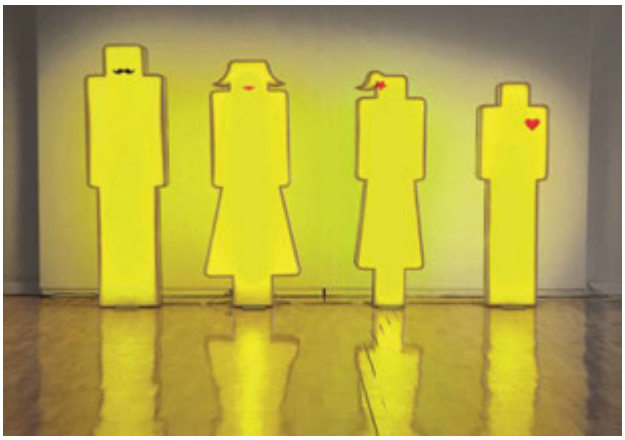


Figure 3. Exhibition 'Exceptionalities' [6].

The exhibition 'Multioriginals' includes prints, photographs, digital prints and videos from the collection of the MSURS by more than fifty artists from different generations. The term 'multioriginal' underlies the exhibition concept, as referring to the reproducible nature of artworks, i.e., the possibility of their multiplication and reproduction. The art practices present the use of a wide range of technologically diverse media in the creation of contemporary art. The use of video, photography and a variety of digital media has made it possible to endlessly multiply products of art.

The project 'Games & Politics' presents 18 non-commercial video games designed by artists, activists, students and designers over the past fifteen years, looking for the answer to the question of exactly how computer games are developing their political potential and how much this art form can offer an alternative to entertainment industry (supported by Goethe-Institute and the Center for Art and Media from Karlsruhe) [8]. Visitors to the exhibition will have the opportunity to try out video games that explore different aspects of contemporary society: political decision making (Democracy 3), gender issues (Perfect Woman), the state that controls (TouchTone and Orwell), war of drones (Killbox), attitude towards refugees (Escape from Woomera), resistance to totalitarian systems (Yellow Umbrella), the power of the media (The Westport Independent), past or current political events (The Cat and the Coup und Madrid), the effects of warfare (This War of Mine) and offer experiences that will surely encourage them to reflect on the role of video games in the modern world.

As a part of the project 'Games & Politics', an interesting programme was organized in the Museum, in a creative and educational manner: present the world of video games and the parallel between art and gaming, young and old generations, as well as a panel discussion on the gaming industry, its influence in the world of politics, culture and entertainment. As shown in Fig. 4, invitation for events was made using social networks as an in important tool to be considered by museums in future.

The 'Light Laboratory' project is an interactive digital video mapping solution that was designed on the MSURS facade as given in Fig. 5.



Figure 4. 'Games & Politics' [6].



Figure 5. 'Light Laboratory' [6].

This project has been organized with the aim of supporting the contemporary art scene as well as developing design and animation, with an emphasis on the latest multimedia production. The 'Light Laboratory' project was an excellent opportunity to present works by contemporary young visual artists who characterize in a specific way of working that is often not available to a wide audience.

IV. Challenges and Opportunities

Technological innovation is reshaping role of museums in digital age, building digital identity, and museums are still seeking for effective steps when it comes to diversity and inclusion nature that could make a real difference in future. Museum as an institution that connect past and future could be an in center of a digital society, as a good example of institution that create innovative model for understanding technology to create digital society. Digital identity of museums must include basic foundation of digital society (agile, dynamic, personalized and understanding user's expectation and becoming recognized as a deeply socializing environment). Museums have a goal as always in past, to provoking and bring social activism in question: What will your digital future be like, and how to live in it (give a real solution to every virtual problem)? Digital reforms are essential to capture the current and future technology changes and ensure their integration in society as a whole to achieve max transformation impact.

In transition to digital age/mindset is likely to transform physical art collections to digitized collection, and ensure a lot of opportunities to access content across traditional boundaries. With novel public digital projects, there is an urgent need to adopt policies for share of cultural content. Museums have to be aware about intellectual property of digital content to ensure leading role in digital environment. Since technology plays a crucial role in ensuring access to museum's exhibitions, projects and digitized collection, it is important to facilitate technical knowledge for employees needed to achieve digital identity and recognize diversity and multidisciplinary aspects for projects success.

Digital transformation will be challenging task for Bosnia and Herzegovina, due to political and legal complexity without efficient use of resources to improve lack of digital legislation and pre-digital policy. Managing digitalization and the transformation of society always involves new challenges, and creates a lack of understanding of the challenges that today's societies are facing like fake news, propaganda, manipulate public opinion etc. Understanding how to trigger digital services adoption will be demanding in emerging market in Bosnia and Herzegovina [9]. With technology as a tool, MSURS will improve visitors experience and find effective steps when it comes to diversity, education and inclusion nature of digital society.

Museums, with technology as a tool, in future will imporve access to art as universal human right, and with facilities as a place for education, axploring art to achive full creative potential. Addressing these challanges succesfully, gives a potential to be part of digital society (multidisciplinary models for innovative projects in future). With art as a foundation, MSURS is devoted to increase its influence across society and to redefine visitors experience for the 21st century using digital platforms. As part of strategy for digital society, future Digital Department will promote MSURS using advanced FC models and to develop high quality digital contetnt and digitized collections. Acces using digital platforms will create significant value in new cretive usage of art and enhance visitors experience, and ensure existance in digital society.

MSURS is developing collaborative projects with international support (British Council, Goethe-Institute), the Center for Art and Media from Karlsruhe, International Museum Organization at Unesco - ICOM, UNESCO etc.). Suay Aksoy (General president of ICOM) first official visit to Bosnia and Herzegovina was to MSURS in Banja Luka, as part of official invatition for presenation in ICOM conference in Kyoto, within the framework of the International Committee for Exhibitions and Exchanges, as a good example of how museums can be empowered to pose vexed questions. In addition, touring exibitions as ambition to work localy as well as interantionaly was organized (touring 'Space, Form, Touch' in Trebinje- Bosnia and Herzegovina, 'Perception' in North Macedonia, 'Was ist Kunst in Bosnia and Herzegovina' in Slovenia etc.). It is important to notice, MSURS is dedicated to free entrance to the exhibitions, workshops and author-led tours.

V. CONCLUSION

Bridging a gap between art and technology is a process where new realities will not be in collision but the enlargement of human existence. Museum/Art landscape in digital society must include technology as an integral part to achieve the following goals: contact with visitors, active participation of visitors, international access supported with technology as well as inclusive, educational and increasingly socially oriented approach. The most successful institutions in digital society will be those that recognize and establish multidisciplinary environment to build innovative visitor oriented model based on technology. Museums art environment (physical and virtual) have to include understanding the impact of content to increase engagement to become integral part of their communities. In digital society, fulfilling visitors' demands create many new challenges and opportunities for the museum to address. Using data driven modeling become relevant in art environment as a future innovation enabler for better social influence and public mission of museums.

REFERENCES

[1] K. B. Letaief, W. Chen, Y. Shi, J. Zhang and Y.A. Zhang, "The roadmap to 6G: AI empowered wireless networks" IEEE Communication Magazine, vol. 57, no. 8, pp. 84–90, August 2019

[2] D. Ilisevic and N. B. Curguz, "Future communication model: Challenges and opportunities for society as a whole", Proceedings of MIPRO 2016/CTI, pp. 702-706, June 2016

[3] J. Schonwalder, M. Fouquet, G. D. Rodosek and I.C. Hochstatter, "Future Internet = Content +Services+Management", IEEE Magazine., Communication, vol. 48, no. 11, pp. 27-33, 2009

[4] M. Zago, P. Nespoli, D. Papamartzivanos, M.G. Perez, F. G. Marmol, G. Kambourakis and G. M. Perez, "Screening out social bots interference: Are there any silver bullets", IEEE Communication Magazine. vol. 57, pp. 98–104, August 2019

[5] H. Aksu, L. Babun, M. Conti, G.Tolomei and A. S. Uluagac, "Advertising in the IoT era: Vision and challenges", IEEE Communication Magazine Vol. 56, pp. 138-144, November 2018

[6] http://www.msurs.net/index.php/en/ (accessed 01.02.2020.)

[7] S. Vujkovic, "Interdisciplinary approaches to creating educational programs in museums - Example of Museum of contemporary art of Republic of Srpska ", Second Professional Scientific Conference Schools and Museums, pp. 193-200, Ljubljana 2018

[8] Museum of Contemporary Art of Republic of Srpska, "Inventory 2019", No 102, pp. 80-89, Bosnia and Herzegovina 2020

[9] D. Ilisevic, N. B. Curguz and D. Budimir, "Embracing the 5G era with appropriate regulation framework ", Telfor, Belgrade 978-1-5386-7171-9/18 IEEE, pp. 1-4, November 2018

SWOT analysis of selected digital technologies in transport economics

M. Jović*, E. Tijan**, D. Žgaljić***, and P. Karanikić****
* University of Rijeka, Faculty of Maritime Studies/Department of maritime logistics and management, Rijeka, Croatia
jovic@pfri.hr
** University of Rijeka, Faculty of Maritime Studies/Department of maritime logistics and management, Rijeka, Croatia
etijan@pfri.hr
*** University of Rijeka, Faculty of Maritime Studies/Department of maritime logistics and management, Rijeka, Croatia zgaljic@pfri.hr
**** University of Rijeka, Department of Biotechnology, Rijeka, Croatia
pkaranikic@uniri.hr

***Abstract* - This paper presents a review of selected digital technologies (document digitalization, Port Community Systems, Warehouse Management Systems and Decision Support Systems) in transport sector, and their impact on transport economics. SWOT Analysis of selected technologies in transport sector from an internal and external economic perspective is conducted. The authors will focus on the effect of above-mentioned technologies in terms of cost savings and sustainable operations.**

Keywords – digital technologies; digital innovations; SWOT analysis; transport economics

I. INTRODUCTION

All decisions related to planning, design, and improvement of transport infrastructure have economic implications. Transport economics includes the issues such as transport location, movement of people and freight/goods, transport demand, transport planning [1] as well as planning regarding the implementation of new technologies/innovations. As digital technologies evolve continuously, stakeholders should always be aware of the changes in order to easily adapt to emerging technologies/innovations and to stay ahead in the market [2].

Informatization which may facilitate successful innovation processes [3], refers to the extent by which a geographical area, an economy or a society is becoming information-based, i.e., increase in size of its information labour force [4]. According to Li et al. [3], informatization can help enterprises to create innovation opportunities, to decrease costs, to increase strategic flexibility by responding to changes and challenges of external environment, etc.

The authors will analyze the following digital technologies: document digitalization, Port Community Systems (PCS), Warehouse management systems (WMS), and Decision Support Systems (DSS). The aforementioned digital technologies can simplify and facilitate logistic processes as well as improve the safety and security onboard ships, and in port areas.

The goal of the research is to analyze strengths, weaknesses, opportunities and threats of these technologies through SWOT analysis. SWOT analysis has a "time dimension", and the comparison and monitoring by SWOT analysis is useful to the enterprises at different time points of the business processes [5]. Transparency and easy access to data are the basis for successful business, which also applies to the transport sector. his paper presents a review of strengths, weaknesses, opportunities and threats of the technologies mentioned above.

II. TRANSPORT ECONOMICS

Mobility and transport have always played an important role in economic, ecological and social development [6]. On the one hand, transport is one of the main factors of economic growth and quality of life; on the other hand, it is one of the main causes of environmental pollution [7]. Furthermore, the growth of international trade and globalization causes an increase in traffic and freight [6]. Transportation comprises of various activities (policy making, planning, designing, etc.) and includes various stakeholders who have to exchange documents in order to execute transport services [8].

Economics and transportation are closely related terms and the transport sector may be considered as an important component of the economy [9]. When transportation systems are efficient, they provide social and economic opportunities and benefits that result in positive multiplier effects such as improved accessibility to markets, employment and additional investments. When transportation systems are incomplete in terms of capacity or reliability, they cause increased costs or missed opportunities. Efficient transportation reduces costs, while inefficient transportation increases costs [9]. As trade and cargo volumes continue to grow internationally,

companies in the transport sector around the globe are looking to new technologies to help manage resources in a more sustainable and cost-effective manner [10].

Seaport, as a part of a transport sector, and a main node of maritime transport [11], is often seen as a determinant of the economic growth, as it contributes to the development of economic sectors and to the generation of economic benefits [12]. Costs, time and predictability are decisive factors for seaport competitiveness [13]. Seaports are essential for the support of economic activities in the surrounding areas [14]. Each seaport, as a part of maritime transport which aims to be competitive, should accept digitalization [15], as transparency and easy access to data are of the utmost importance for successful seaport business [16].

III. DIGITAL TECHNOLOGIES IN THE TRANSPORT INDUSTRY/SECTOR

Transport involves a large number of stakeholders such as ship owners, insurance companies, agents, etc. [17] who have to exchange document in order to execute and complete transport services [18]. With the development of international trade and globalization, traffic and cargo volumes keep increasing. Therefore, numerous stakeholders are compelled to accept changes and turn to more effective practices by introducing technologies that can gather and process massive amounts of information (in a cost-effective way) [19]. On the other hand, although digital trends evolve continuously, according to the international survey conducted by FutureNautics Maritime and Ericsson (2017), the maritime industry, as a part of transport industry, is not fully aware of how digital trends affect their business and the competitiveness of organizations [20]. Maritime transport stakeholders should embrace innovations in order to stay competitive and to keep up with digital trends.

As informatization may facilitate innovation [3], a synergetic development between the informatization and the innovation could enhance the competence of companies involved in the transport [21]. Despite different perspectives among entrepreneurs, academics and policy makers, innovations are defined broadly as the development and use of new ideas and behaviours in organizations and narrowly as implemented technologically new products and processes or significant technological improvements in products and processes [22]. In general, two different types of innovations exist: (a) new market innovations that create a new demand for novel technologies and related products, and (b) low-end innovations that provide technologies with similar characteristics to existing technologies but at a lower cost [23].

Based on the literature review and the experience in several projects related to digitalization in transport, authors have chosen the following digital technologies in transport: document digitalization, Port Community Systems, Warehouse Management Systems and Decision Support Systems. Following is the SWOT analysis of the above-mentioned digital technologies. The authors have chosen the aforementioned technologies due to their strong contribution in accelerating, simplifying and automating transport processes.

A. Document digitalization

Digitalization of business processes and transparency are the most important trends that companies that operate in the supply chain will need to develop considerably in the future [24]. The authors have performed the SWOT analysis of document digitalization in the transport sector as follows:

Strengths: Data handling in digital, structured and standardized way is a major prerequisite for innovative solutions and processes [25]. One of the strengths is the reduction or elimination of paper documents and streamlining business processes: ship operators, masters, and agents are still burdened with having to fill in paper documents, which contain repetitive data, and to distribute them to different government authorities [26]. Another strength is long-term cost savings. Data and/or document distribution to the competent authorities creates confusion caused by data redundancy. These administrative tasks significantly increase shipping costs by utilizing excessive waiting time for cargo processing, also causing delays in the next port of call [27]. Furthermore, digitalization allows stakeholders involved in the business processes to input the necessary data only once, and to reuse the data according to their scope of work and authorization. One of the strengths is also the decreased processing time. For example, the TradeNet system (gathering various parties from the public and private sectors to exchange trade information electronically) is capable of handling more than 30,000 declarations a day, it processes 99% of permits in just 10 minutes and receives all monetary collections through interbank transactions [12].

Weaknesses: Investment costs represent one of the weaknesses of document digitalization. Necessary technical modification depends on the state of existing technologies used in a company, and must be adapted according to the needs of the company. Document digitalization usually implies a major upgrade or replacement of working tools, applications and underlying infrastructure. Furthermore, there is a steady increase in the acceptance and use of digitalized documents for different modalities. However, there are still obstacles for a complete transition to e-documents. The main obstacle is the existence of different authorities in different states who have not yet legally accepted the provision of digital data [28]. Furthermore, document standardization in most cases needs the revision of the document standards used within the company (as standardized data exchange with other parties is necessary). Moreover, it is necessary to digitalize existing documents and create new ones in digital form.

Opportunities: Documents in structured digital form enable accelerated execution time and reutilization and verification of the data. Structured and standardized documents such as Smart Bill of Lading can allow improved data exchange, access to relevant cargo data, simplified notification process and data processing [29]. Digitalized documents can be available for the next stops in the transport chain well ahead of the cargo, which

allows better preparation, planning and forecasting of services. Furthermore, digital signature can vastly speed up document processing and improve document security (simple authentication). Digitalization will spur automation, lead to the development of smart ships and positively impact safety and environmental performance. Digitalization has also enabled more far-reaching concepts, such as Big Data, Internet of Things, Blockchain, and cloud computing, which will provide the shipping industry with new ways to collect, process and exchange valuable data in real time [18].

Threats: Freight forwarders, customs brokers and other logistics service providers already extensively use electronic tools for planning, warehouse management, fleet management, customs declarations and the creation of transport or other documents. Some logistics service providers have even developed their own IT solutions and are beginning to consider themselves more as IT providers. Today, transport documents almost never originate on paper anymore, but electronically. In many cases, however, the transfer of electronic data to another party remains paper-based and needs to be re-typed (manually, with the risk of errors) to transform it into electronic data again [28]. One of the threats is the lack of strategies: when an organization seeks to transform from a manual process to a comprehensive digital platform, it requires a successful leadership strategy which can be influential and impactful in advancing this type of change over the long term [30]. Further initiatives at legislative level appear to be needed, at EU and, potentially, national level, in order to remove current legislative and administrative barriers to the use of electronic transport documents, particularly in regard to interaction with the authorities, for regulatory compliance purposes [31].

B. Port Community System (PCS)

A PCS is a neutral and open electronic platform enabling intelligent and secure exchange of information between public and private seaport stakeholders in order to improve the competitive position of the seaport environment [32]. PCS plays a key role as a seaport business process facilitator, and its importance is growing due to the new trends. PCS can play a major role as countries move towards the Single Window environment) [33]. The following is a SWOT analysis of the PCS:

Strengths: PCS enables connectivity of multiple systems operated by many different seaport organizations: PCS connects the multiple systems operated by a variety of organizations that make up a seaport community [34]. Moreover, PCS enables standardized message exchange among port community members. A paperless and standardized communication is not only a prerequisite for efficient maritime transport operations which include multiple stakeholders, but also for improving supply chain integration, coordination, and performance [35]. One of the strengths is increased transparency: PCS provides full transparency of the cargo movement [36], including dangerous goods and other notifiable cargoes, status information and control, tracking and tracing through the whole logistics chain as well as full range of cargo and the maritime statistics. Another strength is real time information [37], meaning smoother data processing and faster cargo movement.

Weaknesses: One of the weaknesses is the unwillingness of some port community members to integrate the PCS [37]. Furthermore, indirect nature of most benefits can only be realized in the long run [37]; use of the system is not mandatory; fundamental differences in governance models across PCS operating companies [37], etc. Another weakness is the uneven distribution of benefits: because of that, port community members might decide not to integrate with the system. This is especially relevant for small companies and may prevent the development of an advanced technological infrastructure. Moreover, in order to enable a functional usage of this application, it is necessary to analyze the data that must be processed through this application, develop and implement new software solutions for the port management etc.

Opportunities: PCSs could expand to the entire supply chain. This would most probably be achieved through an integration of all ICT systems along the supply chain [37]. Furthermore, cooperation between PCSs might be enhanced. Regional or even global PCSs might be designed. In both cases a further standardization of interfaces and processes would be required. The requirements of and benefits for each company would have to be outlined and agreed on in advance [37]. PCSs could extend their service portfolio in order to further facilitate collaborative planning and inter-organizational data processing [37]. Regional or national governments might enforce regulatory changes in the systems' favor [37]. PCS can serve as an exchange platform between different applications and systems, for example between operations management application in the company and national Single Window environment (for example for approval, notification, authorization, etc.).

Threats: One of the threats is the system vulnerability (depending on the PCS architecture and network exposition) [38]: if the PCS is exposed to third parties through a dedicated web interface, the attacker can exploit common web application vulnerabilities to have access to the PCS. If the PCS is connected with external systems through data exchanges protocols such as API or EDI protocols, the attacker can organize a man-in-the-middle attack by intercepting the data exchanges and modifying them in case data exchanges are not secured enough; or if direct access to the PCS itself out of the port network is possible, the attacker can exploit weak network security measures to have direct access to this system and use credentials he may have stolen through social engineering. One of the threats is also the unwillingness of port community members to share information: port community members might be unwilling to use the system because of skepticism regarding data security [37].

C. Warehouse Management System (WMS)

A WMS is a software tool that allows organizations to control and administer warehouse operations from the time goods or materials enter a warehouse until they move out [39]. The following is a SWOT analysis of Warehouse Management Systems:

Strengths: One of the strengths is the reduction or elimination of paper documents: with a WMS, warehouse personnel do not have to spend time reading or handling paper documents, keying in data or thinking about where they need to go [40]. Another strength is lowering the upfront costs: for example, with a cloud-based WMS solution, there are never any hardware or software upgrade or maintenance fees. There are no IT infrastructure costs or costs for hardware, system, or database administrators [41]. Furthermore fewer mistakes are made, and errors are immediately identified [42]: WMS ensures warehouse control over operations to help ensure continuity, consistency and reliability. Getting the most work done in less time with fewer errors means that there will be less work done researching and correcting errors and billing. This lowers the burden on customer service representatives so that they can focus on providing top quality support [43]. Other strengths are: improved warehouse safety and security, improved supplier relationships [44], and improved communication and information exchange.

Weaknesses: The training required to operate WMS devices along with the cost of the non-cloud based software together can easily exceed the price of the devices [45]. Furthermore, WMSs always require expert knowledge to achieve the maximum benefit [46]. Such a system demands expert knowledge to handle the devices well, which can be a challenging task as finding aptly skilled resources can be tough [45]. Other weaknesses are: the need to rearrange numerous warehouses in order for a WMS to work properly [40]; increased master data maintenance; additional process steps for receipt and picking; hard to resolve problems caused by incorrect processing [47], etc.

Opportunities: WMS can facilitate management and planning of operations. Moving around the warehouse is often challenging for staff. In most cases, an employee would still need a ladder to reach high shelves — such inconveniences reduce the productivity of inventory management and increase the risk of workplace injuries. Autonomous technologies present a way to increase warehouse productivity: automated guided vehicles can calculate the shortest route to any aisle and restock the inventory with no human supervision. Tasks, like picking and packing, are monotonous and tiresome — thus, the odds of human error are higher than in more demanding tasks. IoT and smart warehouse technologies help automate repetitive assignments and allocate the workforce more efficiently. By introducing IoT to the warehouse, store managers will be able to reduce order inaccuracies and inventory damage [48].

Threats: Just-in time warehouse approach is not possible without adequate and reliable information. If data about the cargo is unavailable or available too late during the process, the WMS cannot include this information in the planning process, which can negatively affect the warehouse performance and later stages of logistic processes. One of the threats is the adaptability to change: there is a possibility that logistics operators will not be as fast to respond as the manufacturing and distribution industry. Other threats are the lack of awareness about benefits of WMS among small-scale industries; constant need for software upgrades etc.

D. Decision Support Systems (DSS)

A DSS is a software solution used to support determinations, judgments, and courses of action in an organization or a business [49]. A DSS may include an expert system or artificial intelligence (AI) [50]. The following is a SWOT analysis of the DSS:

Strengths: Notable strengths are reduced decision cycle time, increased employee productivity and more timely information for decision making [51]. Another strength is the improved decision-making effectiveness and interpersonal communication: Model-driven DSS provides a means for sharing facts and assumptions. Data-driven DSS make "one version of the truth" about company operations available to managers and hence can encourage fact-based decision making. Improved data accessibility is often a major motivation for building a data-driven DSS.

Weaknesses: Increased implementation and maintenance costs. On the other hand, the cost of decision-making decreases once a decision support system is installed. Nevertheless, the development and implementation of a DSS requires large investments. Customization may imply higher costs [52]. Furthermore, the DSS requires investing in an information system to collect data from many sources and to analyze them to support the decision making. Some analysis for DSS needs the advance of data analysis, statistics, econometrics and information system, which means a high cost to hire the specialists to set up the system [53]. Another weakness is the difficulty in quantifying all the data: a DSS majorly relies on quantifiable data. Consequently, it is difficult to analyze intangible or indefinable data. In reality, some values cannot be very specific and defined in numbers. Even though a DSS may quantify some of these aspects, the end result must be duly considered by the decision makers. They must use their own judgment when making the final decision [52].

Opportunities: during vessel operation, data measured from different sensors and other external sources (e.g., weather information, route planning, other operational constraints) can be used to give the master valuable advice on how to operate with less use of fuel [54]. Furthermore, data generated through sensors embedded in various things/objects will generate massive amounts of unstructured (big) data on real-time basis that holds the promise for intelligence and insights for dramatically improved decision processes [55].

Threats: DSS can overload decision makers with information and actually reduce decision making effectiveness [53]. Furthermore, too much dependence on DSS may present a threat. Some decision makers develop a tendency to depend too much on computerized decision making without considering uncontrollable factors which can lead to negative consequences. Decision makers may not hone their skills further because of excessive dependence on DSS [52]. One of the threats is the employee's resistance to change: training is the only way to avoid the potential negative issues [53].

IV. CONCLUSION

Economics and transportation are closely related terms, and when transportation solutions are efficient, they can enable economic benefits, for example, decreased costs. In this paper, authors have analyzed strengths, weaknesses, opportunities and threats of selected digital technologies and solutions (document digitalization, Port Community Systems, Warehouse Management Systems and Decision Support Systems) related to transport, and their impact on transport economics. The authors concluded that, although document digitalization brings numerous advantages (cost savings, reduction or elimination of paper documents, data re-use etc.), it is not always a simple process and depends on numerous factors (the state of existing technologies used in a company etc.). Another digital technology, a Port Community System facilitates optimization, management and automation of port and logistics processes. However, one of the weaknesses are increased implementation and maintenance costs of PCS. Furthermore, the success of a PCS depends on actors' willingness to share information.

Warehouse Management System, as another digital solution, can for example, reduce operating costs by increasing warehouse productivity. On the other hand, if cargo data is unavailable during the process, it can negatively affect the warehouse performance and later stages of logistic processes (as WMS system cannot include this information in the planning process). Furthermore, WMS implementation is a complex process as it depends on the level of maturity of currently implemented ICT systems, on the level of automation etc.

Decision Support Systems enable increased employee productivity, more timely information for decision making, faster and more rational decisions, which ultimately results in cost reduction. However, implementation and maintenance costs of DSS are high (adapting the system to the company needs may also require higher costs).

The research proved that although the aforementioned digital technologies and solutions may provide numerous advantages and opportunities in transport, there is still a lot of room for improvements. The implementation of those technologies and solutions is becoming more common and those digital technologies will be upgraded over time, enabling decreased costs of business processes and increased security.

ACKNOWLEDGMENT

This work was supported by "DigLogs – Digitalising Logistics Processes" (Interreg V-A Italy – Croatia 2014-2020) project.

REFERENCES

[1] D. Teodorović and M. Janić, "Transport economics," Transportation Engineering, 2017. [Online]. Available: https://www.researchgate.net/publication/311997791_Transport_Economics. [Accessed: 04-Dec-2019].

[2] G. Fallah, "10 Digital Marketing Trends for 2019 you Should Know," 2019. [Online]. Available: https://www.digitaldoughnut.com/articles/2018/december/10-digital-marketing-trends-for-2019. [Accessed: 04-Dec-2019].

[3] C. Li, S. Han, S. Kumar, and W.-X. Feng, "The influence of senior executive support informatization on radical innovation performance," Ind. Manag. Data Syst., vol. 119, no. 4, pp. 821–839, 2019.

[4] "Informatization," 2019. [Online]. Available: https://www.definitions.net/definition/informatization. [Accessed: 04-Dec-2019].

[5] "Prometna povezanost Srednje Dalmacije | Rural Koncept," 2018. [Online]. Available: http://www.rural-koncept.hr/hr/destinacije/dalmacija/srednja-dalmacija/prometna-povezanost/. [Accessed: 15-Apr-2019].

[6] M. Jović, N. Kavran, S. Aksentijević, and E. Tijan, "The Transition of Croatian Seaports into Smart Ports," in 42nd international convention on information and communication technology, electronics and microelectronics, 2019, pp. 1618–1622.

[7] S. Dimić, D. Pamučar, S. Ljubojević, and B. Dorović, "Strategic transport management models-the case study of an oil industry," MDPI Sustain., vol. 8, no. 9, 2016.

[8] E. Tijan, M. Jović, and P. Karanikić, "Economic and ecological aspects of electronic Transportation Management Systems in seaports," Proceedings of the Maritime and Port Logistics Bar Conference, 2019. [Online]. Available: https://www.bib.irb.hr/1003853. [Accessed: 31-Dec-2019].

[9] M. A. Mosaberpanah and S. D. Khales, "The Role of Transportation in Sustainable Development," Conference: In ICSDEC 2012: Developing the Frontier of Sustainable Design, Engineering, and Construction, 2012. [Online]. Available: https://www.researchgate.net/publication/255991703_The_Role_of_Transportation_in_Sustainable_Development. [Accessed: 04-Dec-2019].

[10] GreenPort, "Smart and sustainable ports," 2016. [Online]. Available: https://www.greenport.com/news101/Projects-and-Initiatives/smart-and-sustainable-ports. [Accessed: 23-May-2019].

[11] E. Tijan, A. Agatić, M. Jović, and S. Aksentijević, "Maritime National Single Window — A Prerequisite for Sustainable Seaport Business," MDPI Sustain., vol. 11, no. 17, pp. 1–21, 2019.

[12] T. A. Jouili and M. A. Allouche, "Impacts of seaport investment on the economic growth," Promet - Traffic - Traffico, vol. 28, no. 4, pp. 365–370, 2016.

[13] "2017 Transportation Management Systems Trends," 2017. [Online]. Available: https://www.vaizva.com/transport-management-systems/2017-transportation-management-systems-trends/.

[14] Ş. C. Gherghina, M. Onofrei, G. Vintilă, and D. Ş. Armeanu, "Empirical evidence from EU-28 countries on resilient transport infrastructure systems and sustainable economic growth," MDPI Sustain., vol. 10, no. 8, 2018.

[15] M. Jović, M. Filipović, E. Tijan, and M. Jardas, "A review of blockchain technology implementation in shipping industry," Pomorstvo, vol. 33, no. 2, pp. 140–148, 2019.

[16] M. Jović, E. Tijan, S. Aksentijević, and B. Sotošek, "The role of Electronic Transportation Management Systems in Seaport Digitalization," 32nd Bled eConference: Humanizing Technology for a Sustainable Society Conference Proceedings, Kranj, Slovenia: University of Maribor Press, 2019. [Online]. Available: https://www.bib.irb.hr/992899. [Accessed: 13-May-2019].

[17] M. Jović, E. Tijan, R. Marx, and B. Gebhard, "Big Data Management in Maritime Transport," Journal of Maritime and Transportation Sciences, 2020. [Online]. Available: https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=338474. [Accessed: 27-Jan-2020].

[18] E. Tijan, M. Jović, M. Jardas, and M. Gulić, "The Single Window concept in international trade, transport and seaports," Pomorstvo : scientific journal of maritime research, 2019. [Online]. Available: https://hrcak.srce.hr/229258. [Accessed: 30-Dec-2019].

[19] Marshall University, "Big Data, Big Data Analytics," 2019. [Online]. Available: https://www.marshall.edu/bigdata/. [Accessed: 23-Aug-2019].

[20] Ultimate Software, “5 challenges to Digital Transformation in the Maritime sector,” 2017. [Online]. Available: https://www.ultimatesoftware.nl/en/5-challenges-to-digital-transformation-the-maritime-sector/. [Accessed: 12-Feb-2019].

[21] K. Bi, H. Ma, and P. Huang, “Informatization and Process Innovation: A China Case Study,” 2013. [Online]. Available: https://scialert.net/fulltext/?doi=jas.2013.1473.1478. [Accessed: 04-Dec-2019].

[22] L. Majamäki and M. Akpinar, “Challenges and success factors in pursuing disruptive innovations: A Finnish high-tech start-up case study,” Finnish Bus. Rev., no. September, pp. 1–11, 2014.

[23] J. Guo, J. Pan, J. Guo, F. Gu, and J. Kuusisto, “Measurement framework for assessing disruptive innovations,” Technol. Forecast. Soc. Change, vol. 139, pp. 250–265, Feb. 2019.

[24] W. Kersten, M. Seiter, B. Von See, N. Hackius, and T. Maurer, Trends und Strategien in Logistik und Supply Chain Management. 2017.

[25] Y. Yoo, “Digitalization and Innovation,” 2010. [Online]. Available: https://www.researchgate.net/publication/48926382_Digitalization_and_Innovation. [Accessed: 12-Dec-2019].

[26] European Maritime Safety Agency, “National Single Window Prototype: an Electronic Solution for Simplifying Administrative Procedures,” 2015. [Online]. Available: http://www.emsa.europa.eu/emsa-documents/latest/item/2317-national-single-window-prototype-an-electronic-solution-for-simplifying-administrative-procedures.html. [Accessed: 10-Jul-2019].

[27] E. Tijan, M. Jardas, S. Aksentijević, and A. Perić Hadžić, “Integrating Maritime National Single Window with Port Community System – Case Study Croatia,” in 31ST Bled eConference - Digital Transformation: Meeting the Challenges Conference Proceedings, 2018, pp. 1–11.

[28] CLECAT, “Digitalisation in transport and logistics,” 2017. [Online]. Available: https://www.clecat.org/media/Position paper CLECAT Digital Transport and Logistics_1.pdf. [Accessed: 08-Apr-2019].

[29] “From the Past to the Future,” 2019. [Online]. Available: https://cargox.io/welcome/. [Accessed: 05-Jul-2019].

[30] M. Sow and S. Aborbie, “Impact of Leadership on Digital Transformation,” Bus. Econ. Res., vol. 8, no. 3, p. 139, 2018.

[31] European Commission, “Towards paperless transport within the EU and across its borders,” 2018. [Online]. Available: http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=15358.

[32] IPCSA, “PCS / Port Community Systems - IPCSA International,” 2014. [Online]. Available: https://ipcsa.international/pcs. [Accessed: 22-May-2019].

[33] “Port Community Systems,” IMO MTCC GMN Digital Week, 2019. [Online]. Available: https://gmn.imo.org/wp-content/uploads/2019/02/IMO-MTCC-Port-Community-Systems-day-1.pdf. [Accessed: 04-Dec-2019].

[34] “PCS / Port Community Systems - IPCSA International,” 2017. [Online]. Available: https://ipcsa.international/pcs. [Accessed: 06-Mar-2019].

[35] “The importance of standardization - BNQ,” 2019. [Online]. Available: https://www.bnq.qc.ca/en/the-importance-of-standardization.html. [Accessed: 15-Mar-2019].

[36] L. Heilig and S. Voß, “Information systems in seaports : a categorization and overview Information systems in seaports : a categorization and overview,” Inf. Technol. Manag., no. December, 2017.

[37] S. Treppte, “The Role and Scope of Port Community Systems in Providing Data that Enhances Supply Chain Risk Management - A Case Study for Freight Forwarders in the Port of Rotterdam,” 2011. [Online]. Available: https://www.erim.eur.nl/fileadmin/default/content/erim/research/centres/smart_port/admin/c_news/110911_st_master thesis_pcss and risk management.pdf. [Accessed: 14-Nov-2019].

[38] “PORT CYBERSECURITY: Good practices for cybersecurity in the maritime sector,” 2019. [Online]. Available: https://www.enisa.europa.eu/publications/port-cybersecurity-good-practices-for-cybersecurity-in-the-maritime-sector/at_download/fullReport. [Accessed: 02-Feb-2020].

[39] “Warehouse management system (WMS),” 2018. [Online]. Available: https://searcherp.techtarget.com/definition/warehouse-management-system-WMS. [Accessed: 04-Dec-2019].

[40] “WMS Pros and Cons,” 2006. [Online]. Available: https://www.ewweb.com/archive/wms-pros-and-cons. [Accessed: 14-Nov-2019].

[41] ORACLE, “What Is a Warehouse Management System (WMS)?,” 2020. [Online]. Available: https://www.oracle.com/applications/supply-chain-management/solutions/logistics/warehouse-management/what-is-warehouse-management.html. [Accessed: 31-Jan-2020].

[42] Mecalux, “Warehouse Management System, Easy Mecalux WMS,” 2020. [Online]. Available: https://www.mecalux.com/software/warehouse-management-system-wms. [Accessed: 31-Jan-2020].

[43] Datexcorp, “What is a Warehouse Management System?” [Online]. Available: https://www.datexcorp.com/what-is-a-warehouse-management-system-wms/. [Accessed: 31-Jan-2020].

[44] “WAREHOUSE MANAGEMENT SYSTEMS FAQS,” 2020. [Online]. Available: https://www.scjunction.com/warehouse-management-system. [Accessed: 31-Jan-2020].

[45] Shiprocket, “Warehouse Management System (WMS) in India – Pros and Cons,” 2019. [Online]. Available: https://www.shiprocket.in/blog/warehouse-management-system-wms-in-india-pros-and-cons/. [Accessed: 31-Jan-2020].

[46] “Warehouse Management System (WMS) in 2020- Pros and cons.” [Online]. Available: https://ithinklogistics.com/blog/warehouse-management-system-wms-in-2020-pros-and-cons/. [Accessed: 31-Jan-2020].

[47] “Advantages and Disadvantages of Warehouse Management,” 2020. [Online]. Available: https://www.erpgreat.com/warehouse/advantages-and-disadvantages-of-warehouse-management.htm. [Accessed: 31-Jan-2020].

[48] “Internet of Things for Smart Warehouses,” 2020. [Online]. Available: https://www.digiteum.com/internet-of-things-for-smart-warehouses. [Accessed: 02-Feb-2020].

[49] “Decision Support System—DSS,” 2019. [Online]. Available: https://www.investopedia.com/terms/d/decision-support-system.asp. [Accessed: 12-Dec-2019].

[50] “Decision support system (DSS),” 2010. [Online]. Available: https://searchcio.techtarget.com/definition/decision-support-system. [Accessed: 12-Dec-2019].

[51] “Advantages of Decision Support System.” [Online]. Available: http://dsssystem.blogspot.com/2010/01/advantages-of-decision-support-system.html. [Accessed: 14-Nov-2019].

[52] “Limitations & Disadvantages of Decision Support Systems.” [Online]. Available: https://www.managementstudyguide.com/limitations-and-disadvantages-of-decision-support-systems.htm. [Accessed: 14-Nov-2019].

[53] “Disadvantages of Decision Support System.” [Online]. Available: http://dsssystem.blogspot.com/2010/01/disadvantages-of-decision-support.html. [Accessed: 02-Feb-2020].

[54] Springer International Publishing, Sustainable Shipping, 1st ed. 20. Springer Nature Switzerland AG, 2019.

[55] “Big Data in Internet of Things: IoT Data Management, Analytics, and Decision Making 2019 – 2024.” [Online]. Available: https://www.marketresearch.com/Mind-Commerce-Publishing-v3122/Big-Data-Internet-Things-IoT-12623349/. [Accessed: 02-Feb-2020].

Shared Services Business Model in ICT environment

T. Žilić
Hrvatski Telekom, Zagreb, Croatia
tomislav.zilic@t.ht.hr

Abstract - Besides technical changes, digital transformation in any business-oriented organization refers to the implementation of more efficient business models. In propulsive and dynamic environment of the ICT industry, companies and organizations should constantly re-evaluate their current business models to be prepared to exist in demanding and dynamic market and environment. This paper is describing one of the possible business models in the ICT environment, and that is a model of shared service centers. This model is about providing shared services within a certain business environment, aiming to be more efficient with ICT technology as additional leverage. Implementation of the shared services model considers both organizational and technological challenges. On the other hand, there are numerous benefits that are being expected as an outcome of implementing this model. This paper is describing a business model of shared service centers, its characteristics and benefits in the ICT business environment. Research based on relevant examples of the ICT shared service centers reveals shared services as an efficient business model in the ICT industry.

Key words: shared services, shared service centers, ICT

I. INTRODUCTION

All big organizations, in today's globalized and fast changing world, are from a business prospective under constant pressure to run sustainable business models that lead to their better performance. Their goal is to find and implement more efficient and more cost effective business models by which organizations make new value. For them, it's always a quest and a fight for improved efficiency and optimization.

The same applies to the ICT industry as well. Even more, as ICT environment is quite dynamic with a big impact of propulsive technology trends and market demands, the changes are even faster thereby making organizations to seek for excellence and efficiency. Effect of ICT technology is big, having products and services that enable and stimulates globalization, collaboration, promptness, accuracy, and many other characteristics of today's business landscape.

The business model of shared services is one of the possible means that big organizations are undertaking as an attempt to achieve efficiency and competitiveness. In big organizations, many business units share common business processes and require a corporate role to ensure their most effective use. In that manner the organizations can create synergies by aligning the internal units that provide shared services [1]. These shared service units are created in a corporate level because of economies of scale and the advantages of specialization and differentiation that these units can create. The sharing of services is driven by the need to improve service provision and to increase efficiency [2].

This paper describes a shared services business model focusing on the environment of the ICT industry. Even though the model itself is applicable to almost any big organization and no matter what industry, it is interesting to see the impact of the technology as the leverage for the shared services model implementation and utilization. Besides describing the concept of shared services, and the ICT industry as a suitable environment to implement the model, the main objective of the paper is a research of the ICT shared service centers in four different industry areas: telecoms, education sector, public services and IT companies. Examples from those various environments will present the model existence in the real sector and based on those successful implementations prove its benefits.

The paper is organized in five chapters. After the introduction as the first one, the model of the shared services is described in the next chapter, and then the environment of the ICT industry. Further on, the focus is on a research based on the analysis of relevant examples of the shared services model implementation. Finally, as a conclusion, the outcome of the research will prove shared services as an efficient and applicable business model in the ICT industry.

II. SHARED SERVICES MODEL

This chapter is going to describe what shared services stand for, and what are the major objectives of such business model. Furthermore, motives, challenges and benefits of the model are discussed.

A. What are Shared Services?

There are quite a few papers, books, articles that are describing the model of shared services and there is no easy way to put it in just a few words. One of those definitions states that shared services are a collaborative strategy in which a subset of existing business functions is concentrated in a new semi-autonomous business unit that has a management structure designed to promote efficiency, value generation, cost savings, and improved service for the internal customers of the parent corporation, like a business competing in the open market [3]. Furthermore, some characteristics which are defining the term of shared services that have been mentioned in

most publications are: the consolidation of processes within the group, delivery of support processes, cost-cutting as a major driver, focus on internal customers, alignment with external customers, separated organizational unit, and operation like a business [4]. Figure 1 depicts change within the organization when implementing shared services model and the role of the shared service center as the entity mandated to provide shared services [5].

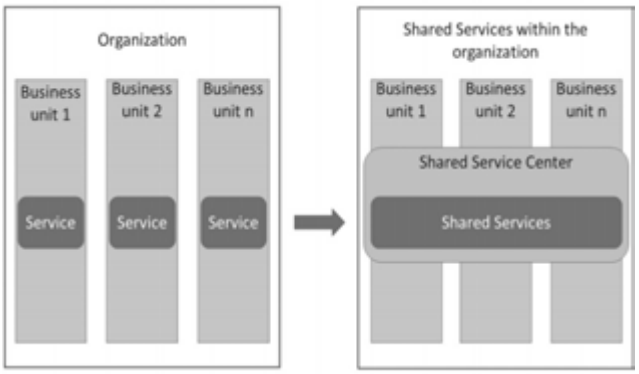


Figure 1: Shared Services Model

A lot of different business functions can be a subject of the shared services model like accounting services, HR services, ICT services, facilities management, legal and insurance services, purchasing, procurement, and some others. As ICT services are one of the shared services business functions with the most benefits and already decent maturity level, that will be the focus area in this paper [6].

Implementing shared services as a new business model is a big and demanding endeavor for any organization. It means organizational transformation, with quite a few organizational and managerial challenges to be considered. It is not just about the services in the scope, but new processes, contractual agreements and even employees are a major part of this change. On the other hand, some major benefits are expected and most often they are being fulfilled proving this as a useful business model.

B. Motives, challenges and benefits

Motives for introducing any new business model are always about improving performance and optimizing in all aspects of an organization. Managers are trying to find a way to improve while increasing competitiveness. All those efforts are driven by the need to increase the effectiveness and value of organizations. Some other drivers for shared services introduction are fundamentally about optimizing people, capital, time, and other corporate resources. Another research [2] depicts motives associated with shared services in following categories: strategic and organizational, political, technical and economic motives. Additionally, in ICT environment, there is also a big impact of technology driven motives in terms of consolidation and optimization of IT systems, application, networks and services.

On the other hand, all those motives mean the same amount of challenges along this journey. Most often, challenges are related to: power struggle within the organization, human resources, standardization of processes and resources, creating a business case for shared services model, shared services governance and implementation strategy, contracting, costs, maintaining service level quality, performance and others [7]. Despite the challenges, expected benefits are far more compelling motives for the model implementation.

Motives for shared services introduction are at the same time the benefits expected as a successful outcome. Hence, anticipated benefits of shared services can be categorized by importance, as follows: economic, technical, process improvement, strategic and organizational, and political benefits [8]. Furthermore, referring to increased efficiency and effectiveness, shared services benefits are related to: best practice processes, economies of scale, greater span of control, lower labor costs, getting the most from investment to technology, standardization, reengineering, integrated procurement, enhanced customer service focus, making the most of specialist skills, improved decision support and so on [9].

Taking into consideration all motives, challenges and benefits of the shared services business model, it is obvious that big organizations can have big expectations from the model introduction. Numerous examples, some of them are presented later in this paper, have many of the benefits named above, therefore indicating model successfulness.

III. SHARED SERVICES IN ICT ENVIRONMENT

After presenting shared services model in general in a previous chapter, the focus in this chapter is on model implementation in the ICT environment. First there is discussion on ICT business environment and impact of ICT technology on business in general. Afterwards, shared services model for the ICT business function is depicted through the role of ICT shared service centers.

A. ICT business environment

Even though the model of shared services is applicable to almost any business environment, this paper focus is on the ICT environment as this is one of the most commonly used areas for model implementations. There are a plenty of examples with successful implementations, especially over the last two decades. The ICT industry is particularly exciting business environment, extremely dynamic, fast growing and fast changing. Technology and consumer trends are constantly developing and improving, having enormous impact on business, globalization, connectivity and society in general. This leads to digitalization and digital transformation of the organizations which is taking place all over the globe and already became a standard expression in any business environment. Technical terms like big data, cloud computing, artificial intelligence, robotic automatization, automated prediction, are becoming more and more popular and present in today's vocabulary. There is no doubt that technology evolution is strongly influencing the evolution of organizations by having a rather big impact on business values. Some of them can be seen in Figure 2 [10].

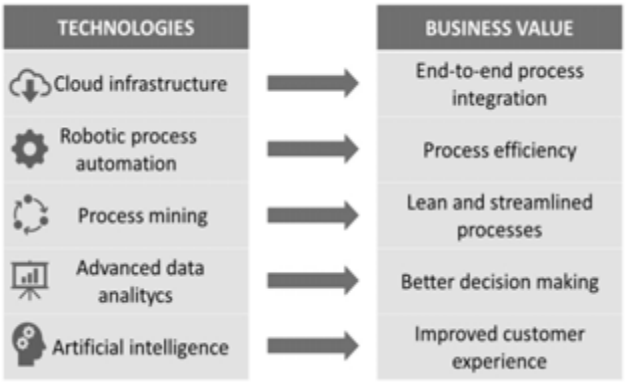


Figure 2: Technology impact on business values

Since some of the crucial prerequisites for shared services model implementation are related to terms like connectivity, collaboration, automation, processing, it is clear that ICT technology is of vital importance for the model applicability and evolution. In other words, technology is a leverage for shared services model utilization in different business environments. Therefore, ICT shared service centers can exist in any industry, as technology is an inevitable asset to any kind of big organization. Later, in the following chapter, a few examples of ICT shared service centers in various industry environments are analyzed, hence proving model successful implementation in different areas.

B. ICT shared service centers

As already mentioned, ICT functions are some of the most suitable candidates for the shared services model. Thus, a lot of examples of ICT shared service centers exists in various organizations in different industry areas. Instead of ICT services, more often used term in literature is IT service, and IT shared service center accordingly, or IT SSC as a shortened idiom. IT shared service features, that are commonly accepted in practice and literature [11] states that the IT Shared Service Centers:

- aim to consolidate the supporting processes within the company,
- focus on the organization's internal customers,
- are semi-autonomous, independent organizational units,
- provide supporting processes for the core processes of the company,
- supply multiple divisions, respectively customers,
- aim to optimize the use of resources within the organization, and
- orient themselves towards external competition.

The scope of the ICT SSCs activities is determined by the technology. The specificities of ICT activities, based on different relationships between IT and business, led to emergence of following shared service center types [12]:

- IT infrastructure – managing a common IT infrastructure, consolidating asset, standardizing processes.
- Technology Expertise – offers advisory and expertise on specific technologies.
- Application Factory – developing and maintaining standardized applications.
- Business Solutions – delivers both business processes and IT solutions.
- Operational Excellence – performs operational business tasks.

Another research [13] identified functions that are being shared from two points of view, a business perspective and a technology perspective. From the business perspective, there is sharing of:

- Business functions – organization core business functions like human resources, finance, IT, etc.
- Processes – a process or several processes in a function like payroll, budgeting, etc.
- Knowledge and expertise.

And from the technology perspective, there is sharing of:

- IT infrastructure – hardware, storage and networks that can be shared.
- IT applications – software and application suites.
- Data and information.

There is no doubt that the organization will face quite a few challenges while implementing shared services. Transformations are always rather demanding and risky undertaking, even more when including the influence of technology in the ICT sector. Most usual challenges of adopting IT shared services are following [14]:

- Ability to deliver IT services – having technical skills to deliver services in scope and establishing trust with customers.
- Communication – common understanding and transparent communication from both business and technology perspective.
- IT Services portfolio – standardization of fragmented IT landscape within the organization.
- Nature of IT services – making IT services communicable and tangible.
- Power and control – finding a balance between an effective and efficient use of resources to avoid resistance against IT SSC emerges.
- Pricing – creating awareness for IT cost and making costs transparent.
- Service Level agreements – a positive necessity to facilitate a dialog between the partners.

When talking about success factors for the shared services model, there are three perspectives to be considered, depending on their relevance: technical, organizational and personal, as shown in Figure 3 [15]. The technical perspective focuses on the establishment of the IT platform as well as IT capabilities. Secondly, the organizational perspective focuses on the organizational environment which encompasses the top management view, funding that will be used to finance the initiative and proper governance guidelines. And lastly, the personal perspective is concerned with the social and psychological aspects, individuals' knowledge and communication skills among the team.

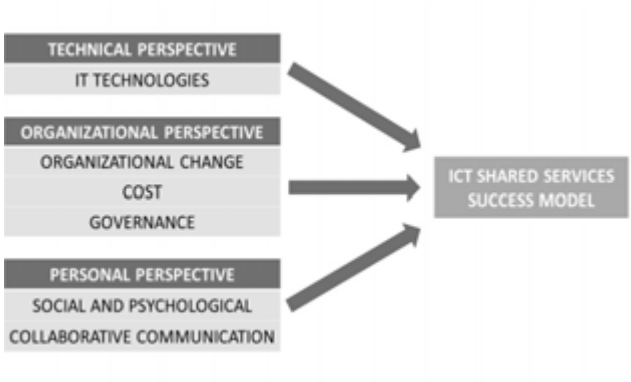


Figure 3: ICT shared services success model

ICT shared services model benefits can be observed from three different perspectives: parental organization, shared service center as a provider, and customer as a shared services receiver [16]. At the corporate group level, parental organization can find following benefits:

- major benefit as cost savings at the group level,
- centralization of common activities and standardization of IT technology and processes,
- SSC can now represent a whole group to external stakeholders.

ICT SSC units themselves, as service providers, have several positive findings as well, as follows:

- focus just on activities moved to their units and become highly specialized, thus more effective,
- resources are better utilized due to decrease of common and repetitive tasks,
- possibility to achieve additional revenue, and
- as they are now representing a whole group, SSCs became more visible and respected entities within the group and towards external partners.

And finally, SSC customers and service receivers, can find following benefits:

- they don't have to focus anymore on part of activities that were moved to SSC,
- efforts and resources from their prospective could be minimized in a matter of SSC activities,
- since SSCs are using synergies at a group level, for example to have common contracts with vendors or any other stakeholders, by economy of scale the cost savings are obviously a major argument for SSC success.

IV. ICT SHARED SERVICE CENTERS EXAMPLES

Research based on of real examples of shared services model implementation are the best way to prove how the model actually works in practice. Following industry areas are selected to present the cases: telecoms, education, public administration and IT companies. Those areas are chosen as there are numerous examples of the ICT shared services centers within them, which is a good basis for different cases comparison. Even though telecoms can be considered as IT companies, within this research there is a distinction with telecoms as a specific part of ICT sector that are mostly incumbents, while IT companies are enterprises in regards of providing software, hardware and related IT services.

A. Telecoms

Telecom operators, as big and often complex organizations, especially the incumbents, are suitable candidates for implementing shared services business model. In nowadays business environment, telecoms are facing many organizational challenges like [17]:

- how to evolve the organization,
- what is the right organizational model for managing complex and diverse group of affiliates,
- how to turn product and innovation organization into an engine of growth,
- how to push performance of the organization,
- what are needed capabilities to remain competitive,
- how to ensure effective management of mergers and integrations.

One of the most successful examples was conducted by Vodafone group which is one of the biggest organizations and brands among telecoms worldwide. During the big transformation program in 2006. Vodafone group introduced shared services business model by creating Vodafone Shared Services organization. The idea was to implement a new future-proof business model, strengthen the company's competitiveness, establish better cost controls, secure the preconditions for acquisitions and innovations, and finally consolidate the company's position in the global market [18]. Today, Vodafone Shared Services is an organization with more than 20000 employees, serving more than 150 organizations group wide, and monthly handling over 4 million transactions. This organization is handling numerous business functions in terms of shared services like human resources, finance, service desk, sales and marketing, and of course ICT services. From the technology perspective, some of the ICT shared services in scope are: technology services like infrastructure and applications, network services, analytics and business intelligence, and communication services [19]. Taking in consideration organization size and services in the scope, Vodafone Shared Services is among the best implementations of the shared services business model in the ICT industry.

Another case among telecoms was performed by Deutsche Telecom group, well known ICT brand and big global organization, which implemented technology shared service centers in 2012. It was a slightly different approach than the previous example, as there were several shared service centers for each technical area or parts of telecom network/service to be more precise. Examples of such newly established entities as telecom network functions were IMS (Internet Protocol Multimedia Subsystem) SSC, IPTV (Internet Protocol Television) SSC, RAN (Radio Access Network) SSC, EPC (Evolved Packet Core) SSC, etc. [16] The idea behind was the same, to implement a more efficient and more cost effective business model within the ICT function.

Many other telecoms like Telenor, Orange, British Telecom, AT&T are successfully using this model within the ICT business function.

B. Education

Another suitable area for ICT shared services is the education sector. There are a lot of examples of the ICT shared service centers in universities all over the world. United Kingdom Universities, as one of the biggest higher education organizations in the world, introduced a shared services model for the ICT function through JISC (*Joint Information Systems Committee)*, an organization entitled to provide digital services and infrastructure to UK universities and colleges [20]. JISC aim is to:

- operate shared digital infrastructure and services,
- negotiate sector-wide deals with IT vendors and commercial publishers,
- provide trusted advice and practical assistance for universities, colleges and learning providers,
- deliver considerable collective digital advantage, financial savings and efficiencies.

Model for shared services delivery in UK higher education is shown in Figure 4 [21].



Figure 4: Shared services in UK higher education

There are many similar examples of ICT shared service centers in higher education sector, especially in the United States, Australia and Europe. Over the years universities found this model to be rather effective for the ICT function.

C. Public administration

Public administration is yet another interesting area for shared services introduction. Public services sector is a bit different environment than the private sector, hence some additional factors like political impact and public pressure should be taken into consideration while implementing the model. Nevertheless, there are quite a few examples of shared services implementations as the governments are trying to carry out the digital transformation of public administration and services.

The government of Canada introduced shared services model as one of the most ambitious attempts in public administration. In 2012. it set up an organization named Shared Services Canada with a long-term, ambitious plan to centralize and consolidate numerous critical IT and telecommunications functions across 44 of its agencies and departments [22]. It took a very focused approach to IT consolidation that involves data centers consolidation, email systems consolidation and IT network consolidation. Today, Shared Services Canada delivers digital services to federal organizations like cyber security, data centers, cloud services, workplace technology, email, telecommunication and networks [23].

Another recent example was conducted by government of Croatia in 2019., through Shared Service Center introduction as the public cloud service [24]. This is one of the key projects of the Ministry of Administration that will integrate state information infrastructure and enable the sharing of information and communication technology as well as the application solutions with the aim of rationalizing them. The project aim is to integrate 300 institutions into the state cloud by 2022.

A lot of other examples of ICT shared service centers exist in public administration all over the world, like the United Kingdom, Ireland, Australia, United Sates, Malaysia, Estonia and many others.

D. IT Companies

The last area of the research is within technology organizations, IT companies engaged in the development of their own software and hardware products, as well as the provision of appropriate services related to the implementation, integration and maintenance of their products. On the other hand, these are companies whose technical solutions potentiate shared services model for other organizations. The importance of digital technology is therefore beyond question in the implementation of shared services models, and some of the latest technological trends that are ICT companies emphasizing in this context are cloud computing, robotic process automation, artificial intelligence, big data, etc.

There are a lot of examples since most of the big IT corporations are using the shared services model, and at the same time providing services and support to other organizations in shared services model usage. Some of them are already considered as pioneers in shared services utilization, for example, IBM one of the most well-known global IT companies that began to apply the shared service center model more than 25 years ago. In a period from 2005. to 2010. IBM achieved efficiency improvement of 4 billion US dollars [25]. Then, there is HP or Hewlett-Packard with an organization called HP Global Business Services, that is providing solutions for deploying shared services models not only internally but also to its external customers [26]. Furthermore, it is worth mentioning SAP as another global corporation that is dedicated to the development and integration of data processing systems and is the world's strongest brand in the context of financial data and systems management. It is SAP, with its Enterprise Resource Planning (ERP) system, that has enabled many other organizations to implement shared services models and is an undisputed leader in that area [27]. Another global giant is Siemens AG, a multinational company headquartered in Germany with a tradition in various technologies over 170 years. Almost twenty years ago, Siemens recognized the potential of the shared services model and founded an internal organization called Global Shared Services to provide shared services for different business functions [28].

There is no question that shared services model is applicable in any big global IT organization. On the other hand, enterprises that have emerged over the past few years, like examples from the global IT startup scene, do not mention the shared services model so much because they often use it from their beginnings and do not undergo the sometimes rather painful phase of the transformation from one business model to another.

V. CONCLUSION

Optimization, consolidation, standardization and efficiency are some of the concepts that run throughout this paper, and those are the basic postulates of the shared services model. ICT technology itself, in the context of the ubiquitous digitization of organizations and society, has proven to be a necessary basis for realizing the shared services model, and a leverage for the further evolution of business organizations. Technology and digitalization have enabled globalization and full-scale economies to emerge. Thus, the development of the shared services model has been following in the footsteps of technology development, and within the last twenty years, it has been realizing shared services potential by using the capabilities of ICT technology.

This paper deals with a shared services business model in ICT environment, looking at the technology impact to the model existence and evolution. General shared services model has been described, focusing on the ICT shared service centers, examining motives, challenges and benefits of model introduction. Finally, the research based on various examples has shown the model applicability in both private and public sectors. In this research it was interesting to look at the features of a shared service center as an effective business model in the ICT technology and industry environment. To conclude, selected case studies from telecoms, educational institutions, public administration and technology companies have shown that the model is very applicable and indicates very good results in achieving greater efficiency of large organizations.

The conclusion is that the shared services model has proven itself, both in the ICT sector and in various other environments. The maturity level achieved through the years of application of the model, as well as its evolution, prove that its life span is not yet over. Numerous scientific and professional papers, as well as many examples of successful application of the model in the ICT industry environment, confirm shared services as a successful business model.

LITERATURE

[1] R.S. Kaplan, D.P. Norton, "The strategy focused organizations", Harward Business School Press, 2001.

[2] M. Janssen, A. Joha, "Motives for establishing shared service centers in public administrations", International Journal of Information Management, Vol. 26, No. 2, pp.102–116, 2006.

[3] B. Bergeron, "Essentials of Shared Services", John Wiley & Sons Inc., Hoboken, New Jersey, 2003.

[4] V. Schulz, W. Brenner, "Characteristics of shared service centers", Transforming Government: People, Process and Policy, Vol. 4 Iss: 3, pp.210 – 219, 2010.

[5] M. Janssen, A. Joha, "Understanding IT governance for the operation of shared services in public service networks", International Journal of Networking and Virtual Organizations, 4(1), 20-34, 2007.

[6] M. Suska, A. Weuster, "Shared Services: Multiplying Success", PricewaterhouseCoopers, 2016.

[7] A. Knol, M. Janssen, H. Sol, "A taxonomy of management challenges for developing shared service arrangements", European Management Journal, 32(1), 91–103, 2014.

[8] S. Miskon, W. Bandara, E. Fielt, "Understanding the benefits of IT shared services: Insights from the Higher Education sector", Proceedings of the Pacific Asia Conference on Information Systems (PACIS 2017), Langkawi, Malaysia, 2017.

[9] Deloitte MCS Limited, "Shared services handbook – Hit the road, A practical guide to implementing shared services", 2011.

[10] M. Suska, A. Weuster, "Shared Services - Digitalise Your Services", PricewaterhouseCoopers, 2019.

[11] V. Schulz, A. Hochenstein, F. Ubernickel, W. Brenner, "Definition and Classification of IT-Shared-Services-Center", AMCIS 2009 Proceedings, Paper 265, 2009.

[12] Capgemini Consulting, "IT Shared Services Center – A powerful organization to achieve cost savings, process efficiency and even business transformation", 2013.

[13] E. Fielt, W. Bandara, S. Miskon, G. Gable, "Exploring shared services from an IS perspective: a literature review and research agenda", Communications of the Association for Information Systems, 34(1), pp. 1001-1040, 2014.

[14] F. Ulbrich, V. Schulz, W. Brenner, W, "Generic Management Challenges of Adopting IT Shared Services", AMCIS 2010 Proceedings, Paper 281, 2010.

[15] N.M. Hashim, N.M. Ali, N.S. Abdullah, S. Miskon, S.H. Huspi, "Success factors model for ICT Shared Services", 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), pp.1-6, 2017.

[16] T. Žilić, V. Čošić, V., "Implementing Shared Service Center in Telecom Environment as More Efficient and More Cost Effective Business Model", MIPRO 2016 Proceedings, 716-719, 2016.

[17] McKinsey & Company, "Reshaping telco organization to meet the industry's new challenges", 2011.

[18] M. Kresak, L. Corvington, F. Wiegel, G. Wokurka, S. Teufel, P. Williamson, "Vodafone answers Call to Transformation", 360° – The Business Transformation Journal, No.2, pp. 54-67, 2011.

[19] Vodafone Shared Services Website, retrieved form: https://www.vodafone.com/what-we-do/services/vodafone-shared-services

[20] JISC Website, retrieved from: https://www.jisc.ac.uk/

[21] I. Herbert, A. Rothwell, "Collaboration and shared services in UK higher education: potential and possibilities", London: Efficiency Exchange, 50pp, 2015.

[22] Computer Economics – Management Advisory, "Shared Services Canada: IT Consolidation on a Grand Scale", 2013.

[23] Shared Services Canada Website, retrieved from: https://www.canada.ca/en/shared-services.html

[24] Government of Croatia Website, retrieved from: https://uprava.gov.hr/print.aspx?id=13434&url=print

[25] D.L. Urso, M.T. Dunham, C. Passi, M. Overstreet, J. Viezel, N. Harling, "Enterprise transformation: The IBM journey to Value Services", IBM Journal of Research and Development, 56(6), 1:1–1:11, 2012.

[26] Hewlett-Packard Development Company, "The HP Global Business Services story", 2012.

[27] SAP Website, SAP Shared Service Framework, retrieved from: https://www.sap.com/products/shared-services.html

[28] Siemens AG Website, retrieved from: https://new.siemens.com/global/en/products/services/gbs.html

Optimization of cargo container loading on railway wagons

S. Aksentijević*, E. Tijan**, M. Jović*** and N. Munitić****
* Aksentijević Forensics and Consulting, Ltd., Viškovo, Croatia
sasa.aksentijevic@gmail.com
** University of Rijeka, Faculty of Maritime Studies/Department of maritime logistics and management, Rijeka, Croatia
etijan@pfri.hr
*** University of Rijeka, Faculty of Maritime Studies/Department of maritime logistics and management, Rijeka, Croatia
jovic@pfri.hr
**** Ministry of the Sea, Transport and Infrastructure, Zagreb, Croatia
natasa.munitic@mmpi.hr

***Abstract* - The aim of this paper is to research the optimization of loading cargo containers on railway wagons during the formation of railway compositions, by matching technical limitations of the wagons, as required by the wagon engineering discipline ("wagon loading schema" – includes cargo position, weight, and axle load, among other parameters), limitations of the railway (maximum speed, railway class, axle load, cargo height etc.), as identified by the "State of the railway network" document issued by the Railway Infrastructure Manager HŽ Infrastruktura Ltd., and the commercial bill of lading. As a part of the research, a mobile Web application has been created and tested with the key pilot stakeholder, confirming the initial hypothesis that the optimization methodology can successfully be used in optimization of cargo container loading and placement on railway wagons, increasing efficiency and decreasing cost and harmful emissions.**

Keywords - railway, wagons, intermodal transport, liberalization, ICT

I. INTRODUCTION

In EU, the attention to efficiency and safety of international railway freight transport has grown in recent years [1]. With Croatian accession to the EU on the 1st of July 2013, a new Railway Act came into force [2], which consistently and integrally regulates the utilization of the railway system in the Republic of Croatia. With the new Railway Act the cargo transport market in Croatia has been liberalized. Croatian accession to the EU has enabled free access of foreign railway cargo operators to national railway infrastructure under the same conditions that apply to national cargo operators. At the moment, there are ten railway cargo transporters registered in Croatia, including HŽ Cargo, previously the only one, and also a state-owned railway cargo operator. Currently, alternative operators have a market share of more than 50 % [3].

A final goal of the act is to achieve harmonization with the European legislation and conditions on the competitive railway services market. The railway operators and infrastructure maintenance and oversight are separated and functioning according to the free market principles, the main prerequisite for the integration of this market segment into the EU market.

II. THEORETICAL FRAMEWORK

According to the Railway Act (OJ No 94/13, 148/13) [2] whose coming into effect has determined a beginning of the railway transport market liberalization in Croatia, the market consists of services and entities. Thus, defined market is operated by the infrastructure manager and operators of service facilities providing the railway services selected as needed in different groups: minimum access package, access to service facilities and to the services provided in these facilities, including track access to service facilities, additional and ancillary services.

Railway infrastructure is defined by the Railway Act as a "public good in general use, which under equal and transparent conditions should be available to all railway undertakings" [4]. The liberalization of the railway market tends to improve the quality of rail services for the railway transport users, at the same time reducing the maintenance costs of railway infrastructure and traffic management for the amount of compensation which is collected from the railway.

Railway Act imposes a non-discriminatory approach to railway infrastructure, and a possibility for legal persons that are not railway transporters to lease sections of infrastructure. The access of railway transporters to terminals and sea and river ports has also been regulated. The transporters have the obligation to conduct business in accordance with market principles and apply the principle of the freedom of service provision. Marked increase of the cargo transport on the terminal and port facilities in the Port of Rijeka has an important role in organization of business processes of cargo operators.

Railway Safety Agency [4] serves as an independent regulatory body ensuring equal market conditions for all the operators. It also establishes the conditions for railway transport operation, monitors railway infrastructure status, railway services market, provides services of the railway infrastructure management and the conditions for usage of

the infrastructure, the principles and procedures that apply to introducing and defining fees for railway services and assigns railway infrastructure capacities, the criteria for issuing and cancelling permits to railway operators, and the services of railway transport of special interest for the Republic of Croatia which a partly state-funded. It also creates a set of criteria according to which operators need to comply with the conditions monitored by the Agency for Railway Safety, that also issues certificates of safety and other related certificates.

Sea (and river) ports may have a critical role in the increase of rail transport due to the fact that a large part of freight is carried by sea [5], [6]. This is the case in the port of Rijeka, located in an important geographic position [7] whereby its capacities must be well positioned and recognized on the traffic map of Europe in order to channel large volumes of cargo to its terminals. Considering the volumes of cargo that are unloaded from ships, the rail has an advantage over trucking because one train can transport a larger quantity of goods [8], it is more environmentally friendly [9], [10], [11], [12] and cheaper [13], [12]. Consequently, ship traffic increase in seaports is directly and positively correlated with the increase of the transport of goods by rail.

With the ongoing and well recognized trend of containerization, even the cargo that was previously typically break-bulk cargo is nowadays placed in containers and transported as such. Therefore, it is of great interest for all cargo operators to optimize the process of container loading and wagon utilization, respecting the state of the railway rules.

III. BUSINESS LOGIC OF CONTAINER LOADING ON WAGONS

When loading cargo on wagons, it is extremely important to optimize loading in relation to the wagon capacity. The current mode is that operators, according to experience and by using simple Excel spreadsheets, create wagon utilization calculations and loading plans. Such primitive work cannot ensure the ideal utilization of wagon capacity. The result is a loss of space and capacity on the wagon, inefficient work and inefficient rail transport. The solution is to create tailor-made software that will independently calculate the loading plan based on cargo and wagon data. This will speed up the loading process and increase the productivity of rail transport. Ultimately it will also contribute to increased competitiveness of the transport route.

The participating partner of TalkNet project [14], Port of Rijeka Authority, has contracted Aksentijevic Forensics and Consulting, Ltd., to provide the business process logic and application development for the optimization of loading containers on wagons. The Talknet project aims to encourage and improve coordination among key stakeholders to strengthen the integration between ports, inland terminals, transport operators and policy makers. Multimodality optimization and eco-innovation solutions are the two main fields of cooperation to establish transnational stakeholders' networks for freight transport

A comprehensive analysis was undertaken in August 2019., that involved the following:

- Identification of stakeholders and railway operators,
- Identification of technical data on utilized wagons,
- Identification of technical data on containers used,
- Business process analysis of wagon cargo loading,
- Translation of business case to application requirements,
- Identification of application configuration and use-cases.

After thorough analysis, a leading cargo rail service provider in the region was chosen for the key pilot action. A management buy-in was obtained (the management was committed to allocating time resources in order to fully map the processes and provide insight into applicable business needs and timely delivery of data, documents and information required to build the application).

The application should present high usability and portability, because the employees working with the cargo wagon composing applications are remote workers and need accessibility outside of the office. As a suitable delivery method, Android application was selected, with the alternative being a Web-accessible application compliant with mobile phone and tablet browsers.

Container train is composed of a number of wagon cars that can be owned by the cargo operator, or can be leased. Each wagon type is characterized by specific technical data, among which most important are:

- Number of axles
- Weight
- Maximum total weight
- Length
- Container loading schema.

Wagons are loaded with containers, and each wagon depending on the type, can carry one or more containers.

Containers are standardized and they can be divided by their length and other characteristics [15]. Most commonly used are 10 ft, 20 ft, 30 ft, 35 ft, 40 ft and 45 ft. intermodal containers. Furthermore, containers can have additional technical capabilities, for example, they can be high format containers (called high cube containers), open top containers, collapsible (foldable) containers, or reefer – refrigerated containers.

A task that railway cargo operators are facing is to form a train using the least number of wagons (effectively, loading each utilized wagon to the maximum of its capacity). During the pilot action, executed as a part of Interreg TalkNet project, it has come to the researchers' attention that cargo operators require an application that would be equally operable using workstations and mobile devices like smartphones and tablets. The envisaged application needs to achieve loading and distribution of the containers that need to constitute one shipment to available train wagon composition to be transported from

point A to B and show them visually/graphically. Requirements for the application configuration that were identified during the process modelling were:

- Ability to enter wagon types and data
- Ability to enter container models and data
- Ability to enter and save train composition (up to 30 wagons each)
- Ability to exchange saved train composition data (e.g. by email, using Google Drive, or other dedicated server storage) with other app users

Furthermore, certain application functionalities were identified that constitute a minimum viable product (MVP) for this purpose. They are as follows:

- Input of the number and type of containers to be transported using bill of lading document
- Load saved train composition to be used
- Automatically distribute containers to wagons according to business rules
- Possibility of a "manual" mode: drag and drop of containers to wagons after the composition has been formed
- Save loaded cargo train composition

The application also needs to follow the basic rules of information security, meaning that different users should have separate usernames and logins, so every user can access only their own data related to train composition and end result, but can use predefined container types common for all users. Configuration data has to envisage wagon types (models), wagon-related data and loading schemas for predefined types of wagons. Output results need to be entered to final Excel loading lists that can be used by the field personnel to facilitate wagon loading, but also for data exchange with ERP and other systems.

During the business modelling, certain characteristics of the wagons needed to be considered by the analysts as inputs for the optimization process. The key pilot partner uses six different types of wagons, and data that has to be entered into the model includes:

- Wagon type name (e.g.: Sggrs(s) 80', Laags...)
- Wagon tare weight (own weight)
- Wagon load weight (in this case, cargo container(s) total gross weight)
- Wagon gross weight (tare weight + container weight with loaded cargo)
- Loading schema (how containers can be distributed on the wagon depending on container type – predefined by the engineers)
- Number of axles
- Max load per axle (should be generally 20 t, but depends on the railway section)
- Wagon "registration number" ("license plate") – a list of registration numbers should be entered in the system, belonging to (property of) an appropriate wagon type

Train composition characteristics that need to be respected for Rijeka container terminals are the following:

- Max. length of train composition in Rijeka: 505 m
- Max container load: 72 TEU
- Max number of wagons: 36 (capable to transport 40' container each).

A creation of train composition also needs to follow certain logic embedded inside the application. User needs to enter the number of wagons to be used in a composition and create a train composition by selecting available wagons in sequence. User has to upload cargo loading list in Excel form in the system. The system has to distribute containers to wagons, applying the following logic:

- Goal: maximum allowed load on each wagon
- Dangerous cargo containers have to be placed single at the wagon
- Refrigerated cargo containers have to be placed at the end of composition, if possible
- Axle load cannot exceed 20 tons per axle
- Warning has to be issued if high cube (HC) container is placed on "C intermodal wagon"
- Containers need to be placed according to wagon's loading schema respecting set goal

The end result of the optimization should be a list of wagons with loaded containers, exportable to Excel, including train composition, tare weight, net weight, and gross weight. Also, the output should provide visual representation of the cargo composition and functionality to manually rearrange cargo containers placed on wagons after optimization; again, with final export to Excel/saveable, and notification that cargo composition is not optimized, but manually rearranged.

IV. SYSTEM DESCRIPTION AND DISCUSSION OF RESULTS

The system Frontend uses Bootstrap and the following Bootstrap modules: jQuery, rangeslider.js, NProgress, FileSaver and SheetJS. Nginx is used to serve static files, and other bidirectional communication is achieved using Socket.IO. The server uses Linux Debian Stretch OS. Backend stability is ensured by using Supervisor module for Node.js (v10.16.3) run within screen instance. Database used is MariaDB v10.1.38. Server is leased VPS and public access is achieved accessing http:\\www.prailway.host.

Accessing the URL, the user is presented with the login mask. The user can enter username and password, select option to be remember on that particular device, initiate login into the system or push the Contact button, invoking default e-mail client to contact the administrator in case of a forgotten password or a need to create a new account. The login mask is shown in Figure 1.

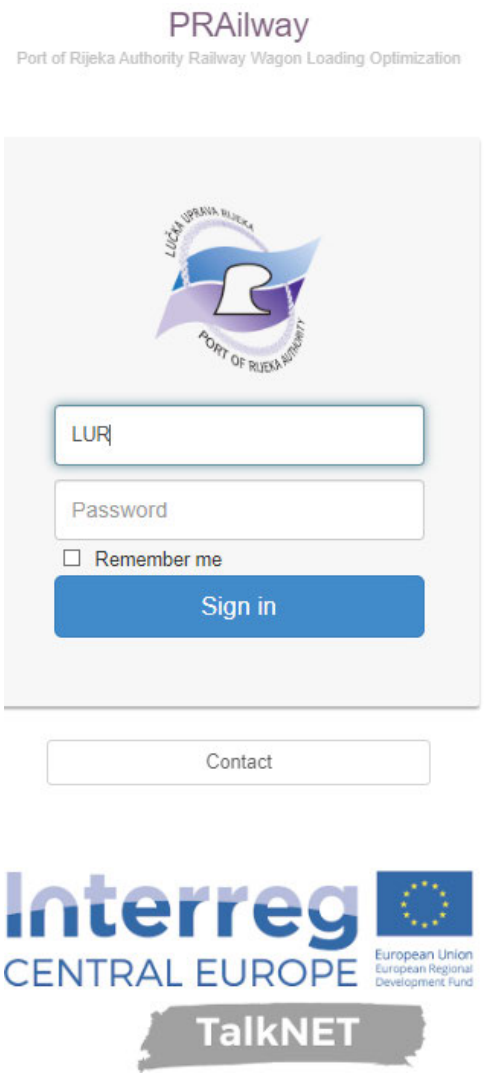


Figure 1: Login mask for Prailway optimization application placed on http:\\www.prailway.host

After logging in, the user is presented with the main panel. There are two different panels for administrators and regular users. Administrators have the capability to add new users, assign access levels and create and reset entry passwords, while regular users can only use the core of the application. By using the functionalities of this panel, the user can load previously saved train composition, exit from the application or create a new train composition.

If the user selects the latter option, he needs to enter the goal train length (maximum allowed train length in meters), and after that, the user can form the composition, depending on the number of available (leased) cargo wagons.

The system contains the embedded data for six wagon types used by the key pilot action partner: LGS-Z, KGS-Z, Rgs-Z, Laags, Sggmr22 and Sggr(s). This screen is shown in Figure 2. The system tracks added wagons in comparison to pre-set allowed train composition length, and after the user has entered the available wagons, he is presented with the mask to upload the Excel bill of lading in form exported from existent cargo operator's ERP.

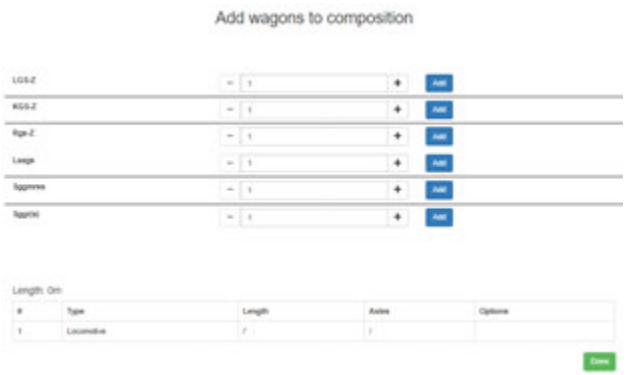


Figure 2: Screen for train wagon composition selection

The application is able to parse the uploaded Excel list and to discern relevant data used for cargo optimization, which includes container type, container weight, type of goods, code, RID code (dangerous cargo) and other relevant data.

After the bill of lading is uploaded to the server, it automatically optimizes the cargo load, considering all constraints, wagon loading schemas, types, characteristics and weight of containers to be loaded, their size and other train composition constraints. The end result is a graphic representation of the optimized train composition, as shown in Figure 3. In this example, there is a large number of wagons that are not necessary to transport the cargo on a particular bill of lading, because they remain empty after the optimization and can be excluded from the composition.

Finally, as required by the project, the user can click (in case that application is used on a workstation) or swipe (in case of touch sensitive devices) on a container loaded to a wagon, and retrieve container info, including size, TEU, container number, client, seal, size, transported goods, NHM, colli, tare weight, weight and sum of weight and tare. The user can then unload the container from the wagon to the server clipboard (or do the same with several containers), and then paste them to another wagon according to personal preferences.

It is evident from Figure 3 that in this particular case, many wagons that were intially planned, after applied optimizations, are in fact not neccessary to transport pre-defined cargo list, and can be exluded from the train composition.

The layout of the wagon unload screen for manual placement after optimization is shown in Figure 4 on the following page.

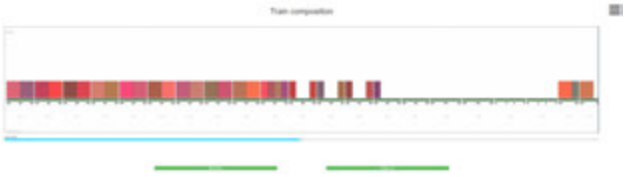


Figure 3: Output of optimization process

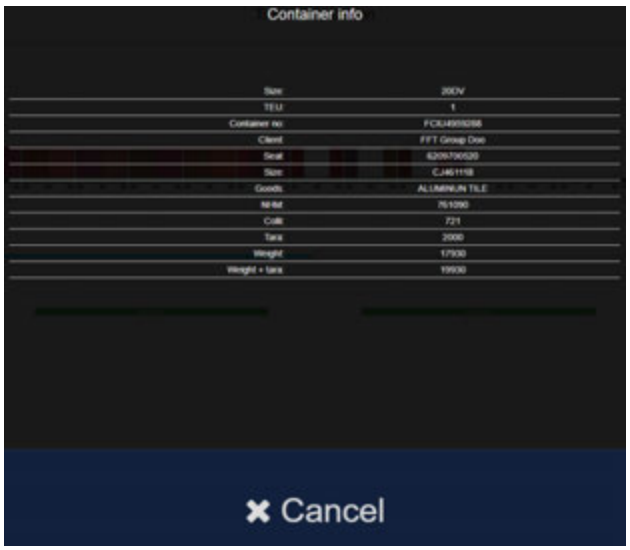


Figure 4: Manual placement mode

Finally, the user can export the formed composition in Excel format, that includes all data contained in the input list, but enriched with wagon composition and container numbers loaded to containers with positions (left, center left, center, center right or right-positioned) according to technical loading schemas. This Excel spreadsheet can be used as a container to be further uploaded to ERP systems, or directly by railway personnel to oversee and coordinate container loading on wagons.

V. CONCLUSION

When loading cargo on wagons, it is important to optimize loading, considering the wagon capacity and state of the railway rules. As a part of TalkNet Interreg Central Europe project, the lack of optimization and structured methodology in optimization of contained loading on wagons has been noted.

After the funding was secured, a key pilot project partner has been identified in form of a railway cargo transport company that provided an insight into the business processes of container cargo transport by using railway.

Business process modeling was composed of several steps that included the analysis of input data and bill of lading, characteristics of used containers (size and type), technical engineering analysis of used wagon types, analysis of the railway infrastructure that imposes limitations on rail traffic and, finally, the analysis of technology and methodology to be used during the optimization process.

After the creation of the functional specification, IT technology analysis process was performed, where adequate server resources and development technologies were acquired and identified. The application has been created, and it was alpha and beta tested, with the last phase being the user acceptance testing. The end result was very satisfying, and a structured application was produced, providing contained loading and placement on different types of wagons that is accessible to all port stakeholders.

The achieved level of implementation and functionality has justified the project goals, as the container placement and loading are now not just reliant on operator's experience and heuristics, but are structured and algorithm-supported. However, the application can be significantly upgraded. For example, at the moment, it is aimed towards railway leading to port of Rijeka and its container terminals. Railway rules are hardcoded in the system. Furthermore, it contains technical loading schemas for wagons types used by the key partner, but not for all possible wagons and tentatively those used by other rail cargo operators.

Possible venues for upgrade are configurable settings for limitations of the railway, and different loading schemas for other wagon types, that would enable the inclusion of other cargo operators. Considering the ongoing implementation of a Port Community System (PCS) in port of Rijeka, there is also a possibility of interfacing this application with the PCS railway module, to achieve additional synergies in the railway cargo optimization arena.

Additionally, the application could be upgraded to support other ways of exchanging data with the surrounding ICT systems. Considering that PCS and container terminal operating systems typically use EDI and XML-format data exchange, such implementation is also recommended for railway cargo loading optimization application as a part of a possible future upgrade project(s).

ACKNOWLEDGMENT

This work was supported by "TalkNet - Transport and Logistics Stakeholders Network" project (Interreg V-A Italy – Croatia 2014-2020) and by University of Rijeka under the Faculty of Maritime Studies projects.

REFERENCES

[1] G. Loprencipe, L. Moretti, T. Pestillo, and R. Ferraro, "Railway Freight Transport and Logistics: Methods for Relief, Algorithms for Verification and Proposals for the Adjustment of Tunnel Inner Surfaces," *MDPI Sustainability*, 2018. [Online]. Available: https://www.mdpi.com/2071-1050/10/9/3145/htm. [Accessed: 01-Jan-2020].

[2] Hakom, "NN br. 94/13, 148/13 Zakon o željeznici," 2013. [Online]. Available: https://www.hakom.hr/UserDocsImages/2014/propisi_pravilnici_zakoni/Zakon o zeljeznici-nesluzbeni procisceni tekst.pdf. [Accessed: 25-Jan-2020].

[3] HŽ Infrastruktura, "Pristup infrastrukturi." [Online]. Available: https://www.hzinfra.hr/?page_id=284&page_id=284. [Accessed: 12-Dec-2019].

[4] "Zakon o sigurnosti i interoperabilnosti željezničkog sustava (NN 82/13, 18/15, 110/15 , 70/17), na snazi od 2015." [Online]. Available: https://www.zakon.hr/z/649/Zakon-o-sigurnosti-i-interoperabilnosti-željezničkog-sustava.

[Accessed: 01-Jan-2020].

[5] R. A. Halim, L. Kirstein, O. Merk, and L. M. Martinez, "Decarbonization Pathways for International Maritime Transport : A Model-Based Policy Impact Assessment," *MDPI Sustain.*, vol. 10, no. 7, 2018.

[6] A. Khaslavskaya and V. Roso, "Outcome-Driven Supply Chain Perspectives on Dry Ports," *Sustainability*, vol. 11, no. 5, p. 1492, 2019.

[7] M. Jović, N. Kavran, S. Aksentijević, and E. Tijan, "The Transition of Croatian Seaports into Smart Ports," in *42nd international convention on information and communication technology, electronics and microelectronics*, 2019, pp. 1618–1622.

[8] Freightera, "Shipping by road or rail: PROS & CONS," 2019. [Online]. Available: https://www.freightera.com/blog/shipping-road-vs-rail/. [Accessed: 25-Jan-2020].

[9] E. Kurtulus and I. B. Cetin, "Assessing the Environmental Benefits of Dry Port Usage: A Case of Inland Container Transport in Turkey," *MDPI Sustainability*, 2019. [Online]. Available: https://www.mdpi.com/2071-1050/11/23/6793/htm. [Accessed: 01-Jan-2020].

[10] Y. Lu, M. Lang, X. Yu, and S. Li, "A Sustainable Multimodal Transport System: The Two-Echelon Location-Routing Problem with Consolidation in the Euro–China Expressway," *MDPI Sustainability*, 2019. [Online]. Available: https://www.mdpi.com/2071-1050/11/19/5486/htm. [Accessed: 01-Jan-2020].

[11] B. Djordjević and E. Krmac, "Evaluation of Energy-Environment Efficiency of European Transport Sectors: Non-Radial DEA and TOPSIS Approach," *MDPI Energies*, 2019. [Online]. Available: https://www.mdpi.com/1996-1073/12/15/2907/htm. [Accessed: 01-Jan-2020].

[12] Q. Hu, B. Wiegmans, F. Corman, and G. Lodewijks, "Integration of inter-terminal transport and hinterland rail transport," *Flexible Services and Manufacturing Journal*, 2019. [Online]. Available: https://link.springer.com/article/10.1007/s10696-019-09345-8. [Accessed: 01-Jan-2020].

[13] Ž. Stević, D. Pamučar, E. K. Zavadskas, G. Ćirović, and O. Prentkovskis, "The Selection of Wagons for the Internal Transport of a Logistics Company: A Novel Approach Based on Rough BWM and Rough SAW Methods," *MDPI Symmetry*, 2017. [Online]. Available: https://www.mdpi.com/2073-8994/9/11/264/htm. [Accessed: 01-Jan-2020].

[14] Port of Rijeka Authority, "Transport and Logistics Stakeholders Network – TalkNet," 2019. [Online]. Available: https://www.portauthority.hr/en/european-projects/interreg-ce-talknet/. [Accessed: 25-Jan-2020].

[15] Profile Euro, "Container Characteristics," 2019. [Online]. Available: https://www.profile-euro.com/for-the-clients/directory/container-characteristics.html. [Accessed: 01-Jan-2020].

Applying Adaptive Neuro-fuzzy Inference System (ANFIS) while Analysing Interdependecies of Tax Burden and Capital Structure of Croatian Hotel Companies

S. Brlečić Valčić*, A. Samodol* and M. Valčić**
* Libertas International University, Zagreb, Croatia
sbrlecic@libertas.hr
asamodol@libertas.hr
** University of Rijeka, Faculty of Engineering, Rijeka, Croatia
mvalcic@riteh.hr

Abstract - **Various conducted studies incorporate the effect taxation has on company's financial policy, in particular in terms of the effect that corporate profit tax has on capital structure. This paper focuses on the effect the overall tax burden has on capital structure (sources of funding) of Croatian hotel companies, due to the fact that tourism as an economy branch has great impact on total Croatian GDP. Total tax burden encompasses corporate profit tax and income tax of hotel companies. In the narrow context, the problem also refers to adjustments, in terms of preferences for capital budgeting methods (new way of capital funding, corporate management, profit payment policy and bank - company relations), as well as to solvency and capital structures which banks require, as parts of contemporary concepts and views on corporate finances. The broader context, apart from GDP growth and rational planning of budget by the current government, refers to financial and market friction. Adaptive neuro-fuzzy inference system (ANFIS), as a type of neuro network based on Takagi–Sugeno fuzzy inference system, has been used in this research as an analysis tool to observe the relations, i.e. impacts, rather than as a universal estimator or optimizor.**

Keywords - capital structure; taxation; hotel companies; ANFIS

I. INTRODUCTION

The aim and objective of each corporate financial policy rely on two major decisions. One of them refers to the quantity of one's own (personal) and borrowed finances within the capital structure, while the other one handles generation and distribution of corporate profit. Economic development should be followed by taxation policy which would serve to back up the context described.

In the last few decades the taxation policy of the Republic of Croatia has been changing, in accordance with the objectives, which, unfortunately, did not rely on development, but rather on patching up the scarce and insufficient budget.

This type of policy has not been very beneficial and did not result in sufficient economic development. Not many studies have been conducted, which would focus on these issues or the development of a particular branch of economy. This paper places focus on the impact of taxation policy in the Republic of Croatia on capital structure of hotel companies, mainly because of the role of the very branch in the GDP structure. In the narrow context, the relationship between tax burden and capital structure has been observed, while in the broader one, it is financial friction that is being monitored.

Financial friction is measured as the difference between margin productivity and yield of capital adapted to capital profits and losses, based on equity and short-term interest rates. It emphasises decreased investments in plants, equipment and inventory, which are the result of the increased gap between earned return and efficient expenditure of resources of loan-dependent companies [1]. The framework of financial friction issues must determine whether hotel companies have goal-oriented policy of selecting their capital structure and which adjustment options there are in case of deviations from the selected policy [2]. It is extremely important to monitor this context with regards to company's lifecycle because it has been proved that financial friction has a direct impact on investment decisions of start-ups [3].

It is equally important to put the development of financial market of a particular country in context [4] since the countries which have less developed financial infrastructure are facing faster development of small companies, rather than that of big ones, and they rely more on their own funding resources [5].

Simultaneously, equity market recorded a drop of regular stock turnover in GDP from 5% to 0.7% at the end of 2018, followed by continuous delisting of joint stock companies from the Zagreb Stock Exchange (ZSE, Trading data and statistics). Stock Exchange is simply not a place where equity is raised, where share structures are developed or where the culture of transparent management is nourished.

Therefore, it is no wonder that citizens and small enterprises decide to invest in their own tourism capacities, in times of extremely low interest rates on savings. In addition to all that, the central state, via their own Croatian Bank for Reconstruction and Development promotes and grants the so-called Loans in Support of the Tourism Industry, which have a positive impact on increased investments of small enterprises and citizens into tourism capacities.

II. THEORETICAL BACKGROUND

The challenge of analysing how taxation affects important financial decisions, including the ones which refer to capital structure, lies in the fact that our tax system is predominantly formal, and it only indirectly tackles basic asset features [6]. Such diverse sources of funding and assets (capital or debts) are faced with different types of taxation treatments, which are not always part of the basic function of asset maintenance. Regarding tax policy alterations, it is vital to take into account the impact that taxation has on start-ups, basically on their lifecycle [7].

Previously conducted similar researches [8] point to conclusions that there is a significant impact of corporate profit tax on capital structure of the company, meaning that lower tax burden causes lower levels of financial leverage and increased equity. These researches suggest consistency with *tradeoff theory of capital structure*.

Furthermore, authors of studies that examined the projection of corporate profit tax annulment [9], have concluded that in case of profit tax annulment, financially challenged companies would be able to accummulate profits more quickly. This type of neutral tax reform would lead to total production increase of 12.2%, which would result in large total equity increase (31.8%) and the increase in the number of companies (34.5%).

Corporate profit tax annulment enables financially challenged companies to retain a larger share of their profit and increase their investments. The possibility of profit retention is particularly relevant to newly founded companies, which are more likely to suffer from various kinds of restraints, unlike average companies.

Capital structure itself depends not only on firm-specific factors, but also on state-specific ones, referring to the state where the firm does business [10]. However, *leverage* tax effect relates to *firm-specific effects*.

In most cases companies manage to reconcile the gap between real and targeted capital structure in the amount of approximately 30 % [11]. The companies of transition economies have the tendency to be adapting their capital structure at the same pace as the countries outside the transition process, and the determinations of targeted capital structure in transition economies resembe the ones observed in the EU states [12].

The capital structure of companies which rely on debt finance is more susceptible to perfomance fluctuations. On the other hand, the ones that depend on internally generated funds address their susceptibility to performance to *cash holdings* [13]. *Cash holding* affects cashflow risk because it limits the amount invested in fixed assets, which, eventually leads to multiple understanding of risk [14]. Companies are not price-sensitive if the amount of debt is below critical value, or in other words, price mechanism does not slow down the demand of short-term financing [15].

The size of a company and a*sset tangibility* have a positive influence on the use of leverage, whereas profitability and growth options have a negative influence on what debt is used for [16]. It also depends on the related business and financial cycles [17], and the financial ones seem to appear more frequently [18] [19]. For that reason the context also refers to the availability of financial resources. Therefore, capital structure can be stable and cyclical, depending on business and financial cycles, wherein stability is usually the result of early planning and cyclical nature of expansion (which usually requires equity) or decreased volume of business (which usually goes hand in hand with increased liability) [20].

The basic nature of research on classical corporate taxation and its effects, focuses on three basic areas: capital policy, debt-equity decisions and choices related to business ownership structure and form of organisation [21]. It is of crucial importance to bear in mind that adjustment risks and expenses of equity accummulation may rationalize dispersion of productivity in a company [22]. Facing collateral restrictions and understanding the impact of financial friction on poor capital allocation and overall productivity is essential. The result of taxation reforms in the Republic of Croatia is lower financial leverage. Lower taxes reduce retention of debt due to reduced interest tax deduction.

III. DATA AND METHODOLOGY

In this research we have been using financial data and business indicators of 32 hotel companies, in the period between 2014 and 2018. Hotel companies have been divided, based on their legal form (Inc. or Llc.), so we have analysed 15 hotel companies which are joint stock companies and 17 limited liability companies.

The effects that tax has on capital structure and business indicators have been monitored taking into account two kinds of taxes, profit tax and income tax.

The initial setup of parameters for each individual model is based on clustering through Self-Organizing Map (SOM).

In order to define the impact that a particular tax has on financial business indicators in the context of capital structure, the analyses have used the share of each tax in the selected parameters; and these are the following: PD/AM (profit tax/depreciation); PDOH/AM (income tax/ depreciation); PD/PP (profit tax/revenues); PDOH/PP (income tax/revenues); PD/UR (profit tax /total expenditure); PDOH/UR (income tax/total expenditure); PD/UI (profit tax /total assets); PDOH/UI (income tax/total assets); PD/DO (profit tax /non-current liabilities); PDOH/DO (income tax/non-current liabilities); PD/KIR (profit tax /capital and reserves) i PDOH/KIR (income tax/capital and reserves).

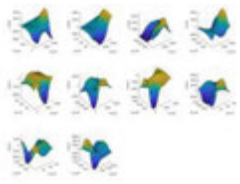
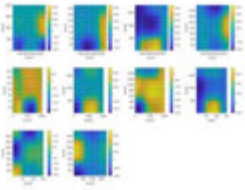
The data set is analysed through Adaptive Neuro-Fuzzy Inference System (ANFIS) (details on the used methodology in Brlečić Valčić, Samodol, 2018) because these tools are frequently used as an universal approximator in modelling nonlinear functions of multiple variables, but also in predicting chaotic time series, etc. (Jang, 1993). Using an input-output dataset, ANFIS delivers a Fuzzy Inference System (FIS), whereby the membership functions parameters are adapted by a backpropagation learning algorithm or combined with the method of least squares (hybrid learning method). Such a setup allows the FIS system to learn from training data (Brlečić Valčić et.al, 2019).

IV. RESULTS

Based on the previous selection of clustering through Self-Organizing Neuro networks, several analysis models have been created.

Model 1 Inc. observes the relation among the following parameters: *Fixed assets (Input 1), Non-current liabilities (Input 2), Liabilities to suppliers or current liabilities (short-term) (Input 3), Revenues (Input 4), Total expenditure (Input 5)* with regard to *Income tax (Output).*

Analysis results are shown in Figures 1 and 2.

Figure 1: 3D illustration of Model 1 Inc.	Figure 2: 2D illustration of Model 1 Inc.
	

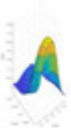
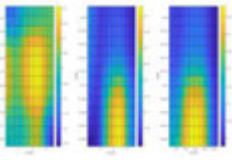
Source: designed by the authors

The largest impact of high-level corporate income tax on this type of hotel companies is the consequence of fixed assets / revenues relation (input1/input 4). Incidently, the analysis implies that at nearly all asset levels high income taxes involve lower revenues, whereas high revenues involve relatively low value of corporate income tax. As far as the impact on capital structure is concerned, or financial structure, the initial analysis has shown that there is no impact of any monitored tax (profit tax and income tax) on capital (in all its forms), i.e. on self-funding. This analysis proves that the impact of income tax on long-term liabilities exists, and it points to a conclusion that medium and high levels of long-term liabilities (Input 2) involve high income tax and, simultaneously, low revenues (Input 4). On the other hand, high long-term liabilities involve lower income tax.

Model 2 Inc. observes the relations among *Fixed asset turnover coefficient* (Input 1), *income tax in depreciation* (Input 2) and *in total assets* (Input 3) with regard to *income tax in capital and reserves* (Output).

Fixed asset turnover coefficient measures revenue in relation to the average value of fixed assets, i.e. it shows how many revenue units can be generated out of one fixed assets unit (kuna). This indicator points to the efficiency of investment policy because fixed assets (investments) are usually the largest segment of company's fixed assets.

Analysis results are shown in Figures 3 and 4.

Figure 3: 3D illustration of Model 2 Inc.	Figure 4: 2D illustration of Model 2 Inc.
	

Source: designed by the authors

All levels of *Income tax in total assets* (Input 3) cause low levels of *Income tax in capital and reserves*, along with low and very high levels of *Fixed assets revenue coefficient* (Input 1). *Income tax in capital and reserves (PDOH/KIR)* is increased at low and medium levels of *Fixed assets*

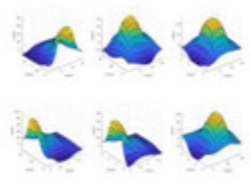
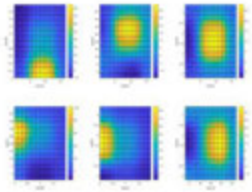
revenue coefficient, along with medium and high levels of *Income tax in depreciation*. The same conclusions can be drawn for the ratio of *Income tax in total assets (PDOH/UI)* (Input 3) and *Income tax in depreciation (PDOH/AM)* (Input 2), with regard to *Income tax in capital and reserves (PDOH/KIR)*. According to the analysis, all levels of *Income tax in total assets* (Input 3) combined with low and very high levels of *Income tax in depreciation* (Input 2) will cause low levels of *Income tax share in capital and reserves*, while medium and high values of *Income tax share in depreciation* (Input 2) combined with low to medium values of *Income tax share in total assets* (Input 3) imply higher levels of *Income tax share in capital and reserves*.

Model 3 Inc. observes the relation among *Current ratio* (Input 1), *Expense coverage* (Input 2), *Interest coverage from EBITDA* (Input 3) and *Profit tax in long-term liabilities* (PD/DO) (Input 4) along with *Income tax in long-term liabilities* (Output).

Current ratio represents the ratio between current assets and current liabilities.

Expense coverage is calculated as a ratio of financial assets and corporate expenses, not including non-financial expenses – depreciation, value adjustment and reserves.

Analysis results are shown in Figures 5 and 6.

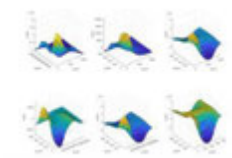
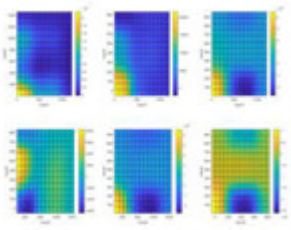
Figure 5: 3D illustration of Model 3 Inc.	Figure 6: 2D illustration of Model 3 Inc.
	

Source: designed by the authors

This analysis indicates that there are very high income taxes on long-term liabilities in almost every combination, apart from: all levels of Current ratio combined with low Interest coverage ratios from EBITDA; low levels of Current ratio followed by low profit tax in long-term liabilities (profit tax/income tax); higher levels of Financial expense coverage followed by low levels of Interest coverage from EBITDA; Very low and higher levels of Financial expense coverage followed by extremely low levels of business expense coverage (profit tax/income tax); Low levels of Interest coverage from EBITDA followed by low, medium and higher levels of (profit tax/income tax).

Model 1 Llc. observes the relation among *fixed assets* (Input 1), *total assets* (Input 2), *capital and reserves* (Input 3) and *share capital* (Input 4) in relation to *profit tax* (Output).

Analysis results are shown in Figures 7 and 8

Figure 7: 3D illustration of Model 1 Llc.	Figure 8: 2D illustration of Model 1 Llc.
	

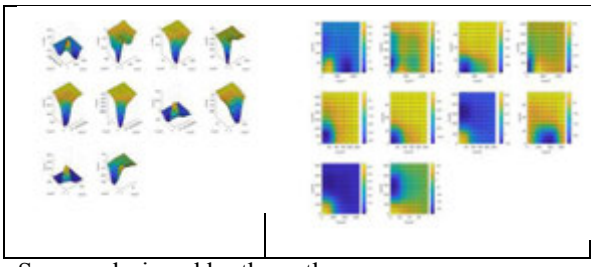
Source: designed by the authors

Higher profit tax is bound with lower levels of fixed and total assets, whereas lower profit tax is, understandably, bound with higher levels of fixed and total assets, due to depreciation. Lower levels of fixed assets, followed by lower levels of capital and reserves value and share capital value, are bound with higher profit tax. Equally, medium and higher values of capital and reserves and of share capital, followed by medium and higher fixed assets values, are bound with lower profit tax. Medium and higher levels of fixed assets, followed by low levels of share capital, indicate very low and very negative values of profit tax. Lower levels of total assets, followed by medium and higher levels of capital and reserves are bound with high levels of profit tax, while low levels of total assets, followed by low levels of capital and reserves indicate negative values of profit tax. Medium and high values of total assets, followed by all levels of capital and reserves values, are bound with medium levels of tax profit. Low levels of total asset value, followed by low levels of share capital value are bound with high values of profit tax. Medium and high levels of total asset value, followed by lower levels of share capital value indicate negative values of profit tax. Medium and higher levels of share capital values, followed by all value levels of total assets indicate low values of profit tax. Low levels of capital and reserves, followed by low levels of share capital are bound with high levels of profit tax values. Low and negative values of profit tax indicate medium and high levels of capital and reserves value, followed by low levels of share capital. Medium and high levels of share capital values on all levels of capital and reserves values are bound with low values of profit tax.

Model 2 Llc. observes the relation among *tangible assets* (Input 1), *current assets* (Input 2), *non-current liabilities* (Input 3), *depreciation* (Input 4) and *corporate income* (Input 5) in relation to *income tax* (Output).

Analysis results are shown in Figures 9 and 10.

Figure 9: 3D Illustration of Model 1 Llc.	Figure 10: 2D Illustration of Model 1 Llc.



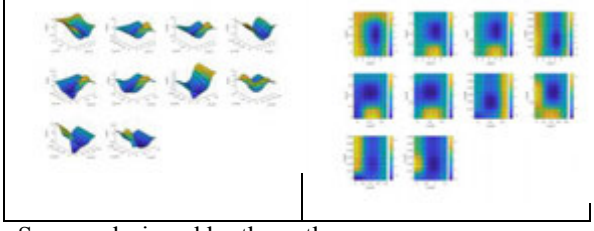
Source: designed by the authors

Lower values of tangible assets followed by low value of current assets imply higher values of income tax. Tangible asset values slightly lower than medium, followed by the same current liability rank, imply very low levels of income tax. All other combinations of tangible and current liability value imply negative values of income tax. Medium to high range of tangible asset value combined with very high rank of revenue value indicates the effect of very low rank of income tax (or around 0). Low value rank of current assets combined with low value rank of revenues implies high value rank of income tax. Medium to high value rank of current assets combined with high value rank of revenues implies low value rank of income tax. Low value rank of non-current liabilities combined with low value rank of revenues implies high value rank of income tax. Medium and high values of non-current liabilities combined with medium and high revenues are bound with very low value of income tax or values around 0. Low values of depreciation combined with low values of revenues are bound with moderately high of income tax values.

Model 3 Llc. observes *current liability coverage* (Input 1), *ROA* (Input 2), *return on invested capital* (Input 3), *profit tax share in depreciation* (PD/AM) (Input 4) and *profit tax share in total assets* (PD/UI) (Input 5) in relation to *profit tax share in non-current liabilities* (PD/DO) (Output).

Analysis results are shown in Figures 11 and 12.

Figure 11: 3D Illustration of Model 1 Llc.	Figure 12: 2D Illustration of Model 1 Llc.



Source: designed by the authors

Current liability coverage is a ratio of financial results cashflow (the sum of net business results and depreciation) and average current liabilities.

ROA is calculated as a net profit / loss and average asset ratio.

Return on invested capital represents the ratio of pre-tax profit and tax, and invested capital (capital and reserves plus financial i.e. interest-bearing liabilities).

Lower values of current liability coverage on all value levels of ROA are bound with values of PD/Do lower than 1, i.e. those which imply current liabilities of values higher than profit tax values. Medium and higher values of non-current liability coverage followed by low levels of return on invested capital indicate levels of PD/DO higher than 5, i.e. much higher values of income tax on non-current liabilities. Values of PD/DO lower than 1 are bound with lower values of current liability coverage, combined with all values of return on invested capital, or all values of current coverage combined with high rank of return on invested capital value. The ratio of current liability coverage and profit tax share in depreciation (PD/AM) cause almost identical effcts on PD/DO. Levels higher than 1 of PD/DO are created by ratio of very low current liability coverage and low to medium values of profit tax share in total assets (PD/UI).

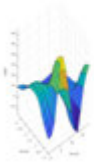
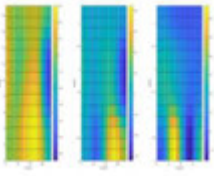
ROA levels in the amounts of 20 – 40%, combined with low levels of return on invested capital assume higher levels of profit tax from non-current liabilities. Lower levels of profit tax from non-current liabilities are bound with lower levels in the amount of 20%, or higher in the amount of 40% of ROA, implying very low or very high levels of return on invested capital. We can draw almost identical conclusions regarding ROA / profit tax share in depreciation (PD/AM) ratio. Medium and high values of return on invested capital combined with low values of profit tax share in depreciation (PD/AM) are bound with higher amounts of profit tax from non-current liabilities (PD/DO from 5 to 10). Equally so, higher amount of PD/DO assumes higher to medium values of PD/AM, combined with low values of return on invested capital. Lower levels of PD/DO (i.e. higher amount of non-current liabilities from profit tax) can be bound with medium to high values of return on invested capital, combined with very high values of PD/AM (it assumes 1.5 times higher depreciation from profit tax).

Low levels of return on invested capital, combined with higher, medium and high values of profit tax in total assets (PD/UI) are bound with higher values of profit tax from non-current assets (PD/DO from 5 to 10). Higher values of non-currrent liabilities from profit tax value (PD/DO lower than 1) can be expected in case of a combination of high values of return on invested capital and all values of profit tax share in total assets (PD/UI). If we monitor the ratio of profit tax share in depreciation (PD/AM) and in total assets (PD/UI), then low levels of PD/AM combined with medium values of PD/UI assume higher amounts of profit tax from non-current liabilities (PD/DO 5 to 10). However, lower values of PD/DO (higher amount of non-current liabilities from profit tax) can be expected if combined with higher profit tax

from depreciation (PD/AM 1.5) at all levels of profit tax from total assets (PD/UI).

Model 4 Llc. observes the relation among *financial coefficient* (Input 1), *rate of return of interest-bearing liabilities* (Input 2), *investment coefficient* (Input 3) with *income tax share in capital and reserves (PDOH/KIR) (Output).*

Analysis results are shown in Figures 13 and 14

Figure 13: 3D Illustration of Model 1 Llc.	Figure 14: 2D Illustration of Model 1 Llc.
	

Source: designed by the authors

Financial coefficient shows the relation of exploitation of borrowed capital (liabilities) and personal capital (capital and reserves).

Investment coefficient shows investment – income ratio, i.e. which amount of their income a company invests in fixed tangible and intangible assets.

Medium to high levels of financial coefficient followed by low to high levels of rate on return on interest-bearing liabilities are bound with much higher amounts of tax on income from capital and reserves (PDOH/KIR 10 to 30). Lower values of tax on income from capital and reserves (PDOH/KIR lower than 1) are to be expected with lower levels of financial coefficient at all levels of rate on return on interest-bearing liabilities.

Very high financial coefficients, accompanied with low investment coefficients assume higher levels of tax on income from capital and reserves. Lower levels of tax on income from capital and reserves (PDOH/KIR lower than 1) can be attached to medium levels of financial coefficient accompanied with low levels of investment coefficient. Lower levels of tax on income from capital and reserves (PDOH/KIR lower than 1) assume the combination of higher rate on return on interest-bearing liabilities and low levels of investment coefficient.

V. CONCLUSION

Even though the previously conducted studies have mostly been monitoring the impact that corporate profit tax has on business, and consequently, capital structure and funding of it, this study has shown that this kind of impact on Croatian hotel companies has to be monitored with regard to two tax categories, i.e. profit tax and income tax (employees' salaries).

Only if we arrange the context in that manner, it can be related to the conclusions of similar studies conducted so far, regarding the following: formality of tax system, which tackles basic asset features only indirectly; lower tax burden which causes lower levels of financial leverage; business performance improvement in relation to lower levels of tax burden reduction; potentially better planning of targeted capital structure, and the impact of financial friction on poor capital allocation and overall productivity.

Equally so, analyses conducted in Croatian hotel companies have brought to surface that, in order to draw more plausible conclusions, it is essential to monitor the impact that individual tax forms have on businesses specifically, taking into account their legal entity types. The same conclusions do not apply to joint stock companies and to limited liability companies.

The conclusions referring to joint stock companies are as follows: corporate profit tax cannot be related to the influence of any of the conclusions from the categories of previously conducted researches, except business performance that refers to liquidity and solvency, regarding its share in long-term liabilities; income tax can be related to all five mentioned categories of similar and related studies; since long-term assets present a large item in these companies, income tax levels affect the amount of income generated by these assets, or, in other words, analyses show that higher levels of income tax will affect lower sales revenues; in relation to this conclusion, indicators within business performance, which can be related to interdependence of tax and capital structure, using selected methodology, are the coefficient of fixed asset turnover, income tax share in depreciation and income tax share in total assets. The above mentioned indicators relate to income tax share in capital and reserves (as a more dominant source of funding); since finance structure is more concentrated on personal funds (capital), the conducted analysis shows that lower value of income tax in capital and reserves value implies either very low or very high coefficient of total assets turnover and income tax share in depreciation. This points to poor wage elasticity, compared to the value of total assets and depreciation, real estate in particular. We could say that there is a latent conflict of basic business activity (services of accommodation and catering) and real estate potential, and renting position of basic activity facilities. The particularity of this business activity tends to be escalating in recent years, in times of workforce deficit and considerable wage increase, which requires more powerful coefficient of total assets turnover. Otherwise, low level of accummulation is going to decrease the personal funds potential; lower levels of income tax share in non-current liabilities are bound with lower values of current ratio and medium and higher values of interest coverage from

EBITDA; the pre-analysis has shown that medium levels of interest coverage from EBITDA normally occur when ROA and ROE have higher values than interest rates, which implies lower financial friction, so we can conclude that, in the context of lower financial friction, lower income tax share in borrowed funds is favourable.

The conclusions referring to limited liablity companies are as follows: corporate profit tax affects the assets of these companies, but also personal funding sources (capital and reserves, share capital) and borrowed funds (non-current liabilities); lower levels of profit tax are bound with medium and higher values of assets and capital and reserves; income tax has an impact on tangible and current assets, depreciation, revenues and borrowed funds (non-current liabilities); higher levels of income tax are bound with lower levels of tangible and current assets, lower revenues and lower levels of non-current liabilities; if profit tax exceeds non-current liabilities, coverage indicator of current liabilities is good, but this implies low value of return on invested capital; analyses indicate favourable value level of non-current liabilities, higher than profit tax value; this also implies higher levels of ROA, and higher levels of ROA are mostly bound with higher levels of ROE, and are higher than interest rates, which implies lower financial friction; income tax share in personal funds (capital and reserves) is bound with financial coefficient, rate of return on interest-bearing liabilities and investment coefficient; lower levels of income tax from capital and reserves are bound with medium range financial coefficient, low level of investment coefficient and higher rate of return on interest-bearing liabilities. This implies that the amount of personal funds should be lower than the amount of income tax.

ACKNOWLEDGMENT

This work was supported by the Libertas International University project "Interaction of financial friction of the hotel industry and total financial friction in the Republic of Croatia".

REFERENCES

[1] Hall, R.E., 2013. Financial Frictions. International Journal of Central Banking. Vol. 9(2), pp 155-163.

[2] Flannery, M.J., Rangan, K.P.,. Partial adjustment toward target capital structures. Journal of Financial Economics 79: 469-506.

[3] Cloyne, J., Ferreira, C., Froemel, M., Surico, P., 2018. Investment, financial frictions and the dynamic effects of monetary policy. Working Paper.

[4] Brlečić Valčić, S, Samodol, A., (2018). Konceptualni model za analizu financijskog razvitka u kontekstu prijedloga učinkovitih mjera monetarne i fiskalne politike. Ekonomska misao i praksa : časopis Sveučilista u Dubrovniku, Vol. 27(2), pp 377-398.

[5] Arellano, C., Y. Bai, and J. Zhang (2012). Firm Dynamics and Financial Development. Journal of Monetary Economics, Vol. 59, pp 533-49.

[6] Auerbach AJ (2002) Taxation and corporate financial policy. In: Auerbach AJ, FeldsteinM(eds) Handbook of public economics, vol 3. Elsevier, Amsterdam, pp 1251–1292.

[7] Haltiwanger, J., Jarmin, R.S., Miranda, J., (2013). Who creates jobs? Small versus large versus young. Rev. Econ. Stat. 95 (2), pp 347-361.

[8] Tzioumis, K., Klapper, L. F., 2012.Taxation and Capital Structure: Evidence from a Transition Economy FinanzArchiv: Public Finance Analysis, Vol. 68(2).

[9] Erosa, A., Gonzales, B., (2019). Taxation and the life cycle of firms. Journal of Monetary Economics. Journal of Monetary Economics, Vol. 105, pp 114-130.

[10] Jong, A., Kabir, R., Nguyen, T., T. (2008). Capital structure around the world: The roles of firm- and country-specific determinants. Journal of Banking & Finance Vol. 32(9).

[11] Asker, J., A. Collard-Wexler, and J. De Loecker (2014). Dynamic Inputs and Resource (Mis)Allocation," Journal of Political Economy, 122(5), pp 1013-63.

[12] Nivorozhkin, E., 2005. Financing choices of firms in EU accession countries. Emerging Markets Review 6: 138-169.

[13] Brav, O., 2009. Access to Capital, Capital Structure, and the Funding of the Firm. The Journal of Finance Vol. 64, No. 1, pp. 263-308.

[14] Acharya, V., S. A. Davydenko, and I. A. Strebulaev. 2012. “Cash Holdings and Credit Risk.” Review of Financial Studies 25: 3572–3609.

[15] Kling, G., 2016. A theory of operational cash holding, endogenous financial constraints, and credit rationing. The European Journal of Finance. Vol. 24(1), pp 59-75.

[16] Lee, H. and Park, K. (2018), Advances in the corporate finance literature: a survey of recent studies on Korea, Managerial Finance, Vol. 44 No. 1, pp. 5-25.

[17] Claessens, S., Kose, M.A., Terrones, M.E.(2012). How do business and financial cycles interact? J Int Econ 87, 178–190.

[18] Borio, C. (2014): The financial cycle and macroeconomics: what have we learnt? J Bank Financ 45, pp 182–198.

[19] Borio, C., Disyatat, P., Juselius, M. (2016): Rethinking potential output: embedding information about the financial cycle. Oxf Econ Pap 69, 655–677.

[20] Al-Zoubi H. A., O’Sullivan, J. A., Alwathnani, A. M., 2018. Business cycles, financial cycles and capital structure. Ann Finance, 14, pp 105–123.

[21] Moll, B., (2014). Productivity Losses from Financial Frictions: Can Self-Financing Undo Capital Misallocation?. American Economic Review. Vol. 104(10), pp 3186-221.

[22] Čičak, J., Vašiček, D. (2019), Determining the Level of Accounting Conservatism through the Fuzzy Logic System, Business Systems Research, Vol. 10 (1), pp 88-101.

Synergy of innovation procedures and communication skills as a success predictor in IT supported management

Jasmina Dvorski*, Andrija Bernik**, Danijel Radošević***
* Faculty of Teacher Education, University of Zagreb, Croatia
** University North, University Centre Varaždin, Croatia
***Faculty of Organization and Informatics, Varaždin, Croatia
jasmina.dvorski@skole.hr, andrija.bernik@unin.hr, danijel.radosevic@foi.hr

Abstract - **This article is based on communication and its influence on the organization. Through our literature review we found that innovative communication is the strategic ability to plan, execute, coordinate and even evaluate stakeholders of the communication process. Communication also can be used as a mean to spread and redefine comparative advantages in order to achieve a successful business organization. The aim of this research was to determine the importance of communication and find which forms in which ways and types of communication are used by their CEO. Through the Google Forms online questionnaire, 202 respondents were interviewed to confirm the hypothesis: *"only specific forms of communication stimulate innovation and create synergy of innovation processes and communication".* The research was ongoing for 14 days and it was conducted in Croatian business organizations. The results have shown that the success of business organizations can relate to the level of innovative procedures and communication skills. The hypothesis is confirmed and also described in depth in doctoral thesis by one of the authors.**

Key words – communication, idea, innovation, innovative communication, know – how.

I. INTRODUCTION

Creativity and innovation are one of the key components of the vision and mission of business organizations. They are mutually interwoven and create synergy: a) creativity as the creation of new ideas, independent of their usability, b) innovation as a functional use of creative ideas to create new, useful products and services. Creativity is the foundation for innovation, and innovation is the visibility of the applied creative idea. Numerous scholars, with emphasis on the initiator of innovation theory, Peter F. Drucker [1], in his work *Innovation and Entrepreneurship,* confirmed that innovation is closely related to creativity and freedom, Similarly, Velimir Srića's [2] approach to the innovativeness and creativity in modern management is described in the book "*Creativity and Innovation Management*", in which he analyses creativity and innovation, integrating the global economic, communicational approach. An interesting is the project of the University of Zagreb, financed by the funds of the Adris Foundation - "*Innovation, Communication and Management in Science: Development of Generic and Transfer Skills of Young Researchers*", within which generic and transfer skills development workshops have been launched with the aim of training the PhD candidates for the effective inclusion in business and scientific activities with a emphases on permanent employability. The project was completed within the international conference International Transferable Skills in Doctoral Education. The Republic of Croatia puts innovation in the forefront as a factor of development, business productivity and sustainable/permanent employability. If innovations and communication in the business organization have a systematic approach, supported by science and technology, proactive innovation policies, assertiveness, with leading IT supported management by the principles of innovative communication, proactive climate, stimulating culture and leadership in the market will not be absent. The awareness of innovation culture is questionable and it is again necessary to question the new paradigms of the field of communication as well as job opportunities and worldwide IT industrialization [3]. The task for the tendencies of the 21st century ("opportunity" entrepreneurship) is investing in to the educational policies and encouraging the business environment that brings in the innovation climate and relevant knowledge and skills for the materialization of productive market contribution [4] [5]. Business organizations that recognize knowledge and innovative solutions significantly differ from business organizations that do not recognize such business opportunities as well as the necessity of an entrepreneurship. The guiding principles of the European Union are reflected in the HORIZON 2020 project through which innovative projects are funded, regardless of whether they are technological or social innovations. *"Social innovations*" have contributed to the creation of new socio-economic entities and bring up the importance in the form of changes and advancements in business organizations and society [6]. By joining the European Union, the Republic of Croatia has adopted the Europe Growth Strategy (Europe 2020) with five areas of growth and innovation (the remaining four are employment, climate / energy, education, social inclusion). The *National Strategy for Innovation of the*

Republic of Croatia 2013-2020 is also implemented in Croatia (since May 2012) with long-term orientation on the development and systematic promotion of innovation as the core value of the economy's performance, but also of society. It aims are to build an effective innovation system that will direct the Croatian economy to society-of-knowledge based activities. Through the activities of the Strategy, the aim is to exploit the potential of the Republic of Croatia at the territorial position, resources and tradition in industrial production, as well as its ability to innovate and creative as the basic factors for the economy's launch. In the year 2020., it is expected that it will be improved: innovation, legislative and a fiscal framework, established communication and cooperation models, scientific research and business sector with aims for developing new products, services, business processes and technology, as well as implementing path of how to apply scientific and research work in the economy and society.

II. SYNERGY OF THE INNOVATION PROCEDURES AND COMMUNICATION SKILLS

Business communication is used within organizations to promote products, services or organizations. It is known that it is a vital part of business, organization and IT support management as they constantly exchange messages and information important for business organization. Creative and innovative business communication significantly helps business organizations to achieve market success and image. Innovative communication between employees, towards the open market, leads to more successful work, and to achieving the desired reputation and earnings, which is the goal of every business organization. Brand managers of many business organizations want to improve business communication and thus improve quality management. It is not enough, for example, that employees only know how to write a business letter or actively use administrative knowledge. For good communication, employees need to be motivated and acquainted with the basics of effective/innovative communication, which helps to expand the vision and mission of the business organization. Therefore, if changes are to be made to increase the performance of a business organization's communication productivity, then business communication needs to improve. An example, the following business segments: lifelong learning, encouraging self-realization of the employees as complete person etc. It is quite certain that it is important to create a stimulating, proactive atmosphere in which it is possible to work in accordance with the values that are self-imposed, with the support of management leadership. It is also important to encourage two-way communication and knowledge sharing as a focus on achieving results that are of major importance for the entire organization. In successful organizations, employees are aware that communication can expand their knowledge and their personal and professional perspectives and create new value of knowledge. For creative innovative and communicative to open, stimulating atmospheres in a business organization, it is necessary to: 1. *Define the problem, see where the business organization is and what it wants to achieve in the open market; 2. To appreciate and critically think about possible suggestions of employees or to involve a larger number of employees, which will bring a wider range of possible creative solutions.* One of the preconditions for the synergy of the innovation processes and communication skills is the creative leadership of a business organization, ready for innovative strategies, plans, coordination, monitoring and evaluation of employees who will through the innovative creative communication processes achieve innovative communication (Pfeffermann et al. [7]).

III. METHODOLOGY

For the purpose of the Dvorski doctoral thesis [8] a research was conducted using an adjusted (for business organizations in Croatia) and translated questionnaire of the author: Zerfass and Ernst [3] [9]: *Forschungsprojekt "Kommunikation als Erfolgsfaktor im Innovationsmanagement"*. Survey was done using Google Forms and it was sent via email. The target groups were CEOs of the communication organizational units (N = 202). The questionnaire measures communicational determinants of innovative communication management. In a statistic analysis qualitative and quantitative data analysis methods were used. Socio-demographic characteristics (independent variables) were examined: *the number of respondents, gender, age, position in the business organization*. Dependent variables are: *information, opinions and attitudes of respondents about the processes and results of these processes in the field of innovative communication.* The questionnaire, translated from the German language / section related to innovative communication /, was sent to the respondents via e-mail. They were given 14-day fulfillment time. The questionnaire was designed with closed-ended questions (Lickert, yes / no) and analyzed using the SPSS program using the following methods: Pearson correlation coefficient, t-value, HI-square test. The aim of the this research was to determine the importance of communication and find which forms in which ways and types of communication are used by their CEO that their organizations become recognized by innovativeness on the market. It has been set a *Hypothesis:* Only specific forms and types of communication stimulate the innovativeness on behalf of communication IT supported management in a business organization.

IV. FINDINGS

Analyzing demographic data of age and gender of respondents (N = 202), it's obviously that 65.3% of respondents are male. The largest number of respondents is in the age group 41 to 50 y.o., i.e., almost half of the respondents are in that age group, which points to the greater part of the CEOs responsible for innovative communication work. The employees have a rich business experience and business excellence, so younger employees can learn from their business experience in the field of innovative communication.

TABLE 1. THE MOST IMPORTANT INNOVATIVE PROCEDURES FOR A BUSINESS ORGANIZATION (N = 202)

	Total			1,00	2,00	3,00	4,00	5,00
	N	M	SD	%	%	%	%	%
Originality and recognition of our products / services on the market	202	3,44	1,39	15,8%	8,9%	17,8%	30,7%	26,7%
Teamwork	202	3,60	1,39	13,4%	9,4%	15,3%	27,7%	34,2%
Possibility of employee interaction (interpersonal relations)	202	3,62	1,31	9,9%	11,9%	16,3%	30,2%	31,7%
Flexible Leadership	202	3,53	1,38	12,4%	14,9%	10,9%	31,2%	30,7%
Mission and vision	202	3,54	1,40	13,9%	9,9%	18,3%	23,8%	34,2%
Reputation / image of a business organization	202	3,54	1,47	16,3%	10,4%	12,4%	24,8%	36,1%
Two-way communication processes	202	3,53	1,38	12,9%	12,9%	12,9%	30,7%	30,7%
Processes that "Bring" Innovation	202	3,48	1,34	10,9%	14,9%	18,3%	27,2%	28,7%
Awarding	202	3,11	1,31	13,9%	20,3%	25,2%	22,3%	18,3%
Availability of information	202	3,33	1,19	7,9%	16,3%	29,2%	27,7%	18,8%
Risk capability	202	3,00	1,14	9,9%	22,3%	38,6%	16,8%	12,4%

Also, 35.6% examined CEOs think that *innovation of any kind is of great importance in their workplace*, suggesting that all business organizations have to put innovation as their main goal and recognition in their basic postulates (vision, mission).

The strategic importance of innovation has been tested by association terms of providing key words to respondents about innovative activities of their business organization (proactivity, change, market competition, imagination, innovation, creativity): What associates you in the innovative actions in your business organization? “Creativity”, was the most important term chosen by 71.78% of the respondents, then followed by term "proactivity" 54.46% (Figure 1) which means that creativity and proactivness directly contribute to innovation processes and the advancement of these processes through communication.

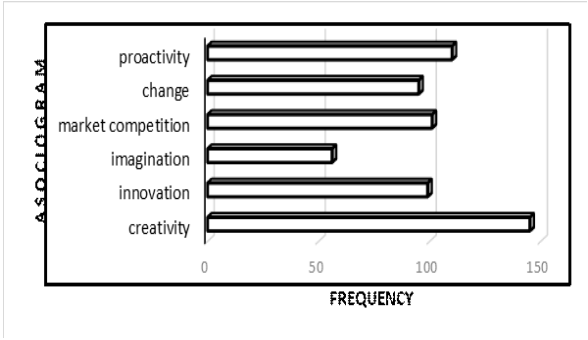


Figure 1. Asociogram innovation and creativity (N = 202)

Table 1 shows that "team work" and "mission and vision" (34.2%) and "reputation of a business organization" (36.1%) were selected as the most important forms of innovation procedures for a business organization. Other offered statements were: originality and recognition of our products on the market (26.7%), employee interaction possibilities (interpersonal relations) 31.7%, flexible leadership (30.7%), two-way communication (30.7%), innovation-generating processes (28.7%), awarding (18.3%), availability of information (18.8%), risk capability (12.4%).

Table 2 shows the attitude of the respondents towa*rds the goals of innovative communication. Cost reduction (66, 83%) and customer satisfaction* (45.54) were statistically significantly selected, which does not surprise, because respondents were from the five major Croatian business organizations (telecommunications, energetic).

TABLE 2. GOALS THAT BUSINESS ORGANIZATION WANTS TO ACHIEVE THROUGH SYNERGY OF THE INNOVATIVE PROCESS AND COMMUNICATION WITH INTERNAL AND EXTERNAL PARTNERS (N = 202)

Goals	**f**	**%**
expense reduction	135	**66,83%**
career guidance (specialization)	32	15,84%
speed and strategic flexibility	66	32,67%
traffic growth	75	37,13%
risk reduction	46	22,77%
higher quality	83	41,09%
faster way to market	37	18,32%
customer satisfaction	92	**45,54%**
reduction of investment costs	36	17,82%
employee satisfaction	17	8,42%

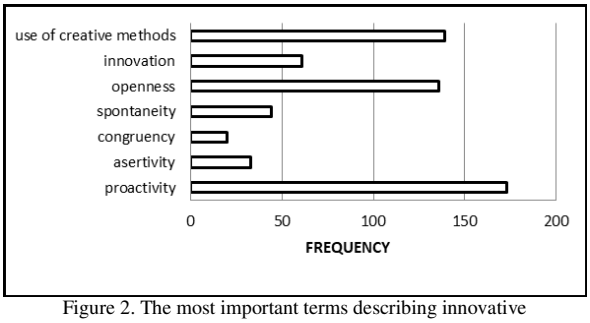


Figure 2. The most important terms describing innovative communication procedures (N = 202)

In order to get a closer look of the thinking direction of respondents (N = 202), they were offered with terms which in their opinion most described innovative communications (Figure 2). The majority of respondents (f = 173, 85.64%) opted for proactivness, and at least for congruency (9.90%).

It is good to point out that the notion of using innovative approach give a high degree of consistency (f = 139, 68, 81%) selected as far as possible by the definition of innovative communication. As tested business organizations are trying to invest in strategic activities where innovative communication is an essential factor, a proactive approach (opposite to reactive) is fully consistent with respondents' responses. Proactivity carries the dimension of responsibility for taking risks. Respondents (N = 202) also estimate how "...innovative communication in the future will gain importance" (85.6%, M = 1.14, SD = 352). It also describes readiness of the business organizations involved in the survey to think about the future in a positive and optimistic way. Returning to the basic human ability postulates - thinking of the ability to communicate, respondents believe that innovative communication, as one of the core human activities, through the encouragement of creativity and expressiveness in the working place, with the goal of realizing themselves as a complete person, will greatly contribute to create an image of the successful business organization.

V. CONCLUSION

This article gives a brief insight into innovative communication skills and the impact that it has on their organizations. Data gathering, on 202 participants in 14-day period, was made on the basis of Zerfass and Ernst survey, which was adopted for the purpose of our experiment. Data analysis of dependent variables was made with SPSS in order to determine the importance of communication and find which forms in which ways and types of communication are used by their CEO.

The results can be summarized as follows. Only specific forms and types of communication stimulate the innovativeness on behalf of communication IT supported management in a business organization - i. e. innovative communication (professional, proactive, creativity, innovation, imagination, market competition, development, change) was the initial hypothesis statement which is confirmed. Creativity and proactivity is of great importance to all respondents in order to create inspirational climate and thus create an atmosphere for innovative communication. The innovation procedures and communication skills have to be in synergy in order to design innovative communication at work. Team work, clear mission and vision of the business organization, reputation / image of the business organization, are closely related in order to achieve the main goal: customer satisfaction, cost reduction and market recognition.

The visible effects of the synergy of the innovation procedures and communication skills are reflected in the new form of business organization, efficient and indivisible communication within and outside the business organization, raising reputation and public trust, and introducing new ideas and knowledge. The synergy of innovation and communication is a unique source for business organization efficiency. Innovation can be learned and nurtured as knowledge, skill, encouragement of critical thinking, finding ways and strategies for innovative communication, creating an inexhaustible source of ideas and creating a positive working environment.

The perception of innovative business organization which were involved in this research, continuously have to improve clearly-defined and market-oriented policies which are supported by all interested stakeholders in society, so they are able to have an image of a successful business organization at the global market. Symbiosis between the innovation procedures and communication skills in a business organization should reflect the organization's strategy and recognition at the market, and in that sense, it is necessary to provide certain resources (communication education), human potential, define strategic knowledge which is essential for success on+ the global market.

REFERENCES

[1] P.F. Drucker, "Inovacije i poduzetništvo: praksa i načela", Zagreb, Globus, 1992.

[2] V. Srića, "Creativity and Innovation Management: a Storytelling approach", Kindle edition; on line book, 2015.

[3] A. Zerfass, and S., Huck, "Innovation, Communication, and Leadership: New Developments in Strategic Communication", U: International journal of strategic communication 1 (2), 107-122. Lawrence Erlbaum Associates, Inc., 2007.

[4] I. Peña-López, "Innovating Education and Educating for Innovation", The Power of Digital Technologies and Skills, 2016.

[5] T. Fotiadis, "Strategic Marketing for High Technology Products: An Integrated Approach", Routledge, 2018.

[6] J. Howaldt, C. Kaletka, A. Schröder, and M. Zirngiebl, "Atlas of social innovation–New practices for a better future", TU Dortmund University, Dortmund, 2018.

[7] N. Pfeffermann, T. Minshall, and L. Mortara, "Strategy and Communication for Innovation", Heidleberg, New York, Dordrecht, London, Springer, 2013.

[8] J. Dvorski, "Odrednice inovativne komunikacije u poslovnom okruženju", Zagreb, Faculty of Philosophy, University of Zagreb (doctoral dissertation), 2017.

[9] A. Zerfass, and N. Ernst, "Kommunikation als Erfolgsfaktor im Innovationsmanagement. Ergebnisse einer Studie in deutschen Zukunftstechnologie-Branchen". Leipzig: Universiteat Leipzig, 2008.

Cloud-based Services Approach as Accelerator in Empowering Digital Transformation

Hyrmet Mydyti*, Jaumin Ajdari *, Xhemal Zenuni*
* SEE University/Faculty of Contemporary Sciences and Technologies, Tetovo, North Macedonia
hm28315@seeu.edu.mk; j.ajdari@seeu.edu.mk; xh.zenuni@seeu.edu.mk;

***Abstract* - Digital Transformation is the process of using digital technologies that impacts and drives significant business and organization improvements and customers' experience. Cloud Computing is one of key digital enablers of Digital Transformation of the products and processes. The main benefits of cloud services in empowering Digital Transformation are impact on cost, work efficiency, elasticity, agility, scalability and security. These benefits encourage and drive businesses and organizations in adapting their strategies to a new digital era.**

This research provides insights, for organizations and businesses, in terms of how cloud-based services approach empowers Digital Transformation. The approach that we are going to elaborate is based on cloud-based services, increasing capacity and enhancing functionality on demand services, with the aim of giving businesses and organizations an understanding of how to transform digitally. Digital Transformation of the organizations and businesses using the cloud-based computing services approach is the aim of this paper.

Keywords - Cloud services; Digital Transformation; Organizations; Businesses, Digitalization;

I. INTRODUCTION

Nowadays, it has become highly important to transform digitally business/organization (in the further text the entity) because it ultimately affects different aspects of the entity as the chances increase to adopt emerging technologies such as social media, mobile technologies, internet of things, cloud technologies, big data analytics, etc. [17]

Managers have goals and interests to create opportunities for their entities to adapt their strategy to the digital reality, by integrating new technologies in their business models, which puts more importance on the topic of processes and operations management [1]. Entities find Cloud Computing attractive and consider it as the key enabler of Digital Transformation, because of provisioning IT efficiency and business agility. Through IT efficiency, the power of modern computers is utilized more efficiently through highly scalable hardware and software resources. And, through business agility, IT can be used as a competitive tool through rapid deployment, parallel batch processing, use of compute-intensive business analytics and mobile interactive applications that respond to user requirements in real time [7]. Entities have a common aim to store large-scale data with low investment cost and enable instant access to this data through the implementation of Cloud Computing [11].

The cloud-based services approach in empowering Digital Transformation has proved to be highly important for this paper.

The analyses and studies of the role of cloud-based services approach in empowering Digital Transformation is to alleviate the process on understanding, encouraging and motivating entities in adapting their strategies to a new digital era. Besides exploring the benefits of cloud services, it includes the impact on cost, work efficiency, elasticity, agility, scalability and security.

The rest of this paper is organized as follows: Section 2 presents a theoretical framework and basic terms related to Cloud Computing. Section 3 provides studies on important aspects of Cloud-based services in empowering Digital Transformation. This section includes an overview of models of the cloud and benefits of cloud computing. Section 4 presents an understanding of Digital Transformation and cloud-based services approach. Finally, section 5 provides the research conclusion of the study and recommendation on future research.

II. THEORETICAL FRAMEWORK

There is a wide literature regarding Cloud Computing and Digital Transformation. In this paper, we have selected some research papers that are focused to basic terms related to Cloud Computing. It brings the scientific data and information from the scholarship together to achieve the main goals of this research.

A. Definitions

The consulted research papers provide definitions on Digital Transformation and Cloud Computing as a concept in general.

Reis J. et al. [1], in their paper define Digital Transformation as "Use of new digital technologies, such as social media, mobile, analytics or embedded devices, in order to enable major business improvements like enhancing customer experience, streamlining operations or creating new business models. As such, the Digital Transformation goes beyond merely digitizing resources and results in value and revenues being created from digital assets". In their research, by using quantitative methodological approaches, two different perspectives of Digital Transformation termed as "Digital Transformation" and "Digitalization" have been ascertained. Furthermore, with the qualitative methodological approach, they demonstrate the strong connection of Digital Transformation to the industry and

the necessity of adoption of business strategies to a new digital reality.

Moreover, Ziyadin S. et al. [2], in their paper introduce the concept of Digital Transformation as "is formed by the merger of personal and corporate IT environments and encapsulate the transformational effect of new digital technologies such as social, mobile, analytical, cloud technologies and the Internet of Things (SMACIT). In a broader sense, Digital Transformation is presented as the integration of digital technologies and business processes in a digital economy".

Neware R. and Khan A. [3], in their paper define Cloud Computing as "a model for enabling convenient, on-demand network access to shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort for service provider interaction".

On the other hand, Islam M. M. et al. [4], in their paper elaborate Cloud Computing as "a modern computing technology where software and hardware infrastructure of an enterprise can be placed over a network to access later in on demand basis via internet instead of having them locally within the enterprise".

B. Characteristics, deployment models and service models of Cloud Computing

Researchers have shown different perspectives related to the Cloud Computing service models, deployment models and characteristics.

Mazumdar A. and Alharasheh [5], in their paper list the key characteristics of Cloud Computing such as on-demand self-service through a secure portal, scalability and elasticity, pay per use, ubiquitous access and location-independent resource pooling. The on-demand self-service characteristic is elaborated as an independent service, without interacting with service provider, in provisioning server, network, and storage capabilities. Scalability and elasticity characteristics are elaborated as increasing or reducing resources elastically to maintain cost efficiencies.

The authors, Rashid A. and Chaturvedi A. [6], in their paper list the deployment models of Cloud Computing as private cloud, public cloud, community cloud and hybrid cloud. Private cloud is considered as a preference of an entity in order to retain high security. Public cloud is considered as a model of general public and industries but retain low security. The community model is regarded as a highly secured model and the hybrid model as a medium secured model.

On the other hand, Neware R. and Khan A. [3], in their paper through their study elaborate three main service models such as IaaS, PaaS and SaaS. IaaS (Infrastructure as Service) is tackled as an on-demand service of virtualization by providing virtual machine, storage infrastructure and network. PaaS (Platform as service) offers application or environment to create application by providing virtual machine, operating system, application and development structure. SaaS (Software as Service) provides on-demand cloud-based foundation for software to the end user, as a complete package.

C. Limitations and Advantages in Implementing Cloud Computing for Digital Transformation

We have analyzed a few research papers regarding this part that will help us in the process of understanding limitations and opportunities of cloud-based services approach in empowering Digital Transformation.

Avram M. [7], in his paper discusses the advantages and barriers of Cloud Computing in an enterprise. In the context of Digital Transformation, the author considers involving technology to improve the economical (or economic) processes. The author recommends that small and mid-size companies should start using cloud computing services. Reduced cost is considered as the most important advantage. Listed barriers to Cloud Computing adoption in the enterprise include security and privacy, connectivity and open access, reliability, interoperability, economic value, changes in the IT organization, and political issues due to global boundaries.

Moreover, authors Islam M.M. et al. [4], in their paper based on the survey methodology, in the perspective of Digital Transformation, explored different limitations of Cloud Computing and presented the ongoing potential solutions towards those problems. The fundamental problems identified and elaborated are data privacy, security, vendor lock-in, interoperability, service availability, absence of unified SLA, performance instability, network bottleneck, lack of scalable storage and reputation fate sharing. It is emphasized that Cloud Computing is widely adopted by the SMEs for its low cost and large enterprises rely on their own infrastructure. Potential solutions listed for the issues are providing complex encryption method of data for more security, usage of unified API, making a unified SLA, a proper scheduling technique, a storage system with unified interface for all cloud providers and hybrid solutions.

Another crucial study, relevant to this paper was by Rahimli A. [8]. The paper is based on quantitative methodology and the data analysis is performed by using SPSS package. Through this paper, the author evaluates the factors that influence entity decisions to adopt Cloud Computing in Malaysia. The factors to be evaluated are the need for Cloud Computing, cost effectiveness, security effectiveness of Cloud Computing and reliability. The analysis, by using SPSS package, and results indicated that security, cost, need, reliability and adoption decision have significant and positive effects. The factors under study represent important aspects of Cloud Computing in empowering the Digital Transformation approach.

Furthermore, authors Petac E. and Petac A-O. [9], in their paper come up with a proposal of an integrated solution that is based on open source software by taking care of all aspects of Private Cloud. This solution is a proposal model for digital business, by underlying Digital Transformation, based on their unique security and right scale needs and concerns.

As the heterogeneity in a joint Cloud Computing environments is a severe challenge, Umar Z. and Emmanual E. [10], in their paper provide a Digital Forensic Framework that can detect intrusion within heterogeneous joint clouds and was developed with the

architectural model and algorithm that can handle the joint clouds heterogeneity and complexity during inter-clouds resources management.

D. Digital Transformation and Cloud Computing for Empowering Digital Business or Organization

Researchers have shown different perspectives on Cloud Computing and Digital Transformation related to the empowering digital business. In the circumstances of successful implementation of Cloud Computing, successful digitalization practices to be elaborated are in the field of agriculture, public administration, manufacturing, urbanism, transport industry and retail.

Ozdogan B. et al. [11], in their paper focus their study in digital agriculture and Cloud Computing is considered playing a vital role in the digital agriculture revolution. The paper has utilized the case method to evaluate the current circumstances of digital agriculture applications in Turkey. Cloud Computing is emphasized as a basic infrastructure that enables intelligent farming implementations such as scalable calculations, software, data access and storage services. And, through Cloud Computing, large-scale data can be stored with low investment cost and instant access to this data becomes possible.

Moreover, Alonso J. et al. [12], in their paper provides a framework to cloudify public services (and the underpinning processes), built upon the analysis of the different stakeholders' (citizens, public sector and the cloud industry) attitude towards Cloud Computing. The presented approach provides the necessary means to enable the implementation of the Transformational Government through the adoption and use of Cloud Computing technologies, supporting the modernization of public administration with ICT and empowering Digital Transformation in what authors have named the Transformational Cloud Government.

Furthermore, Abdelaal M. and Zaki M. [13], in their paper presented the current state of academic research in digital business transformation strategies. It reveals the reasons why companies digitally transform and illustrates companies' internal motivations, as well as their external triggers. The content of Digital Transformation is also framed and distinguished between the business dimensions that are being transformed and the areas that are being impacted by the transformation. Finally, multiple dimensions regarding the content of strategy formulation in the digital era have been identified, and comments on the strategy process have been made. In this context, it has been emphasized that the smart factories and national initiatives such as Industrie 4.0, possibly the largest worldwide trend in manufacturing, require cloud-based systems and networked machines to optimize production.

On the other hand, Serrano W. [14], in his paper presented the concept of Digital as a Service: DaaS, which virtualizes Smart Infrastructure and Cities on four levels: Systems, Transmission, Server and Management. Any complete Digitalization can be implemented independently of its associated physical infrastructure in a Cloud environment. DasS would enable an interoperable Virtual Digital Infrastructure (VDI). Digital as a Service would be its enabler by providing the entire interconnection, integration and virtualization of its Space, Services and Structure (3S).

In their paper, Genzorova T. et al. [15] describe how transport companies can change business model according to digitalization. The case study, as a research methodology, based on agile approach deals with the process of designing the new tool for recording hours spent by the employees for the transport industry, value or supply chain of transport services. For the interest of elaborating what empowers digital business, the authors consider people are the most important asset and hold power to transform a customer's experience. And, human resources are considered playing an active role in helping employees embrace the workplace impacts of technology. The recommendation retains that employees should take part in Digital Transformation actively, not only as an audience.

In addition, in their paper, Hagberg J. et al. [16], develop a framework for digitalization in the retail-consumer interface that consists of four elements: exchanges, actors, offerings, and settings. Drawing on the previous literature, it describes and exemplifies how Digital Transformation impacts each of these elements. On the basis of the previous literature, it has analyzed how exchanges are transformed through changes in communication, transactions and distribution; how actors are transformed through the intermixing of humans and digital technologies, the increased blurring of boundaries, and new actors, roles and relationships; how settings are transformed to include traditional and new settings as well as their intermixing; and how offerings are transformed through changes of products and services, extensions of offerings and new forms of pricing. Thus, the framework contributes to moving the retailing literature beyond discussions of digitalization.

III. Important Aspects of Cloud Computing in Empowering Digital Transformation

Undoubtedly, the cloud-based services approach in empowering digital transformation enables faster access to market. The approach in using cloud-based services on the other hand provides guidance for entities on deploying remotely applications or infrastructure. Moreover, the cloud-based services approach fosters productivity and increases the performance within entities. More importantly, the approach based on cloud-based services provides increased capacity and functionality on demand. Therefore, this paper specifically studies some of the important aspects of cloud-based services in empowering Digital Transformation.

A. An Overview of Models of the Cloud

Five essential characteristics, three service models, and four deployment models compose the cloud model [6].

1) Deployment Models

The deployment models are categorized as follows: Public Cloud, Private Cloud, Hybrid Cloud and Community Cloud.

The public cloud model is the typical cloud-computing model where a service provider offers resources, such as applications and storage, available to the general public, or any organization, over the Internet [5].

The private cloud model offers a high-class computing service that allocates access to a controlled number of people, a defined entity, typically via internal networks or extranets that demand authentication by remote users [5].

The hybrid cloud model is a mixture of the private and public cloud model and even local infrastructure. Though the clouds are mixed up, still each has its individual identity and consequently supporting multiple deployments [6].

The community cloud model involves sharing by two or more entities with shared concerns [5].

Finally, vendors, Cloud Service Providers (CSPs) (e.g., Google, Microsoft, Amazon), provide to their clients the modalities of Cloud Computing resources and services that are dynamically consumed based on customers' request depending on a specific business model [6].

2) Service Models

Based on the computing needs of the clients and different levels of cloud computing, its services can be divided into three key services Software as a Service (SaaS), Platform as a service (PaaS) and Integration as a Service (IaaS) [18].

Infrastructure as a Service (IaaS) is a way of delivering cloud Computing infrastructure as a fully outsourced on-demand service, unlike buying servers, software, datacenter space or network equipment. IaaS aims to focus on ascertaining value to network planners that require to organize specific lower level resources in support of Platform or Software offerings [18].

Platform as a service (PaaS) is a way of delivering Cloud Computing platform and solution stack as a service. It alleviates deployment of applications without the cost and complexity of buying and managing the fundamental hardware and software layers offering all of the facilities necessary to support the entire life process of creating and delivering web applications and services. Platform services provide flexible frameworks with only a few limitations and are able to accommodate some degree of optimization [18].

Software as a Service (SaaS) is a way of a delivering, running and maintaining Cloud Computing application software, operating systems and other resources. Software services are highly standardized and adjusted for efficiency. However, they can only provide minimal customization and extensions [18].

B. Benefits of Cloud Computing

Apart from the facilities that the approach provides with the capabilities, Cloud Computing offers benefits in empowering Digital Transformation as impact on cost and work efficiency, elasticity and agility, scalability, reliability, storage capacity, security for entities. The study of the benefits of cloud-based services approach is to ease the process on understanding and encouraging entities in adapting their strategies to a new digital era.

1) Cost and Work Efficiency

Cost-saving and efficiency is considered a potential advantage of cloud computing in empowering Digital Transformation. Cloud-based services reduce the costs that entities need to spend on infrastructure for their needs [8]. Entities only need to spend what they use and when they may need to increase usage, as it is known as "pay-as-you use' cost structure [5].

Entities benefit from cloud services by improving efficiency and therefore lowering the maintenance cost. Cloud Computing provides large amount of computing power [7], this eliminates the need for operational costs such as power, air conditioning and administration in on-premises infrastructure. Moreover, small entities benefit from compute-intensive business analytics as it was available only to largest corporations [7]. By utilizing the tremendous abilities of Cloud Computing, entities can deploy applications a lot quicker and cheaper compared to the cost of setting up a whole IT infrastructure and services themselves. This facility offers flexibility and eradicates the need for initial capital expenditure [5].

Finally, the cost effectiveness has positive and significant effects on the entities' decision to adopt Cloud Computing [8]. In addition, the accessibility from any place of the data is appealing for entities. This results in creating virtual entities which is very cost-efficient. In this context, the accessibility and unification of professionals around the world become easier. Furthermore, easy manageability saves money by eliminating integration of software since maintenance and updates are performed by Cloud Computing providers.

2) Elasticity and Agility.

Agility is a fundamental requirement for Cloud Computing in empowering Digital Transformation. Computing resources can be rapidly and elastically provisioned and released based on the requirement of the consumer. Consumers consider these resources as if they were unlimited and can be bought in any quantity at any time [6]. Thus, some of the common characteristics a cloud should have, are the elastic capacity and the illusion of unlimited resources. Agility requires that both parts, the consumer and the Cloud Computing provider, cooperate to offer self-management capabilities.

On the other hand, elasticity signifies the provision of services in elastic and adaptable methods, enough to let users to demand the service in near real-time without the need to design new capabilities and resources at times of peak loads. The services required are measured in Service Level Agreements (SLAs) by specifying the capacities and capabilities of the Cloud Computing providers' services in advance. In this context, different levels of service can be provided at different prices - related to uptime (uninterrupted service provision), storage requirements, access to applications, bandwidth (speed of communications) and numbers of users and devices to be supported [5].

Finally, an entity has to continuously re-originate its business models. Cloud provides the required infrastructure, platforms and computing capabilities that support entities stay agile and elastic. In addition, Digital

Transformation implicates traditional work implementations with processes that support agility.

3) Scalability

Besides fostering cost effectiveness, more importantly, Cloud Computing provides high levels of flexibility, scalability, and reusability and high-value speed-to-market solutions [12]. Scalability is a very important characteristic of Cloud Computing in empowering Digital Transformation. The services are very scalable and resilient because the modification and enhancement of the services are very easy and fast. Required bandwidth, processing speed and data storage or number of licenses can be easily increased in a very short time. In addition, in order to meet peak requirement at any time, resource provisioning appears to be unlimited to customers and the consumption can rapidly rise [18]. Moreover, the infrastructure of Cloud Computing is very scalable, and the resources are adaptable. Definitively, new nodes and servers can be added easily by cloud providers to the cloud with smooth modifications to the cloud infrastructure and software [19].

Finally, in the perspective of Digital Transformation of entities, the scalability of business operations or IT scalability is recognized by many entities as one of the major drivers for adoption of Cloud Computing. Moreover, cloud-based services allow entities both to scale up and scale down where entities scale up in less favorable economic circumstances, and scale down their costs, in relation to IT provision - unlike in-house IT provision which involves ongoing costs even when underused [5].

4) Reliability

Reliability is also considered a very important characteristic to the success of Cloud Computing in empowering Digital Transformation. Cloud Computing proves to be more reliable that in-house IT infrastructure. By using multiple redundant sites, reliability is achieved in Cloud Computing. High reliability makes the cloud a perfect solution for disaster recovery and business critical tasks [6].

Cloud Computing providers provide entities with service level agreements (SLA), and it is strictly followed by assuring 99% availability, with the minimum downtime [8],[7]. Generally, each aspect of reliability should be carefully handled when engaging with a CSP. This process is negotiated as part of the SLA and is tested in failover drills [7].

5) Storage Capacity

Different from traditional storage capacity which is not scalable and can be costly due to the fact that it is closely tied to the local computing facilities in use, cloud service is very popular for its computing and storage capacity at lower cost. [20].

"Cloud has larger storage capacity and can handle much more data with no infrastructure investments' [21]. Cloud-computing services, with the intention in enabling digital transformation, offer storage facilities for clients for their operational requirements. As a result, this offering enables product or service customization [5].

6) Security

Security is the biggest concern and the most important characteristic for adoption of Cloud Computing in empowering Digital Transformation. The decisive aspect for potential clients, as entities, is the security and Cloud Computing providers protect the clients' private data, communications and application use. Separation of data for different clients, delivery of efficient replication and recovery mechanisms to robust recovery from failure are sensitive and important aspects that Cloud Computing providers must ensure [5]. In the context of security, the protection of user's files and data is mandatory by cloud service provider [8].

On the other hand, an entity has to take several steps before moving into the cloud. Firstly, the client/entities, ensure(s) if it is technically in accordance with the existing cloud system. Secondly, very importantly, the client/entities, ensures it is not violating any national security law or break the customer data privacy policy. Thirdly, the internal permit executing the workloads on the cloud environment. Finally, entities prepare their business process where consequently on-demand cloud products can be obtained every time they are required [4].

In addition, the implementation of a Private Cloud provides a high level of Cloud Computing security. Private Cloud setup is implemented by providing more control over the company's data and ensuring security, however, still with greater potential risk for data loss due to natural disasters. Different from other types of Cloud Computing models, the entity that is implementing the private cloud is responsible for running and managing the IT resources [9].

Potential solutions exist for cases where the security of data is a concern. The solution to security concern uniquely can be provided through complex encryption method to data. The cloud provider can search and compute the encrypted data but cannot see what the actual data is. This method needs high bandwidth and processor cycle. Using of hardware based security can be more effective in respect to bandwidth and processor speed. However, it causes higher cost [4].

Finally, displacing particularly a database to the cloud provides numerous advantages of enlarged protection from threats such as data breaches, disasters, and system shutdown. Multiple backups can be generated and there is a lower risk of system failure, especially where large amounts of data are implicated.

IV. Understanding Digital Transformation and Cloud-based Services Approach

"The society as a whole is facing a fast and radical change due to the maturation of digital technologies and their ubiquitous penetration in all markets; companies are facing an increasingly tougher competition due to globalization by the increased demand from customers. Because of this, companies are seeking to survive and attain competitive advantages by going digital" [1].

As Digital Transformation of entities is presented as a main issue, nowadays the adoption of strategies of the entities to a new digital reality continues to be an issue of

concern. Apart from that, companies are paying attention to the adaptation of processes and operations management in order to show progress in the adaptation to the new digital reality.

A. Digital Transformation Characteristics

Evolutionary process manifests that the Digital Transformation phenomenon is a continuous process, over a period of time. Digital Transformation is considered an evolutionary process since it evolves with time, and the impacts bring a radical change to the entity. Likewise, digital technologies as key drivers of Digital Transformation are evolving [17].

B. Digital Transformation Drivers

The digital capabilities and digital technologies are two key Digital Transformation drivers. In order to achieve the best Digital Transformation results, relevant skill set and culture – digital capabilities should be incorporated with digital technologies. Digital Technologies, manifests that at the foundation of all Digital Transformation endeavors are digital technologies. Digital technologies build opportunities that entities leverage. These opportunities have the potential to transform business models, operational processes and customer experiences. Thus, the entity takes advantages from the impact of Digital Transformation [17].

C. Digital Transformation Impact

The key Digital Transformation impact is value creation. This value is achieved by both the entity and its customers. Factors included to the value achieved are as follows: operational efficiencies, improved customer experiences; enhanced business models; strategic differentiation, competitive advantage, improved stakeholder relationships, cost savings, etc., [17].

Finally, Digital Transformation, by combining the elements of the characteristics, drivers and impact, can be defined as follows: "DT is an evolutionary process that leverages digital capabilities and technologies to enable business models, operational processes and customer experiences to create value" [17].

Our approach focuses on opportunities that cloud-based services offer instead of challenges. The entities are struggling with the challenges in implementation and practice because of several limitations that cloud-based services impose. We suggest that this innovative technology is issued by small and big entities. In this paper, the main benefits of cloud-based services are delivered and elaborated. The main benefits of cloud services in empowering Digital Transformation are impact on cost and work efficiency, elasticity and agility, scalability, reliability, storage capacity, security for entities. Our approach, by taking in consideration these benefits of cloud-based services, encourages and motivates entities to adapt their strategies to a new digital era. Definitively, Cloud Computing is an opportunity that has the potential to transform business models, operational processes and customer experiences. Correspondingly, we think that Cloud Service Providers must tackle and eliminate the limitations that cloud computing has, especially the security aspects which big entities consider as highly important. In our paper, we provided arguments and analyses of many successful studies on Digital Transformation, benefits and advantages of cloud computing, and on successful implementations of cloud computing in empowering digital entities.

V. RESEARCH CONCLUSION

This section contains the conclusions of the research study. The reflections for completing this research and recommendations for future studies have been included.

In this research paper, a comprehensive research of cloud-based services is done, and different insights, perspectives, approaches and practices are provided, as well as several important aspects of cloud-based services empowering Digital Transformation are provided.

The major goal of the literature review section was to understand the researchers' perspectives on Cloud Computing and Digital Transformation. The selected papers broaden the understanding and experiencing of readers on key perspectives on cloud-based services as an opportunity to the Digital Transformation journey of entities.

The paper introduces the innovative approach by proving that Cloud Computing presents a revolutionary solution and the most modern trend towards Digital Transformation by providing entities with studies on benefits and offerings based on their needs. The fundamental benefits identified are cost and work efficiency, elasticity and agility, scalability, reliability, storage capacity and security. Cloud Computing is being adopted by many entities for its benefits. Researcher consider the implementation of Cloud Computing in small and mid-size entities less complex compared to large size entities. More importantly, cloud-based services release consumers from maintenance of hardware and software and benefit from the advantages.

The main aim of the tackled aspects is to attract entities in using cloud technology as a catalyst of Digital Transformation. Entities, researchers and cloud-based service consumers will be the key beneficiaries of this modern technology.

Finally, this paper helps mainly entities understand the positive impact of cloud-based services approach in empowering Digital Transformation of their entity.

Cloud Computing offers many benefits for large and small size entities. Normally, there are concerns, but in the future, those will not pose limitations; researchers and Cloud Service Providers (CSPs) will give solutions to each concern reflected.

In the future, the aim is to receive feedback from entities, in a long term, regarding the advantages and disadvantages of cloud-based services. In this way, the new research, focusing on new perspectives and the intentions of cloud-based services in pushing Digital Transformation, will be based on empirical studies.

Furthermore, we will address security aspects and ways to reduce risks related to Cloud Computing.

REFERENCES

[1] Reis J., Amorim M., Melão N., Matos P. 2018. Digital Transformation: A Literature Review and Guidelines for Future Research. In: Rocha Á., Adeli H., Reis L.P., Costanzo S. (eds) Trends and Advances in Information Systems and Technologies. WorldCIST'18 2018. Advances in Intelligent Systems and Computing, vol 745. Springer, Cham. DOI:https://doi.org/10.1007/978-3-319-77703-0_41.

[2] Z. Sayabek & S. Suieubayeva & A. Utegenova (2020). Digital Transformation in Business. S. I. Ashmarina et al. (Eds.): ISCDTE 2019, LNNS 84, pp. 408–415, 2020. https://doi.org/10.1007/978-3-030-27015-5_49

[3] R. Neware & A. Khan. (2018). Cloud Computing Digital Forensic challenges. Conference: IEEE 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018), At Coimbatore, India. DOI: 10.1109/ICECA.2018.8474838

[4] M.M. Islam & S. Morshed & P. Goswami (2013). Cloud Computing: A Survey on its limitations and Potential Solutions. International Journal of Computer Science Issues. 10. 159-163.

[5] A. Mazumdar & H. Alharahsheh. (2019). Insights of Trends and Developments in Cloud Computing. South Asian Res J Eng Tech; Vol-1, Iss- 3 (Oct-Nov, 2019): 98-107. DOI:10.36346/SARJET.2019.v01i03.003.

[6] A. Rashid & A. Chaturvedi. (2019). Cloud Computing Characteristics and Services: A Brief Review. International Journal Of Computer Sciences And Engineering. 7. 421-426. DOI:10.26438/ijcse/v7i2.421426.

[7] M-GAvram. (2014). Advantages and Challenges of Adopting Cloud Computing from an Enterprise Perspective. Procedia Technology 12 (2014) 29–534. DOI10.1016/j.protcy.2013.12.525.

[8] A. Rahimli. (2013). Factors Influencing Organization Adoption Decision On Cloud Computing. International Journal of Cloud Computing and Services Science (IJ-CLOSER). 2. DOI:10.11591/closer.v2i2.2111.

[9] P. Eugen & A-O. Petac. (2019). The Challenge of Private Cloud for the Digital Business. Analele Universitatii Ovidius Constanta. XVI. 373-379.

[10] Z. Umar & E. Emmanuel. (2019). A Framework for Digital Forensic in Joint Heterogeneous Cloud Computing Environment. Journal of Future Internet. 3. 1-11. DOI:10.18488/journal.102.2019.31.1.11.

[11] B. Ozdogan & A. Gacar & H. Aktaş. (2017). Digital Agriculture Practices in The Context of Agriculture 4.0. Journal of Economics, Finance and Accounting. 4. 184-191. DOI:10.17261/Pressacademia.2017.448.

[12] J. Alonso & M. Escalante & L. Orue-Echevarria Arrieta. (2016). Transformational Cloud Government (TCG): Transforming Public Administrations with a Cloud of Public Services. Procedia Computer Science. 97. 43-52. DOI:10.1016/j.procs.2016.08.279.

[13] M. Helmy Ismail Abdelaal & M. Khater & M. Zaki. (2018). Digital Business Transformation and Strategy: What Do We Know So Far?. DOI:10.13140/RG.2.2.36492.62086.

[14] W. Serran. (2018). Digital Systems in Smart City and Infrastructure: Digital as a Service. Smart Cities. 1. 134-153. DOI:10.3390/smartcities1010008.

[15] T. Genzorová & T. Corejova & N. Stalmašeková. (2019). How Digital Transformation can influence business model, Case study for transport industry. Transportation Research Procedia. 40. 1053-1058. DOI:10.1016/j.trpro.2019.07.147.

[16] J. Hagberg & M. Sundström & N. Egels-Zandén. (2016). The digitalization of retailing: an exploratory framework. International Journal of Retail & Distribution Management. 44. 694-712. DOI:10.1108/IJRDM-09-2015-0140.

[17] R. Morakanyane & A. Grace & P. O'Reilly, Phillip. (2017). Conceptualizing Digital Transformation in Business Organizations: A Systematic Review of Literature. DOI: 10.18690/978-961-286-043-1.30.

[18] A. Prasanth. (2012). Cloud Computing Services: A Survey. International Journal of Computer Applications. 46. 975-8887.

[19] I. Odun-Ayo & F. Adesola & V. Samuel. (2018). Cloud Computing and Open Source Software - Issues and Developments.

[20] T. Kamalakannan & K.S. Senthil & M. Shanthi & D. Radhakrishnan. (2019). Study on Cloud Storage and its Issues in Cloud Computing.

[21] S. Chinu, K. Sakshi, V. Amandeep, K. Harish. (2018). Chapter 7 - A Hybrid Computational Intelligence Decision Making Model for Multimedia Cloud Based Applications. University Institute of Engineering and Technology, Panjab University, Chandigarh, India. https://doi.org/10.1016/B978-0-12-813314-9.00007-4

Performance Analysis of Aruba wireless local network in Croatian Pension Insurance Institute

A. Skendžić*, B. Kovačić ** and L. Ljubičić***
* College of Information Technology – VsiTe, Zagreb, Croatia
** University of Rijeka/Department of Informatics, Rijeka, Croatia
*** Croatian Pension Insurance Institute, Zagreb, Croatia

E-mail addresses: askendzic@vsite.hr, bkovacic@inf.uniri.hr, luka.ljubicic@mirovinsko.hr

Summary - This paper presents performance analysis of wireless local area network (WLAN) within the building of the Central Office of Croatian Pension Insurance Institute in Zagreb. The network platform on which the measurement was performed is based on implemented Hewlett Packard Enterprise Aruba wireless system which supports IEEE 802.11 ac (WiFi 5) standard. The observed system topologically counts 51 access points (AP) spread over 6 floors of the physical location together with the associated network system infrastructure as part of the CES local network. The measurement methodology which was used was included number of clients by location of deployed access points, access point loads, wireless network usage, most popular applications used on the network, popular network destinations, measurement of wireless network bandwidth and wireless network total data usage. Collected data were taken in period of 6 months.

Keywords - wireless network, wireless LAN performance Analysis, Aruba wireless

I. INTRODUCTION

Well-built network infrastructure is prerequisite for successful business of any modern company. Depends of business requests, the main goal is to find optimal solution. As business request are growing so in proportion in communication infrastructure system is dimensioned according to business requests and business needs. Local Area Networks (LAN) are presenting backbone of company's IT infrastructure [1], they are increasing productivity and support competitiveness for new application technologies while for sophisticated parts of the LAN/WAN network system infrastructure high reliable, speed and security is needed.

On the other side, todays wireless local networks (WLAN) are *de facto* standard for data transmitting [2] in the role of local, wired network infrastructure expansion. Lower implementation costs, availability and performance (e.g. speed) are just some features in wireless local networks benefits and using. However, wireless local networks are demanding specific are planning (coverage area) and using additional network communication equipment [3]. In business environment with higher number of users specific requests are set especially in user/device network authentication and authentication processes, separate guest/employee account access and device provisioning, access control and management. These requests are some of mechanisms which are using in wireless communication networks solutions.

Central Office of Croatian Pension Insurance Institute (forward: Institute) has deployed wireless local network infrastructure (WLAN) which is realized by Aruba wireless network infrastructure [4]. Institute wireless network topology is consists of 51 access points (AP) at 6 building floor levels with 3 main Service Set Identifiers (SSIDs):

1) **HZMO_Djelatnici –** network access for Institute employees.

2) **HZMO_Gosti –** network access for Guests.

3) **HZMO_Print –** which is serving network floor printers.

Analysis and performance measurement were based on Aruba Air Wave system for analysis and reporting which collected information's about clients number, the most commonly used network applications, type of Clients operating system, Access points loads, wireless network bandwidth and data usage. Gathering all information together can help IT team to improving the network

and for planning action on the future. Collected data include period from September 1st, 2019. until February 1st , 2020.

Performance measures are taken at Institutes location. Some of previous academic paper works in Croatia which are based on network performance measuring exploration are paper works [5], [6] and [7]. Performance measuring of wireless local network gives information about network state and it offering answers about improvement needs with the goal of service quality keeping especially speed, availability and security.

II. ARUBA WIRELESS LOCAL NETWORK TOPOLOGY

Wireless network topology at Central Office of Croatian Pension Insurance Institute is based on two Hewlett Packard Enterprise (HPE) Aruba network controllers series 7200 which are in active-active mode. Controllers have redundancy and they are connected on Cisco ASA with one side and on another to the distribution network switches HPE Aruba series 2500 with PoE+ support. However, control and management system Aruba ClearPass is located on dedicated server of Institute Data center. Aruba WLAN topology is shown on Figure 1.

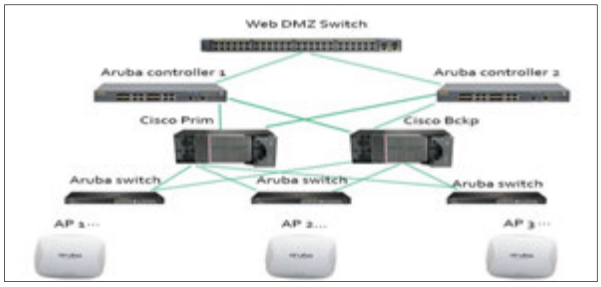


Figure 1. Aruba WLAN topology. Source: authors

According to Gartner 2019 [8], Hewlett Packard Enterprise company (Aruba) – HPE is presenting market lider of advanced network security solution segment for business users in which Aruba – HPE ClearPass network security system management belongs to. All ClearPass network management system activities are based on three steps:

- user identification,
- appending security rules,
- protection.

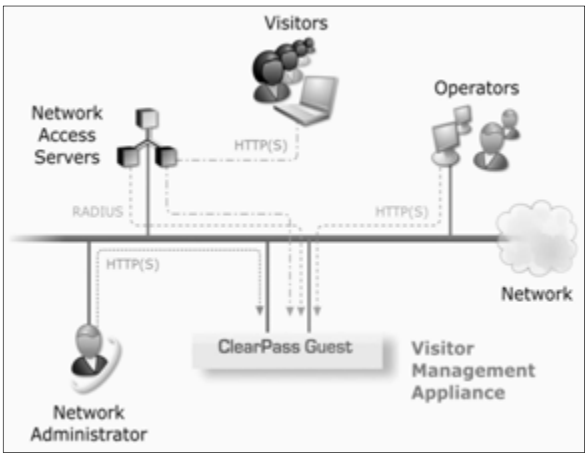


Figure 2. ClearPass Reference Network Diagram. Source: arubanetworks.com

For data analytics collecting Aruba Air Wave Wireless management is used which is set on dedicated server [9]. Total of 51 Access points are deployed including series: 46 devices of Aruba AP 315 and Aruba AP 325 series (indoor) and 5 devices series Aruba AP 277 (outdoor). Indoor and outdoor Access points technically are working at both radio frequency specters 2.4 Ghz and 5 Ghz because of supported IEEE 802.11ac (WiFi 5) network standard [10]. For performance measurement both 2.4 Ghz and 5 Ghz specters were used.

III. ARUBA NETWORK PERFORMANCE ANALYSIS AND METHODOLOGY

Data measurement of wireless local network of Croatian Pension Insurance Institute is done by Aruba Air Wave utility in half year period starting from September 1st, 2019 until February 1st, 2020. Research goal was to get measured data about Access points load, number of wireless network users, popular services (applications) which are used on network, popular network destinations and network bandwidth. Additional measurement data is related on operating system Device Type and information about Time Spent By Device Type.

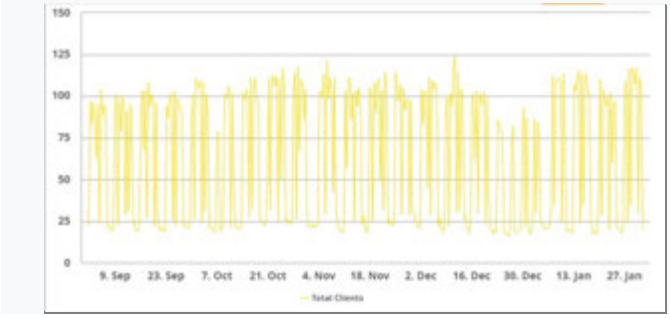


Figure 3. Number of connected Clients during research period. Notice highest peak between **2. Dec** and **16. Dec**. Source: authors

Maximum number of connected clients as shown at Figure 3 reached maximum value at December 11th with **122 connected** clients at 10:55 AM.

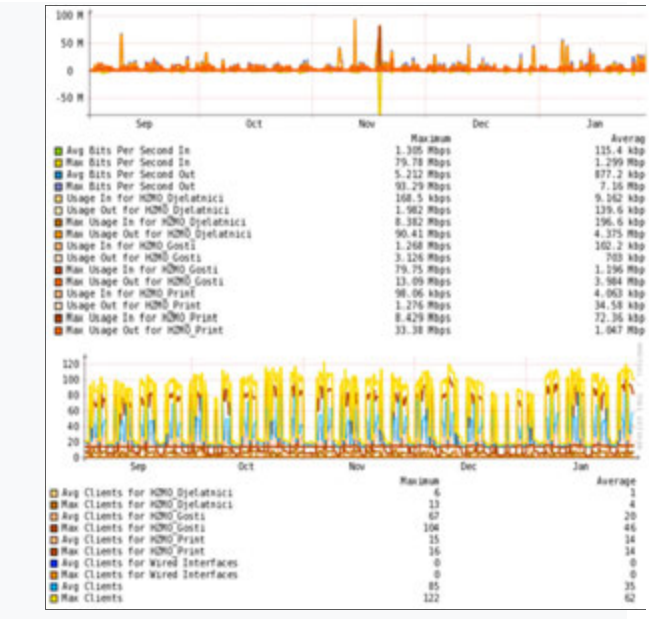


Figure 4. Maximum and average network usage (bandwidth) with maximum and average number of network Clients by SSID. Source: authors

Figure 4 presents minimum and maximum average of wireless local network usage in kbps and Mbps measure units. Network usage refers to bandwidth measuring of each SSID (Total In network traffic and Total Out network traffic) **HZMO_Djelatnici**, **HZMO_Gosti** and **HZMO_Print**. Maximum measured Total Out network traffic (Max Out) was **93.29 Mbps** achieved at the beginning of November 2019. Maximum measured value on SSID HZMO_djelatnici was **90.41 Mbps** (Max usage - Out) while maximum measured average (Average – Out) was **3.98 Mbps** on SSID HZMO_Gosti.

Figure 4 also shows maximum and average number of total connected Clients by each SSID. SSID HZMO_Gosti has reached maximum value of **104** connected Clients with average value of **46**. Maximum average of connected Clients was **85** clients at January 2020.

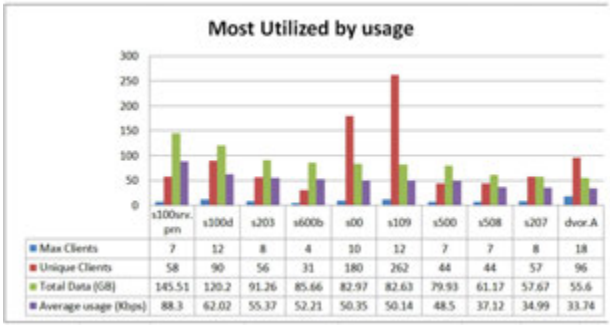


Figure 5. Most utilized Access point by Usage. Source: authors

Figure 5. shows the most loaded Access points according to the total amount of transferred network data (In and Out data). Access point **s100srv.prn** has the highest load of **145.51 GB** data transfer and it is the highest data loaded Access point at all. The main reason of highest data load is that **AP s100srv.prn** is serving printers which are used for daily printing. At the other side, **AP s109** has the biggest number of unique Clients connectivity (Unique Clients**) total 261** Clients because of the fact that nearby is situated meeting room of Assistant director.

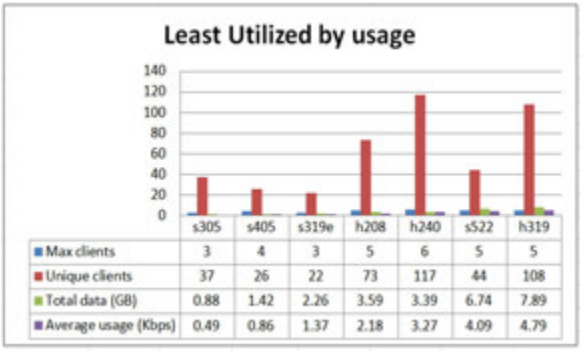


Figure 6. Least Utilized Access point by Usage. Source: authors

Figure 6 shows the least utilized Access points according to total data transfer. The least utilized Access point was AP s305 with total of **0.88 GB** data transferred and average usage of **0.49 Kbps**.



Figure 7. Most Utilized AP by Maximum Concurrent Clients. Source: authors

When we talk about the highest loaded Access points according to maximal number of concurrent users the highest loaded Access point was AP dvor.A with **18** concurrent devices. On the other side, AP s109 recorded the highest number of unique devices MAC addresses, total of **262**, while AP s100d recorded highest network traffic, total of **102.2 GB** as shown at Figure 7.

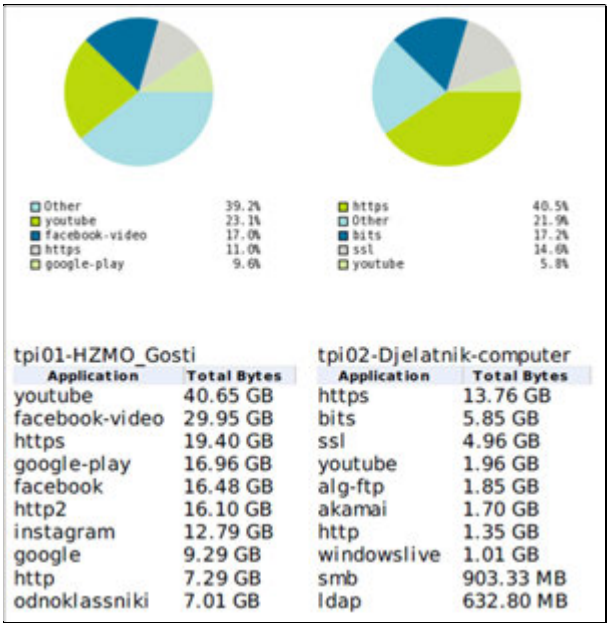


Figure 8. Top 10 Applications by type of users. Source: authors

Figure 8 shows the most frequently used applications (services) by type of users on wireless local network. For user type **tpi01-HZMO_Gosti** total of **39.2%** refers on specific applications (Other) while using YouTube shows share of **23.2%** of total wireless network traffic. Https is on 3rd place with 11%. For user **tpi02-Djelatnik-computer** https has share of 40.5%, category Other is on second place with 21.9% and bits with 17.2%.

TOP 10 applications according to total quantity of network traffic refer to all three SSIDs: HZMO gosti, HZMO djelatnici and HZMO Print. At first position is YouTube with total of **57.39 GB** of network data traffic, while the last one is Google with **10.90 GB** of network data traffic. Social network Facebook is at 5-th place with **20.90 GB**.

According to users category Figure 7 shows that category tipi01 – HZMO Gosti has achieved the highest network traffic to YouTube application **40.65 GB in total**, while category tpi02-Djelatnik-computer has achieved network data traffic of **13.76 GB** in period September 1st, 2019 until February 1st, 2020.

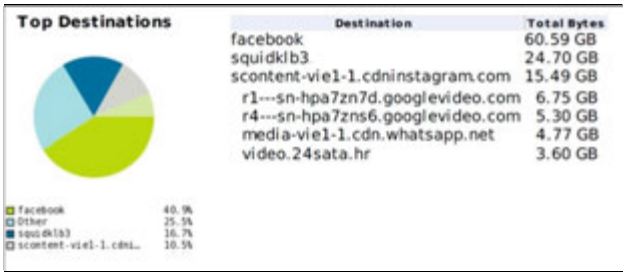


Figure 9. Top destinations (All SSID). Source: authors

The most popular network destinations of this research according to share of total amount of network data are: 1) Facebook with **60.59 GB** of data and share of **40.9%**, 2) squidklb3 with **24.70 GB** of data and share of **16.7%**, 3) scontent-viel-1.cdninstagram.com with **15.49 GB** of data and share of **10.5%**. The most popular network destinations are shown in Figure 9.

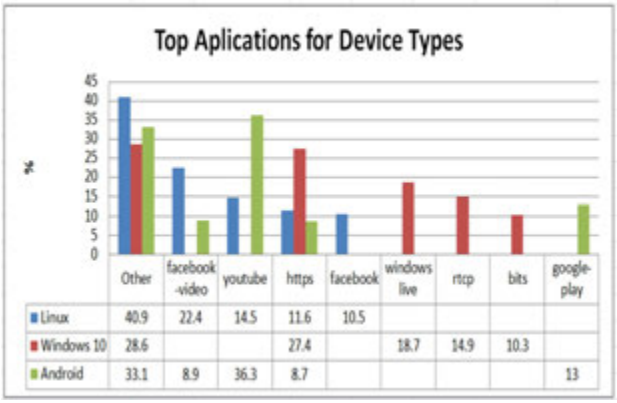


Figure 10. Top Applications for Device Types. Source: authors

Figure 10 shows the top 9 applications according to the type of OS. YouTube application has total share of **36.3%** which is accessed from mobile phone devices (Android OS), while Window 10 OS has most access of **27.4%** to https and other **28.6%**. Linux devices have accessed mostly to other **40.9%** and https **27.4%**. It is necessary to point out that access to some services is restricted by firewall rules and security policies.

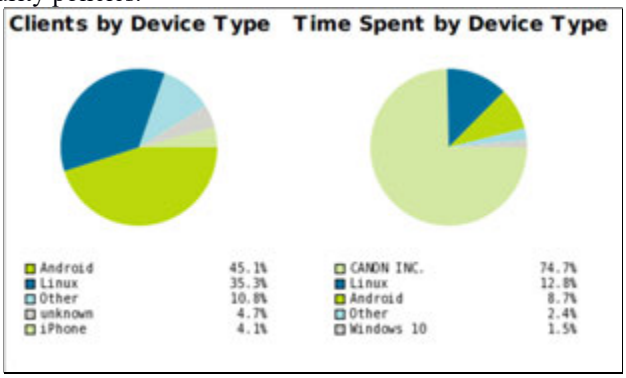


Figure 11. Clients by Device Type and Time Spent by Device Type. Source: authors

Android OS on wireless local network according to type of OS is represented by share of **45.1%**, Linux **35.3%**, others **10.8%** and iPhone with **4.1%**. Devices which have the longest time spent on network are: CANON with share of **74.7%**, Linux **12.8%**, Android **8.7%**, other **2.4%** and Windows 10 **1.5%**. Higher share level of Time Spent by Device Type was done by CANON devices because of connected printers on SSID HZMO_Print as shown at Figure 11.

IV. CONCLUSION

This paper work analyzed WLAN performances at Central Office of Croatian Pension Insurance Institute with the purpose of analyzing wireless network traffic with special indication on the highest loaded critical point of Aruba system. Collected data were done by Aruba Air Wave report utility as integrated tool of Aruba system in period form September 1st, 2019. until February 1st, 2020. The results of research were presented as follows:

- Access point with highest network data traffic load: AP s100srv.prn with total **145.5.1 GB** of data and average usage of **88.3 Kbps**.
- Access point with the highest number connected Clients (Max Clients): AP Dvor.A with total of **18** clients and AP with highest number of unique MAC addresses: AP s109 with **169** unique Clients.
- Access point with the lowest data network traffic: AP s305 with total network traffic of **0.88 GB** and average usage of **0.49 Kbps**.

The most used applications on wireless network (Including all SSID-s) are:

- YouTube with total of **57.38 GB** of network traffic, considering that the highest amount of wireless network traffic refers on SSID HZMO_Gosti, **40.65 GB** total.
- Https with total of **43.16 GB** of wireless network traffic.
- Facebook-video with total of **39.59 GB** of wireless network traffic.

Average number of connected network clients through the period of 6 months was **35**, while maximum number of connected clients was **122**. The most used applications on wireless network (all SSIDs) were:

- YouTube application with **57.38 GB** of network traffic, mostly used by SSID HZMO_Gosti (guest access) with total of **40.65 GB** of network traffic.
- Https with total of **43.16 GB** of network traffic.
- Facebook-video with total of **39.59 GB** of network traffic.

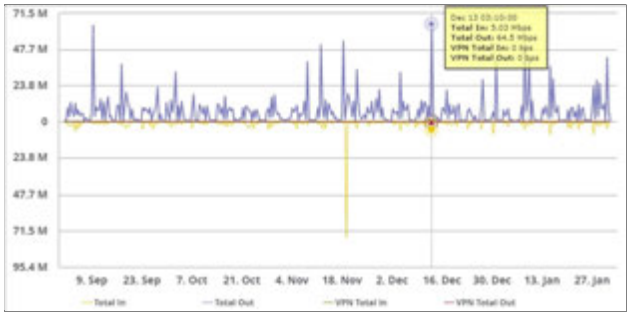


Figure 12. Total Data Usage. Total In and total Out of wireless network traffic. Notice highest peak on 13. Dec similar to 09. Sep.

As shown at Figure 12 total wireless network traffic peaks (In and Out) reached 64.5 Mbps for inner and 5.03 Mbps for outer network traffic at December 13th, 2019.

The result of this study will be used for presenting Institutes top level management for wireless network recommendations to extend Access Point devices at congested and high density area which have high load and serving high number of users. Also, this study could present snapshot of wireless network usage and wireless network state at Institute for period of 6 months activity.

REFERENCES

[1] Integragroup, https://www.integragroup.hr/usluge-i-rjesenja/infrastruktura/mrezna-infrastruktura (14. veljače 2020.)

[2] Skendžić, A.: *Sigurnost infrastrukturnog načina rada bežične mreže standarda IEEE 802.11,* Zbornik Veleučilišta u Rijeci, Vol. 2 (2014), No. 1, pp. 163-176, dostupno na https://hrcak.srce.hr/file/190380 (14. veljače 2020.)

[3] HAKOM izvješće 2010, *Mreže i usluge nove generacije, regulatorni aspekti mreža sljedeće generacije te uporaba digitalne dividende* https://www.hakom.hr/UserDocsImages/2015/komunikacijske_mreze_i_usluge/Projekt%20Pogled%20ubudu%C4%87nost_izvjesce_2010.pdf (16. veljače 2020.)

[4] Arubanetworks, https://www.arubanetworks.com/products/networking/ (05. veljače 2020.)

[5] Zrinec, L. (2016). *Analiza performansi računalne mreže uporabom network performance monitor-a*, Završni rad, FPZ. Dostupno na: https://repozitorij.fpz.unizg.hr/islandora/object/fpz%3A429/datastream/PDF/view (18. veljače 2020.)

[6] Grgurić, G. (2013). *Analiza performansi bežičnih mreža temeljenim na IPv4 i IPv6 protokolima*, Diplomski rad, FPZ.

[7] Ban, I. (2014). *Izvedba i analiza performansi bežične veze na otvorenom*, Završni rad, Sveučilište Sjever, dostupno na: https://repozitorij.unin.hr/islandora/object/unin:299 (18. veljače 2020.)

[8] ArubaNetworks, https://blogs.arubanetworks.com/spectrum/aruba recognized-by-gartner-as-a-leader-in-2019-magic-quadrant-for-the-wired-and-wireless-access-infrastructure/ (18. veljače 2020.)

[9] Arubanetworks, https://www.arubanetworks.com/products/networking/management/airwave/ (20. veljače 2020.)

[10] CISCO, https://www.cisco.com/c/dam/en/us/products/collateral/wireless/aironet-3600-series/white-paper-c11-713103.pdf (21. veljače 2020.)

A Study of Coordination Challenges in Digital Policy Implementation and Evaluation in Finland

Omoregie Charles Osifo*
* University of Vaasa, School of Management, Vaasa, Finland
Charles.osifo@univaasa.fi

Abstract - **This paper focuses on identifying the natures of coordination challenges in digital policy implementation and evaluation in Finland. The research adopts a qualitative approach, because relevant literature and reports are reviewed to ascertain these challenges. The review result shows that coordination challenges in digital policy implementation and evaluation in Finland are related to issues such as autonomy and independence, poor policy planning and approval, financial limitations, inadequate expertise and personnel shortages, orientation towards gains (profit), developmental inequalities among regions; lack of awareness among some stakeholders etc. In addition, the review result shows that recommendations for addressing these challenges are related to issues such as more collaboration, openness, partnership building and networking, improving awareness, enhancing digital and computer education, increasing budgetary allocations, more transparency, encouraging local participation and accountability etc.**

Keywords – Coordination; Challenges; Digital Policy, Implementation; Evaluation; Finland

I. INTRODUCTION

A clear conceptual framework for this research report is presented via the figure below.

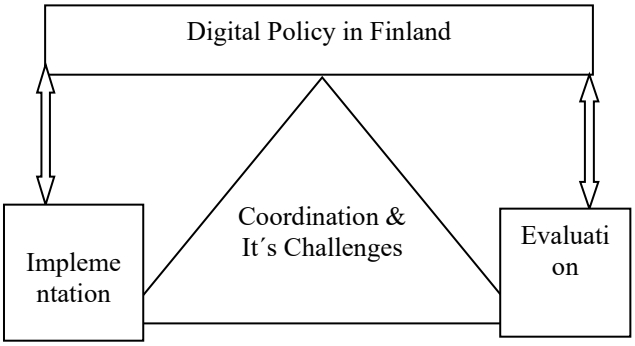


Figure 1. A Conceptual Overview

Finland is a Nordic welfare state with high level of human wellbeing and technological/digital developments. Finland is ranked as a leader in digital technology among countries such as Denmark, Sweden, Netherlands and United States, but issues still exist in relation to perfection rating [1] and [2]. These issues are mainly related to coordination challenges in digital policy implementation and evaluation in Finland.

Digital policy mainly relates to the acts and framework of putting into use and encouraging the opportunities made possible by digitalization. In these acts and framework, regulation of digital and electronic communication or ICT, digital infrastructure (e.g. data centers), network and information security, frequency policy, broadband utilization and access and many others are core issues of concern and deliberation. According to the World Bank, digitalization plays a crucial role in the fight against many social, environment, economic, political challenges in our present world [3].

Digitalization as a concept is tied to so many things, because of its broad nature; the growth in digital technologies such as internet services has changed and transformed our modern world from different perspectives. In addition, to this change is the way phenomena (e.g. poverty, migration, corruption) that exist in our world are transformed/influenced and evaluated/viewed. The popular definition of digitalization according to [4], is the process of integrating digital technologies into everyday life through the digitization of everything that can be digitized. However, a focused and more encompassing meaning is derivable from digitalization as an academic concept. Digitalization as an academic concept, would help us to understand and study the transformative roles (effects), policy strategy, implementation, innovativeness, cost, successes, and challenges.

Differentiating between digitalization, digitization and digital transformation could be challenging. Digitization concerns the changing of things into digital forms that can be processed via computer. Digitalization is the act of using and welcoming things that have been transformed to digital forms. The digital transformation part concerns the effect(s) of digitization [5]. Digitalization, digitization and digital transformation are related, because they are all part of the change idea. According to [1, 3, 6, 7] the digital revolution has transformed the traditional ways of

doing things in many aspects of human existence. Digitalization is strongly reflected in the ways people now communicate with one another; the digitalization of communication systems has affected the mode of communication worldwide. According to [8, p.34] and [9], digital communication possesses some relevant advantages over analogue communication, because of improved noise immunity and processing simplicity.

Coordination as a concept on the other hand plays crucial role in digital policy implementation and evaluation. Coordination terms and theories have been developed in various fields to coordinate the functioning between levels, organs, components, and objects, because of their usefulness to both public, nonprofit and private sectors. If interdependence is missing, then the essence of coordination is not discussed. Coordination is the interrelation of interactions, functions, structures, resources, and materials in the development and management of institutions, products and services. The concept of coordination can be made more tangible, if critical examination is carried out about the design of main coordination instruments and their underlying mechanisms [10, 11, 12].

In many societies around the world, studies have shown that implementation and evaluation of digital policy often encounter challenges and most of such challenges are related to the issue or topic of coordination. "Implementation" concerns mainly with execution of ideas or policies and "evaluation" mainly deals with the assessment of ideas or policies in management studies. According to [13, p.1], "*evaluation is the act of making assessment or judgment in relation to the amount, value, or reliability of something or a phenomenon of interest. Evaluation can act as a link between, expectation and reality, intervention and feedback, need and fulfilment, interest and sacrifice, past and present, future and reform, cost and performance, etc.*"

II. STATEMENT OF RESEARCH AND METHODOLOGY

The inabilities of different tiers and levels (local, regional, central, and international) of governments, experts, institutions and other stakeholders to coordinate digital transformation policy/strategies in different sectors and evaluate the implementation outcomes and progress in many societies and institutions have limited the outcomes of different reform ideas. The key coordination challenges that exist in the digital transformation of many societies are related to autonomy and independence; policy planning and approval; financial inadequacies; expertise/skill limitations; developmental inequalities and lack of awareness [3, 6]. This current research motivation is driven by the occurring changes in public sectors (e.g. e-government) and private sectors (e.g. digital economy), which are being influenced by key emerging global, national, social, economic, environmental, technological, and political events and developments (e.g. innovative ecosystems). "Change is said to be the only constant in life" and individuals, groups, institutions, nations that refuse to change with time and situation often encounter the survival challenge(s) or lose their competitive advantage(s).

The integration of global economy and cross-border sharing of ideas/intelligence in relation to science, education, administration, and security have led renowned public management expert (Christopher Pollit) to identify "place" as one endangered species of public management, because one administrative or economic or political or social or technological idea can now be applied at the local, regional/state, federal/national and global levels simultaneously. Change management in the public, private and non-profit sectors has strong background in contemporary global issues such as technological innovations and their safeties. Change management (CM) is a popular discourse in many disciplines (e.g. public administration and business administration) today. Change management is often referred to as the general strategy adopted in teaching and helping institutions, governments, groups, societies and individuals to pursue and realize change.

According to [12], the effects of digital transformation (e.g. ICT tools) can be explained in relation different variable such as political variables, administrative variables, socio-economic variables, and socio-cultural variables. In addition, the successes and challenges can be tested through different innovation projects such as digital government and digital economy.

The main objective of this research is to identify coordination challenges in digital policy implementation and evaluation; and examine the recommendations in relation to strategies and ways of addressing coordination challenges in digital policies implementation and evaluation in Finland.

This research report tries to answer the following questions:

1. What are the Natures of Coordination Challenges in Digital Policy Implementation and Evaluation in Finland?
2. What are the Recommendations for Addressing these Challenges in Finland?

To answer the above questions, a qualitative review of some selected literature and reports is adopted as methodology. The qualitative method and analysis are suitable for this kind of research, because of the high level of flexibility that it supports and possibility to answer nominal and ordinal research questions.

III. DIGITAL POLICY AND COORDINATION CHALLENGES

Digital policies are key part of innovative societies and systems. Digital policies are meant to be implemented and successes, goals, results and challenges are meant to be evaluated. According to [3], governments

in most developed countries have been urged to develop and implement digital policy strategies, in order to realize digital transformation; this became imperative, because of the key contribution of digital technologies as a strategic enhancer in creating open, participatory and trustworthy public sectors in improving social inclusiveness and governmental accountability, and to connect government and non-government actors and develop new approaches to add to national development and long-term sustainable growth.

Digital policies formulation and implementation (digitalization) are highly reflected in all the models of New Public Management (NPM) reform ideas and other private sector reform. New Public Management (NPM) reform ideas such as the Market Model, Flexible Government Model, Deregulating Government Model, Participatory State Model and E-Governance Model emphasizes the roles of digital policies in societal transformation.

The market model maintains that the assumptions of private sector model are the best way of managing the public sector, where digital tools enhance efficiency. The participatory state model stipulates the need for democratic and collective mechanisms to be key features in governance by understanding what the public wants and finding ways toward fulfilling them; digital tools help in the realization of this objective. The flexible government model emphasizes the need for government and its agencies to make appropriate policy responses to environmental changes rather than just responding in their habitual ways to what are inherently novel changes; digitalization is an essential part of this change/innovation. Deregulating government model stipulates the need for the reduction of formalised rules and regulations and the elimination of rules that would cause red tape to public service functions; digital tools are essential here. E–Governance model of reform is quite a new approach to reforming the public sector. This model maintains that the adoption of information and communication technologies (ICTs) in the management of public service is essential for functional development, increasing growth and reducing unnecessary delay (red tape). [15, p.1546.]

Emphases on implementing and evaluating digital policies are not only relevant to the public sector, but also the private and non-profit sectors from the angles of corporate governance, hybridity and service efficiency for examples. Government to Business (G2B) services has good benefits for companies in areas of cost cutting [16, 15, p.1546-1547].

Digital government like digital economy emphasizes the adoption of digital tools (software, hardware, applications etc.) and framework in providing and executing relevant services, objectives and goods/products in both private sector and public sector; some of these tools include social media, cloud computing and other computer and mobile technologies/applications such as e-voting and e-participation tools. Specifically, digital government through the e-government framework is about the provision of public services to citizens, institutions and other stakeholders; while digital economy is economy premised on digital computing technologies or executing business activities such as marketing through the internet and World Wide Web (www).

Digitalization is part of innovative policy that focuses on promoting openness, inclusiveness and value creation for the citizens, society and world in general. This includes areas such as digital economy and digital product/services, digitalization and public sector, smart city development and management, machine learning as a tool for improving scientific performance, artificial intelligence as part of present and future private and public sectors and society, energy transition and efficiency, and sustainable development.

Focusing on “Smart city” for example, digital policies are a key issue of importance, because smart city is an idea of urban development that entails social, economic, and environmental parts of a city, which help to create public value for citizens. In the functioning of smart city, technology occupies a significant part, because the smooth/good functioning of a smart city depends on technology. The technology needed for high-level development such as smart city is Information and Communication Technology (ICT), which helps to promote and realize interaction and collaboration between citizens, good service delivery, smoot running of smart systems, and development and maintenance of physical and social infrastructures needed in a smart city. The significance of digital policies has been made manifest in issues relating to population, city size, connectivity and infrastructure etc.

In the topic of “Digital Economy” (e.g. sales), digital policies implementation and evaluation are also highly significant in realizing set objectives and goals. These have manifested in issues such as quality, pricing, adulteration, customers’ satisfaction, privacy, piracy, clouding etc.

Since coordination has to do with interrelation of interactions, functions, structures, stakeholders, and resources in the development and management of the public and private sectors; therefore, tendencies of encountering challenges are very high. Coordination challenges in digital policy implementation and evaluation can be linked to political, administrative, socio-economic, socio-cultural, technological, demographic and other variables or factors.

In this research coordination challenges in digital policy implementation and evaluation are studied within the scope of digital government and digital economy as innovation projects in Finland. The results from the

qualitative review of some selected literature and reports on the nature of coordination challenges are presented in the next section. In addition, are recommendations in relation to strategies and ways for addressing these challenges.

IV. Findings from selected Literature and reports on the case of Finland

In this section, findings from the selected literature and reports on the coordination challenges that exist in Finland in digital policy implementation and evaluation are presented under relevant issues. The identified natures of coordination challenges are presented along with the innovation project(s) they have occurred.

A. Identified Coordination Challenges in Digital Policy Implementation Based on Studied Literature and Reports

In table 1 below, the identified coordination challenges in relation to digital policy implementation in Finland and the data sources are presented.

Table 1: Identified Coordination Challenges in Finland (Implementation)

Nature of Coordination Challenges	Project(s) that Challenges have Occurred
Autonomy and Independence	Digital government
Inadequate expertise and personnel shortages	Digital government and Digital economy
Financial limitation	Digital government and Digital economy
Lack of awareness and Issues of participation and trust	Digital government and Digital economy
Rules and regulations (e.g. GDPR)	Digital Government and Digital Economy
Ethical Issues	Digital government
Addiction and monotony	Digital Government and Digital economy
Security (e.g. hacking, virus and malware) and privacy issue	Digital government and Digital Economy
Orientation towards gains (profit) and Competition	Digital economy
Developmental inequalities	Digital government
Illiteracy and Non-Acceptance	Digital government
Big data or data deluge	Digital government
Ambiguity and complexity	Digital government and Digital Economy
Poor policy planning and approval	Digital government
Language and beliefs	Digital economy and Digital government

Sources: [1, 2, 3, 7, 10, 15, 17, 18, 19, 20, 21, 22]

The natures of the identified coordination challenges in digital policy implementation in Finland occur according to the selected literature and reports, because of different factors or variables such as socio-cultural, socio-economic, technological, demographic, structural, administrative, legal etc. The issues of autonomy and independence are key natures of coordination challenges in digital policy implementation in Finland. According to the literature and reports, Finland is a country where different regional governments or administrations enjoy high level of administrative autonomy and independence. These challenges affect negatively to a great extent the "who (e.g. decision-maker), when (e.g. time), where (e.g. exact location), and how (e.g. cost and finance)" of digital policy implementation, which have been experienced in project such as digital government. Inadequate expertise and personnel shortages are other identified natures of coordination challenges. According to the literature and reports the digital and technological sectors in Finland suffers personnel shortages of up to 50 percent and often extremely high skill personnel are hired from abroad to manage some core digital infrastructures and projects. The shortages in budgetary allocations across different sectors in Finland also serve as coordination challenge in digital policy implementation, because many institutions often encounter financial difficulties in collaborating with other institutions on the development of digital infrastructures and platforms. This challenge has been experienced in the digital government and digital economy projects.

Lack of awareness and issues of participation and trust are other natures of identified coordination challenge according to the literature and reports used; this is because many citizens and stakeholders are unaware of details or formulation of digital policies. These challenges affect the participation and trust of citizens and other stakeholders in digital policy implementation. Rule and regulations (e.g. GDPR) are other natures of identified coordination challenges. New rules and regulations such as the General Data Protection Regulation (GDPR) affects free flow of information sharing and interactions between individuals and other

stakeholders such as security agencies in combating crimes. Ethical issues such as misconduct affects to a little extent the interaction between citizens and other stakeholders in implementing digital policies in Finland; especially as it concerns the projects of digital government and digital economy. This is because ethical issues occurrence leaves a huge gap of distrust in the process of coordinating to implement digital policies according to the studied literature and reports. Addiction and monotony to digital tools and platforms such as internet and social media among especially youths are other coordination challenges in digital policy implementation in Finland. The right feedbacks in improving digital policies are often not ascertained due to these challenges. Security (e.g. hacking, virus and malware) and privacy issues are other identified natures of coordination challenges. These challenges affect citizens and other stakeholders´ data and consent, which later affects the willingness to participate in digital policies improvement and implementation.

Orientation towards gains (profit) and competition are other natures of coordination challenges in digital policy implementation in Finland. These challenges have been experienced in projects such as digital government and digital economy according to the studied literature and reports. The private sectors high orientation for making profit often affect their interest in collaborating and interacting with public sector on the best and cheapest way to implement and improve digital policies. Developmental inequalities among different regions in affects the interaction of stakeholders in implementing uniform digital policies across Finland, digital infrastructure in Urban centers are more solidified than the rural areas. Illiteracy and non-acceptance are other natures of coordination challenges. Digital and technological illiteracy among many members of the elderly population in Finland affects their usage of digital tools (e.g. e-voting software), which later affect individual and institutional acceptance of digital policies implementation. Another coordination challenge in digital policy implementation in Finland is the issue of big data or data deluge that emanates from large population or big size of cities and municipalities. The inability to critically study every viewpoint affects negatively the implementation of digital policies according to the studied literature and reports.

Ambiguity and complexity in the design of digital infrastructure and tools also affect coordination challenges in digital policy implementation in Finland, because the usage of tools becomes difficult across the whole Finnish population. These challenges also affect negatively the acceptance of digital policies among individuals and institutions. Poor policy planning and approval are other identified natures of coordination challenges in digital policy implementation in Finland. The non-acceptance of digital policies among many stakeholders is mainly due to poor policy planning in taking their interest into serious consideration during formulation; in addition, is the long delay often experienced before policy approval. Finally, the other coordination challenges in digital policy implementation in Finland concern the issue of language and beliefs. The migrants´ population difficulties in speaking and understanding the Finnish often affect their contribution during policy formulation. Also, among many foreigners in Finland is diversity in religious and cultural beliefs. These differences in comparison to the popular make collaboration often difficult.

B. Identified Coordination Challenges in Digital Policy Evaluation Based on Studied Literature and Reports

In table 2 below, the identified coordination challenges in relation to digital policy evaluation in Finland and the data sources are presented.

Table 2: Identified Coordination Challenges in Finland (Evaluation)

Nature of Coordination Challenges	Project(s) that Challenges have Occurred
Autonomy and Independence	Digital government and Digital economy
Inadequate expertise and personnel shortages	Digital government and Digital economy
Financial limitation	Digital government and Digital economy
Comparative issues and phenomena	Digital government
Design complexity and ambiguity	Digital economy
Performance disparity	Digital economy
Data size and Fatigue and Lag	Digital government and Digital economy
Limited transparency and participation	Digital government and Digital economy

Sources: [1, 2, 3, 7, 10, 15, 17, 18, 19, 20, 21, 22]

Unlike the case of coordination challenges in digital policy implementation in Finland, which has numerous natures of challenges, the case of coordination challenges in digital policy evaluation has lesser natures of challenges. Autonomy and independence are also core natures of coordination challenges in digital policy evaluation. In Finland, regional governments or administrations enjoy high level of administrative autonomy and independence. These challenges affect negatively to a great extent the evaluation framework of

digital policies. These challenges according to the studied literature and reports exist in realizing the full objectives of the digital government project in Finland. Inadequate expertise and personnel shortages are also coordination challenges in digital policies evaluation in Finland according to the studied literature and reports. These challenges affect manpower issues in achieving the objectives of digital government and digital economy. Financial limitation is also a nature of coordination challenges in digital policy evaluation in Finland according to the studied data. This challenge affects negatively the ability to collect relevant data from individuals and other key stakeholders in order to assess and make relevant reforms in digital government and digital economy projects.

Comparative issues and phenomena are other natures of coordination challenges in digital policy evaluation in Finland according to the studied literature and reports. This challenge mainly arises, because of development inequalities among different regions of Finland. Design complexity and ambiguity are other natures of coordination challenges in digital policy evaluation. These challenges arise, because of the inability of most individuals and stakeholders to understand the functioning of digital tools such as the internet, which thereby, make the collection of right data to improve digital policies difficult. Performance disparity is another nature of coordination challenge in digital policy evaluation in Finland. This challenge occurs because digital tools (e.g. internet) and infrastructure function in varied capacity in different households, institutions, municipalities, and cities. Data size, fatigue and lag are other natures of coordination challenges in digital policies evaluation. These challenges occur, because of large population size, city size and other demographic and environmental factors in digital government and digital economy projects. Finally, limited transparency and participation are other natures of coordination challenges in digital policy evaluation. These challenges affect negatively the willingness of individuals and other stakeholders to give the relevant data in improving digital policies in Finland. According to studied literature and reports these challenges have occurred in the digital government and digital economy projects, which affect negatively the realization of their set objectives.

C. Recommendations for Addresssing Coordination Challenges in Digital Policy Implementation Based on Studied Literature and Reports

In table 3 below, the recommendations for addressing coordination challenges in digital policy implementation in Finland and the data sources are presented.

Table 3: Recommendations for Coordination Challenges in Finland (Implementation).

Nature of Coordination Challenges	Recommendations to Addressing the Identified Coordination Challenges
Autonomy and Independence	More collaboration, openness, partnership building, networking and Improving awareness
Inadequate expertise and personnel shortages	Enhancing digital/technological and computer education
Financial limitation	Increasing budgetary allocations
Lack of awareness and Issues of participation and trust	More transparency, Creating grassroots awareness and Digital and computer education
Rules and regulations (e.g. GDPR)	Adoption of legal and ethical frameworks, Updating of digital infrastructure, Computer education, Openness and Accountability
Ethical Issues (e.g. misconduct)	Improving integrity and compliance framework
Addiction and monotony	Increasing monitoring and users 'regulation
Security (e.g. hacking, virus and malware) and privacy issue	Updating digital infrastructures and tools, More collaboration and partnership building and Enhancing digital and computer education
Orientation towards gains (profit) and Competition	Improving integrity and compliance framework
Developmental inequalities	Increasing budgetary allocations to rural areas and poorer municipalities.
Illiteracy and Non-Acceptance	Improving transparency and digital education
Big data or data deluge	Renewing digital infrastructures and tools, Supporting digital and computer education and Training more personnel and expert
Ambiguity and complexity	Improving and simplifying the design of digital infrastructure and tools

Poor policy planning and approval	Encouraging more participation, Adopting recommendations, More transparency, Controlling cost and Managing resources
Language and beliefs	Enhancing integration framework, Encouraging more transparency and participation, Improving roles of representatives, Cross-cultural understanding, Flexibility, Experimentation (equilibrium and large-scale piloting) and Boosting motivation

Sources: [1, 2, 3, 7, 10, 15, 17, 18, 19, 20, 21, 22]

According to the studied literature and reports, recommendations concerning ways and strategies for addressing coordination challenges in digital policy implementation are clearly stated in table 3 above. The recommendations include: More collaboration, openness, partnership building and networking and Improving awareness; Enhancing digital/technological and computer education; Increasing budgetary allocations; More transparency, Creating grassroots awareness; Improving integrity and compliance framework; Renewing digital infrastructures and tools, Supporting digital and computer education and Training more personnel and expert; Improving and simplifying the design of digital infrastructure and tools; Encouraging more participation, Adopting recommendations, More transparency, Controlling cost and Managing resources; and Enhancing integration framework, Encouraging more transparency and participation, Improving roles of representatives, Cross-cultural understanding, Flexibility, Experimentation (equilibrium and large-scale piloting) and Boosting motivation.

D. Recommendations for Addresssing Coordination Challenges in Digital Policy Evaluation Based on Studied Literature and Reports

In table 4 below, the recommendations for addressing coordination challenges in digital policy evaluation in Finland and the data sources are presented.

Table 4: Recommendations for Coordination Challenges in Finland (Evaluation).

Nature of Coordination Challenges	Recommendations to Addressing the Identified Coordination Challenges
Autonomy and Independence	Enhancing networking, Transparency and Partnership building
Inadequate expertise and personnel shortages	Computer and digital training
Financial limitation	Increasing budgetary allocation
Comparative issues and phenomena	Adoption of comparative method, online method, & statistical parameters
Design complexity and ambiguity	Methodological clarity and focus and Replication in design infrastructure
Performance disparity	Enhancing networking and accountability, Adequate planning, Monitoring and Periodic review
Data size and Fatigue and Lag	Employing and training more personnel and digital expert, Adoption of ethical framework (e.g. equity, transparency and participation) and Updating digital infrastructure and tools.
Limited transparency and participation	More interviews and Surveys

Sources: [1, 2, 3, 7, 10, 15, 17, 18, 19, 20, 21, 22]

Also, according to the studied literature and reports, recommendations concerning ways and strategies for addressing coordination challenges in digital policy evaluation are clearly stated in table 4 above. The recommendations include the following: Enhancing networking, Transparency and Partnership building; Computer and digital training; Increasing budgetary allocation; Adoption of comparative method, online method, & statistical parameters; Methodological clarity & focus & Replication in design infrastructure; Enhancing networking and accountability, Adequate planning, Monitoring and Periodic review; Employing and training more personnel and digital expert, Adoption of ethical framework (e.g. equity, transparency & participation) and Updating digital infrastructure and tools; and Interviews and Surveys

V. CONCLUSION

This qualitative research tries to identify coordination challenges in digital policy implementation and

evaluation in Finland. In addition, recommendations concerning the strategies and ways via which these challenges can be addressed are also identified and presented. The study of coordination challenges in this research report is withing the scope of digital government and digital economy innovation projects. The study of coordination challenges in digital policy implementation and evaluation is for the main purpose of realizing the objectives of digital government and digital economy projects. Qualitative review of some selected literature and reports is adopted in answering the two research questions: (1) what are the natures of coordination challenges in digital policy implementation and evaluation in Finland? And (2) what are the recommendations for addressing these challenges in Finland?

It was discovered from the studied literature and reports that coordination challenges in digital policy implementation in Finland include issues such as autonomy and independence; inadequate expertise and personnel shortages; financial limitation; lack of awareness and issues of participation and trust; rules and regulations (e.g. GDPR); ethical issues (e.g. misconduct) etc. Findings from the studied literature and reports on the nature of coordination challenges in digital policy evaluation in Finland include issues such as autonomy and independence; inadequate expertise and personnel shortages; financial limitation; comparative issues and phenomena; design complexity and ambiguity etc.

On the recommendations concerning strategies and ways for addressing these challenges, it was discovered from the studied literature and reports that coordination challenges in digital policies implementation in Finland relate to issues such as more collaboration, openness, partnership building and networking and Improving awareness; Enhancing digital/technological and computer education; Increasing budgetary allocations; More transparency, and Creating grassroots awareness etc. Also, according to the studied literature and reports, recommendations concerning ways and strategies for addressing coordination challenges in digital policy evaluation include issues such as Enhancing networking, Transparency and Partnership building; Computer and digital training; Increasing budgetary allocation; Adoption of comparative method, online method, & statistical parameters; Methodological clarity & focus & Replication in design infrastructure etc.

REFERENCES

[1] European Commission, "eGovernment in Finland" Publications Services," 2015.

[2] Yle, "Index: Finland Third-Most Advanced Digital Economy in Europe," Retrieved February 6, 2020. from https://yle.fi/uutiset/osasto/news/indexfinlandthird-mostadvanced_digitaleconomyineurope/10220114

[3] OECD, "OECD Comparative Study: Digital Government Strategies for Transforming Public Services in the Welfare Areas," 2016.

[4] BusinessDictionary, "Digitalization", Retrieved February 6, 2020. From http://www.businessdictionary.com/definition/digitalization.html

[5] S. Khan, "Leadership in the Digital Age: A Study on the Effects of Digitalisation on Top Management Leadership" Stockholm University, No.54, 2016.

[6] OECD, "Highlights from the OECD Science, Technology and Industry Scoreboard 2017-The Digital Transformation: Finland", 2017.

[7] M. Gavet, "The Digital Revolution is Destroying our Democracies: It Doen´t Have to be That Way," World Economic Forum 2018, 2017.

[8] B. G. Peters. The Future of Governing. Kansas: University Press of Kansas, 2001

[9] M. S. Roden, Analog and Digital Communication Systems. Los Angeles, Ca: Discovery Press, 2003.

[10] O. Viinamäki, "A Theory of Coordination and Its Implication on EU Structural Policy: A Comparative Study of the Challenges for Coordination in Structral Funds in Finland, Ireland and Sweden. Administrative Science No. 132. Vaasa: University of Vaasa. 2004.

[11] G. Boella, and L. van der Torre, "Coordination and Organization: Definitions, Examples and Future Research Directions," in Electronic Notes in Theoretical Computer Science, vol. 150, 2006, pp. 3–20.

[12] E. Beuselinck, K. Verhoest & G. Bouckaert, Reforms of Central Government Coordination in OECD Countries for Cross-National Unifying Processes?; In (Eds.) Kuno Schedler and Isabella Proeller: Cultural Aspects of Public Management Reform, 2007, pp. 77–109. Amsterdam: JAI Press.

[13] O. C. Osifo, "Evaluating Governance and Management in Africa: A Utilitarian Perspective," Working Papers 8, Public Management 5, University of Vaasa, 2015

[14] L. H Gutiérrez and L. F Gamboa, "Determinants of ICT Usage among Low-Income Groups in Colombia, Mexico, and Peru," The Information Society, vol. 26, iss. 5, 2010, pp. 346–363.

[15] O. C. Osifo, "Examining Digital Government and Public Service Provision: The Case of Finland," in Mipro 2018 – 41st IEEE International Convention, Ed. Karolj Skala. Rijeka: Croatia Society for Information and Communication Technology, Electronics and Microelectronics – MIPRO, 2018, pp. 1545-1550

[16] C. L. Johnson, "A Framework for Pricing Government E-service," Electronic Commerce Research and Applications, Vol. 6, Iss. 4, 2007

[17] Ministry of Finance, "Final Project Report for SADe: Development Plans and Action Plan for eServices and Administration between 2009-2012", 2009.

[18] UNCTAD, "The Digital Economy Report 2019. VALUE CREATION AND CAPTURE: IMPLICATIONS FOR DEVELOPING COUNTRIES," United Nations, 2019

[19] A. Anttiroiko, Issues of Information Society in Finland; In (Ed.) Toshio Obi: E-governence, 2007, pp. 19-22. Amsterdam: IOS Press.

[20] J. Julin, E-governence Experience in Finland; In (Ed.) Toshio Obi: E-governence, 2007, pp. 64-66. Amsterdam: IOS Press.

[21] A. Anttiroiko, P. Kuusisto, and J. Seppälä, Democtratizing E-governement in the City of Tampere; In (Eds.) A. Anttiroiko and Antti Kasvio: E-city, Analysing Efforts to Generate local Dynamism on the City of Tampere, 2004, pp. 389-407. Hershey: Idea Group Publishing.

[22] European Commision, "Europe's Digital Progress Report (EDPR) 2017 Country Profile Finland," 2017

Port Community System feasibility analysis – Case study Split

I. Torlak*, E. Tijan**, S. Aksentijević*** and A. Jugović**
* Jadroagent Rijeka, Rijeka, Croatia
torlak.ivan@gmail.com
** University of Rijeka, Faculty of Maritime Studies/Department of maritime logistics and management, Rijeka, Croatia
etijan@pfri.hr, ajugovic@pfri.hr
*** Aksentijević Forensics and Consulting, Ltd., Viškovo, Croatia
sasa.aksentijevic@gmail.com

***Abstract* - In order to increase the competitiveness of Croatian seaports, seaport cluster stakeholders should strive to digitalize and optimize their business processes. A significant part of such digitalization and optimization is constituted by the implementation of a Port Community System. In this paper, the authors will define the theoretical framework of Port Community Systems, and will determine the basic economic indicators and benefits of investing in the implementation of the Port Community System. Potential costs of project development and the costs of maintaining the system within the forecasted period of use will be determined. The potential savings from the digitalization of business processes related to cargo processing will be determined, based on the Port of Split case study. Digitalization of Croatian seaports will create the potential for increasing seaport competitiveness and traffic volumes. Since self-financing of such projects is extremely demanding for individual seaports, Port Community System implementation should be achieved by synergistic efforts of the national and regional authorities, seaports and other involved stakeholders, and by using available EU funding.**

Keywords – Port Community System; feasibility analysis; case study; Port of Split; digitalization

I. INTRODUCTION

Investments in seaports can be divided in two basic groups, traditional investment in port infrastructure, such as construction of warehouse and piers, etc., and investments in improvement of business processes that include heterogeneous maritime ICT systems like, VTMIS (Vessel Traffic Monitoring & Information Systems), ISPS (International Ship and Port Facility Security Code) enhancement systems, IoT (Internet of Things platforms), and WSN (Wireless Sensor Networks).

Seaports are essential for the support of economic activities in the surrounding areas [1]. Each port which aims to be competitive, should accept digitalization [2]. Digitalization and stakeholder connectivity (forwarders, maritime agents etc.) may enable numerous advantages such as: the business facilitation, the reducing of mistakes and impediments. Port community systems (PCS) can be defined as holistic, geographically bounded information hubs in global supply chains that primarily serve the interest of heterogeneous collective of port related companies [3]. In seaports, a communication system such as PCS holds an important role in reducing the operational costs. PCS helps the competent authority to take the lead by providing a logistics solution to private operators, encouraging them to share information that ultimately leads to lowered logistics costs, faster delivery/pickup in the import/export chain, and higher customer satisfaction. Bringing all users together enhances efficiency of the physical flow of freight, drives economic growth, and as a secondary effect, assists in reducing externalities such as pollution, congestion, and land use impacts [4].

Considering that each investment must pass through process of financial and economic evaluation, this paper discusses the economic aspects of PCS implementation in Croatian cargo ports. State Port Authorities in Croatia as non-profit organizations and organizations that manage concessions for all port basins have a major role as project coordinators and, partially, investors in development activities. There must be firm assurances that this kind of public investment presents a positive social benefit, so the government, and in this case the port authority, must find a suitable solution to decrease overall implementation costs. Enabled by the possibilities of newly developed technology, digital initiatives are increasingly more present in port business processes. This type of business innovation is now seen as a game changer. Analyzing the services provided by the ports and maritime sector, the same types of developments are also pursued in the named business sectors [5].

II. FINANCING THE DESIGN, IMPLEMENTATION AND MAINTENANCE OF THE PCS

Each PCS is specific, each country has individual legal regulations and therefore each port community develops its own PCS according to its needs [6]. Furthermore, seaports differ by the number and type of data being exchanged, the actors (businesses subjects) which are exchanging the data and messages, the owners of systems which are exchanging data and messages (for example, private systems or systems managed by government bodies), by way of maintenance and funding of these systems, the potential of evolution of existing systems, etc. [7]. In smaller countries that do not have strong and economically developed ports, large and demanding projects such as building and implementing ICT port systems, i.e. a PCS, requires synergy between a wide range of authorities governed by public law, economic subjects involved in work processes, and scientific community. The European Port Community System

Association defines PCS as a neutral and open electronic platform enabling intelligent and secure exchange of information between public and private stakeholders in order to improve the efficiency and competitive positions of the seaport communities [8].

In Croatian seaports, stakeholders mostly own and operate separate information and communication systems, except in the port of Ploče, where the PCS was implemented in 2012 [9]. At the moment, PCS is not implemented in the Port of Rijeka. Implementation was initially planned for 2008. Four valid offers were received and vendor was selected and started with initial analysis and implementations. Project was suspended in 2011, partly due to the development of the Croatian Integrated Maritime Information System (CIMIS) and change of focus.

The effectuation of PCS has been continued mid-2017 and in 2021, the Port of Rijeka PCS system should be fully developed. The project is being developed in partnership with the European Union through CEF – Instrument (Connecting Europe Facility – CEF), and part of that project is the Contract for technical support in design and implementation of PCS for the project *"Upgrade of the Rijeka Port Infrastructure – Port Community Information System"*. The project is financed in partnership with the Republic of Croatia and the European Union in the following ratio: 85% from the European funds and 15% from the budget of the Republic of Croatia via The Ministry of the Sea, Transport and Infrastructure, i.e. Port of Rijeka Authority. The expected value of the entire project is 1.7 million euro [10]. The agreed price of work and installation, i.e. the entire development costs of the Port of Rijeka PCS is 8.003.100 HRK [11].

The importance of this investment is reflected on the Port of Split because of the similarities between these two ports, so the design of the Port of Rijeka PCS system can be used and adapted to the Port of Split, as further explained below.

When in 2021 (after the end of the Port of Rijeka PCS development) the Croatian Ministry of the Sea, Transport and Infrastructure takes over and adapts the source code of the Port of Rijeka PCS system to the needs of the Port of Split, the authors expect that the adjustment costs should be around 12% of the implementation costs of the system in the Port of Rijeka, based on the adjustment experience of similar PCS solutions. Therefore, the actual adjustment costs of the PCS system for the Port of Split would be 960.372 HRK.

The project holder, in this case the Port of Split Authority, can obtain a product without initial operative problems because the system will go through the testing phase in the Port of Rijeka. The mentioned costs include the takeover of the project from the Port of Rijeka and its adjustment to the specific needs of the Port of Split, under the coordination from the Ministry of the Sea, Transport and Infrastructure.

The adjustment and development of the PCS system in the Port of Split will require technical support in order to develop the system as rapidly as possible. From the authors´ experience, those costs will be equal to the technical support costs for the initial development of the Port of Rijeka PCS system. The agreed price for the technical support in the project for the initial Port of Rijeka PCS system is 2.297.312,50 HRK [12], and this amount will be used as a reference for the Port of Split PCS system development. This includes all the connected costs, such as the cost of foreseen compliance with the National Single Window platform, analysis of the documents flow, and other issues which were (and will be) defined in the development phase of the Port of Rijeka PCS system.

The experience of more developed ports demonstrates that positive economic and financial evaluation of the project is the basic prerequisite for making the decision about the development and introduction of the PCS system. Financing such projects which bring economic benefits to public institutions and legal persons can be done through various models, but the model that is used in the development of the Port of Rijeka PCS system is surely the most attractive one due to the fact that significant European fund grants can be obtained. Since the investment in the Port of Rijeka is supported by the European Union and the Republic of Croatia through co-funding via the CEF fund, the project for the Port of Split PCS system should also be developed using the same or similar funding sources.

The most profitable development model, which enables additional savings in purchase and installation of hardware and program licenses, is the one where Port of Split PCS is implemented using the existing hardware of the Port of Rijeka PCS system. The investment costs would then include the adjustment costs of the equipment of the Port of Rijeka PCS system existing capacities. The important issue is ensuring the safe data storage system. Backup data storage would be stationed in Rijeka, while in case of need for disaster recovery, the Port of Rijeka and Ploče PCS could serve as redundant platforms.

According to projections in the bill of quantities for the Port of Rijeka PCS system maintenance after 31/12/2020, the costs of annual maintenance are 313.824 HRK with VAT, and the costs are relevant to system support licenses for the server and operating system, as well as for data links and SMS gateway.

The costs, which can be planned for the period of 5 years are annual maintenance, hardware adjustment and technical assistance (Table I). Surely the most profitable solution would be that the Ports of Split and Rijeka, as well as other ports, have a standardized maintenance system besides the unique hardware, which could be delivered by the local suppliers.

TABLE I. COSTS OF IMPLEMENTATION AND FIVE-YEAR MAINTENANCE (PCS SPLIT, IN CROATIAN KUNAS)

Hardware adjustment	**960.372**
Technical assistance	**2.297.312**
Maintenance	**1.569.116**
Total	**4.826.800**

III. THE IMPACT OF THE PCS SYSTEM ON THE USERS

The basic role of cargo ports is to provide services to end users, ship-owners and cargo owners. Various concession holders also offer their services in the port area, while the control of the business processes is done by the competent institutions. In the Port of Split, providing services to ship-owners and cargo owners is carried out by 67 different legal entities, who conduct their business under the management of the Port of Split Authority which grants concessions. There are 17 shipping agencies, 12 cargo forwarding, 15 port area and warehouses concessionaires, 2 fumigation and pest control companies, 6 ship chandlers, 6 fuel bunker supply companies, 6 cargo superintendence, 1 pilot service, 1 mooring company, and 1 harbour towage company in that area. Institutions that are under the state governance are Customs, Police, Harbourmaster´s Office Split, Sanitary Inspection, and Port of Split Authority.

During the implementation it is necessary to perform process re-engineering in order to achieve a high level of optimization and avoid multiple data entry. The success of the re-engineering will be limited by the current legislation and the sole will of participants who will need to change their established business methods. The re-engineering should be used for the process standardization in all ports of the Republic of Croatia, however the methodology of doing the same business processes may differ from port to port. The interpretation of the procedures not only differs from port to port, but different officials within the same port interpret legislation and procedures differently, which often leads to one institution implementing different procedures in different shifts. According to [13], by re-engineering only a part of the administrative processes, 48,5% of time needed to perform tasks within the port cluster can be saved, which can also bring financial savings to all companies which run their businesses in the Port of Split. Except for the financial savings, the use of PCS in the port cluster would also enable the standardization of electronic documents and business procedures. This way, multiple data entry would be avoided, which would consequently minimize errors during the data entry process.

Shipping agent needs about 7 hours to complete the tasks related to ships' arrival. Cargo forwarder needs around 6 hours per bill of lading, for ships with more bills of lading, e.g. container ships, more time is needed because more stakeholders are included in those processes. The time needed for concession holders in port areas and warehouses to perform customs procedures is very difficult to determine, but time saving will surely be achieved by facilitating communication with stakeholders and customs. Disposition of cargo would be delivered to the customs through the system. In the process of obtaining licenses and permission for bunker delivery on ships, it takes 3 hours for obtaining licenses and 30 minutes for procedures per bunker delivery. Ship chandlers spend 45 minutes per delivery of goods due to customs procedures, for example in 2017 in the Port of Split 218 ship supply were made in international transport [14]. Concession holders participated in surveys, and according to their experiences, the time needed to perform the same processes is listed in Table II below.

TABLE II. AVERAGE TIME FOR ADMINISTRATIVE PROCESSES OF EACH SERVICE

Activity	WH	Index	SH
Shipping agent	7	0,485	3,395
Forwarder	5	0,485	2,425
Stevedores	0,5	0,485	0,2425
Pilot	1	0,485	0,485
Customs Office	2	0,485	0,97
Total hours	15,5		7,5175

WH - Working hours,

SH - Saved hours,

Index - Possible savings with PCS.

Data was analysed only for the processes which are mandatory in procedures related to the cargo processing during the vessel call in port analysis does not include processes performed by companies, and which are not included in every call e.g. towing services, Harbour master´s Office which processes most data through CIMIS and Border Police information management system.

From the period of 2013 until 2017 the Port of Split had an average of 642,6 calls from merchant ships per year.

AAC - average annual number of calls in Split cargo port,
THwh - Total working hours on each call,
TH - Total number of working hours.

$$AAC * THwh = TH \quad (1)$$

$$642{,}6 * 15{,}5 = 9.966{,}5$$

The previous calculation shows that the mentioned participants annually spend 9.960,5 administrative commercial hours on average. As it has already been stated, the full use of the PCS system can be obtained by re-engineering administrative and legal processes and then the time savings can amount up to 48,5%. The mentioned data will be used as an index in the calculation below.

TH - Total number of working hours,
INDEX – 0,485,
Tsh - Total saved hours per average year.

$$TH * Index = THsh \quad (2)$$

$$9.966{,}5 * 0{,}485 = 4.833{,}75$$

According to the above calculation, the use of the PCS system and process re-engineering in the cargo Port of Split in an average year can lead to savings of a minimum of 4.830,75 hours.

Since the motive for this research was to perform a preliminary analysis of cost-effectiveness of the PCS system for the entire port cluster, the next step will be to determine how much time can be saved by its implementation. The total savings will be calculated in

such a way that the expected number of saved working hours is multiplied by the average gross hourly wage in the transport and storage sector. The average hourly wage is set by the data form the Croatian Bureau of Statistic (CBS). The hourly wage for May 2020 in the previously stated sector was 59,22 HRK [15] gross.

Tsh - Total saved hours per average year,
AHW - Average hourly wage,
GS - Gross savings for 1 year.

$$Tsh * AWH = GS \quad (3)$$

$$4833{,}75 * 59{,}22 = 286.254{,}67$$

As it is obvious from the calculation, expected savings in one year of use for participants that were part of this research are 286.254,70 HRK. In the 5-year period, the amount of savings is 1,431,273,5 HRK. If the average annual traffic growth in the port of 0,64% is calculated as well, then the possible savings in the observed part of administrative and legal affairs in the 5-year period would be 1.440.433,65 HRK.

According to these calculations, using the PCS system for the processes included in this research, savings can be obtained in the amount of 1.440.433,65 HRK, which constitutes 29,80% of total investment with 5-year maintenance. If the Port of Split PCS system is financed by the same model as in the Port of Rijeka, where the project is developed and financed in partnership with the European Union through CEF – Instrument in the ratio of 85%-15%, adjustment and implementation of the PCS system for the Port of Split would cost 4.826.800 HRK, and the amount which would be financed by the Republic of Croatia would be 724.020 HRK, as shown in the Table III. Finally, all participants who will be included in the Port of Split PCS system operation obtain savings in the observed processes in the amount of 778.576 HRK, which makes 107.53% of the investment, i.e. that part of the investment which was supposed to be financed from the Croatian sources. It is important to emphasize that the mentioned amounts make the minimal expected savings since the calculations involve only one part of the business process which will be included into the PCS system. The complete calculation will be possible after all business processes are defined which will include the Port of Split PCS. From everything stated above, it can be concluded that the PCS system is an indispensable investment which will ensure a faster development of business in the port and increase competitiveness among cargo ports in the Adriatic, while the port development will consequently carry more cargo and more business opportunities to cluster participants.

TABLE III. INVESTMENT AND SAVINGS RATIO FOR THE REPUBLIC OF CROATIA (IN CROATIAN KUNAS)

Business subject savings	1.440.433,65
Investment for the Republic of Croatia	724.020
Economic benefits of the investment	778.576
Index of cost-effectiveness	1.04

These are mainly cost saving benefits arising from reduced operating expenses and reduced overall inland transport costs. Much of these savings are realized outside of the immediate port area and do not produce direct revenue to the area. However, the reduced costs could result in improvement of the ports competitive position and in the attraction of additional cargo to the port, thereby producing revenues and benefits noted in the above section. The transportation savings can also improve competitiveness of the port area, which can result in the expansion of the markets of the industries in the area [16].

A challenging period awaits the Port of Split Authority, since the authorities in charge of the port must create business plans which will optimize business methods and aim to increase the volume of business activities in the future. The optimization of business methods represents the transition from ″AS IS″ to ″TO BE″ condition [17].

- "AS IS" state represents continuing business operations with a lower level of ICT technology implementation, i.e. continuing the conservative business methods of the port cluster in the Port of Split, maintaining increased administrative and legal costs, which in the long term can result in the loss of market share and less traffic in the port.
- "TO BE" state, which includes the implementation of the PCS system and process re-engineering, should bring financial savings and accelerate administrative, legal and commercial business procedures related to the ships and cargo. Increasing the level of use of ICT technologies is one of the segments which help port systems to provide the most quality services to the users, with the aim of increasing the competitiveness of the port business.

IV. DISCUSSION

The operative PCS system generates expenses which are related to the management of the system operation and periodic (annual) maintenance. Financing these expenses will be crucial to the use of the full potential that the PCS system offers to the port cluster. If the development and implementation of the system are financed in the anticipated way, by co-financing 85% from the European funds, the use of the system should be free of charge for all users in the first five years of exploitation, and the system should be managed by the Port of Split Authority, which should have capable personnel for those needs.

Annual costs for the system maintenance should be circa 313.824 HRK if the system is hosted on the Port of Rijeka PCS system hardware, the mentioned amount regards the hardware licenses and the system software as well as data traffic. Service desk employees' costs are not listed, but they should be determined by an annex to the contract for the Port of Rijeka PCS system maintenance.

A legal framework that prescribes the business methods for all economic subjects in the port area of ports open for public transport requires participants to hold a concession or approval for work from the competent port

authority. This excludes public institutions that deal with process surveillance and coordination.

Concession holders or holders of approval for work must pay a fee whose amount is set by the contract between the Port Authority and the holder. Since the Ministry of the Sea, Transport and Infrastructure does not foresee paying for the use of the PCS system, and the Port Authority already charges holders a fee for conducting their business in the port area, the PCS system should be the contribution of the Port Authority to raising the quality of port services. If the use of the PCS system is free for all participants, this will surely motivate all potential users to start using it for all the benefits it offers. The Port of Split Authority should present the PCS system to all participants in the port cluster through seminars and workshops before the implementation and release of the system, and motivate cluster participants in that way. Unfortunately, at the moment it is not possible to force the stakeholders to use the PCS system.

The authors believe that a system which is free for users is an optimal model of exploitation, enabling users' access to modern information technologies without new expenses, which can often be de-stimulating. If the users recognize the advantages and adjust to the new business methods accordingly, and adjust internal organizational business processes, they can obtain significant financial savings. Those who manage to use the given opportunities and make savings by re-engineering internal business processes can offer their services at a lower price on the market in order to increase competitiveness of their own business or lower the price of services and increase profitability. Finally, the effects of business informatization will increase the competitiveness of the traffic route and the port.

V. CONCLUSION

The purpose of this paper was to determine the economic effects of the development and implementation of the PCS system based on the example of the Port of Split, which could continue as a follow-up of the CEF instrument project whose aim is to develop an information infrastructure in the Port of Rijeka through the project ″Upgrade of the Rijeka Port Infrastructure – Port Community Information System″. As the basis for the calculation a medium number of calls was determined that was made in the period from 2013 to 2017 and a potential number of saved working hours for all those included in the administrative and legal business activities. The basic prerequisite for achieving the savings can be obtained in the case of reengineering of business processes, but that reengineering can be achieved only as far as the current legal provisions allow it. The results show that savings of 1.440.433,65 HRK can be made in the 5-year period, which is 29.80% of 4.826.800 HRK, which is the amount of the expected adjustment, implementation and maintenance costs of the Port of Split PCS system. The Croatian side would finance this project in the ratio of 15%, which represents the amount of the investment of 724.020 HRK. Maintenance costs in the period of 5 years are expected to be 1.569.116 HRK. Development, implementation and maintenance costs are determined using premise that the Port of Split system will be running on Port of Rijeka hardware, which would enable significant savings in the initial development of the project.

Since the port is being managed through the public institution of the Port of Split Authority, which in its essence is a non-profit organization, and it makes significant income from the activities of companies that are concession holders in the port area, the Port of Split Authority has the obligation to enable all participants the use of the PCS system in their business activities to help the development of the entire port community. This way the Port of Split Authority can add to its function in the community and become the economic development holder in the traffic route and business competitiveness of its surroundings.

ACKNOWLEDGMENT

This work was supported by "DigLogs – Digitalising Logistics Processes" (Interreg V-A Italy – Croatia 2014-2020) project and by University of Rijeka under the Faculty of Maritime Studies projects.

REFERENCES

[1] Ş. C. Gherghina, M. Onofrei, G. Vintilă, and D. Ş. Armeanu, "Empirical evidence from EU-28 countries on resilient transport infrastructure systems and sustainable economic growth," MDPI Sustain., vol. 10, no. 8, 2018.

[2] M. Jović, M. Filipović, E. Tijan, and M. Jardas, "A review of blockchain technology implementation in shipping industry," Pomorstvo, vol. 33, no. 2, pp. 140–148, 2019.

[3] F. Srour, M. V. Oosterhout, P. V. Baalen and R. Zuidwijk, "Port Community System Implementation: Lessons Learned from International Scan," Transportation Research Board 87th Annual Meeting, vol. 08-2041, 2008.

[4] V. Carlan, C. Sys, T. Vanelslander, "How port community systems can contribute to port competitiveness: Developing a cost–benefit framework," Transportation Business & Management, vol. 19, pp. 51-64, 2016.

[5] Port Authority Split, "Operational report for 2017," retrieved October 30, 2018, from: https://portsplit.hr

[6] E. Tijan, M. Jardas, S. Aksentijević, and A. Perić Hadžić, "Integrating Maritime National Single Window with Port Community System – Case Study Croatia," in 31ST Bled eConference - Digital Transformation: Meeting the Challenges Conference Proceedings, 2018, pp. 1–11.

[7] E. Tijan, A. Agatić, and D. Cisic, "Single Window Concept In Croatian Seaport Clusters," CITEM Conference on International Trade, Education and Marketing 2012., retrived April 24, 2019 from https://www.researchgate.net/profile/Edvard_Tijan/publication/284169724_Single_Window_Concept_In_Croatian_Seaport_Clusters/links/5c5c5ae6a6fdccb608af307c/Single-Window-Concept-In-Croatian-Seaport-Clusters.pdf

[8] International Port Community Systems Associations, "How to develop Port Community Systems," 2015, retrieved September 1, 2019, from http://www.ipcsa.international/armoury/resources/ipcsa-guide-english-2015.pdf

[9] M. Jović, "Digital transformation of Croatian seaports," in 32nd Bled eConference: Humanizing Technology for a Sustainable Society Conference Proceedings / Doctoral Consortium, 2019, pp. 1147–1164

[10] Ministry of the Sea, Transport and Infrastructure, retrieved November 6, 2018 from http://promet-eufondovi.hr/news/potpisani-ugovori-za-razvojne-projekte-luke-rijeka-financirane-kroz-cef-instrument/

[11] Isporuka i instalacija PCS sustava, Ugovor broj: I 11-18 L.U., Lučka uprava Rijeka, pp. 5, April, 2019

[12] Tehnička pomoć u dizajnu i implementaciji "Port Community System", Ugovor broj: 60/18 L.U., Lučka uprava Rijeka, pp. 5, 2018.

[13] E. Tijan, "Integral model of eletronic exchange of data in see port clusters," PhD Thesis, Faculty of Maritime Studies, International University of Rijeka, 2012.

[14] Amec Rijekatekstil d.o.o., 2018, retrieved November 20, 2019 from www.amec.hr

[15] Croatian Bureau of Statistic, retrieved November 20, 2018 from https://www.dzs.hr/Hrv_Eng/Pokazatelji/ZAPOSLENOST%20i%20place/PLACE.xlsx

[16] N. Guler, "Economic evaluation of port investments," Journal of Maritime & Transportation Sciences, vol. 40, no. 1, pp. 205-221, 2002.

[17] C. Vaghi, L. Lucietti, "Cost benefits of speeding up reporting formalities in maritime transport," Transportation Research Procedia, vol. 14, pp. 213-222, 2016.

Digital business models in the logistics services

A. Agatić*, T. Poletan Jugović*, E. Tijan***, A. Jugović****
* University of Rijeka, Faculty of Maritime Studies/Department of Maritime and Transportation Technology, Rijeka, Croatia
agatic@pfri.hr
** University of Rijeka, Faculty of Maritime Studies/Department of Maritime and Transportation Technology, Rijeka, Croatia
poletan@pfri.hr
*** University of Rijeka, Faculty of Maritime Studies/Department of Maritime Logistics and Management, Rijeka, Croatia
etijan@pfri.hr
**** University of Rijeka, Faculty of Maritime Studies/Department of Maritime Logistics and Management, Rijeka, Croatia
ajugovic@pfri.hr

***Abstract* – Logistics operators are facing various challenges in provision of the logistics services: digitalization, changes in delivering and working processes, changes of routes in the international trade, changes in commerce including e-commerce growth and networking between companies. The digitalization is currently changing logistics services and logistics operators are forced to adapt or create new business models. In this paper, authors research the challenges i.e. influencing factors which logistics operators are facing in the provision of the logistics services. Furthermore, authors will define the groups of logistics operators currently present at the market, from the aspect of digitalization. The focus of this paper is to analyse digital business models of logistics services.**

Keywords – digitalization; logistics operators; logistics services; business models

I. INTRODUCTION

The digitalization process strongly influences logistics i.e. logistics services and changes the existing business models. The logistics services could be defined as activities involving the execution of one or many logistics functions (transport services, warehousing, customs, information services, etc.) towards the ordering party by the service provider, based on contractual provisions [1]. The global logistics market is set to expand from $8 trillion in 2015 to $15 trillion by 2023 [2]. Logistics operators offer wide range of logistics services to demanding customers which expect logistics services to be fast, transparent, reliable, flexible, and provided with minimal costs. Furthermore, customers are aware of the possibilities of various digital technologies. Traditional business models in which digital technologies are minimally used are rapidly transforming to digital business models. Digital business models are based on the digitalization i.e. the use of digital technologies to provide logistics services [3]. Large logistics operators are currently adapting the existing business models to remain competitive among the new "digitalized" logistics operators. The digitalized logistics operators, which are new entrants on the market, were the first in using the advantages of digital technologies (Cloud, Internet of Things - IoT, Big Data, etc.) to introduce digital business models. Competition between traditional and digitalized logistics operators leads towards more specialization and consequently to implementation of the various new digital business models.

In this paper, the authors will provide a brief review of the influencing factors which are expected to determine the future of logistics services. Furthermore, the authors will define the basic groups of the logistics operators according to digitalization level in their business models. The goal of this paper is to analyze the digital business models used in provision of logistics services, which are expected to be increasingly used in the future.

II. TRENDS IN THE LOGISTICS SERVICES

In this chapter the authors analyze trends in the transport and logistics (T&L) industry, which present the challenges for logistics operators and will affect the provision of logistics services.

The consultant company PricewaterhouseCoopers (PwC) conducted a survey of trends in the transport and logistics industry (T&L) with the aim of identifying the factors which will influence the transport and logistics market in the Central and Eastern Europe. According to the provided survey there are five factors which will shape the future of logistics services [4]:

1. Digitalization: changes the providing of logistics services and enables redesigning of existing and developing of new services, redesigning of existing and implementation of new business models, new transaction types, new marketplaces and new revenue sources. The provided survey showed that logistics service providers are motivated to make changes in their business and implement digital technologies: 54% to increase the

This work has been financially supported by University of Rijeka under the Faculty of Maritime Studies projects.

revenue, 16% to increase the profit and 11% to ensure better customer experience. Digitalization has enabled logistics service providers to create new services in addition to traditional logistics services such as transportation. For example, data management and analytics along the supply chain are new services, which enable better planning, problem forecasting and improve delivery reliability [5]. Furthermore, digitalization has enabled online ordering, online tracking and online payments, which improves the customers experience. Digitalization will make the entire logistics supply chain – from suppliers, purchasing, producers, warehousing, commissioning, distribution, logistics and trading to the end-customer – monitored and optimized in real time [6]. Thus, logistics service providers are aware of the importance to invest in digitalization of logistics services. For example, DHL launched the "Strategy 2025," in which they announced the investment of 2 billion euro ($2.19 billion) to speed up the digitalization process [7]. It is estimated that digitalization in the logistics industry will increase up to 1.5 trillion USD value by 2025 [8].

2. Software driven process change: automation due to development of Artificial Intelligence (AI) and Robotic Process Automation (RPA) solutions will replace human workforce and make tracking calculation or claims management faster. For example, Gleematic is the A. I. robotic process automation (RPA) software used in Singapore. It can control computers and complete jobs traditionally done by white-collar workers [9]. Furthermore, predictive maintenance will stabilize the delivery times and ensure constant availability of the fleet. Thus, logistics services will be smoother, safer, more reliable and customer friendly. Logistics operators will be able to avoid unnecessary maintenance costs and mistakes in simple, repetitive processes, which will improve control in the overall process of the logistics services. It is expected that the global Intelligent Transportation System (ITS) market is expected to grow from USD 46.16 billion in 2018 to USD 64.92 billion by 2023 [10]. The global predictive maintenance market (using the digital technologies to determine condition of the equipment and machines, while reducing time and costs [11]) is expected to grow by 37%, reaching value over 10.9 billion USD in 2022 [4].

3. Machine driven process changes: improvement of the machine-driven process changes can increase the efficiency of deliveries and warehousing. Logistics services could be improved through more efficient warehousing using new transport technologies. Digital technologies which are expected to be leading in next generation of warehouse automation include: driverless cars, drones, image recognition technologies, the Internet of Things (IoT) and machine learning [12]. Furthermore, automation of the core operations will substitute the human workforce and make logistics services more reliable, faster and efficient. For example, using the robots in material handling process (pick and place, palletization, packaging) [13]. Global Material Handling Robots market size is expected to surpass 13.8 billion USD in value by 2026 [13]. To users and employees, machine driven process changes will mean interacting less with people resulting in greater availability of flexible logistics services. Overall costs of use may be one of the main factors to make users shift from traditional engines and in long term, such solutions may reduce the impact of using fossil fuels from the cost and environmental perspectives.

4. Shifts in the international trade: the future of the international trade will be characterized by development of new transport and trade corridors. For example, the corridors between China and the EU within the ‘’One Belt One Road Initiative-OBOR’’ are expected to develop more over the next few years. The OBOR development will lead to lower costs of transport and enable logistics operators to increase cooperation and create new services. This shift in the international trade will reduce costs and provide larger accessibility to products and goods from different foreign markets. Another example is the Northern Sea Route. The development of the Northern Sea Route will create shifts in trade between Asia and Europe, diversion of trade within Europe and decrease the traffic through Suez channel [14]. Thus, it will change the provision of logistics services. New corridors will shorten the transit and delivery time. Furthermore, new transport and trade corridors will integrate the infrastructure facilities, policies, institutions, and investments to spur wider socio–economic development [15].

5. Changes in commerce: eCommerce growing across regions, coupled with increased optimization in logistics, could create a push for sharing economy and value chain integrations between logistics companies, eCommerce and producers. According to the provided survey, 55% of logistics operators are planning to form new strategic alliances or joint ventures, while 38% plan new mergers and acquisitions, and 32% plan to collaborate with entrepreneurs and startups. Logistics operators could increase their revenues through the creation of a new services or extension of existing services to support eCommerce. Furthermore, networking between companies will reduce the costs and enable resource sharing. For example, costly and complicated processes like last-mile delivery and reverse logistics mean that online retailers may be constantly adapting to meet customer needs and stay competitive. This highlights the importance of collaboration between logistics services providers and eCommerce companies. This can contribute to minimization of unexpected transportation costs [16]. In 2019, eCommerce share of total global retail sales was 14.1% and it is expected to increase 2% a year through 2023. The eCommerce is expected to value more than $6.5 billion USD by 2023 [17].

III. INNOVATIVE BUSINESS MODELS IN THE PROVISION OF LOGISTICS SERVICES

Digitalization process in the logistics services resulted in division of the logistics operators into two main groups [18]:

1. Incumbent logistics operators: the logistics operators (e.g. Maersk, DHL, Panalpina) which provide a

wide range of different logistics services through developed business networking. The digitalization and appearance of the digitalized logistics operators definitely moved incumbent logistics operators towards the digitization of their services and change of existing business models. Still, digitization of their business models is slower than the digitalized logistics operators.

2. Digitalized logistics operators: these companies are actually digital startups (e.g. Flexport, Freighthub, Freightos) which are transforming traditional provision of logistics services by using digital technologies (Big Data, Internet of Things - IoT, Cloud technology, etc.) to network all the stakeholders/parties involved in providing logistics services via online digital platforms. The digitalized logistics operators are developing business models for one or more segments of the logistics services. Digitalized logistics operators enabled added value to their customers: ease of access, price transparency and real time integrated services [19].

Digitalized logistics operators made a breakthrough in logistics services' business models by implementing digital business models. The main advantage of digitalized logistics operators is the ability to process numerous transactions and handle large amounts of data online, all in a short time [20]. Furthermore, digitalized logistics operators forced the incumbent logistics operators to adapt their business models [20].

Hereafter, the authors provide a review of the digital business models which are expected to transform the logistics services [19], [21], [22].

Matching demand and supply: logistics operators providing services via online digital platforms, enabling shippers to search for cargo space and capacity providers such as carriers, thus enabling automation of routing and pricing. Freightos is an example of a company which provides instant freight quotes and enables online capacity booking immediately at a given rate. Freightos functions as a SaaS-Software as a Service technology. By using the Freightos' platform, logistics companies can manage and upload contracts and other documents. The platform also provides additional information about fees, legal issues, penalties, ship location and availability, weather data, price, schedules from shippers, shipowners data, etc. Another example is the NYSHEX [23], which specialized in matching demand and supply by providing the contracting services in the ocean shipping via online digital platform. NYSHEX provides online contract services for CMA CGM, Cosco Shipping, Hapag-Lloyd, One, HMM and Maersk [23]. Using the NYSHEX platform, shippers and carriers increase planning accuracy. A large majority (96.4%) of the contracts executed on the NYSHEX platform in 2019 were fulfilled, meaning that carriers accepted the cargo for the shipping while cargo owners and non-vessel-operating common carriers (NVOs) delivered the containers they contracted [24].

Connectivity through platforms: consolidation and integration of the information across the supply chain. The goal is to enable end-to-end visibility via the online digital platform. Single platform enables company's automation of trade and logistics processes with their partners. For example, FourKites, a pioneer company in end-to-end visibility, developed the platform by combining machine learning with real-time data to provide complete visibility of the entire process of logistics services to the worlds' largest "Fortune 500" companies and third-party logistics firms across 70 countries [25]. The platform is also optimized for mobile phones. FourKites uses a network of more than four million GPS/electronic logging devices (ELD) and covers all transport modes: ocean, rail, intermodal, over-the-road, LTL and parcel [25]. Another example, Infor Nexus platform, connects entire supply chain-suppliers and manufacturers, brokers, logistics services providers and banks, through a single cloud-based platform. Infor Nexus platform united the physical and financial supply chain processes in a unified cloud-based platform, providing visibility of the entire supply chain [26].

Forwarding Freight Digitally: digital freight forwarders (DFFs) use the online digital platform to offer more logistics services than in the previous digital models. The DFFs goal is to provide a seamless user experience of shipping goods while collecting information on a single platform. Platforms are replacing manual processes and paperwork with instant price quotations, simple and standardized document management, simple communication process, and easy access to real time data that can be used to track shipments. This digital business model consists of two sub models:

1. Digital forwarders with external operations: the logistics operators having extensive operational capabilities in-house i.e. group of experts which can offer know-how and physical assets to local agencies and forwarding service providers. Furthermore, integral part of this model are online digital platforms. Digital forwarders with external operations enable their contractors to avoid operational complexity while taking the direct control over the process. For example, the Flexport provides a digital online platform for direct booking with carriers and contracting with local forwarders [27]. Furthermore, customers online digital platform is managed by the group of experts (custom expert, data expert, etc.) which are carrying out the whole service process. Through the online platform, the transaction costs of email and phone communications are eliminated while enabling customers more visibility into entire supply chain. The Flexport clients spend 30% less time to find the shipment information, while 88% of clients are satisfied with easy access to reporting and analytics which helps them make more data-informed decisions [27]. Another example is Transmetrics, which implemented an automated forecasting tool based on Artificial Intelligence and specifically customized for freight forwarders. The tool predicts customer demand per asset type and geographical area, weeks in advance. This enables forwarding companies to plan and procure the right capacities, reserve the space needed, at the right locations and the right time [28].

2. Digital forwarders with internal operations: the companies offering comprehensive services. These companies are actually incumbent logistics operators which are digitalizing their business models. They operate on a large geographical area and offer more methods of transport as well as local customer service. For example, Kuehne+Nagel, implemented "myKN" online digital platform which enables planning and managing shipments in one place: shipment information, quote request, booking and tracking of shipments [29]. Furthermore, Kuehne+Nagel implemented "Zebraxx" online digital platform for small and medium-sized partners who do not have the necessary IT infrastructure [29]. Customers can track their goods with their mobile phones, from the moment they are handed over in the warehouse until delivery [29].

Customers as competitors: The customers of logistics operators are also entering the logistics services market. These new logistics operators see the opportunity to gain control of the complete online customer experience by implementing digital technologies. Amazon, for example, is moving into air and ocean freight forwarding. Amazon is developing the Prime Air service - fully electric drones that can fly up to 15 miles and deliver packages under five pounds to customers in less than 30 minutes, since 75 to 90 percent of purchased items are under that weight limit [30]. Another example is Cargomatic, provider of the online marketplace which instantly connects shippers with a network of Cargomatic owned trucks, trucking companies and owner-operators. Shippers post local loads and truckers deliver them. The Cargomatic handles all the particulars in between including GPS tracking, real-time updates and alerts [31]. Another example is Deliv, company which provides same day delivery services [32]. Deliv uses digital technologies of omni-channel platforms and GPS-enabled smartphones to manage the drivers' fleet [32].

Niche logistics operators: Digitalization opens space for different niche logistics operators which provide specific logistics services, substituting the standard logistics services. For example, 3D printing has a potential to substitute shipments by printing required goods on demand and directly on site. Further technical evolution of 3D printing has a potential to increase number of printable products and reduce the need for conventional logistics [33]. The parcel lockers are another example of potential niche logistics service. A parcel locker is a secure container that enables to receive large packages. Parcel lockers are especially practical for packages that do not fit into standard mailboxes. Rather than making multiple delivery attempts, shippers can leave packages in a parcel locker, where they will be held for the recipient to collect [34]. Customers can choose from an increasing number of pick-up stations, which can be considered as an individualized niche service. For example, the Package Nexus Smart Locker provides an optimal solution for unattended package delivery [35]. It is a cloud-based electronic package delivery system. The Package Nexus Smart Lockers give customers 24-hours access to deliver and retrieve packages while tracking the shipment. Thus, some standard logistics services could be substituted by niche logistics services that fulfill specialized requirements for the specific customers groups [33].

Digital business models are changing business operations, enabling logistics operators to be more informed and to enhance forecasting [21]. Furthermore, logistics operators can automate shipment booking process, reduce manual carrier booking and provide instant updates on schedules and changes. Beside the changes for the logistics operators, digital business models also make changes for customers [21]. Customers can for example obtain instant price quotations. Furthermore, digital business models provide seamless end-to-end documentation handling for customers, as well as easy shipment tracking and automatic invoicing.

IV. CONCLUSION

The future of logistics will be determined by several influencing factors, which will consequently affect the providing of logistics services: digitalization; automation of processes; automation of warehousing and new transport technologies; development of new transport and trade corridors; eCommerce growth, etc. Digitalization has already changed the provision of logistics services through the adaptation of existing and implementation of new business models, development of new services, new transaction types, new marketplaces and new revenue sources. The digitalization process has separated the logistics operators in two main groups: incumbent logistics operators and digitalized logistics operators. New digitalized logistics operators have utilized the advantages of digital technologies to enter the logistics market and have thus implemented new digital business models. Incumbent logistics operators were forced to respond with adaptation of their existing business models. The digital business model of matching demand and supply enables the logistics operators which provide services via online digital platforms to connect the shippers in search of cargo space with capacity providers such as carriers. Connectivity through platforms is a digital business model of logistics operators who are specialized in consolidation and integration of the information across the supply chain. The goal is to enable end-to-end visibility via the online digital platform. Forwarding Freight Digitally is a business model provided by logistics operators (digital freight forwarders) which use the online digital platform to provide a seamless customer experience of shipping goods while collecting information on a single platform. The platform reduces manual processes and paperwork with instant price quotations. This digital business model consists of two sub models: Digital forwarders with external operations who have groups of experts which can offer know-how to local agencies and forwarding service providers, all via online digital platform. Digital forwarders with internal operations are actually incumbent logistics operators offering comprehensive services on a large geographical area and with more transport methods by using the possibilities of online digital platforms. Customers as competitors is a digital business model implemented by customers of logistics operators, with the tendency to enter the logistics services market. These new

logistics operators see the opportunity in digital technologies to gain control of the complete online customer experience. Niche logistics operators are an example of the recent digital business model which aims at providing specific logistics services, substituting the standard logistics services. For example, 3D printing has a potential to substitute shipments by printing required goods on demand and directly on site, reducing the need for conventional logistics. Digital business models in logistics services make changes in operations of logistics operators and in customers experience: information accessibility, enhanced forecasting, shipment booking automation, instant updates on schedules and changes., instant price quotation, seamless end-to-end documentation handling, easy shipment tracking and automatic invoicing.

REFERENCES

[1] S. Ziyadin et.al., ''Differentiation of logistics services on the basis ABC analysis'', E3S Web of Conferences, Vol. 159, The 1st International Conference on Business Technology for a Sustainable Environmental System (BTSES), 2020, pp. 1-8, [Online]. Available: https://www.e3s-conferences.org/articles/e3sconf/pdf/2020/19/e3sconf_btses2020_04034.pdf [Accessed: 17-Apr-2020]

[2] Turvo, [Online]. Available: https://turvo.com/company [Accessed: 17-Apr-2020]

[3] N. Meyer, D. Horvat, M. Hitzler and C. Doll, '' Business models for freight and logistics services'', Fraunhofer Institute for Systems and Innovation Research ISI, Econostor, 2018, [Online]. Available: https://www.econstor.eu/bitstream/10419/179111/1/1023085038.pdf [Accessed: 17-Apr-2020]

[4] PricewaterhouseCoopers, ''Five Forces Transforming Transport & Logistics; PwC CEE Transport & Logistics Trend Book 2019'' [Online]. Available: https://www.pwc.com/hu/hu/kiadvanyok/assets/pdf/transport-logistics-trendbook-2019-en.pdf [Accessed: 17-Apr-2020]

[5] F. Heistermann, M., ten Hompel and T. Torsten Mallée, ''Digitisation in Logistics'', BVL International; Bundesvereinigung Logistik (BVL) e.V., 2017, [Online]. Available: https://www.bvl.de/en/positionpaper-digitisation [Accessed 24-May-2020]

[6] DHL, ''Digitalization – a potential revolution for the logistics industry'' [Online]. Available: https://www.dhl.com/global-en/home/about-us/delivered-magazine/articles/2017/issue-5-2017/digitalization-a-potential-revolution-for-the-logistics-industry.html [Accessed: 17-Apr-2020]

[7] DHL, ''Strategy 2025: Deutsche Post DHL Group accelerates growth in core businesses and invests EUR 2 billion in digital transformation'' [Online]. Available: https://www.dpdhl.com/en/media-relations/press-releases/2019/strategy-2025-deutsche-post-dhl-group-accelerates-growth-in-core-businesses-and-invests-eur-2-billion-in-digital-transformation.html [Accessed 24-May-2020]

[8] World Economic Forum, ''Delivering change: digital transformation in logistics'' [Online]. Available: https://reports.weforum.org/digital-transformation/delivering-change-digital-transformation-in-logistics/ [Accessed: 17-Apr-2020]

[9] Gleematic, [Online]. Available: https://www.robotics-process automation.com/singapore/ [Accessed 24-May-2020]

[10] Markets and Markets, ''Intelligent Transportation System Market'' [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/intelligent-transport-systems-its-market-764.html [Accessed: 17-Apr-2020]

[11] Deloitte, ''Making maintenance smarter; Predictive maintenance and the digital supply network'' [Online]. Available: https://www2.deloitte.com/us/en/insights/focus/industry-4-0/using-predictive-technologies-for-asset-maintenance.html [Accessed: 17-Apr-2020]

[12] Blue Yonder, ''Warehouse of the Future: Digitalization and New Technologies Driving Next Generation Warehouse Automation'' [Online]. Available: https://blog.blueyonder.com/warehouse-of-the-future-digitalization-and-new-technologies-driving-next-generation-warehouse-automation-part-2-of-2/ [Accessed 24-May-2020]

[13] GlobeNewswire, ''Global Material Handling Robots Market'' [Online]. Available: https://www.globenewswire.com/news-release/2020/03/30/2008157/0/en/Global-Material-Handling-Robots-Market-Size-to-surpass-USD-13-8-Billion-in-value-by-2026.html

[14] World Trade Institute, ''Opening up the northern sea route will change the face of trade" [Online]. Available, https://www.wti.org/institute/news/479/opening-up-the-northern-sea-route-will-change-the-face-of-trade/ [Accessed 24-May-2020]

[15] A. S. M. Abdul Quium, ''Transport Corridors for Wider Socio–Economic Development'', Sustainability 2019, 11(19), [Online]. doi: https://doi.org/10.3390/su11195248

[16] CAI Transportation Blog, ''Three ways e-commerce is changing the logistics industry '' [Online]. Available: http://blog.capps.com/3-ways-e-commerce-is-changing-the-logistics-industry [Accessed 24-May-2020]

[17] Shopifyplus, ''What Is the Future of eCommerce? 10 Insights on the Evolution of an Industry'' [Online]. Available: https://www.shopify.com/enterprise/the-future-of-ecommerce [Accessed 24-May-2020]

[18] PricewaterhouseCoopers, 'Shifting Patterns; The future of the logistics industry'', 2016, [Online]. Available: https://www.pwc.com/hu/hu/kiadvanyok/assets/pdf/transport-logistics-trendbook-2019-en.pdf [Accessed: 17-Apr-2020]

[19] DHL Trend Research, ''Logistics Trend Radar; Delivering insight today, creating value tomorrow'', Version 2018/2019, [Online]. Available: https://www.dhl.com/content/dam/dhl/global/core/documents/pdf/glo-core-trend-radar-widescreen.pdf [Accessed: 17-Apr-2020]

[20] Oliver Wyman, ''How start-ups are digitalizing logistics'' [Online]. Available: https://www.oliverwyman.com/our-expertise/insights/2017/sep/oliver-wyman-transport-and-logistics-2017/innovations/how-start-ups-digitalize-logistics.html [Accessed: 20-Apr-2020]

[21] Boston Consulting Group-BCG, ''The Digital Imperative in Freight Forwarding'' [Online]. Available: https://www.bcg.com/publications/2018/digital-imperative-freight-forwarding.aspx [Accessed 20-Apr-2020]

[22] Freightos, ''The Digital Age Is The New Competitive Arena For 3PLs'' [Online]. Available: https://www.freightos.com/digital-age-new-competitive-arena-3pls/ [Accessed 25-Apr-2020]

[23] NYSHEX, [Online]. Available: https://www.nyshex.com/ [Accessed 20-Apr-2020]

[24] JOC.com, ''NYSHEX; Volume tripled, membership doubled in 2019'' [Online]. Available: https://www.joc.com/maritime-news/container-lines/nyshex-volume-tripled-membership-doubled-2019_20200129.html [Accessed 21-Apr-2020]

[25] FourKites, [Online]. Available: https://www.fourkites.com/platform/ [Accessed 21-Apr-2020]

[26] Infor Nexus, [Online]. Available: https://www.infor.com/products/infor-nexus [Accessed 22-Apr-2020]

[27] Flexport, ''The new Flexport Platform'' [Online]. Available: https://www.flexport.com/our-tech-platform/ [Accessed 22-Apr-2020]

[28] Transmetrics, [Online]. Available: https://transmetrics.eu/product/freight-forwarding/ [Accessed 22-Apr-2020]

[29] Kuehne+Nagel, ''Technology and Innovation; The Ease of Doing Business'' [Online]. Available: https://2018-annual-report.kuehne-nagel.com/in-fokus/customer-technology-people/people [Accessed 22-Apr-2020]

[30] The Verge, ‘’Amazon’s new transforming Prime Air delivery drone‘’ [Online].
Available: https://www.theverge.com/2019/6/5/18654044/amazon-prime-air-delivery-drone-new-design-safety-transforming-flight-video [Accessed 21-Apr-2020]

[31] Cargomatic, [Online]. Available: https://cargomatic.com/ [Accessed 21-Apr-2020]

[32] Deliv, ‘’Technology ‘’ [Online]. Available: https://www.deliv.co/about/# [Accessed 20-Apr-2020]

[33] Hofmann, F. Osterwalder, ‘’Third-Party Logistics Providers in the Digital Age: Towards a New Competitive Arena’’, Logistics 2017, 1, 9, [Online]. Available: https://www.mdpi.com/2305-6290/1/2/9 [Accessed 21-Apr-2020]

[34] Expert Market, ‘’Parcel Lockers’’ [Online]. Available: https://www.expertmarket.com/postage-meters/what-is-a-parcel-locker [Accessed 01-May-2020]

[35] Package Nexsus, ‘’Smart Lockers’’ [Online]. Available: https://packagenexus.com/smart-lockers/ [Accessed 01-May-2020]

Using fuzzy logic in analysing and modelling the reflection of monetary and fiscal conditions on GDP per capita in Croatia

A. Samodol*, S. Brlečić Valčić*, A. Ostojić*
* Libertas International University, Zagreb, Croatia
asamodol@libertas.hr
sbrlecic@libertas.hr
aostojic@libertas.hr

Abstract* - In this paper, the monetary conditions are observed through loans to the central government, loans to the local government, loans to non-financial associations, excess liquidity and the primary liquidity rate. The fiscal conditions are observed through total revenue, total expenditure, deficit/surplus, domestic and foreign debt. Modelling of the reflection of thus defined monetary and fiscal conditions on the Gross domestic product (GDP) per capita is observed using data from 2002 to 2018. With clustering, the established problem required the creation of 4 independent models. The models were created by using fuzzy logic due to incongruity, class limits and the subsequent uncertainty and truth bias that may have appeared in the established problem. The results of research conducted in this way point that the reflection of monetary and fiscal conditions on GDP per capita should be observed through three individual and one main model. Such approach indicates an interrelationship of individual variables with foreign debt, deficit/surplus, total expenditure and finally GDP per capita. These models can serve the monetary and fiscal authorities in the analysis and modelling of an effective interaction of monetary and fiscal conditions aimed at increasing GDP per capita.

***Keywords* -** *monetary conditions, fiscal conditions, GDP per capita, fuzzy logic*

I. INTRODUCTION

The European Fiscal Board published an assessment of the general orientation of the fiscal policy in the Eurozone, with consideration of forecasts, emphasising that the Eurozone is going through a period of weaker growth. Considering the economic and geopolitical uncertainty, economically more developed countries are advised to take a neutral fiscal stand.

Those member states that have not yet reached good fiscal positions are advised to construct fiscal buffers in accordance with the regulations of the Stability and Growth Pact. Countries with a high share of public debt are especially warned to make an effort to orient it toward a strong downward trajectory. As opposed to that, large member states whose budget surplus exceeds the needs of the Pact should use a part of the available fiscal space to invest in and support economic growth.

In terms of monetary conditions, a strong, long-term aversion to risk and financial vulnerability continues to grow after years of low interest rates; strengthening of disinflation pressures is increasing difficulties in debt serving; the space of monetary policy for combating a fall is narrowing and finally, adverse shocks are taking place, which are causing more long-term effects than those recorded in the past.

It is thus without question that the approach to monetary and fiscal policy interaction should be changed. It should be tailored to each member state, and finally, the objectives of each individually tailored interaction should follow trends with consideration of: increase in demand within and outside the Eurozone (domestic and foreign demand); reduction of risks that may be caused by increased demand; further decrease in unemployment; possible tightening of financing terms for high-debt countries; the need to strengthen the production sector; uncertainty of the international trade system; positive expenditure dynamics; more available income, based on fiscal incentives; further employment growth; lack of immunity of labour market to global trends; inflation rate at a lower level than expected; the fact that companies absorb higher unit costs of labour through lower profit margins; lower trend of selling prices in the production sector and higher in the service industry and slow investment growth.

The research part of the paper, based on selected data in the period from 2002 to 2018, using fuzzy logic, analyses the existing conditions, i.e. reflects on effects of monetary and fiscal policies on GDP per capita.

II. THEORETICAL BACKGROUND

Central banks generally have similar general goals, but different operational and indirect goals. In order to monitor the effectiveness of measures and techniques they use in achieving these goals, it is necessary to approach the analysis in a different way from the one that was established in previous years. Namely, a lot of economic processes result from the action of several factors, of which each has different term goals. Therefore, the chronology itself should be observed in a different way, for example, as a range of components acting on different frequencies. Some considerably outdated research [1] has already proven, for example, that if we observe the monetary policy measures through a longer period of time and in an accurate way, the restrictive monetary policies help control inflation.

The standard sticky-price models imply incredibly large output and inflation reactions to interest rate changes in distant future. Such results lead to a widespread concern and doubt in the limited use of such frameworks in creating and analysing possible effects of future policies.

Recently, researchers have suggested different micro-founded modifications of the standard modelling framework that reduce future effect on economic activity. Dynamics of thus modified models is consistent with the standard New Keynesian model [2].

Production and price increase react differently within the two models based on open and closed economy. In case that the economy is subject to a negative shock of monetary policy in a closed economy, the production decreases, and inflation initially rises, and then falls. On the other hand, in an open economy, in case of a significant decrease in liquidity, there is a relatively persistent reduction of production and increase in inflation. A positive shock in the factor of total productivity increases production and reduces inflation equally in closed and in open economy [3].

Through credit channels, the shocks of monetary policy or macroeconomic shocks impact the lending activity of banks both in terms of supply and demand [4]. However, commercial banks mostly rely on their own financial circumstances (market power, capital, risk, liquidity, profitability) in setting a strategy of defence from external shocks. The increase in loan placements impacts economic growth [5]; however, the effect is not as strong as expected [6], which also depends on the type of loan placement.

Unconventional and unorthodox monetary policy can lead to an improved approach to crediting by decreasing the risk level of asset class and by well-conceived policies, which, even during crisis periods enable access to crediting.

Some authors [7] say that for the assessment of fiscal policy sustainability, a sufficient condition is that the public debt is on a stationary trajectory, while others [8] point to the necessity of abandoning the conventional approach to fiscal stance, because they vary and are not as constant as it is supposed in most standard econometric tests of fiscal sustainability. The ability to prove fiscal solvency, in the form of strong positive conditioned relationship between public debt and primary fiscal balance is much more successful with more developed economies and those whose debt level is lower than 50 % of GDP [9]. The assessment of fiscal policy sustainability cannot be static. In other words, a static assessment often moves from an assumption that non-interest government spending will fall or stay the same in accordance with the current budget policies and continuation of favourable cyclic factors with guaranteed stability of public debt (Blanchard et al., 1990). This is supported by research that takes large time series of data into consideration, such as the one done by [10] that took into consideration a period from 1795 to 2012. They point to evidence that governments react more seriously and effectively to a decrease in budget deficit when the debt level is high, as opposed to low.

Therefore, the conventional deficit measures, whether real, cyclically corrected or adjusted to inflation, are a bad indicator of the real fiscal stance of the government. An effective public debt and deficit management should be observed through four key factors [11]. The first one relates to the probability of inflation increase in the future, which can take place in economies in which the volume for income increase by generating money compared to an increase in tax is not limited, and by measures that will impact a decrease in inflation.

The second factor relates to the assessment of consistency of a specific programme of spending and taxes, i.e. an adjustment that must be made in the plans of spending and tax burdening in order for the government to remain solvent. The third factor relates to the fact that an increase in budget deficit will only supress private spending, which depends on the state of the economy, time horizon and whether the increased deficit is temporary or permanent. The fourth factor relates to the exclusion of an existing measure from the fiscal system along with the development of a model that serves to analyse this impact and suggests a good decision in a specific economy.

Sustainability assessment cannot always be limited to medium term; however, government liabilities to specific programmes have consequences to the fiscal policy that reach far into the future. Therefore, the anticipated future pressures, such as social expenditure due to aging population, should be included in each of the measures aimed at latent sustainability issues [12].

Fiscal sustainability enhancement in emerging economies can be achieved by defining financial

policy priorities of the state and constructing adaptive institutional patterns of the budget system based on budget management, which enable the optimisation of the relationship between system elements and the cyclical nature of the socio-economic development [13].

In the last 25 years, views on coordinating policies in the interaction between monetary and fiscal policies have changed. The consequences (whether positive or negative) that have resulted from the cooperation/non-cooperation between an individual government and its central bank have instigated research into measures and the interrelationship of monetary and fiscal policy measures, aimed at achieving a stable monetary anchor. This primarily relates to an optimal level of stability of prices and coordination of monetary and fiscal policies necessary to achieve it [14]. When it comes to the stabilisation of government debt, the monetary and fiscal authorities are involved in a dynamic conflict.

Fiscal authorities tend toward limiting public debt via the usage of fiscal instruments, while the monetary authorities tend toward limiting public debt on an established level, along with tendencies to keep the increase of money stock as close to the aimed value as possible [15].

Within contemporary research, recommendations for an effective interaction between monetary and fiscal conditions include [16]: targeted inflation of 2 % and operational independence of the central bank; study into and assessment of the consequences of the feasibility of specific policies, especially those in circumstances of slow growth and low inflation; direction of automatic stabilizers toward cyclical variations of deduction of investment tax, real estate tax, VAT and substitution of estimated income tax based on data on the one collected in the previous year; specification of conditions by the government in case of implementation of fiscal expansion; during the business cycle, to keep the share of net public investment in the GDP at a consistent, sustainable level, and subsequently collect tax, considering the potential of economic growth; establishment of an Office for Budget Responsibility aimed at assessing fiscal policy and analysis of the shortcomings of governmental financial initiatives; a five-year plan of public sector balance as a percentage of the GDP; setting the upper limit to government spending at 47 % of the GDP; avoidance of any temptation to use foreign currency interventions for any purpose, except smooth adjustment.

III. DATA SET AND METHODOLOGY

Within the monetary framework, based on selected data, two types of bank loans are confronted, excess liquidity of banks (as surpluses of banks' liquidity above the amount necessary to maintain the obligatory reserves) and the banks' primary liquidity rate, as indicators of a bank's liquidity.

Excess liquidity in HRK (Croatian kuna) is defined as a difference between the achieved average on a bank's settlement account and in the cash register (until October 2008), and from November 2008 as a difference between the achieved average on the settlement account in the period of maintaining the part of the obligatory reserves in HRK and the minimum of the necessary average on the settlement account according to the calculation of the part of the obligatory reserves in HRK.

The primary liquidity rate is a share expressed as a percentage of the monthly average of daily statuses of excess liquidity in banks in HRK in a monthly average of daily statuses of deposits, which form the basis for calculating the obligatory reserves.

The fiscal framework (general government) confronts budget income and expenditure, budget deficit and the method of financing budget expenditure through domestic and foreign debt.

The source of data for the table showing nonfinancial accounts of the government is the Croatian Bureau of Statistics (CBS). The methodological basis for data elaboration is the European System of National and Regional Accounts (ESA) and the Manual on Government Deficit and Debt (MGDD). The reach of institutional sector of the government is defined by the Decision on Statistical Classification of Institutional Sectors (Official Gazette, No 1/2015) that is officially adopted and published by the CBS, and which is based on the ESA 2010 methodology, according to which the government is divided into: central government, social security funds and local government. The principle of calculation upon recording revenue and expenditure is defined as one of the basic methodological principles of ESA 2010, whereby for specific tax revenues, the method of time adjustment of money transactions is applied, while other tax revenues are shown on pure cash principle. Furthermore, this statistical approach applies methodological adjustments defined by the provisions of the ESA 2010 and MGDD. The most important ones among them for the general government sector are the treatment and recording of investments, public-private partnerships and concessions, capital transfers in cases of statistical debt assumption, payment of guarantees and capital injections as non-financial transactions. Likewise, this methodology includes special guidelines in recording revenue from issuing the right to emit greenhouse gas, cost of procuring military equipment, revenue from issuing UMTS licences and in implementing revenue corrections based on super-dividend test as well as

corrections related to EU fund transactions (CNB, Bulletin, 255, year XXV, November, 2019.)

Initially, the development of fuzzy sets theory was initiated by noticing the lack of effectiveness of traditional techniques and analysis systems in problems with complex or poorly defined variables so they could be realistically described in classical differential equations. The usual thread passing through the problems of this type is the incongruity of class limits and the subsequent inaccuracy, uncertainty and truth bias. The fuzzy set concept offers a way out of such a reality through reaction that serves as a starting point for the development of theories that exhibit modelling ability in such circumstances [17].

Due to this, it was selected in this research, and the theory of the concept itself has already been described in [18].

Although there are a great number of different membership functions, two very frequently applied membership functions were used for the purpose of this paper. For all input variables the generalized bell membership function (gbellmf) was used, expressed as

$$f(x) = \frac{1}{1 + \left|(x-c)/a\right|^{2b}} \tag{1}$$

dependent upon coefficients a, b and c, where generally $b > 0$, and c defines the centre of the curve.

For output variable the triangular membership function (trimf) was used, expressed as

$$f(x) = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{b-a}, & a \le x \le b \\ \dfrac{c-x}{c-b}, & b \le x \le c \\ 0, & c \le x \end{cases} \tag{2}$$

where parameters a and c allocate the points of the triangular function on the x-axis, while parameter b allocates the peak of the triangle.

Fuzzy sets with linguistic rules expressed as

IF (x is LVx) THEN (y is LVy) (3)

where LVx and LVy are the values of linguistic variables defined by fuzzy sets over a range of values of input sets X and Y, respectively.

Each observation shown in Table 1 and Table 2, with consideration of a specific year represents one fuzzy rule.

All values are scaled from 1 to 5, whereby 1 represents the lowest value and 5 the highest value. An equal number of linguistic variables is assumed for all input and output variables. In this case, linguistic variables can be said to belong to the set

$$\{LV_1, LV_2, ..., LV_m\}. \tag{4}$$

The clustering method of self-organising neural network (SOM) [19] was used to select parameters in a specific model.

IV. RESULTS

The "blood work" itself of the suggested problem, i.e. the relatedness of the suggested variables to the GDP per capita, shown in Picture 1, indicates a high complexity and impossibility of direct connection.

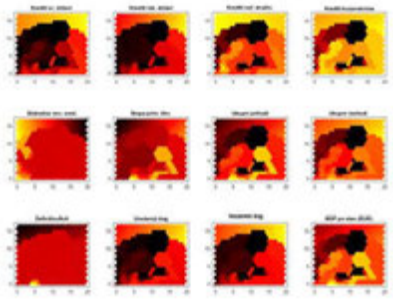


Figure 1: Clustering of suggested parameters by using SOM
Source: authors

Therefore, the problem of reflection of monetary and fiscal conditions on the GDP should be observed through 4 separated models shown in Table 1.

Model 1-financing government debt - loans to central government (LCG) - loans to local government (LLG) - domestic debt (DD) **output – foreign debt (FD)**	*Model 2-currency policy* - excess liquidity (EL) - primary liquidity rate (PLR) **output - deficit/surplus (D/S)**
Model 3-crediting the private sector *- loans to non-financial associations (LNA)* *- loans to households* ***(LH)*** ***output – total expenditure (TE)***	*Model 4-reflection of monetary and fiscal conditions on GDP per capita* - loans to households (LH) - primary liquidity rate (PLR) - total revenue (TR) - total expenditure (TE) **output – GDP per capita (GDP)**

Table 1: Models for the analysis of monetary and fiscal conditions on GDP per capita
(Source: authors)

Namely, models 1 – 3 represent the specificity of Croatian monetary and fiscal conditions that should be understood and related before processing, results and interpretation of model 4, which represents the reflection of monetary and fiscal conditions on GDP per capita.

Model 1, shown in Picture 2, shows the result of "financing general government debt", in such a way that selected variables are confronted: "loans to central government and social security funds" and "loans to local government", with consideration of the monetary component "domestic debt" and the fiscal component "foreign debt".

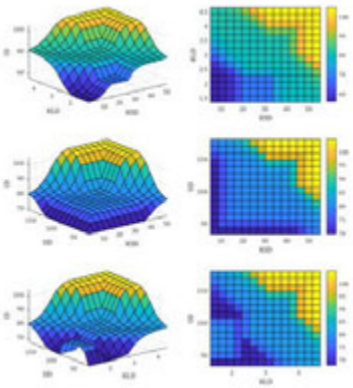


Figure 2: Model 1
Source: authors

The model describes different possibilities of financing and refinancing the existing general government debt.

For example, lower levels of foreign debt (60 to 70 billion) require the levels of loans to the central government from 10 to 20 billion, loans to the local government from 1.5 to 3 billion, domestic debt up to 50 billion and the level of loans to the local government up to 3 billion. The combination of around 80 billion of foreign debt and 50 and over billion of loans to the central government is, of course, valid.

Such a significant substitution of foreign financing sources would lead to a greater possibility of placing the surpluses of bank liquidity (HRK 30 billion), which are just slightly lower than the total bank loans to the central government (36 billion) at the end of 2018. On the other hand, and in keeping with the new trends in the interactions of fiscal and monetary conditions, thus targeted conditions would need to take into consideration that the assessment of the sustainability of the fiscal policy cannot be static, but perspective. In other words, a static assessment often lies on the assumption that non-interest spending of the general government will decrease or remain unchanged in accordance with the current budget policies and the continuation of favourable cyclic factors, along with a guaranteed stability of public debt. The analysis indicates that in our case we are dealing with a stasis without the willingness to change it.

A perspective policy requires an assessment of consistency of a specific programme of spending and taxation, i.e. an adjustment that needs to be implemented in plans of spending and tax burdening so the government would remain solvent. Therefore, it primarily requires the study and assessment of consequences of the feasibility of individual policies in conditions of insufficient growth and low inflation.

Model 2, shown in Picture 3, shows the results and consequences of the currency policy of the central bank, the monetary anchor of which is the currency exchange rate, and not the money supply.

Due to an intensive implementation of final operations of buying foreign currency in order to prevent the strengthening of Croatian kuna against the euro, in the period from 2016 to 2018 excess liquidity of banks tripled, and the primary liquidity rate doubled.

The model confronts "excess liquidity", "primary liquidity rate" within the monetary framework and "deficit/surplus" within the fiscal framework.

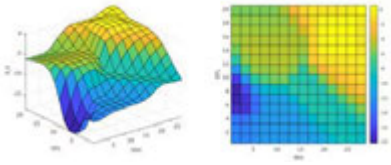
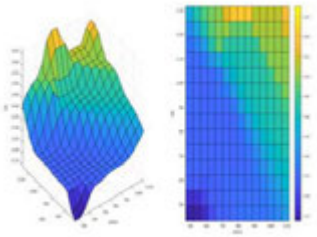


Figure 3: Model 2
Source: authors

According to the model, a surplus requires a high level of excess liquidity from 18 to 30 billion in set conditions, and also a high level of primary liquidity rate from 14 – 18 billion. Likewise, a surplus is possible with the combination of excess liquidity from around 20 billion and a high liquidity rate of about 20 %, which, again, points to the paradox contained in the low crediting activity of banks on the one hand, and a continued high daily liquidity of banks on the other.

In relation to newer, previously mentioned research, and since we can say that Croatian economy is open in the case of, obviously highly necessary, significant decrease of liquidity, this could impact the reduction of an already low level of production and an increase in inflation, which would be acceptable up to a level of 2 %.

Model 3, shown in Picture 4, represents the activity of bank lending to the private sector. The model connects "loans to households" and "loans to non-financial associations" within the framework of monetary components and "total expenditure" of the general government within the fiscal framework.



Picture 4: Model 3
Source: authors

The model shows that the lower values of total expenditure are related to loans to non-financial associations moving from 50 to 90 billion, with loans to households moving from 50 to 80 billion at the same time. Likewise, loans to households can assume values up to 120 billion, but with a simultaneous value of loans to non-financial associations up to 55 billion. This possibility of substitution (decrease) of loans to households and increase of loans to non-financial associations for about 40 billion points to the inertia of the monetary authority and the lack of utilizing other measures and instruments of the monetary policy.

High levels of loans to non-financial associations (110 billion) and loans to households (over 130 billion) are related to high levels of total expenditure (over 160 billion).

The above model clearly indicates the need for micro-founded modifications of the standard modelling framework based on the effects of a positive impact on economic activities. The necessity of increasing loan placements to companies in our case requires an unconventional and unorthodox monetary policy that can lead to an enhancement of approach to crediting by decreasing risk levels of the asset class and by well-conceived policies, which enable access to loans even in times of crisis, to efficient micro, small and medium enterprises, thus substituting inefficient debtors.

Finally, the Model 4, shown in Picture 5 unites the variables, for which the initial analysis established to have a direct impact on the GDP per capita, namely from the standpoint of monetary conditions "loans to households" and "primary liquidity rate" with the fiscal conditions of "total revenue" and "total expenditure" of the general government.

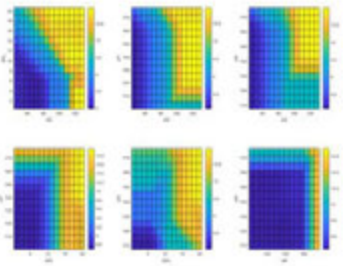


Figure 5: Model 4
Source: authors

It also suggests high levels of loans to households (from HRK 110 to 120 billion) with high levels of primary liquidity rate (18 to 20 %) and total revenue level above the total expenditure (around 170 billion). If we relate the analysis of this model to the guidelines of contemporary research into the interaction of monetary and fiscal conditions, the monetary policy should follow the direction of automatic stabilizers toward cyclic variations of tax deductions for investments, real estate tax, VAT and the substitution of the estimated income tax based on data on the collected amount in the previous year, along with a five-year plan of the public sector balance as a percentage of the GDP and setting an upper limit to government spending to 47 % of the GDP.

Before concluding this paper, it should be emphasised that clustering by using SOM in Model 4 does not include loans to non-financial associations (enterprises), which speaks volumes about the condition of Croatian economy and about what the movement of GDP per capita is founded on. An extremely high liquidity of banks does not serve as a crediting potential directed towards companies but dominantly towards households. On the other hand, the monetary authorities with the mission to defend further strengthening of the HRK currency against the euro by buying foreign currency additionally increases the available liquidity surpluses of banks. Finally, by continually raising the level of expenditure on the "growth" model established in such a way, the general government is maintaining an inappropriate share of expenditure in the GDP and is "killing" the potential of GDP per capita growth through a different policy of general government spending.

V. CONCLUSION

The results of research conducted in this paper point to a conclusion that the reflection of monetary and fiscal conditions on GDP per capita should be observed through three individual and one main model.

The results of thus established models show that lower levels of foreign debt (60 to 70 billion) require lower levels of loans to the central government from 10 to 20 billion, loans to the local government from 1.5. to 3 billion, domestic debt up to 50 billion and the level of loans to the local government up to 3 billion. Certainly, the previous combination is not the only one, but the model describes different possibilities of financing and refinancing general government debt.

According to the model shown, a surplus requires, in established conditions, a high level of excess liquidity from 18 to 26 billion, and also a high level of the primary liquidity rate from 14 to

18 %. However, a surplus is also possible with a different combination of excess liquidity (around 20 billion) and a primary liquidity rate of around 20 %, which points to a paradox consisting of a low crediting activity of banks on the one hand, and a high daily liquidity of banks on the other.

Lower levels of total expenditure are related to loans to non-financial associations moving from 50 to 90 billion with a simultaneous movement of loans to households from 50 to 80 billion. Likewise, loans to households can assume a value of 120 billion, but with simultaneous values of loans to non-financial associations up to 55 billion. This possibility of substituting (decreasing) loans to households and increasing loans to non-financial associations for about HRK 40 billion points to additional and unused unconventional measures of monetary policy.

Higher levels of GDP per capita require high levels of loans to households (from 110 to 120 billion) with high levels of primary liquidity rate (18 to 20 %) and total revenue at a level above total expenditure (around 170 billion). Such a result of the main model that shows GDP per capita as a reflection of monetary and fiscal policy, points to a misalignment and tentativeness of monetary and fiscal policy makers in the Republic of Croatia.

Therefore, this research can serve monetary and fiscal authorities in establishing more creative and effective policies.

ACKNOWLEDGMENT

This work was supported by the Libertas International University project "Interaction of financial friction of the hotel industry and total financial friction in the Republic of Croatia".

REFERENCES

[1] Aguiar-Conrariaa, L., Azevedob, N., Soaresc, M., J., (2008). Using wavelets to decompose the time–frequency effects of monetary policy. *Physica A: Statistical Mechanics and its Applications, 387* (12), pp 2863-2878.

[2] Nakata, T., Ogaki, R., Schmidt, S., Yoo, P., (2019). Attenuating the forward guidance puzzle: Implications for optimal monetary policy, Journal of Economic Dynamics & Control, 105, pp 90-106.

[3] Wondemhunegn, E.M., (2019). Business cycles in Ethiopia under alternative monetary policy rules, African Journal of Economic and Management Studies, 10(3), pp 299-313.

[4] Nguyen, C.P., Le, T.H., Su, T.D., (2020). Economic policy uncertainty and credit growth: Evidence from a global sample, Research in International Business and Finance, 51

[5] Wulandari, R., (2012). Do credit channel and interest rate channel play important role in monetary transmission mechanism in Indonesia?: A structural vector autoregression model, Procedia - Soc. Behav. Sci. 65, pp 557–563.

[6] Ashcraft, A.B., (2006). New evidence on the lending channel. Money Credit Bank, pp 751-775.

[7] Velinov, A., (2015). Assessing Fiscal-Policy Sustainability: On the Different States of the Debt-to-GDP Process, Public Finance Analysis, 71(4), pp 415-439.

[8] Polito, V., Wickens,M. (2012). A model-based indicator of the fiscal stance, European Economic Review, 56(3), pp 526-551.

[9] Mendoza, E. G., Ostry, J.D., (2007). International Evidence on Fiscal Solvency: Is Fiscal Policy "Responsible"?, Journal of Monetary Economics, 55, pp 1081–1093.

[10] Lo Casico, I., (2015). A wavelet analysis of US fiscal sustainability. Economic Modelling, 51, pp 33-37.

[11] Buiter, W.G., (1985). A Guide to Public Sector Debt and Deficits, Economic Policy, 1(1), pp 13–79.

[12] Blanchard, O., Chouraqui, J. C., Hagemann, P.R., Sartor, N., (1990). The sustainability of fiscal policy: new answers to an old question, OECD Economic Studies 15, pp 7–36.

[13] Harbar, Z., Harbar, V., Sobchuk, S., Menchynska, O., (2019). Fiscal policy under economic transformation, Advances in Economics, Business and Management Research, volume 95, 6th International Conference on Strategies, Models and Technologies of Economic Systems Management (SMTESM 2019), 95, pp 206-211.

[14] Canzoneri, M., Cumby, R., Diba, B., (2010). The interaction between monetary and fiscal policy. Friedman, B., Woodford, M. (Eds.), Handbook of Monetary Economics, 3, pp.935–999.

[15] Anevlavis, T., Papavassilopoulos, G., Engwerda, J., Van Arle, B., (2018). Debt Stabilization In The Presence Of Endogenous Risk Premia: A Dynamic Game Approach. Macroeconomic Dynamics, 23(7), pp 1-33.

[16] Russell Jones, R., Llewellyn, J., (2019). Reducing Inequalities. National Institute Economic Review, Vol. 250(1), pp 75-82.

[17] Bojadziev, G., Bojadziev, M., (2007). Advances in Fuzzy Systems: Applications and Theory — Vol. 23, World Scientific Publishing Co. Pte. Ltd.

[18] Brlečić Valčić, S. Bagarić, L., (2017). Return on strategic effectiveness - the need for synchronizing growth and development strategies in the hotel industry using revenue management, Ekonomska istraživanja, 30(1), pp 1631-1654.

[19] Brlečić Valčić, S. (2016). Relationship between different business value components within the oil and gas industry, Ekonomski vjesnik : časopis Ekonomskog fakulteta u Osijeku, 29(1), pp 241-256.

The Role of Perception in the Adoption of Digital Platforms in Agriculture

K. Tomičić-Pupek*, I. Pihir* and M. Tomičić Furjan*
* University of Zagreb / Faculty of Organization and Informatics, Varaždin, Croatia
ktomicic@foi.unizg.hr; ipihir@foi.unizg.hr; mtomicic@foi.unizg.hr

Abstract - **In these times of global disruptions, digital product distribution platforms are playing a major role in driving the global economy. In the development of such platforms, the scope, architecture, realization technology as well as the channels used are important. However, a well-accepted/designed platform does not guarantee profitable results in a digital business model. There are various soft factors that can significantly affect the reach of a platform. Due to its role in food distribution, the specific lifecycle of products and seasonality, agriculture is an industry where timely reach is crucial. In this paper, we investigate the factors influencing users' perceptions toward platforms in agriculture. We consider both the perspective of the customers/buyers, and the perspective of the producers/sellers. We also examine the perception factors that can influence one's intention to join and adopt digital platforms. The identification of relevant perception factors is based on a literature review and empirical data from publicly available sources. Our goal in future research is to confirm or revise factors that impact perception, and to assess the impact of perception on the design and functionality of digital platforms, as well as its impact on the development of digital business models.**

Keywords – digital platform, perception, adoption, agriculture

I. INTRODUCTION

The digital transformation of agriculture requires the use of Industry 4.0 technologies, which transform this industry into Smart Agriculture or Agriculture 4.0 [1], [2], [3]. Scientific research projects carried out in cooperation with the IT industry has resulted in the identification of the potential of digital transformation, which differs from previous efforts to digitize (i.e. to convert analog to digital contents) or digitalize (i.e. making business digital through digital exchange formats relating to business content, automation processes and introducing changes to business models). Digital transformation, unlike its predecessors, requires a more comprehensive approach. We can define it as an effort to organise continuous improvements to user-oriented digital business models and the application of modern technical potentials that enable businesses to be technologically and digitally mature. Disruptive events combined with technological advances and innovations are "transforming economies, societies, and our lives" [4].

These new technical potentials use and combine Industry 4.0 technologies such as the Internet of Things (IoT), Big Data, Artificial Intelligence (AI), Cloud Computing (CC), Blockchain and other emerging technologies. These technologies support the development of agriculture, and establish an interconnection between producers/sellers and customers/buyers of agricultural products on digital platforms. All of this together digitally transforms agriculture in all phases; from the initial phases of farming and food production all the way to packing and food distribution. It supports the entire supply chain [2], [5], [1].

Based on a report by Haberli et al. [6] on the perception of farmers, the agricultural sector in general is positive towards the expected benefits of integrating information technologies. Farmers are willing to invest jointly with governments in order to create benefits in agriculture and support the digital transformation of this sector [5], [2], [6], [1].

Considering the recent disruptions in global markets, digital platforms for the distribution of agricultural products are a suitable context for us to examine. Utility perception factors regarding digital platforms for agricultural products in rural areas represent perceived values, and influence customers' decisions to use such platforms. In this context, we use the following definitions and terms:

- Digital platform for the distribution of agricultural products: This is a unique digital space that brings together consumers and producers to buy, sell and distribute agricultural products. The term "platform" refers to a unique online website that facilitates the trade of products and/or services. This includes linking auxiliary or concomitant products or services of other providers regardless of the technical infrastructure.
- Utility: This refers to a measurable property of an object in relation to planned effects and the perception of the achieved effects.
- Perception: It refers to a subjective assessment of the objective properties of an object of interest.
- Utility perception of a digital platform for the distribution of agricultural products: This is a subjective assessment that measures the achievement of planned effects connected to the usage of a platform for the distribution of agricultural products.

With all of this in mind, this paper researches and analyzes literature, and includes literature case study

analysis and other sources in order to determine perception factors about digital platforms in the field of agriculture. Particular focus is put on food distribution; however, it should be noted that the conceptual model of perception factors could be considered through all stages of production and distribution across the entire supply chain. This is due to digital platforms having various levels of support. Factors relating to the adoption and perception of digital platforms could help researchers, industry and agricultural producers to transform their businesses at all production stages.

The paper is organized as follows. After the introduction, the methodology is presented. Our literature review constitutes the theoretical background, and prepares the reader for the results, which are presented as a conceptual model of initial utility perception factors for digital platforms focusing on agricultural products in rural areas. The paper ends with future research possibilities and the conclusion.

II. METHODOLOGY

The identification of relevant perception factors is based on a literature review and empirical data from publicly available sources. According to Siddh et al [7], the growth of empirical data and empirical studies is well marked in the agricultural sector. It should be noted though that the researchers emphasize the need to direct similar research within a "intra-functional and intra-firm scope at the organizational and supply chain level."

The literature review was conducted rather widely in order to find articles of interest for our research. Two scientific databases (Web of Science and Scopus) were used in May 2020 for the research. The search was based on a combination of keywords (described in Table I). The search focused on peer-reviewed articles in journals and conference proceedings. The search results from Web of Science (WoS) included 11 papers, while Scopus resulted in only 7 papers.

TABLE I. SEARCH STRATEGY (2015-2019)

WoS search strategy	**WoS Hits**
TOPIC: (platform) AND TOPIC: (perception) AND TOPIC: (food distribution); Refined by: PUBLICATION YEARS= (2019 OR 2016 OR 2018 OR 2015 OR 2017) AND DOCUMENT TYPES= (ARTICLE); Timespan=All years.	Results found: 11 Sum of the Times Cited: 51
Scopus search strategy	**Scopus Hits**
TITLE-ABS-KEY (platform AND perception AND food AND distribution) AND DOCTYPE (ar) AND PUBYEAR > 2014	Results found: 7 Sum of the Times Cited: 46

Source: Authors

Additional analysis was performed after the articles were collected. After merging the results from Scopus and WoS, we excluded one duplicated entry. In the end, 17 publications which were found in both databases were used for the qualitative analysis.

Because we are focusing on perception in the use of digital platforms, we have developed the following criteria for the selection of papers: (1) an article is considered relevant if it covers reports on buyer or producer factors influencing their intention of using digital platforms, and (2) process or technology-related aspects are mentioned.

After applying these criteria, only 4 publications remained for further qualitative analysis [7], [8], [9], [10]. The small number of publications allowed us to analyse the content using qualitative methods. After that, we broadened our search to platforms like Google scholar and ResearchGate in order to gather publications that are similar in topic. We applied the same selection criteria and found more significant publications addressing perception and/or willingness as topics of interest. These publications served as the basis for our conceptual model [11], [12], [13], [16].

Our broader research is oriented on issues relating to the design and development of digital platforms, which means that empirical data also plays a major role in gathering data. Therefore, we have followed and observed several regional "e-market" sites and social media groups for agricultural food distribution over the last two months. This has been conducted to identify some common trading scenarios. Even though we could not apply longitudinal scientific research methods to track regional "e-market" sites and social media groups (due to the COVID-19 pandemic), some experiences from the hype of their acceptance in public seem to be worth investigating in detail.

III. THEORETICAL BACKGROUND

Digital transformation can be characterized by 5 key determinants [14], among which customer centricity[1] and talent, capability and capacity strengthening[2] are seen as important factors that affect value for users and benefits for producers within a platform-based business model [15]. Platforms provide a virtual place where customers can get access to products and services, and where producers can deploy their services or products. Based on this premise, the implementation of online trading platforms is seen as one of the main objectives of digital transformation in agriculture [1]. Digital transformation in agriculture can also include other technologies that are either related to or integrated within the platform. This is shown in a survey by Lezochea et al. [2] through the analysis of key impacts and challenges that organizations face when applying them. In the next three subchapters, some of these impact parameters and challenges are elaborated on in more detail.

A. Perception as an impact factor

Jayashankar et al. [16] explored factors influencing farmers' willingness to adopt IoT by investigating the role of trust in technology, perceived value and risk. Taking into consideration general technology trends and concerns (like IoT and farmers' concerns about data ownership), the authors based their hypothesis on strong theoretical inputs on the transition from goods-dominant logic to service-dominant logic in agriculture, perceived utility, new combinations of technological solutions for productivity improvements and other value-based adoption model

[1] Customer centricity means that the customer is aware of his/her rights and obligations, has certain expectations, gains experience and holds attitudes that he/she shares with others.

[2] Talent, capability and capacity strengthening means that the employees know how technologies can be used, and they understand how they can be used to improve performance.

elements [16]. An interesting study on user packaging interaction by Mumani and Stone [8] shows that various touchpoints like purchasing, checking out at the cashier, handling, opening, storage, disposal and other interactions with the product are considered important in understanding buyer perceptions. Furthermore, examining the attributes of these touchpoints like graphics, information volume, and size are also relevant for understanding how buyers evaluate products.

B. Process parameters

Process parameters and measures like duration, costs, resource allocation and other performance indicators still play a significant role in assessing operational excellence. The recent global disruption caused by the pandemic has resulted in noticeable changes to the paradigm of buying goods earlier. The supply of food through digital platforms is increasing in various forms; from e-market sites to social platform groups. It is putting customers and producers on to new journey paths. New digital business models are resulting in business improvements and increased efficiency. This indicates that digital transformation is also oriented toward business transformation. In order to achieve production and operational goals, a balanced combination of technological solutions must be designed [17].

According to Siddh et al. [7] in their literature review, around 33% of research publications focus on information management, followed by 16% on sustainability management and 13% on logistic management. Some common issues that are examined as quality factors include product and raw materials, hygiene, nutrition and resources.

C. Technology-driven change

Digital transformation can be a change driven by technology. Technology concepts and applications have been thoroughly researched in that context [18], [19], [20], [21]. The EU Parliament [22] identifies 3 technology categories:

a) Technologies with a high impact on the agri-food value chain: Internet of things (IoT), Automation and Robotization, Artificial Intelligence (AI) and Traceability and Big Data;

b) Technologies with a medium impact on the agri-food value chain: Blockchain, GNSS (Global Navigation Satellite System) and Virtual Reality;

c) Technologies with a low impact on the agri-food value chain: Broadband networks, Information and Communication Technology and Platforms for e-Business.

According to Tariq et al. [9], social media platforms play a relevant role in forming the consumer's attitude, and thereby, their willingness to conduct online impulse buying. Platform features and information cues are critical in efforts to reach targeted consumer segments. Suchiradipta and Raj [23] investigated the usefulness, advantages, purpose, difficulties, and social media preferences in the agricultural sector. Their research shows that social media and platforms contribute to effective communication at various levels. According to a study from 2018 that included 29 interviews with agricultural knowledge providers in New Zealand [24], the sources of change in perception toward platforms include externally driven demands (like food safety and sustainability), or internal ones (such as the need to implement digital technologies within organizations). The same authors also investigated how knowledge providers in the agricultural sector have responded to emerging digital technologies. They concluded that the examined population can be evaluated at an early stage of response maturity and emphasized the need for strategic thinking in organizations in this sector [25].

IV. THE CONCEPTUAL MODEL OF INITIAL UTILITY PERCEPTION FACTORS FOR DIGITAL PLATFORMS FOCUSING ON AGRICULTURAL PRODUCTS IN RURAL AREAS

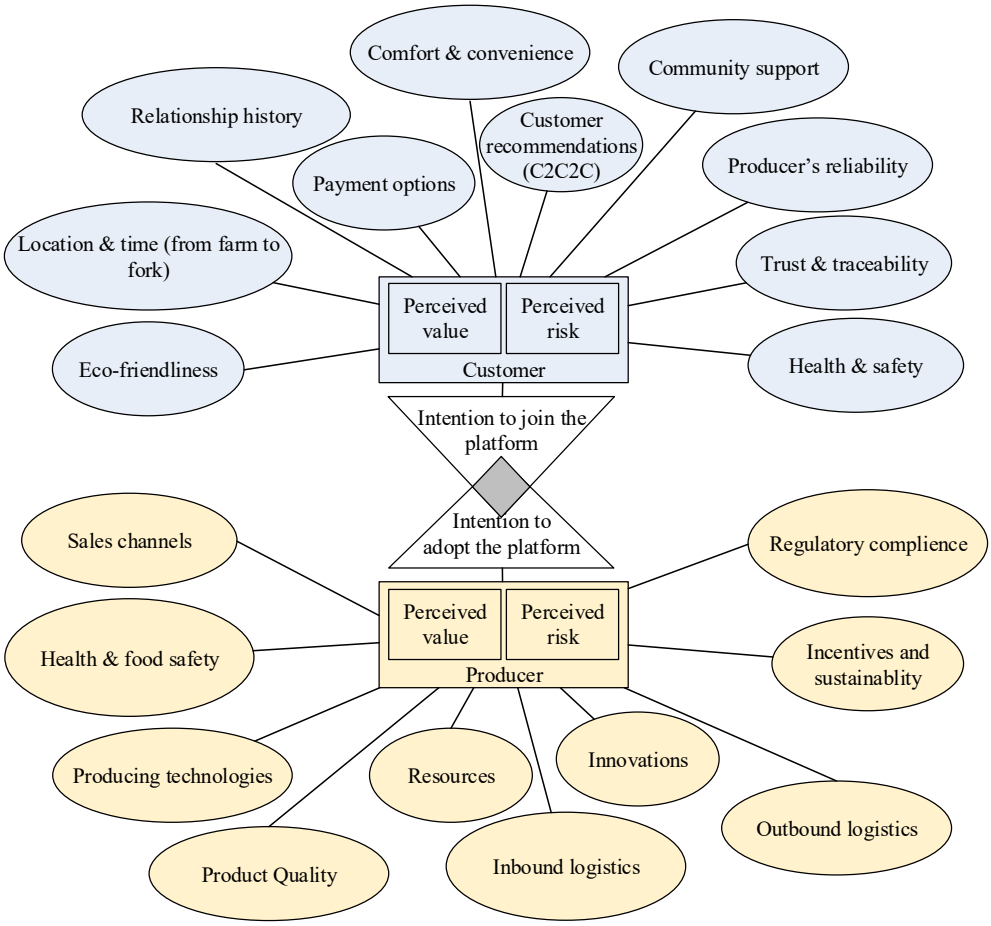
The operational efficiency and effectiveness of the process supported by modern digital technologies is just one of the factors influencing the customer's decision on how to buy agricultural products. The process of generating a product or service itself can be very efficient and effective; however, these features stem from the perspective of the process owner and his/her understanding on how the process should be formed. From the customer's perspective, it seems that the focus has shifted from being on how the process is operationally good, and more toward how well the process solves the customer's problem. Therefore, focusing on the customer by understanding the customer's expectations, needs and goals, increases the chances that the product will be accepted in the market or that the service will come to life. In short, the process is considered successful because it solves the customer's problem, and not because the delivery process is efficient and effective. In order to investigate the factors influencing the gap in customer and producer expectations, a conceptual model of initial utility perception factors from the perspective of customers and the perspective of producers can be made. This can then be used to reflect on ways to bridge the gap.

In order to examine utility perception factors regarding digital platforms for agricultural products in rural areas, we developed a conceptual model of initial utility perception factors.

It has been designed by relying mostly on the value adoption model [11], the holistic narrative of perceived value and the concepts of perceived risk and trust with IoT technology from Jayashankar et al. [16], Tariq et al.'s research model [9], Almquist's Elements of Value model [12] and the B2B Elements of Value [13].

Figure 1 shows the factors from the perspective of customers and producers. These factors can be in convergent or divergent relationships with each other, which influences the perceived value or perceived risk. This is applicable to both sides, i.e. to the producer and the customer. For the customer, the balance between perceived value and risk determines his/her intention of joining the platform in order to buy. For the producer, it impacts their intention to adopt platforms in order to sell. For each group (consumers and producers), 10 factors have been identified that determine the perception of

consumers and producers regarding the utility of a platform for the distribution of agricultural products. The factors are listed in Table II and Table III. The factors are classified within each group (customer or producer). For each factor, the process and technological key terms are described.



Source: Authors

Figure 1. Conceptual model of initial utility perception factors for digital platforms focusing on agricultural products in rural areas

TABLE II. INITIAL FACTORS OF PERCEPTION REGARDING THE UTILITY OF THE PLATFORM FOR THE DISTRIBUTION OF AGRICULTURAL PRODUCTS - CUSTOMERS

Group	Factor	Process key terms	Technological key terms
Customer	Eco-friendliness	Reducing the eco-footprint, avoidance of artificial additives, seasonality of products, eco-friendly packaging options	Open data on ecological footprints
	Location & time (from farm to fork)	Buying locally fresh products in the shortest time possible time from the farm (or field in use)	Blockchain
	Relationship history with the producer	Existing relationships with the producer, existing experiences (positive and negative)	CRM systems
	Payment options	Price, Price Volatility, Payment security, avoiding cash	Contactless payment, card payments, digital wallets (Revolut)
	Comfort & convenience	Remote ordering, time saving, reliability of delivery	Social media and platforms
	Recommendations (C2C2C)	Choosing a supplier based on recommendations and recommending it to others based on their experience (Customer-to-Customer-to-Customer, C2C2C)	Social media and platforms
	Community support	Purchasing from local producers to support the sustainability and development of the local economy, local production	Social media and platforms
	Producer's reliability	Product quality, delivery time, price	ERP systems, Blockchain
	Trust & traceability	Trust in the production process, traceability of products	Blockchain
	Health & safety	Disruption challenges such as epidemics and pandemics	Open data on disruptions

Source: Authors

TABLE III. INITIAL FACTORS OF PERCEPTION REGARDING THE UTILITY OF THE PLATFORM FOR THE DISTRIBUTION OF AGRICULTURAL PRODUCTS – PRODUCERS

Group	Factor	Process key terms	Technological key terms
Producer	Sales channels	Customer-preferred sales channels versus producer-preferred channels, channel integration	Social media and platforms
	Health & food safety	Quality of products, production process and supply chain in the delivery of health products to the end customer	Blockchain
	Production technologies	Green deal, operationally efficient and effective production process, process monitoring, climate change	IoT, drones
	Product Quality	The quality of the production process and the final product	Blockchain
	Resources	Human labour, machines, robots; a culture of continuous investment in new skills, knowledge and capacity to increase the efficiency of the production process	Robotics
	Inbound logistics / Supply Chain	Suppliers of raw materials (seedlings and seeds, fertilizer), time spent on improving the supply chain	Platform-based Supply Chain Management
	Innovations	Innovation potential, innovation process monitoring, innovation implementation, consulting services	Innovation Management System (IMS)
	Outbound logistics / Distribution chain	Downloading and preparing customer orders, delivery planning, packaging, delivery	Social networks and platforms
	Incentives and sustainability	Long-term sustainability and capacity increase, incentive system (not in the scope of the project, but affecting the readiness of producers)	Drones, digital certificates
	Regulatory compliance	Regulatory compliance requires administrative approval	Digital Certificates

Source: Authors

Different perception factors may have different impacts on the perceived value and risk. Further research is needed to confirm, revise or reject the perception factors. The customer's intention of joining a digital platform in order to buy products is the first issue impacting a platform's reach, and other factors as well as experiences could be significant with regards to customer retention on a platform.

Regarding producers, the perception factors which impact a producer's intention to adopt a platform could also be connected to the intention to develop an additional channel for product distribution. Also, other impact factors regarding platform retention need to be systemized.

Confirmed and revised perception factors impact the design of digital business models, which encompass personalized approaches to various customer segments and channels. They also impact efforts to create profitable value propositions, efficient cost structures and revenue streams.

This has an impact on the platform architecture, especially on the necessary functionalities like producer-buyer matching and filtering, presentation of offerings, delivery and payment options and even technologies used for platform implementation.

V. FUTURE RESEARCH

As our current research focuses on identifying and investigating utility perception factors of digital platforms for agricultural products, our future research will be oriented toward testing the model and examining correlation weights. This will be conducted via a focused survey with producers and consumers. Furthermore, due to the limited research on this topic, we note that there is a need to create appropriate propositions for designing digital platforms for food distribution. This can be conducted by assessing the impact of the design and functionality of certain digital platforms, and by assessing the impact of digital business models, clear value propositions and operating business models on food distribution. Within that scope, the role of technology providers should also be examined.

VI. CONCLUSION

We have investigated perception factors influencing the intention of consumers and producers to join and adopt digital product distribution platforms focusing on agricultural products. Based on the literature review and empirical data from publicly available sources, we identified 10 perception factors for each of the collaborating parties: a) consumers: eco-friendliness, location & time, relationship history with the producer, payment options, comfort & convenience, recommendations, community support, producers reliability, trust & traceability and health & safety; b) producers: sales channels, health & food safety, production technologies, product quality, resources, inbound logistics/supply chain, innovations, outbound logistics/distribution chain, incentives and sustainability, and regulatory compliance. Understanding these perception factors may contribute to determining their intention of using a digital platform, the reach of the platform and the volume of users. In our future research, we will gather user inputs to confirm or revise the identified perception factors and to assess their impact on the design and functionality of digital platforms.

We believe that the proposed conceptual model is a valid contribution that allows for the simultaneous presentation of perception factors for producers and consumers in one place. The conceptual model should be confirmed and the influence of individual factors should be investigated, but we believe that the model can help in the following manner. First, it assumes that perception factors are possible barriers regarding the consumer and producer's intention to join or adopt digital platforms for trading agricultural products. Secondly, it can serve as a means to create mutual understanding and facilitate better matching between producers and consumers. Moreover, the model acts as a tool for prioritizing key features that producers/sellers want to emphasize when communicating their strengths; and has the capacity to provide the value

proposition that customers expect. Lastly, the model could support producers in choosing the most suitable digital platform.

ACKNOWLEDGMENT

This research has been conducted as a part of the wider research within the project "Competence Centre for Digital Transformation of the Food Industry in Rural Areas." The project is co-funded by the European Union through the European Regional Development Fund (ERDF).

REFERENCES

[1]. R. Zaripova, M. Tyurina, O. Rocheva, A. Chupaev, A. Sharifullina, "Major Trends in the Digital Transformation of Agriculture," In: Advances in Economics, Business and Management Research 2019, "New Silk Road: Business Cooperation and Prospective of Economic Development", vol. 131, pp. 271-275. Atlantis Press, France, 2019.

[2]. M. Lezochea, J. E. Hernandez, M. del Mar Eva Alemany Diazc, H. Panettoa, J. Kacprzyk, "Agri-food 4.0: A survey of the supply chains and technologies for the future agriculture," in Computers in Industry, vol 117, 2020.

[3]. F. Mazzetto, F., R. Gallo, P. Sacco, "Reflections and Methodological Proposals to Treat the Concept of "Information Precision" in Smart Agriculture Practices," in Sensors, vol 20(10), 2847, 2020.

[4]. M. Chui, M. Evers, J. Manyika, A. Zheng, and T. Nisbet, "The Bio Revolution: Innovations transforming economies, societies, and our lives," McKinsey Global Institute, May 2020, https://www.mckinsey.com/mgi/overview

[5]. K. Agrawal, N. Kamboj, "Smart agriculture using IOT: A futuristic approach," in International Journal of Information Dissemination and Technology, vol. 9(4), pp. 186-190, 2019.

[6]. C. Haberli Junior, T. Oliveira, M. Yanaze, E. E. Spers, "Performance, farmer perception, and the routinisation (RO) moderation on ERP post-implementation," in Heliyon, vol. 5(6), pp. 1-13, 2019

[7]. M. M. Siddh, G. Soni, Rakesh Jain, M. Kumar Sharma, V. Yadav, "Agri-fresh food supply chain quality (AFSCQ): a literature review," in Industrial Management & Data Systems, vol. 117 (9), pp. 2015-2044, 2017.

[8]. A. Mumani, and R. Stone, "State of the art of user packaging interaction (UPI)," in Packag Technol Sci, vol. 31, pp. 401-419, 2018.

[9]. A. Tariq, C. Wang, Y. Tanveer, U. Akram, and Z. Akram, "Organic food consumerism through social commerce in China," in Asia Pacific Journal of Marketing and Logistics, vol. 31(1), pp. 202-222, 2019.

[10]. P. Vaz-Fernandes, and S. Caeiro "Students' perceptions of a food safety and quality e-learning course: a CASE study for a MSC in food consumption," in International journal of educational technology in higher education, vol. 16:37, 2019.

[11]. H. Kim, H. Chan, and S. Gupta, "Value-based adoption of mobile internet: an empirical investigation," in Decision Support Systems, vol. 43, pp. 111-126, 2007.

[12]. E. Almquist, J. Senior, and N. Bloch, "The elements of value," in Harvard Business Review, September 2016 https://hbr.org/2016/09/the-elements-of-value

[13]. E. Almquist, J. Cleghorn, and L. Sherer, "The B2B Elements of Value," in Harvard Business Review, March–April 2018. https://hbr.org/2018/03/the-b2b-elements-of-value

[14]. I. Pihir, K. Tomičić-Pupek, and M. Tomičić Furjan, „Digital Transformation Playground - Literature Review and Framework of Concepts" in Journal of Information and Organizational Sciences, vol. 43, pp. 33-48, 2019.

[15]. B. Marr, Tech Trends in practice. Wiley, UK, 2020.

[16]. P. Jayashankar, S. Nilakanta, W. J. Johnston, P. Gill, and R. Burres, "IoT adoption in agriculture: the role of trust, perceived value and risk," in Journal of Business & Industrial Marketing, vol. 33(6), pp. 804-821, 2018.

[17]. M. Tomičić Furjan, K. Tomičić-Pupek, I. Pihir, "Understanding Digital Transformation Initiatives: Case Studies Analysis," in Business Systems Research, vol. 11(1), pp. 125-141, 2020.

[18]. I. A. T Hashem, V. Chang, N. B. Anuar, K. Adewole, I. Yaqoob, A. Gani, E. Ahmed, and H. Chiroma, "The role of big data in smart city," in International Journal of Information Management, vol. 36(5), pp. 748-758, 2016.

[19]. R.T. Kreutzer, „Treiber und hinterguende der digitalen transformation". D. Schalmo et al. (Hrsg.) Digitale Transformation von Geschaeftsmodellen. Springer, pp. 33-58, 2017.

[20]. K. Schwab, The fourth industrial revolution. UK, Portfolio penguin, 2017.

[21]. E. Ismagilova, L. Hughes, Y. K. Dwivedi, and K. Ravi Raman, "Smart cities: Advances in research - An information systems perspective" in International Journal of Information Management, vol. 47, pp. 88-100, 2019.

[22]. M. Pesce, M. Kirova, K. Soma, M.-J. Bogaardt, K. Poppe, C. Thurston, C. Monfort Belles, S. Wolfert, G. Beers, and D. Urdu, "Research for AGRI Committee – Impacts of the digital economy on the food-chain and the CAP," EU parliament report: European Parliament, Policy Department for Structural and Cohesion Policies, Brussels, 2019.

[23]. B. Suchiradipta, and S. Raj, "The online culture of agriculture: exploring social media readiness of agricultural professionals," in CSI Transactions on ICT, vol. 6, pp. 289–299, 2018.

[24]. K. Rijswijk, L. Klerkx, and J. A. Turner, "Digitalisation of agricultural knowledge providers: the case of New Zealand," 13th European International Farming Systems Association (IFSA) Symposium, Farming systems: facing uncertainties and enhancing opportunities, Chania, Crete, Greece, pp. 1-14, July 2018

[25]. K. Rijswijk, L. Klerkx, and J. A. Turner, "Digitalisation in the New Zealand Agricultural Knowledge and Innovation System: Initial understandings and emerging organisational responses to digital agriculture," in NJAS - Wageningen Journal of Life Sciences, vol. 90–91, pp., 2019.

Social entrepreneurial intention: Does the classroom matter?

I. Kedmenec*
* University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
ikedmenec@foi.hr

Abstract - **Due to the COVID-19 pandemic universities had to adopt distance learning in just a few days. The paper aims to determine whether distance learning approach resulted in different outcomes for the students who participated in a mandatory course Social Entrepreneurship in comparison to the outcomes achieved in the previous academic year in traditional classroom learning environment. The course is focused on strengthening social entrepreneurial intention, social entrepreneurship desirability and social entrepreneurship feasibility. The results show that there are no statistically significant differences in these variables at the end of the course between the student groups who had different modes of learning. Regardless of the learning mode, the course did not strengthen social entrepreneurship desirability, nor social entrepreneurial intention. However, it did strengthen knowledge on financing opportunities for social enterprises. What is more, the results show that the students who participated in the distance learning course perceived the implementation of the course Social Entrepreneurship as better and their level of engagement during the course as higher in comparison to the students who participated in the classroom.**

Keywords - social entrepreneurial intention; social entrepreneurship education; longitudinal; on-line learning; e-learning; distance learning

I. INTRODUCTION

Social entrepreneurial intention is an attractive research topic. However, social entrepreneurial intention models that include the impact of education are rarely empirically tested [1]. Further on, longitudinal studies [2] and studies on distance learning in entrepreneurship education are rare [3], while in social entrepreneurship education they are almost non-existent.

Firstly, the paper examines whether social entrepreneurship course (SEC) increases social entrepreneurial intention (SEI) of the students and its antecedents: social entrepreneurship desirability (SED) and social entrepreneurship feasibility (SEF). Secondly, the paper analyses whether the SEC implementation mode (in the classroom or distance learning) changes the impact of the SEC on SEI and its antecedents. Finally, the research examines whether the students' perception of the course implementation, its outcomes and students' engagement during the course differs due to the SEC implementation mode (in the classroom or distance learning).

The second part of the paper presents a literature review on social entrepreneurial intention, social entrepreneurship education and distance learning, and develops the belonging hypotheses. The third part of the paper explains the applied methodology including data collection, measures, experiment and methods. Chapter four contains the detailed review of the obtained results while the fifth chapter provides a discussion of results and concludes the paper.

II. SOCIAL ENTREPRENEURIAL INTENTION AND SOCIAL ENTREPRENEURSHIP EDUCATION

Social enterprises are devoted to solving social problems and their purpose is to make a positive social impact, not to maximize return to owners [4]. Social entrepreneurs go through a three-component process [5]: they identify a stable, but unjust equilibrium that causes the exclusion, marginalization, or suffering of a group of people which on its own does not have financial means or political clout to achieve the transformation; they create a social value in this equilibrium; they create a stable ecosystem around this new equilibrium that alleviates the suffering of the targeted group of people. Some authors emphasize that social entrepreneurship results in systemic, large-scale transformation [6]. Social entrepreneurs usually use market forces to achieve sustainability of the new equilibrium [7]. In the European Union social enterprises provide social services and work integration for vulnerable groups of people [8].

Social entrepreneurial intention is a self-acknowledged conviction by a person that they intend to set up a social enterprise in order to create a social change through innovation [9], [10]. In order to ensure the understanding of the term social enterprise, in this study social entrepreneurial intention is defined as the intention to found and own business that would solve social problems, have participative decision making and reinvest its profits for the realization of some social mission. According to the theory of entrepreneurial event [11] the antecedents of entrepreneurial intention are entrepreneurship desirability and entrepreneurship feasibility. Following that line of thought, the antecedents of social entrepreneurial intention [12] are social entrepreneurship desirability and social entrepreneurship feasibility. Social entrepreneurship desirability refers to the personal attractiveness of starting a business that would solve social problems, while social entrepreneurship feasibility refers to a degree to which one perceives herself capable of starting a business that would solve social problems [11], [12].

Social entrepreneurship education aims at strengthening both social entrepreneurship desirability and social entrepreneurship feasibility hoping to enhance social entrepreneurial intention of students. Previous research found a positive relationships between participation in entrepreneurship training and social entrepreneurial intention [13]. There is a statistically significant positive association between the "know what" component of social entrepreneurship education and both the desirability and the feasibility of social entrepreneurship, while the "know how" component of social entrepreneurship education is statistically significantly positively associated with the feasibility of social entrepreneurship [14]. However, the research about distance learning in social entrepreneurship education is rare.

E-learning refers to all computer and Internet-based activities that support teaching and learning, both on-campus and at distance; it also includes administrative and academic uses of information and communication technologies supporting teaching and learning [15]. E-learning can be incorporated in a course with different levels of intensity [16], [17]: as a classroom aid, in a blended learning mode or in a fully online course. As a classroom aid e-learning supports traditional learning within a classroom setting, while in a blended learning mode the time spend in face-to-face learning is combined with distributed learning systems [18]. A fully online course is a distance learning course in which the lectures and all other activities are performed online without any classroom time.

Faculty of Organization and Informatics had previous experience with the development of an online study program in which the courses were performed 50% online and 50% in the classroom [19]. However, the COVID-19 pandemic demanded all of the study programs to switch to the distance learning mode in just a few days.

The research followed the progress of students enrolled in Social Entrepreneurship course (SEC) in two academic years: 2018/2019 and 2019/2020. In the academic year 2018/2019 the SEC was implemented in a classroom setting, while in 2019/2020 the SEC was implemented in a distance learning mode. The first hypothesis tests whether the two analyzed groups of students statistically significantly differ in SED, SEF and SEI at the beginning of the SEC. Since the two groups belong to the two consecutive academic years and the criteria for enrolling the course were the same for both groups, the following hypothesis with its belonging subhypotheses is posed:

H1: There is no statistically significant difference in SED (H1a), SEF (H1b) and SEI (H1c) at the beginning of the SEC between the two groups.

Since the meta-analyses demonstrated that there is no clear decline in educational effectiveness when using distance education technology [20], nor in the level of student satisfaction [21]. the following hypotheses are posed.

H2: There is no statistically significant difference in SED (H2a), SEF (H2b) and SEI (H2c) at the end of the SEC between the two groups.

H3: In the both groups SED (H3a), SEF (H3b) and SEI (H3c) at the end of the SEC increased in comparison to their levels at the beginning of the SEC.

H4: The students' perception of SEC implementation (H4a), SEC outcomes (H4b) and their level of engagement (H4c) does not differ between the two groups.

III. METHODOLOGY

The data were collected in the summer semester of two academic years 2018/2019 and 2019/2020, always at the beginning of the course Social Entrepreneurship as well as at the end of the lectures, prior to determining the students' final grades. In the academic year 2018/2019 the course was implemented in the classroom setting. In the academic year 2019/2020 due to the COVID-19 pandemic the majority of the course, except the first two weeks, was implemented as distance learning. The same is true for the questionnaire implementation. In the academic year 2018/2019 the participants filled out the paper version of the questionnaire in the classroom, while in the academic year 2019/2020 the students filled out the on-line version of the questionnaire. The pre-test and post-test questionnaires were administered under the same conditions in the each analyzed academic year.

The participation in the study was completely voluntary and anonymous. The anonymity was ensured by the usage of codes chosen by students who were asked to write down the same code on the first and on the second questionnaire. To decrease the risk of losing one's code, in the academic year 2018/2019 the students were asked to write a code on their first survey and to write it again on a piece of paper that they packed in the envelopes. Then they wrote their full name on the envelopes and signed the envelopes on the part where the envelopes are sealed. The reason for that was to ensure that no one would open the envelopes before the post-test questionnaire administration. Before administering the post-test questionnaire, the researchers gave to the students the envelopes so they could remind themselves of the code which then they could write on the second survey. This procedure enabled pairing pre-test and post-test questionnaires. The researchers decided to use this method because in the previous similar research, when the students were asked to keep the code with themselves, most of the students could not remember the code for completion of the post-test questionnaire. In the academic year 2019/2020 the students were asked to write a code in their first survey and to take a photograph of the code with their smart phones and to keep the photograph until the end of semester.

A. Sample

The population in the academic year 2018/2019 consisted of 81 third-year students of the undergraduate study program Economics of Entrepreneurship at the Faculty of Organization and Informatics, University of Zagreb. After excluding the incomplete questionnaires, the sample consisted of 68 respondents (84%). Participants (N=68) were on average 21,38 years old (SD

=0.847) at the beginning of the course and 88.2% of the students (n=60) were female.

In the academic year 2019/2020 73 students were enrolled in the course Social Entrepreneurship. After excluding the incomplete questionnaires, the sample consisted of 67 respondents (92%). Participants (N=67) were on average 21.73 years old (SD = 1.274) at the beginning of the course and 86.6% of the students (n=58) were female reflecting the sex structure of the third-year students' population of the study program Economics of Entrepreneurship.

B. Measures

The questionnaire collected data about social entrepreneurial intention and the perception of the course Social Entrepreneurship. The statements used in the research were rated on a five-point Likert scale with the range from 1 – I strongly disagree to 5 – I strongly agree. The used statements are available in Appendix 1.

Previous research showed that Croatian students often confuse social entrepreneurship with corporate social responsibility [22]. Since the term social entrepreneurship may be perceived differently by the different respondents it was important to use the measurement scales that measure social entrepreneurial intention, social entrepreneurship desirability and social entrepreneurship feasibility without using the term social entrepreneurship. Thus, in order to measure social entrepreneurship desirability, social entrepreneurship feasibility and social entrepreneurial intention measurement scales which do not use the term social entrepreneurship in their statements were used [14], [23]. Instead, the statements describe the main characteristics of social entrepreneurship emphasizing the provision of a solution to a social problem at the expense of profit maximization, participative decision-making and reinvestment of the profit into social mission. Setting a timeframe is important when measuring entrepreneurial intention [24]. Therefore, in all statements measuring SEI a timeframe of 10 years is present.

Social entrepreneurship desirability was measured using the statements about the advantages and pleasure that starting a company that solves social problems would bring to the student and whether such a company represents an attractive professional choice [25]. Social entrepreneurship feasibility was measured as a degree to which one perceives herself capable of starting a business that would solve social problems including easiness of starting a business that solves a social problem, gathering a team of people and raising the necessary finances [25].

After the completion of Social Entrepreneurship course the students were also asked about their perception of the course's different aspects. They had to evaluate SEC implementation, SEC outcomes and their own level of engagement.

C. Experiment

The course Social Entrepreneurship is a one-semester course and it is mandatory for the study program Economics of Entrepreneurship at the Faculty of Organization and Informatics. Since it is mandatory, it is especially adequate for longitudinal research since the students who enter differ in terms of their interest for solving social problems. The course applies a variety of teaching methods. Table 1. presents the teaching methods that were used in the academic years 2018/2019 and 2019/2020. Since the summer semester of the academic year 2019/2020 was implemented as distance learning, some of the exercises that were implemented the previous year were not performed in the same capacity. However, some new tasks for the students were added. The majority of the tasks and the course's content was the same in both examined years and the differences were not substantial. In the academic year 2019/2020 the lectures were in the form of synchronous learning which means that the lecturers gave live presentations with a direct feedback from the students.

TABLE I. TEACHING METHODS USED IN THE SOCIAL ENTREPRENEURSHIP COURSE IN THE ACADEMIC YEARS 2018/2019 AND 2019/2020

Teaching method	2018/2019	2019/2020
Lectures	YES, all students	YES, all students
Reading the materials prepared by the lecturer and answering in writing the questions based on the materials	YES, all students	YES, all students
Volunteering or interviewing a social entrepreneur	YES, all students	YES, a minority of the students
Telling stories about activists	YES, some students	YES, some students
Assuming the roles of real social entrepreneurs and answering the questions posed by the moderators in Social Entrepreneurship Conference organized during the class	YES, some students	YES, some students
Organizing Committee for the Social Entrepreneurship Conference organized during the class (organization of coffee breaks, flyers for the event and music)	YES, some students	NO
Watching a documentary on Mondragon Corporation and answering the appropriate questions	YES, all students	YES, all students
Tracking one's own social impact for a week	YES, some students	YES, some students

D. Methods

The variables SED, SEF, SEI, SEC implementation, SEC outcomes and students' level of engagement were calculated as arithmetic means of their belonging statements. To test the hypotheses independent samples t-test and paired samples t-test were used. Independent samples t-test was used to compare means of different variables between the students from the two academic years. Paired samples t-test was used to compare means of different variables at the beginning and at the end of the SEC for the same group of students. The tests were performed at the 5% level of significance.

IV. RESULTS

Reliability analysis results for SED, SEF and SEI are presented in Table 2. Cronbach's alpha is satisfactory for SED and SEI, but it is not adequate for the SEF construct. Thus, the analysis of SEF was done on its separate components.

TABLE II. CRONBACH'S ALPHA FOR SED, SEF AND SEI

Factor	Beginning of the SEC	End of the SEC
SED	0.797	0.802
SEF	0.536	0.481
SEI	0.780	0.770

The results of the independent-samples t-test show that there is no statistically significant difference (p=0.886) between the groups from two academic years in social entrepreneurship desirability at the beginning of the SEC. Both groups perceive social entrepreneurship as slightly desirable ($\bar{x}_1$ =3.68, SD_1=0.74, $\bar{x}_2$ =3.69, SD_2=0.88), confirming the hypothesis H1a. Further on, there is a statistically significant difference (p=0.886) between the two groups in one component of social entrepreneurship feasibility at the beginning of the SEC. That is the ability to gather a team of capable people to start a company that would solve some particular social problem ($\bar{x}_1$=3.37, SD_1=0.83, $\bar{x}_2$=3.70, SD_2=0.97). There is no statistically significant difference in the remaining components of SEF. Both groups are indecisive (p=0.535) whether they agree or disagree that it would be easy for them to start and manage a company that solves some particular social problem ($\bar{x}_1$=2.74, SD_1=0.73, $\bar{x}_2$=2.82, SD_2=0.87). The similar results are obtained (p=0.433) for the statement about knowing all the possibilities for financing a company whose mission is to solve social problems ($\bar{x}_1$ =2.76, SD_1=0.92, $\bar{x}_2$ =2.64, SD_2=0.90). Thus, the hypothesis H1b is partially confirmed. The results also show that there is no statistically significant difference (p=0.433) between the groups from two academic years in social entrepreneurial intention at the beginning of the SEC, confirming the hypothesis H1c. The students from both groups on average do not agree nor they disagree that they have intention to enter social entrepreneurship in the next ten years ($\bar{x}_1$=2.91, SD_1=0.76, $\bar{x}_2$=2.79, SD_2=0.94).

The results of the independent-samples t-test show that there is no statistically significant difference (p=0.498) between the groups from two different academic years in social entrepreneurship desirability at the end of the SEC. Both groups perceive social entrepreneurship as slightly desirable ($\bar{x}_1$ =3.72, SD_1=0.79, $\bar{x}_2$ =3.62, SD_2=0.89) confirming the hypothesis H2a. There is no statistically significant difference in the components of SEF. Both groups on average do not agree, nor they disagree (p=0.963) whether they would be able to gather a team of capable people to start a company that would solve some particular social problem ($\bar{x}_1$=3.46, SD_1=0.95, $\bar{x}_2$=3.45, SD_2=1.10). Both groups are also indecisive (p=0.786) whether they agree or disagree that it would be easy for them to start and manage a company that solves some particular social problem ($\bar{x}_1$=2.72, SD_1=0.83, $\bar{x}_2$=2.76, SD_2=0.91). Regarding the statement about knowing all the possibilities for financing a company whose mission is to solve social problems there is no statistically significant difference between the two groups (p=0.245). Both groups on average mostly agree with this statement ($\bar{x}_1$=3.99, SD_1=0.76, $\bar{x}_2$=4.14, SD_2=0.78). Thus, the hypothesis H2b is confirmed. The results also show that there is no statistically significant difference (p=0.234) between the groups from two different academic years in social entrepreneurial intention at the end of the SEC. The students from both groups on average do not agree nor they disagree that they have intention to enter social entrepreneurship in the next ten years ($\bar{x}_1$ =2.88, SD_1=0.82, $\bar{x}_2$=2.71, SD_2=0.91). Thus, the hypothesis H2c is confirmed.

The results of paired-samples t-test for the academic year 2018/2019 show that there is no statistically significant difference (p=0.601) between SED at the end of the course ($\bar{x}_2$ =3.72, SD_2=0.79) and SED at the beginning of the course ($\bar{x}_1$=3.68, SD_1=0.74). The results for the academic year 2019/2020 show the same. In that academic year there is no statistically significant difference (p=0.601) between SED at the end of the course ($\bar{x}_2$=3.69, SD_2=0.88) and SED at the beginning of the course ($\bar{x}_1$=3.62, SD_1=0.89). The hypothesis H3a is rejected.

Regarding SEF components, in the academic year 2018/2019 there is no statistically significant difference (p=0.484) between the ability to gather a capable group of people to start a business that would solve a particular social problem at the end of the SEC ($\bar{x}_2$=3.46, SD_2=0.95) and the same ability at the beginning of the SEC ($\bar{x}_1$ =3.37, SD_1=0.83). The same was found for the academic year 2019/2020 (p=0.052). The ability to gather a capable group of people to start a business that would solve a particular social problem at the end of the SEC ($\bar{x}_2$=3.45, SD_2=1.06) was not higher in comparison to its level at the beginning of the SEC ($\bar{x}_1$=3.70, SD_1=0.97). In the same manner, there was no statistically significant difference in easiness of starting and managing a company that solves some particular social problem both in the academic year 2018/2019 (p=0.890) and 2019/2020 (p=0.520). In the academic year 2018/2019 the students were on average indecisive whether they agree with the statement both at the end of the SEC ($\bar{x}_2$=2.72, SD_2=0.83) and at the beginning ($\bar{x}_1$ =2.74, SD_1=0.73). In the academic year 2019/2020 at the end of the course the students remained at the same level of indecisiveness ($\bar{x}_2$=2.76, SD_2=0.91) as they were at the beginning of the course ($\bar{x}_1$ =2.82, SD_1=0.87). The only component in which there was a statistically significant difference in both academic years was the knowledge about all the possibilities for financing a company whose mission is to solve social problems. In the academic year 2018/2019 the knowledge on financing was higher (p=0.000) at the end of the SEC ($\bar{x}_2$ =3.99, SD_2=0.76) than it was at the beginning of the SEC ($\bar{x}_1$ =2.76, SD_1=0.912). In the academic year 2019/2020 the knowledge on financing was also higher (p=0.000) at the end of the SEC ($\bar{x}_2$=4.14, SD_2=0.78) than it was at the beginning of the SEC ($\bar{x}_1$=2.68, SD_1=0.91). The hypothesis H3b is partially confirmed for one aspect of SEF and that is the knowledge on financing a company whose mission is to solve social problems.

The results of paired-samples t-test for the academic year 2018/2019 show that there is no statistically significant difference (p=0.864) between SEI at the end of the course ($\bar{x}_2$=2.89, SD_2=0.82) and SEI at the beginning of the course ($\bar{x}_1$=2.90, SD_1=0.76). The results for the academic year 2019/2020 show the same. There is no

statistically significant difference (p=0.309) between SEI at the end of the course ($\bar{x}_2$=2.70, SD_2=0.91) and SEI at the beginning of the course ($\bar{x}_1$=2.79, SD_1=0.94). The hypothesis H3c is thus rejected.

Reliability analysis results for SEC implementation, SEC outcomes and students' level of engagement are presented in Table 3. Cronbach's alpha is satisfactory for all presented constructs.

TABLE III. CRONBACH'S ALPHA FOR STUDENTS' PERCEPTION OF SEC COMPONENTS

Factor	Cronbach's alpha
SEC implementation	0.861
SEC outcomes	0.814
Students' level of engagement	0.712

The results of the independent-samples t-test show that there is a statistically significant difference (p=0.002) between the two groups in the students' perception of SEC implementation. Although both groups strongly agree that the SEC was implemented well, surprisingly, students who participated in the distance learning mode of SEC perceived SEC implementation as even better ($\bar{x}_2$=4.71, SD_2=0.35), in comparison to the students who participated in the SEC in the classroom ($\bar{x}_1$=4.46, SD_1=0.55). The hypothesis H4a is rejected.

There was no statistically significant difference between the groups (p=0.912) in their perception of the SEC outcomes confirming the hypothesis H4b. Both groups mostly agree that they have gained useful knowledge and skills ($\bar{x}_1$ =4.39, SD_1=0.59, $\bar{x}_2$ =4.39, SD_2=0.55).

There is however a statistically significant difference (p=0.008) between the two groups in students' perception of their level of engagement during the SEC. Although both groups mostly agree that their level of engagement was high, students who participated in the distance learning mode of SEC perceived their level of engagement as higher ($\bar{x}_2$=4.41, SD_2=0.59), in comparison to the students who participated in the SEC in the classroom setting ($\bar{x}_1$=4.11, SD_1=0.71). The hypothesis H4c is rejected.

V. CONCLUSION

The advantages of this study are that it is empirical and longitudinal which is rare in the research of social entrepreneurial intention. Also, it presents an example of distance learning in social entrepreneurship education.

Both groups of students perceived social entrepreneurship as slightly desirable at the beginning of the Social Entrepreneurship course. At the same time both groups cannot agree, nor disagree with the statements that describe the feasibility of social entrepreneurship. Further on, the students from both groups on average do not agree nor they disagree that they have intention to enter social entrepreneurship in the next ten years.

Regardless of the teaching mode, the Social Entrepreneurship course did not strengthen social entrepreneurship desirability, nor social entrepreneurial intention. However, it did strengthen knowledge on financing opportunities for companies whose mission is to solve social problems, which could contribute to social entrepreneurship feasibility.

Both groups of students strongly agree that the SEC was implemented well. Moreover, both groups mostly agree that they have gained useful knowledge and skills and that their level of engagement was high. However, the students who participated in the distance-learning course perceived the implementation of the course Social Entrepreneurship as better and their level of engagement during the course as higher in comparison to the students who participated in the classroom.

Three main questions arise from this research. Firstly, how can social entrepreneurship education adapt in order to strengthen social entrepreneurship desirability and social entrepreneurship feasibility and social entrepreneurial intention? Secondly, what is the impact of social entrepreneurship education in comparison to other factors of influence such as the existing development level of social entrepreneurship in a given country? Thirdly, do we measure the wrong constructs when evaluating the effects of a single mandatory course? Should we measure the possible changes in empathy, self-efficacy and general outlook on business? Maybe the goal of strengthening social entrepreneurial intention with a single mandatory course is simply excessive and not necessary. Further research should work on providing the answers to these questions.

Limitation of the research is that it is based on the comparison between the students from two academic years. Thus, the results should be interpreted with caution and future research should test the obtained results on a larger sample.

REFERENCES

[1] V. Ahuja, A. Akhtar, and O. P. Wali, "Development of a comprehensive model of social entrepreneurial intention formation using a quality tool," *J Glob Entrepr Res*, vol. 9, no. 1, p. 41, Dec. 2019, doi: 10.1186/s40497-019-0164-4.

[2] R. S. Shinnar, D. K. Hsu, and B. C. Powell, "Self-efficacy, entrepreneurial intentions, and gender: Assessing the impact of entrepreneurship education longitudinally," *The International Journal of Management Education*, vol. 12, no. 3, pp. 561–570, Nov. 2014, doi: 10.1016/j.ijme.2014.09.005.

[3] M. Al-Atabi and J. DeBoer, "Teaching entrepreneurship using Massive Open Online Course (MOOC)," *Technovation*, vol. 34, no. 4, pp. 261–264, Apr. 2014, doi: 10.1016/j.technovation.2014.01.006.

[4] P. L. Cochran, "The evolution of corporate social responsibility," *Business Horizons*, vol. 50, no. 6, pp. 449–454, 2007.

[5] R. L. Martin and S. Osberg, "Social entrepreneurship: the case for definition," *Stanford social innovation review*, vol. 5, no. 2, pp. 27–39, 2007.

[6] SAID Business School, "Skoll Centre for Social Entrepreneurship," 2013. http://www.sbs.ox.ac.uk/ideas-impact/skoll/about-skoll-centre-social-entrepreneurship.

[7] J. Mair and I. Martí, "Social entrepreneurship research: A source of explanation, prediction, and delight," *Journal of World Business*, vol. 41, no. 1, pp. 36–44, Feb. 2006, doi: 10.1016/j.jwb.2005.09.002.

[8] European Commission, "Communication from the Commission to the European Parliament, the Council, The European Economic and Social Committee and the Committee of the regions, Social Business Initiative Creating a favourable climate for social enterprises, key stakeholders in th," 2011.

[9] L. C. Prieto, S. T. A. Phipps, and T. L. Friedrich, "Social entrepreneur development: An integration of critical pedagogy, the theory of planned behavior and the ACS model," *Academy of Entrepreneurship Journal*, vol. 18, no. 2, p. 17, 2012.

[10] A. T. P. Tran and H. Von Korflesch, "A conceptual model of social entrepreneurial intention based on the social cognitive career theory," *Asia Pacific Journal of Innovation and Entrepreneurship*, vol. 10, no. 1, pp. 17–38, Jan. 2016, doi: 10.1108/APJIE-12-2016-007.

[11] A. Shapero and L. Sokol, "The social dimensions of entrepreneurship," in *Encyclopedia of entrepreneurship*, C. Kent, D. Sexton, and C. Vesper, Eds. Englewood Cliffs: Prentice-Hall, 1982, pp. 72–90.

[12] J. Mair and E. Noboa, "Social Entrepreneurship: How Intentions to Create a Social Venture are Formed," in *Social Entrepreneurship*, J. Mair, J. Robinson, and K. Hockerts, Eds. New York, USA: Palgrave Macmillan, 2006, pp. 121–136.

[13] A. Chinchilla and M. Garcia, "Social Entrepreneurship Intention: Mindfulness Towards a Duality of Objectives," *Humanist Manag J*, vol. 1, no. 2, pp. 205–214, Apr. 2017, doi: 10.1007/s41463-016-0013-3.

[14] I. Kedmenec, M. Rebernik, and P. Tominc, "Social Entrepreneurship Education and Its Association with Perceived Desirability and Feasibility of Social Entrepreneurship among Business Students," *Croatian Journal of Education : Hrvatski časopis za odgoj i obrazovanje*, vol. 18, no. 4, pp. 1035–1065, Dec. 2016, doi: 10.15516/cje.v18i4.1774.

[15] A. Bates and A. Sangra, *Managing technology in higher education: Strategies for transforming teaching and learning*. USA: John Wiley & Sons, 2011.

[16] T. Bates and G. Poole, *Effective teaching with technology in higher education: foundations for success*. Jossey-Bass, 2003.

[17] A.-M. Nortvig, A. K. Petersen, and S. H. Balle, "A Literature Review of the Factors Influencing E-Learning and Blended Learning in Relation to Learning Outcome, Student Satisfaction and Engagement," vol. 16, no. 1, p. 10, 2018.

[18] R. M. Bernard, E. Borokhovski, R. F. Schmid, R. M. Tamim, and P. C. Abrami, "A meta-analysis of blended learning and technology use in higher education: from the general to the applied," *J Comput High Educ*, vol. 26, no. 1, pp. 87–122, Apr. 2014, doi: 10.1007/s12528-013-9077-3.

[19] N. Kadoić and N. Begičević Ređep, "Methodology for online study program development - case study ePITUP," in *Proeceedings of 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2015, pp. 948–953.

[20] M. Allen, E. Mabry, M. Mattrey, J. Bourhis, S. Titsworth, and N. Burrell, "Evaluating the Effectiveness of Distance Learning: A Comparison Using Meta-Analysis," *Journal of Communication*, vol. 54, no. 3, pp. 402–420, 2004, doi: 10.1111/j.1460-2466.2004.tb02636.x.

[21] M. Allen, J. Bourhis, N. Burrell, and E. Mabry, "Comparing Student Satisfaction With Distance Education to Traditional Classrooms in Higher Education: A Meta-Analysis," *American Journal of Distance Education*, vol. 16, no. 2, pp. 83–97, Jun. 2002, doi: 10.1207/S15389286AJDE1602_3.

[22] J. Perić and A. Delić, "Social entrepreneurship in Croatia: Do regional disparities influence young people's perception of social entrepreneurship as a potential career path?," *Ekonomski vjesnik : Review of Contemporary Entrepreneurship, Business, and Economic Issues*, vol. XXVII, no. 1, pp. 81–92, Oct. 2014.

[23] I. Kedmenec, M. Rebernik, and P. Tominc, "Development of Social Entrepreneurial Intention Model," in *Proceedings of the 6th Global Conference on Managing in Recovering Markets, GCMRM 2015*, Maribor, Jul. 2015, pp. 373–382, Accessed: Nov. 02, 2019. [Online]. Available: https://www.bib.irb.hr/770211.

[24] N. Krueger, "Entrepreneurial intentions are dead: Long live entrepreneurial intentions," in *Understanding the Entrepreneurial Mind*, A. L. Carsrud and M. Brännback, Eds. New York, NY: Springer New York, 2009.

[25] I. Kedmenec, "The Impact of Individual, Social and Cultural Factors on the Social Entrepreneurial Intentions of Business Students," PhD thesis, University of Maribor, Maribor, 2015.

APPENDIX 1.

Social entrepreneurship desirability (SED)
1. It would bring me great pleasure to start a company whose mission would be to help the solving of social problems.
2. Starting a company that solves some social problem represents an attractive professional choice.
3. To own a company that emphasizes its social mission has more advantages than disadvantages for me.
Social entrepreneurship feasibility (SEF)
1. I would be able to gather a team of capable people if I decided to start a company that would solve some particular social problem.
2. It would be easy for me to start and manage a company that solves some particular social problem.
3. I know all the possibilities for financing a company whose mission is to solve social problems.
Social entrepreneurial intention (SEI)
1. Within the next 10 years, I would like to start a company whose primary goal would be the solving of some social problem that is particularly important to me, instead of profit maximization.
2. Within the next ten years, I am going to start a company in which decisions are made by considering the needs of all those involved (the employees, local community, clients and others), even if that entails decreased profits.
3. Within the next ten years, I intend to start a company that would invest most of its profits into solving a certain social problem of a particular interest to me (such as poverty, environmental pollution etc.).
Students' perception of the SEC implementation
1. Lecturers were available for direct communication.
2. The lecturers explained the material and key concepts in a clear and understandable way.
3. I received meaningful feedback on my progress / work.
4. The lecturers were well organized and prepared for lectures / teaching activities.
5. Lecturers encouraged discussion and critical thinking.
6. The lecturers were ready to answer the students' inquiries.
7. Lectures, teaching activities and assignments were in line with the objectives of the course.
Students' perception of SEC outcomes
1. Course performance encouraged collaboration and teamwork.
2. The course encouraged active student engagement.
3. Attending the course helped develop my interest in the topic of social entrepreneurship.
4. Through this course I have acquired useful knowledge and skills.
5. I think that I would be able to practically apply the knowledge acquired in the course.
6. Activities and tasks helped me understand key concepts.
Students' perception of their level of engagement during the SEC
1. I fulfilled the agreed tasks on time.
2. I regularly prepared for classes and teaching activities.
3. I gained a deeper insight into the topic through my own engagement and commitment.

Digital transformation of monitoring customer behaviour in the cars sales

Ivan Radoš*, Miljenko Hajnić*, Ivona Radoš*
*Algebra University College, Zagreb, Croatia
ivan.rados@racunarstvo.hr, hajnic.m@gmail.com, ilakic123@gmail.com

Abstract - **Monitoring customers' behaviour to improve the quality of services and goods is the daily routine of any successful company. Quality of firm services can be improved with effective planning, smart targeting potential customers and reducing operational expenses to the adequate level. Digitalization of available data and analysing those data can bring new information and insights that are invisible on a first sight. This paper brings results of the quantitative research based on collected data from car buyers and car dealers in the Republic of Croatia during 5 years period. By correlating variables of the vehicle price, customer's gender, distance between customer's location, vehicle length and place where the vehicle was bought we provided several answers that any marketing department may use for car sales market researches and preparing customers targeting. Knowing customers, their behaviour and needs makes it easier for firms to predict future trends. Since the data itself was provided by one large car sales company, the paper will focus on the car sales. In this paper, we analysed available data using computer programs and concluded that the relations between correlating variables give useful information for all participant in the car sales market and that information can lead to reducing firm expenses on car sales marketing campaigns, reaching more potential customers and increasing firm income.**

Keywords – ***customer behaviour, digitalization, predictions, car sales, marketing***

I. INTRODUCTION

Most business today base their prediction of future trends on many different parameters. The key to finding the right values within data is the definition of customer behaviour and identifying key factors in user profiling itself. Shopping online is a rising worldwide trend. Most companies now days use a variation of tools to track their clients and to satisfy the legal norm in the processes that require legal administration. Firms are generating a huge amount of data daily, and it is hard to make an analysis based on scattered data.

This paper aims to try to answer several questions that may help the firm's marketing department in the future targeting of potential customers. Proper customer targeting may decrease marketing budget, may expand the business beyond the firm's current boundaries and may lead to saving customer's time in the car buying process. If a customer spends less time on visiting dealers and on choosing the desired car, the firm may expect a higher amount of customer's money spent on the car purchase. To achieve needed answers, we quantitatively analysed the previously collected data from car dealers and put variables of vehicle length, customer's gender, car price and distance between customer and dealer's location in different correlations.

The focus of the conducted research was the interpretation of correlating data results in the available data scope: customer's address, customer's gender, dealer's location - the point of sale and price of a purchased car. The goal of this research is to find what customer's, dealer's and purchase's data can be used for smart targeting of a future potential customer and provide the base for the future researches in the field of car sales business processes digitalisation.

In the first part of the article, the focus will be on the known customer behaviours within the car sales. In the cited section of the article, the focus will also be on the familiar behaviour of the user when choosing a product and everything that precedes its selection. The central part of the paper presents the results of the research. The results show that the impact of distance from the point of sale has a significant effect. The difference in influence is also visible through gender elaboration. In the concluding part of the paper, the critical research results are summarised, which companies dealing with motor vehicle transfer can use to improve the sales of their companies by designing improved customer profiling.

II. LITERATURE REVIEW

Car as a product for the road transportation of goods and people became one of the necessities for living and business. Cars are segmented into two categories, for private use and business use. In recent time, cars are more and more treated almost like family members. As a result of this fact, the customers' perception of cars is changing as well as customers' decisions while buying cars become more complex. The buying decision is a process of evaluation of available attributes and selection of the most desirable options of the product. The customers' buying decisions depend on various causes, motives and influences.

Customers' decision styles vary [8]. Following literature review brings the relevant and latest works in the field of customers decision processes related to the buying car.

Customers' behaviour and decisions are highly influenced by several factors. For example, customers' educational level, age, income level, taste and many other factors [2]. Four variables, brand image [10], brand trust [1], product quality and price [3], have a positive and

significant influence on customers' decision when buying cars.

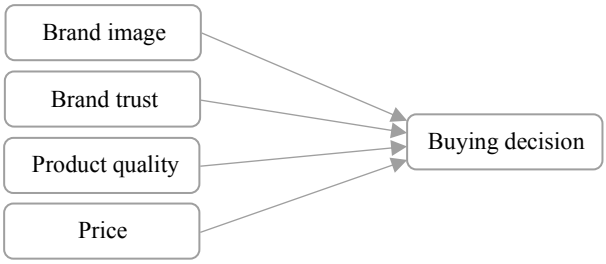


Chart 1. Model of Buying Decision [2]

According to the study [2], the price has the greatest impact on the customers final decision, compared to the other variables resulted in consumer's decisions.

The intention of the customer in the decision process when buying a car is to find all cars' attributes that will fulfil customers' needs. To achieve this, the customer must evaluate all attributes of each option, compare and process that information, and then choose the best or most wanted available option [5]. One of the interesting results of this study is that consumers rather affect each other than to reason individually.

When buying cars customers make decisions based on the well-explained information and they will evaluate car attributes in the decision-making process [7]. The evaluation process of car attributes could be a long and heavy task for customers. The study [7] says that, in general, consumer assigns different levels of attributes priority and the three most important attributes consumers take into account when evaluating car alternatives are the price, the operating cost and the car quality. After the evaluation of car attributes, the consumer makes a decision. Chart 2 describes a human's activities in the decision-making process.

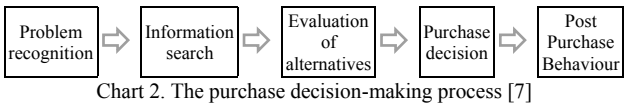


Chart 2. The purchase decision-making process [7]

The analysis [4] reports that new technology, more space and larger vehicle are the customers' top reasons when buying cars. Chart 3 shows the most important factors for making a decision which vehicle to buy where is evident that first place takes vehicle price.

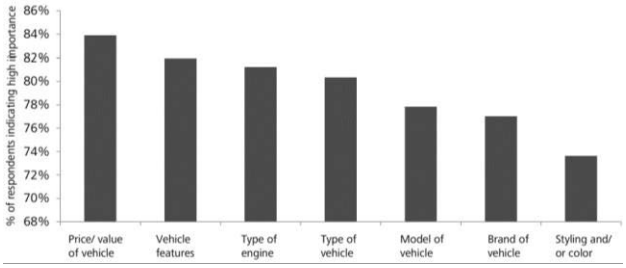


Chart 3. Impact of factors on the decision of which vehicle to purchase [4]

About 50% of respondents spend more than 10 hours on the evaluation of the car's attributes. Customers prefer to search the internet and read the car's reviews. This report also states that the internet is rarely used to contact the dealers during the customers' research phase.

III. THE RESEARCH METHODOLOGY

A. The Research Purpose /the Research Goals

The main purpose of this quantitative and descriptive research was to yield types of publicly available data and collected customers' data that can be used in the marketing sector for offline or online campaigns and targeting potential car buyers. It could be very helpful knowing, for example, that male car buyers are willing to spend more money on larger cars but in the same time, the male buyers are not willing to travel more than 20 kilometres from home to buy a car.

The goal of this paper is to provide the additional questions that marketing analysts may use in their future market research developing new selling strategies and planning efficient marketing campaigns.

Main research questions were defined as follows:

1. *Is there a relation between the price of the bought vehicle and the customer's distance to the dealer?*
2. *Can a prediction of customers' total amount of money for buying a vehicle be made based on the distance of customers from the selling location?*
3. *Does customer's gender influences the buying decision of a vehicle length?*

B. The Research Sample and Methods

In this study, we used real data collected in the period from 2014-01-22 until 2019-10-11 by a questionnaire developed for one of the largest car sales company in Croatia. The questionnaire was tested among a sample of 10826 customers. Collected data were anonymised before publishing them while keeping all necessary attributes for study purpose [6]. Information and knowledge about customers preferences and behaviour are important for the design of targeted sales campaigns.

This study tested relevant variables identified in previous research [9] on car purchase and change of ownership. Variables used in this research were the customer's address, customer's gender, dealer's location - the point of sale and price of a purchased car. Research was conducted on two groups of consumers: men make up 76.44% of the total number of customers, while women make up 23.56%.

Processing of data was done through the computer program Microsoft Visual Studio which uses programing language C# and relational database management system SQL.

Customers' residences addresses were converted into geographical coordinates (longitude and latitude) by reversible geocoding API provided by Google online services. In the same way, dealers' location where cars purchases were made, were transformed into decimal latitude and longitude values of geographical location.

For calculating distance between customers and dealers, we tried to use a formula for spherical law of cosines, which determines the great-circle distance between two points on a sphere given their geo-coordinates but that calculation wasn't appropriate for our research because time spend for passing the distance

between two points in urban areas is not even close to the time needed for passing equal distance in suburban areas. Instead of using reversible geocoding API from Google, we used other Google's API which calculates the distance between two addresses and which gives results measured in seconds needed for passing the distance using a car in general traffic conditions. For every distance calculation, we recorded both values of calculated distances, represented in kilometres and seconds, for later analysis.

Technical data (vehicle length) was obtained through historical data collected during the import of the vehicle's data into the database. The tools available in programing language Python were used for data visualisation and visual interpretation.

The subject analysis also presents gender relations according to the distance between buyer and seller, in terms:

> ***Personal vehicle*** belongs to a category whose maximum permissible mass does not exceed 3,500 kg and which are designed and constructed for the carriage of not more than 8 passengers, not counting the driver's seat. Personal vehicles may be combined with a trailer with a maximum authorised mass not exceeding 750 kg.
>
> ***OLS*** Ordinary least squares method; OLS (Ordinary Least Squares) statistical method for making a prediction model using linear regression.

C. *The Research Results*

The results of the research are presented according to the research questions, as follows:

1. *The relation between the price of the bought vehicle and the customer's distance to the dealer*

Quantitative analysis yielded a result of correlating variables customers' gender and distance between customers' and car dealers' locations. A detailed view of the average length travelled from the buyer's house to the dealer's point of sale is shown in Table 1.

Gender	Distance (km)
Male	82.40
Female	54.65

Table 2. The average distance from home address to seller location in km, *N = 10.826*

Further analysis of the data showed that, in general, buy more expensive vehicles (almost 19% more expensive than female respondents). The detailed view can be seen in Table 3

Gender	Average price
Male	103.985.40 kn
Female	87.641.06 kn

Table 3. The average price within gender, *N = 10.826*

By correlating results from Table 2 and Table 3 we can see that average female customer spend around 27% more money per kilometre of the distance between customers home and car dealer, as shown in Table 4.

Gender	Amount per km
Male	1.261,95
Female	1.603,67

Table 4. Distance price ratio in kn, *N = 10.826*

The probability functions of the price variable, and the distance in seconds, are approximately in uniform distribution frames.

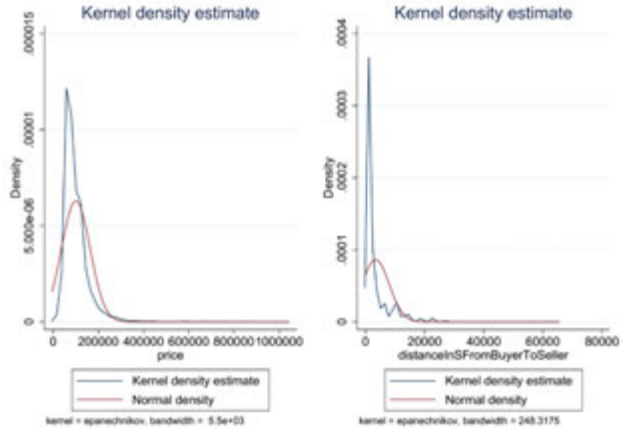


Chart 4. *Probability functions of the price variable, and the distance in seconds; N = 10.826.*

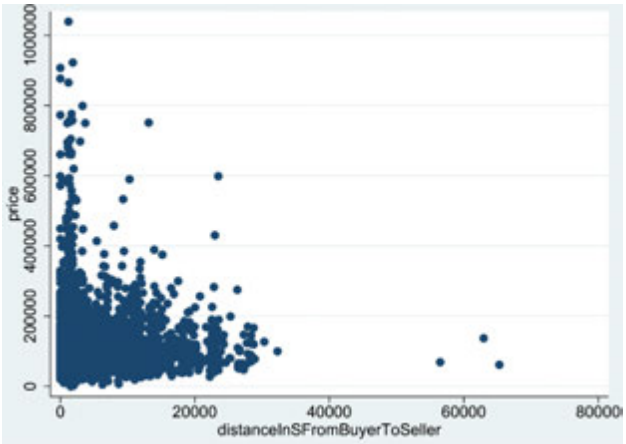


Chart 5. *Distribution of price variable, and the distance in seconds; N = 10.826*

2. *Prediction based on the relation between car price and distance of customers from the selling locations*

Estimation of a simple linear regression model by an OLS estimator, where the price is a dependent variable and distance from the customer in seconds is an independent variable (for each year separately, time-slice data). The following tables (Table 5 and Table 6) show the OLS elaboration by years. _cons variable represents value of the constant in the regression model. *N variable* represents the number of observations in the sample. *R* variable represents the coefficient of determination in the sample while *adj.*R^2 corrected coefficient of determination in the sample

year	2014	2015	2016	2017	2018
distance	1.109**	0.689*	0.849*	0.724	0.856*
_cons	88391.3***	90360.0***	95322.2***	102585.1***	108404.4***

	(1961.0)	(1396.8)	(1665.2)	(1845.0)	(2128.6)
N	436	811	1733	1833	1456
R^2	0.029	0.007	0.005	0.002	0.003
adj. R2	0.027	0.006	0.005	0.002	0.003

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, ***, $p < 0.001$

Table 5. *Linear regression model by an OLS estimator; N = 10.826*

In Table 6 variable *N* represents the number of observations in the sample. R^2 represents the coefficient of determination in the sample while $adj.R^2$ corrected coefficient of determination in the sample. Variable *male* represents the total number of included male customers in in the sample. Female customers are not shown here because gender is binary grouped, and a boolen value is used to denote gender (men are true, women are false).

year	2014	2015	2016	2017	2018
distance	1.088**	0.636*	0.843*	0.708	0.808*
	(0.341)	(0.300)	(0.391)	(0.370)	(0.405)
male	7837.1	8202.5**	16254,2***	15485,3***	18427,0***
	(4306.1)	(2622.9)	(2484.3)	(3065.7)	(3387.3)
_cons	82230.8***	84181.1***	83291.7***	91615.0***	95213.9***
	(3931.9)	(2250.5)	(2259.0)	(2787.7)	(2851.4)
N	436	811	1733	1833	1456
R^2	0.037	0.016	0.023	0.014	0.019
adj. R2	0.032	0.014	0.022	0.013	0.017

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6. *Elaboration table by coefficients; N = 10.826*

In both cases (Table 5 and Table 6), the coefficient of the variable distance is significant and positive, ie with an increase in distance by one second, the price increases (from 0.63 to 1.10 kunas; depending on the age). The coefficient of the male variable is always significant and positive and indicates that male customers on average buy more expensive cars than women (they use the opportunity to buy cars more online, about 80% of the sample are men, there are slight variations over the years).

3. The relation between customers' gender and vehicles' length

Through the analysis, the total range of vehicles' length was defined, and the shown in Table 7. *Longest vehicle* represents the longest total length of the sold vehicle measured in millimetres while the *Shortest vehicle* represents the shortest total length of sold vehicle measured in millimetres.

Longest vehicle	7340 mm
Shortest vehicle	2695 mm

Table 7. Range of vehicles purchased in the observed period, *N = 10.826*

Observing the data on the length of the vehicle, the analysis yielded values of vehicles' length in correlation with customers' genders. We read the following average values in Table 8.

Vehicle bought by	Average vehicle length
Male customer	4380 mm
Female customer	4210 mm

Table 8. Gender in correlation with the average length of the vehicle, *N = 10.826*

IV. CONCLUSION

Many studies have been conducted to explore the most important attributes related to customers buying decision and highly likely more studies will include all those questions in the future. Consequently, the results of these studies suggest many new research questions. Consumers pre-buy behaviour is the aggregation of all decisions that the consumer makes before buying a car. In real life, customers evaluate cars' attributes differently and their perception, behaviour and experience have different effect on purchase decisions. From the study results, many companies may find interesting how male and female customers respond to various car purchase factors.

During this research, several valuable insights were gained. The observed period and geolocation of the research (Croatia) showed that there is indeed a significant correlation between included variables: customers' and dealers' locations, customers' genders, vehicles prices and vehicles' length. Gained results from this research can be used by analysts for car selling market researches and targeting potential new car buyers.

We concluded that male customers generally travelled greater distances, almost 50% more than female customers, to purchase a motor vehicle. On the sample of 10,826 cars sold research showed that male customers, on average, spent a significantly higher amount to buy a car. By correlating results, we can see that the average female customer spent around 27% more money per kilometre of the distance between customers home and car dealer.

Observed data of the gender and length of the vehicle, yielded a conclusion that male customers bought longer vehicles than female customers. Difference of 170 millimetres of vehicle's length indicates that future smart targeting and preparing offers should be different for male and female customers. Combining that insight with result that male customers spent more money on bought cars, car selling market analysts can prepare appropriate offers to potential car buyers, rationalize marketing expenses of a firm and get more customers by more suitable offers.

The results of this research can be also used to predict the future behaviour of existing customers, which through advertising, marketing and other sales activities will influence the increase of customers who are currently less represented.

REFERENCES

[1] S.F.S. Alwi, B. Nguyen, T.C. Melewar, Y.H. Loh, M. Liu, "Explicating industrial brand equity: Integrating brand trust, brand performance and industrial brand image", 2016, Industrial Management & Data Systems, Vol. 116 No. 5, pp. 858-882, DOI: 10.1108/IMDS-09-2015-0364

[2] A. Amron, "The Influence of Brand Image, Brand Trust, Product Quality, and Price on the Consumer's Buying Decision of MPV Cars", 2018, European Scientific Journal, Vol.14, No.13, DOI: 10.19044/esj.2018.v14n13p228

[3] J. Beneke, & R. Flynn, T. Greig, M. Mukaiwa, "The influence of perceived product quality, relative price and risk on customer value and willingness to buy: A study of private label merchandise", 2013, Journal of Product and Brand Management, 22, DOI:10.1108/JPBM-02-2013-0262

[4] Deloitte Touche Tohmatsu Limited (DTTL), "Driving through the consumer's mind: Steps in the buying process", 2014, Available at: https://www2.deloitte.com/content/dam/Deloitte/in/Documents/ma

nufacturing/in-mfg-dtcm-steps-in-the-buying-process-noexp.pdf [accessed: 26.05.2020.]

[5] T. Hansen, "Perspectives on consumer decision making: An integrated approach. Journal of Consumer Behaviour", 2005, 4 (6):420–437, DOI: 10.1002/cb.33

[6] M. Jändel, Decision support for releasing anonymised data, Elsevier, Computers & security 46 (2014) 48-61

[7] T. Laurence, C. Machari, (2008), "Consumer Behaviour for Purchasing Cars Task 1.4. Clever Clean Vehicle Research ", http://www.belspo.be/belspo/ssd/science/Reports/CLEVER%20WP1.4%20Consumer%20behaviour%20for%20purchasing%20cars.pdf [accessed: 26.05.2020.]

[8] B. Mittal, "Facing the Shelf: Four Consumer Decision-making Styles", 2017, Journal of International Consumer Marketing, DOI: 10.1080/08961530.2017.1318732

[9] I. Radoš, T. Babić, Digital transformation as a process of using digital technologies for monitoring and designing the user experience, 2020

[10] J. Sasmita, N.M. Suki, "Young consumers' insights on brand equity", 2015, International Journal of Retail & Distribution Management, 43, 276-292, DOI: 10.1108/IJRDM-02-2014-0024.

“Financial” Aspects of Spotify Streaming Model

Joško Lozić*, Goran Vojković* i Marin Milković*
*Sveučilište Sjever, Varaždin, Croatia
josko.lozic@unin.hr
goran.vojkovic@unin.hr
marin.milkovic@unin.hr

***Abstract* - The aim of this paper is to analyze the business model and financial results of the Spotify music streaming platform. The digitalization of production processes has directly affected the development of new media industries. The old media industries developed in the model of mass media and limited media markets. New media industries are associated with technological progress and the digitalization of production processes. The digitalization and convergence of production and distribution processes has directly affected the complete paradigm shift in the development of new media industries. The music industry was among the first to enter the process of transformation. The production and distribution of physical sound carriers has been replaced by streaming music platforms. The purchase of physical sound carriers in "brick and mortar" stores has been replaced by subscriptions on the streaming platform. Spotify is the largest global streaming music platform. It generates the highest revenue and has the largest number of subscribers. The strength and impact of the Spotify platform on the global market is evidenced by the emergence of the term “spotification”, which is used for all similar business development models.**

Key words - music industry, Spotify, streaming industry

I. INTRODUCTION

Spotify is an audio streaming platform for broadcasting music. It is organized as a platform to access the Internet and choose how to use the service. There are two basic ways of using: a) an approach that allows a limited time of using the service accompanied by advertisements and b) a subscription agreement that allows personalization of the service without advertising content. Personalization of the service or subscription relationship means creating your own music lists with songs that the user wants to listen to, recommendations to other users as well as receiving recommendations from users with whom he communicates. In practice, „streaming“ service – refers not simply to a distribution technology, but also to a technology of surveillance (as it allows precise monitoring of individual media consumption), and to a business model on subscriptions [1]. The Spotify application enables monitoring of podcasts with notifications and limited communication between users and subscribers. Although it has some characteristics similar to radio stations from the "old" media industry, such as a podcast or a music number on demand. Spotify is organized in a platform economy model and is fully oriented to the high technology market and the "new" media industry. The organization of production in the platform model is divided into two basic forms. The platform can be organized as a moderate product development process or a moderate ecosystem building process [2].

Production-oriented platforms are based on engineering design, the production of physical products, and the use of economies of scale and economies of scale [3]. The ecosystem perspective portrays platforms as the hubs of wider business systems that facilitate, coordinate, and control exchanges between multiple organizations [4]. There was s shit in the early twenty-first century to the personal computers and mobile digital playback devices such as Apple iPod as the prevalent ways of consuming music [5]. Ever since this service was launched in 2008, Spotify has also frequently served as a metaphor for a model that is supposedly is not limited to music, but replicable in a wide range of media industries and cultural forms [1]. By 2010, it already had 10 million users. The first headquarters were registered in Luxembourg, but business operations were managed from Stockholm, Sweden. Around year 2010-11 a shift was clear in the Swedish discourse about new business models. The analogical meaning of Spotify drifted away from the advertising-based model, towards the subscription-based model [1].

The application is adapted for all operating systems and includes a wide range of content browsing capabilities. Music numbers can be searched by author, song, album, record label and other. The user application meets the standards of iOS, Android, Linux, and is used equally successfully on MacOS and Windows computers. In addition, the adaptation of the application to all types of smartphones and PDAs, or to all operating systems they use, has directly affected the global growth and popularity of the corporation. In the fourth quarter of 2019, Spotify had 248 million visitors within a period of 30 days or one month. Spotify categorizes with digital service platforms without a proprietary device that, as its core service, mediates service exchanges exchange between commercial content producers and end-customers [6].

II. LITERATURE REVIEW

This digitalization of music has evolved through a series of transformative phases; from vinyl’s to CDs, from CDs to mp3s and from mp3s to streaming services. Each such phase has involved a rearranging of the socio-technical infrastructure of music, affording new modes of production and consumption, similar to previous shifts in music technology [7]. Simon (2019) gives an overview of the development of technology within the music industry, ranging from the development of radios, through the

industry of vinyl's, sound recording tapes, CDs, computer memory discs to modern streaming services. Up to 2000 the music industry was focused on physical sound carriers. Sales of sound carriers grew until 1984, culminating in 2006 [8]. Each of the stages of development introduced new changes in social relations and the way of consuming musical content.

With the development of smartphones, MP3 audio players, tablets, laptops and other technologically advanced devices, users have easy access to music and other media content. More importantly, users are given access while on the move which was a significant step forward in the development of the music industry [9]. Digital music provides the ability to overcome some of the limitations of physical recordings by extending their length and expanding access to recordings. However, what was once a social / cultural aspect of listening to music has diminished by allowing access to large databases of already recorded music [10]. The development of music technology has led to an increase in mobility and accessibility and has gradually given consumers increased control over music content. Many of the changes associated with the digitization of music precede digitization in other industrial sectors [11] [12]. Listening to music has turned from mainly a stand-alone activity into everyday activities [13]. Digitized music and streaming services allow you to listen to music during free activities outside the home such as running [14]. Soundtracking involves choosing and listening to music to accompany various everyday practices such as commuting, working, hanging out and driving [7].

The development of streaming services like Spotify, Deezer and Apple Music has also meant that consumers can access large catalogues of music on the go [15]. The streaming model was first used in big-data processing of data broadcast in real time such as stock market data, weather forecasts, sports results, etc. [16]. Streaming services are different from these traditional products. The song that is broadcast via the service is not permanent, because access ends after the subscription expires [17]. Jean Baudrillard distinguish collection from the inferior activity of accumulating, noting that objects have two possibilities: they can be utilized or they can be possessed [18]. Streaming has eluded the boundary between music promotion and music consumption, producing "a new relation between exposure and sale, united within one and the same service" [19]. Playlists have become a central form of music consumption on streaming platforms. By signing in to Spotify, you will immediately be confronted with a playlist wall, tailored to your listening history, time of day, and other specific characters [20] [21].

Music streaming services afford people the ability to create and name their own playlists, or take pre-existing playlists crated by professionals play list editors and add, edit or remove tracks [22]. The reality is that dematerialization of music archives into streaming services shapes, but also is shaped by, the transnational drive for new markets by media companies like Spotify, Pandora, and Netflix to deliver catalogue content to bigger consumer markets [23]. In the mature post-disruptive supply chain, where the social peer-to-peer sharing aspect of Spotify are seemingly downplayed in favor of commercial playlist and customized playlist based on Spotify user data, the major's direct control of what music the consumer is exposed to actually increases compared to the pre-disruptive regime [24]. Music streaming services are technologically modern corporations that operate on one of the platform economy models. They use and develop software applications, but they cannot be classified in the software industry [25] [26]. Music streaming services, such as Spotify and Apple Music, have potential to disrupt the social dynamics of music consumption. For free or in exchange for a subscription fee, music streaming services facilitate anytime, anywhere access to vast catalogues of licensed music at little or no cost using an Internet-enabled device, such as a smartphone [27]. Music streaming services such as Spotify and Apple Music, have gone from a niche to the dominant mode of music distribution [28].

Spotify encourages its users to sign up with a Facebook profile to make it easier to connect with users with whom they share the same musical tastes [12] [26] [29]. By placing music numbers on existing servers and distributing them on the Internet, they use a form of cost model known as zero marginal cost [30] [31]. The zero marginal cost model, in its full expression, is applied to organizations that distribute a product from their own production as digital content. The zero marginal cost model could be applied to the Netflix platform, in the part where the products for which Netflix is the producer are distributed. Spotify uses a similar model although it pays rights to license holders to perform music numbers. We define Spotify as a digital service or platform that mediates the service between end providers and customers, something similar to Netflix or Hulu [6]. Streaming technology and combining content on-line and off-line models represent the third wave of digital disruption [32]. Sun refers to Pollack (2011) [33] and emphasizes how Spotify was acclaimed as "exactly what music fans had been waiting for, fulfilling the long-sought dream of a 'Celestial Jukebox' – a service that makes every song always available, freely and legally" [34]. Spotify itself has usually not been on the frontline of innovation, but followed trends set by other actors in the music, entertainment and technology industries [1].

III. KEY FINDINGS

In 2008, Spotify started with work, and in just six years of existence, i.e. in 2014, it has already exceeded the total revenue of one billion Euros. The third phase started in March 2014 following the removal of the AppFinder. In contrast to the previous phase, the boundary resource configuration now sought to consolidate the service scope across external devices while simultaneously facilitate the generation of external subscriber interfaces [6]. Aileen Lee was the first to use the term unicorn to denote promising high-tech start-ups. Investors popularly called them "working horses" because they promised high revenue growth and above-average profits for investors [35]. Subsequently, the name was given an expanded meaning and "unicorn" marked a start-up project that in a very short time generated revenue or capitalization of one billion dollars. Common to all the corporations that were given that name was that they were, in some way, connected to Silicon Valley, or that, like Spotify, they were created on the basis of the use of high technology. Spotify is not a lonely Swedish unicorn, there are several Swedish

"unicorns" such as Skype, Klarna, Avito, Mojang, Apotea and Truecaller [23].

Table 1 Revenue and cost of streaming service Spotify (mill. Eur)

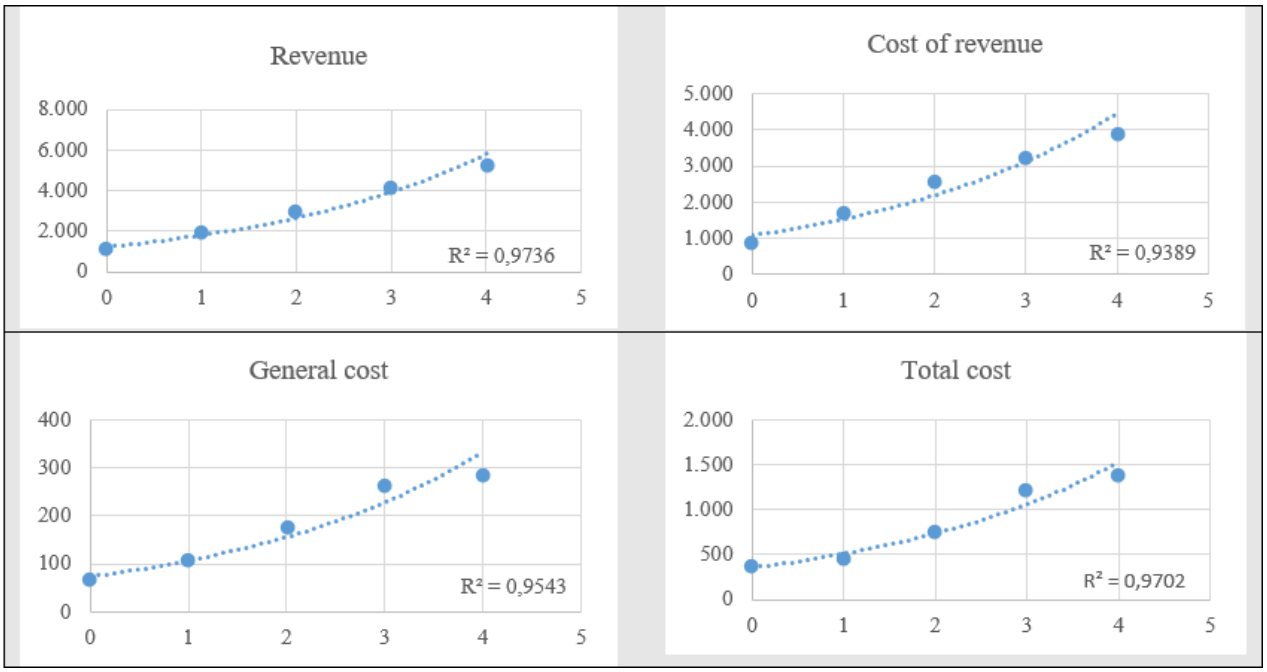
	2014.	*2015.*	*2016.*	*2017.*	*2018.*
Revenue	1.085	1.940	2.952	4.090	5.259
Cost of revenue	911	1.714	2.551	3.241	3.906
Gross profit	174	226	401	849	1.353
Research and development	114	136	207	396	493
Sales and marketing	184	219	368	567	620
General and administrative	67	106	175	264	283
Cost	365	461	750	1.227	1.396
Operative loss	**-191**	**-235**	**-349**	**-378**	**-43**

Source: own illustration

Table 1 provides an overview of the corporation's financial performance over the past five years. In the analyzed period, total revenue increased by almost four times or by 384.70%. Cost of revenue increased by 328.76%, which enabled the growth of gross profit by 677.59%. Direct operating expenses include all costs related to the acquisition and payment of copyright. The costs are paid according to the number of broadcasts, but also according to the average number of total published music content. There is constant controversy about how to pay for copyright and who owns the copyright if the music number is broadcast on a streaming service. Gross profit margin increased from 16.04% in 2014, to 25.73% in 2018.

Other operating costs in the analysis include general operating costs and finally, as the sum of all costs, total operating costs. General operating costs include the costs of managing and maintaining the music platform, i.e. all those costs that maintain the system, and are not directly related to the core business. Most of these costs include the salaries of management and employees of the corporation. In the analyzed period, general expenses increased by 422.30%, which was less than the growth of direct costs and less than the growth of total revenues. General expenses accounted for 6.18% of total revenue in 2014, before falling to 5.38% of total revenue in the last analyzed period. One of the factors in the successful operation of a corporation is reflected in the growth of total costs. Total costs, except for direct operating costs, increased by 382.70% in the analyzed period, which is far below the total revenue increase of 484.70%. Total operating expenses amounted to 33.64% of total revenue in the first analyzed period, to fall to a higher 26.64% in 2018.

The coefficients of determination of the four analyzed items from the Spotify account and profit are shown in Figure 1. All four coefficients are greater than 95% of the interpretation, and the total revenue coefficient has the highest interpretation coefficient and is 97.36%. Table 2 shows the exponential trend equations for the four analyzed items. The results of the research show an average increase in the total income of the corporation in the analyzed period of 47.74%. The average revenue growth, in the last five years, is five percentage points higher than the increase in



Picture 1 Coefficient of determination (Spotify)
Source: own ilustration

Table 2 Exponential trend equation

Revenue	Cost of revene	General cost	Total cost
$Y = 1219*1,4774^x$	$Y=1077,2*1,4260^x$	$Y=73,117*1,4614^x$	$Y=353,91*1,4422^x$

Source: own ilustration

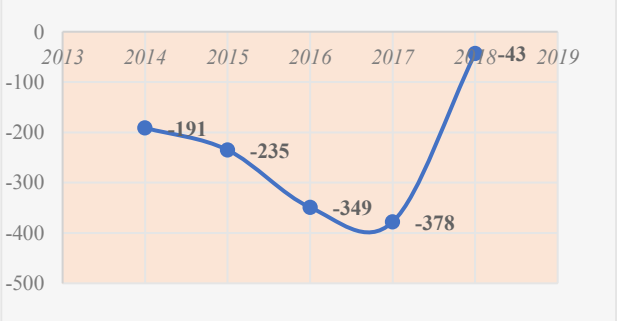
direct operating expenses. Direct operating expenses increased on average by 42.60% and account for the largest cost in total costs and are the most important indicator of the movement of the corporation's total operations. The average increase in general operating expenses was 46.14%, which is one percentage point less than the average increase in total revenues, but the general cost is a relatively small item compared to other types of costs and does not significantly affect the overall operating result.

The financial result is directly related to the number of service users. Spotify generates most of its revenue from the subscription relationship. In relative terms, subscriptions make up about 90% of revenue, while advertising revenue is only about 10%. In the last quarter analyzed, Spotify had 248 million users who visited using the app. As a monthly user, every visit of an unregistered visitor who views or listens to some content on the application in the time recorded in milliseconds is recorded. Users who pay a subscription and have personal accounts opened on the app are called premium users. The number of such users increased in the analyzed period of five years from 28 million to 113 million. Data on the increase in the number of users are shown in Table 3.

Table 3 Premium and monthly Spotify users (mill.)

Period	**Premium users**	**Mounthly users**
Q3 2015	28	82
Q3 2016	36	113
Q3 2017	62	149
Q3 2018	87	191
Q3 2019	113	248
	$Y=27,459*1,4438^x$	$Y=84,191*1,3150^x$
	$R^2=0,9842$	R2=0,9975

Source: own illustration



Picture 2 Gain/loss before tax (mill. Eur)

Source: own illustration

The reasons for the successful operation of the corporation and the increase in revenue of 384.70% in the last five years can be directly related to the increase in the number of premium customers. The number of premium users, i.e. users who pay some form of subscription, increased at an average rate of 44.38% per year. At the same time, the number of users who came to the app at least once a month without buying a subscription was increasing at an average rate of 31.50% per year. The average growth rate of regular users was higher than the growth rate of occasional users, which ensured revenue growth and stability of the platform. In cyberspace, people collect lists rather than objects, and those lists serve as a form of personal expression that derives from but also supersedes the record collection [36]. In comparing music-streaming practices with earlier ideals for music collection one immediately encounters the dilemma that digital formats and streaming services make it impossible to "collect" music as such because the format offers music thorough subscription rather than ownership [18].

Indicators of financial result and balance of income / expenditure before interest and taxes, recorded a negative trend from 2014 to 2017. Losses realized after the payment of direct and general expenses increased from 191 million Euros to 378 million Euros. In the last analyzed period, losses were reduced by 88%, and the total loss was $ 43 million. An increase in the number of subscribers of 44.38% and an increase in revenue at a rate of 47.74% indicate the end of the negative trend in the financial result. The growth phase, or star phase as it is called, is slowly coming to an end and the corporation is approaching a mature phase. However, Spotify should be careful with estimates of future financial results. Fleischer (2020) warns of the fundamental precarity of technological starts-up, including Spotify. The precarity results from their dependency on venture capitalist willing to cover the losses, as these start-up (with few expectations) are not profitable. In order to attract the capital, they need to sustain what is commonly known as hype or buzz [1] In a study of digital media, a relevant approach might therefore be to „follow the hype" [37].

IV. CONCLUSION

The music industry has changed its revenue model. Music streaming platforms generate most of the revenue of the music industry. Brick and mortar stores of physical sound carriers are very rare and serve only a few more vinyl collectors.

Future research in the music industry will focus on two fundamental directions. On the one hand, research will seek to determine the structure and level of revenue by categories of the music industry, and on the other hand, research will be focused on researching individual ecosystems of music streaming platforms. Long-term research in the field of music industry will focus on the analysis of revenues of global music corporations that will generate an increasing share of revenues from copyright and synchronization. As a special musical niche, the sale of vinyl and all forms of income associated with that niche will appear.

Spotify has grown into the largest global music streaming platform. It generates revenues of more than 5 billion euros and has more than 100 million premium users. In 2018, operating losses were EUR 43 million and decreased sharply compared to last year. The increase in revenue and number of customers gives hope that it will stabilize its business and become a profitable corporation.

REFERENCES

[1] Fleischer, R. (2020). "Universal Spotification? The shifting meanings of Spotify as a model for the media industries." Popular Communications. https://doi.org/10.1080/15405702.2020.1744607

[2] Skog, D.; Wilmelius, H.; Sandberg, J. (2018). „Digital Service Platform Evolution: How Spotify Leveraged Boundary Resources to Become a Global Leader in Music Streaming". Proceedings of the 51st Hawaii International Conference on System Sciences., pp.4564-4573.

[3] Thomas, L.D.W.; Autio, E; Gann, D.M. (2014). "Architectural Leverage: Putting Platforms in Context", Academy of Management Perspectives 28(2), pp. 198–19.

[4] Gawer, A. (2014). "Bridging differing perspectives on technological platforms: Toward an integrative framework", Research Policy 43(7), pp. 1239–1249.

[5] Hesmondhalgh, D.; Meier, L.M. (2018). „What the digitalisation of music tells us about capitalism, culture and the power of the information technology sector". Information, Communication & Society, 21:11, 1555-1570.

[6] Skog, D.; Wilmelius, H.; Sandberg, J. (2018). „Digital Service Platform Evolution: How Spotify Leveraged Boundary Resources to Become a Global Leader in Music Streaming". Proceedings of the 51st Hawaii International Conference on System Sciences., pp.4564-4573.

[7] Fuentes, C.; Hagberg, J.; Kjellberg, H. (2019). „Soundtracking: music listening practices in the digital age". European Journal of Marketing., pp. 483-503. ISSN: 0309-0566.

[8] Simon, J.P. (2019). „New players in the music industry: lifeboats or killer whales? The role of streaming platforms. Digital Policy, Regulation and Governance, pp. 525-549. DOI 10.1108/DPRG-06-2019-0041.

[9] Denegri-Knott, J. (2015), "MP3", Consumption Markets and Culture, Vol. 18 No. 5, pp. 397-401., doi.org/10.1080/10253866.2015.1048962.

[10] Arditi, D. (2017). "Digital Subscriptions: The Unending Consumption of Music in the Digital Era". Popular Music and Society. Volume 41, (3), pp. 302-318. doi.org/10.1080/03007766.2016.1264101.

[11] Dholakia, N., Reyes, I. and Bonoff, J. (2015), "Mobile media: from legato to staccato, isochronal consumptions capes", Consumption Markets and Culture, Vol. 18 No. 1, pp. 10-24., doi.org/10.1080/10253866.2014.899216.

[12] Irene, C.L.NG. (2014). Creating New Markets in the Digital Economy: Value and Worth. Cambridge University Press.

[13] Hagberg, J.; Kjellberg, H. (2017). "Digitalized music: entangling consumption practices", (Eds). Cochoy, F.; Hagberg, J.; Peterson McIntyre, M.; Sörum, N. Digitalizing Consumption: how devices shape consumer culture, Routledge, pp. 167-189.

[14] Kerrigan, F.; Larsen, G.; Hanratty, S.; Korta, K. (2014). "Gimme shelter': experiencing pleasurable escape through the musicalisation of running", Marketing Theory, Vol. 14 No. 2, pp. 147-166.

[15] Sinclair, G. and Tinson, J. (2017), "Psychological ownership and music streaming consumption", Journal of Business Research, Vol. 71, pp. 1-9.

[16] Maheshwari, A. (2019) Digital Transformation: Building Intelligent Enterprises. Wiley & Sons. ISBN 978-1-119-54083

[17] Hiller, S. R.; Walter, M.J. (2017). The rise of Streaming Music and Implication for Music Production. Rev Netw Econ; 16(4): 351–385. doi.org/10.1515/rne-2017-0064.

[18] Hagen, A.N. (2015). "The Playlist Experience: Personal Playlist in Music Streaming Services". Popular Music and Society, 38:5, 625-645.

[19] Kjus, Y. (2016). „Musical exploration via streaming services: The Norwegian experience". Popular Communication, 14, 127–136.

[20] Prey, R. (2019). Background by Design: Listening in the Age of Streaming, in Naxos Musicology International, 1:1., https://www.researchgate.net/publication/337195472.

[21] Moazed, A.; Johnson, N.L. (2016). Modern Monopolies: What It Takes to Dominate the 21ST-Century Economy. Apllico, LLC. ISBN 9781250091895.

[22] Webster, J. (2019). „Taste in the platform age: music streaming services and new forms of class distinction". Information, Communication & Society.

[23] Burkart, P.; Leijonhufvud, S. (2019). "The Spotification of public service media". The Information Society.35:4, 173-183.

[24] Kask, J.; Oberg, C. (2019). "Why majors surge in the post-disruptive recording industry". European Journal of Marketing. Vol.53, No.3, pp. 442-462.

[25] Lozić, J.; Milković, M.; Lozić, I. (2017). "Economics of platforms and changes in management paradigms: Transformation of production system from linear to circular model". Economic and Social Development 26th International Scientific Conference on Economic and Social Development – "Building Resilient Society". Zagreb, str. 125-136. ISSN 1849-7535.

[26] Parker, G.G.; Van Alisyne, M.W.; Choudary, S. P. (2016). Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You. W.W. Northon & Company. ISBN 978-0-393-24913-2

[27] Morris, J. W.; Powers, D. (2015). „Control, curation and musical experience in streaming music services". Creative Industries Journal, 8(2), 106–122.

[28] Webster, J. (2019). „Taste in the platform age: music streaming services and new forms of class distinction". Information, Communication & Society.

[29] Shilds, B. (2017) Social Media Management: Persuasion in Networked Culture. Oxford Universtiy Press. ISBN 978-0-19-029633-9.

[30] Rifkin, J. (2015). The Zero Marginal Cost: The Internet of Things, the Collaborative Commons, and the Eclipse of Capitalism. St. Martin Griffin. ISBN 978-1-137-28011-4.

[31] Lozić, J. (2019a) „Zero marginal cost in magazine industry: changing of cost paradigm in „new" magazine industry".44th International Scientific Conference on Economic and Social Development. Split., str. 125-136. ISSN 1849-7535,

[32] Evans, D.S.; Schemalensee, R. (2016). Matchmakers: The Economics of Multisided Platforms. Harvard Business Review Press. ISBN 978163391728.

[33] Pollack, N. (2011). „Spotify's Celestial Jukebox" [online]. Wired UK.

[34] Sun, H. (2019). Digital Revolution Tamed: The Case of the Recording Industry. Palgrave Macmillan. ISBN 978-3-319-93021-3.

[35] Lozić, J. (2019b). Menadžment ekonomije platformi. Sveučilište Sjever, Centar za digitalno nakladništvo. ISBN 978-953-7809-93-5.

[36] McCourt, T. (2005). "Collecting Music in the Digital Realm." Popular Music and Society 28.2: 249–252.

[37] Eriksson, M.; Fleischer, R.; Johansson, A.; Snickars, P.; Vonderau, P. (2018). Spotify teardown: Inside the black box of streaming music. Cambridge, Massachussets: MIT Press.

Value of Innovation Platforms in Agriculture

Martina Tomičić Furjan, Larisa Hrustek, Igor Pihir
Faculty of Organization and Informatics, Varaždin
University of Zagreb
martina.tomicic@foi.hr; larisa.hrustek@foi.hr; igor.pihir@foi.hr

***Abstract* – Innovation through modern digital technologies ensures the survival of businesses in today's ever-changing working environment. Therefore, it should be seen as a comprehensive process of development and application of new competitive digital business models, which will elaborate the impact of the proposed innovation to business success and progress. New ideas and innovations have to be managed, in order for them to achieve their main role as a valuable trigger or contribution to the business development. An Innovation Management Systems (IMS) can be designed as a platform within which all relevant stakeholders of the innovation process can operate in real-time and at the same operational level. In this paper the management of digital innovations in agriculture is presented, with an accent on problems, currently used technologies as well as innovation platforms as potential technology solutions for better management within this sector. Moreover, the value of innovation management systems for some identified stakeholders is analysed and described, giving a comprehensive view of benefits it brings to them.**

Key words - innovation; innovation platform; innovation management system; agriculture; digital technologies

I. INTRODUCTION

A general literature overview, seen in [1] on global developing trends showed that the traditional approach to structuring the global GDP is a basis for determining four current and potential future growth poles of the modern global economy:

- Agriculture, which future perspectives are connected to digital modernization of agricultural production
- Industry, which future perspectives are connected to formation of Industry 4.0
- Innovations (R&D), which future perspectives are connected to high tech production
- Service sphere, which future perspectives are connected to development of e-commerce

At the same time the level of digital technologies application and the adoption is not advancing at the same speed in all industries [2]. Moreover, some industries of the most important economic importance, such as agriculture, which is the industry this paper has in focus, are slowly adopting new concepts of work, and environmental, climatic, market and organizational challenges in this area are increasing. Due to its important role in the food chain, the transformation of agriculture is expected [3] and growth in the application of digital technologies is welcomed [4]. Researchers agree that digital innovation [4] and knowledge sharing can have a profound impact on the transformation of the agricultural sector, although this transformation is lagging behind, especially in local communities and on smaller agriculture businesses.

In Europe, continuous policy efforts are aimed at strengthening the digital market, in particular the digitization of the agricultural sector through the Cork 2.0 declaration "A better life in rural areas"; EU Commission proposal for Common Agricultural Policy beyond 2020; EU Communication "Digitizing Europe Industry"; or Digital Single Market (DSM) initiative [3].

The fundamental role of digital technologies is especially important in the innovation process, where the goal of agricultural businesses, driven by the wave of the Industry 4.0 paradigm, is to integrate digital technologies into business processes in agricultural production [3]. The European Commission, in the presented common agricultural policy after 2020 [5] defines policy goals such as increased competitiveness, balance of the food chain, landscape and biodiversity conservation, generational renewal, health and food protection, successful resistance to climate change and financial stability, which should be achieved. Due to the limitation of other resources, this is strongly dependable on the transformation of agricultural processes through the application of technologies and innovation.

Value creation in the agricultural sector is rather linked to other business segments and sectors than to individual farms [4], and the increasing complexity of farm business and interaction with other stakeholders requires digital adaptation and application of communication, selling and distribution platforms.

Analog to Industry 4.0., the agriculture is also undergoing a transformational cycle by redesigning internal and external collaboration and interaction across the whole value chain and processes [6].

II. DIGITAL INNOVATION IN AGRICULTURE

A. Technological innovation in the agriculture

The agricultural sector has rightly been identified as one of the most important industries as it ensures the survival of mankind. Food needs are growing, and it is certainly worth mentioning that in some parts of the world food needs are not being met. Global health disruptions also affect the understanding of the significance of

agriculture locally and regionally raising the pressure on producers to become more efficient and agile.

The problems and challenges in the agricultural sector are diverse in nature and complexity by themselves. One of the most significant challenges is that of an ecological and climatic nature, which relates to global warming and unpredictable weather conditions [2]. The problem of the occurrence of unknown animal and plant diseases and the occurrence of emergencies can disrupt any cycle of agricultural production or completely interrupt it. Another set of challenges is related to markets and demand volatility. Inadequate policies and strategies for agricultural management, changing quantities and prices in the agricultural production sector, inventory management of goods and other performance related issues cause financial and other risks and responses in this sector.

Difficulties in applying innovative technologies in agriculture arise because farms face complex decision-making problems, related to the challenges mentioned above. To optimize agricultural processes and operations, farmers must consider various input parameters that can result in sudden oscillations of the entire agricultural system [7].

In addition to the capacity and readiness of the agricultural enterprise, institutional infrastructure is one of the important factors in understanding innovation activity. In particular, an agricultural policy that affects farmers' income could affect the innovation process. A strong link between public support for innovation research and programs that support farm incomes has been identified. This could explain the investment in innovation in the agricultural sector [7].

Innovation through technologies should be one of the main factors changing the agricultural sector, and within the agricultural scenario, the classification of innovation can be divided into the following three categories: a) policy issues and strategies behind the creation and adoption of innovations; b) process innovation and product innovation; and c) the impact of innovation on economic operators and the market [7].

B. Use of digital technologies for innovations in the agriculture

The agricultural sector is characterized by poor adaptation and adoption of digital technologies and innovative tools, which leads to a gap between innovative tools on the market and their exploitation and application in agriculture [2]. The challenge of introducing new innovative ways of doing business parallel to daily workloads for SME's in agriculture minimises the opportunities of promising implementations. There are many factors which influence the readiness of agricultural SME's to go digital, but this seems to be a complex research area with a huge potential.

Organizational problems or challenges in the agricultural sector can be related internally to the capacity of farmers or agricultural enterprises and related to the knowledge and skills for adoption and application of digital technologies [7].

Digital technologies can be the basis for the development of the agricultural sector and increase its competitiveness in the market. Digital technologies should lead to solving challenges [8] related to environmental, climate, food and preservation sovereignty. Intelligent digital solutions help address such challenges [8] and transform the agricultural sector in terms of economic, social and environmental sustainability [3].

Innovation and use of digital technologies can increase agricultural production, reduce the consumption of resources, time and money, improve working conditions for farmers, preserve and improve the characteristics of agricultural conditions (soil), significantly reduce risks (erosion, disease) or increase resource availability (water) [9].

Robotics, sensors and Big Data analytics are reported as the three pillars of platform digital agriculture [3], and cyber-physical systems represent the basis for the development of innovative solutions for monitoring and managing processes in agricultural enterprises [10]. In addition, Internet of Things (IoT), artificial intelligence (AI) and Blockchain are also reported as being leading promising digital technologies in the entire value chain, from suppliers of input products to consumers [11]. Robotics and robotic systems have made a significant contribution to mastering routine and manual operations in agriculture, and the contribution is visible in the replacement of people in a number of non-professional occupations [8].

Large amounts of data are constantly generated in agriculture, and this technology can be widely used through the digital platforms of the agricultural complex [8], because otherwise, if they cannot be turned into useful tools to support farmers, they have little value [3]. The importance and role of these technologies is studied and explained in crop management systems, livestock, and water management [10].

Internet of Things (IoT) enables communication and transmission of information via the Internet directly between instruments like equipment, devices and goods [8]. Most common case known is monitoring of soil parameters and weather conditions to enable smart irrigation solutions, protection against pests and pesticide use [10]. Examples of pesticide use include unmanned aerial vehicles and ground vehicles with the goal of reducing risks thanks to real-time detection systems and agricultural tools for real-time precision handling, while drones and smart glasses enable reality tracking and tracking and sampling of processed soils to identify problems and enable management of possible outcomes [10].

All mentioned technological innovations are important for companies to strengthen their production processes and organizational structures by using automation and other similar technologies in business processes of the agricultural sector [3], while platforms and applications enable data collection, business monitoring, exchange of knowledge and experience and development of specific agricultural innovations, branches or activities.

III. INNOVATION PLATFORMS IN AGRICULTURE

Digital applications and platforms can potentially drastically change the way knowledge is processed, communicated, accessed and used [3], especially in the agricultural sector where the full transformation of business can be enabled through the application of digital technologies. Digital platforms for farmers provide decision-making opportunities that were not previously possible and simplify the conduct of business processes, leading to radical changes in farm management [3]. The importance of web platforms in agriculture is emphasized by different authors [2] who have developed a method for predicting future shortcomings in agriculture (machine failures), and predictive maintenance is provided through a web platform.

IMS in the agricultural sector are important for several reasons. The first reason is that the adoption of innovations, especially the ones related to digital technologies in agriculture is slow, except for the general application of new technologies in the automation of operations. Another reason is that the challenges in this sector are growing and decision-making is becoming more complex. The third reason is that the development of agricultural production has a perspective, which is confirmed by the fact that the needs for food are increasing, and the supply chain is becoming increasingly unstable. The digital transformation of the agricultural sector, with an emphasis on the application of digital technologies and the development of new ideas and innovations in agricultural processes and products, can contribute to the creation of a promising industrial branch, as a guide to the economy.

Innovation platforms in agriculture sector can be implemented to encourage the adoption of agricultural innovations and stakeholder interactions within a value chain [12] and help in their management. The European Commission has developed a comprehensive system of agricultural knowledge and innovation, abbreviated AKIS, which is designed as multiple process of generating, exchanging, and using knowledge in the field of agriculture that interacts between the heterogeneous stakeholders involved. The AKIS platform includes various stakeholders, the most important being farmers, advisors and researchers. The needs and interactions of stakeholders and access to digital agriculture differ significantly, which is influenced by various factors such as the type of activity, identified problems and others [3].

The use of platforms and multiple stakeholder engagements can result in improved interactions between platform members and their local environment, proactive participation in platform meetings, and a significant increase in farmers' knowledge about the challenges. The use of platforms can also contribute to enrichment of agricultural production and smart business. Also, innovation platforms can facilitate the collective understanding and use of ready-made options (climate-smart agriculture options) that provide an answer to potential problems [13].

Innovation platforms can be considered as means to develop and scale innovation, as a result of the idea creation of different stakeholders, and they can facilitate conflict resolution, negotiation, social learning and collective decision-making for coordinated action [12].

The key stakeholders we have identified in the innovation process are presented in table 1.

TABLE 1: STAKEHOLDERS OF IMS

Stakeholder	Role description
Idea creator	An idea creator is any participant in an innovation process who creates an idea or innovation related to a process, product, or performance of an activity. In an agricultural enterprise, it can be any employee of the enterprise who creates an idea or proposes an innovation.
Expert	Expert represents external knowledge, i.e. research groups, institutions or universities that through their scientific contribution, research and projects contribute to the development of the field and help solve problems or develop ideas. In an agricultural enterprise, these can be external partners with whom the enterprise cooperates and carries out research projects in the field of agriculture. One of the examples of such cooperation is the implementation of experiments in farming where the results of varieties of different seed houses are tested, and everything takes place within the cooperation of the Ministry of Agriculture, i.e. the Sector for Professional Participation in Agriculture, seed distribution companies and agriculture.
Owner	Owner is the party that identifies the problem and emphasizes other actors on the platform to present their ideas. It can be the owner of an agricultural enterprise or the manager of a certain segment of agricultural production or process. Given the complexity of the agricultural sector, problems can be identified in different business segments.
Member of experiment	A member of an experiment is a member of a team who works on a particular idea or innovation and participates in its development, which should ultimately result in the development of a prototype process model, product prototype, or a new business concept. These can be experts in the same field or teams that combine knowledge from different fields.

The value IMS can bring to each of the identified stakeholder is shown in the next chapter and can help in defining their benefits.

IV. VALUE OF IMS FOR STAKEHOLDERS

The value of IMS for identified stakeholders is presented through the Value proposition canvas (VPC) [14] in figures 1, 2, 3 and 4. VPC consists of 2 parts, The Customer and the Value part. The **Customer** side deals with customers' jobs - what they want to do; pains - what are the problems they are facing when doing their jobs; and gains - what they expect to get as a benefit. The **Value** side describes the product or service which is offered - in case of this paper the IMS; pain relievers - how they help in neutralizing pains; and gain creators - how they create wanted benefits.

IMS platform and innovation process enable the idea creator (Figure 1) to create and develop ideas. Idea creator

can be any participant who creates an idea for a specific problem of the agriculture business. In addition to creating the idea, the basic job of the idea creator is to participate in the development of innovations, work on patents and participate in the experiment of using the idea. IMS platform enables idea creators to identify existing ideas, exchange ideas with other stakeholders, and play with ideas, ie create ideas related to agricultural products. This is possible through integration with other stakeholders, involvement in the development of agricultural businesses through the platform and the application of innovative methods in the development of the idea. The problems that an idea creator may encounter on the platform are the dissemination of own ideas and non-recognition of other good ideas, and in order to reduce the problems, the categorization of ideas and ready-made forms for creating ideas will help.

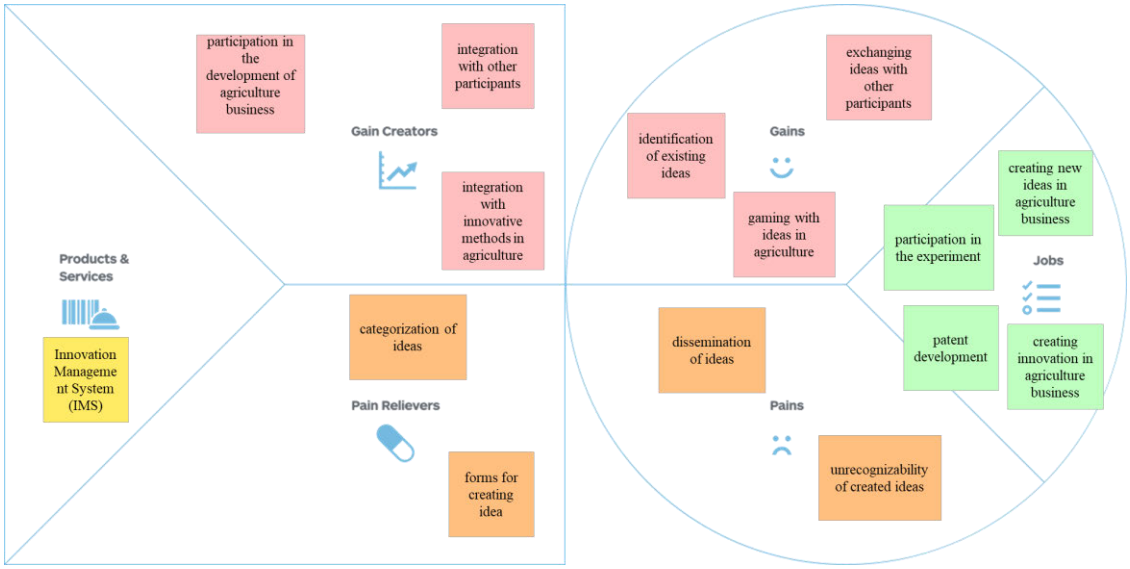


FIGURE 1. VPC FOR IDEA CREATOR

Experts (Figure 2) are external participants on the IMS platform and they are in charge of adding and creating knowledge units on the platform to make relevant knowledge units available to the farm, participating in creating solutions to specific farm problems and commenting and giving feedback on ideas created. The value that external experts achieve through engagement on the platform is contact and cooperation with private farms, which is the basis for further research in the agricultural sector, but also the contribution of science to the agricultural sector.

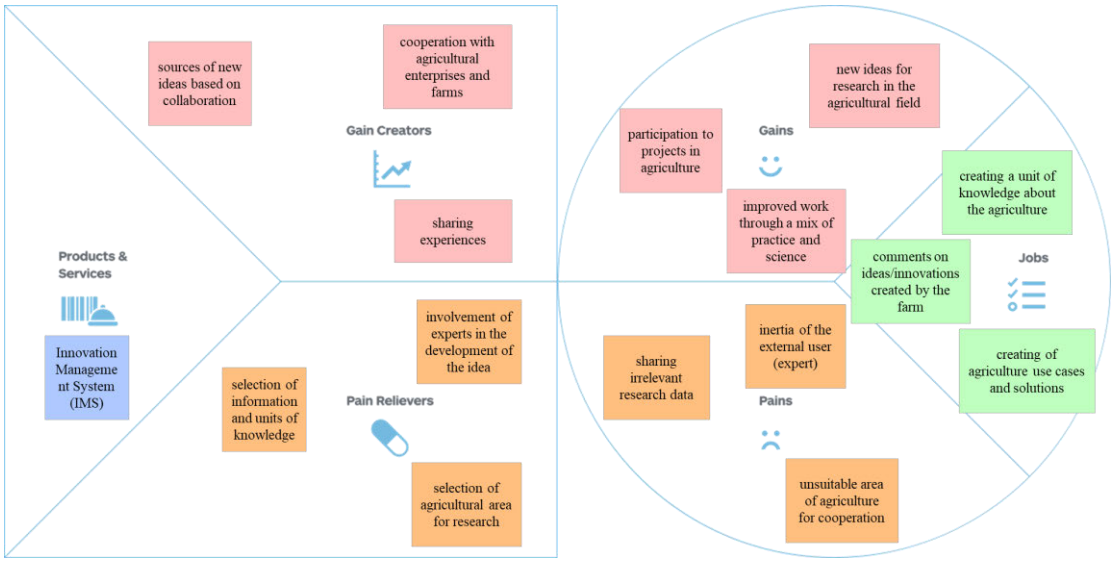


FIGURE 2. VPC FOR EXPERT

Owner (Figure 3) can be a business owner or a farm manager, who identifies the problems facing, initiates problem-solving activities in terms of generating ideas. The platform can allow the owner to launch initiatives to solve farm problems and manage ideas created by other stakeholders.

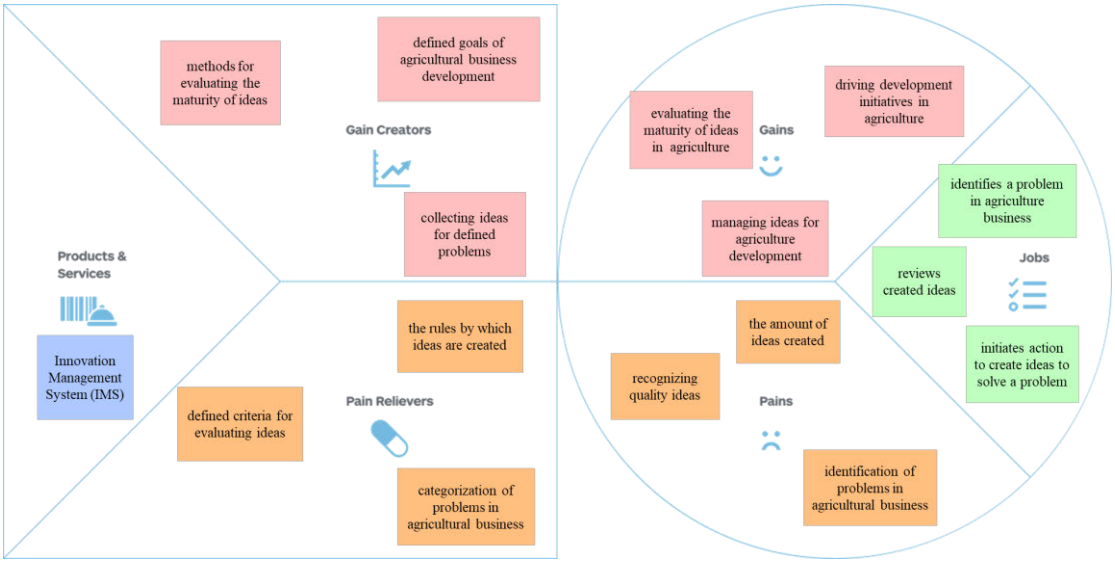


FIGURE 3. VPC FOR OWNER

A member of experiment (Figure 4) participates in the elaboration of the selected idea/innovation, by commenting on the idea and adding relevant content for the development of the idea. Through the platform, acquirement of new knowledge in collaboration with external experts, upgrade of ideas and contribution to the development of selected ideas for further elaboration is enabled.

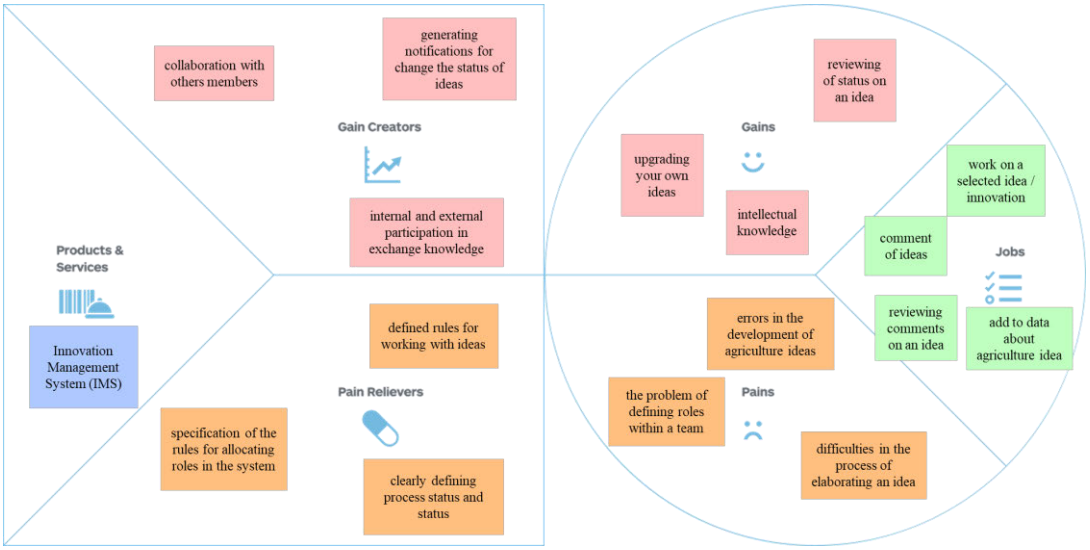


FIGURE 4. VPC FOR MEMBER OF EXPERIMENT

The application of digital technologies can offer many opportunities in agriculture and result in optimized agricultural processes, precise and smart agriculture, and an IMS platform can provide support to the innovation and development process. Collaboration and communication of stakeholders through IMS is one of the mechanisms for the exchange of knowledge and ideas leading to learning and innovation. IMS stakeholders are

provided with learning based on experience and examples of good practice and better results, such as faster decision-making, effective problem solving [11] and increased development potential. Also, the platform enables the creation of collaborative knowledge and the formalization and transfer of knowledge [3]. This form of cooperation encourages mass cooperation, socialization of knowledge and intensive use of information and knowledge, which facilitates and improves knowledge management in the agricultural enterprise. Establishing stakeholder cooperation through IMS is an important factor in understanding innovation activity and as such supports farmers' income and influences the innovation process. It also represents a strong link between support for innovation research and support programs for income of agriculture business [7].

V. CONCLUSION

The digital age brings great potentials of digital technologies usage and opens many opportunities for progress in diverse industries. In the agriculture, it brings new and raising expectations for improvements, but also, challenges are growing as well.

In this paper a short overview on how innovations related to digital technologies have penetrated in the agricultural sector has been shown.

Due to its characteristics of being mainly a traditional industry, highly dependent on environmental conditions, with relatively experienced human resources who dispose with a lot of domain knowledge and skills but lack the ones related to use of digital technologies, agriculture is relatively slow on adopting innovation ideas and implementing them. An innovation management system can help in identifying digital technologies that can bring benefits to the main actors and users and lead the implementation towards success. Digital platforms, as one of Industry 4.0. technologies have been shown as a possible technological solution, due to their ability to bring all stakeholders together and enable their collaboration and easy communication. This can lead to the development of new business models and new opportunities to increase economic value streams.

The value that a platform based IMS can contribute to all identified stakeholders is easily identifiable but is relatively hard to test. Our further research will therefore be directed into collecting more research data and evaluation of the presented value proposal on different case studies. That will for sure enrich and broaden our insights into the research field.

ACKNOWLEDGEMENT

This research has been conducted as part of the wider research in the project Competence Centre for Digital Transformation of the Food Industry in Rural Areas, which is funded by European Union through the European Regional Development Fund (ERDF).

REFERENCES

[1] A.V. Bogoviz, S.V. Shkodinsky, A. A. Skomoroshchenko, I. V. Mishchenko, and T. D. Malyutina, "Scenarios of Development of the Modern Global Economy with Various Growth Poles," Lecture Notes in Networks and Systems 73, 2020, pp. 185-192

[2] H. Lüttenberg, C. Bartelheimer, and D. Beverungen, "Designing Predictive Maintenance for Agricultural Machines," 26th European Conference on Information Systems: Beyond Digitization - Facets of Socio-Technical Change, ECIS 2018, pp. 1-17

[3] J. Ingram and D. Maye, "What Are the Implications of Digitalisation for Agricultural Knowledge?," Frontiers in Sustainable Food Systems, Vol. 4, No. 66, 2020, pp. 1-6

[4] J.M. Sagarna Garcia and D. Pereira Jerez, "Agro-food projects: analysis of procedures within digital revolution," International Journal of Managing Projects in Business,13(3), 2019, pp. 648-664

[5] CAP specific objective: Ensuring viable farm income (European Commission). Available: https://ec.europa.eu/info/sites/info/files/food-farming-fisheries/key_policies/documents/cap_specific_objectives_-_brief_1_-_ensuring_viable_farm_income.pdf

[6] I. Zambon, M. Cecchini, G. Egidi, M.G. Saporito, and A. Colantoni, "Revolution 4.0: Industry vs. agriculture in a future development for SMEs," in Processes, Vol. 7, Issue 1, No. 36, 2019, pp. 1-16

[7] G. Bucci, D. Bentivoglio, A. Finco, andM. Belletti, "Exploring the impact of innovation adoption in agriculture: How and where Precision Agriculture Technologies can be suitable for the Italian farm system?," IOP Conference Series: Earth and Environmental Science, Vol. 275, Issue 1, No. 012004, 2019, pp. 1-7

[8] O.V. Kozhevina, N.V. Salienko, E.A. Starozhuk, V.A. Klueva, I.V. Pavlova, "Digital Transformation of the Industrial Sector: Challenges and Prospects," Lecture Notes in Networks and Systems 129 LNNS, 2020, pp. 530-539

[9] B. Rivza, D. Vasilevska, and P. Rivza, „Impact of digital innovation on development of agriculture in Latvia," Engineering for Rural Development 18, 2019, pp. 682-687

[10] L. Trivelli, A. Apicella, F. Chiarello, R. Rana, G. Fantoni, A. Tarabella, „From precision agriculture to Industry 4.0: Unveiling technological connections in the agrifood sector," British Food Journal, Vol. 121, Issue 8, 2019, pp. 1730-1743

[11] A.M. Ciruela-Lorenzo, A.R. Del-Aguila-Obra, A. Padilla-Meléndez, and J.J. Plaza-Angulo, "Digitalization of agri-cooperatives in the smart agriculture context. Proposal of a digital diagnosis tool," Sustainability (Switzerland), Vol. 12, Issue 4, No. 1325, 2020, pp. 1-15

[12] D. Dabire, N. Andrieu, P. Djamen, K. Coulibaly, H. Posthumus, A.M. Diallo, M. Karambiri, J.M. Douzet, B. Triomphe, "Operationalizing an innovation platform approach for community-based participatory research on conservation agriculture in Burkina Faso," Experimental Agriculture, Vol. 53, Issue 3, pp. 460-479

[13] A. M. Osorio-García, L. Paz, F. Howland, L. A. Ortega, I. Acosta-Alba, L., Arenas, N. Chirinda, D. Martinez-Baron, O. Bonilla Findji, A. M. Loboguerrero, E. Chia & N., Andrieu, „Can an innovation platform support a local process of climate-smart agriculture implementation? A case study in Cauca, Colombia, Agroecology and Sustainable Food Systems," Agroecology and sustainable food systems, 2019, pp. 1-34

[14] A. Osterwalder, Y. Pigneur, G. Bernarda, and A. Smith, „Value proposition design," Hoboken, New Yersey: John Wiley &Sons, 2014.

Decision making on Digital Platforms in Agriculture

Nikola Kadoić*, Katarina Tomičić-Pupek* and Neven Vrček*
* University of Zagreb, Faculty of organization and informatics, Pavlinska 2, Varaždin
nkadoic@foi.unizg.hr, ktomicic@foi.unizg.hr, nvrcek@foi.unizg.hr

***Abstract* - There are many ways to sell and buy products in all fields, using digital technologies. That is the case for the agriculture field as well. Among many possibilities and different ICT solutions in agriculture, selection of the appropriate solution that will result in the satisfaction of both - customers and producers - becomes an important issue, and it is the main research topic of this paper. In this paper, we present the factors that influence the digital platform selection and propose two models for decision making on digital platforms. Decision models can help customers and producers to decide if a specific platform can be useful for their business. The decision model is based on a multi-criteria decision-making method called Dex. Dex is selected because it is very intuitive and less complicated than other multi-criteria methods. On the other hand, Dex can aggregate all crucial factors and include the specifics and subjectivity (perception) of each user (customer or producer). The paper also contains a demonstration of the implementations of two models on the hypothetical cases. In future research, we plan to evaluate the model in practice (and possibly update it if needed) and present it to the scientific and professional community. That can help producers/customers to optimize their business processes but also give directions for the development of new platform functionalities for platform developers.**

Keywords – agriculture, digital platform, decision making

I. INTRODUCTION

Digital platform as a term refers to an online real-time accessible digital space similar to a forum where many stakeholders meet in order to communicate, offer, buy, solve, exchange or share ideas, products, services, or other goods. Due to its primary task of facilitating the trade in a virtual environment trust and perception aspects influence the reach of a platform. Another important attribute of digital platforms is the usefulness that depends on expectations and delivered functionalities regardless of industry type the platform is intended for. In the agriculture sector, the variety of value chains, horizontal and vertical integration [1] and the diversity of touchpoints in performing business models creates a high demand and supply on digital platforms. The impact of digital technologies in the agriculture sector has been mostly reported by presenting case studies that incorporate digital technologies in agriculture [2]. Different stakeholders have different roles as platform users and in that sense, the decision on joining or adopting a digital platform becomes significantly important for surviving in a highly competitive environment. Decision making on joining or adopting digital platforms in agriculture is a complex process by itself that needs to be carried out by actors with usually little digital competencies. In order to ease the selection and decision process, we investigated factors that influence the perceived value or perceived risk and thereby impact the decision on joining and adopting digital platforms. This research was performed within a project which aims to achieve a digital transformation of the food industry in rural areas. The transformation efforts in the area of agriculture and food distribution include a widespread application of knowledge and innovation in the production of agriculture products and services. It is aimed at impacting the readiness of producers for challenges regarding economic growth and increased competitiveness of the rural and agricultural sectors. The project is a combination of a total of 8 research and development project activities whereby following three activities are dealing with: Research of models and development of a platform for the application of IoT in agriculture in rural areas, Software for decision support in managing the adaptability of business systems, and Distribution platform for food products according to the "Platform-based business" model. Within the scope of the project, intuitive multi-criteria decision models for platform selection can be seen as a valuable contribution.

Section II contains the theoretical background of the topic. Section III discusses the selection of the multi-criteria method that is the most appropriate for decision making on the digital platforms in agriculture. Finally, Section IV contains decision models and their demonstration.

II. THEORETICAL BACKGROUND

In order to develop a theoretical model for supporting the decision making on digital platforms in agriculture, we relied on a conceptual model of initial utility perception factors previously designed within the same project [3]. The conceptual model of initial utility perception factors is oriented on food distribution platforms, and for the development of the decision model in this research phase, it is sufficiently complex to test our assumption that a solid decision making model can be designed.

For each actor in the decision making process (consumers and producers), 10 factors have been identified. The factors impact the perception of actors regarding the value gained and risks upon which the utility of a platform is subjectively assessed. The factors are listed in Table I and Table II. The factors are classified within each group (customer or producer) and described.

TABLE I. FACTORS OF PERCEPTION – CUSTOMERS PERSPECTIVE

Factor	Description
Eco-friendliness	Reducing the carbon footprint [4], avoiding ecologically unsafe additives, but also caring about the seasonality of products as well as waste load reduction (like packaging options [5]).
Location & time (from farm to fork)	The farm-to-fork paradigm consists of the intention and efforts to buy local fresh products in the shortest time possible time from the farm (or field) to use [6].
Relationship history with the producer	Existing experiences relationships with the producer in terms of responsiveness, expertise, flexibility, commitment and other aspects [7], [8]
Payment options	Price, price volatility, security issues, avoiding cash, and other attributes of payment processes are necessary but not the most important decisive factors. Additional and hidden costs like fee for platform usage [9] also play a role in weighing benefits and costs.
Comfort & convenience	Usefulness and enjoyment, Ease of use [9] affect the subjective sense of value. Also, product information availability can influence the intention to order products and services remotely taking into account benefits like time-saving and reliability of delivery [10].
Recommendations (C2C2C)	Previous product reviews can be understood as recommendations or warnings which can affect the buying decisions of consumers when choosing a platform or specifically a supplier within a platform [10]. In social media-based platforms, recommendations should be based on actual experience and can have a vortex similar behavior (Customer-to-Customer-to-Customer, C2C2C).
Community support	Local and regional centricity as a sense of affiliation [7] like purchasing from local producers with the goal of supporting the sustainability and development of the local economy, local production, and similar elements related to the purpose like social responsibility according to Almquists elements of value [8].
Producer's reliability	Many elements can be significant when it comes to producers' reliability. Some of them are related to economic or performance functional elements of value like product quality, delivery time saving, cost reduction [8].
Trust & traceability	The role of trust [11] through the whole value chain including preparation, production process, harvesting, delivery, and other elements with an acceptable level of traceability.
Health & safety	Even without disruption challenges like epidemics and pandemics health-related awareness of consumers regarding the safety of products and services and their impact on the quality of life loyalty and satisfaction [10].

TABLE II. INITIAL FACTORS OF PERCEPTION – PRODUCERS PERSPECTIVE

Factor	Description
Sales channels	The distinction between customer-preferred sales channels and producer-preferred channels is important for understanding how these channels intersect and allow touchpoints in terms of collaboration options. This factor includes various aspects from food retailing and restaurants to creating new value chains with customer experience journey management [4].
Health & food safety	Health-related awareness of producers regarding the safety of products and services influencing the consumers' perception of their quality of life are impacting loyalty and satisfaction [10]. The overall quality of products, production, supply chain, and delivery to the end customer should also be traceable.
Production technologies	Enhancement of production technologies under the phrase "Green deal" and a smaller carbon footprint. It is oriented towards operational efficiency and effectiveness through the whole production process, including increased control and precision as well as insuring growing potential for more or less complex interfaces in production eco-systems [4].
Product Quality	The quality of the production process and the final product should be communicated through the platform in the appropriate form and by enough product information [10].
Resources	"Consolidation of the value chain as a single player can do many steps in the value chain" [4] can be achieved by implementing digital technologies to empower and support human labor. Robotics and innovation culture to increase the efficiency of the production process should be available regardless of producers' digital competencies.
Inbound logistics / Supply Chain	"Local Supply Capacity and Constraints to Regional Local Agriculture Expansion" was addressed by Werner et al. [6]. Other aspects of this factor include relationships and channels to suppliers of raw materials (seedlings and seeds, fertilizer), time spent on improving the supply chain in order to respond to growing demands locally.
Innovations	The innovation in agriculture can be a factor that according to [12] guarantees adaptation to a new paradigm of the global economy. The lifecycle of innovation should be supported by a distribution platform in order to facilitate the identification of innovation potentials, the innovative design, innovation implementation and monitoring as well as other activities.
Outbound logistics / Distribution chain	Outbound distribution chains have a variety of issues regarding options for ordering, downloading online orders and preparing customer orders, delivery planning, packaging, and the final delivery. Intermediate actors collaborate via other business models like restaurants [6], hotels or tourism enterprises which opens a wide space for improvement potentials.
Incentives and sustainability	By understanding customers' elements of value that trigger them to purchase, long-term sustainability is viewed as an eco-systems and life-changing factor [7]. Incentives for efficiency and capacity increase which come from the governmental or administrational level can affect the readiness of producers to adopt new technologies in their business.
Regulatory compliance	According to [13] control is intended to reduce uncertainty by laying structures that increase the likelihood of anticipated outcomes. Regulatory compliance in agriculture from our point of view is relevant in relation to health and safety issues and time spent by the producer to ensure the compliance.

III. DECISION MAKING ON DIGITAL PLATFORM IN AGRICULTURE

A. Selecting a suitable decision-making method

Decision making on digital platforms in agriculture is undoubtedly a multi-criteria decision-making problem. In the previous section, we presented the list of factors that influence the digital platform selection (i.e. joining or adoption). To be able to successfully deal with ten criteria in the decision-making problem we need a suitable decision-making method. So, a logical research question that appears now is related to the selection of multi-criteria decision-making method that can aggregate all the factors on one side, but on the other side, we need to select a method that is intuitive for the users, not very complex and simple to understand for the users. The idea is to propose a decision-making model that can be successfully used by average agriculture customers and producers who do not have the scientific expertise in the decision-making field.

There are many multi-criteria decision-making methods available for this problem analysis. Some of them are: the analytic hierarchy process (AHP) [14], [15], the analytic network process (ANP) [16], [17], Electre, Topsis, Promethee [18], simple additive weighting (SAW), Vikor [19], Dex [20]–[22], and SNAP [23]–[25]. Due to high-level expertise in multi-criteria decision-making (MCDM) field, we were deeply analyzed those methods.

The results are presented in Table 3. The goal was to select the method that can cover all the factors, but which is as simple for users as possible. Since we knew the average users' profiles in terms of expertise in decision-making methods and characteristics of the analyzed methods, using the qualitative analysis, we selected the Dex method as the one that can be the most successfully used for modeling the decision-making process of the selection of digital platforms in agriculture.

TABLE III. THE ANALYSIS OF THE MCDM METHODS WITH RESPECT TO APPLICATION IN AGRICULTURE FIELD AT THE LEVEL OF AVERAGE USERS

Characteristics	Dex	AHP	ANP	Electre	Topsis	SNAP
Covering all important aspects of the problem	+	+	+	+	+	+
Qualitative factors	It is purely qualitative method – supports decision making in agriculture without need for quantification	Quantification of the qualitative factors using Saaty scale and pairwise comparisons	Quantification of the qualitative factors using Saaty scale and pairwise comparisons	Quantification of the qualitative factors using constructed scales which changes the original factors' values	Quantification of the qualitative factors using constructed scales which changes the original factors' values	Quantification of qualitative factors using DEMATEL scale
Complexity of the method and decision models for final users	After the models are created, the user can intuitively use it	After the model is created, the user must make a high number of pairwise comparisons	After the model is created, the user must make many pairwise comparisons (higher than in AHP)	It covers 10 steps which can be understood by users with high mathematical background; it does not ensure selection of the best alternative	The method covers many steps which can be understood by users with high mathematical background	Very complex, no support for alternatives' level; need to be combined with e.x. composite index
Incorporating the subjectivity (perceived factors)	Highly possible through creating the decision rules and discretization	Possible, but not intuitive	Possible, but not intuitive	Constructed scales are intuitive, but they limit the factors' values	Constructed scales are intuitive, but they limit the factors' values	Possible, but not at the level of the alternative
Simulating human thinking	Highly	Medium, since there are many possible problems in terms of inconsistency	Medium, since there are many possible problems in terms of inconsistency	Creating constructed scales	Creating constructed scales	No support for alternatives' level

Our choice of applying Dex method is based on following:

- Among the mentioned methods, only Dex is a purely qualitative method and factors that influence the decision on the digital platform in agriculture are in most cases qualitative attributes,
- The methodology of the Dex is still complex, but after we propose the general models for decision-making, the implementation of the models will be much simpler for users than in the case of implementation of models that would be created using other methods
- Using Dex, we can aggregate all the factors. The same can be achieved by using other methods which means that Dex is equally strong as other methods in terms of aggregating multiple factors into the decision,
- The Dex is more intuitive than the other mentioned multi-criteria decision-making methods,
- The steps of the Dex are more similar to the human thinking than the steps of other mentioned multi-criteria decision-making methods, and that is very important since the decision models are intended to be applicable by not decision-making experts and professionals.

B. *Dex method*

Decision Expert (Dex) is a qualitative multi-criteria decision-making method. It is also a hierarchical method (problem structure is similar to the structure in the AHP method). Dex is implemented in a software called Dexi and that software is used in this paper.

There are several main steps in Dex method [21], [26], [27]:

- Creating the hierarchical tree of decision making problem. Here, the model is modeled through tree which has several interpretations: decomposition, dependence, and aggregation. In our case, we will define two hierarchical trees, one for producers, and the other for customers. At the lowest levels of the hierarchical trees, there will be factors identified in the previous section. They will be then grouped into higher-level elements, and finally aggregated into the main element at the highest level of the hierarchical tree.
- Creating the scale for each element in the hierarchy. Here, we will define the qualitative scale for each element in the tree. In most cases, values L (low), M (medium) and (high) are used. However, at the lowest level, it is possible to have only two values (L or H), and at higher and the highest level, it is possible to have more than three values (ex. VL (very low) and VH (very high)). In this step, we can also create the rules which describe how are the values from the natural scale of the factors transformed into qualitative values in the Dex (V, L, H): ex. for the factor *Eco-friendliness* we have to have the meanings of the V, L, and H on its natural scale. The input values of the alternatives by certain criteria are determined by the discretization of the continuous value space. This process can be done using the threshold or using the percentile.
- Creating the decision rules. Here we have to create the rules that describe the conclusion mechanism for the elements in the tree that are not at the lowest level of the hierarchy. decision-making rules represent the basic mechanism of conclusion and decision-making in the DEX method [28]. Decision rules are defined at the level of aggregated criteria and at root level decisions that describe which value will take the criterion (on its scale) for each combination of criteria values from the level below.
- Once a hierarchical model has been created, and after the rules of decision are defined, the final step is the evaluation of alternatives. In our case, the alternatives are concrete digital platforms in agriculture

The Dex models in this paper can be helpful to decision-makers (produces and customers) in evaluating the platforms and deciding to use some of them or not.

IV. Decision Models and their Implementation

A. *A decision model for producers*

The hierarchical tree of the decision model for the producer consists of 14 elements. Ten of them are the factors presented in Section II, and three of them are aggregated elements. The tree has been structured as a three-level hierarchy, and the main structuring method was the qualitative analysis. The tree is presented in Figure 1.

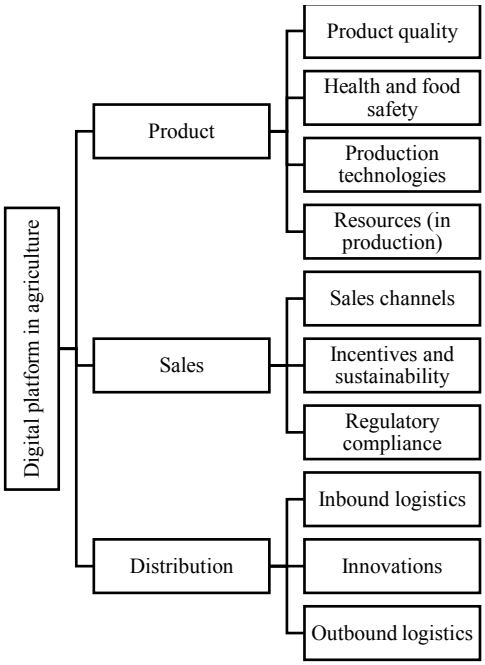


Figure 1. The hierarchical tree of Dex model (producers)

The values of all factors are set to values L, M, and H. Since each producer can perceive each factor value for some digital platform differently, it was not needed to define the discretization rules for factors' values. When evaluating the platforms, the user will use its perception to express the factors' values.

The decision rules for element *Product* are partly presented in Figure 2. Since that element aggregates four factors, and each of them is described with three values, the number of decision rules is 81 (3^4).

Decision rules Product

medium — Use scale orders — Use weights

	Product quality	Health and f	Production techno	Resources	Product
65	high	medium	low	medium	**low**
66	high	medium	low	high	**low**
67	high	medium	medium	low	**low**
68	high	medium	medium	medium	**medium**
69	high	medium	medium	high	**medium**
70	high	medium	high	low	**medium**
71	high	medium	high	medium	**medium**
72	high	medium	high	high	**high**
73	high	high	low	low	**low**
74	high	high	low	medium	**medium**
75	high	high	low	high	**medium**
76	high	high	medium	low	**medium**
77	high	high	medium	medium	**medium**
78	high	high	medium	high	**high**
79	high	high	high	low	**medium**
80	high	high	high	medium	**high**
81	high	high	high	high	**high**

Figure 2. Decision rules for element Product

Similar decision rules are defined for elements *Sales*, *Distribution,* and root element *Digital platform in agriculture*. Those rules can be modified, but they are recommendations of investigators in the field based on their expert opinions.

Finally, in the third step of the Dex method, there is an evaluation of the alternatives. Here, we present the application of the model on three hypothetical cases (evaluation of three platforms, Figure 3).

Option	DP 1	DP2	DP3
. Digital platform in agriculture	medium	medium	high
.. Product	medium	high	medium
... Product quality	medium	high	high
... Health and food safety	medium	high	medium
... Production technologies	medium	high	medium
... Resources (in production)	medium	medium	medium
.. Sales	medium	medium	high
... Sales channels	medium	medium	medium
... Incentives and sustainability	medium	low	high
... Regulatory compliance	medium	medium	high
.. Distribution	medium	medium	high
... Inbound logistics	medium	medium	high
... Innovations	medium	low	medium
... Outbound logistics	medium	high	high

Figure 3. Evaluation of three digital platforms

The user (producer) has to input the values for the elements at the lowest level of the hierarchy (in Figure 3 those are the ones with three dots in the title), and the software, using the previously defined decision rules, calculates the values of higher-level elements. The digital platform DP3 achieves the qualitative grade *high* and is the best among the three platforms. It is recommended to use DP3.

B. *A decision model for customers*

The qualitative analysis as the structuring method is used for creating the hierarchical tree in the Dex model for customers. Again, there are 14 elements of the hierarchy tree. Ten factors from Section II are clustered into three higher-level elements *Producer*, *Product,* and *Sales*. The hierarchical tree of the Dex model is presented in Figure 4.

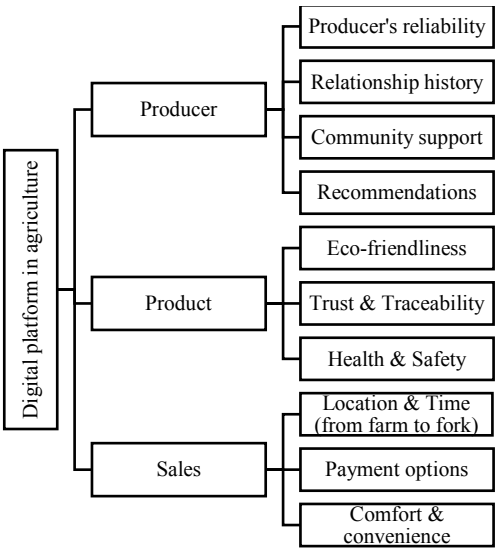


Figure 4. The hierarchical tree of Dex model (customers)

The values of all factors for customers are set in a same way as in the case of producers. The decision rules for element *Product* are partly presented in Figure 5. Since that element aggregates three factors, and each of them is described with three values, the number of decision rules is 27 (3^3).

Decision rules Product

high — Use scale orders — Use weights

	Eco-friendliness	Trust & Traceability	Health & Safety	Product
11	medium	low	medium	medium
12	medium	low	high	medium
13	medium	medium	low	medium
14	medium	medium	medium	medium
15	medium	medium	high	medium
16	medium	high	low	medium
17	medium	high	medium	medium
18	medium	high	high	high
19	high	low	low	medium
20	high	low	medium	medium
21	high	low	high	high
22	high	medium	low	medium
23	high	medium	medium	medium
24	high	medium	high	high
25	high	high	low	medium
26	high	high	medium	high
27	high	high	high	high

Figure 5. Decision rules for element Product

Similar decision rules are defined for elements *Sales*, *Producer,* and root element *Digital platform in agriculture*. Those rules can be modified, but they are recommendations of investigators in the field based on their expert opinions. Figure 6 presents the evaluation of three digital platforms from the perspective of the customers.

Option	DP 1	DP2	DP3
. Digital platform in agriculture	medium	medium	medium
.. Producer	medium	medium	low
... Producer reliability	medium	high	high
... Relationship history	medium	high	medium
... Community support	medium	medium	low
... Reccomnendations	medium	medium	medium
.. Product	medium	medium	high
... Eco-friendliness	medium	medium	medium
... Trust & Traceability	medium	medium	high
... Health & Safety	medium	low	high
.. Sales	medium	low	high
... Location & Time	medium	low	high
... Payment options	medium	low	medium
... Comfort & convenience	medium	low	high

Figure 6. Evaluation of three digital platforms

The user (customer) has to input the values for the elements at the lowest level of the hierarchy (in Figure 6 those are the ones with three dots in the title), and the software, using the previously defined decision rules, calculates the values of higher-level elements. Here, all platforms achieve the same final value at the root element, so analysis at the second level of the hierarchy is needed. Since the DP3 has better values per two criteria than DP2, DP3 would be a better choice than DP2.

V. LIMITATIONS AND FUTURE RESEARCH

The limitations of our research are dealing with coverage of identified perception factors and the lack of evaluation of our decision model in practice. In future research, the evaluation of two proposed decision models (for producers and for customers) need to be conducted and revised with the inclusion of all relevant actors in order to speed and ease up the decision process and to ensure that the attribute domains (i.e. values) for certain factors cover the lack of experience with a certain platform and decision

framworks. More involvement of scientific and professional community as well as of end-users of digital platforms is vital in order to tie up this research more to architectural design and development of functionalities for platform developers.

VI. CONCLUSION

The main goal of this paper was to present the models which can be helpful in terms of decision making on digital platforms in agriculture. Among many possibilities in that area, both – producers and customers – hardly decide which of the platform to use. Using many (or all) of them is almost always impossible since there is a lot of work in updating the information about the products, quantities, prices, etc.

We analyzed several multi-criteria decision-making methods and the conclusion is that the most appropriate method for creating the decision models is Dex - due to its complexity, level of the intelligibility by not experts in decision-making field and possibility, the individuality of perception and aggregating the multiple factors.

Finally, we presented two models. One can be applied by producers, and the other by customers. They can help them evaluate the platform and decide which platform to use. In future research, we plan to evaluate the model in practice (and possibly update it if needed) and present it to the scientific and professional community. That can help producers/customers to optimize their business processes but also give directions for the development of new platform functionalities for platform developers.

ACKNOWLEDGMENT

This research has been conducted as a part of the wider research within the project *Competence Centre for Digital Transformation of the Food Industry in Rural Areas*. The project is co-funded by the European Union through the European Regional Development Fund (ERDF).

REFERENCES

[1] M. Pesce *et al.*, "Research for AGRI Committee - Impacts of the digital economy on the food chain and the CAP," *European Parliament, Policy Department for Structural and Cohesion Policies,* no. February, 2019.

[2] N. M. Trendov, S. Varas, and M. Zeng, "Digital technologies in agriculture and rural areas," *Briefing Paper FAO*, p. 26, 2019.

[3] K. Tomičić-Pupek, I. Pihir, and M. Tomičić Furjan, "The Role of Perception in the Adoption of Digital Platforms in Agriculture," in *Proceedings of Mipro*, 2020. /accepted for publishing

[4] M. Chui, M. Evers, J. Manyika, A. Zheng, and T. Nisbet, "The Bio Revolution," *McKinsey Global Institute*, no. May, 2020.

[5] A. Mumani and R. Stone, "State of the art of user packaging interaction (UPI)," *Packaging Technology and Science*, vol. 31, no. 6, pp. 401–419, Jun. 2018.

[6] S. Werner *et al.*, "Prospects for New England Agriculture: Farm to Fork," *Agricultural and Resource Economics Review*, vol. 48, no. 3, pp. 473–504, Dec. 2019.

[7] E. Almquist, J. Senior, and N. Bloch, "The 30 Elements of Consumer Value: A Hierarchy," *Harvard Business Review*, 2016. [Online]. Available: https://hbr.org/2016/09/the-elements-of-value. [Accessed: 04-Jul-2020].

[8] "What B2B Buyers Really Care About." [Online]. Available: https://hbr.org/2018/03/the-b2b-elements-of-value. [Accessed: 04-Jul-2020].

[9] H. W. Kim, H. C. Chan, and S. Gupta, "Value-based Adoption of Mobile Internet: An empirical investigation," *Decision Support Systems*, vol. 43, no. 1, pp. 111–126, 2007.

[10] A. Tariq, C. Wang, Y. Tanveer, U. Akram, and Z. Akram, "Organic food consumerism through social commerce in China," *Asia Pacific Journal of Marketing and Logistics*, vol. 31, no. 1, pp. 202–222, 2019.

[11] P. Jayashankar, S. Nilakanta, W. J. Johnston, P. Gill, and R. Burres, "IoT adoption in agriculture: the role of trust, perceived value and risk," *Journal of Business and Industrial Marketing*, vol. 33, no. 6, pp. 804–821, 2018.

[12] M. de F. Oliveira *et al.*, "Innovations in Sustainable Agriculture: Case Study of Lis Valley Irrigation District, Portugal," *Sustainability*, vol. 11, no. 2, p. 331, Jan. 2019.

[13] C. Agyekumhene, J. R. de Vries, A. van Paassen, P. Macnaghten, M. Schut, and A. Bregt, "Digital platforms for smallholder credit access: The mediation of trust for cooperation in maize value chain financing," *NJAS - Wageningen Journal of Life Sciences*, vol. 86–87, pp. 77–88, Nov. 2018.

[14] M. G. Marković, N. Kadoić, and B. Kovačić, "Selection and prioritization of adaptivity criteria in intelligent and adaptive hypermedia e-learning systems," *TEM Journal*, vol. 7, no. 1, 2018.

[15] N. Kadoic and I. Kedmenec, "Using ICT tools for decision making support in local government units," in *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2016, pp. 1533–1538.

[16] N. Kadoić, "Characteristics of the Analytic Network Process, a Multi-Criteria Decision-Making Method," *Croatian Operational Research Review*, vol. 9, no. 2, pp. 235–244, 2018.

[17] N. Kadoić, N. Begičević Ređep, and B. Divjak, "Decision Making with the Analytic Network Process," in *SOR 17 Proceedings*, 2017, pp. 180–186.

[18] N. Kadoić, N. Begičević Ređep, and B. Divjak, "E-learning decision making: methods and methodologies," in *Re-Imagining Learning Scenarios*, 2016, vol. CONFERENCE, no. June, p. 24.

[19] C.-H. Liu, G.-H. Tzeng, and M.-H. Lee, "Strategies for improving cruise product sales using hybrid 'multiple criteria decision making' models," *The Service Industries Journal*, vol. 33, no. 5, pp. 542–563, 2013.

[20] M. Bohanec, N. Kadoić, and N. Begičević Ređep, "Qualitative Multiple Criteria Models with Cycles : A Preliminary Study with Method DEX," in *Scientific program of the 24th International Conference on Multiple Criteria Decision Making*, 2018.

[21] M. Bohanec and V. Rajkovič, "DEX : An Expert System Shell for Decision Support," *Sistemica*, vol. 1, no. 1, pp. 145–157, 1990.

[22] M. Bohanec, "Qualitative MultiCriteria Modelling Method DEX: Approach, Recent Advances and Applications," in *Book of Abstracts, 16th International Conference on Operational Research, KOI 2016*, 2016.

[23] N. Kadoić, B. Divjak, and N. Begičević Ređep, "Integrating the DEMATEL with the analytic network process for effective decision-making," *Central European Journal of Operations Research*, vol. 27, no. 3, pp. 653–678, Sep. 2019.

[24] N. Kadoić, N. Begičević Ređep, and B. Divjak, "A new method for strategic decision-making in higher education," *Central European Journal of Operations Research*, no. Special Issue of Croatian Operational Research Society and Collaborators, Oct. 2017.

[25] M. Dzeko, N. Kadoic, and Z. Dobrovic, "Metamodeling SNAP, a Multi-Criteria Method for Effective Strategic Decision Making on e-Learning Issues," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2019, pp. 849–853.

[26] M. Bohanec, M. Žnidaršič, V. Rajkovič, I. Bratko, and B. Zupan, "DEX Methodology: Three Decades of Qualitative Multi-Attribute Modeling," *Informatica*, vol. 37, pp. 49–54, 2013.

[27] V. Đurek, N. Kadoić, and Ž. Dobrović, "Digital Maturity of Higher Education Institution: A meta model of the Analytical Network Process (ANP) and Decision Expert (DEX)," in *Proceedings of Central European Conference on Information and Intelligent Systems 2018*, 2018, vol. 1, pp. 223–230.

[28] M. Mihelčić and M. Bohanec, "Approximating incompletely defined utility functions of qualitative multi-criteria modeling method DEX," *Central European Journal of Operations Research*, vol. 25, no. 3, pp. 627–649, Sep. 2017.

Inovacije i tehnološki napredak u poslovanju morskih luka i njihov utjecaj na gospodarstvo

G. Mudronja*
* Sveučilište u Rijeci, Pomorski fakultet, Zavod za logistiku i menadžment u pomorstvu i prometu, Rijeka, Hrvatska
mudronja@pfri.hr

***Sažetak* - Ulaganje u istraživanje i razvoj te poticanje primjene novih tehnologija važan su segment poslovanja morskih luka budući da utječu na održavanje konkurentske pozicije te samo napredovanje poslovanja morskih luka. Inovacije stvaraju poticaje za dugoročni rast jer stvaraju mogućnosti za poboljšanje produktivnosti. Razvoj novih tehnologija i procesa može rezultirati većom zaradom i potencijalnim ekonomskim rastom gradskog ili regionalnog gospodarstva. Cilj ovoga rada je analizirati inovacije koje se primjenjuju u poslovanju morskih luka te definirati razvojne mogućnosti poslovanja morskih luka u budućnosti iz aspekta tehnološkog napretka. Nadalje, u radu se analizira i utjecaj primjene novih tehnologija u morskim lukama na gospodarski rast. Rezultati istraživanja ukazuju da se većina inovacija u pomorskoj industriji vezuje uz inovacije u dizajnu i izradi plovila, inovacije na području goriva te inovacije u razvoju luka i lučke infrastrukture. No, morske luke imaju također značajne mogućnosti implementacije suvremenih tehnologija u poslovanju u budućnosti kao što su implementacija oblaka *(eng. cloud),* implementacija *blockchain* tehnologije te primjena dronova. Primjena novih tehnologija utječe na ekonomski rast gospodarstva, no istovremeno da bi se nove tehnologije mogle implementirati, gospodarstvo mora imati određene predispozicije kao što su obrazovano stanovništvo, niska stopa nezaposlenosti i usmjerenost na industrijalizaciju.**

Ključne riječi - istraživanje i razvoj; tehnološki napredak; morske luke; gospodarski rast

I. UVOD

Gospodarska vrijednost morskih luka sastoji se od direktnog utjecaja na stvaranje dodane vrijednosti i novih radnih mjesta te indirektnog i induciranog utjecaja koji nastaju zbog postojanja luke na određenom prostoru i njenog poslovanja. Poslovanje morskih luka utječe na konkurentnost grada ili regije u kojima se luka nalazi, no zbog prirode posla vrlo je lako prebaciti poslovanje iz jedne u drugu luku, stoga bi morske luke trebale biti visoko motivirane za stalno poboljšanje svoje konkurentnosti. Općenito, poslovanje u morskim lukama sklono je promjenama te novim trendovima što je indirektno povezano s praćenjem nove tehnologije i ulaganjem u inovacije. Inovativni kapaciteti u poslovanju postaju sve češće pokazatelji gospodarskog uspjeha poduzeća i sektora odnosno luka i sektora brodarstva [1]. Korisnici morskih luka ne zahtijevaju samo prevoženje tereta na najefikasniji način, već je potrebno da i rukovanje podacima bude sofisticirano te da se u poslovanju primjenjuju suvremene tehnologije. Morske luke prate razne trendove od povećanja kapaciteta, inovacija u dizajnu i izradi plovila, inovacija na području goriva do primjene 5G mreže i Interneta stvari *(eng. Internet of Things).* Primjena novih tehnologija doprinosi kvalitetnom prometnom sustavu što je bitan čimbenik za poslovanje morskih luka, a posljedično i za rast gospodarstva. Globalizacija i gospodarski rast stvaraju pozitivne efekte na okolinu te generiraju više razine prihoda čime promet dobiva na važnosti budući da spaja proizvodnju i potrošnju koje se većinom nalaze na različitim stranama svijeta.

Ovaj rad financiran je sredstvima projekta Pomorskog fakulteta Sveučilišta u Rijeci pod nazivom „Kompleksni multiplikativni učinci morskih luka i njihov utjecaj na gospodarstvo“.

II. MORSKE LUKE I S NJIMA POVEZANE INOVACIJE I TEHNOLOŠKI NAPREDAK

Razvoj prometnog sustava i njegova učinkovitost ovise o primjeni novih tehnologija što ujedno stvara preduvjet za gospodarski rast i međunarodnu mobilnost. Integracija svjetskog tržišta koja uključuje slobodan prijevoz robe čini promet vrlo važnim gospodarskim sektorom. Mnoge su koristi transporta i tehnološkog razvoja koji ga prati, no s druge strane postoje i negativni utjecaji. Određeni aspekti smatraju tehnologiju uzrokom mnogih problema, dok je drugi vide kao potencijalni odgovor na nove potrebe i nove izazove kao što su smanjenje negativnog utjecaja na okoliš i korištenje alternativnih izvora energije, ali i poboljšanje učinkovitosti u poslovanju morskih luka. Inovacije u pomorstvu mogu se klasificirati u sljedeće skupine [1]:

- Inovacije u dizajnu i izradi plovila
- Inovacije na području tehnologije goriva
- Inovacije tehnologije brodskih motora
- Inovacije u razvoju infrastrukture i luke
- Integracija i prijenos tehnologije
- Inovacije na ostalim tehnološkim područjima

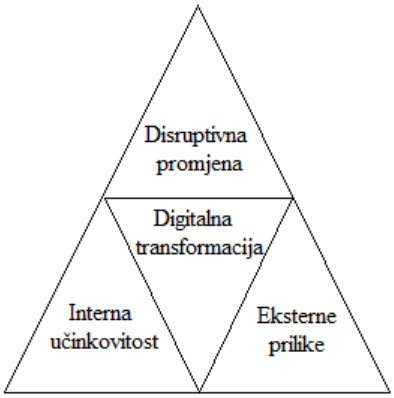
Konstrukcija plovila glavno je područje na kojemu se odvijaju inovacije vezane uz dizajn i izradu plovila. Inovacije su usmjerene na razvoj novih metoda gradnje, odabir materijala, konstruiranje oblika i veličine plovila budući da će plovila koja imaju manju masu i niži aerodinamični otpor imati manju potrošnju goriva, a samim time će se smanjiti i emisija stakleničkih plinova. Također se primjenjuje računalno potpomognuto projektiranje čija svrha je dobivanje rješenja za minimiziranje težine materijala plovila. Gorivo i njegova primjena u pomorstvu bitan su čimbenik vezan uz brigu o klimatskim promjenama te su inovacije na području

tehnologije goriva usmjerene upravo u tom smjeru. Četiri su glavna dijela na kojima se mogu raditi poboljšanja u aspektu goriva: (1) električni pogon i (2) vodik koji su povezani s obnovljivim izvorima energije te su izvori energije koji ne onečišćuju okoliš te (3) bio gorivo i (4) ugljikovodično gorivo koji spadaju pod alternativna goriva koja proizvode manju emisiju stakleničkih plinova te omogućuju smanjenje ovisnosti o neobnovljivim izvorima fosilnih goriva. Što se tiče inovacije vezanih uz tehnologiju brodskih motora, one su također vezane uz brigu o okolišu, tj. smanjenje emisije štetnih plinova, smanjenje gubitaka energije u motorima s unutarnjim izgaranjem te smanjenje buke motora.

Razvoj infrastrukture i razvoj luke te inovacije vezane uz njih mogu osigurati odvajanje prometa od drugih aktivnosti u gradu te time pridonijeti zaštiti kvalitete života građana. No, infrastruktura također može svojim vizualnim karakteristikama smanjiti kvalitetu života što se može ublažiti korištenjem novih materijala i tehnologija. Što se tiče integracije i prijenosa tehnologije važno je za istaknuti da inovacija sama po sebi nije dovoljna, već je potrebna njena implementacija od strane svih sudionika u poslovnom procesu što zna biti izazovno. Nadalje, nova tehnologija zahtijeva nove načine upravljanja i nova pravila koji se u literaturi još nazivaju i *orgware*. Pojam se definira kao mrežna sastavnica informacijskih sustava koja obuhvaća sve metode, propise i specifikacije kojima se usklađuje djelovanje svih sastavnica u poslovnom sustavu. Sustav *orgware* u aspektu poslovanja morskih luka obuhvaća nove ugovore i modele između brodara, brodarskih poduzeća, špeditera, transportnih poduzeća, operatera terminala, države i ostalih sudionika na tržištu. Inovacije na ostalim tehnološkim područjima obuhvaćaju većinom kontrolu prometa i unaprjeđenje njenih sustava kako bi se poboljšali prometni tokovi te smanjilo zagađenje okoliša [1].

Primjena Jedinstvenog sučelja za formalnosti u pomorskom prometu *(eng. Maritime National Single Window)* tehnološki je napredak koji je značajno promijenio proces razmjene informacija između sudionika u pomorskom prometu i u poslovanju morskih luka. Koncept omogućava svim sudionicima koji su uključeni u poslovni proces unošenje podataka samo jednom [2]. Time se eliminira potreba da jedan sudionik dostavlja iste podatke više puta različitim sudionicima poslovnog procesa. Jedinstveno sučelje za formalnosti u pomorskom prometu koristi jednu točku unosa podataka koja je u osnovi jedinstveni i umreženi informacijsko – komunikacijski sustav. Osim administrativnih postupaka koji spadaju pod područje Jedinstvenog sučelja za formalnosti u pomorskom prometu, komercijalni postupci također se trebaju provoditi na učinkovit način što se postiglo korištenjem koncepta Informacijskog sustava dionika u lukama *(eng. Port Community System)*. Koncept se može definirati kao elektronička platforma koja omogućuje inteligentnu i sigurnu razmjenu informacija između javnih i privatnih sudionika poslovnog procesa, kako bi se poboljšao konkurentski položaj morskih luka. Informacijski sustav dionika u lukama optimizira, upravlja i automatizira lučke procese putem jednokratne pohrane podataka te putem povezivanja transportnih i logističkih lanaca [3]. Navedeni koncept utječe na smanjenje troškova, omogućuje brži protok informacija, efikasniju i bržu isporuku tereta što u konačnici može pozitivno utjecati na gospodarski rast [4].

Digitalna transformacija kao proces uvođenja digitalnih tehnologija u poslovanje važna je za poslovanja morskih luka budući da pojednostavljuje poslovne procese i omogućava veću transparentnost poslovanja. Shema 1 prikazuje ciljeve kojima doprinosi digitalna transformacija. Digitalnom transformacijom poboljšava se način rada te se ponovno planiraju unutarnji procesi što doprinosi boljoj internoj učinkovitosti. Nadalje, javljaju se i eksterne prilike u postojećoj poslovnoj domeni, npr. privlačenje novih kupaca, pružanje novih usluga i slično, a sve zajedno vodi ka disruptivnoj promjeni, odnosno digitalizacija uzrokuje potpunu promjenu poslovnih procesa [5].



Shema 1. Prikaz ciljeva digitalne transformacije [5]

Inovacije i tehnološke promjene ne moraju uvijek biti veliki i revolucionarni izumi, već se procesom inoviranja također može smatrati i diskontirani proces malih poboljšanja. Geerlings [6] smatra da bi ukupan rezultat malih inovacija i poboljšanja mogao biti značajniji od velikih revolucionarnih otkrića. Nadalje, nakon što se inovacije dogode, one ne moraju odmah biti prihvaćene od okoline te ne moraju automatski biti superiornije od starih tehnologija. Stoga se u morskim lukama kao i u ostalim sektorima u proces učenja uvode poboljšanja koja novu tehnologiju adaptiraju kako bi se mogla koristiti u poslovanju. Morske luke rijetko primjenjuju velike revolucionarne inovacije te se češće javljaju manja poboljšanja na već spomenutim područjima dizajna, goriva, brodskih motora i infrastrukture.

III. RAZVOJNE MOGUĆNOSTI POSLOVANJA MORSKIH LUKA

Inovacije u pomorskom sektoru većinom su orijentirane na propulziju plovila, odnosno dizajn plovila, tehnologiju brodskih motora, gorivo i slično što znači da su usmjerene na postojeće tržište. Manji broj inovacija bavi se povećanjem korištenja informacijsko – komunikacijske

tehnologije, a najmanje se inovacije bave stvaranjem manjih mreža za prijevoz robe unutar grada pomoću unutarnjih plovnih putova i kanala. Općenito, pametna tehnološka rješenja omogućuju lukama da postanu efikasnije u pogledu protoka informacija [7]. Razvojne mogućnosti morskih luka i logističkog sektora mogu se razvijati u različitim pravcima, a neki od njih su sljedeći:

- Primjena 5G mreže
- Primjena koncepta Interneta stvari *(eng. Internet of Things)*
- Primjena koncepta Interneta vozila *(eng. Internet of Vehicles)*
- Primjena oblaka *(eng. cloud)*
- Primjena *blockchain* tehnologije
- Primjena dronova

Jedna od razvojnih mogućnosti morskih luka je primjena 5G mreže koja može odgovoriti na zahtjeve automatizacije i umjetne inteligencije koji se pojavljuju kod operativnih zahtjeva. Mreža 5G može se primjenjivati u svakodnevnim operacijama morskih luka i terminala jer pruža mogućnost vrlo brzog prijenosa podataka te može podržati digitalnu transformaciju industrije kontejnerskih brodova povezujući logistička poduzeća i luke. Luka Hamburg testirala je 5G tehnologiju te su rezultati pokazali kako će 5G mreža pružiti pomoć inženjerima prilikom optimiziranja i praćenja cjelokupnog sustava. Primjena 5G mreže omogućuje preduvjete za razvoj drugih tehnologija kao što je koncept Internet stvari koji služi za povezivanje fizičkih objekata. Primjena koncepta Interneta stvari u morskim lukama može ići u smjeru njegovog postavljanja na lučku opremu koja bi putem senzora prenosila podatke. Time bi se povećala transparentnost i učinkovitost lučkih operacija, ali i stvorila velika količina podataka koje luke i terminali mogu iskoristiti za optimizaciju svojih procesa i poboljšanje usluga [8].

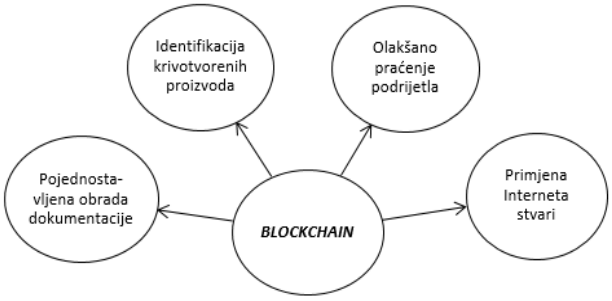
Internet vozila tehnologija je čiji senzori mogu biti postavljeni na vozila, ali i lučke objekte i parkirališta. Time bi omogućili informacije o optimalnom putu, očekivanim troškovima te dostupnim mjestima. Također bi tehnologija omogućila povećanu sigurnost u prometu te manje sudara vozila. Luka Valencija zajedno s kompanijom MSC Mediterranean Shipping Company uvela je koncept Interneta vozila te su kamioni MSC-a koji posluju u Španjolskoj opremljeni uređajima koji ih prate u stvarnom vremenu. Time pružaju informacije luci Valencija u predviđanju potencijalnih gužvi te omogućuju učinkovitije upravljanje lučkim prometom [8].

Korištenje oblaka u pomorskom sektoru može pomoći lukama i terminalima da se povežu u većoj mjeri te da prijeđu na automatizirane procese. Također se mogu primijeniti na platformama za planiranje ruta koje omogućuju korisnicima dijeljenje informacija u stvarnom vremenu te pružaju mogućnost primjene platforme za operativne sustave Internet stvari. Primjer je luka Antwerpen koja koristi automatizaciju temeljenu na oblaku u sklopu svojeg projekta *„Smart Port Vision“* [9].

Blockchain tehnologija smatra se podobnom za implementaciju u poslovanju morskih luka i u logističkim procesima zbog velikog broja sudionika uključenih u njihovo poslovanje. Ona nudi inovativnu platformu za novi decentralizirani i transparentni mehanizam transakcija u industriji i poslovanju [10]. No, tehnologija se još nije u potpunosti implementirala te su njene mogućnosti u fazi testiranja i predviđanja. Potencijalna područja implementacije *blockchain* tehnologije mogu se razvrstati u četiri skupine [11]:

- Pohrana i prijenos dokumentacije tereta
- Vidljivost procesa
- Smanjenje vremena za komercijalna odobrenja
- Automatizacija operacija

Korištenjem *blockchain* tehnologije za pohranu i prijenos dokumentacije tereta omogućile bi se točne informacije o dokumentaciji diljem logističkog lanca čime bi se također utjecalo na bolju transparentnost cjelokupnog logističkog procesa. Poboljšanje vidljivosti procesa omogućuje poboljšanu koordinaciju između svih sudionika u poslovanju čime se ubrzava planiranje prijevoza tereta kao i carinjenje. Smanjenjem vremena za komercijalna odobrenja poboljšalo bi se trgovinsko financiranje, tj. omogućila bi se vidljivost cjelokupnog procesa sudionicima koji se ne nalaze u luci (npr. banke, osiguravajuće kuće) što bi bilo korisno u rješavanju uskih grla [11]. Kvaliteta interakcija među sudionicima u morskim lukama određuje njihov poslovni uspjeh [12]. Sudionici u poslovanju također su suočeni s učestalom izradom i distribucijom papirnatih dokumenata brojnim upravnim tijelima [2] što bi *blockchain* tehnologija mogla minimizirati mogućnošću automatizacije operacija korištenjem Interneta stvari i pametnih ugovora. Neke od prednosti primjene *blockchain* tehnologije prikazane su na shemi 2. Iako *blockchain* tehnologija pruža brojne prednosti, njeni nedostaci su nepostojanje jedinstvenog temeljnog standarda, teško savladivi koncepti tehnologije te potreba za programiranjem intervencija čak i u najjednostavnijim oblicima implementacije [13].



Shema 2. Potencijali primjene *blockchain* tehnologije u morskim lukama [5]

Primjena dronova u poslovanju morskih luka prvenstveno se odnosi na praćenje i kontrolu emisija plinova s plovila budući da je s dronovima lakše doći na morsku lokaciju [14]. No, dronovi imaju potencijal za daljnji razvoj u logistici kao primjerice obavljanje manjih operacija isporuke paketa i upravljanje skladištem. Trgovački lanac

Wallmart u posljednjih nekoliko godina radi na testiranju dronova u svojim skladištima s namjerom da se dronovima u budućnosti upravlja skladištima. Nadalje, Amazon posjeduje patent za dostavu paketa dronovima koji odgovaraju na razne ljudske geste. Dronovi u kombinaciji s kvalitetnim logističkim softverom mogu poboljšati učinkovitost poslovanja i unaprijediti poslovne procese morskih luka i logističkog sektora [15].

Korištenje novih tehnologija i razvoj i primjena inovacija uvelike mogu unaprijediti poslovne procese pomorskog sektora, no također postoje i mnoge tehnološke, ekonomske, institucionalne i političke zapreke za uspješno uvođenje inovacija i novih tehnologija. Prema autorima Geerlings i Wiegsmans [1] potrebno je obratiti pozornost na paradoks koji kaže da tehnološka rješenja mogu biti politički poželjnija jer traže najmanje iznose državnih intervencija. S druge strane, da bi se tehnologija smatrala isplativom mora biti ekonomski korisna za kupce i dobavljače. Ukoliko se najštetniji načini prijevoza unaprijede, postat će isplativiji te će se smanjiti njihovi operativni troškovi. Tada dolazi do situacije da se privlači promet od manje štetnih prema štetnijima načinima prijevoza što je kontraproduktivno. No, s druge strane mnogo je koristi koje nastaju uvođenjem novih tehnologija u poslovanje kao što je veća transparentnost cjelokupnog procesa, niži troškovi, bolja usklađenost među sudionicima, smanjenje uskih grla i slično te sve navedeno dovodi do boljeg poslovanja morskih luka, a dugoročno i do pozitivnog utjecaja na gospodarski rast.

IV. UTJECAJ MORSKIH LUKA NA GOSPODARSTVO

Morske luke doprinose rastu bruto domaćeg proizvoda te su važan čimbenik modernog društva budući da se proizvodnja i potrošnja mnogih proizvoda nalaze na različitim dijelovima svijeta. Uloga luka je povezivanje proizvodnje i potrošnje te stvaranje globalnih lanaca opskrbe. Ekonomska vrijednost luke može se definirati u okvirima vrijednosti koja je ostvarena u određenom razdoblju zbog prisutnosti luke, kao što je iznos dodane vrijednosti i broj zaposlenih ljudi u luci [16]. Luke su važan čimbenik gospodarstava jer smanjuju troškove prijevoza, povećavaju zaposlenost te stvaraju ostale popratne djelatnosti. Većina svjetske trgovine u nekom segmentu ima veze s aktivnostima vezanim uz pomorstvo te mnoge morske luke imaju koristi od povećane međunarodne trgovine. Proces globalizacije, integracija svjetskog tržišta te gospodarski rast doprinijeli su značaju prometa kao jedne od važnijih gospodarskih grana.

Osim spomenutih direktnih učinaka koje morske luke stvaraju na gospodarstvo, vezanih uz stvaranje dodatne vrijednosti i stvaranje novih radnih mjesta, one također stvaraju i indirektne i inducirane učinke. Indirektni učinci nastaju kada se ostvareni dohodci od rada u morskim lukama te poslovna zarada i porezni prihodi naknadno troše u okruženju npr. u mjestima za zabavu i restoranima čime generiraju nova radna mjesta i poslovnu zaradu. Inducirani učinci vezani su uz poboljšanje kvalitete usluga u prometnom sustavu te poboljšanje infrastrukture na određenom prostoru zbog egzistiranja luke [17].

Globalizacija je fenomen koji je uvelike utjecao na poslovanje morskih luka budući da resursi i dobra nisu uvijek na istoj lokaciji proizvodnje i potrošnje, stoga je pomorski prijevoz jedan od načina njihovog povezivanja. Globalno kretanje tereta uključuje obale i oceane, unutarnje plovne puteve, željeznicu, ceste i zračni prijevoz. Glavni čimbenik koji pokreće povećanje i širenje lučkih djelatnosti je upravo globalizacija proizvodnje jer su proizvodni sustavi postali sve više globalno isprepleteni te se sirovine, poluproizvodi, ali i gotovi proizvodi pronalaze na različitim stranama svijeta [18]. Kako bi luke i u budućnosti ostale konkurentne i prilagođavale se procesu globalizacije, potrebno je da svoje poslovanje prilagođavaju internim i eksternim izazovima koji su navedeni u sljedećoj tablici.

TABLICA I. INTERNI I EKSTERNI IZAZOVI U MORSKIM LUKAMA KOJI SE JAVLJAJU ZBOG GLOBALIZACIJE [19] [20]

Interni izazovi	Eksterni izazovi
Rast poslovanja	Novi konkurenti
Kontinuirane inovacije	Nova tehnologija
Implementacija najnovijih *know-how* načina rada	Novi poslovni modeli
Implementacija najnovijih tehnologija za rad terminala	Promjene na tržištu
Zapošljavanje visokoobrazovane radne snage	Promjene rizika
Novo organizacijsko razmišljanje	Regulatorni izazovi
Pružanje lučkih usluga orijentiranih na korisnike	Promjene atraktivnosti posla
Oslobađanje od birokratskih pravila i postupaka	Geopolitički razvoj
Razvoj vlastite stručnosti za upravljanje lukama i razvoj lučke strategije	

Kao i u svakoj gospodarskoj grani pa tako i u pomorstvu, morske luke moraju pratiti trendove kako bi ostale konkurentne na tržištu. Prethodno je navedeno da je većina trendova u morskim lukama povezana s digitalizacijom i automatizacijom te je usmjerena na razvoj tehnologije i inovacija, što se može vidjeti i iz prethodnih smjernica koje su većinom usmjerene upravo na poboljšanje pružanja usluga primjenom novih tehnologija. Ulaganje u istraživanje i razvoj povećava mogućnost postizanja višeg tehnološkog standarda što rezultira uvođenjem novih

proizvoda ili proizvodnih procesa te potencijalno višim razinama prihoda i gospodarskim rastom. S druge strane, razvoj novih ideja i nove tehnologije dugotrajan je proces koji i nakon završetka ne mora biti pozitivno prihvaćen od okoline. Razvoj i prihvaćanje nove tehnologije uvelike ovisi o kulturnim i organizacijskim čimbenicima područja u kojima se odvija. Nisu sva područja sklona inovacijama te postoje određeni „društveni filteri“ koji stvaraju preduvjete za njihovu implementaciju [21]. Demografska i obrazovna dostignuća preduvjet su za implementaciju inovacija, odnosno područja s obrazovanijim stanovništvom bit će sklonija inovacijama od onih s manje obrazovanim stanovništvom. Nadalje, područja s visokom zaposlenošću sklonija su inovacijama od područja s niskom zaposlenošću, kao i područja koja su u većoj mjeri usmjerena na proizvodni i uslužni sektor, a kojima pripada i pomorski sektor.

V. ZAKLJUČAK

Razvoj i primjena inovacija i novih tehnologija važni su čimbenici poslovanja morskih luka jer utječu na njihovu konkurentsku poziciju i razvoj poslovnog procesa. Zbog prirode posla korisnici mogu vrlo lako prebaciti teret iz jedne luke u drugu, ukoliko nisu zadovoljni s uslugom, što je još jedan od razloga zašto se morske luke moraju truditi pružati kvalitetnu uslugu temeljenu na suvremenoj tehnologiji. Inovacije u morskim lukama većinom su usmjerene na već postojeća tržišta i obuhvaćaju diskontirani proces malih poboljšanja na područjima dizajna i izrade plovila, tehnologije goriva, tehnologije brodskih motora, razvoja infrastrukture i same luke, integracija i prijenosa tehnologije te na ostalim tehnološkim područjima. No, razvojne mogućnosti morskih luka mogu obuhvaćati i inovacije koje su usmjerene više na informacijsko – komunikacijske tehnologije. Primjena 5G mreže u morskim lukama može omogućiti digitalnu transformaciju morskih luka i vrlo brzi prijenos podataka što pruža dobre temelje za primjenu koncepata Interneta stvari i Interneta vozila. Navedeni koncepti obuhvaćaju senzore koji se postavljaju na opremu i vozila putem kojih se prenose podaci te omogućuju bolju transparentnost poslovnog procesa čime se također i ubrzava poslovanje. Primjena oblaka u poslovanju morskih luka također se vezuje uz poboljšanje dijeljenja podataka u stvarnom vremenu te potporu za prelazak morskih luka na automatizirane procese. *Blockchain* tehnologija omogućila bi bolju pohranu i prijenos dokumentacije tereta čime bi se poboljšala vidljivost procesa i koordinacija između svih sudionika. Smanjilo bi se vrijeme za komercijalna odobrenja te poboljšala automatizacija operacija. Dronovi u poslovanju morskih luka također su jedna od razvojnih mogućnosti te se mogu koristiti za praćenje i kontrolu emisije plinova s plovila na moru, ali i za logističke procese upravljanja skladištem i isporuke paketa. Primjena inovacija i novih tehnologija u morskim lukama mogu uvelike unaprijediti poslovne procese čime se utječe na smanjenje troškova poslovanja, usklađivanje svih poslovnih procesa, bolju transparentnost i smanjenje uskih grla. Sve navedeno poboljšava poslovanje morskih luka kao i njihov financijski rezultat, što posljedično može utjecati na gospodarski rast.

LITERATURA

[1] Geerlings, H. and Wiegmans, B., "Technological Innovations", in Geerlings, H., Kuipers, B., and Zuidwijk, R. (eds) Ports and Networks, Strategies, Operations and Perspectives. London and New York: Taylor & Francis Group, pp. 332–347., 2018.

[2] Tijan, E., Agatić, A., Jović, M. and Aksentijević, S., "Maritime National Single Window – A Prerequisite for Sustainable Seaport Business", Sustainability, 11(17):4570, str.1-21, 2019.

[3] International Port Community Systems Association (IPCSA), "Port Community Systems", 2014., Dostupno: https://ipcsa.international/pcs

[4] Tijan, E, Jardas, M., Aksentijević, S. and Perić Hadžić, A. "Integrating Maritime National Single Window with Port Community System – Case Study Croatia " in Proceedings of the 31st Bled eConference – Digital Transformation: Meeting the Challenges Conference Proceedings, University of Maribor Press, Bled, Slovenia, pp. 1–11., 2018.

[5] Aksentijević, S. "Collection of the main disruptive innovations, main innovations in BIG DATA management and main automation innovations that could affect the area in the next five years", Digitalizing Logistic Processes (DIGILOGS), University of Rijeka, Faculty of Maritime Studies, 2019., neobjavljeno

[6] Geerlings, H., "Meeting the Challenge of Sustainable Mobility", Heidelberg i Berlin, Njemačka: Springer Verlag, 1998.

[7] Jović, M., Kavran, N., Aksentijević, S. and Tijan, E., "The Transition of Croatian Seaports into Smart Ports" in MIPRO 2019 42nd International Convention Proceedings, Opatija, Hrvatska, 2019.

[8] The Motorways of the Sea Digital Multichannel Platform, "Insight: Four Trends of ports to watch in 2019", 2019. Dostupno: https://www.onthemosway.eu/insight-four-trends-of-ports-to-watch-in-2019/?cn-reloaded=1.

[9] Port Technology, "Ports in the Cloud: The Next Step in Automation?", 2018. Dostupno: https://www.porttechnology.org/news/ports_in_the_cloud_the_next_step_in_automation.

[10] Tijan, E., Aksentijević, S., Ivanić, K. and Jardas, M., "Blockchain Technology Implementation in Logistics", Sustainability, 11(4):1185, pp. 1-13, 2019.

[11] Francisconi, M., "An explorative study on blockchain technology in application to port logistics", Master thesis, Delft University of Technology; Delft, Netherlands, 2017. Dostupno:

https://repository.tudelft.nl/islandora/object/uuid%3Ab1de98dd-f1a7-456a-b8cb-4d87dc2c4f9f.
[12] Tijan, E., Jović, M. and Karanikić, P., "Economic and Ecological Aspects of Electronic Transportation Management Systems in Seaports", in Proceedings of the Maritime and Port Logistics Bar Conference, Bar, Crna Gora, pp. 132-138, 2019.
[13] Jović, M., Filipović, M., Tijan, E. and Jardas, M., "A Review of Blockchain Technology Implementation in Shipping Industry", Pomorstvo: Scientific Journal of Maritime Research, 33(2), pp. 140-148, 2019.
[14] Xia, J., Wang, K. and Wang, S., "Drone scheduling to monitor vessels in emission control areas", Transportation Research Part B. Elsevier Ltd, 119, pp. 174–196, 2019.
[15] Capterra, "Drones in Logistics: A How-To Guide for Beginners", 2018. Dostupno: https://blog.capterra.com/drones-in-logistics-how-to-guide/.
[16] Van den Bosch, F. A. J., Hollen, R. M. A. and Volberda, H. W., "How Ports Create Strategic Value for their Country", in Geerlings, H., Kuipers, B., and Zuidwijk, R. (eds) Ports and Networks, Strategies, Operations and Perspectives. London and New York: Taylor & Francis Group, pp. 38–53., 2018. Dostupno: http://elibrary.worldbank.org/doi/pdf/10.1596/1813-9450-6259.
[17] Talley, W. K., "Port Economics", London and New York: Routledge, Taylor & Francis Group, 2009.
[18] Hesse, M., "Global Chain, Local Pain: Regional Implications of Global Distribution Networks in the German North Range", Growth and Change, 37(4), pp. 570–596, 2006.
[19] Buck, R., "Trade Globalisation or Investment Regionalisation? The Impact on European Seaports", ESPO Conference, 2018. Dostupno: https://www.espo.be/media/31-05 1130 Rene Buck.pdf
[20] Inoue, S., "Challenges of the World's Ports in the Globalization Era", The 5th Young Leaders Training & Research Program. Ulaanbaatar, Mongolia, 2010. Dostupno: https://www.neaef.org/public/neaef/files/documents/publications_pdf/young_leaders/5th/Inoue Satoshi 1.pdf
[21] Rodríguez-Pose, A., "Innovation Prone and Innovation Averse Societies: Economic Performance in Europe", Growth and Change, 30(1), pp. 75–105, 1999.

Digitalna transformacija pomorskog transporta kao dijela Plave ekonomije

M. Jović*, A. Agatić**, i A. Jugović***
* Sveučilište u Rijeci, Pomorski fakultet/ Zavod za logistiku i menadžment u pomorstvu i prometu, Rijeka, Hrvatska
jovic@pfri.hr
** Sveučilište u Rijeci, Pomorski fakultet/ Zavod za logistiku i menadžment u pomorstvu i prometu, Rijeka, Hrvatska
agatic@pfri.hr
*** Sveučilište u Rijeci, Pomorski fakultet/ Zavod za logistiku i menadžment u pomorstvu i prometu, Rijeka, Hrvatska
ajugovic@pfri.hr

***Sažetak* - U ovom znanstvenom radu se analizira važnost digitalne transformacije Plave ekonomije, s naglaskom na digitalnu transformaciju pomorskog transporta, koji predstavlja jednu od važnih aktivnosti Plave ekonomije te osim brojnih dionika uključuje i brojne resurse (luke, obalna područja, itd.). Pojašnjeni su pojmovi i koncept digitalne transformacije i Plave ekonomije te su definirane aktivnosti Plave ekonomije. Nadalje, u radu je pozornost posvećena utjecaju pomorskog transporta na ekonomiju te novim tehnologijama kao pokretačima digitalne transformacije.**

Ključne riječi – digitalna transformacija; plava ekonomija; pomorski transport

I. UVOD

Svjetski oceani i mora imaju veliki potencijal doprinijeti opskrbi hranom, energijom i prirodnim resursima [1]. Gotovo 80 posto globalne trgovine robom prevozi se morskim putem [2].

Plava ekonomija je koncept u nastajanju koji potiče efektivnije upravljanje oceanima ili „plavim“ resursima [3]. Uz Plavu ekonomiju, često se spominje pojam Plavi rast, koji podrazumijeva ostvarenje gospodarskog rasta temeljenog na iskorištavanju morskih resursa, a istodobno sprečavajući njihovu degradaciju, prekomjernu upotrebu i zagađenje. [4]. Stoga zaštita oceana i mora postaje ključna u postizanju globalnog cilja održivosti [5].

Digitalna transformacija je proces preoblikovanja poslovnih modela uslijed usvajanja i korištenja digitalnih tehnologija koji je prisutan u brojnim industrijskim i društvenim aktivnostima [6], [7]. Posljednjih godina poduzeća u gotovo svim industrijama provela su niz inicijativa za istraživanje novih digitalnih tehnologija i iskorištavanje njihovih potencijala [8]. Digitalna transformacija je važna za učinkovitost, pojednostavljenje procesa, smanjenje troškova i bolje korištenje resursa i postojeće infrastrukture [9]. Isto se postiže primjenom digitalnih tehnologija, te napose transformacijom ključnih poslovnih aktivnosti čime se utječe na proizvode, procese, organizacijske strukture i koncepte upravljanja [8].

Transparentnost i jednostavan pristup podacima osnova su za uspješno poslovanje. Stoga, istraživački problem proizlazi iz povećanih troškova i izgubljenog vremena zbog zastarjelih postupaka i neadekvatnog provođenja i praćenja poslovnih procesa. Cilj istraživanja u ovome radu je analizirati važnost digitalne transformacije Plave ekonomije kao i novih tehnologija koji predstavljaju pokretače digitalne transformacije, s naglaskom na pomorski transport.

II. PLAVA EKONOMIJA I DIGITALNA TRANSFORMACIJA

Svrha Plave ekonomije, kao koncepta održivog korištenja mora i oceana, jest ekonomski razvoj, poboljšanje kvalitete življenja i rada te očuvanje ekosustava [10]. Upravo stoga ovom radu, naglasak je stavljen na ekonomski razvoj i očuvanje ekosustava.

Pod Plavu ekonomiju spada više aktivnosti: tradicionalne (poput ribarstva, vađenja nafte i plina), kao i nove, brzorastuće (poput plave biotehnologije), koje pokazuju veliki potencijal za daljnji ekonomski rast, otvaranje novih radnih mjesta i inovacije [11]. Aktivnosti koje spadaju pod Plavu ekonomiju su [12], [13]:

- Korištenje prirodnih živih resursa: prehrambeni proizvodi iz djelatnosti ribarstva i akvakulture te farmaceutski i kemijski proizvodi kao dio morske biotehnologije. Povećane potrebe za sigurnom i nutritivno vrijednom hranom te inovativnim farmaceutskim proizvodima potiču održivo korištenje živih resursa mora i oceana.
- Korištenje prirodnih neživih resursa: vađenje mineralnih sirovina, energetskih resursa i desalinizacija vode. Aktivnosti su potaknute porastom potražnje za mineralnim sirovinama, obnovljivim izvorima energije i pitkom vodom.
- Transport i uslužne djelatnosti na moru, oceanima i obalnim područjima: pomorski transport i turističke aktivnosti. Pomorski transport kao dominantan u robnoj razmjeni bilježi kontinuirani rast te čini važnu aktivnost Plave ekonomije s obzirom da uključuje brojne resurse (luke, obalna područja, pomorska dobra, itd.) i brojne dionike pomorske trgovine. Turizam bilježi rast te posljedično snažnije djelovanje na obalna područja. Obuhvat i obim aktivnosti, osobito pomorskog transporta, značajno utječe na Plavu ekonomiju.
- Industrijske aktivnosti na obalnim područjima: izgradnja brodova i plutajućih objekata, izgradnja čamaca za zabavu i sport, brodski strojevi,

brodska oprema, popravak i održavanje brodova i čamaca. Inovativne tehnologije koje se koriste pri gradnji brodova i tehnološki napredna oprema pridonose održivosti obalnih područja te održivosti mora i oceana.

- Očuvanje mora i oceana: monitoring mora i oceana, sekvestracija ugljika, zaštita obalnog područja i gospodarenje otpadom. Aktivnosti su potaknute povećanom potrebom zaštite mora, oceana i obalnih područja.

Digitalna transformacija trenutno je važan trend koji prodire u brojne industrijske i društvene domene [7]. Prema [14], digitalna transformacija je upotreba novih digitalnih tehnologija kako bi se omogućilo poboljšanje poslovanja (primjerice, poboljšanje korisničkog iskustva, pojednostavljenje operacija ili stvaranje novih poslovnih modela). [15] definiraju digitalnu transformaciju kao proces obnavljanja i ponovnog inženjeringa poslovanja radi digitalizacije poduzeća.

Prema [16], s jedne strane, digitalna transformacija olakšava promjene u poslovnim modelima i pruža nove mogućnosti ostvarivanja prihoda. S druge strane, digitalnu transformaciju pokreću nove digitalne tehnologije, koje omogućavaju poboljšanje performansi poduzeća.

III. METODOLOGIJA

Autori su započeli istraživanje korištenjem kombinacije ključnih riječi „Plava ekonomija i digitalna transformacija“, „Plava ekonomija i pomorski transport“ i „Digitalna transformacija i pomorski transport“ i alternativnih ključnih riječi „Plava ekonomija i tehnologije“ i „Digitalna transformacija i transport“ (naslov, sažetak i ključne riječi). U tu svrhu uglavnom su korištene baze podataka Web Science, Google Scholar, ResearchGate i SpringerLink. Pretraživanje članaka provodilo se prema vremenskim ograničenjima (2015-2020) i uglavnom je uključivalo članke iz časopisa i radove sa konferencija. Kako bi se osiguralo da mogući korisni nalazi iz različitih područja nisu isključeni, autori nisu ograničili upite na određeno polje ili indeks.

IV. POMORSKI TRANSPORT I NOVE TEHNOLOGIJE KAO POKRETAČI DIGITALNE TRANSFORMACIJE

Transport je multimodalni, multi-problemski i multi-spektralni sustav, jer uključuje različite kategorije i aktivnosti, kao što su donošenje politika, planiranje, projektiranje, izgradnja i razvoj infrastrukture [17]. S jedne strane, transport (uključujući sve transportne grane) je jedan od glavnih čimbenika gospodarskog rasta i kvalitete života; s druge strane, jedan je od glavnih uzroka onečišćenja okoliša [18].

Kako je ranije već navedeno, najveći udio roba u svijetu prevozi se morem, s 80% volumena i 70% vrijednosti ukupnog tereta u svijetu [19]. S ekonomskog aspekta, prednosti pomorskog transporta su brojne od: velikog pojedinačnog kapaciteta, ogromnog ukupnog kapaciteta čitave prometne grane, niskih troškova i niskih cijena prijevoznog puta, do fleksibilnosti potražnje, dugog životnog vijeka i mora kao slobodnog puta koji ne zahtijeva velika ulaganja u infrastrukturu [20].

Prema procjeni Konferencije Ujedinjenih naroda za trgovinu i razvoj (UNCTAD-*United Nations Conference on Trade and Development)* očekuje se prosječni godišnji rast pomorske trgovine od 3,5 % do 2024. godine [21]. Pomorski transport ima ključnu ulogu u međunarodnoj trgovini robom Europske unije (EU). U 2015. godini, vrijednost pomorske trgovine EU robom sa zemljama koje nisu članice EU procijenjena je na otprilike 1,777 milijardi eura, što čini oko 51% trgovine robom u EU [22]. Sveukupno, pomorski transport je u 2017. godini u EU činio 6% radnih mjesta, 12% BDV-a i 16% dobiti u ukupnoj Plavoj ekonomiji [13].

Nove tehnologije, koje se nazivaju i pokretačima digitalne transformacije, na primjer Internet stvari (*Internet of Things*), umjetna inteligencija (*artificial intelligence*), proširena stvarnost (*augmented reality*), virtualna stvarnost (*virtual reality*) itd.), prisutne su u pomorskoj industriji, omogućujući povećanu učinkovitost brodskih operacija [23]. Očekuje se da će globalni prihodi od pomorskog teretnog prometa porasti na preko 205 milijardi USD do 2023. godine, a temelj za to je korištenje digitalnih tehnologija odnosno digitalna transformacija [30]. U digitalnoj transformaciji tržišta pomorskog teretnog prometa, segment softverskih rješenja u 2017. godini procijenjen je na 15,3 milijarde dolara. Tehnologije i digitalna transformacija pomorskog teretnog prometa optimizirali su poslovanje tvrtki, čime se ostvaruje svrha Plave ekonomije (ekonomski razvoj) [31].

Na primjer, s pojavom Interneta stvari, ugrađeni sustavi u fizičkim objektima postali su spojeni na svjetsku komunikacijsku mrežu, pružajući podatke i primajući podatke s drugih uređaja i informacijskih sustava [24]. Nove tehnologije i koncepti donose društvene promjene i postaju važni u ekonomskim procesima [25], [23]. Prema istraživanju (Technology Industry Innovation) koje je proveo KPMG International Cooperative (2018) [26], 17% ispitanika u svijetu identificiralo je Internet stvari kao vodeću tehnologiju u pokretanju digitalne transformacije.



Slika 1. ClassNKNAPA GREEN softver

Uporabu najnovijih digitalnih tehnologija u pomorskom prometu [23] su kategorizirali u osam područja: autonomna vozila i robotika (*autonomous vehicles and robotics*); umjetna inteligencija; Veliki podaci (*Big Data*); virtualna stvarnost, proširena i mješovita stvarnost (*augmented and mixed reality*); Internet stvari; računalstvo u oblaku i na rubu (*the cloud and edge computing*); digitalna sigurnost (*digital security*); 3D tisak (*3D printing*) i aditivni inženjering (*additive engineering*). Prema njihovom istraživanju, kada se ispituju različiti industrijski sektori, morske su luke sektor koji pruža velike mogućnosti za studije. Nadalje, prema njihovom istraživanju, domene poput Umjetne inteligencije i Velikih podataka su detaljno obrađene dok su domene poput robotike i Interneta stvari, koje mogu imati značajan utjecaj na brodske operacije, nerazvijene i zaslužuju veću pozornost.

V. VAŽNOST DIGITALNE TRANSFORMACIJE U POMORSKOM TRANSPORTU I NJEZIN UTJECAJ NA PLAVU EKONOMIJU

Kao sto je već ranije spomenuto, pomorski transport predstavlja važnu aktivnost Plave ekonomije budući da uključuje brojne resurse i dionike. S razvojem međunarodne trgovine i globalizacijom, količina prometa i tereta, a samim time i informacija se povećava. Stoga su brojni dionici prisiljeni prihvatiti promjene i okrenuti se učinkovitijim praksama uvođenjem tehnologija koje mogu prikupljati i obrađivati velike količine informacija (na ekonomičan način) ili koje mogu pomoći u očuvanju ekosustava [28].

Digitalna transformacija u pomorskom transportu mogla bi značajno utjecati na Plavu ekonomiju, uključujući ekonomski razvoj, poboljšanje kvalitete življenja i rada te očuvanje ekosustava [10], a s obzirom na sve prethodno istaknute značajke pomorskog transporta. S obzirom na kontinuirani pritisak kojeg gospodarske aktivnosti generiraju na svjetske resurse i neophodnu potrebu za smanjenjem štetnih emisija, digitalna transformacija može pridonijeti održivom razvoju svjetske ekonomije. Procjenjuje se da će, ukoliko se ne pristupi digitalnoj transformaciji, postojeće poslovne prakse doprinijeti globalnom jazu od 8 milijardi metričkih tona između ponude i potražnje prirodnih resursa do 2030. godine što će dovesti do 4,5 bilijuna dolara izgubljenog ekonomskog rasta do 2030. godine [29].

Predviđa se da će digitalna transformacija pomorskog teretnog prometa zabilježiti stopu rasta od ~ 10% između 2019. i 2027. godine (CAGR *Compound Annual Growth Rate*- složena godišnja stopa rasta). Nadalje, očekuje se da će globalna digitalna transformacija pomorskog teretnog prometa dostići ~ 38,4 milijarde američkih dolara do 2027. godine u usporedbi s 18,2 milijuna USD u 2019. godini [31]. Povećanje broja trgovinskih sporazuma kao i veća važnost tehnologija kao što su Blockchain i umjetna inteligencija za bolje i fleksibilnije upravljanje trgovinom, potiču aktivniju digitalnu transformaciju pomorskog teretnog prometa [31].

Kao što je već ranije navedeno, nove tehnologije predstavljaju pokretače digitalne transformacije i trebale bi pomoći u procesu smanjenja štetnih emisija, ali i omogućiti pametne operacije, razmjenu informacija, integrirani logistički lanac, gdje je poslovanje sigurno i ekonomično [32]. Jedan od primjera takvih tehnologija koji utječe na smanjivanje troškova, a samim time pomaže i u ispunjavanje svrhe Plave ekonomije, jest i CargoX Blockchain platforma za stvaranje pametne teretnice (stvaranje teretnice u digitalnom obliku). CargoX omogućuje povezivanje prijevoznika, uvoznika, izvoznika i ostalih dionika u uravnotežen sustav. U tom pogledu, svrha CargoX jest uklanjanje potrebe za posrednikom, pružanjem alata za sigurno dijeljenje dokumenata, značajno smanjujući troškove i vrijeme isporuke, nudeći visoku razinu sigurnosti i transparentnosti [33], [34].

Veliki podaci su također jedna od tehnologija koja može utjecati na smanjenje troškova. Naime, brodovlasnici i brodari istražuju korištenje analitike Velikih podataka kako bi umanjili troškove goriva. Niski troškovi goriva mogu nadoknaditi niske cijene prijevoza na tržištu. Na primjer, ClassNKNAPA GREEN softver (Slika 1) koji su zajednički razvili ClassNK i NAPA nudi rješenje praćenja i optimizacije performansi u stvarnom

vremenu pomoću kojeg se može smanjiti potrošnja goriva [35].

Svrha Plave ekonomije nije samo ekonomski razvoj, nego i očuvanje ekosustava. Primjerice, smjernice Međunarodne pomorske organizacije po pitanju ograničenja emisije sumpora na brodovima za najmanje 50% do 2050. godine, potaknut će povećana ulaganja u tehnologije koje omogućuju dionicima da prate i analiziraju teretne brodove radi operativne održivosti i održivosti okoliša [36]. Osim toga, uz pomoć tehnologije kao što su Veliki podaci, brodari mogu odrediti optimalnu brzinu za potrošnju goriva, uzimajući u obzir čimbenike kao što su troškovi goriva tarife i redovi vožnje. Smanjenje brzine rezultirat će optimizacijom potrošnje goriva i značajno će pridonijeti smanjenju emisije štetnih plinova [37].

Važno je napomenuti kako digitalizacija i digitalne tehnologije olakšavaju suradnju među dionicima u pomorskom prometu, s obzirom da postoji veliki broj dokumenata koje je potrebno razmijeniti kako bi se izvršile usluge prijevoza [38], [39], [40]. Kada su sustavi upravljanja prometom učinkoviti, oni pružaju ekonomske prilike i koristi koje rezultiraju pozitivnim multiplikacijskim efektima, kao što su: bolji pristup tržištima, odnosno pozicioniranje na novim tržištima, zapošljavanje i dodatna ulaganja [41].

VI. ZAKLJUČAK

Važnost digitalne transformacije sve je više prepoznata u različitim industrijskim i društvenim domenama. S obzirom na postojanje različitih definicija, digitalnu transformaciju moguće je definirati kao transformaciju poslovanja nekog poduzeća uz pomoć novih tehnologija, ne bi li poduzeće ostvarilo veće prihode, odnosno poslovna poboljšanja. Nove tehnologije se stoga nazivaju i pokretačima digitalne transformacije, a neke od njih su: Internet stvari, umjetna inteligencija, proširena stvarnost, itd.

Svrha Plave ekonomije, kao koncepta održivog korištenja mora i oceana, jest ekonomski razvoj, poboljšanje kvalitete življenja i rada te očuvanje ekosustava. Upravo stoga su se autori u ovom radu usredotočili na ekonomski razvoj i očuvanje ekosustava.

Pod Plavu ekonomiju mogu se razvrstati nekoliko aktivnosti vezanih za eksploataciju mora i oceana: korištenje prirodnih živih resursa; korištenje prirodnih neživih resursa; transport i uslužne djelatnosti na moru, oceanima i obalnim područjima: pomorski transport i turističke aktivnosti; industrijske aktivnosti na obalnim područjima: izgradnja brodova i plutajućih objekata, izgradnja čamaca za zabavu i sport, brodski strojevi, brodska oprema, popravak i održavanje brodova i čamaca; očuvanje mora i oceana: monitoring mora i oceana, sekvestracija ugljika, zaštita obalnog područja i gospodarenje otpadom.

Pomorski transport predstavlja jednu od važnih aktivnosti Plave ekonomije s obzirom na brojne dionike i resurse. Zbog postojanja velikog broja dionika, kao i porasta prometa, javila se potreba za uvođenjem tehnologija koje omogućuju prikupljanje i obrađivanje velikih količina informacija na ekonomičan način, ali i tehnologija koje će omogućiti očuvanje ekosustava.

Osim naglašene važnosti razmjene informacija na ekonomičan način, uz smanjenje štetnih emisija, važno je smanjiti i zagađenja koja uzrokuju sami brodovi. Smjernice Međunarodne pomorske organizacije po pitanju ograničenja emisije sumpora na brodovima pomoći će u očuvanju ekosustava. Kako bi se to postiglo, bit će potrebno uložiti u tehnologije koje omogućuju dionicima praćenje i analizu teretnih brodova.

Iz svega navedenog moguće je zaključiti kako je Plava ekonomija usmjerena ka održivom gospodarstvu, a digitalna transformacija (odnosno digitalne tehnologije koje su glavni pokretači digitalne transformacije) omogućuju ispunjavanje svrhe Plave ekonomije.

LITERATURA

[1] S. W. K. van den Burg, J. Aguilar-Manjarrez, J. Jenness, and M. Torrie, "Assessment of the geographical potential for co-use of marine space, based on operational boundaries for Blue Growth sectors," *Mar. Policy*, vol. 100, no. June 2018, pp. 43–57, 2019.

[2] L. Ababouch, "Fisheries and Aquaculture in the Context of Blue Economy," *Feed. Africa*, vol. 2, no. 21-23 October, p. 13, 2015.

[3] The Commonwealth, "Blue economy," 2020. [Online]. Available: https://thecommonwealth.org/blue-economy. [Accessed: 19-Mar-2020].

[4] "The Blue Economy and Blue Growth | OpenChannels: Sustainable Ocean Management and Conservation," 2019. [Online]. Available: https://www.openchannels.org/top-lists/blue-economy-and-blue-growth. [Accessed: 06-Jun-2019].

[5] S. Niavis, T. Papatheochari, T. Kyratsoulis, and H. Coccossis, "Revealing the potential of maritime transport for 'Blue Economy' in the Adriatic-Ionian Region," *Case Stud. Transp. Policy*, vol. 5, no. 2, pp. 380–388, 2017.

[6] J. Jeansson and K. Bredmar, "Digital Transformation of SMEs: Capturing Complexity," in *32nd Bled eConference. Humanizing technology for a sustainable society*, 2019, pp. 523–541.

[7] J. Gray and B. Rumpe, "Models for the digital transformation," *Softw. Syst. Model.*, vol. 16, no. 2, pp. 307–308, 2017.

[8] C. Matt, T. Hess, and A. Benlian, "Digital Transformation Strategies," *Bus. Inf. Syst. Eng.*, vol. 57, no. 5, pp. 339–343, 2015.

[9] "Towards paperless transport within the EU and across its borders," 2018. [Online]. Available: http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=15358.

[10] The World Bank, "What is the Blue Economy," 2017. [Online]. Available: https://www.worldbank.org/en/news/infographic/2017/06/06/blue-economy. [Accessed: 20-Mar-2020].

[11] European Parliament, "Ocean governance and blue growth:

Challenges, opportunities and policy responses," 2019. [Online]. Available: https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI%282019%29635557. [Accessed: 23-Mar-2020].

[12] World Bank Group, "Blue Economy Development Framework; Growing the Blue Economy to Combat Poverty and Accelerate Prosperity," 2016. [Online]. Available: http://pubdocs.worldbank.org/en/446441473349079068/AMCOECC-Blue-Economy-Development-Framework.pdf. [Accessed: 20-Mar-2020].

[13] European Commission, "The EU Blue Economy Report 2019," 2019. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/676bbd4a-7dd9-11e9-9f05-01aa75ed71a1/language-en/. [Accessed: 22-Mar-2020].

[14] K. Hausberg, P., Liere-Netheler, K., Packmohr, S., Pakura, S., Vogelsang, "Digital Transformation in Business Research: A systematic literature review and analysis," *Druid18*, 2018.

[15] C. J. Saul and H. Gebauer, "Digital transformation as an enabler for advanced services in the sanitation sector," *Sustain.*, vol. 10, no. 3, pp. 1–18, 2018.

[16] I. Kovynyov and R. Mikut, "Digital Transformation in Airport Ground Operations," *NETNOMICS Econ. Res. Electron. Netw.*, no. March, pp. 1–30, 2018.

[17] M. H. Almasi, A. Sadollah, S. Kang, and M. R. Karim, "Optimization of an improved intermodal transit model equipped with feeder bus and railway systems using metaheuristics approaches," *Sustain.*, vol. 8, no. 6, 2016.

[18] S. Dimić, D. Pamučar, S. Ljubojević, and B. Dorović, "Strategic transport management models-the case study of an oil industry," *Sustain.*, vol. 8, no. 9, 2016.

[19] A. Khaslavskaya and V. Roso, "Outcome-Driven Supply Chain Perspectives on Dry Ports," *Sustainability*, vol. 11, no. 5, p. 1492, 2019.

[20] D. Naletina and E. Perkov, "The economic importance of maritime shipping with special reference on Croatia," 2017. [Online]. Available: https://www.researchgate.net/publication/324389577_THE_ECONOMIC_IMPORTANCE_OF_MARITIME_SHIPPING_WITH_SPECIAL_REFERENCE_ON_CROATIA. [Accessed: 20-Mar-2020].

[21] United Nations Conference on Trade and Development, "Review of Maritime Transport 2019," 2019. [Online]. Available: https://unctad.org/en/PublicationsLibrary/rmt2019_en.pdf. [Accessed: 20-Mar-2020].

[22] Eurostat Press Office, "World Maritime Day - Half of EU trade in goods is carried by sea - Rotterdam, Antwerp and Hamburg busiest cargo ports," *News release 184*, no. September, pp. 0–3, 2016.

[23] M. Fruth and F. Teuteberg, "Digitization in maritime logistics—What is there and what is missing?," *Cogent Bus. Manag.*, vol. 4, no. 1, Nov. 2017.

[24] D. Beverungen, O. Müller, M. Matzner, J. Mendling, and J. vom Brocke, "Conceptualizing smart service systems," *Electron. Mark.*, vol. 29, no. 1, pp. 7–18, Mar. 2019.

[25] J. Vrchota, M. Mařiková, P. Řehoř, L. Rolínek, and R. Toušek, "Human Resources Readiness for Industry 4.0," *Journal of Open Innovation: Technology, Market, and Complexity*, 2019. [Online]. Available: https://www.mdpi.com/2199-8531/6/1/3. [Accessed: 12-Dec-2019].

[26] KPMG, "The Changing Landscape of Disruptive Technologies: Tech disruptors outpace the competition," 2018. [Online]. Available: https://assets.kpmg/content/dam/kpmg/pl/pdf/2018/06/pl-The-Changing-Landscape-of-Disruptive-Technologies-2018.pdf. [Accessed: 09-Sep-2019].

[27] "IDC Forecasts Worldwide Spending on the Internet of Things to Reach $745 Billion in 2019, Led by the Manufacturing, Consumer, Transportation, and Utilities Sectors," 2019. [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=prUS44596319. [Accessed: 09-Sep-2019].

[28] Marshall University, "Big Data, Big Data Analytics," 2019. [Online]. Available: https://www.marshall.edu/bigdata/. [Accessed: 23-Aug-2019].

[29] World Economic Forum, "How can digital enable the transition to a more sustainable world?" [Online]. Available: http://reports.weforum.org/digital-transformation/enabling-the-transition-to-a-sustainable-world/. [Accessed: 27-Mar-2020].

[30] International transportation, "Maritime industry navigates digital transformation to reshape supply chain." [Online]. Available: https://www.internationales-verkehrswesen.de/maritime-industry-report/. [Accessed: 27-Mar-2020].

[31] Transparency Market Research, "Digital Transformation of Maritime Freight Market." [Online]. Available: https://www.transparencymarketresearch.com/digital-transformation-of-maritime-freight-market.html. [Accessed: 20-Mar-2020].

[32] Springer International Publishing, *Sustainable Shipping*. Springer Nature Switzerland AG, 2019.

[33] "From the Past to the Future," 2019. [Online]. Available: https://cargox.io/welcome/. [Accessed: 05-Jul-2019].

[34] M. Jović, M. Filipović, E. Tijan, and M. Jardas, "A Review of Blockchain Technology Implementation in Shipping Industry," *Sci. J. Marit. Res.*, vol. 33, no. 2, pp. 140–148, 2019.

[35] Trelleborg Marine Systems, "Use of big data in the maritime industry," 2018. [Online]. Available: https://www.patersonsimons.com/wp-content/uploads/2018/06/TMS_SmartPort_InsightBee_Report-to-GUIDE_01.02.18.pdf. [Accessed: 06-Aug-2019].

[36] R. Ascutia, "Digital transformation of maritime shipping accelerating, says report," 2019. [Online]. Available: https://www.portcalls.com/digital-transformation-of-maritime-shipping-accelerating-says-report/.

[37] M. Zghurovsky and Y. P. Zaychenko, *Big Data: Conceptual Analysis and Applications*. 2019.

[38] M. Jović, N. Kavran, S. Aksentijević, and E. Tijan, "The Transition of Croatian Seaports into Smart Ports," *MIPRO Digit. Econ. Soc. Proc. 2019*, Feb. 2019.

[39] E. Tijan, A. Agatić, M. Jović, and S. Aksentijević, "Maritime National Single Window — A Prerequisite for Sustainable Seaport Business," *MDPI Sustain.*, vol. 11, no. 17, pp. 1–21, 2019.

[40] E. Tijan, M. Jović, and P. Karanikić, "Economic and ecological aspects of electronic Transportation Management Systems in seaports," *Proceedings of the Maritime and Port Logistics Bar Conference*, 2019. [Online]. Available: https://www.bib.irb.hr/1003853. [Accessed: 31-Dec-2019].

[41] M. A. Mosaberpanah and S. D. Khales, "The Role of Transportation in Sustainable Development," no. September 2013, pp. 441–448, 2013.

Specifičnosti Uberovog modela prijevoza s primjenom u Republici Hrvatskoj

Kristian Jugović*, Alen Jugović* i Svjetlana Hess*
* Sveučilište u Rijeci/Pomorski fakultet, Rijeka, Hrvatska
jugovickristian@outlook.com, ajugovic@pfri.hr, shess@pfri.hr

***Sažetak* - Konvencionalne taksi službe u Hrvatskoj posluju više od sto godina, a 2015. godine se na hrvatskom taksi tržištu pojavio i Uber, višenamjenska platforma namijenjena prijevozu putnika koja funkcionira putem mobilne aplikacije. Korištenjem mobilnih aplikacija uz fleksibilnost taksi službe postiže se jednostavnije, brže i modernije obavljanje prijevoza putnika, a to je posebno značajno u vremenu kada se sve više ljudi koristi raznim načinima gradskih i međugradskih usluga prijevoza. U ovom se radu posebno analizira Uber kao prijevoznička taksi služba te njen utjecaj na cijenu i kvalitetu taksi usluga. Ujedno su istaknuti segmenti koji čine Uber drugačijim od ostalih ponuđača taksi prijevozne usluge te način na koji tvrtka djeluje na klasično taksi tržište, prvenstveno s aspekta utjecaja na korisničke navike i konkurenciju.**

Ključne riječi - organizacija taksi usluga; Uber; prijevoz; tržište

I. UVOD

Analiza utjecaja usluge tvrtke Uber na tržište prijevoznih usluga, usporedba sudionika i utjecajnih čimbenika te osvrt na smjer razvitka ovakvog načina usluživanja korisnika osnovni je cilj ovog rada. Uber je američka tehnološka tvrtka osnovana 2009. godine s ciljem razvoja aplikacije kojom bi se mogao naručiti prijevoz a ova je ideja zaživjela 2011. godine. Putem aplikacije se moglo zatražiti luksuzne crne automobile u nekoliko gradskih područja u San Franciscu (sjedištu Uber-a) dok danas ta aplikacija predstavlja složeni logistički sustav koji rabi mobilnu tehnologiju. Tvrtka je prisutna u preko 600 gradova u cijelom svijetu. Službeno su Uber-ove usluge i aplikacija postale dostupne od 2011. godine i tadašnja aplikacija je imala jednu opciju, a ta je opcija bila narudžba crnog luksuznoga automobila čija cijena je bila 1,5 puta veća od cijene taksija. Ta je ideja dovela do toga da je danas Uber najpoznatiji taksi brand te jedna od najuspješnijih i najbrže rastućih start-up tvrtki u svijetu. Treba u tome istaknuti i značajnu ulogu aplikacije koja se koristi na pametnim telefonima jer upravo ona mijenja način na koji se razvija ovakva vrsta tržišta. U drugom dijelu rada postavlja se pitanje da li je ova usluga legitimna te kako na to gledaju građani i konkurencija, a ponajviše zakonska regulativa. Jedan od rezultata razvitka ovakvih vrsta usluga su efekti koje digitalne platforme stvaraju u razvijanju digitalne ekonomije kao nove velike grane razvijanja tehnoloških sektora.

II. PREGLED RAZVOJA USLUGE UBER

Prema istraživanju autora Clewlow i Mishra, prijevoz na poziv narastao je na oko 250 milijuna korisnika u prvih pet godina a ujedno navode kako u većim gradovima 21% odraslih korisnika koristi ovakvu vrstu vožnje, dok dodatnih 9% koristi uslugu sa prijateljima iako nemaju instaliranu aplikaciju [1]. Mislilo se da je najčešći razlog tome izbjegavanje vožnje osobnim automobilom zbog alkoholiziranoga stanja osoba, ali prema podacima kompanije Uber na popularnost usluge je najviše utjecala mobilna aplikacija putem koji korisnici potražuju putovanje i odabiru vrstu usluge.

Svoju veliku popularnost Uber duguje inovativnoj aplikaciji koja je i temelj poslovanja kompanije. Karakterizira je jednostavno korisničko sučelje sa svim najnovijim informacijama i značajkama jednostavnim za shvaćanje krajnjem korisniku. Platforma je dizajnirana tako da korisnik ima apsolutno sve potrebne podatke na jednom mjestu te je na taj način aplikacija upotrebljiva svim uzrastima. Uz jednostavno sučelje kvaliteta i cijena usluge su dodatni parametri koji postavljaju tvrtku na prijevozničko tržište iznad drugih klasičnih prijevoznika. Uber povećava udobnost i smanjuje troškove prijevoza taksijem, pa tako jedna od usluga UberX osigurava čak 20-30% manje cijene u odnosu na tradicionalne taksi usluge [2]. Wallstem [3] pronalazi dokaze koji ukazuju da se dolaskom Ubera na tržište smanjio broj pritužbi za taksi usluge po putovanju.

Sukladno periodu u kojem je pokrenuta, uslugu ponajviše koristi generacija Y i Z. Generacija Y ili često nazivana "*Millenials*", trenutno je najveći korisnik Ubera. "Milenijalci" su kao potrošači puno drugačiji nego prijašnje generacije, oni nisu zainteresirani za fizičko posjedovanje stvari, ta skupina samo želi pristup stvarima. Oni ne kupuju glazbu, "stremaju je" (stream predstavlja reproduciranje sadržaja multimedije uz plaćanje mjesečne naknadu za korištenje usluge), oni ne kupuju automobile već su ljubitelji tzv. "ride hailing-a (poglavlje IV.). Oni su ljudi okrenuti internetu [4]. Uzevši to u obzir Cannon i Summers [5] ističu da je "*ride hailing*" oborio klasičnu taksi industriju koja je tradicionalno pod velikom regulativom.

III. OSNOVNE ZNAČAJKE KLASIČNE TAKSI USLUGE

Zakon ponude i potražnje najbitniji je sa svakoga ekonomskoga aspekta. Naime bez jednog ne postoji drugo niti obrnuto. Prometna ponuda je ukupna prometna

infrastruktura i suprastruktura koja stoji na raspolaganju za obavljanje prometnih usluga. Prometna potražnja za prijevoznim uslugama je ukupnost zahtjeva koje korisnici usluga žele i mogu ostvariti uz određenu cijenu i u određenom razdoblju [6]. Usluga klasičnog taxi prijevoza koja je započela još početkom 17. stoljeća uvelike je promijenila način kretanja ljudi i dobara te počela definirati drugačiji logistički sistem. Prijevozništvo u obliku taksi usluge postalo je bitna komponenta u životu ljudi u smislu obavljanja svakodnevnih zadataka i jednostavnog prijevoza od točke A do točke B. Za većinu svjetskih tržišta smatra se da je prijevoznička industrija strogo regulirana jer je određeno mnogo detalja oko načina na koji davatelj usluge može dobiti dozvolu za rad, koju cijenu može postaviti, koje korisnike primiti, kakve sve standarde sigurnosti mora zadovoljiti i sl. [7]. Standard sigurnosti u ovoj industriji odnosi se ponajviše na zaštitu korisnika kroz ispravnost vozila kojima davatelj usluga obavlja djelatnost, a to su odrednice za automobil kao što su visina, duljina, godište [8]. Kroz gledište reguliranja cijena Zuluaga D. [9] iznosi da se cijene prilagođavaju očekivanom prometu, dobu dana, raspoloživom broju vozača te potražnji putnika. Ovo je učinkovit sustav jer promjenom cijena se olakšava privremeni višak opskrbe ili nestašica. Također olakšava konkurenciju između pružatelja usluga. Najupečatljivija značajka taksi usluge osim propisanog svjetlećeg znaka na krovu vozila je uređaj za mjerenje, točnije taksimetar. Prema Hrvatskom zavodu za mjeriteljstvo svaki prijevoznik u taksi industriju dužan je imati ovaj uređaj upaljen tokom pružanja usluge. Taksimetar ugrađen u taksi vozilo na temelju signala generatora električnih impulsa proporcionalnih prijeđenom putu izračunava prijeđeni put i na temelju prijeđenog puta, dodataka za prtljagu i /ili vremena prema važećoj tarifi izračunava i pokazuje cijenu vožnje [10].

U novije vrijeme proveden je i dio deregulacije odnosno smanjenog uplitanja zakona u poduzetničku sferu ponajviše u pogledu liberalizacije tržišta. Tako je, primjerice, u Zagrebu 2010. godine gradska skupština omogućila i drugim firmama obavljanje prijevozničke usluge bez potrebe izdavanja koncesije. Takav potez postavio je temelj za konkurentno tržište koje za svrhu ima povećanje kvalitete usluge i manje cijene.

IV. ZNAČAJKE "RIDE HAILING-A"

Usluga "*Ride hailing-a*" se može definirati kao aktivnost potražnje za vozilom sa vozačem u svrhu prijevoza na neko određeno mjesto u nekom određenom vremenu [11]. U većim gradovima sa razvijenim javnim prijevozom dolazi do problema koji se naziva "problem zadnje milje" koji predstavlja poteškoću dostizanja do krajnjega odredišta, odnosno od zadnje stanice ili neke druge lokacije do krajnjeg odredišta korisnika kao što su dom, posao i sl. [12]. Naime zbog postojanja usluge kakvu nude Uber i slični prijevoznici (Lyft i dr.) ljudi su skloniji upotrebljavati javni prijevoz u kombinaciji sa takvom vrstom usluge jer im donosi rješenje problema zadnje milje [13]. Usluge na zahtjev zamjenjuju tradicionalne poslovne modele brže nego što je itko mogao predvidjeti, a omogućuju trenutno pružanje roba i usluga. Ono što je počelo kao najbrži način da se pozove taksi, tzv. uberizacija, danas je promijenilo način na koji tvrtke percipiraju potražnju, osiguravaju ponudu i istu isporučuju. U svakom danom trenutku korisnik ima mogućnost odabira vrste usluge koja mu je potrebna i koja mu odgovara, a za to plaća naknadu. Tim pristupom, iskorištavanjem tuđega dobra uz određenu naknadu dolazi se do reduciranja broja vozila na nekome području.

Usluge na zahtjev temelj su nove ekonomije, ubrzane ekonomije. Usluge na zahtjev su preokrenule cjelokupnu industriju vezanu uz način na koji korisnik dolazi do proizvoda ili usluge. Kompanije koju su se prilagodili takvom načinu rada doživjele su veliki rast u usporedbi na one kompanije koje takav način nisu još preuzele. Ključ uspjeha ovakvog poslovnog modela je i dostupnost pametnih telefona i internet usluga u svijetu. Pametni telefoni nikad nisu bili dostupniji široj populaciji, njihova cijena tokom godina toliko je pala da se pametni telefoni nalaze gotovo u svakom domu. Na taj način proizvod i usluga dostupni su 24 sata na dan tijekom cijele godine. Iznimno lako korištenje, korisničko sučelje koje stvara dobro, lako i zabavno iskustvo eliminirajući pritom bilo kakve prepreke između interakcije korisnika i usluge te konstanta poboljšanja ključ su uspjeha kod stvaranja takve vrste usluge. Usluge na zahtjev funkcioniraju na tri temeljna načela: (1) privlači ih prvenstveno relativno niska cijena jer su za pružatelje usluga uklonjeni nepotrebni posredni koraci, (2) velika brzina pružanja potrebnih usluga ili robe i (3) lakoća kontaktiranja korisnika od strane pružatelja usluga. Pojačana konkurencija jedan je od najvećih uzroka promjena, a očekuje se ogromno povećanje konkurencije s ulaskom novih igrača na tržište ili korištenjem alternativnih poslovnih struktura. Ono što je potaknulo upravo odabir ovih servisa leži u samoj inovativnosti, jednostavnosti i transparentnosti koju su tehnološke kompanije postavile kao kriterij i bazu za privlačenje korisnika. Glavna odlika toga je informacijsko-komunikacijska tehnologija koja predstavlja temelj poslovanja ovih kompanija.

Uber Technologies Inc. 2011. godine prvi je pokrenuo trend naručivanja taksi usluge putem aplikacije koja je dostupna na više mobilnih platformi te time omogućila najjednostavniji pristup toj usluzi. Pristizanjem "*ride hailing*" servisa, transparentnost i otvorenost postali su glavni kriteriji koji korisnici preferiraju. Konačno, nakon godina postojanja uvedena je GPS tehnologija u prijevoz, tako da korisnik zna sve i nije u mogućnosti da bude prevaren odnosno da podaci budi manipulirani od strane prijevoznika. Takav način rada oduševio je korisnike Ubera te su se oni masovno priklonili takvoj vrsti prijevoza što je uzrokovalo pad za potražnjom kod klasičnih taksi prijevoznika koji su praktički imali monopol ili duopol. Tržište je dolaskom servisa postalo više konkurentnije te je na kraju korisnik od toga najviše prihodovao. Nadalje, jedno od najbitnijih načela je i recenzija usluge, dobar "glas" je ono čime vozači "zaslužuju" vožnje. Prema tome, da bi vozač mogao voziti mora imati dobre ocjene jer u protivnom neće moći dobivati vožnje tj. sustav mu ih neće prosljeđivati. U interesu je tvrtke da svaka usluga bude na najvišem nivou, samim time imaju i zahtjeve od vozača koji su bitan čimbenik te usluge budući da posluju i pod imenom kompanije. Vozače se ocjenjuje na temelju više kriterija:

komunikativnost, humor, kvaliteta usluge (vozačke sposobnosti) te mnoge druge pozitivne stvari koji usluge čine hvalevrijednom. Vozača se također može ocijeniti i lošom ocjenom što, ako je više takvih ocjena, neće dobro utjecati po vozača jer je firma postavila standard po kojem vozač mora imati minimalnu ocjenu 4,6 kako bi nastavio primati vožnje. U suprotnom sa manjom ocjenom od 4,6 vozač će biti sankcioniran i neko vrijeme neće moći obavljati uslugu prijevoza za Uber. Na taj način kroz aplikaciju korisnik ima mogućnost poboljšavati kvalitetu usluge i kroz komentare je na neki način pomoći unaprijediti. Da bi usluga mogla biti isplativa i nuđena po cijeni na način da bude konkurentnija od klasičnih taksi prijevoznika, a i rasprostranjenija potrebno je provesti znatno ublažavanje ograničenja i regulacija uslugama kao što su Uber i slične platforme [14]. To sugerira da takve tvrtke imaju prednost u troškovima u odnosu na tradicionalne tvrtke jer uspješno izbjegavaju skupe regulacije. Ispunjavanje takvih zahtjeva je skupo što se na kraju odražava i na cijenu usluge. Kombinacijom inovativnih tehnologija, stvaranjem novih trendova i usluga, mogućnost prisutnosti na svim mobilnim uređajima te zaobilaženja skupih i striktnih regulacija ovakav poslovan model čini konkurentnijim, jeftinijim, privlačnim i rasprostranjenijim.

Uber je dolaskom na tržište Hrvatske napravio velike razlike između usluga tradicionalnih taksi prijevoznika i samoga sebe. Naime, kompanija ima izrazito otvoren pristup sa tarifama te je vrlo privlačna mladima koji uz tehnologiju imaju jako lagan dostup funkciji. Jedan od bitnih čimbenika uz cijenu je i transparentnost. Kako bi kompanije poslovala legalno prema hrvatskom Zakonu o fiskalizaciji bila je potrebna jedinstvena preinaka u samoj aplikaciji. Naime, sustav odnosno aplikacija je spojena na poreznu službu te je prijevoznik prihvaćanjem vožnje dužan izdati račun koji je sa plaćanjem odmah fiskaliziran. Na taj način omogućen je legalan način rada na teritoriju Republike Hrvatske. Ulaskom na tržište kompanija je praktički potresla cijeli prijevozнički sektor svojim novim i inovativnim poslovnim modelom koji se temeljio na jednostavnosti, transparentnosti i efikasnosti. Uber je naime prva kompanija koje je unijela te segmente na taksi tržište. Do tada svaka usluga bila je vezana za taksimetar, pozivanje na telefon i na neutemeljene rute koje su korištene na način da se pokuša maksimizirati profit. Uvođenje tehnologije baziranoj na GPS-u ili Galileu uvelike se promijenio način na koji se korisnici oslanjaju na korištenje usluge time što se iz korijena izmijenio način obračuna cijena te na koji se način one definiraju a i ukazane su korisniku prije konzumiranja usluge. Na taj se način napravio odskok u percepciji na koji korisnik sagledava uslugu i kako joj prilazi.

Način formiranja cijena, koji ujedno predstavlja uspješnost poslovanja, izrazito je bitan čimbenik poslovanja. Cijene se mogu definirati tako da se uzme niz čimbenika koji utječu na profitabilnost te se uvrštavanjem tih čimbenika tj. troškova dolazi do završne cijene koju definira davatelj usluge. Cijena se formira tako da bude konkurentna, dostupna ali i profitabilna. S obzirom na gotovo konstantno povećanje cijena Zakon o prijevozu u cestovnom prometu [15] bi trebao dovesti do liberalizacije taksi tržišta, odnosno do bolje usluge za korisnike. Kilometri naravno nisu jedini čimbenik formiranja cijena, kod Ubera je izrazito izražen i prije spomenuti zakon ponude i potražnje. U samoj osnovi tvrtka djeluje savršeno u pogledu ekonomista što se tiče cijena. Cijena je u njihovom slučaju čisti produkt ponude i potražnje. Kada je vozača malo, a upita odnosno korisnika puno cijena se podiže, ukoliko je vozača od kosnika više tada cijena pada. Uber funkcionira na način frekventnosti, to znači da algoritam ovisno o ponudi i potražnji stalno formira drugačiju cijenu koja je dostupna unutar aplikacije. Cijena se mijenja s obzirom na rast ili pad potražnje za prijevozom i brojem vozila koji se nalaze u trenutnom području. Uber je prošao kroz velike promjene na koje je morao pristati da postane legalan u Hrvatskoj. Naime, s novim Zakonom o prijevozu u cestovnom prometu [15], kao i Pravilnikom o posebnim uvjetima za vozila kojima se obavlja javni cestovni prijevoz i prijevoz za vlastite potrebe [16], Uber se po pitanju statusa izjednačio s hrvatskim taksi službama.

V. SWOT ANALIZA

Korištenjem alata swot (strenght-snaga, weaknesses-slabosti, opportunities – prilike, threats – prijetnje) analize navode se svi aspekti koji bi mogli utjecati na poslovanje kompanije i to tako da su prikazani parametri na koje bi se trebalo fokusirati i voditi poslovanje sukladno smjernicama koje analiza daje. Takvim načinom fragmentiranja čimbenika na manje dijelove omogućen je uvid u trenutnu situaciju pa se preciznije mogu donijeti odluke prilikom razmišljanja o ulasku ili ekspanziji pojedine usluge na određenom tržištu. Sa swot analizom mogu se raščlaniti i preciznije analizirati stavke koje pogoduju i one koje eventualno odmažu u poboljšanju poslovanja i stvaranja nove vrste usluge, te dobiti smjernice za privlačenje novih korisnika. Za usluge Ubera analiziraju se nadalje četiri osnovne kategorije: snaga, slabosti, prilike i prijetnje.

SNAGA

- Prepoznatljivost branda
- Kvaliteta usluge
- Kartično plaćanje
- Jednostavno povećanje flote
- Inovativnost
- Mala odgovornost prema kooperantima
- Brza prilagodba novim tržištima
- Niska cijena usluge
- Upotreba i fleksibilnost

Prepoznatljivost branda bitan je čimbenik koji utječe na voljnost korisnika da koristi određenu uslugu i uzme u obzir korištenje usluge kompanije prije ostalih operatora. Kvaliteta usluge najbitniji je čimbenik koji korisnik analizira prilikom korištenja usluge gdje uspoređuje vrijednost svoga novca sa vrijednosti usluge koju plaća, tako da održavanje kvalitete uvelike pospješuje poziciju na tržištu. Laka upotreba svodi se na jednostavno

korištenje aplikacije koja stavlja uslugu na raspolaganje svim korisnicima koji posjeduju pametni telefon.

SLABOSTI

- Ideja je lako primjenjiva te ju je lako kopirati
- Direktan odnos kompanije sa korisnicima odvija se preko kooperanata
- Model je nepredvidiv
- Privatnost korisnika je upitna
- Odnos kompanije prema kooperantima
- Održavanje flote vozačima je teže zbog niskih cijena
- Otkazivanje vožnje od strane vozača
- Skokovi u cijeni

Mogućnost lakoga kopiranja omogućava ostalim sudionicima implementaciju istog načina usluge u svoje poslovanje te potencijalno oduzimanje tržišnog udjela. Održavanjem odnosa sa klijentima preko kooperanata može predstavljati problem na način da reputacija i kvaliteta ovise o osobi koja nije direktan zaposlenik tvrtke te ju je moguće lakše narušiti.

PRILIKE

- Razvoj električnih i autonomnih vozila omogućuje smanjenje troškova
- Usluga se može lako proširiti na dijelove gdje klasični taksi prijevoznici ne mogu doći
- Lakši pristup stranom tržištu
- Omogućavanje modela prijevoza bolesnika i ostalih posebnih vrsti prijevoza može utjecat na povećanje profita
- Uvođenje pretplate i posebne pogodnosti za stalne korisnike
- Marketing i PR (public relation – odnos sa javnošću)

Ulaganjima koja se odnose na istraživanja u sektoru autonomnih vozila pomogla bi reducirati potrebu za ljudskim kadrom, vozačima. Širenje spektra usluge moguće je putem osiguravajućih društava ili putem drugih kanala u svrhu stvaranja jedinstvene vrste usluge vezane isključivo za zdravstvo ili zdravstveni turizam. Mogućnost uvođenja pretplate predstavlja priliku za konstantniji pritok novčanih sredstava

PRIJETNJE

- Prosvjedi vozača radi malih primanja mogu obeshrabriti dolazak novih vozača i socijalno osjetljivih korisnika
- Povećanje konkurentnosti reflektira manjim profitom
- Deregulacija od strane države može izazvati krucijalne promjene načina operiranja servisa pa i njegovo gašenje
- Loši vozači štete brandu
- Nezakonite radnje kooperanata
- Presude suda da su vozači zaposlenici a ne kooperanti uzrokovale bi drastično povećanje troškova i odgovornosti
- Razvoj autonomnih vozila drugih kompanija

Kooperanti kao prijevoznici nemaju ugovor sa kompanijom te se na njih ne odnose određena prava koja bi imali kao zaposlenici te postoji opasnost od štrajkova i odstupanja od korištenja usluge sa strane socio-ekonomskih osjetljivih osoba. Regulacija je temelj uređivanja tržišta te je kao takav važan čimbenik koji određuje na koji način kompanija posluje sa kakvim obvezama prema strankama i državi. Vozač je dio usluge te on svojim ponašanjem može promijeniti percepciju kakvu korisnik ima prema samome brandu i usluzi.

Zaključak swot analize ukazuje da bez obzira na prikazane slabosti i prijetnje, kao što su lakoća kopiranja i implementacije ideje, održavanje odnosa, mogućnosti nezakonitih i neetičnih radnji od strane kooperanata te nezadovoljstva vozača, postoje i brojne snage i prilike, primjerice jaka prepoznatljivost branda, inovativnost koju kompanija implementira u svoje usluge, budući tehnološki ciljevi i velike prilike za proširenje usluga. Širok je spektar mogućnosti koje tvrtka ima u smislu osmišljavanja i stvaranja usluga te širenja svoga poslovanja na različite sfere kako bi olakšala i na neki način unaprijedila standard života sadašnjih i potencijalno novih korisnika.

VI. ZAKLJUČAK

Uber je danas prepoznatljiv kao jedna od najuspješnijih i najbrže rastućih start-up tvrtki u svijetu. Cijena usluge ovisi o potražnji, vremenu i udaljenosti vožnje te varira tijekom povećane potražnje. ITS mobilna aplikacija koja korisnicima omogućava lakše i brže naručivanje vozila nego što to pružaju taksi prijevoznici, pruža i veću transparentnost te bolji uvid u podatke o načinu funkcioniranja usluge aludirajući time na cijenu prijevoza i prijevoznika koji izvršava uslugu. Ovakav način plasiranja na tržište dovodi do mogućnosti bržeg i efikasnijeg načina stvaranja trendova i promjena, bolje konkurentnosti, manjih cijena, više inovacija u raznim područjima te ostvarivanje bolje povezanosti sa korisnikom. Takva globalizacija usluga stvara izazove kako u domaćoj tako i globalnoj ekonomiji, ali i politici kao regulatoru zaduženom za jednako tržišno natjecanje svih sudionika po istim pravilima. Regulativa mora biti tako usmjerena da pruža mogućnost tvrtkama da budu inovativne i kreativne u kreiranju usluga kako bi krajnji korisnik imao korist od toga uzimajući u obzir i razvijanje društva u nekom daljem pogledu posebice u zaštiti socijalnih i radnih prava. U Hrvatskoj se Uber od 2018. godine, donošenjem Zakona o prijevozu u cestovnom prometu, po pitanju statusa izjednačio s hrvatskim taksi službama. Uber je bitan segment razvoja društva, protoka informacija i povezivanja ljudi. Tvrtka koja uvijek teži daljnjem razvoju ne prestaje iznenađivati svoje korisnike kvalitetnim i konstruktivnim promjenama koje donosi. Ulaganjem u čovjeka i njegove ideje te implementacijom tehnologije mogu se ostvariti i najzahtjevniji ciljevi koje

tržište postavlja, a to je stalno zadovoljavanje ponude i potražnje.

LITERATURA

[1] R. R. Clewlow and G. S. Mishra, " Disruptive transportation: the adoption, utilization, and impacts of ride-hailing in the United States", Institute of Transportation Studies, University of California, Davis, 2017.

[2] https://www.uber.com, (31.01.2020.)

[3] S. Wallsten, "The competitive efects of the sharing economy: how is Uber changing taxis?" Technology Policy Institute, 2015.

[4] B. N. Greenwood and S. Wattal, "Show me the way to go home: an empirical investigation of ride sharing and alcohol related motor vehicle homicide, MIS Quarterly, Vol. 41, 2017, pp.163–188.

[5] S. Cannon and L. H. Summers, "How Uber and the sharinge economy can win over regulators", Harvard Business Review, October 13, 2014; dostupno na https://hbr.org/2014/10/how-uber-and-the-sharing-economy-can-win-over-regulators

[6] S. Hess, Planiranje prometne potražnje, Pomorski fakultet u Rijeci, Rijeka, 2010.

[7] J. P. Toner, "Regulation in the taxi industry", working paper 381, Institute of Transport Studies, University of Leeds, Leeds, UK, 1992.

[8] http://www.podaci.net/_gHRV/propis/Pravilnik_o_mjeriteljskim/P-mtztut03v1158.html (15.01.2020.)

[9] D. Zuluaga, "Taxi and private hire vehicle regulation: a briefing", Institute for Economic Affairs, 2017.

[10] https://dictionary.cambridge.org/dictionary/english/ride-hailing (02.02.2020.)

[11] N. Sadowsky and E. Nelson, "The impact of ride-hailing services on public transportation use: a discontinuity regression analysis", Economics Department Working Paper Series. 13., 2017.; dostupno na https://digitalcommons.bowdoin.edu/econpapers/13

[12] H. Wang and A. Odoni, "Approximating the performance of a "last mile" transportation system", Transportation Science 50(2), 2014. pp.659-675.

[13] K. Gyódi, "Uber vs licensed taxi drivers: a war between technological standards", University of Warsaw, Faculty of Economic Sciences, Digital Economy Lab, No. 2, 2017.

[14] S. Silverstein, "Everything you need to know about Uber prices in 21 cities", Business Insider, 16th October 2014; dostupno na: http://www.businessinsider.com/uber-vs-taxi-pricing-by-city-2014-10

[15] Narodne novine, 41/2018

[16] Narodne novine, 50/2018

ICTLAW

International Conference on

INFORMATION AND COMMUNICATION TECHNOLOGY LAW

Steering Committee

Chairs: Tihomir Katulić, University of Zagreb, Croatia
Hrvoje Lisičar, University of Zagreb, Croatia

Members: Dražen Dragičević, University of Zagreb, Croatia
Nina Gumzej, University of Zagreb, Croatia
Marko Jurić, University of Zagreb, Croatia
Tihomir Katulić, University of Zagreb, Croatia
Hrvoje Lisičar, University of Zagreb, Croatia
Lucija Vejmelka, University of Zagreb, Croatia
Goran Vojković, University of Zagreb, Croatia

Program Committee

Members: Tihomir Katulić, University of Zagreb, Croatia
Hrvoje Lisičar, University of Zagreb, Croatia
Marko Jurić, University of Zagreb, Croatia

PAPERS

Application of the General Data Protection Regulation in Schools: A Qualitative Study with Teachers, Professional Associates and Principals

L. Vejmelka*, T. Katulić * M. Jurić* i M. Lakatoš**
* University of Zagreb, Faculty of Law, Zagreb, Croatia
** URIHO- Foundation for Vocational rehabilitation and Employment of Disabled Perosns, Zagreb, Croatia
lvejmelka@pravo.hr, tkatulic@pravo.hr, mjuric@pravo.hr, mia.lakatos1@gmail.com

***Abstract*- Current changes in legislation and the GDPR implementation challenges require specific compliance efforts of educational children institutions. The aim of this qualitative research is to find out the opinion of teachers, professional associates and principals on the implementation of the GDPR and issues these institutions face. Data were collected using the focus group method (4 in total) and analyzed with a simple content analysis. The results show that teachers and practitioners generally are not familiar with application of the GDPR in school settings, while the experiences of the principals are different. Some of them are fully or partially informed and part of them are not familiar with the application of the GDPR in schools. A part of the participants expressed the need for additional training in this regard. The findings of the research contribute to the understanding of experts from schools on the application of the Regulation and the national law and provide insight into their experiences in the implementation of the new framework. The results of the research can contribute to the development of educational content for professionals in educational settings that would facilitate the application of the GDPR precisely in its application to children, a vulnerable group with their specific needs.**
Key Words – qualitative research, GDPR, schools

I. INTRODUCTION

Children are a particularly vulnerable group in society and their well-being in different domains is in the focus of the researchers and professionals of various profiles [1,2]. Special attention is directed towards legal protection of children rights. The Convention on the Rights of the Child, a universal and widely used document on children's rights in the world, in Croatia ratified in 1991, in Article 16 explicitly stipulates that no child shall be subjected to arbitrary or unlawful interference with his or her privacy. Safety and privacy are a universal rights of the child, and these issues should be taken into account especially by those who care for the children on a daily basis as institutions within the educational sector. Article 3 of the Convention obliges private and public services and institutions to harmonize their activities with the standards of the competent authorities, including the one on security matters [3].

Advent of information society services such as search engines, social networks, content streaming providers and other forms of electronic commerce and online communication has had a profound influence on societal development, allowing the faster closure of the digital divide and access to education and business services and opportunities unheard of even a decade ago. These days, as higher education is becoming a service to be disrupted, even the elementary and secondary levels of education, traditionally a purview of government are feeling the effects of changes information technology is causing in the fabric of the society. [4]

The rise of the postindustrial, information society on the back of the Internet revolution has not been without new and profound challenges and dangers, especially to the most sensitive societal groups – children and the elderly – who are struggling to grasp the rules of the new online environment, an environment which is increasingly relying on collecting and processing personal data both as the resource and as the commodity that fuels the development of new innovative products and services.

Where the traditional model of paying for products and services such as software, processing power or online platform service used to be an online financial transaction, the new Internet economy is increasingly relying on exchanging products and services for users' personal data which is then monetized through various online behavioural marketing schemes. [5]

This transition into a personal data driven economy is especially perilous for the societies' most sensitive groups which by themselves will not be able to recognize and stop harmful practices such as overcollection of personal data or use of data for incompatible secondary purposes. Often these subjects are unable by themselves to seek effective protection of their rights and freedoms through established methods of legal redress. [6]

This research was part of the project "Application of the General Data Protection Regulation in childcare institutions funded by University of Zagreb in 2018, led by Assistant professor Tihomir Katulić.

The European Union has, to its credit, been steadily paving the way to better, modern legislation and to perception of personal data protection as a fundamental right to begin with. Following the adoption of the Treaty of Lisbon and the Charter of Fundamental Rights which came into effect in 2010 completing the establishment of personal data protection right as a fundamental right in the EU parallel to privacy, the next step in development of the common European legal framework of data protection was to replace the aging Data Protection Directive from 1995 with a modernized law that would take into account all the technological and legal changes and advancements in the past 25 years. [7]

This development became a Regulation, the General Data Protection Regulation (GDPR), that was finally adopted in 2016 and came into effect in 2018. The Regulation holds within its 99 articles and 173 recitals further development of key concepts established in earlier texts such as the roles of data controllers and data processors, the development of explicit data subject rights such as the right to be informed about data processing, the right to access own data, the right to rectify outdated and incorrect personal data, the „right to be forgotten" or the right of data subjects to demand data controllers remove data they have no longer legal basis to process, but also adopts a more modern approach of ensuring security of processing through information security practices which are now also becoming the matter of EU legislation. [8] GDPR gives special attention to children and how they need particular protection because they may be less aware of the risk of sharing their personal data [9].

The control mechanisms the Regulation develops are well known to information security experts, from *privacy by design and by default*, through requisite use of technical and organizational protection measures and provisions regarding security incident discovery and reporting to national supervisory bodies and data subjects themselves. [10]

COMPLIANCE CHALLENGES IN PUBLIC EDUCATION SYSTEM IN CROATIA

Measures and obligations of data controllers make perfect sense in the context of big data industries, especially communications, financial and health industry. Especially because they are traditionally reliant on collecting and processing huge amounts of personal data and performing detailed profiling on the users of their services. While these measures are welcome in the context of increased tendencies of mass government surveillance, especially with the advent of smart cities and Internet-of-things (IoT) technologies usually performed under public authority or obligations of central and local government bodies under national law, they do create implementation and compliance challenges for smaller controllers such as small business companies, underfunded or underdeveloped public institutions and civil society organizations who usually do not posess adequate resources, knowhow and manpower to address these issues. Among these, especially vulnerable are local health and education facilities, the latter being the object of our research. These organizations, especially in former Eastern Block and transition societies traditionally suffer from the lack of funding and systematic support from local government that establishes and funds them. Meeting stringent GDPR obligations for them is a significant challenge while they struggle to offer services to especially vulnerable populations.

In these organizations, most of the burden of compliance efforts has been delegated to their management left to figure out how to use their modest resources, usually without any significant support from local government or competent central government institutions.

DATA PROTECTION IN PUBLIC EDUCATION

The main activities of public education institutions are centered around providing education services to different categories of students ranging from nurseries and kindergartens to elementary and secondary education.

The Croatian educational sector for children in numbers shows the coverage of children in primary and secondary education in Croatia. At the beginning of the 2019 school year, 316 104 children were enrolled in 2115 elementary schools in Croatia. [11]. There were 148 466 high school students in 739 secondary schools enrolled the same school year [12].

Schools collects various personal data: Information on active and past students and their family members, information on employees and their family members, information on visitors, co-workers, partners, etc. [9] In school institutions, personal data can be collected directly from respondents or family member, by electronic communication, by access official databases, and may also be submitted by ministries and other organizations through specialized information systems. Personal data in schools are collected, stored and processed on a daily basis, which entails, with knowledge of different regulations, and the handling of digital competencies and specific IT skills. IT support and information security are often not the focus of educational institutions, which poses an additional challenge when implementing the complex requirements of the new General Data Protection Regulation.

In today's world, where information and data are highly valued, group environments like schools, where most children reside on a daily basis, represent the ability to gather a large amount of information relatively quickly and easily in a short time. However, at the same time, there is a risk of inappropriate and illegal use of personal data, which should be minimized in order to protect the safety and well-being of the child.

In this paper, our research focus was centered on the efforts of professionals in the education settings in Split-Dalmatia County in implementing the GDPR in the schools.

RESARCH OVERVIEW

The aim of this preliminary qualitative research is to find out the opinion of teachers, professional associates and principals of the schools in Split-Dalmatia County on the implementation of the GDPR and issues these institutions are experiencing. Research questions for the teachers, professional associates and principals are set up with a purpose of better understanding:

1. How do they perceive their knowledge about GDPR and did they participate in a formal education? (Table I, Table II)
2. What do they think about process of implementation of GDPR in their institution? (Table III, Table IV, Table V)
3. What specific challenges did they experienced with implementation of the GDPR in their institution? (Table VI, Table VII).

II. METHODS

Data were collected in the fall of 2018, using the focus group method (4 groups in total), lasting a maximum of 30 minutes, respectively. Research was conducted by experienced researcher who participated in short training on GDPR in education prior to the field research.

Sample: of the four focus groups, two were conducted with teachers and professional associates (10 participants in total) and two with principals (8 participants in total). All of the participating teachers and professional associates were female, while 7 out of 8 principals were male. To ensure confidentiality, further socio-demographic data about participants was not collected, encouraging honesty in answering questions.

Ethics: Approval of the principal of the school to conduct the research in their school was mandatory. Informed consent was collected from every participant prior to the research, as well as consent for audio recording of the focus group. Participants were assured that their personal information would not be used and that it would not be possible to relate their answers to their identities. They were also assured that only members of the research team would have access to the collected data, which will be destroyed after the finalization of the project.

Data analysis: Data are analyzed with a simple content analysis, which. included the following steps:
1. Minimal language editing of all open-end question responses
2. Underlining of specific statements of the respondents
3. Extraction of the underlined statements of the respondents
4. Separating statements into different categories
5. Defining sub-categories (if applicable)
6. Assigning a frequency of the respondents' statements within a category

Content analysis was used for the categorization and classification of answers on open-end questions. The method of simple content analysis enabled the systematic overview of textual information by its frequencies in specific category. The results are expressed in the number of occurrences. Presented participants' statements in the results section, are chosen to describe the categories in more detail.

III. RESULTS AND DISCUSSION

The qualitative content analysis, that enabled the systematic overview of data assigned to specific categories was performed on responses of teachers and principals with purpose of assessing their opinion about GDPR and its implementation in their schools.

TABLE I KNOWLEDGE ABOUT GDPR AND THE DEGREE TO WHICH THEY ARE INFORMED

Category	Codes	Frequency
Level of knowledge about GDPR in general (12)	Teachers' lack of knowledge about GDPR	4
	Basic knowledge of principals about GDPR	6
	Advanced knowledge of principals about GDPR	2
Lack of interest about GDPR (9)	Lack of teachers' general interest in GDPR	9
Level of awareness of the risks of data misuse (10)	Teachers' lack of knowledge about the risks of data misuse	4
	Teachers' basic knowledge about the risks of data misuse	6
Level of awareness of internal rules on personal data protection (11)	Teachers' lack of knowledge about the internal rules on data protection	6
	Teachers' basic knowledge about the internal rules on data protection	3
	Non-familiarity with data protection officer	2

Content analysis of the first research question produced four categories in the first block of the results: (1) Level of knowledge about GDPR in general, (2) Lack of interest about GDPR, (3) Level of awareness of the risks of data misuse and (4) Level of awareness of internal rules on personal data protection. (Table I)

Results shows that only principals have basic (*f*=6, *RSS2: Partly, we are familiar in a way that we are trying to apply the regulation in the everyday functioning of the School.)* and in less cases, advanced (*f*=2) knowledge about GDPR while teachers show lack of knowledge in this field (*f*=4, *NSS1: I am not familiar with the regulation and do not know much about it.*). In the same time teachers expressed lack of general interest for the topic of GDPR (*f*=9, *NOS1: I personally am not much but for the reason that I was not very interested… I am more oriented towards teaching, not papers*) which was the most frequent category of the first research question. Also, research findings implicate that in school setting there is lack of knowledge about risks of data misuse (*f*=4), personal data protection (*f*=6, *NO3: if there is, we do not know it exists*) and that they are not familiar who is data protection officer in their school (*f*=2) or they have basic information about risks of data protection (*f*=6) and internal rules of data protection (*f*=3). For comparison it is interesting to note that a Eurobarometar research on more than 27000 Eu citizens shows that general population in 67% know about the General Data Protection Regulation and 57% of respondents know about their national data protection authorities [13].

TABLE II EXPERIENCES WITH FORMAL EDUCATION AND MOTIVATION FOR FURTHER EDUCATION ABOUT GDPR

Category	Codes	Frequency
Experiences with formal education about GDPR (21)	Non-participation of teachers and principals in formal education about GDPR	9
	Participation of principals in formal dana protection education	4
	Non-participation of data protection officer in further education about GDPR	4
	Participation of data protection officer in formal education	4
Interest in future educations about GDPR (6)	Teachers' interest in future education	3
	Principals' interest in future education	3

Table 2 presents experiences with formal education about GDPR in education and their interest for future education in the field. Results shows that teachers generally don't participate in education about implementation the GDPR in school settings (*f*=9) as the same as some of data protection officers in the school (*f*=4) who did not participate in available education. On the other hand, some principals expressed that they've participated in formal GDPR education (*f*=4), as well as some data protection officers (*f*=4). Part of the teachers (*f*=3) and principals (*f*=3) expressed their interest in further education in this field. It is important to enhance that it is not clear whether there was a lack of institutional support in organizing educations within the field or there was a lack of interest from the implementers, in this case principals, teachers and professional associates. Scarce research in the field of data protection in educational sector emphasizes the complexity of issues schools face in the process of GDPR implementation. Rosmaini, Kusumasari, Lubis and Lubis [14] state that this includes competent employees, technology, task and structure, as well as procedure and policy aligned with relevant and related personal data protection act.

TABLE III IMPLEMENTATION OF GDPR: CLARITIY OF INSTITUTIONAL GOALS

Category	Codes	Frequency
Clarity of defined goals of personal data protection in the institution (8)	Lack of goals of personal data protection strategy	2
	Basic level of clarity of defined personal data protection goals	3
	Intermediate level of clarity of defined personal data protection goals	2
	Advanced level of clarity of personal data protection goals	1

Content analysis of the second research question produced four categories regarding/in relation to implementation of the GDPR in schools: (1) Clarity of defined goals of personal data protection in the institution, (2) Existence of data protection guidelines and regulations in the institution, (3) Existence of data protection officer in school and (4) Security and data protection measures (7).

Table 3 shows that there are not always clearly defined goals of personal data protection within institution. Some of the participants said that the goals are not defined (*f*=2, *NSS3: We have no clear goals*) or that they are familiar with them on a basic level (*f*=3). Two schools show intermediate level (*f*=2) and two schools are identified as positive practice example and shows advanced level of clarity of goals of data protection (*f*=2, *ROS1: We are on our way … In practice, this has complicated our lives, but I know that we should work that way. We are in the process of hiring a legal person to take care of this*.).

In some schools, GDPR implementation started with video surveillance, as a part of security measures (*f*=4), while one school physically locks the data in a secured space (*ROS2: We use physical data lock…cabinets, firebox.*). Two schools reporting the lack security data protection measures (*f*=2).

TABLE IV IMPLEMENTATION OF GDPR: SECURITY

Category	Codes	Frequency
Security and data protection measures (7)	Non-existent security data protection measures	2
	Security cameras	4
	Securing data in a locked room	1

Table 5 shows that teacher's opinion is that there is no data protection guidelines and regulations on school level (*f*=8, *NS5: We are not adequately educated or know that there are some guidelines*), which is confirmed by the statements of three principals (*f*=3). Other principals report that guidelines and regulations exists (*f*=4), or that this field is partially regulated (*f*=1, *RSS2: Partially, within the regulations governing the protection of archives and records*). One institution reports gradual introduction of data protection guidelines in the institution (*f*=1), and again one school shows advanced level of development of internal data protection regulations within institution. National governments of some European countries prepared detailed guidelines of implementing the GDPR in schools within positive practices and policies regarding personal data security [15]

TABLE V IMPLEMENTATION OF GDPR: REGULATION AND DATA PROTECTION OFFICER

Category	Codes	Frequency
Existence of data protection guidelines and regulations in the institution (20)	Lack of data protection guidelines in the institution – Teacher perspective	8
	Guidelines and regulations exist- Principal perspective	4
	Gradual introduction of data protection guidelines in the institution- principal perspective	1
	Lack of internal data protection regulations- principal perspective	3
	Internal data protection is partially regulated	1
	Advanced level of internal data protection regulations within the institution	1
Existence of data protection officer in school (13)	Administrative staff	10
	External expert	1
	Data protection officer is assigned in the institution	3

Content analysis of the last research question examining challenges of GDPR implementation in school setting produced four interesting categories: (1) Lack of systematic introduction of GDPR in the education, (2) Lack of a universal model of information provision for educational employees, (3) Challenges regarding consent for usage of personal data, (4) Teacher resistance to the introduction of a privacy statement regarding the use of personal data. Experts agree that institutions should adopt internal policies and assure the implementation of the privacy measures in order to demonstrate compliance with Regulation and meet the goals of effective data protection [16].

TABLE VI CHALLENGES: LACK OF SYSTEMATIC PREPARATION

Category	Codes	Frequency
Lack of systematic introduction of GDPR in the educational sector (9)	Lack of clear information how the GDPR is implemented in their school	3
	Lack of a clear procedure for implementing GDPR at school	6
Lack of a universal model of information provision for education sector employees (9)	Basic information of teacher on teacher council	2
	Basic information of principals on principals meeting	1
	Information available on school information panel	2
	Online documents available	1
	Information on media	1
	Other sources of information	2

Table 6 shows categories that identify a lack of systematic preparation for the implementation of GDPR in schools. Participants report a lack of clear information on how the GDPR is implemented in their school (f=3), as well as a lack of protocol for implementation (f=6, *NSS2: I know there should be a person in charge of taking care of that information, but we don't have that person*). At the same time, educational employees are being informed about GDPR through various "ad hoc" models of information/ means, including teacher council (f=2), principals meeting (f=1), school information panels (f=2), available online documents (f=1), media (f=1) and some of them mention other unofficial sources of information, such as tv and other media and professional conferences (f=2).

Results of the last research question regarding specific challenges participants experienced during GDPR implementation were particularly interesting. Table 7 shows that teachers and principals experience challenges regarding consent for personal data use. Some of them believe that they have a general knowledge of needing to collect the consent in certain situations (f=4), while others identify a lack of uniform procedure within and between schools as the challenging part of the collection of the consent (f=5, *NOS5: One school works like this - another school works differently*). A number of participants' experience ambiguities in the implementation of specific provisions about consent for personal data usage (f=4, *NOS1: I wish that the law resolves that we do not need to collect signatures*). Furthermore, they are resistant about implementation of privacy statement for teachers and disliked this procedure (f=5., *NO3: Before (the GDPR) we didn't need to sign with a signature that I would not talk about children from school either at home or outside school or anywhere*).

TABLE VII CHALLENGES: CONSENT AND PRIVACY STATEMENTS

Category	Codes	Frequency
Challenges with consent for usage of personal data (17)	General knowledge of the specific situations of consent for the use of personal data	4
	Lack of uniform procedure about collecting consents for the use of personal data	5
	Resistance to collection of consent for the use of personal data	5
	Ambiguities in the implementation of specific provisions about consent for personal data usage	4
Teacher resistance to the introduction of a privacy statement (5)	Teacher resistance to the introduction of a privacy statement regarding the use of personal data	5

Informed consent and strong ethical and legal standards are highly important, especially for the children population, considering their age and possibility to understands the risks and responsibilities. Informed consent of the child and parent is obligatory in a number of situations where institutions are obliged to pre-collect personal data, but some of the public activities planned in the school year program could resolve easily with the parent consent for planned activities on the beginning of the school year. Authors of this manuscript strictly oppose asking parents to sign blank consent forms for the child without actually informing them on the subject of the consent, as is the practice in some institutions (*NO3: in some schools they all sign up at the beginning of the school year).*

IV. RESEARCH IMPLICATIONS AND LIMITATIONS

This preliminary qualitative research is significant considering the lack of the research in the field of implementation of GDPR in education. Preliminary results indicate the need for further research that will explore in detail the opportunities and challenges of implementing GDPR in educational settings. There is also an indisputable need to improve the systematic and professional introduction of new regulations in educational institutions. It is certain that both teachers and principals need better support, as well as systematic access to information and training in the field of data protection.

Furthermore, the findings of this research may direct future research in this field, as well as be useful to experts planning and implementing educational content in the education sector.

Although the research design planned more complex qualitative analysis predicting that the participants answers would be sufficiently informative and meaningful to be appropriate for thematic analysis, this did not turn out to be correct. Given the very modest and simple responses of the research participants, it was decided to carry out a simple content analysis. These preliminary research results can be a starting point for planning future qualitative research in the field of GDPR application in school settings. Although the authors of this paper are aware that the response frequencies in the categories are too low to allow any statistical data handling, they nevertheless point to a need for more detailed planning of research in this field. Research findings, but also general and very restrained and short responses from research participants indicate that there is a significant need to inform, educate and support school staff in GDPR implementation in their professional practice.

V. CONCLUSION

The results show that teachers and professional associates are generally not familiar with the application of the GDPR in school settings, while the experiences of principals are different. Some of them are fully or partially informed and part of them are not familiar with the application of the GDPR in schools. This research confirms a lack of the systematic preparation for the implementation of the GDPR in some Croatian schools. Motivation for further education and personal responsibility for continuous professional growth, especially in the field of protection of rights and data protection is inevitable part of the ethical and effective practices. At the same time, it is the responsibility of the system to make lifelong education available to the educational employees and to ensure investment in these sectors so that they can properly and adequately collect, handle and store personal data, within the framework of GDPR in educational institutions.

Although they work with children and young people, a group that is particularly sensitive when it comes to data security, this research shows that they do not have enough knowledge, experience and skills required in the specific situations in their public action.

It should be emphasized that this paper contributes to the understanding of the implementation of the Regulation in the school context in Croatia, which is especially significant given the lack of sources of literature in the domestic context in this field. Based on these results, it seems necessary to develop accessible educational content for educational employees to ensure an adequate application of GDPR in relation to children. Particular recommendations for developing educational content for employees in school setting are:

- Educations should address the needs of the employees in school setting meaning that content should be applicable in their everyday professional practice
- Educations should be accessible to different categories of employees (teachers, principals, professional associates) with specific content related to their different professional roles within schools
- Guidelines and standards for the implementation of GDPR in school settings should be clear and accessible - all the actors should be informed about their content.

Preliminary results of the presented research can be useful in planning and conducting research in this field, especially since there is an unquestionable need for evidence-based practice in the field of GDPR implementation in the school context.

REFERENCES

v[1] M. Ajduković and M. Šalinović (ur). Indikatori dobrobiti djece. Zagreb: UNICEF ured za Hrvatsku i Ministarstvo za demografiju, obitelj, mlade i socijalnu politiku, 2016.

[2] J. Bradshaw, A. Keung, G. Rees and H. Goswami. "Children's subjective well-being: International comparative perspectives". *Children and Youth Services Review*, 33, 2011. 548–556.

[3] United Nations. Convention on the Rights of the Child, 1989. available at https://www.ohchr.org/en/professionalinterest/pages/crc.aspx

[4] S. Audsley, F. Kalyani, M. Bronwen, B. Robinson and K. Varney. An Examination of Coursera as an Information

Environment: Does Coursera Fulfill its Mission to Provide Open Education to All?, The Serials Librarian: From the Printed Page to the Digital Age, 2013. 65:2, 136-166, p. 157.
[5] N. Parlov, Ž. Sičaja and T. Katulić.: GDPR – Impact of General Data Protection Regulation on Digital Marketing. Annals of disaster risk sciences. 1, 2, 2018., 105-116
[6] Information Commissioner's Office Guide to Data Protection, available at https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-of-access/
[7] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Official Journal L 281 , 23/11/1995 P. 0031 - 0050
[8] T. Katulić. "Transposition of EU Network and Information Security Directive into National Law", MIPRO 2018 41st International Convention Proceedings Skala, Karolj (ur.). Rijeka : MIPRO, 2018. 1328-1333
[9] A. Calder. EU GDPR: *A Pocket Guide, School's edition*. IT Governance Publishing Ltd. 2018.
[10] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union, L 119/1
[11] Croatian Bureau of Statistic. Basic schools end of 2017/2018 school year and beginning of 2018/2019 school year. First release. 2019. Available at https://www.dzs.hr/
[12] Croatian Bureau of Statistic. Upper secondary schools end of 2017/2018 school year and beginning of 2018/2019 school year. First release. 2019. Available at https://www.dzs.hr/
[13] EU. Special Eurobarometer 487: Charter of fundamental rights and General Data Protection Regulation. Available at https://data.europa.eu/euodp/en/data/dataset/S2222_91_2_487_ENG
[14] E. Rosmaini, T.F. Kusumasari, M. Lubis and A.R. Lubis, 2018, J. Phys.: Conf. Ser. 978 012037. doi :10.1088/1742-6596/978/1/012037
[15] IT Governance Privacy Team, EU General Data Protection Regulation (GDPR) - An Implementation and Compliance Guide, Third edition, 2019, ITGP.
[16] Data protection: toolkit for schools
Guidance to support schools with data protection activity, including compliance with the General Data Protection Regulation (GDPR), 2018., Department for Education. Available at https://www.gov.uk/government/publications/data-protection-toolkit-for-schools

IoT Devices and the Need to Inform Utility Users of Collecting, Controlling and Processing of Personal Data

G. Vojković, Ph.D. * and M. Milenković, LL.M*
* University of Zagreb, Faculty of Transport and Traffic Sciences/Chair of Transport Law and Economics, Zagreb, Croatia
gvojkovic@fpz.hr
mmilenkovic@fpz.hr

Abstract - **The introduction of smart devices into households presents new challenges in terms of privacy and protection of personal data, given the wide set of data they collect. Replacing various meters that measure electrical energy, water or natural gas consumption, and with smart meters bring many benefits, such as the ability to read data automatically, an approach to managing energy distribution based on real-time energy consumption, although on the other hand it presents a new security risk. For example, by introducing a smart meter for electricity or water which can be monitored to even obtain information on how many household members are present in the household, and when the household is empty, and what are their life habits. Under the current consumer protection framework in the EU, service providers are under no obligation to notify customers that new consumer connected devices capture a significantly larger set of data about them and pose a potential privacy risk. In this paper, we consider the basic dangers to privacy and personal information brought by using IoT devices, and outline proposals to change the legal framework to inform consumers about the capabilities of these utilities to control devices bi-directional communication, coordination and the very need to monitor the data collected.**

Keywords – IoT, privacy, EU, regulation, data

I. INTRODUCTION

In last several years various measuring and other devices are being replaced by smart devices, as part of the IoT model. Electricity meters, water consumption meters and other different meters are becoming smart meters - allowing remote reading, alerting (e.g. at unusually high-power consumption) and with the addition of new functionality, e.g. measuring consumption at a certain point in time.

What is IoT? There are many definitions currently in use. In one of our previous papers, we chose one because of its practicality: "The Internet of Things, commonly abbreviated as IoT, refers to the connection of devices (other than typical fare such as computers and smartphones) to the Internet." The business media following the development of IoT often references different products and services such as interconnected cars, smart home appliances and even kitchen appliances and medical devices such as heart monitors. [1]

So far, there have been a few relevant market research efforts to estimate the size and importance of the IoT market. One of them has estimated that the number of internet-connected devices will outnumber humans interfacing the internet to the ratio of 4-to-1 by year 2020. Naturally, many of these devices will belong to IoT paradigm. [2] Another research we referenced in earlier papers referenced the fact that the average smart home in the United States now employs the services of eleven smart devices. [3]

This and other research have interesting ramifications and we can safely assume that IoT products and services have already transcended the development phase and the use by early technology adopters and enthusiasts and has entered a widespread everyday use.

In our everyday lives, the basic television sets have been replaced by increasingly connected and interactive Smart TV variety, households increasingly adopt products like smart speakers – virtual assistant and in some areas the general population is not even aware that utility companies, such as their water, gas or electricity supplier, has replaced their traditional meters to new devices employing IoT. [4]

Like all devices connected to the Internet, IoT devices are prone to malware attacks. These attacks have traditionally increased in number and severity. According to research conducted by infosec companies such as F-Secure, in the first half of 2019, their global honeypot network detected a threefold increase of information security incidents bringing a total to over 2.9 billion events. The same research underlines that most of the attacks were conducted by means of malicious software such as internet worms targeting TCP ports responsible for, among other things, IoT communication. [5]

The current IoT paradigm employs the use of a large number of devices produced by various manufacturers developed for hundreds of different uses even if they sometimes share the basic hardware platform and

communication hardware. The heterogeneous nature of these devices, most of which battery-powered and interconnected via wireless network interfaces, makes discovery and treatment of information security incidents a very demanding effort. The modern IoT technology can even create its own communication network without the help of dedicated infrastructure by utilizing its embedded wireless communication. [6]

Some IoT devices are not protected by sufficient access control. [7] However, the IoT cannot purely focus on network security. An application layer must be included. [8]

Replacing common metering devices used in households with smart devices is a reality in a larger part of the world today. The introduction of smart meters with the ability to read data remotely has become routine instead of the current "dummy" devices which needed to be physically accessed and periodically read.

The IoT trends have influenced electronic industry in Croatia and Slovenia as well. A local manufacturer has introduced IoT technology into metering devices, i.e. Iskraemeco AM550, an electricity meter that has been seen in use in the Republic of Croatia. This device provides two-way (''energy'') measurements, active energy and power, 4Q reactive energy & power, apparent energy & power, instantaneous value of voltage, current, power factor, frequency and power and an absolute measurement of active energy & power. [4] [9]

Additionally, this device has more possibilities of reading data. Optical port, RJ11 (for in-house display), wired and wireless M-bus, WAN/NAN Communication modules – PLC G2/G3, also point-to-point 2G/3G/4G. [9]

II. COLLECTION OF PERSONAL DATA BY SMART DEVICES

Unlike previous "dummy" devices, smart devices have two important differences:

a) they collect substantially more data,

b) the ability to read data remotely is a common feature.

As an example, we can take the aforementioned measurement of household electricity consumption (which is often the first device to be installed as a smart device). With a standard meter, we can read the consumption of the past time, if we stand in front of the device (that is, on a monthly reading, spending over the last month).

With smart meters, it is possible to read consumption at a specific time, which provides us with very sensitive personal and private information. Such a device monitors the power consumed at the exact time and it is possible to find out when tenants/users are usually in the household, or when they spend time outside the household. Such information may be valuable to the third malicious parties. In addition, with a little knowledge of the average consumption of the appliance, it is possible to reconstruct when tenants go to sleep, when they shower, whether they use the oven, what are their customs regarding the heating of the object etc.

Specifically, these are very sensitive private information. Collected information can also lead to the conclusion of how many tenants/users are currently in the household. Also, such information may be of interest to insurance companies, life insurance may be interesting to know if a person eats late or has other unhealthy lifestyle habits that can be indirectly detected.

Weaver wrote: "Residential utility customers have a legitimate expectation to preserve individual and behavioral privacy with regard to energy-related or water consumption data collected by the utility. (…) Furthermore, there is a deep concern that inadequate cyber security measures surrounding the digital transmission of smart meter data will expose such data to misuse by authorized and unauthorized users of the data." [10]

What a "smart" memory meter looks like was described back in 2010. [11]. Figure 1. shows power segments (appended with labels from activity logs) per day for one of the households.

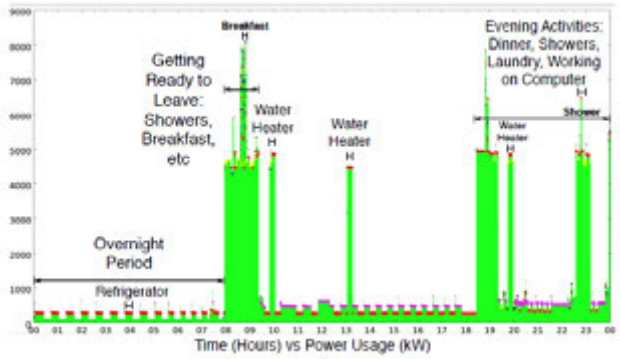


Figure 1. Example day-long second-level power trace [11]

Figure 2. reveals when users/consumers were in one of the households over the course of a month with weekends highlighted.

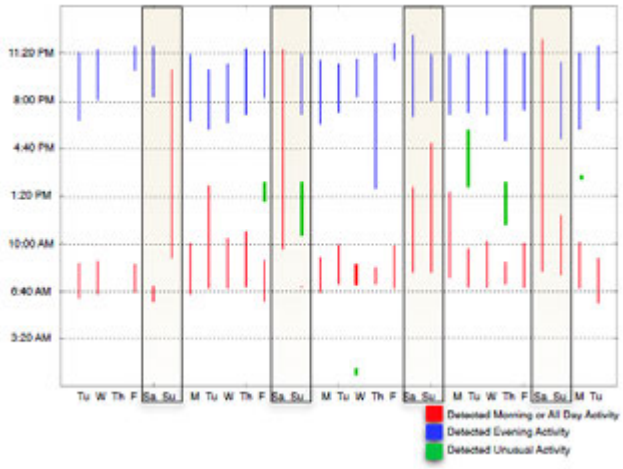


Figure 2. Identification of human presence with high [11]

Authors summarizing that analysis demonstrate how easy it is to identify private information from smart meters. [11]

Let's remind ourselves, when we talk about service providers, we are talking about companies that collect data from tens or hundreds of thousands of users! In addition, when it comes to malicious hacking, a malicious party can access a large number of measuring instruments in a very short period of time, because their security systems are similar and often defaulted (see infra).

Even before the "smart devices" era, in 2012, British Information Commissioner's Office has already warned how data from different measurement devices represent personal information: "Data concerning the water consumption for a particular address will be tenants' personal data due to the fact that data determines how much that individual will be charged off." [12] This can be applied to smart devices.

Information Commissioner's Office states another important fact regarding the collection of various consumer spending data: "Also, if necessary, the water utility company is likely to be able to easily obtain the name of, if not the occupier, then at least the registered owner of the property." Utilities are usually connected and closely related to the local government, often linked to management and ownership. Therefore, it is very easy to link and misuse different consumption metrics with personal and proprietary data.

General Data Protection Regulation (further: GDPR) does not specifically mention IoT devices, but according to material scope in the Art. 2. of the GDPR includes IoT devices: "This Regulation applies to the processing of personal data wholly or partly by automated (…). [13]

GDPR in Recital 6 acknowledges: "Rapid technological developments and globalization have brought new challenges for the protection of personal data. The scale of the collection and sharing of personal data has increased significantly. Technology allows both private companies and public authorities to make use of personal data on an unprecedented scale in order to pursue their activities." [13]

For data collected by different meters it is important Recital 39 of the GDPR: ""Natural persons should be made aware of risks, rules, safeguards and rights in relation to the processing of personal data and how to exercise their rights in relation to such processing." [13]

Regarding the ability to read IoT devices remotely, the service provider must apply appropriate safeguards in accordance with rectal 83 of the GDPR: "In order to maintain security and to prevent processing in infringement of this Regulation, the controller or processor should evaluate the risks inherent in the processing and implement measures to mitigate those risks, such as encryption. Those measures should ensure an appropriate level of security, including confidentiality, taking into account the state of the art and the costs of implementation in relation to the risks and the nature of the personal data to be protected." [13]

An IoT device, such as a power or water meter, is a substantially different device from a "dummy" meter. It enables the collection of a whole range of very sensitive information about the service user. Recital 60 of the GDPR states: "The controller should provide the data subject with any further information necessary to ensure fair and transparent processing taking into account the specific circumstances and context in which the personal data are processed." [13]

Do service providers inform the user of a substantially different, wider, and more accurate range of data collected by IoT meters and other devices in comparison to "dummy" devices? For example, in Croatian business practice, a service user who receives an electricity meter only signs the record that the meter has been replaced.

IoT devices can easily be used for "profiling". By the GDPR definition, profiling means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person. In such case, according Recital 60 and the Art. 13. the data subject should be informed of the existence of profiling and the consequences of such profiling. [13]

Also, the question about the devices for measuring energy consumption should be looked beyond the GDPR itself, taking into account the overall capabilities of these devices, not just the set of functions used by the service provider.

Relationship between users/consumers of utilities and similar services and providers of these services cannot be viewed only through the view of the GDPR, but also through regulations governing consumer protection. Users should therefore be aware of the possibilities provided to the service provider by the IoT measuring device, in order to get acquainted to what they give their consent when using such devices. Also, they should be aware that such devices are potentially vulnerable by the third parties.

Appropriate and quality data protection on the servers of the service providers does not necessary mean much if an unauthorized person can easily read the device located in front of the household of any potential service user.

III. INFORMING USERS ON THE IMPACT OF IOT DEVICES ON THEIR PRIVACY

The basic relationships between the various utility providers and consumers are governed by consumer protection regulations. The basic consumer protection regulation in the European Union is Directive on consumer rights. [14]

Scope of Directive on consumer rights is noted in the Art. 3 "This Directive shall apply, under the conditions and to the extent set out in its provisions, to any contract concluded between a trader and a consumer. It shall also apply to contracts for the supply of water, gas, electricity or district heating, including by public providers, to the extent that these commodities are provided on a contractual basis." [14]

Directive on consumer rights regulates in detail consumer information rights. According the Art. 1, "The purpose of this Directive is, through the achievement of a high level of consumer protection, to contribute to the proper functioning of the internal market by approximating certain aspects of the laws, regulations and administrative provisions of the Member States concerning contracts

concluded between consumers and traders." When 'consumer' means any natural person who, in contracts covered by this Directive, is acting for purposes which are outside his trade, business, craft or profession; and 'trader' means any natural person or any legal person, irrespective of whether privately or publicly owned, who is acting, including through any other person acting in his name or on his behalf, for purposes relating to his trade, business, craft or profession in relation to contracts covered by this Directive. [14]

In the Art. 5. "Information requirements for contracts other than distance or off-premises contracts" (distance or off-premises contracts have additional rules which are not important for this paper) Directive on consumer rights quote that before the consumer is bound by a contract other than a distance or an off-premises contract, or any corresponding offer, the trader shall provide the consumer with the information in a clear and comprehensible manner, the main characteristics of the goods or services, to the extent appropriate to the medium and to the goods or services. Such obligation shall also apply to contracts for the supply of water, gas, or electricity, where they are not put up for sale in a limited volume or set quantity.

Directive on consumer rights does not regulate smart devices and utilities connected with IoT. We believe that consumers should have the right to be informed about smart devices that will be implemented in their households in the future, which measure services such as water, gas, or electricity. This problem is not fully covered by the GDPR. "To summarize its material scope, the GDPR applies to any processing of personal data. The Regulation will become relevant for companies as soon as any data processing takes place. The (material) scope is interpreted in a very broad manner in order to ensure a high level of protection." [15]

Here, we shall mention the previous acquaintance of the user with the possibilities of the device, specifically the acquaintance with the criteria. In our opinion, the consumer should give his consent to any new IoT "smart" device that the utility provider would like to implement in its property for the purpose of simplifying service metering. The trader must inform the consumer about the functionality and the relevant interoperability of device, especially meter.

The capabilities of IoT devices are as we already mentioned, substantially wider than the actual range of data that provider is processing at some point! Security of IoT devices is generally not satisfactory, 2019 reports mention increasing attacks on IoT systems. "Internet of Things (IoT) has become a primary target for cybercriminals. The repeated security incidents on IoT devices represent a rising trend for IoT attacks. The proliferation of connected devices in consumer, enterprise, and healthcare organizations, and their internal vulnerabilities, have created a security blind spot where cybercriminals can launch a Zero-day attack to compromise devices like webcams, smart TV, routers, printers, and even a smart home." [16]

As an example. Portals for IT-experts noted in 2019. about malware called Silex. "Silex works by trashing an IoT device's storage, dropping firewall rules, removing the network configuration, and then halting the device. It's as destructive as it can get without actually frying the IoT device's circuits." [17]

There is interesting example from Russia, where government has recently mandated that all electricity meters be replaced by online smart meters. It open new "black market", Russian criminals started modifying and selling customized firmware for such devices. These modified smart meters are marketed on black market as a means to save on monthly residential bills for electricity, water, and gas."[18] Authors of same document predict that in the future, hacking smart meters may become lucrative for criminals.

Let us ask ourselves a question: is today already possible to order vendor cheating software to show a lower consumption of benchmarks, and/or could the same software be used to "increase" the bill to a non-sympathetic neighbor? Or to gather information about when someone is in the household? Obviously, it is possible. If a device can be hacked, it can be misused, regardless of which set of data is collected by the provider. A hacker or a user of such services (who commissioned the hacking service on the black market) may use the device and data for completely different purposes than those collected by the service provider. Therefore, we believe that the installation of these devices should be governed by consumer protection regulations, not by the GDRP!

And let us also take into consideration another fact, Moore in his book "Crossing the Chasm" shows model which describes the market penetration of any new technology product in terms of a progression within the types of consumers it attracts throughout its useful life:

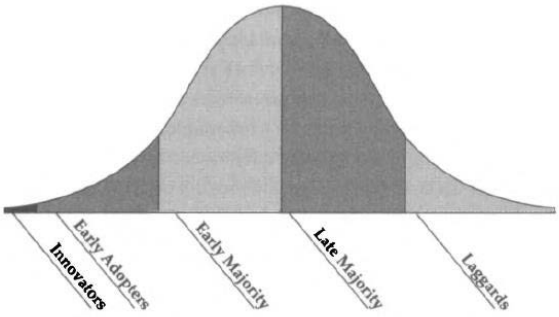


Figure 3 Technology adoption life cycle by Geoffrey A. Moore [19]

The chart above clearly illustrates why IoT devices are vulnerable. Although it cannot be directly applied to metering devices (their introduction depends on the provider), it can be applied to all the other IoT devices. Initially used by enthusiasts and advanced users who are usually familiar with technology, security systems, and hacking devices. And then later, as the device becomes standard, it is increasingly bought by users who lack specialized knowledge and skills, further increasing the vulnerability of these devices.

In order to gain its full affirmation, it is necessary to ensure the highest possible degree of security and privacy

in the use of IoT devices and the proper operational implementation of IoT.

There are several potential threats:

- Security threats: refers to the prevention of financial harm that can be experienced by users and IoT service providers, as well as the loss of reputation and payment of provider penalties
- Privacy abuse: unethical usage of users' data and users' habits
- Operational threats: data loss, contract breach and inaccurate data.

To make building and using IoT platforms safe to use, the American non-profit organization OWASP (Open Web Application Security Project) analyzed IoT technology and listed 10 IoT vulnerabilities in its 2018 report**:**

1. weak, Guessable, or Hardcoded Passwords,
2. insecure Network Services,
3. insecure Ecosystem Interfaces,
4. lack of Secure Update Mechanism,
5. use of Insecure or Outdated Components,
6. insufficient Privacy Protection,
7. insecure Data Transfer and Storage,
8. lack of Device Management,
9. insecure Default Settings,
10. lack of Physical Hardening. [20]

In short, IoT meters are sensitive to a whole range of potential threats. They are also vulnerable outside the scope of the GDPR, as they can be abused by the third parties, especially if they are not adequately protected or properly maintained by the utility provider, or the devices are not maintained properly. Problems can also occur outside of the service provider itself - for example, cessation of upgrades by the manufacturer.

Therefore, we believe that the user should be aware of such potential threats and should give prior consent to the use of IoT devices.

IV. PROPOSAL FOR NEW EU-REGULATION

We believe that Directive on consumer rights should be supplemented. Possibilities and threats of IoT devices should be listed in the recital. Furthermore, in the Art. 5. of the Directive (Information requirements for contracts other than distance or off-premises contracts), it states that consumer information should be supplemented by a new point which should read:

"The utility provider is obliged to explain to the recipient of the service when placing an IoT device (measuring device) into the household, including:

- what are the purposes of measuring of such a device,
- where can it be used and,
- what capabilities of this type of reading the device has."

Also, with each implementation of the new device comes an instruction on the device, which previously stated what are the main technical capabilities of the device that is being implemented into the household.

In Britain, Guy Herbert of NO2ID says: "Smart meters are presented as an environmental and power-saving initiative. But it's a highly surveillant model. It can tell how many showers you have had, when Evidence of the race to monetize the data from smart meters is already emerging." [21]

Moreover, what we would like to point out is the reasonable fear that once the data is released, it can ''fall into the wrong hands'' and in that field the big data market will open very quickly, which will potentially be very dangerous for citizens and another possible attack on their personal and private information. What can be even worse if a third party who is not even a utility operator reaches this large amount of data and knows the devices at the technical level because the devices themselves are not protected from intrusion by third parties and they can simply resell such data or compromise the privacy of the individuals who have been hacked.

V. CONCLUSION

Due to the reasons outlined above, it is clear that utility data collection methods for electricity, and also gas or water usage of data for home users using smart meters could represent an unnecessary and unreasonable invasion of privacy. IoT devices, and especially here we emphasize measuring devices, pose a serious privacy and security risk, and users of such devices should most certainly be informed of what kind of data could such devices read. Although at first glance this obligation appears to be covered by the provisions of the GDPR, however, the issue of notifying end-users is much broader.

According to the GDPR, 'processing' means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means. [13] About that controller has obligation to inform a natural person.

Therewith, IoT ecosystem often has insufficient protection, and they are often left with default passwords, and it is also to be expected that many consumers won't even have many years of technical support and regular security upgrades. In such cases, potential for abuse is much broader than the data set collected by the operator and the obligation to notify the user. Therefore, the authors believe that European consumer protection regulations should cover both IoT and metering devices in such a way to impose an obligation of informing users of the capabilities of such devices, privacy threats, and measures to be taken by the users for their own safety; primarily changing initial device codes. Nevertheless, when introducing smart meters, the users should be explicitly aware of the capabilities and purpose of these devices.

As Tan noticed: "IoT ecosystem involves multiple parties in multiple jurisdictions in multiple countries." [22] Thus, the system only becomes more complicated and

prone to attacks by the third malicious parties and exploitation of all user information collected by the system through smart meters.

Recognizing the fact that IoT devices are increasingly being used by users who are not experts, and utilities do not even ask users if they would like to install such devices in their households, therefore we propose the aforementioned amendments to the Directive on consumer rights which would oblige utilities introducing smart devices, and even sellers of smart devices, to inform consumers of all the functionalities of these devices, the ability to collect data, the ability to read data from these devices, and the purpose for which data will be used, given that the most of the read data is actually unnecessary for the utility provider when issuing the invoice. The obligation to provide detailed information on device capabilities and their potential impact on privacy should be explicitly defined for metering IoT devices - which are installed without the option to refer the users to the capabilities of the device, and sometimes without even referring to the fact that the usual scale has been replaced by IoT device.

VI. REFERENCE

[1] A. Meola, "What is the Internet of Things (IoT)? Meaning & Definition", 2018, Available: https://www.businessinsider.com/internet-of-things-definition [Accessed 17 2 2020]

[2] M. Hung, "Leading to IoT" Available: https://www.gartner.com/imagesrv/books/iot/iotEbook_digital.pdf [Accessed 17 2 2020]

[3] "Risk or Reward:What lurks within your IoT?", 2017 Available: https://assets.kpmg/content/dam/kpmg/pl/pdf/2018/02/pl-Raport-KPMG-Risk-or-reward-What-lurks-within-your-IoT.PDF [Accessed 17 2 2020]

[4] G. Vojković, M. Milenković and T. Katulić, "IoT and Smart Home Data Breach Risks from the Perspective of Croatian Data Protection and Information Security Law", in *Proceedings of theENTRENOVA -ENTerprise REsearch InNOVAtion Conference*, Rovinj, Croatia, 2019

[5] F-Secure, "Attack Landscape H1 2019", 2019 Available: https://blog-assets.f-secure.com/wp-content/uploads/2019/09/12093807/2019_attack_landscape_report.pdf [Accessed 25 2 2000]

[6] H. Peterson, E. Baccelli, M. Wählisch and J. H. Schiller, "The Role of the Internet of Things in Network Resilience", in *Internet of Things. IoT Infrastructures. First International Summit, IoT360*, Rome, Italy, 2014

[7] R. Sagy, "Internet of things security: new connections, new threats", IBM, Las Vegas, 2016

[8] H. Suo, J Wan, C. Zou and J. Liu, "Security in the internet of things: a review", in *2012 4th International Conference on Computer and Automation Engineering (ICCAE 2012)*, Mumbai, India, 2012

[9] Manufacturer page, "Iskraemeco", 2019 Available: http://www.iskraemeco.com/files/5514/3982/5764/AM550.pdf [Acceseed: 18 2 2020]

[10] K. Weaver, "How Smart Meters Invade Individual Privacy," SkyVision Solutions, Available: https://smartgridawareness.org/privacy-and-data-security/how-smart-meters-invade-individual-privacy/ [Accessed 25 2 2020]

[11] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet and D. Irwin, "Private Memoirs of a Smart Meter", in *BuildSys 2010*, Zurich, Switzerland., November 2, 2010

[12] Information Commisioner Office, "Determining what is personal data, v1.1, 20121212" Available: https://ico.org.uk/media/for-organisations/documents/1554/determining-what-is-personal-data.pdf [Accesed: 19 2 2020]

[13] "Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation),L 194/1 OJEU

[14] "Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011 on consumer rights, amending Council Directive 93/13/EEC and Directive 1999/44/EC of the European Parliament and of the Council and repealing Council Directive 85/577/EEC and Directive 97/7/EC of the European Parliament and of the Council", L 304/64

[15] V. Paul and A. Von Dem Bussche, "The EU General Data Protection Regulation (GDPR)", Springer International Publishing, 2017

[16] R. Srinivas, "10 IoT Security Incidents That Make You Feel Less Secure" CISOMAG, 10 January 2020 Avaiable: https://www.cisomag.com/10-iot-security-incidents-that-make-you-feel-less-secure/ [Accessed 23 2 2020]

[17] C. Cimpanu, "New Silex malware is bricking IoT devices, has scary plans," ZDNet, 25 June 2019 Available: https://www.zdnet.com/article/new-silex-malware-is-bricking-iot-devices-has-scary-plans/ [Accessed 23 2 2020]

[18] S. Hilt, V. Kropotov, F. Mercês, M. Rosario and D. Sancho, "The Internet of Things in the Cybercrime Underground", Trend Micro Research, 2019

[19] G. A. Moore, "Crossing the Chasm, 3rd Edition: Marketing and Selling Disruptive Products to Mainstream Customers", HarperBusiness, 2014

[20] OWASP, "OWASP Internet of Things," 2018 Available: https://www.owasp.org/images/1/1c/OWASP-IoT-Top-10-2018-final.pdf [Accessed 25 2 2020]

[21] P. Collinson, "Is your smart meter spying on you?", The Guardian, 2017 Available: https://www.theguardian.com/money/2017/jun/24/smart-meters-spying-collecting-private-data-french-british [Acessed: 20 February 2020]

[22] S. Tan, " How safe is your personal data collected by your smart devices?", The Business Times, 2018 Available: https://www.businesstimes.com.sg/opinion/how-safe-is-your-personal-data-collected-by-your-smart-devices [Acessed 20 2 2020]

Use of security settings on social networks of elementary and high school students in the Split-Dalmatia County

R. Matković*, L. Vejmelka, Ž. Ključević***
* Public Health Institute of Split-Dalmatia County , Split, Croatia
** University of Zagreb, Faculty of Law, Zagreb, Croatia
roberta.matkovic@nzjz-split.hr, lvejmelka@pravo.hr, zeljko.kljucevic@gmail.com

***Abstract* - The use of the Internet and modern technologies provide various benefits in daily life and are almost impossible to avoid. Many forms of social behavior have "moved" into the virtual world. In the same time, potential dangers and risks for children and young people and also adults got their online dimension. During the 2016 NZJZ SDŽ conducted quantitative study to identify patterns of behavior on the Internet with a particular focus on Internet security, participation in cyberbullying, Internet addiction and parental controls when using the Internet. The study involved elementary school students (N=286) and high school students (N=539) making a representative sample for the Split-Dalmatia County. The online survey questionnaire was applied. The aim of this manuscript is to present a part of the results of the mentioned research with focus on the use of security settings on social networks of children and young people. Almost an equal proportion of elementary and high school students know how to use additional security settings on social networks (elementary school 85,7% and high school 85%). When asked if they use additional security settings on social networks, high school students use security settings more often than elementary school students (elementary school 69,2% and high school 74%). Furthermore, 22,4% of elementary school students and 25,6% of high school students have a publicly available profile on social networks. This habit is more commonly reported by boys (boys 28,8% and girls 16,3% in elementary schools and boys 36,2% and girls 9,7% in high schools).**

Keywords* - *internet usage, social networks, security settings, elementary school students, high school students

I. INTRODUCTION

The Internet and mobile technologies are a part of everyday environment. According to Global Digital Overview for 2019, 724,6 million people or 86% of population in Europe have access to the Internet [1] while the Croatian Bureau of Statistics data from 2019. show that 82% of population in Croatia use the Internet [2]. Monitoring that has been done in recent years clearly indicates a steady increase in both Internet users and active users of social media and mobile subscribers [1], and the data that 51,2% of all Internet traffic is done via mobile phones [3] indicates that the Internet is most commonly used via mobile phones. According to the already mentioned Global Digital Report for 2019, compared to the same source for 2018, an increase of 7,6% of the general population of Europe, an increase of active users of social media on mobile devices by 4,5% and an increase of active users of social media by 3,2 % [1].

It is the modern technologies that improve their performance year after year, the availability of Internet signal and the low cost of mobile Internet that influence the constant growth of users by changing their daily habits. The Internet offers many advantages, including the availability of information, facilitated communication and entertainment, but also disadvantages such as cyberbullying, misuse of information and unauthorized content sharing as well as overuse of Internet content that has a number of psychological consequences, including addiction.

Today, children and young adults are growing up with the widespread use of the Internet and mobile technologies. Croatian Bureau of Statistics (2019) states that children and young people in the 16-25 age group are the most frequent Internet users and that 99% of them use the Internet [2].

A. Use of social media

Social media are defined as “websites and applications that allow users to create and share content and participate in social networking” [4] or “that allow people to communicate and share information”. The social media revolution was launched with the founding of Facebook in 2003. Although the original idea of the founding was for the purpose of networking Cambridge University students, the idea quickly gained global interest and found its purpose except in networking and communication also in the business and marketing world. Since then, numerous social media with the same or similar characteristics have appeared on the Internet, such as Twitter, Instagram, Snapchat, etc.

There is almost no Internet user who doesn't have at least one profile on some social media. By creating a profile, users enter their first and last name, date and year of birth, the school they attend, list the sports club they cheer for, the musician or band they listen to, leave locations they have visited, post photos, videos, etc. All this information is crucial in user networking. In this way, users manage the impression of themselves, present themselves as they would like to be seen by others.

B. Children on social media

Children and young people communicate on social media about different aspects of life. They use social media for entertainment, learning, informing and communication with their peers and adults. There is a number of research on the use of social media. Thus, EU Kids Online states that in 2010 44% and in 2014 63% of social media users are children and young people in the 11-16 age group [5]. For 2018, we have data that 90% of young people in Croatia in the 16-24 age group use social media [6]. These data suggest that the popularity of social media among children and young people is unquestionable, and that the number of users in this age group is constantly increasing.

Social media have changed the communication paradigm since they appeared. Their original purpose was to connect users with each other, and with the increasing occupation of the internet space they became an integral part of the business and marketing world as well as their private lives. By creating profiles on social media, children and young people share their private information, photos and thus create an impression of themselves. Using the Internet and social media also entails numerous risks. On the Internet, children are also exposed to those who impersonate themselves and choose a child or a young person as a victim of various harmful online activities, whether it is sexual exploitation, shaming, exclusion, etc. The perpetrators may be children or adults. Social media security settings provide at least some level of privacy for the child and can protect the child's well-being in the virtual environment. Social media are also used by children and young people to connect with others of similar interests or traits, such as the school they attend, the club they cheer for, the musician they listen to, which gives them a false sense of connection, closeness and friendship. By sharing information about themselves, children and young people very often become victims of verbal violence, ridicule, sexual predators, etc.

There are numerous consequences of social media on daily life. The Internet and the media are becoming major factors in socialization, suppressing school and family and influencing the formation of values and behaviors of young people as well as the creation of their leisure time [7]. In support of this is the data according to EU Kids On-line, where 50% of children and young people aged 11-16 state that they rather spend time on the Internet than in real interaction [5]. Networking users are called "friends", and the concept of friendship has changed in general. The confidence of children and young people depends on the number of friends and likes on the post. Participating in initiatives, expressing opinions and pointing out views are facilitated. Social media have also changed emotional relationships. They make it possible to spy on and control partners, increase jealousy and special attention in a sentimental relationship is given to the publishing of a relationship status on Facebook [8]. Social media in the lives of children and young people have almost erased a boundary between the real and the virtual world. These two worlds are very intertwined for them, relationships continue and flow from one world to another and we can say that the Internet is certainly a virtual reality for new generations.

Everything that children and young people “set on” the Internet, every post, photo, every piece of information or location, this is all information that can endanger their safety. There are numerous websites that display social media from second to second. In one minute Facebook has 360 new users, 150 000 messages have been sent, 292 980 statuses have been posted, over 100 000 friend requests have been sent, over 136 000 new photos have been published [9]. One such "digital jungle" is endangering the privacy and security of users. The protection of private content is impaired due to the possibility of hacking, but also the media illiteracy of the users themselves, who often impulsively present facts and retell their own events for which social media are not a safe place [7]. Thus, with many advantages and known disadvantages of using the Internet and social media, there are "new" risks, such as compromised privacy and security, cybercrime and the prevalence of inappropriate content [10]. Social media offer the ability to protect profiles and set security settings. According to EU Kids Online, children often do not understand what privacy means or they are not familiar with security settings they can use [5]. Exposure to online risks can have long lasting and intense negative consequences for every person, and especially for young people, who are not always critical about the risks and are often not even aware of them [10]. Therefore, additional efforts need to be made to bring safety settings closer to children in order for their privacy to be protected.

The topic of this paper is the presentation of the use of security settings on social media by children and young people in elementary and high schools in Split-Dalmatia County. The descriptive research presented here is a part of a broader research project on the use of the Internet by children and young people in this county. Since this is a descriptive presentation of the data, it is not necessary to set hypoteses of the research.

II. METODHS

The aim of this quantitative study is to identify patterns of behavior on the Internet with a particular focus on social networks security settings.

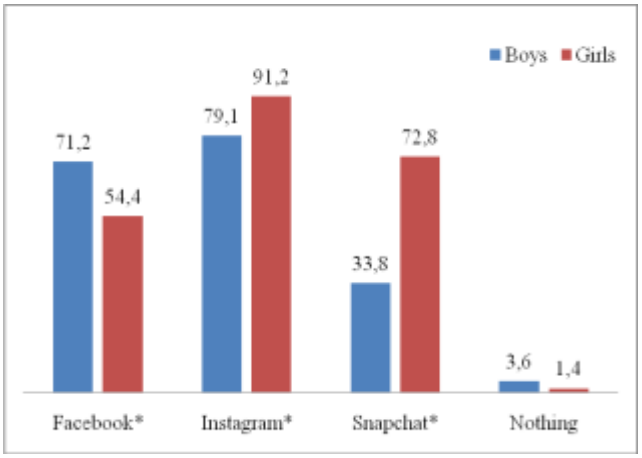
The study involved 16 elementary (N = 286) and 11 high schools (N = 539) in a representative sample consisted of the 3% of the population. Gender distribution shows 48,6% of boys and 51,4% of girls from elementary schools, and 59,9% of boys and 40,1% of girls in high schools. The average age of elementary school students is 14 years and the average age of high school students is 16 years.

An online questionnaire was used to collect the data, filled in individually by the students during their school / classroom stay. The questionnaire contained several fields, a *Socio-Demographic Questionnaire* designed for the purpose of this survey, as well as a *Questionnaire on child safety on the Internet* that included questions about parental control and the characteristics of interactions on social networks, and the use of settings on social networks, which will be presented in this paper. The survey also used a LaRose and Tsai [11] adapted *Online activity questionnaire for children* that contained 16 variables that examined time spent on the internet in individual activities (eg, social networks, gambling, chat room, internet forums, bloging, instant messaging…) answered on a 6-point scale (0 = never to 5 = more than 4 hours). The research was conducted in collaboration with teachers of computer science or professional assistants and teachers in schools. Detailed instructions and guidelines for conducting the research were sent to the research associates. The online questionnaire form is adapted for mobile phones use. The survey lasted a maximum of 20 minutes, all contemporary ethical standards for the research with children were implemented and the research team was fully available for any additional questions.

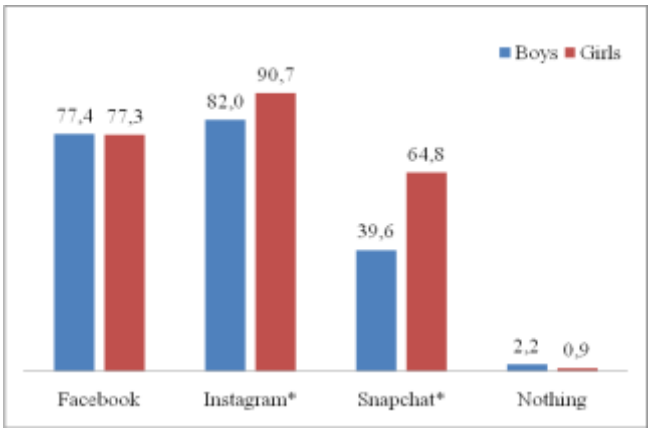
III. RESULTS AND DISCUSSION

According to the research results, children and young people most commonly use Instagram, 85,3% in elementary and 85,5% in high schools (Pictures 1 and 2). Another popular social media is Facebook, used by 62,6% of elementary school students and, more often, 77,4% of high school students.

According to internal data from the Public Health Institute of Split-Dalmatia County collected on a representative sample of high school students in Split (N = 1088), the results show that in 2013, when the research was conducted, Facebook was the most popular social media among high school students in Split, with 94,4% of users, Instagram had 9,9% of users, and no student had a Snapchat. Recent research results indicate a decline in the popularity of Facebook among children and young people, as well as an increasing trend of Instagram users in the above mentioned age group.



PICTURE 1. SOCIAL NETWORK PROFILES OF ELEMENTARY SCHOOL STUDENTS



PICTURE 2. SOCIAL NETWORK PROFILES OF HIGH SCHOOL STUDENTS

Similar results are shown by HR Kids Online research, which also points out that Facebook is still the most used social media in Croatia between the ages of 9-17, while Instagram is ranked second [12]. Without intention to compare the results, we can conclude that social media are a very lively media that struggles daily for new users and for retaining old users with new and improved options.

If we compare profiles according to gender we can find differences. Female students are more likely to have an Instagram profile and this gender difference is statistically significant (Instagram - Elementary School Students: $\chi^2 = 8,239$; $p = 0,004$ and Instagram – High School Students: $\chi^2=7,911$; $p=0,005$), as well as Snapchat (Snapchat - Elementary School Students: $\chi^2 = 43,672$; $p = 0,000$ and Snapchat – High School Students: $\chi^2 = 32,845$; $p = 0,000$). As for Facebook, elementary school boys are more likely to have a Facebook profile (Facebook - Elementary School Students: $\chi^2 = 8,613$; $p = 0,003$), which can be explained by playing virtual games provided by the Facebook platform.

Only 2,4% of elemetary school students and 1,7% of high school students do not have a single profile on social

media, which supports the view that the vast majority of the Internet users are social media users. Also, the high proportions of having profiles in student population leads to the conclusion that students have two or more profiles at the same time.

As a benchmark for assessing Internet users' habits, it is certainly time spent on the Internet. The research participants were offered fifteen different services and time frames within which they evaluated the time spent on each of these services, on two devices - a mobile device and a computer. Only the time spent on social media on a mobile device and a computer in elementary and high school students is shown here.

Students are more likely to use mobile phones to browse content on social media, which is expected given the full benefits of the mobile Internet, but also due to the fact that a large number of students own this device. Four and more hours are spent on social media by 15% of elementary school students (Table 1) and 20,6% of high school students (Table 2).

On the other hand, 5,2% of elementary students (Table 3) and 6,9% of high school students (Table 4) spend 4 hours or more on a computer. Gender differences in both sub-samples were not statistically significant. Research by Vejmelka et al. [10] on a representative research of high school students shows that students spend between 3 and 4 hours a day on social media.

TABLE 1. TIME SPENT ON SOCIAL MEDIA ON A MOBILE DEVICE IN ELEMENTARY SCHOOL STUDENTS

	Boys		Girls		Total	
	n	%	n	%	n	%
Never	37	26,6	41	27,9	78	27,3
Up to 30 min.	53	38,1	38	25,9	91	31,8
Up to 1 h	12	8,6	17	11,6	29	10,1
1-3 h	19	13,7	26	17,7	45	15,7
4-8 h	3	2,2	13	8,8	16	5,6
>8 sati	15	10,8	12	8,2	27	9,4
Total	139	100	147	100	286	100

TABLE 2. TIME SPENT ON SOCIAL MEDIA ON A MOBILE DEVICE IN HIGH SCHOOL STUDENTS

	Boys		Girls		Total	
	n	%	n	%	n	%
Never	56	17,3	22	10,2	78	14,5
Up to 30 min.	97	30	57	26,4	154	28,6
Up to 1 h	57	17,6	41	19	98	18,2
1-3 h	56	17,3	42	19,4	98	18,2
4-8 h	25	7,7	26	12	51	9,5
>8 sati	32	9,9	28	13	60	11,1
Total	323	100	216	100	539	100

TABLE 3. TIME SPENT ON SOCIAL MEDIA ON A COMPUTER IN ELEMENTARY SCHOOL STUDENTS

	Boys		Girls		Total	
	n	%	n	%	n	%
Never	59	42,4	76	51,7	135	47,2
Up to 30 min.	48	34,5	46	31,3	94	32,9
Up to 1 h	13	9,4	11	7,5	24	8,4
1-3 h	10	7,2	8	5,4	18	6,3
4-8 h	3	2,2	2	1,4	5	1,7
>8 sati	6	4,3	4	2,7	10	3,5
Total	139	100	147	100	286	100

TABLE 4. TIME SPENT ON SOCIAL MEDIA ON A COMPUTER IN HIGH SCHOOL STUDENTS

	Boys		Girls		Total	
	n	%	n	%	n	%
Never	142	44	101	46,8	243	45,1
Up to 30 min.	103	31,9	70	32,4	173	32,1
Up to 1 h	38	11,8	23	10,6	61	11,3
1-3 h	14	4,3	10	4,6	24	4,5
4-8 h	4	1,2	3	1,4	7	1,3
>8 sati	22	6,8	9	4,2	31	5,8
Total	323	100	216	100	539	100

Time estimation spent on any service, including social media, is an important piece of information, but at the same time, mobile devices and the mobile Internet make it impossible for children and young people to more realistically estimate the time they actually spend on the Internet. They do not have the "internet time" because their mobile device is constantly available to them, "notifications" keep coming and in fact, children and young people are constantly available on the Internet and they spend much more time on using the Internet and social media than they are aware of. According to the results of a National Research Conducted by the Child and Youth Protection Center of, almost every other adolescent immediately checks social media when a "notification" comes, one in three checks them several times a day, and every fifth adolescent in a social media checks constantly, regardless of whether the notification comes or not [13].

One in five elementary school students and one in four high school students state that they have an open profile on social media, publicly available to everyone (Tables 5 and 6). Their personal information is exposed to the public, and students with publicly available profiles are at risk. These data suggest that prevention activities through informing children about the harmful consequences of unprotected data and practical teaching about data protection should begin in elementary school before children “create” their profiles on social media. There is a significant gender difference in high school students (χ^2 = 0,000), boys have publicly open profiles more often (Tables 5 and 6), which is consistent with studies on adolescent risk behaviors in general, such as

intoxication, driving without seat belt fastened, problematic gambling, engaging in fights, etc. [14, 15]

We asked our research participants if they know how to use security settings and whether they use them on the social media on which they have created profiles (Tables 5 and 6). Students in both sub-samples alike, 85% of them say that they know how to use security settings. Here we can bring up a shortcoming of this research because we categorically asked students if they knew how to use security settings and we did not have available instruments to measure what they meant by security settings. The same shortcoming can also be applied to answers to the question if participants use security settings at all.

TABLE 5. PUBLIC AVAILABILITY OF PROFILES, KNOWLEDGE AND USE OF SECURITY SETTINGS ON SOCIAL MEDIA IN ELEMENTARY SCHOOL STUDENTS

	Boys		Girls		Total	
	n	%	n	%	n	%
Social media profile publicly available to everyone	40	28,8	24	16,3	64	22,4
Knowledge how to use additional security settings on social media	114	82,0	131	89,1	245	85,7
Use of additional security settings on social media	87	62,6	111	75,5	198	69,2

TABLE 6. PUBLIC AVAILABILITY OF PROFILES, KNOWLEDGE AND USE OF SECURITY SETTINGS ON SOCIAL MEDIA IN HIGH SCHOOL STUDENTS

	Boys		Girls		Total	
	n	%	n	%	n	%
Social media profile publicly available to everyone	117	36,2	21	9,7	138	25,6
Knowledge how to use additional security settings on social media	267	82,7	191	88,4	458	85,0
Use of additional security settings on social media	226	70	173	80,1	399	74,0

Every third elementary school student and every fourth high school student does not use security settings on social media. According to HR Kids Online, one in five children between the ages of 9 and 17 is completely or mostly unable to change their privacy settings, for ex. on social media. According to the same research, one in four children between the ages of 9 and 14 and every third child between the ages of 15 and 17 are completely or mainly concerned about their privacy on the Internet [12].

IV. CONCLUSION

Children and young people are exposed to the dangers of the digital world through social media and public, unprotected profiles where they publish various information about themselves and their environment. By staying in a secure "real" environment while being surrounded by "virtual" friends they don't give importance to the Internet security. According to the results of the research presented, some children and young people who have publicly open profiles, know how to use the security settings, but do not use them and they didn't set them on their profiles on social media. Unfortunately, compromised privacy and security on social media can compromise their security both online and in real life, but children and young people are often unaware of the dangers they face. These data suggest that prevention activities through informing children about the harmful consequences of unprotected data and practical teaching about data protection should begin before children “create” their profiles on social media. A key role in safety is played by adults, parents and the environment, teachers and professionals who should certainly include the topic of the Internet and behavior on social media in their daily communication with children. It is important to emphasize that these results point to the need to develop more specialized prevention programs that will include the use of social security settings and digital tools. The practical application of the results is possible in the direction of developing guidelines, info flyers and campaigns with practical and specific guidelines for setting up the privacy settings for different social networks and digital tools. It is recommended to establish cooperation with service providers, telecommunications companies and social media administrations in promoting the use of privacy settings. It is an indispensable responsibility of the service providers, i.e. the social media themselves, to make the security settings more accessible and transparent and to promote their safe use. And in general, continuous and accessible education of parents, teachers and students about the possibilities and importance of using privacy settings is necessary. All these actors need to be warned about the risk of publicly disclosed personal information and the possibility of protection that includes adequate privacy settings on user profiles. Certainly, the academic community has a task to explain the importance of maintaining privacy on the Internet through additional research efforts and to contribute to a safer environment through reaching and preserving an appropriate level of children's well-being on the Internet and social media.

LITERATURE

[1] S. Kemp, "Digital 2019: Global Digital Overview." Datareportal. Accessed January 15, 2020. [Online]. Available: https://datareportal.com/reports/digital-2019-global-digital-overview

[2] Croatian Bureau of Statistics, „Usage of Information and Communication Technologies (ICT) in Households and by Individuals, 2019, First Results," 2019. Accessed January 15, 2020. [Online]. Available: https://www.dzs.hr/Hrv_Eng/publication/2019/02-03-02_01_2019.htm

[3] J. Clement, (2020, Feb 13) „Share of mobile internet traffic in global regions 2020." Statista. Accessed Feb. 15, 2020. [Online]. Available: https://www.statista.com/statistics/306528/share-of-mobile-internet-traffic-in-global-regions/

[4] T. Dokman, M. Kuzelj and D. Malnar, "Društvene mreže u ulozi modernog oružja – percepcija doktoranada," *Polemos*, vol. XXI, no. 41, pp. 133-150, 2018. Accessed January 15, 2020. [Online]. Available: https://hrcak.srce.hr/207140

[5] S. Livingstone, „EU Kids Online." Accessed January 15, 2020. [Online]. Available: https://lsedesignunit.com/EUKidsOnline/html5/index.html?page=1&noflash

[6] Shaping Europe's digital future, "Digital Economy and Society Index Report 2019. Use of internet services," Accessed January 15, 2020. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=59977

[7] I. Kunić, M. Vučković Matić and J. Sindik, "Korištenje društvenih mreža kod učenika osnovne škole. The use of social networks among elementary school population," *Sestrinski glasnik*, vol. 22, no. 2, pp. 152-158, 2017. Accessed January 15, 2020, doi: 10.11608/sgnj.2017.22.032. [Online]. Available: https://doi.org/10.11608/sgnj.2017.22.032

[8] E. Družeta and N. Zgrabljić Rotar, "Uloga Facebooka u sentimentalnim odnosima mladih," *Medijske studije*, vol. 8, no. 16, pp. 78-97, 2017. Accessed January 15, 2020, doi: 10.20901/ms.8.16.6. [Online]. Available: https://doi.org/10.20901/ms.8.16.6

[9] "Social Media in Real Time- Interactive infographic," Social Hospitality. Accessed January 15, 2020. [Online]. Available: https://socialhospitality.com/2016/01/social-media-in-real-time-interactive-infographic/

[10] L. Vejmelka, N. Strabić and M. Jazvo, "Online aktivnosti i rizična ponašanja adolescenata u virtualnom okruženju," *Društvena istraživanja*, vol. 26, no. 1, pp. 59-78, 2017. Accessed January 15, 2020. [Online]. Available: https://doi.org/10.5559/di.26.1.04

[11] R. LaRose & H.-y. S. Tsai, „Completion rates and non-response error in online surveys: Comparing sweepstakes and pre-paid cash incentives in studies of online behavior," Computers in Human Behavior, vol. 34, pp. 110–119, 2014. Accessed January 15, 2020. [Online]. Available: https://doi.org/10.1016/j.chb.2014.01.017

[12] HR Kids Online, "Nacionalno komparativno istraživanje o medijskim navikama djece i njihovih roditelja te sigurnosti djece na internet." Accessed January 15, 2020. [Online]. Available: http://hrkids.online/post/fifth-press/

[13] Poliklinika za djecu i mlade grada Zagreba, "*Društvena online iskustva i mentalno zdravlje mladih*." Accessed January 15, 2020. [Online]. Available: https://www.poliklinika-djeca.hr/istrazivanja/rezultati-nacionalnog-istrazivackog-projekta-drustvena-online-iskustva-i-mentalno-zdravlje-mladih/

[14] Hrvatski zavod za javno zdravstvo, " Europsko istraživanje o pušenju, pijenju i uzimanju droga među učenicima ESPAD 2015," 2016. Accessed January 15, 2020. [Online]. Available: https://www.hzjz.hr/wp-content/uploads/2016/10/HR_ESPAD_2015_RGB_3.pdf

[15] T. Maglica, "Obilježja roditeljstva i odnos roditelja prema kockanju kao prediktori kockanja muške djece," Ph.D. dissertation, Edukacijsko rehabilitacijski fakultet, Svučilište u Zagrebu, Zagreb, Hrvatska, 2018. Accessed January 15, 2020. [Online]. Available: https://urn.nsk.hr/urn:nbn:hr:158:234646

Visual Analysis of Similarity and Relationships Between Legal Texts

J. Opiła*, T. Pełech-Pilichowski*
*AGH University of Science and Technology, Cracow, Poland
jmo@agh.edu.pl
tomek@agh.edu.pl

Abstract – **Visualization of texts of legal acts, based on an assumption of high level of their informativeness, goes beyond the process of visualization of a text in a given file format. The large dimensionality of databases of legal texts, complex relationships between them (including multi-level and often time dependent references), ambiguity of terms used in various legal acts, specificity of regulations in different countries of EU and a number of other factors make difficult effective and ergonomic visualization of information to a user, including various user groups/profiles (for example, lawyers, judges, entrepreneurs). The article presents the issue of processing legal texts as non-standard texts, difficult to process with the natural language processing (NLP) tools. The aim of such tools is to achieve a high processing efficiency for input data containing texts written in everyday (common) language, mainly English one. Problems of visualization of texts of legal acts as a graphical presentation of the values and relations of attributes in analyzed legal text are discussed. Issues of "attributing" of legal texts are outlined to give a possibility of their further processing aimed at information and knowledge extraction, in particular for the purposes of analyzing a similarity of texts and their visualization.**

Keywords - Legal texts, Legal acts, EU Law, Visualization, Attributization, NLP, Analysis of text similarities.

I. INTRODUCTION

Automated analysis of legal text is an important and promising branch of practical application of numerical algorithms aimed at information retrieval from large input datasets. Such data may represent, for example, a set of documents, a domain/field of law, a collection of objects selected from input data etc. In general, considering a process of legal text analysis, it can be viewed as a text processing, capable of obtaining valuable results as an effect of usage of Natural Language Processing (NLP) algorithms [12]. Nevertheless, legal texts differ from typical ones (common English written) for which a number of algorithms has been developed. Such differences are visible in a language being analyzed and processed, including internal/external relationship structure for description of connections between parts of documents and between objects. Thus, it is valuable to evolve approaches, methods, algorithms and paradigms for legal text processing, in particular – considering legal acts written in various national languages.

For purposes of digital processing, legal texts have to be preprocessed. Such a process is dedicated to transform raw input data into structured form expressed by attributes related to a processed text. In addition, it can be used for removing redundant data (expressed as multiplied sentences, definitions, references etc.). Another important task is an identification of a network of internal relationships inside the analyzed textual object as well as external ones between objects. Furthermore, such approach, differ from common (typical) text processing, gives a potential opportunity of obtaining a supplementary (additional) information, valuable for instance for comparing legal texts, identification of mistakes and contradictions, detection of unauthorized changes in legal texts, finding intentional and authorized changes but also having involuntary and pernicious side effects (for example unmeant cancellation of related Act). As a result of attributing followed by a sophisticated information retrieving, large numerous sets of values describing structure of relationships may be produced. Such a big sets of data are very difficult for manual processing and analysis. In order to achieve stable, usable, user-friendly and ergonomic presentation of produced results dedicated visualization algorithms have to be worked-out and then employed for a display of aggregated information and, in the end, comprehensive show of gathered knowledge [15].

In the course of research, it was found that a modified Gantt diagram may be useful for that goal. While 2D version of the chart is well known and widely used in Project Management, due to complexity of relations in system of legal Acts, three-dimensional version ought to be elaborated. Furthermore, in order to visualize temporal and hierarchical relations between Acts, a directed graph should be employed. As a result, hybrid visualization design is proposed based on a Gantt modified design and directed graph scheme.

Therefore, usability of modified Gantt and directed graph designs were tested for visual reproduction and analysis of dependencies and similarities in the texts of Legal Acts. For the sake of simplicity, it was decided to restrict research to Parliamentary Acts only.

Included graphics were produced using *graphviz* – OpenSource graphing utility [9] and *Inkscape* – OpenSource vector graphics program [10].

This paper is funded by the National Science Centre, Poland under project no. 2017/27/B/HS5/01376

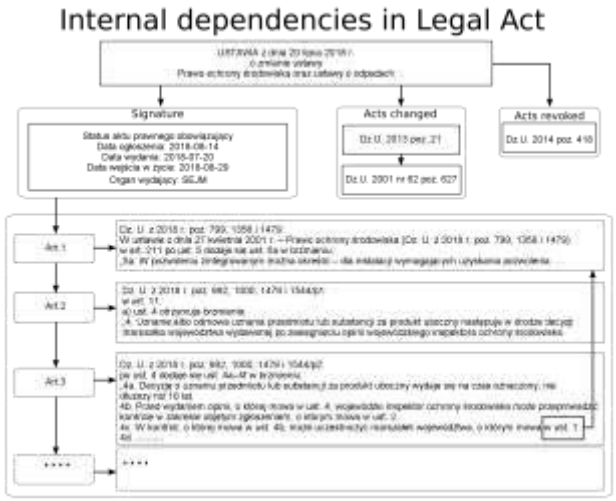


Figure 1. Schematic view of dependencies in selected Polish Legal Act shown as directed graph. Image for illustratory purposes only so text was not translated. Prepared using visualization toolkit graphviz. Source: own study.

II. DEPENDENCIES BETWEEN LEGAL ACTS

Dependencies (relationships) between legal texts are complex, multilevel and multidimensional. They are complex because they have to cover a number of legal texts of different status (rank), size (in a number of characters, words etc.), branch of the law (domain/(sub)domain) and legal acts prepared/created by different authors (in particular, legislators). Relationships between legal texts are multilevel because an act should be strongly connected with different objects placed in different level of dependencies (relationships). In addition, a part of an act (for example, a paragraph) can refer to a particular text (or vice versa), but another/further part (a paragraph) of the same legal text can have a relation to completely different law domain/legal text/user. From this point of view (especially for an algorithmic processing) an object may be viewed as input data (in particular, text data enriched by attributes) for an algorithm devoted to obtain output data (a result) based on a finite sequence of predefined operations.

Processing of legal texts differs from typical processing of text written in common (and English) language which is widely based on employing Natural Language Processing (NLP) solutions, methods and approaches. Legal texts, as an input dataset for further processing with dedicated algorithms, have to be treated as specific objects. Legal texts contain ambiguities (for example, a term can be described by different definitions in a number of legal acts), complex and cause-effect dependencies between objects (objects can be considered, for example, as sentences and paragraphs). Legal acts change in time because of a necessity of adapting regulations to changes in particular domains and an environment (i.a. adapting law to technology). But a change in a text from a set objects (a domain/subdomain) can entail changes in related documents (from the same domain/subdomain), both in content (modifications, translations, clarifications, but also adding/removing of data) and relationship structure.

A relationship's structure may be viewed as a description of a (sub)domain referred to a branch of the Law or its subset. Note, that a subset as an input data intended for further processing, can be identified, for example, as a part of a regulation (consistent or inconsistent), a set of paragraphs, a set of legal texts etc. Relationships can be viewed as internal and/or external. They directly indicate dependencies between objects. Internal relationships related to an object (a dataset) can describe connections (both direct and indirect) between subsections, parts, components and other attributes of the object. By default, an internal relationship may be a mapped typical (common) structure of a document, i.e. it can cover a division into sections, headers, paragraphs etc. Nevertheless, such a simple structure is not adequate for advanced digital processing aimed at information retrieving. External relationships indicate dependencies between objects – of the same and different type, level, domain/subdomain/field, range, language and many others.

Algorithms for legal texts processing have to be dedicated (and also adjusted) to input datasets (objects). Relationship structures and dimensionalities of datasets representing particular law (sub)domains are, in details, different. Moreover, such differences can be also observed between corresponding texts from the same (sub)domain but applied in general (for particular country, province, state etc.). Thus, a processing algorithm efficient for a particular text (or a set of texts) does not necessarily have to be suitable to a similar text, even of the same type.

An analysis of a legal texts can be focused on the use of algorithms for syntactics analysis (for example, counting of words, applying N-grams methods [16]) or semantic one [17] – more advanced, complicated and complex. The goal of a semantic analysis is to retrieve information from data, especially human-like meaning of processed words, sentences and paragraphs. Acquisition of information can be viewed as a data enriched with meaning. Considering typical texts (written as common English language) such aim is tough but promising. On the other side, considering retrieving of meaning from input data in the form of legal text is a hard task. Meaning of legal text is not just an information get as a result of an analysis with a dedicated algorithm. According to relation between law and a legal system, in fact, meaning obtained as a result of digital analysis, can be viewed as factor which covers both information obtained as an effect of processing of legal text and information/knowledge existing and/or rising as an effect of interpretation by lawyers (opinions of judges, Supreme Court judgements, implicit (typically, unwritten) rules of social coexistence, experience of lawyers etc.).

One of important aims of digital legal text processing is to identify a similarity between objects and, in particular to detect internal and external dependencies. Such an identification may support analysis of changes in texts, automated detection of relationships as well as support for legislation process, including translations to local languages and adaptation to national legal systems. Considering legal text processing conditions, briefly outlined above, a similarity analysis between texts can be treated as a comparison of results of both the syntactic (partial) analysis and semantic one [17]. Such analyses are, typically, an estimation of a factor (a value, a coefficient) indicating a level (degree) of similarity. An analysis of dependencies between texts can be also achieved with visualization process as a method for processing input data

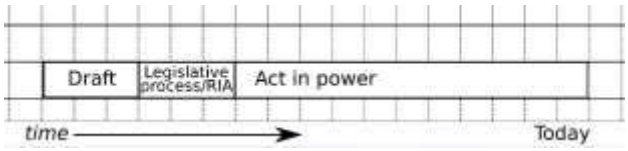


Figure 2. Sample Gantt chart of a Legal Act's lifeline. 2D visualization. Source: own study.

into a graphics – in order to obtain an ergonomic presentation of information acquired from data.

III. VISUALIZATION OF DEPENDENCIES

As mentioned above, carefully thought-out dependency visualization schemes can provide significant assistance in analysis of the text of legal act and in turn Text of Law. This applies to all its aspects, from the visualization of mutual relations between source acts (texts of legal acts/Legal Acts), through the visualization of references, to the visualization of the semantic structure of individual acts of making or applying the law.

During the legislative process, but also the application of law, it is extremely important to detect pathological dependencies in order to eliminate ambiguity of law. This can be achieved in several ways, one of which is a thorough analysis of the text of the draft legal act. Unfortunately, the dependencies inside the legal system are so confusing that even a qualified team may overlook something. Therefore, new analytical tools are needed – they can be, on the one hand, systems for the automatic analysis of the text of a draft legal act, operating on the basis of complex ontological systems, and on the other, systems for visualizing the structure of the legal act, taking into account the mutual multilevel dependencies with other acts. Both of these tools should be used simultaneously, because they show a high degree of synergy. The analysis results developed by automatic systems are best communicated to the user with the use of visualization, while the raw structure visual analysis can help detect unusual abnormal relationships difficult to detect by automatic systems. Overlooking of the delayed date of entry into force of the provision or, on the contrary, the individual expiry of the statutory provision before the end of the entire act may be a good example.

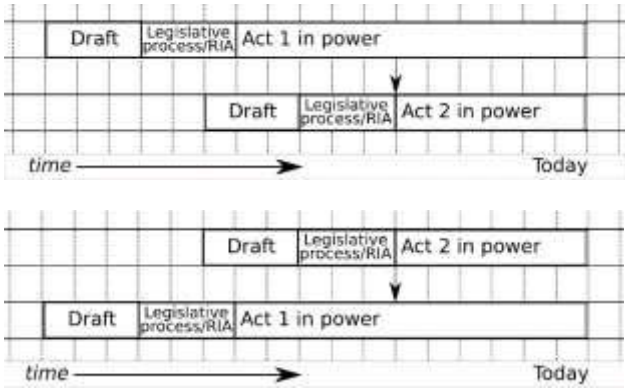


Figure 3. Sample Gantt chart of temporal dependencies between texts of two legal acts. The "Act 2" properly references "Act 1"(upper) but in the lower picture, Act 1 references Act 2 temporarily improperly. Reference number is not depicted. Source: own study.

Applied technologies, visualization techniques and designs should be the subject of intensive studies of perception, informativeness and efficiency, and should be subject to continuous development. While in some applications it will be effective to use graph theory to visualize various relationships, in others, for example, to visualize temporal relationships, it would be better to use the graph metaphor Gantt. In many cases, it may be expedient to use 3D graphics techniques, including interactive technologies, including augmented and virtual reality.

The great advantage of the graph is the direct indication of abandoned nodes / leaves (with no access and/or exit), blind edges (no target) or empty (i.e. void, as an indicated target has ceased to exist or never existed) as well as edges from nowhere (no starting node). As mentioned previously, dependencies within the legal system, for example between legal acts, may be, inter alia: explicit, indirect, implicit or sometimes hidden:

- explicit – the legal act explicitly cites another legal act clearly indicating its reference number,
- implicit/indirect – these dependencies result from general principles of law, for example, every legal act implicitly depends on a higher-ranking act, e.g. acts are subject to the Constitution,
- hidden – these are dependencies (or lack thereof) that are deliberately but rather accidentally created during the legislative process or when applying the law. It may be a concept not defined in the text of a legal act but having the usual linguistic interpretation, or on the contrary – defined in the text of a legal act, but having a different meaning than the commonly accepted.

If the nodes (vertices) take role determining the structure of the text of the law, and the leaves (leaf – terminal node) contain specific semantic units (the minimum piece of the text denoting the separated whole – the object, relation or action) then the visual analysis of the graph of the text of the law Act should quickly indicate nodes without leaves or solitary leaves, not related to other elements by any relation shown as graph edges. On the other hand, subsequent versions of a given semantic unit can be formed into a FILO stack (First In, Last Out), where the most current record is at the top of the stack and its original value is at the bottom (see Fig. 1).

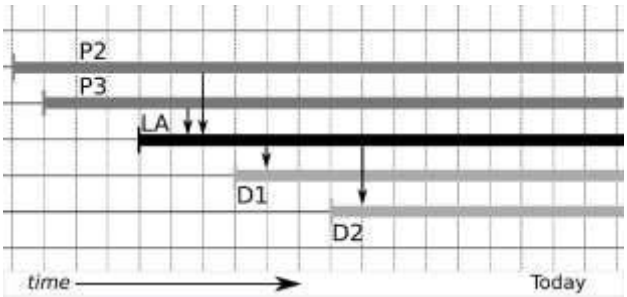


Figure 4. Several Acts connected with relationships of dependence. 2D visualization. Source: own study

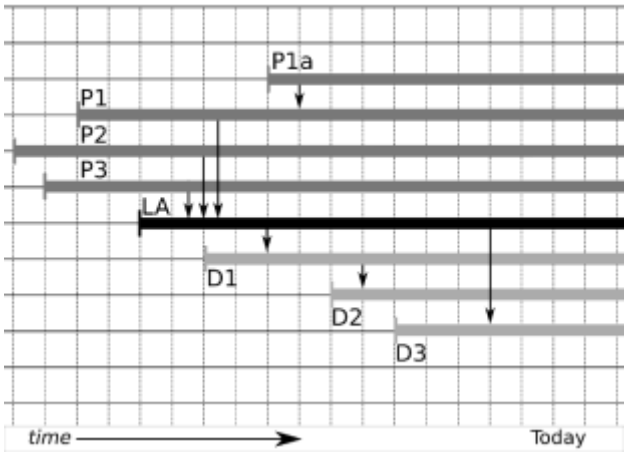


Figure 5. Several Acts connected with relationships of dependence. 2D visualization. Act P1a introduces amendation of the original Act P1. Source: own study

It should be noted that the data base of legal acts must operate in the data warehouse regime. As a rule, this type of database allows only adding new elements with an appropriate time stamp and their reading out at any time. To illustrate this process and indicate the current state of the legal text a properly adapted Gantt chart, often used in project management, is suitable. However, it must be adapted to the specifics of the issue, by adding a visualization of changes in the "lifeline" of each text of the legal act – creation, modification, suspension, merging with another text or derogation (see Fig. 2). Thanks to the linear structure it is possible to easily capture a snapshot of the right text at any time step.

Further considerations for the simplification of discourse were limited to legal acts in the rank of the Parliamentary Act only.

Figure 2 presents visually the most important stages of the legislative process for the case of a typical Act using slightly modified Gantt design. After submitting a draft of the bill to Parliament (in the manner provided for in the Constitution), the legislative process begins, in which it is established practice to examine the effects of legal regulation (*Regulatory Impact Analysis* – RIA). After the Act is passed and signed by the person in charge, usually The President, it enters into force and is valid until the derogation or expiration, if it occurs.

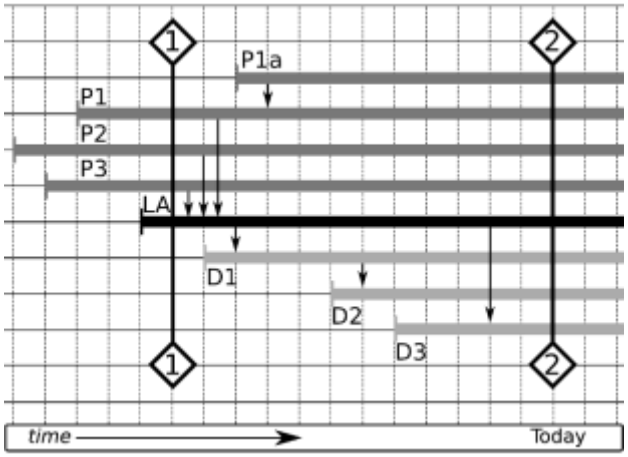


Figure 6. Two timesteps (1-1 and 2-2)are shown as crossections of the Gantt chart. Vertical grid symbolizes midnights. This image is a base for later shown directed graphs. Source: own study

However, it is a very rare case that given Act does not depend on any other Act but The Constitution only. Most often Acts depends on at least one other Act. This is illustrated in the Fig. 3. The upper visualization shows that Act 2. refers Act 1 (shown using arrow). In this case it is clear that reference is proper, as Act 1. is in charge when Act 2. is being enacted. On the other hand, lower image shows that Act 2. is referred improperly in Act 1., namely it is referred before Act 1. went into power thus, there is a period when Act 1. is not valid. This simple visualization helps detect serious inconsistency in the regulation.

Figure 4 shows even more complex situation. Five Acts are depicted there. As Act "LA" is under analysis it is emphasized by solid black. It depends on "P2" and "P3" Acts and is referred by Acts "D1" and "D2". Hierarchy of dependency is shown using arrows. Now it is clear, which Acts have to be checked doing regulatory impact analysis. Changes in "LA" may affect meaning of "D1" and "D2", while improper amendation of "P2" and "P3" may left "LA" useless.

Another case is shown in the Figure 5. Still discussed is dependency of Act LA on some Acts (P1, P1a, P2, P3) and influence on others (D1, D2, D3). Act LA refers to Acts P1, P2 and P3 directly, but Act D2 indirectly refers to Act LA, as it refers Act D1 which directly refers to LA. Inspecting the visualization one may easily spot this and thus conclude that some change to LA may affect D2 in unpredictable and involuntary way. There is also another clue concerning influence of amendation of Act P1 by Act P1a. Thanks to visualization it is clearly visible that change of single article or even double tiret introduced by the text of amendation may heavily influence other Act not necessarily thematically connected with Act amended. Act. For example, a slight change in the definition of a "public road" in the Polish road traffic law [1][3] many years ago has affected Art. 89 of the Code of Offenses in an unforeseen manner[1] [4] what forced the Court to ask prejudicial question to the Constitutional Tribunal of Poland and started intense discussion in Polish legal press [5]. It must be stressed that in reality Acts are interconnected far more than shown in simple examples above.

The Text of the Law is dynamic and changes more often or less often and it is an ordinary and unstoppable process, because the law must be adequate to the constantly changing society and technology. Fortunately, all changes are at certain time points, i.e. at midnight. The acts shall enter into force at midnight on the first day of validity and shall apply until midnight on the last day of validity. Properly designed visualization of dependencies between Acts should take this into account. On every timestep must be possible to credibly reproduce complete hierarchy of dependencies up to the Constitutional Law. This is illustrated in Fig. 6 which is derived from previously

[1]In the previous regulation, the term "road" was limited only to "roadways", but in the new regulation, to the "road" belong, a.o. pavements.

discussed Fig. 5. Two elements are added symbolizing timesteps at cross sections 1 and 2 (denoted using rhomboids). Thus, although it may be easily concluded that Act LA depends on P1, P2 and P3 Acts but it cannot be easily stated if dependency is of direct, indirect, equal or not equal rank. This may be solved using directed graph as shown in Fig. 7.

Now it may be easily spotted that dependency of Act LA on Acts P1, P2 and P3 is of the same rank, i.e. primary and dependency is explicit and direct. Things get worse at cross section 2 as hierarchy of dependencies gets deeper. Resulting graph is shown in the Fig. 8. Now Act LA depends directly on Acts P2, P3 and indirectly on amendation of Act P1 ie. P1a. Furthermore, Acts D1 and D3 refers Act LA directly and explicitly while Act D2 explicitly but indirectly. This scheme helps understand complexity of the whole hierarchy and explicitly shows that amendation P1a of Act P1 may affect not only directly dependent Act P1 but Acts LA, D1, D2 and D3 as well.

Number of references in typical Act may become large in complex or amended many times Acts like Tax Law. Thus, relevant graph accordingly may considerably grow exceeding dimensions suitable not only for computer display but for printout too. In this situation, it is necessary to implement an interactive interface that would allow for to zoom in, zoom out and move the graph and differentiate selected features, edges or nodes, using the color and width of the lines. As suggested by Fig. 5, 6, 7 and 8, the entire hierarchy of legal acts is three-dimensional in nature, which is why in particularly complex cases the best solution may be to use interactive Virtual Reality 3D systems (VR headset and gloves). However, one more problem arises namely prior data preparation for visualization. Due to the enormous size and complexity of Legal Acts databases it is impossible to prepare it by hand so usage of automated systems is obligatory but not sufficient. Of course, advanced database systems may easily report all dependencies just by using database structure and properly formed SQL queries. However, this method does not allow for detection of more complex dependencies including semantic similarity, indirect, implicit, hidden or approximate relations. Thus, complete analytic system must employ semantically aware software.

IV. Requirements for computer systems

Prospects of semantical analysis of the Text of The Law were thoroughly discussed in the past [7][13][14] but still a bunch of unsolved problems remains. Usually in hypertext-oriented systems extensive XML tagging is used exploiting a variety of XML schemes like LegalXML, LexML or Metalex [7]. Although ontology based thus semantic aware standards are available [8][9], current systems are still based mainly on formal structure of legal documents, thus not allowing for sophisticated analysis of the semantic meaning of the legal texts. In fact, such systems ought to employ Natural Language Processing (NLP) procedures and deep Machine Learning (ML) algorithms. The output should contain data required for visualization of the network of dependencies in a selected time, eg. all Acts related to given legal category, eg. illegal money transfers, money laundering or tax fraud, including detailed relations between them and other Acts connected directly or

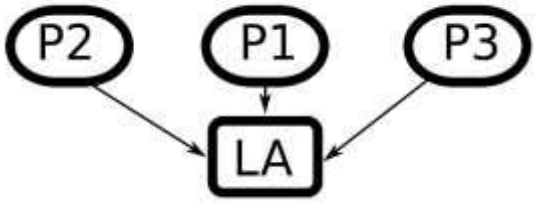


Figure 7. Act LA depends directly and explicitly on Acts P1, P2 and P3. Dependency is primary. Source: own study

indirectly. Such intelligent systems would speed-up RIA analysis of drafts of legal Acts and ensure that they do not affect other Acts in involuntary and destructive way.

Computer systems for legal text processing have to comply a necessity of dedicated data preprocessing application and further data processing aimed at information retrieval. Data preprocessing, as an important step for internal classification, noise removing (from input texts containing, for example, direct or indirect multiplications) and selection of valuable information, has to be focused on input data attributing and identification of homogeneous subsections of texts, thus also parametrization of internal and external relationship structure. Computer systems have to make available functionalities of adaptation of data processing properties to changes in input dataset structure, stored content and also effects of changes in syntactic, semantic and dimensionality. Such parametrization of legal texts should allow to apply advanced data processing algorithms, including the ones inspired by nature (evolutionary systems, artificial immune systems etc. [13]).

Designing and developing of computer systems for legal text processing, focused on information and/or knowledge processing (instead of widely used syntactic processing) should comply a high level of self-adaptation and functionalities of fixing (adjusting) a number of parameters in relation to specificity of legal texts being

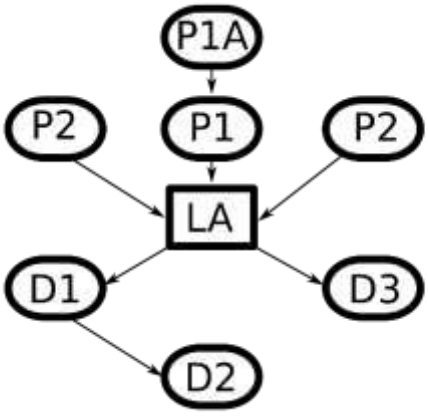


Figure 8. Act LA depends directly on Acts P2, P3 and indirectly on amendation of Act P1 ie. P1a. Furthermore, Acts D1 and D3 refers Act LA directly and explicitly while Act D2 explicitly but indirectly. Source: own study

processed. Considering a large complexity of such a task, which is also time-consuming and dependent on an availability of input data (including historical datasets, important for self-learning systems), a system (complex or domain one) should be designed as dedicated. It is an effect rising from an application of various algorithms and approaches for attributing and identification of consistent subsections in a processed dataset in relation to aim of processing (legislation supporting, identification of inconsistences in a set of legal acts etc.)

V. CONCLUSION

Complex internal and external relationships inside a text and between legal texts can be expressed with a number of approaches. Considering efficient (informative) ones, it may be done with properly designed visualization designs, dedicated similarity measures and also sophisticated visualization of data and relationships. In order to gain this aim, it is necessary to introduce a number of pre-processing levels focused on removing redundant data (i.a. multiplied definitions contained in a set of processed legal acts, references, texts/sentences used for describing a structure of a legal act) and clarification of information included.

In the paper, conditions for processing of legal text, aimed at relationship identification have been outlined, indicating a necessity of applying dedicated (novel) approaches of text processing. Advantages, difficulties and shortcomings of visualizations of legal texts have been presented. Further work will focus on a problem of attributing of legal texts and applying distance-based detection algorithms for identification of similarity between legal texts and also their structure using advanced visualization systems.

ACKNOWLEDGMENT

Authors thanks to prof. W. Cyrul for encouragement and support.

REFERENCES

[1] J. Opiła, T. Pełech-Pilichowski,. (2014) "Problems with the storing and presentation of information" in: W. Cyrul "Information technology and law", Kraków Jagiellonian University Press, ISBN: 978-83-233-3658-7, pp. 27–40,

[2] Act Road Traffic Law (Ustawa Prawo o ruchu drogowym), Dz.U. 1997 nr 98 poz. 602, URL: http://prawo.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU19970980602, (accessed 2020.02.04)

[3] "Act on road safety and order on public roads" (Ustawa o bezpieczeństwie i porządku ruchu na drogach publicznych) Dz.U. 1961 nr 53 poz. 295, Art.3.1, URL:

[4] Code of Ofenses, (Kodeks Wykroczeń), URL: http://prawo.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU19710120114, (accessed 2020.02.04)

[5] "Leading a child's hand on the pavement is an offense?" (Prowadzenie dziecka za rękę po chodniku wykroczeniem?), URL: Infor.pl, https://www.infor.pl/prawo/wykroczenia/charakterystyka-wykroczen/3053040,Prowadzenie-dziecka-za-reke-po-chodniku-wykroczeniem.html, (accessed 2020.02.04)

[6] W. Clark, W.N. Polakov, F.W. Trabold, "The Gantt chart, a working tool of management", New York, The Ronald press company, 1922, availabe at URL: https://archive.org/details/cu31924004570853, accessed 2020-02-14,

[7] W. Cyrul, J. Duda, J. Opiła, T. Pełech-Pilichowski, „Informatization of the text of the law" (pl: „*Informatyzacja tekstu prawa*"), LEX, Wolters Kluwer SA, Warsaw 2014,

[8] Resource Description Framework (RDF), URL: https://www.w3.org/RDF/, (accessed 2020-02-06)

[9] OWL 2 Web Ontology Language, Document Overview (Second Edition), 2012-12-11, URL: https://www.w3.org/TR/owl2-overview/ (accessed 2020-02-06),

[10] *graphviz* – Computer graphing suite of programs, URL:https://graphviz.org, (accessed 2020-01-27),

[11] Inkscape - computer Open Source program URL: https://inkscape.org/ , (accessed 2020-02-06),

[12] Y. Masterman, "Language, Cohesion and Form (Studies in Natural Language Processing)", Cambridge University Press, 2010

[13] T. Pełech-Pilichowski, W. Cyrul, "Computer-Aided Legislation Based on Immune-Like Processing of Legal Texts", In: Logic in the Theory and Practice of Lawmaking, M.Araszkiewicz and K. Płeszka edt., Springer, 2015

[14] Semantic Processing of Legal Texts (Where the Language of Law Meets the Law of Language), Editors: E. Francesconi, S. Montemagni, W. Peters, E. Tiscornia, D. (Eds.), Lecture Notes in Artificial Intelligence, Springer, 2010,

[15] J. Opila, "Role of visualization in a knowledge transfer process", Business Systems Research; ISSN 1847-8344, 2019 vol. 10 no. 1, p.164–179, URL: https://hrcak.srce.hr/file/319710 (accessed 2020-02-14).

[16] M. Damashek, "Gauging similarity with n-grams: Language-independent categorization of text." *Science* 267 , no. 5199 (1995): 843-849

[17] C. Goddard, "Semantic Analysis: A Practical Introduction (Oxford Textbooks in Linguistics)", Oxford University Press, 2011

Privacy Policy Understandability Analysis of Croatian Electronic Publications

Marta Alić
Zagreb University of Applied Sciences, Zagreb, Croatia
marta.alic@tzv.hr

Abstract - **Information transparency in a field of data and privacy protection is not recent topic. For a long time it was approached from the perspective of software engineering and information system design, in which it was usually categorized as a non-functional requirement in relation to the functionality of the software, considered primarily as a quality issue and "second class" concept. But with enforcement of General Data Protection Regulation it has come in focus of research more prominently as a requirement in engineering privacy compliance. And for privacy instrument to be transparent, not only factor of informativeness, giving necessary information, needs to be adressed, but it's quality is also a prerequisite. Understandability is one of quality factors in achieving transparency. In paper, using readablity formulas, understandability of privacy instruments, privacy policies, of Croatia's most visited electronic publications is put to analysis.**

Keywords – transparency, readability, understandability, electonic publications, privacy policy

I. INTRODUCTION

With the growth of communication channels and the increasingly pronounced trend of digital transformation, there is a growing need to provide organizational systems with built-in principles of "Privacy by Design" [1].

In this context, (digital) privacy is today considered as the right and ability of persons to communicate through digital channels, while controlling the dissemination of his or her personal information. To do so essential prerequisite is being informed about who and how participates in the interaction, what information is exchanged and what the information is used for.

Unauthorized collection and use of personal data is considered a risk to the freedom of individuals, who are an active stakeholders in the protection of privacy. Therefore, they have broad rights to information transparency of data collected and processed about them, in order to make informed decisions and to intervene in data processing in accordance with the given rights (withdrawal from data processing, request for deletion of data).

Publication of privacy policies serves as a basic (ex ante) tool of transparency. Although much research has focused on demonstrating usability of these tool [2] [3][4] on various technologies and interfaces [5][6] [7] [8], they have often been shown to be ineffective [9] [10] [11], due to their complexity. and incomprehensibility [12] [13] [14][15]. It is estimated that an Internet user in the U.S. should set aside 244 hours per year, or 10 days, per year to read all of the privacy policies on the pages he or she visits [16].

Requirements of using "clear and plain language" and intelligibility stated in Article 12.1 of [17] are closely linked in meaning that information "should be understood by an average member of the intended audience" [18].

So for a tool to be usable, it not only should present necessary and complete information, it should present them to be understandable.

A. Understandability and readability

Understandability, the quality of comprehensible language or thought [19], is connected to the provision of information and the linguistic quality of an instrument – privacy policy.

It can be measured with readability metrics as a measures of how easy it is for a user to read and understand a specific text.

There are several well-established readabillity formulas, objective assessments of the text weight [20], available for this purpose. As a mathematical equations obtained by regression analysis, formulas represent an assessment of the difficulty or complexity of the text for reading, but also for understanding. The level of readability is expressed either by a number which represents the weight of the text itself, on a descriptive scale from e.g. 'very easy' to 'very difficult', or a number that represents the level of education needed to understand the text, with very high prediction accuracy.

As early as the 1920s, educators discovered a way to use the vocabulary weight and sentence lengths to predict the difficulty of text levels. They fit that method into a formula that shows the readability of the text, and their success has been confirmed by application of different formulas to the present day. Great progress in readability formula development was achieved in the 1950s, when new formulas were developed and the use of more than 40 different readability formulas is mentioned in the literature today [21], although existence even more than 200 published formulas in the 1980s was mentioned [20].

Among the most frequently mentioned readability formulas used can be singled out: Flesch Reading Ease, Flesch – Kincaid Grade Level Index, SMOG, Fog Test, Fry Formula, Dale–Chall Formula etc.

The most commonly validated formulas are those in English language, and there are modifications of formulas for other languages, such as German, French, Dutch, Danish, Chinese, Russian, Swedish, Vietnamese, Korean, Hindi, Hebrew, etc. [22]

In Croatian, work of [23] has been most significant in validating readability formulas and it will be used as a methodology for this paper.

II. METHODOLOGY

A. Sampling

Research focused transparency tools of Croatian most popular electronic publications, „editorially designed program contents published daily or periodically via the Internet by providers of electronic publications for the purpose of public information and education [24]. Research sample was gathered from the list of 50 most visited websites in Croatia on Alexa [25] and SimilarWeb [26] in News and Media category. In April 2020 sampled were privacy policies of 11 publications: Index.hr, 24sata.hr, Vecernji.hr, Jutarnji.hr, Telegram.hr, GlasIstre.hr, Net.hr, Dnevnik.hr, NoviList.hr, Tportal.hr i N1info.hr. On three of listed publications (RTL.hr, Narod.hr, Poslovni.hr) published privacy policies haven't. been found. Publications that are owned by the same publisher, such as HANZA MEDIA d.o.o (SlododnaDalmacija.hr, Gloria.hr), share same privacy policy text, so readability results for Jutarnji.hr can relate to them.

B. Readability Formulas

Readability formulas used for this research are: the Simple Measure Of Gobbledygook (SMOG), the Flesch Reading Ease (FRE), the Flesch-Kincaid grade level (FKGL), the Automated Readability Index (ARI) and the Lasbarhetsindex Swedish Readability Formula (LIX).

While SMOG, FKGL and ARI express the years of schooling required to understand a particular text, FRE and LIX formulas correspond to the descriptive weight scale of the text.

The most commonly used (computer) readability formula is Flesch Reading Ease (FRE), named after the author Rudolf Flesch, an American citizen of Austrian origin, who in the atmosphere of democratization of education in the United States developed a formula that was supposed to help determine the weight of children textbooks. His formula has a simple measure and, other than in education, had a great success in the field of journalism [27].

Flesch published his first readability formula in 1943 and since than there have been revisions. Today this formula has been simplified and reduced on two variables, the average word length and the average sentence length, and reads:

$$FRE = 206{,}835 - 0{,}846\ wl - 1{,}015\ sl,$$

where wl is the word length expressed by the number of syllables (word length), and sl is the length of the sentences expressed by the number of words (sentence length).

Croatian variant made by [28] based on contrastive analysis of English and Croatian corpus of 100,000 words of various types of texts and publications published after 1995, increases the index by 50:

$$FREhr = 206{,}835 - 0{,}846\ wl - 1{,}015\ sl + 50.$$

Readability is expresses on a scale from 0 to 100, where each weight category is explained descriptively; 0 is the index of maximum weight, and 100 is the index of simplicity or lightness of the text: 90-100:Very Easy; 80-89 Easy; 70-79: Fairly Easy; 60-69: Standard; 50-59: Fairly Difficult; 30-49: Difficult; 0-29 : Very Confusing.

The score between 60 and 70 is largely considered acceptable. Scores between 90.0 and 100.0 are considered easily understandable by an average 5th grader, scores between 60.0 and 70.0 are considered easily understood by 8th and 9th graders and scores between 0.0 and 30.0 are considered easily understood by college graduates in USA.

The Flesch-Kincaid grade level (FKGL) is variant of the same method, using the average sentence length and the average word length as an input to calculate number of grades needed to understand the text:

$$\text{Number of Grades} = (0{,}39\ ASL) + (11{,}8\ ASW) - 15{,}59,$$

where ASL is average sentence length, calculated by dividing the total number of words by the total number of sentences, and ASW is the average number of syllables per word, calculated by dividing the total number of syllables by the total number of words in the tested sample.

For texts on Croatian this formula is validated [23] to:

$$\text{Number of Grades} = (0{,}39 \times ASL) + (11{,}8 \times ASW) - 22.$$

The Simple Measure Of Gobbledygook or SMOG formula is published in 1969. By G. Harry McLaughlin [29]. It is a manual method of readability testing, resulting with the index that represents the number of years of schooling required for the reader to read, and then understand, a particular text without difficulty.

The SMOG formula consists of four steps and is calculated as follows: 1) three samples of text are selected, each of which consists of 10 consecutive sentences, from three different parts of the written material (a total of at least 100 words); 2) the total number of words with three or more syllables in those 30 sentences is calculated; 3) the second root of the obtained number is calculated; 4) the number 3 is added to the result. When the tested text contains less than 30 sentences, methodology differs.

Simplified formula for English:

$$SMOG = 3 + \sqrt{3+}\ \text{syllables}$$

is modified to Croatian [23]:

$$SMOGhr = 2 + \sqrt{4+}\ \text{syllables}$$

to reflect the same level of schooling as the English language formula.

The Lasbarhetsindex Swedish Readability Formula (LIX) is a readability measure to calculate the difficulty of reading a foreign text. The formula was developed by Swedish scholar Carl-Hugo Björnsson [30]. The LIX readability formula is as follows:

$$LIX = A/B + (C \times 100)/A,$$

where A represents number of words, B is number of sentences and C is number of long words (more than 6 letters).

Formula result is also expressed descriptively on a scale from the easiest to the most difficult text to understand: <24 very easy; 25-34 easy; 35-44 mediocre; 45-54 difficult; > 55 very difficult.

It can be modified for Croatian text as [23]:

$$LIX = A/B + (C \times 100)/A + 2$$

where A represents number of words, B is number of sentences and C is number of long words (more than 7 letters), and the existing descriptive scale made for the original formula is retained.

The automated readability index (ARI) is a readability test derived from ratios representing word difficulty (number of letters per word) and sentence difficulty (number of words per sentence), formulated by authors, in 1967 [31].

Formula outputs a number that approximates the schooling age needed to understand the text:

$$ARI = 4{,}71 \times (ch / w) + 0{,}5 \times (w / s) - 21{,}43.$$

For texts in the Croatian language, it would read [23]:

$$ARI = 4{,}71 \times (ch / w) + 0{,}5 \times (w / s) - 23{,}23,$$

where ch is total number of characters; w is total number of words and s is total number of sentences.

C. *Tools used*

To calculate the total number of words and the number of characters, the "wordcount" option in Word application from *Microsoft* 365 was used. To calculate the total number of syllables, a program was used to calculate these variables available on the Internet [32], which was selected as the most reliable after comparing the results of native speakers of English and Croatian variables [23]. A text analysis program available on the Internet [33] was used to calculate the number of words of more than seven characters, and the calculation of four-syllable words in Croatian texts was done by hand by a author, professor of the Croatian Language and Literature.

The computer program Excel in *Microsoft* 365 was used to calculate the formulas and average values of the length of words and sentences.

III. RESULTS

Analysis of readability with two formulas that result with indexs of readability on the descriptive scale, FRE and LIX, shows scores of poor readability on both scales (Table 1). On LIX scale all publications scored more than 55 as "very difficult" to comprehend. On FRE scale, inversely proportional than LIX, more than half of publications (55%) scored as „difficult" to read (range from 30 to 49). Other 45% (5 publications) scored as „very confusing" (range from 0 to 29).

On both scales, privacy policy on Vecernji.hr scored as most readable, while Tportal.hr has most unreadable privacy notice.

Average score on FRE scale is 30,87, just above the limit for „difficult" as overall assesment.

Average number of years of schooling (grades) required to understand privacy policies of sampled electronic publication is 16: these documents are easily understood by college graduates (Table 2).

Again, policies on Vecernji.hr are showing most readable results on FKGL scale (13 grades) and ARI scale (14 grades), while on SMOG scale least schooling (15 grades) is needed to comprehend privacy policy on Jutarnji.hr. As this publication shares policy text with SlobodnaDalmacija.hr and Gloria.hr, same amount of schooling is needed to understand privacy information on that publication.

TABLE 1 RESULTS OF FRE AND LIX ANALYSIS

Publication	FRE	LIX
N1info.hr	39,99	63,31
Vecernji.hr	40,11	63,13
Jutarnji.hr	30,38	68,62
Telegram.hr	26,81	66,02
GlasIstre.hr	31,48	70,26
Net.hr	30,22	68,28
Dnevnik.hr	27,42	73,17
24sata.hr	32,27	68,48
Indeks.hr	29,87	66,51
NoviList.hr	28,69	73,12
Tportal.hr	22,38	73,77

TABLE 2 RESULTS OF FKGL, ARI AND SMOG ANALYSIS

Publication	FKGL	ARI	SMOG
N1info.hr	13,85324	14,25285	15,36932
Vecernji.hr	12,91864	13,99437	15,96148
Jutarnji.hr	14,4425	15,30115	14,7047
Telegram.hr	15,39315	16,17538	14,83216
GlasIstre.hr	16,41515	17,35697	15,28821
Net.hr	15,37806	16,76065	15,04159
Dnevnik.hr	17,68367	19,45612	16,0767
24sata.hr	15,51773	17,01181	16,22876
Indeks.hr	15,48476	16,63768	15,28821
NoviList.hr	17,53129	18,85691	15,56981
Tportal.hr	17,34838	18,40428	16,67479

Most years of schooling (18 by FKGL scale and 19 by ARI scale) is needed to comprehend privacy rules on Dnevnik.hr and on Tportal (17 by SMOG).

IV. DISCUSSION

Transparency in data protection is set as a control mechanism [34] and a factor for trust [35] [36], but also a key prerequisite for the participation of data subjects in the data processing [37], ie "intervenability", the possibility of intervention.

Transparency tools or technologies are a combination of technological solutions and legal or procedural frameworks [34], which aim to reduce the existing high information asymmetry between data subjects and data controllers [38], as respondents cannot always determine which data on them controller collects and, consequently, for what purposes they are used, which can have detrimental effects on the privacy of respondents [39]. Reference [40] therefore defines them as "tools that can provide an individual with a clear view of aspects relevant to his (personal) data and privacy".

Although intervenability, as the next step in privacy protection, goes beyond the requirements of transparency, both principles are mutually defined through the international standards ISO/ IEC 29100: 2011 [41] and the legal regulations of the Personal Data Protection Regulation (GDPR) [17]. In order to meet the requirements of the principles of data protection and privacy, transparency should be meaningful and useful, ie meet the requirements in relation to the possibility of intervenability as a goal of privacy. In doing so, mechanisms for achieving intervenability should provide respondents with clear, prominent, easily understandable, accessible tools for consuming the right to choose in relation to the processing of their personal data at the time of data collection, at first processing or within a reasonable time [25].

Research results shown that prerequisite for intervenability, transparency, in context of electronic publications, the main source of information for more than 62 percent of Croatian citizens [42], hasn't been satisfactory in achieving set goals in terms of clearness and understandability.

In the last census in 2011, the distribution by education shows that most of the Croatian population has completed high school (11 or 12 years of schooling). Share of highly educated population (15 to 17 years of schooling), average needed to comprehend researched privacy policies, is only 15.4% of total population.

V. CONLUSION

Transparency can be viewed as a regulatory or voluntary requirement. Regulatory requirements are generally about the compliance between system requirements and regulatory constraints, enforced by law, such as GDPR. Judging by the results of research, this requirement, in case of electronic publications in Croatia, is set as a compliance mechanism primarily, rich in informativeness, but poor in quality. And in this context, it's role in intervenability of data subjects and their privacy protection is questionable.

REFERENCES

[1] A. Cavoukian and others, "Privacy by design: The 7 foundational principles," *Information and Privacy Commissioner of Ontario, Canada*, 2009.

[2] P. Murmann and S. Fischer-Hübner, "Tools for Achieving Usable Ex Post Transparency: A Survey," *IEEE Access*, vol. 5, pp. 22965–22991, 2017.

[3] F. Schaub, R. Balebako, and L. F. Cranor, "Designing Effective Privacy Notices and Controls," *IEEE INTERNET COMPUTING*, 2017.

[4] J. Gluck *et al.*, "How Short Is Too Short ? Implications of Length and Framing on the Effectiveness of Privacy Notices," *Symposium on Usable Privacy and Security (SOUPS)*, no. Soups, 2016.

[5] S. Wachter, "Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR," *Computer Law and Security Review*, 2018.

[6] S. Patil, R. Hoyle, R. Schlegel, A. Kapadia, and A. J. Lee, "Interrupt now or inform later?: Comparing immediate and delayed privacy feedback," *Conference on Human Factors in Computing Systems - Proceedings*, vol. 2015-April, pp. 1415–1418, 2015.

[7] P. G. Kelley, L. Cesca, J. Bresee, and L. F. Cranor, "Standardizing Privacy Notices : An Online Study of the Nutrition Label Approach," *CHI 2010 - The 28th Annual CHI Conference on Human Factors in Computing Systems, Conference Proceedings*, pp. 1573–1582, 2010.

[8] M. Harbach, M. Hettig, S. Weber, and M. Smith, "Using Personal Examples to Improve Risk Communication for Security and Privacy Decisions," *CHI 2014- The 28th Annual CHI Conference on Human Factors in Computing Systems, Conference Proceedings*, 2014.

[9] R. E. F. Ellow, F. O. Clip, P. R. F. Ellow, and F. O. Clip, "Privacy Harms and the Effectiveness of the Notice and Choice Framework," *JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY*, 2014.

[10] A. Rao, F. Schaub, N. Sadeh, A. Acquisti, and R. Kang, "Expecting the unexpected: Understanding mismatched privacy expectations online," *SOUPS 2016 - 12th Symposium on Usable Privacy and Security*, no. Soups, pp. 77–96, 2019.

[11] L. F. Cranor, "Necessary but not sufficient: Standardized mechanisms for privacy notice and choice," *Journal on Telecommunications and High Technology Law*, vol. 10, pp. 273–308, 2011.

[12] P. Raschke, A. Küpper, O. Drozd, and S. Kirrane, "Designing a GDPR-compliant and usable

privacy dashboard," in *IFIP Advances in Information and Communication Technology*, 2018.

[13] N. Asanka, G. Arachchilage, and J. Slay, "Designing Privacy for You : A User Centric Approach For Privacy Designing Privacy for You," in *HCI International*, 2017.

[14] S. Wilson *et al.*, "The Creation and Analysis of a Website Privacy Policy Corpus," in *Proceedings ofthe 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 1330–1340.

[15] J. R. Reidenberg *et al.*, "Disagreeable privacy policies: Mismatches between meaning and users' understanding," *Berkeley Technology Law Journal*, vol. 30, no. 1, pp. 39–88, 2015.

[16] A. M. Mcdonald and L. F. Cranor, "The Cost of Reading Privacy Policies," *A Journal of Law and Policy for the Information Society*, pp. 1–22, 2008.

[17] European Commission, *General Data Protection Regulation*. 2016.

[18] Article 29 Working Party, "Guidelines on transparency under Regulation 2016/679," pp. 1–31, 2018.

[19] "WordNet," *Princeton University*. [Online]. Available: https://wordnet.princeton.edu.

[20] W. Dubay, *The Principles of Readability*, no. April. 2014.

[21] C. C. Doak, L. G. Doak, and J. H. Root, *Teaching Patients with Low Literacy Skills*. 1996.

[22] B. Zakaluk and S. J. Samuels, *Readability: Its Past, Present, and Future*, no. 6. 1988.

[23] S. Brangan, "Razvoj formula čitkosti za zdravstvenu komunikaciju na hrvatskom jeziku," 2011.

[24] *Zakon o elektroničkim medijima/Electronic Media Act. (NN 153/09, 84/11, 94/13, 136/13)*

[25] "Top Sites in Croatia," *Alexa*. [Online]. Available: Top Sites in Croatia.

[26] "Top sites ranking for all categories in Croatia," *SimilarWeb*. [Online]. Available: Top sites ranking for all categories in Croatia.

[27] R. Flesch, "A new readability yardstick," *Journal of Applied Psychology*, vol. 32, no. 3, 1948.

[28] S. Brangan, "KVANTITATIVNA PROCJENA TEŽINE TEKSTA NA HRVATSKOM JEZIKU," vol. 1, pp. 35–58, 2014.

[29] G. H. Laughlin, "SMOG Grading —– a New Readability Formula," *Journal of Reading*, vol. 12, no. 8, 1969.

[30] C.-H. Björnsson, *Lix för tysk skoltext*. Stockholm: Pedagogiskt centrum, 1968.

[31] E. A. Smith and R. J. Senter, "Automated readability index.," *AMRL-TR. Aerospace Medical Research Laboratories (6570th)*, pp. 1–14, 1967.

[32] "WordsCount: Syllable Analysis," *WordsCount*. [Online]. Available: http://www.wordscount.info/wc/jsp/clear/syllable.jsp.

[33] "Text Statistics Analyser," *Using English*. [Online]. Available: https://www.usingenglish.com/resources/text-statistics/.

[34] T. Pulls, *Privacy-Preserving Tools Privacy-Preserving Transparency-Enhancing Tools*. 2012.

[35] S. Crane, H. Lacohée, and S. Zaba, "Trustguide — trust in ICT," vol. 24, no. 4, pp. 69–80, 2006.

[36] C. Andersson *et al.*, "Trust in PRIME," pp. 552–559, 2005.

[37] G. Danezis *et al.*, *Privacy and Data Protection by Design - from policy to engineering*, vol. abs/1501.0, no. December. 2015.

[38] B. W. Schermer, "The limits of privacy in automated profiling and data mining," *Computer Law & Security Review*, vol. 27, no. 1, pp. 45–52, 2011.

[39] R. Calo, "The Boundaries of Privacy Harm," 2010.

[40] M. Hansen, "Marrying transparency tools with user-controlled identity management," in *IFIP International Federation for Information Processing*, 2008.

[41] ISO/IEC, "ISO/IEC 29100:2011 Information technlogy - Security techniques - Privacy Framework; Technical report."

[42] Ipsos, "Medijske navike u Republici Hrvatskoj," 2019.

EE

International Conference on

ENGINEERING EDUCATION

Steering Committee

Chairs: Predrag Pale, University of Zagreb, Croatia
Branimir Pejčinović, Portland State University, United States
Juraj Petrović, University of Zagreb, Croatia

Members: Leslie Martinich, Competitive Focus, Austin, TX, United States
Lueny Morell, InnovaHiEd Institute, Mayaguez, Puerto Rico
Predrag Pale, University of Zagreb, Croatia

Program Committee

Members: Tomislav Jagušt, University of Zagreb, Croatia
Lueny Morell, InnovaHiEd Institute, Mayaguez, Puerto Rico
Branimir Pejčinović, Portland State University, United States
Juraj Petrović, University of Zagreb, Croatia
Ana Sović Kržić, University of Zagreb, Croatia

PAPERS

Closing the Gender Gap in Engineering: Students Role Model Program

E. Vidal*, E. Castro*, S. Montoya* and K. Payihuanca*
*Universidad Nacional de San Agustín de Arequipa/Departamento Académico de Ingeniería de Sistemas Informática, Arequipa, Perú
evidald@unsa.edu.pe, ecastro@unsa.edu.pe, smontoya@unsa.edu.pe, kpayihuanca@unsa.edu.pe

***Abstract*— A report of UNESCO highlights that in higher education, only 35% of students who enrolled in majors related to STEM are women. Stereotypes, cultural barriers, lack of confidence, lack of knowledge of career opportunities, and lack of vision of possibilities are more than enough reasons for students who do not pursue university studies in areas of Engineering or drop out University. The Professional School of Software Engineering at Universidad Nacional de San Agustín de Arequipa was not the exception to this reality. To make an impact and change our reality, the authors started a "Students Role Model Program". The objective was to create a community of female students that became role models who share their motivation to STEM with new female students and to take active participation in the recruitment process. The main activities were divided into two branches: a) Exposure to role-models at the student level and b) Participation in the international community Django Girls. We believe that the main contribution of this work is to show the structure of our program which can be replicated in different engineering majors that seek to improve the retention of their female students.**

Keywords—diversity, gender gap, inclusion, mentorship

I. INTRODUCTION

A recent report of UNESCO highlights gender inequalities in the teaching of science, technology, engineering and mathematics (STEM). In higher education, only 35% of students who enrolled in majors related to STEM are women. Additionally, women leave STEM in disproportionate numbers during their studies. The report also presents some kinds of interventions that help increase women's engagement with STEM education: supportive learning environments, mentoring and role models [1].

Studies have shown that mentors can help women to motivate them and improve their self-confidence to achieve their goals. Mentors have an impact on students' persistence [1]. Education and gender equity are an integral part of the 2030 Agenda for Sustainable Development adopted by the United Nations (UN) in 2015 [2].

Stereotypes, cultural barriers, lack of confidence, lack of knowledge of career opportunities, and lack of vision of possibilities are more than enough reasons for students who do not pursue university studies in areas of Engineering or drop out University. The Professional School of Software Engineering was not the exception to this reality. In a previous study published in 2016 [3] we showed that in the application processes, in a range of six years (2010 to 2015), the number of women was only 20% compared to men (Fig. 1). To make an impact and change our reality, the authors started a mentoring program called "Student Role Model Program" (SRMP). This paper describes the background of the program, structure, main activities, and the initial results. We have improved the number of women enrolled to the **School of Software Engineering**. We believe that the main contribution is that our experience could be replicated and adapted to the context of other institutions.

The rest of the paper is organized as follows: Section II presents some related work. Section III presents the initial context, Section IV describes the details of the Program and the initial results. Finally, we present our conclusions.

I. BACKGROUND

The impact of mentoring in academic performance has long been studied before. An important reference is Redmon [4] who highlights the impact on student identity and social integration into academic life. Also, the study presents the key elements to design an effective mentoring program and the essential elements to administrate a mentoring program. The study of Wilson et al [5] describes a strategy for improving diversity and retention in undergraduate STEM disciplines. The study ensures that each undergraduate participant has mentoring relationships with faculty, staff, and their peers. The model is based on the integration of mentoring and academic interventions into a structured research program. In a more specific work Parker, Hall & Kram [6] affirm that peer mentors can draw on more recent and relatable experiences, and students are most of the time more comfortable approaching peers.

Our work is based on the structure proposed by Redmon for designing and administrate a mentoring program, but it is focused on a female peer-mentor where most of the activities are developed by the students. According to Redmon [4], there are three key elements to design a Mentoring Program: institutional commitment, goal formulation and organizational structure.

Institutional Commitment is related to financial, administrative and resource support. Goal formulation defines the expectations of the program and the regular analysis that must be done to measure the impact of the program. The organizational structure helps to organize the administrative and operational activities and assign specific tasks.

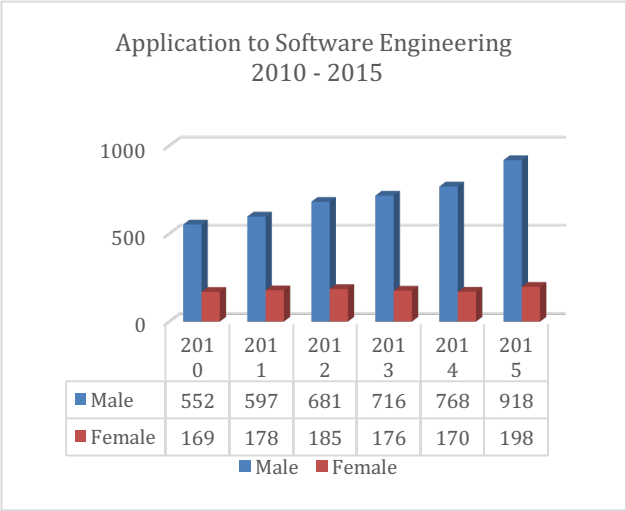


Fig. 1 Applications to Software Engineering had a low number of women from 2010 to 2015 compared to men [11].

Once the mentoring program is designed, it is important to consider how to administer it. Redmon [4] affirms that there are three main activities: a) *recruitment*: identifying the population to be served based on the needs of the institution, then the presentation of the program and participant profiles b) *education*: training the mentors according to the goals of the program and give the mentors enough basic information that increases the effectiveness of the mentoring process and c) *recordkeeping and program evaluation*: it aids with the collection of data for analysis of the progress of the program

II Student Role Model Program

A. Design

The SRMP program was founded in January 2016. The main goal was to improve women's application to Software Engineering and to create a community of female students that allows us to integrate students from the first to the fifth year and create a supportive, friendly and collaborative link between them through the development of specific activities. There is important institutional support when it refers to outreach majors and there is specific funding for student academic and extracurricular activities.

Since the program is focused only on the Software Engineering School, the organization is very simple. There is a principal coordinator who is a faculty, in charge of doing the coordination with the institution when funding or resources are required. We also have two female faculty who are in charge of teaching students who will have the role of peer-mentors and to give them support to deal with not regular situations, and finally, we have the student-mentors. The SRMP initiated with 10 students. The main activities were divided into two branches: a) exposure to role-models at the student level and b) participation in an international community: Django Girls. Django Girls [10] is an international non-profit organization that empowers women related to technology. It was founded in Europe and has currently spread to more than 518 cities in more than 94 countries.

B. Recruitment

We organized a launch event to present the program. All female students and professors were invited. The launch event was motivational and to explain the objectives of SRMP.To maintain the dynamism and constant communication, a Facebook group was created. In this group, news of interest about the program are published, such as scholarships, news related to women who inspire, events, and congresses. The achievements of our students are highlighted as well.

C. Exposure to peer Role-Model

Having role-models at the peer level allows students to have a closer view on their achievements. One way to boost students' confidence is by giving them responsibilities to support their classmates from the first semesters. In 2016, teaching assistant positions for second-year students were created for programming courses: ten men and ten women to support first-year students. Mixed teams were created to help each of the laboratory groups. The objective was not only to motivate the outstanding students (the mentors) of the second year but to provide a close role model for the first-year students.

Another activity carried out since 2016 is the participation of female students in vocational fairs where high school students are invited to visit the University to know about the different majors. Our students are in charge of explain the curricula, they explain the kind of projects they develop and they also shows different kind of equipment they use in their learning process like Virtual Reality hardware or Robots (Fig 2). We have observed since female students are involve in this activity more female high school students show interest in talk with the students and ask questions.(Fig 3)

D. Participation in International Communities.

Judge and Hurst [8] present the concept of capitalization that means a form of proactive attitude that involves participating in volunteer opportunities that provide professional growth and development. This can include attending relevant presentations or joining organizations. According to Holland et. al [9] students who engage in these kinds of activities are expected to be more willing to act as mentors to their peers because of their increased levels of involvement and experience.

In a joined effort, female students from Software Engineering at Universidad Nacional de San Agustin and Systems Engineering at Universidad Católica de Santa María (UCSM) founded Django Girls Arequipa in 2017. From 2017 to date, Django Girls has organized 03 workshops and almost 02 talks a year. There are two main activities developed by Django Girls: one-day workshops and talks. Each workshop has between 60 and 80 women who develop their first web application using HTML, CSS, Python, and Django.

We have found that being part of an international organization increases our students' confidence. That confidence is demonstrated in the answer feedback of the survey given at the end of every activity. A total of 420 women participated in the activities carried out, between

workshops and talks. In 2018 Django Girls Arequipa was promoted to Django Girls Perú.

Talks are short and last one or two hours. Professional women from STEM are invited to share their experiences. Topics are motivationally oriented. Some other topics are related to career development.

Django Girls Arequipa has been part of important events to inspire and attract more women to the Software Engineering field. One of them was Tech Lady Hackathon [11] organized by the Embassy of the United States, 69 girls attended. Django Girls Arequipa gave three workshops in this event.



Fig. 2 Tech Lady Hackathon Arequipa - an event organized by the Embassy of the United States and hosted by Universidad Nacional de San Agustín. The Django Girls Arequipa team gave three workshops.

V INITIAL RESULTS

Fig. 3 shows the results of the application process from 2016 to 2019. We can see an increase in the interest of women in the Software Engineering field.

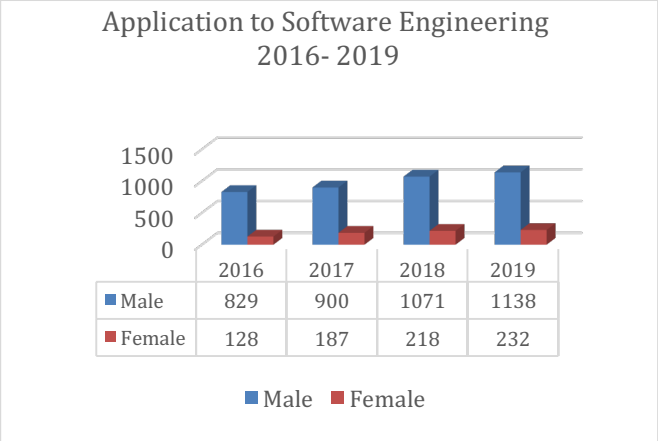


Fig. 5 Application to Software Engineering has increased from 2016 to 2019[1].

Compared to Fig 1. we can observe that in 2016 the number of women who applied to Software Engineering was lower than in 2015. We attribute this result to the beginning of the program and the impact should be noticed in the next application process in 2017. We also observe that the number of women interested in the field has increased year after year. Even though it is true that there is an improvement, the gap between men and women is still huge. This lets us consider new actions to take to get more women involved.

We consider important to highlight the point of view of the students involved in different activities; the feedback was obtained from the survey of the experience of every student.

Suling Vera (CS1 Assistant): “Sharing your knowledge and experiences with new classmates in the course helps you to understand that sharing what you have learned goes one step beyond just learning. Giving support to your peers helps them to have more confidence”.

Montoya Muñoz Solansh Jaqueline (Senior year student – founder of Django Girls Arequipa): "Being part of Django Girls means the opportunity that I have to support the development of many women in STEM. I consider very important that women know a role model that leads them to study or continue studying a STEM career. In Django Girls, I have met many girls like me who like technology, and I have realized how important it is our participation in this area. I think that all women who interact in some way with technology should be in some organizations like Django Girls to reduce the gender gap."

Payihuanca Mamani Keisi Yamile (Senior year student – founder of Django Girls Arequipa): "Being part of an organization like Django Girls contributed to my development not only professional but also personal. With Django Girls, I learned to perform and organize events, the strict management of tasks with dates and responsibilities. I met many inspiring women who shared a goal in common, reducing the gender gap in our industry and increasing the rate of participation of women in technology. I would definitely recommend students and women in general to be part of communities, not only international but local ones too, or initiatives in their own university because I think it is important to look for experiences that help us understand the broadness of the world we live in and take the path to achieve real changes. "

Maria Quijlla (First-year – CS1): “Many times the fear of not being able to express doubt about programming to a classmate or teacher because of the fear of “what people will say” or "what a shame to ask him this", made me keep the doubt. It caused me not to be able to move forward. Having female assistants who knew about programming made me feel confident enough to ask about even the smallest details. I can imagine that they might also had the same doubts as me when they were first-year students”.

Dania Cahuana (First-year – CS1): "In my opinion, when I took CS1, having a female peer assistant helped me a lot because I was more confident when I had to ask questions when my program went wrong. A peer assistant has a closer view of your doubts since she took the CS1 course the year before.

V CONCLUSIONS

After four years of working with the SRMP, we reinforce the importance of having the support of the institution when it refers to resources and funding. It helps to coordinate and host external activities like those developed with Django Girls. Regarding the participation of students in international communities, we can say that the university is the ideal scenario to boost peer role model initiatives. Students had to create a specific team and to involve faculty and professionals that can guide and advise their projects. Additionally, it is important to establish links with other universities to expand collaborative work and resources.

Keeping the community active is a key factor. Consultations must be resolved promptly, maintaining a good interaction with the team. Communities grow according to networking, therefore, interaction with members, assistants, and stakeholders is important.

It is important to highlight that in many of the activities, we have focused on the active participation of first and second-year students. This is because the highest chances of dropping out occur in the first years of the career. The activities aimed at this critical group were the participation as a Laboratory Assistant to support the CS1 and CS2 courses, and the invitation to be part of Django Girls Arequipa. Senior students were more involved in the career fair.

We have also seen that showing students through talks or inviting industry professionals, the possibilities of professional development make them feel more involved in their career and have a clear picture of the different options that exist for them, not only at the local or national level but in an international level as well.

We believe that work at Django Girls is making a difference for EPIS (Escuela Profesional de Ingeniería de Sistemas) students. The students feel welcomed and accompanied not only academically by their classmates, but they have also found a space to develop as a community that gives them opportunities for improvement through the various activities offered by Django Girls.

Students involved in SRMP have made many women interested in technology and trust their potential. In addition to the friendship, ties created, not only in the group in charge but also among the attendees, which enlarges their work contacts and establishes ties between other women and/or men for a common purpose.

REFERENCES

[1] UNESCO. Cracking the code: girls' and women's education in science, technology, engineering and mathematics (STEM), 2017.

[2] United Nation. Transforming our World: the 2030 Agenda for Sustainable Development. A/RES/70/1. New York, United Nations, 2016

[3] E. Vidal and E. Castro. "Women in software development": Towards the retention of female students in Systems Engineering: Experience and lessons learned. In Proceedings of the LACCEI International Multi-conference for Engineering, Education and Technology, 2016.

[4] S. P. Redmond Mentoring and cultural diversity in the academic setting. Am Behav Sci 34:188–200, 1990

[5] Z. S. Wilson, L. Holmes, K. Degravelles, M. R. Sylvain, L. Batiste, M. Johnson, M., and I. M. Warner, I. M. Hierarchical mentoring: A transformative strategy for improving diversity and retention in undergraduate STEM disciplines. *Journal of Science Education and Technology*, *21*(1), 148-156, 2012.

[6] P. Parker, D. Hall, K. Kram. Peer coaching: A relational process for accelerating career learning. *Academy of Management Learning & Education,7,*487-503. doi:10.5465/AMLE.2008.35882189, 2008

[7] Universidad Nacional de San Agustin de Arequipa. http://www.unsa.edu.pe. Last Acces January 2020

[8] T. Judge and & G. Hurst. Gapitalizing on one's advantages: Role of core self-evaluations. *Journal of Applied Psychology, 92,* 1212-1227. doi:10.1037/0021-9010.92.5.12, 2007

[9] J. M. Holland, D. A. Major, V. J. Morganson, and K. A. Orvis. Capitalizing on opportunity outside the classroom: Exploring the supports and barriers to the professional development activities of computer science and engineering majors. *Journal of Women and Minorities in Science and Engineering, 17,* 173-192. doi:10.1615/JWom- enMinorScienEng.2011002913, 2011

[10] DjangoGirl. https://djangogirls.org/. Last access January 2020

[11] Techlady Hackaton https://innovation-weekend-peru.com/techlady-hackathon-arequipa/. Last access January 2020

[11] Universidad Nacional de San Agustín de Arequipa. Oficina de Admisión. http://admision.unsa.edu.pe/orientacion-estadistica/

[12] Universidad Nacional de San Agustin de Arequipa Oficina Imagen Institucional. http://www.unsa.edu.pe/mas-de-10-mil-escolares-en-vii-feria-de-orientacion-profesional-de-la-unsa/. Last Access March 2020

Integrating Industry Seminars within a Software Engineering Module to Enhance Student Motivation

G.J. Collins
University College London/School of Management, London, United Kingdom
g.collins@ucl.ac.uk

Abstract - **Engineering students increasingly demand and require coverage of emerging technologies to prepare themselves for subsequent research and employment. Industry and professional bodies are also concerned that engineering education doesn't always prepare students adequately for the world of work.**

The software engineering postgraduate professional practice module at University College London is designed to provide real-world experience, before students commence their industry research projects. Industry speakers are invited from a range of organizations, including ThoughtWorks, IBM, Form3, Verne Global, and Fujitsu.

Seminars include: DevOps, microservices, cloud-native architectures, machine learning and quantum technologies. Before each topic is covered, students are asked their understanding of the subject matter, via questionnaires. This information is shared with industry speakers to ensure the content of presentations is compatible with students' prior knowledge. It has proved valuable to allow time for discussions to facilitate professional networking, which particularly benefits female students.

Students have indicated they highly value the real-world project examples delivered by industry experts. This suggests that integrating industry seminars can enhance engineering education and motivate students by covering leading-edge technologies and practices. However, this requires considerable time in coordinating and co-developing seminars, and such initiatives need to be adequately resourced to be effective.

Keywords - engineering education; professional networking; industry seminars; motivation

I. INTRODUCTION

There is shortfall of engineers in the UK; this problem of recruiting qualified engineers and researchers extends across Europe [1, 2]. One way to improve this situation is to ensure university teaching better equips students for the demands of industry. Another is to deliver topics that students have indicated they want to study; to help improve retention and increase applications from those less represented in society.

The Royal Academy of Engineering (RAEng) recommends to reduce this shortfall of engineers there should be increased industry-academic engagement [3]. They outline that industry involvement in teaching provides real-world examples and can increase motivation in the subject, by highlighting relevance and practical examples. The RAEng also recommend that project management skills are taught as an integral part of engineering courses, delivered by business school staff within universities. This is the case at University College London (UCL), where the professional practice lectures are taught by a member of staff based in the UCL School of Management.

Fung [4] argues that there should be no conflict between the needs of industry and that of university research. Both require inquiry and critical thinking. UCL's Connected Curriculum [5] instills that students learn more effectively through research and inquiry. One approach to inculcate this philosophy is not only integrating research within the curriculum but to integrate research and development project examples via guest speakers from industry.

II. MODULE DESIGN

The software engineering postgraduate professional practice module is delivered within the Department of Computer Science, which is a division of the Faculty of Engineering Sciences. The module is designed as a series of 10, 2-hour, afternoon industry seminars, with a project management lecture and tutorial of 2 hours the following day. Students study this module during the spring term, immediately before commencing their research projects.

This module provides an opportunity to cover projects that all students have expressed an interest in. Male students have mentioned delivery pipelines and technical approaches. Female students have outlined interests particularly in health and sustainability [6]. Increasingly, students have mentioned over the years an interest in research that has a positive impact on the environment. These topics provide an ideal opportunity to show the relationship with the UN's sustainability goals.

The module design places emphasis on considering a wide range of users and their views to improve solutions. Discussing previous students' research projects is a way to show the different approaches students have taken and reinforces the need to consider the different perspectives of stakeholders. An initial lecture on inclusion and diversity also highlights the importance of neurodiversity: different ways of thinking and communicating about a problem. The module aims to foster inclusion and sustainability throughout. It is aligned to the Engineering and Physical Sciences Research Council's (EPSRC's) Framework for Responsible Innovation [7] which is also embedded in the Horizon 2020 initiatives for inclusive and sustainable research and innovation [8]. The teaching includes the EPRSc's Anticipate, Reflect, Engage and Act (AREA) framework, to ensure students continually evaluate the ethical and societal impact of their research projects. The UCL Department of Computer Science has received the highest category within the UK university sector Athena SWAN (Scientific Women's Academic Network) award; recognition for the promotion of diversity and furthering research opportunities and careers for women engineers [9].

By integrating the project management lectures with the industry seminars, the aim is to create a cohesive module. The project management lectures include case studies and outline previous student group research projects. Students are keen to see what previous cohorts of students have achieved in their research projects, and the environmental and societal impact these have made. The initial examples cover student research projects using machine learning to reduce air pollution exposure for pedestrians and cyclists [10]. They demonstrate how students have elaborated goals, planned their research, and provide a way to highlight further research opportunities. The aim is also to widen the knowledge of the students to help create what some have called "T-Shaped engineers" [11], where they not only understand the technical aspects, but are better able to manage projects, and appreciate the wider implications of their research for society.

III. Ensuring Student Motivation

By establishing students' understanding of topics via questionnaires, before coverage within the module, seminars can be better tailored to the students' needs.

As an example, Fig. 1., outlines a question before the architecture topic is delivered. This is not to catch students out but to consider what in their opinion is the most appropriate answer. The answers to these questions formed a basis for discussions, for the project management and technical challenges. These discussions were continued within their electronic module forum using Moodle.

For the question "How likely are you to document architecture decisions" Fig. 1. the choices were:

1. Although there are some industry approaches regarding architecture decisions I would not necessarily look at these, but would at least document the problem, a solution. and an alternative.
2. With further help or working as a team I feel confident I could adapt an appropriate method, document a solution, outline an alternative and describe why our chosen solution was adopted.
3. I am aware of Decision Centric Architecture Reviews (DCAR) or an alternative method and may reference this and adapt this method so that the time involved in using this is not an overhead to the project. As well as the solution and an alternative I may include forces in favor of our solution as well as those against.

Students are likely to give answers aligned to their own way of thinking and their own viewpoint. Students are also less likely to feel defensive over their answers, than if they were asked after a teaching session [12]. Explaining the purpose to students of these questionnaires also helps to build a shared endeavor of learning, in that they can appreciate that the teaching material is being tailored to their understanding and their future research and employment needs.

The answers to these questions and follow-on discussions helped determine what background material the industry speaker would need to cover. They also help tailor the notes for the project management lectures. Tailoring the content to students' answers ensures that leading edge examples are incorporated and that well understood concepts are not repeated.

By discussing these answers and the proposed project management content with the industry speaker, this ensures a more consistent structure for the module.



Figure 1. Answers provided by the students attending the professional practice module academic year 2017-18. Example shows the results to one section of the questionnaire, for the topic architecture, covering architecture decisions. Participants = 41. Class size 44, including 5 students from the UCL School of Management. 15 students opted for choice 1, 23 opted for choice 2, and 3 for choice 3.

By sharing this information with the industry speakers, they also have a better understanding of the level they need to aim at, and are more likely to feel they are working as a team with the academic staff. This builds a better basis for creating a more coherent vision for the module and building relationships that may lead on to joint university and industry research collaborations.

The industry seminars cover current research and development projects and are designed to help students when they start their research projects. The topics are designed to cover project management as well as technical issues.

The industry seminars academic year 2018/19 were from speakers from a range of organizations:

1. HSBC, Advances in (High Performance Computing) HPC for exposure analytics
2. Verne Global, HPC for AI applications and sustainable data center design
3. NetApp, Building the Hybrid Cloud, DevOps: Agility without sacrificing enterprise stability
4. Fujitsu, Quantum Computing
5. Form3, Evolving Systems Design: from unreliable (Remote Procedure Call) RPC to resilience using Linkerd. Coverage of containers and schedulers and microservice architecture
6. Instana, Producing Actionable Information from Monitoring Data. Automatic Agents, covering detection algorithms, neural networks, and dynamic graphs
7. RedHat, Containers, Kubernetes, OpenShift and OpenStack
8. IBM, Information Architecture, Delivering Innovation
9. Digital Asset, Making Contracts Smart. Integration of business logic within the distributed ledger
10. ThoughtWorks, Microservices, NGINX, Microservices, demonstration of a denial-of-service attack (DoS attack)

There are also one-hour tutorials to supplement the industry seminars and project management lectures to encourage students to discuss in more detail the issues raised in both sessions. Project management sessions cover project examples and case studies, using a "flipped lecture" [13] approach, to create an active learning opportunity, by devoting class time to analyzing and evaluating the information. The scenario and background information are provided to students online before the class, leaving class time to discuss the material in detail.

Project management lectures 2018/2019 delivered by UCL School of Management lecturer:

1. Introduction, including student air pollution reduction projects using machine learning
2. Case study of a software engineering project
3. Agile project management
4. Retrospectives and reviews, using project examples
5. Boundary communications case study
6. Software architecture decisions, multi-objective optimization
7. Risk and opportunity, use of machine learning in Artificial Intelligent Operations (AIOps)
8. Scheduling practical exercise
9. Teamwork, inclusion, and diversity
10. Sustainable IT. Process improvement at scale and within agile development projects.

Industry seminars are interlinked with the project management lectures. For example, one seminar illustrates how machine learning can be used to identify problems earlier within software development cycle. These concepts are reinforced within the project management stream showing how machine learning can be used to reduce alerts, which can cause fatigue for engineers responsible for software delivery pipeline, and how this time can instead be used for more creative work. Students' research projects, which use machine learning to establish lower pollution routes for pedestrians are outlined as well. Related examples are also covered showing how AI can be applied to provide optimal routes for distribution during disaster relief. Other examples highlight the benefits within medicine; how machine learning can provide a higher level of accuracy in diagnosis within cancer screening programs.

At every stage in planning the module there is a reflection on how the content will engage and motivate the class. It is argued that the way the instructor frames the work is a major factor for student motivation [14]. To inspire students, introductory topics are based on previous students' research projects and how these have made a positive impact on society. To motivate the class further examples of how students during the previous academic year have carried their research forward within industry and research programs at UCL are also highlighted.

IV. CHALLENGES

Sourcing inspiring speakers takes considerable time. Students have outlined they particularly appreciate speakers that demonstrate interactive examples. The industry speakers are selected for their enthusiasm for their subject and their ability to answer technical questions clearly. Invariably, this means that the speakers are passionate about their work and committed to helping their teams pioneer advances in research. To improve the module students are encouraged during the final week of the module, to complete an anonymous online feedback form, within their own time.

"The lecturer managed to secure talks from incredible industry speakers who shared up-to-date, relevant

knowledge on current topics" Professional practice student 2017-18

The students find seminars with live coding particularly engaging; the NGINX seminar included an example of the prevention of a denial of service attack (DoS attack):

"I especially appreciated the talks from NGINX and Chef, which were two companies with whose products I'm intimately familiar with." Professional practice student 2017-18

The industry speaker from NGINX specifically created a website, so that the provision of this software could be demonstrated. Here after outlining security concepts a volunteer from the audience was asked to initiate a DoS attack. As students have as their main interest software engineering, it was not surprising that talks that had live coding examples were the most popular (c.f. modern interactive museum exhibits).

They also appreciated the relevance to their forthcoming research projects and software development careers.

"The speakers where very interesting. Many of the topics touched on in the project management part were important for our future careers..."

"[The lecturer] finds ways to motivate student participation and he finds amazing guest speakers from industry."

Professional Practice students 2018/19

However, one student mentioned within the anonymous feedback 2017/18 and again 2018/19 that having a written assessed component for this module would be a better approach, rather than rely fully on the examination.

"I think having the lecturers from industry was both very interesting and really useful. Some of those lectures have been my highlights at UCL. Maybe an assignment that allows students to select 3 of the visiting talks then to write a short report…That way you would need a 100% exam."

"I believe that an exam is not a good way to asses this module…I think coursework would be more appropriate as it would take a longer time to put thought into the valuable topics and find references and supporting claims."

Professional Practice students 2017-2018 and 2018/19 respectively.

This feedback is being taken in to consideration in redesigning the assessment of the module for future years, to include assignments as well as a final examination. The assessment of the students' research projects is based on their research approach, reports, and presentations within their follow-on module.

V. PLANNING SEMINARS

Researchers have recognized that collaborating to build effective relationships with industry speakers is time consuming [15]. When inviting an industry speaker, a follow-up meeting is necessary, either at the university or more usually at their software development site, to discuss coding practices and the proposed seminar. An outline of the module is provided to the industry speaker, with a summary of main objectives and that students are keen to understand real-world scenarios and the dilemmas teams may face. Industry speakers need to be aware of the students' skill level and languages they use for coding. There also needs to be a discussion how their topic interrelates to the other modules and research at UCL. Reviews of the slides and content occur well in advance of the industry seminar. These regular communications, Fig. 2., also help avert any problems, as industry speakers are more likely to feel part of the team, delivering the common objectives of the research module.

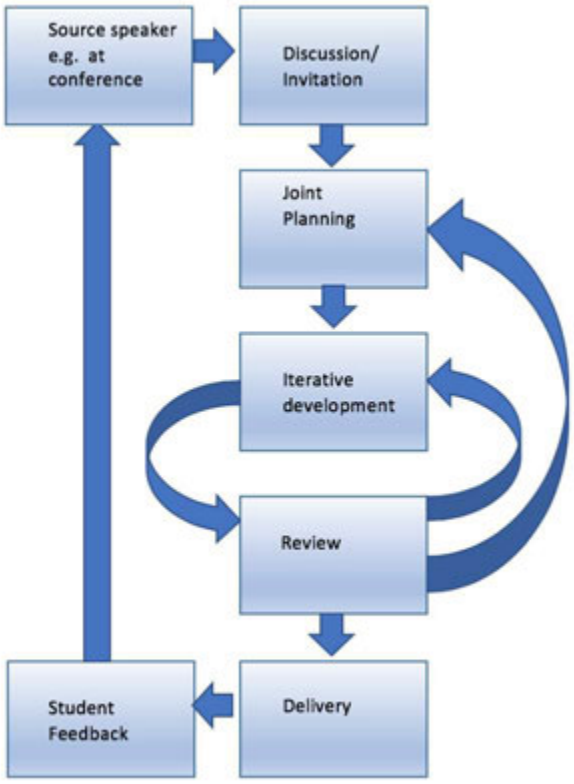


Figure 2. The stages of planning and iterative development of industry seminars for the UCL professional practice module. The review process involving the industry speaker and academic feeds into the development of the seminar and future planning. Student feedback determines future topics.

VI. DISCUSSION

Business-university research collaborations provide significant benefits to their participants. Yet from the Dowling report [16] it seems that industry or universities are not reaping the full benefits of these collaborations. Key to these collaborations are the people involved and building long lasting and trusted relationships. These require a mutual understanding of the goals of both sides and a willingness to collaborate to achieve a successful common vision. To facilitate this enough time needs to be allocated for those involved within their work schedules.

The advantages of integrating industry speakers are that they provide a real-world perspective and frequently opportunities for research. Industry speakers often invite students and staff to forthcoming conferences on a complementary basis. The research students regularly accept and value these invitations, which due to budget limitations they may not otherwise be able to attend. Speakers also outline any further opportunities to contact them and their organization, or visits to see their development and research environment. Having this period of discussion, which breaks down the initial barriers of communication, may be one approach that can help students develop their professional networks.

Davey [17] suggests that there are two important aspects to professional networks the instrumental, the assistance and advice, and the psychosocial including social aspects. Instrumental support can also be considered as tangible support [18], actions that are useful for career development. Instrumental networking, the formation of ties with others explicitly for professional goals, are important for cooperation at work as well as career development. Students often have well developed social networks but limited professional networks. By providing adequate time for discussions with industry speakers this provides an opportunity for students to develop their professional networks. This is particularly important for female students who are less likely to have the support of professional networks [19]. However, previous research by the author has outlined that female students are keen to engage with industry if given the opportunity [6]. Casciaro et al. have shown [19] that females often perceive that instrumental networking can be a morally compromising, by creating an unfair career advantage for themselves. This research has shown that facilitating such networking opportunities, these barriers which often limit female students in their careers, may be broken down. By providing the opportunity at the end of the speaker seminars and speakers initiating the networking and the further work visits, this may be one way that academic modules can support all students, and address the imbalance reported in female professional networking.

Having different perspectives outlined also helps students appreciate the need to engage and listen to their colleagues and the different views within their teams. Within the project management lectures the importance of this process is emphasized. If, for example, during a technical discussion the team eventually reject an idea, then the team member who proposed this idea at least feels as though their views are listened to, and if discussed in detail will also understand the reasons why. This is reinforced during the lecture covering architecture decisions. Here the method of proposing an architectural solution, the pros and cons are recorded, and any methods accepted or rejected are also explained.

How industry speakers are integrated effectively in delivering engineering modules is an area of research that needs to be addressed. May-Newman and Cornwall [21] outline problems where an industry speaker does not have the pedagogic experience, or where the academic does not have the practical experience. Conferences are an essential way to keep up to date but are also a rich source of speakers who can communicate effectively. Collaborating and co-developing seminars resolves the latter issue; it also ensures coverage of the syllabus. By integrating industry speakers within a module this addresses the problem of the academic not having practical experience; and a range of speakers addresses that a real-world expert delivers every topic. After all, it cannot be expected that a lecturer is an expert in every area.

Research has shown that inviting and collaborating with industry speakers is time consuming. Fung [4] has pointed out this should be adequately resourced. This is emphasized by the RAEng that this should be more a joint venture. However, experience reports [15] indicate that this activity is often left to one lecturer in a module. Having a single point of contact is a risk; a lecturer may be absent or may be assigned to another role. Whether these industry speaker invitations should be a centrally organized or not is interpreted differently within departments in the UCL Engineering Sciences faculty. It however needs to be recognized that often a speaker will provide a seminar as they have a professional relationship to the academic that has first invited them. The problem with communicating solely through a central system is this professional relationship may be lost if the original party is not involved.

VII. CONCLUSIONS

Strategic links between universities and industry provide myriad benefits. For universities, the staff and students, opportunities to address real-world applications and challenging research problems. This provides an opportunity for leading edge research opportunities for students. Industry-into-academia initiatives can further enhance this, via visiting professorships and visiting speakers. The faculty of Engineering Sciences at UCL benefit from both aspects. This can also provide a conduit

for ad hoc mentoring and career advice. This is further supported by providing additional time after industry seminars for further questions and discussion. This activity seems to particularly support those less represented in the engineering profession to engage in networking.

Sharing the information prior to talks and feedback from talks is essential in developing talks at the appropriate level. Also, it a way to help speakers feel they are part of a shared process in developing the skills and knowledge of future engineers. The closer the collaboration, the more likely there will be successful research academic-industry opportunities identified in the future, for both research staff and students. Another positive outcome is the practical experience provided to staff and students for the visits as well as the career networking opportunities. In addition to this are the events and conferences that staff and students are invited to, often on a complementary basis. Sharing the knowledge helps also to deliver the ethos of the connected curriculum; integrating teaching and research. It also provides a valuable example to students of academic-industry collaboration, something we are trying to encourage within their research projects.

The industry speakers within the module provide a wide range of perspectives. This reinforces the idea that within teams, workshops and in projects we need to consider different views and perspectives. It also helps students understand that diversity in teams is an asset if we wish to create solutions for the whole of society. If we wish that students develop engineering solutions for the benefit of society then we need support from industry to help develop students' research experiences.

Integrating industry speakers within engineering research modules has been shown to help prepare students for their forthcoming research projects as well as world of work. The students' feedback from this module is that industry speakers are highly valued and provide an important adjunct to the project examples and technology coverage. However, adequate time needs to be allocated for this to be effective. Senior managers need to be aware of the considerable development time required. There are the overheads of sourcing speakers, understanding in detail what students require, planning and co-developing the seminars, and reviewing the material. The rewards for students of well-designed modules are immense; the increase in motivation and preparation for their research projects as well as increased networking opportunities. With humanity facing ever more complex problems we need to ensure that students have the best preparation for research. If students are introduced to real-world problems and dilemmas within their research studies this will provide a better basis for them to solve the increasingly complex engineering problems and help them create a more sustainable future.

ACKNOWLEDGMENT

The author acknowledges the kind support of colleagues, Nicolas Gold and Emmanuel Letier, contributing industry speakers, and the UCL software engineering professional practice research students.

REFERENCES

[1] https://www.randstad.co.uk/career-advice/skills/uk-engineering-facing-skills-crisis-where-are-the-jobs/

[2] Harvey Nash/KPMG CIO Survey 2019 https://www.hnkpmgciosurvey.com/charts/home.php

[3] Effective Industrial Engagement in Engineering Education https://www.raeng.org.uk/publications/reports/effective-industrial-engagement-in-engineering-edu

[4] D. Fung, A Connected Curriculum for Higher Education, UCL Press, 2017.

[5] https://www.ucl.ac.uk/teaching-learning/connected-curriculum-framework-research-based-education

[6] G. Collins, "Challenges for teaching sustainability and promoting diversity within a software engineering course," in Proceedings of the 45th European Society for Engineering Education (SEFI) Conference, 18-21 September 2017, Azores, Portugal, pp. 41-48. https://www.sefi.be/conference/2017-azores/

[7] https://epsrc.ukri.org/research/framework/

[8] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation

[9] https://www.ucl.ac.uk/computer-science/about/women-computer-science

[10] G. Collins, H. Varilly, and T. Yoshinori, "Pedagogical lessons from an international collaborative big data undergraduate research project," in Proceedings of the 2015 European Conference on Software Architecture Workshops (ECSAW'15), ACM, New York, Article 32, pp. 1-6.

[11] D. L. Johnstone, "Scientists Become Managers – The T-Shaped Man," IEEE Engineering Management Review, Vol. 6 (3), pp. 67-68, 1978.

[12] M. Goldsmith, "Try feedforward instead of feedback," The Journal for Quality and Participation, vol. 26 (3), pp. 38-40, 2003.

[13] J. Novak, B. Kensington-Miller, T. Evans, "Flip or flop? Students' perspectives of a flipped lecture in mathematics," International Journal of Mathematical Education in Science and Technology, vol. 48(5), pp. 647-658, 2017.

[14] M. Vansteenkiste, J. Simons, W. Lens, K.M. Sheldon, and E. L. Deci, "Motivating learning, performance, and persistence: the synergistic effects of intrinsic goal contents and autonomy-supportive contexts," Journal of Personality and Social Psychology, vol 87(2), pp. 246-260, 2004.

[15] C. Armatas and T. Papadopoulos, "Approaches to work integarted learning and engaging industry in vocational ICT courses: evaluation of an Australian pilot program," International Journal of Training Research, vol. 11(1), pp. 56-68, 2013.

[16] https://www.raeng.org.uk/publications/reports/the-dowling-review-of-business-university-research

[17] L. Davey, "Everyone's Network Should Provide Two Things," Harvard Business Review, 30th September 2016.

[18] N. K. Semmer, et al., "The Emotional Meaning of Instrumental Social Support," International Journal of Stress Management, vol 15(3), pp. 235–251, 2008.

[19] T. Casciaro, F. Gino, and M. Kouchaki, "The contaminating effects of building instrumental ties: how networking can make us feel dirty," Administartive Science Quarterly, vol. 59 (4), pp. 705-735, 2014.

[20] K. May-Newman and G. B. Cornwall, "Teaching medical device design using design control," Expert Review of Medical Devices, vol. 9(1), pp.7-14, 2012.

From Framework Programs to Teaching: Integrating Experience from European Research Projects in Teaching Engineering

W. Brenner, N. Adamovic
TU Wien, Institute of Sensor and Actuator Systems, Wien, Austria
werner.brenner@tuwien.ac.at

Abstract - **Considering the international approach to research and teaching is essential for a research-oriented institution to function successfully, continuously develop and remain internationally visible. This paper provides an overview of the most important aspects to consider in order to successfully integrate European research projects into engineering teaching methods. This helps to facilitate the project's dissemination and utilization. Involvement of students in research projects results in a wide range of formal, non-formal and informal learning activities and effects for both students and researchers. Furthermore, this paper presents examples originating from EC's Framework Program 7 (FP7) ASSEMIC (micro handling and assembly for hybrid MEMS), FP7 SolarDesign (Photovoltaics), Horizon 2020 (H2020) CABRISS (sustainability) and H2020 EMMC-CSA (simulation).**

Keywords – European research projects; teaching engineering; dissemination; management skills; research skills

I. INTRODUCTION

The key elements of a project's ecosystem are established and applied as part of the project to facilitate its dissemination and utilization. This task is primarily relevant to define the dissemination strategy and its realization during the project. The target groups of dissemination include not only private large companies and small and medium-sized enterprises (SMEs) who act as manufacturers, application developers, system integrators, tools and methodology providers, and so forth, but also standardization bodies, such as universities and training and teaching institutions. An awareness and dissemination plan that specifies milestones, messages, target audiences and dissemination tools must be developed and applied.

Dissemination can be achieved through a variety of channels, and among these are teaching and training. Making use of connections to other networks that address related topics is highly recommended. The aim is always to raise the awareness of the project and the applied methodology, thus emphasizing the significance and effect of scientific and technological innovation. Third-level teaching institutions try to provide solid basic knowledge in combination with specific knowledge that can be chosen by each student subject to their personal interests. Internationality is promoted in teaching, research and service activities. This university-wide strategy is essential to the success, development, profile and attractiveness of the university for international students. Successful participation in European programs and fostering contacts in transnational networks contribute to the successful positioning of research teams and graduates at an international level. This includes forming bilateral university partnerships that take advantage of regional and subject-based strengths, promoting the mobility of students, teaching staff and researchers, and fostering visibility and attractiveness of the university.

Committed teaching with close links to relevant research activities guarantees that a country remains internationally competitive as a research location. Additionally, students are encouraged to learn about current research projects [1]. Funded European projects are usually characterized by a high level of enthusiasm within the consortium. Experiencing this spirit of success can be valuable for involved students.

II. INTEGRATING RESEARCH PROJECTS INTO TEACHING ENGINEERING

This paper provides examples of how Electrical Engineering/Information Technology students at the Technical University of Vienna (Technische Universitaet Wien) become acquainted with European research projects that are currently being coordinated by lecturers. Students at both the master and doctoral level can participate. The philosophy is characterized by the attempt to move away from:

- repeatedly-used exercise examples; and
- presenting well-known, well-documented standard theory.

The goal was to concentrate on current topics of high relevance for Europe and beyond. In the context of consortia meetings, there is the opportunity to get to know successful researchers and to learn from their experience in both linking their individual fields of expertise with the societal challenges and on the other hand managing European projects. Studies [9] have shown that students have the following expectations when participating in such research projects:

- Gain experience in a field of personal interest;
- Increase the chances of gaining employment with one of the project's consortium members;
- Work in a professional "real world" setting;
- Gain experience in teamwork and the related soft skills;
- Gain skills in communications, presenting results, persuasiveness and facilitation;
- Learn about quality standards and measures to guarantee them;
- Learn how to reach compromises;
- Gain better possibilities in the future labor market; and
- improve English language skills.

Integrating current research projects into teaching engineering requires substantial additional efforts from lecturers because the topics change constantly. Therefore, lecturers must be aware that the content will change year after year, depending on the development of the projects. Consequently, lectures must be updated accordingly.

III. Direct Involvement of Students in the Research Process of Projects

A. *Identification of participants and management of teams*

Students are directly involved in selected scientific topics, including the search for innovative solutions. The main prerequisites here are motivation, high academic achievement and creativity. In ideal and simple-to-implement cases, university lecturers are also in charge of coordinating a European project or at least have decision-making authority in the project. This facilitates the integration of current projects in the lectures. If such personalities are not available in the faculty, it is best to hire external coordinators as guest lecturers or visiting professors. In this case, they will design a part of a course, which probably will be blocked in nature. Lecturers and leaders of selected European projects must be willing to cooperate intensively over longer periods so that the extent of student involvement in the projects can be determined in advance. This also includes the questions of:

- Which courses to include; and
- Who can participate – all or only individuals?

Lecturers and coordinators will define the qualification requirements and communicate them to potential participants in advance. Selection criteria can include:

- Academic record (past grades);
- Personal preferences and level of commitment;
- Communication skills;
- Level of proficiency in English; and
- Personal motivation (e.g. enjoyment of new challenges).

Fig.1 shows the basic process of how lecturers identify potential participants for project work.



Figure1. Procedure for identification of students for involvement in European projects

The experiences presented in this paper have been gained over the past 20 years and represent a small excerpt from a variety of topics.

Lecturers always had direct and active involvement as coordinators or work package leaders of FP7 or H2020 projects. In this way, the definition of the topics, the depth of the expected work and the intended contact points with the project consortium have been uncomplicated. Students who are registered for three selected courses at the end of the master studies Electrical Engineering and Information Technologies (TU Wien) have the opportunity to overtake specific tasks in the current European research projects. This involvement has always been voluntary.

It may happen that not all students are suitable or willing to actively participate in a special group of the lecture. This can be the case if the students have known the standard requirements for years and have to worry about being overwhelmed with unknown project work. Then the standard lecture will take place in parallel. Experience shows that around two-thirds of the participants in the three courses included take part in project work. Regardless of the method of implementation, participating students will receive a certificate at the end of the semester. In individual cases, an additional work contract is offered, especially if the expected amount of work clearly exceeds the frame of the course and the quality of delivered work is outstanding.

In order to practice project work in a realistic manner, students form small groups of 4 to 5 participants. On average, 15 students per year were involved. For comparison purposes: around 110 students graduate in Electrical Engineering and Information Technology each year. When forming groups, the lecturer largely takes into account the wishes of the participants, while paying attention to an interdisciplinary group composition.

Similar to real project work, the lecturer and the participants define and agree on responsibilities and expected output together in a kick-off-meeting. During the semester, groups that usually deal with different topics have weekly group meetings. Participants report the progress and the lecturer can assess the engagement of the participants continuously. Additionally, the lecturer

organizes contacts with senior researchers involved in the project. If a project meeting takes place near the university in the semester, the students get the opportunity to have direct contact with the international consortium. As long as related costs are moderate, the lecturers' institutes or the European projects cover them. The working language is always English. The students' ability to communicate in English has improved significantly in the 20 years under review, but every additional exercise is preparation for the future, often international, field of activity of an engineer.

As always in university teaching, the result depends on the intensity of supervision and the commitment of the participants. In the 20 years of implementation, both patentable solutions emerged that could find their way into the documentation of the projects, but also only average elaborations.

The assessment must take into account the students' subject-specific knowledge base and the agreed scope of work. Thus, it is necessary to evaluate each case individually. Moreover, it turned out that throughout the years student satisfaction has always been very high. Nevertheless, the collaboration between project manager, lecturer and student is always a new and fresh experience. This is mainly due to the multidisciplinary and unconventional approach students demonstrate.

Involvement of students with topics from an EU-funded project make some legal framework conditions necessary. Legal departments must prepare intellectual property (IP) agreements and non-disclosure agreements in advance. It is advisable to involve only legal experts in drawing up such contracts, to ensure that all eventualities are covered. Different organizations and individuals (e.g. researchers with or without an employment contract, students, etc.) may have very different IP interests.

A detailed IP strategy that is tailored to and acceptable for each institution and respectively each individual should be provided at the beginning of the project to avoid disputes later. Students do not have the status of employees and thus are not obliged to transfer ownership of innovations they generate. They remain owners of their solutions and are free to use or exploit them. Additionally, there is some risk that students might take legal action on project funds after having assessed their performance share in the overall project.

B. Example 1: Photovoltaic (PV) Driven Street Lamp

Students with knowledge in simulation were involved in the development of a PV-driven street lamp. With FP7 SolarDesign in the background, one offered task was to simulate the expected energy output of a PV-driven streetlamp (see Fig. 2) according to the position of the sun in different seasons. Valid standards for the required brightness in public spaces had to be obtained and interpreted: “Road lighting” CEN/TR13201-1, 2, 3, 4. Fig. 3 illustrates the main training content, i.e. understanding the interdependence between design, valid standards and the minimum efficiency of PV materials to guarantee the requested luminescence on the street.



Figure 2. Simulated solar street lamp – “Solar Hub” (designed and developed by SIARQ Advanced Solar Design).

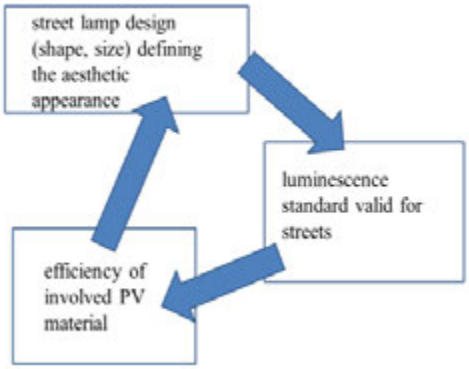


Figure 3. Example of the interdependence of design, standards and PV efficiency

The shape and size of the street lamp’s reflector define its aesthetic appearance. Thus, the diameter is limited to avoid a clumsy or bulky appearance. This restricted surface area defines the available PV-bearing area, and consequently, the electrical energy that can be converted into light. FP7 SolarDesign had to solve the research problem to increase the efficiency of the PV material and especially its temperature coefficient to guarantee the standardized luminescence on the street. The developed model met the needs of the designers/architects. Specific user-defined shapes of modules can be easily achieved, while respecting the electrical requirements. Students were involved in simulating the energy output of PV integrated into the surface of different streetlamp designs. Fig. 4 shows the results of the simulation to predict the daily power yield of the prototype in order to design the system to be fully self-sufficient.

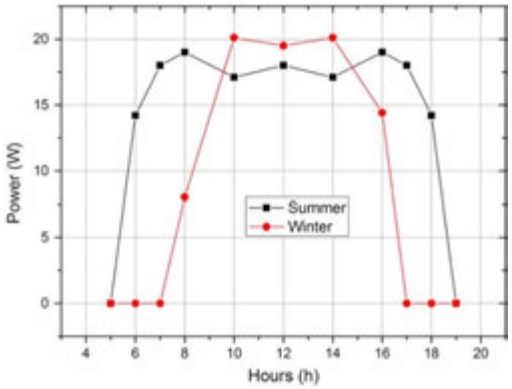


Figure 4. Simulation results – “Solar Hub”. Maximum power point at different times of the day, for a summer (June) and winter (December) in Barcelona, Spain [2].

A light-emitting diode (LED) lighting system is part of the bottom of the device, and the rest of the components (battery, electronics, etc.) are integrated into the dome. This design enables power generation throughout the daytime period and utilizes the diffused light. The shape of the lamp favors lower sun positions, while the power collection is more critical in wintertime.

In this example, experienced students were involved. The solutions that were developed were highly professional in their quality and became part of the report of the research project as well as part of a scientific publication.

C. *Example 2: Micro-handling and Micro-assembly*

Microelectromechanical systems (MEMS) have directly evolved from the mature semiconductor technologies, borrowing from them many microfabrication technologies, design methodology and working principles. However, to provide combined mechanical, fluidic or optic functionalities, it is often necessary to integrate and assemble individual micro components into hybrid microsystems. Micro-handling and -assembly represent some of the most serious difficulties in microsystems manufacturing. This defined the scientific objectives of FP7 ASSEMIC.

A variety of unresolved questions in the field of micro-handling and micro-assembly provide the basis for a variety of demanding topics for students' training, including:

- Design of flexible, modular micro and nano-handling robots;
- Design of micro actuators;
- Integrated micro sensors;
- Design and function of end effectors;
- Tools for machining MEMS components;
- Principles of handling and assembly of MEMS and, considering the ongoing miniaturization of MEMS: solutions for handling and assembly of nanoelectromechanical systems (NEMS); and
- Control techniques.

Characteristic challenges are:

- The fragility or sensitivity of micro parts, calling for advanced micro-handling technologies;
- The need to endure harsh working conditions, requiring special bonding and packaging solutions; and
- Overcoming parasitic forces and undesired adhesion effects derived from the scaling of laws.

Fig. 5 shows an example resulting from a students' creativity session focused on "solutions for micro-handling". In the next step, identified potential solutions are clustered on a flip chart and provide the basis for further discussion and selection. As all tasks are in the micro dimension, students get a basic knowledge of the applicable micro-structuring processes too.

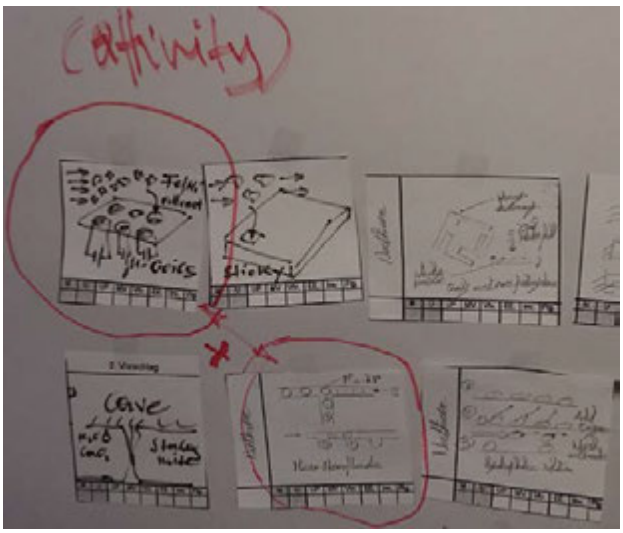


Figure 5. Example: clustered solutions for micro-handling (the photo shows the practiced working method for finding solutions on the flip chart)

Involved students develop innovative concepts for the handling of micro components and concentrate on special chapters, such as the avoidance of collisions of tools in the chamber of the scanning electron microscope. Again, reports and presentations form the basis for an assessment of the students' performance.

IV. Acquisition of Related/Soft Skills

A. *Example 3: Sustainability*

Sustainability is a multidisciplinary topic in engineering education. Students need to be able to evaluate and apply information from a wide range of disciplines including economics, engineering, environmental science and social sciences [3]. The World Commission on Environment and Development defined sustainability as "development that meets the needs and aspirations of the present generation without compromising the ability of future generations to meet their own needs" [4]. H2020 CABRISS aims at developing the technologies for recycling even large quantities of PV waste, thus allowing a reduction of the operating expenses. Being responsible for the PV waste take-back can be costly since solutions to decommission and recycle such waste are currently not technically mature [5]**.** This fact defines many thematic areas for the students to address, one of which is the waste electrical and electronic equipment (WEEE) directive of the European Union. This type of task enables the students to understand the valid regulations and their areas of application. Thus, the influence on all phases of the product life cycle is studied. The recast WEEE directive states that when supplying a new product, distributors are responsible for ensuring that such waste can be returned to the distributor at least free of charge on a one-to-one basis under the condition that the equipment is of equivalent type and of same function as the supplied equipment [6]. Thus, PV modules distributors will have an increasing role in the PV

collection network. Table I gives an overview of the responsibilities of the producers.

TABLE I. ESSENTIAL RESPONSIBILITIES OF PV PRODUCERS ACCORDING TO WEEE [6]

	Responsibilities according to WEEE
1	Yearly report of the PV modules put on the national market: WEEE register
2	Organizing and financing the take-back and waste management of the photovoltaic modules
3	Ensuring a financial guarantee of the disposal operations when photovoltaic modules are considered as household WEEE
4	Achieve mandatory collection and recycling targets
5	Informing end customers on how to dispose their End-of-Life (EoL) PV modules
6	Informing treatment facilities of the product's composition, including the potential use of hazardous materials

Students acquainted themselves with the Life Cycle Assessment (LCA), performed according to ISO 14040/14044, which is the internationally most recognized evaluation method to compare the environmental impacts of different production processes and technologies. This covers the systematic compiling and examining of all inputs and outputs of materials and energy and the associated environmental impacts of a product "from cradle to grave" [7].

The PV recycling system starts with the collection, the pre-sorting and transportation of End-of-Life PV modules. The next step is a recycling plant where disassembling and separation of the main components (frame, wires, eventually glass, cell material etc.) are carried out. Material products of these manufacturing steps can be sold to the market as basic materials and pre-products. They could also stay in the recycling loop (as it will be assumed here). Since the focus is on specific valuable materials like silver, indium and silicon semiconductor technologies, it can be assumed that the frame metals (mainly Al) and glass will be delivered to the respective markets and not stay in the loop.

Students learned that the efficiency of the recycling process is influenced by the choice of the location for the recycling plant. The use of synergies in production is important to minimize both transportation and investment costs. The results show that products from the investigated lab-scale processes are generally not competitive on the market. However, most of the industrial-scale processes are suitable to develop attractive business models. The major part of the End-of-Life PV consists of glass, aluminum and polymers. The share of mass is 97 % in case of c-Si technologies and even higher in case of thin film technologies. The economic value of these materials is very dominant.

B. Example 4: Project Management

This comprises sharing a broad knowledge of the institutional framework of the EU as well as its mechanisms. In focus are the necessary skills in management of EU project issues, which combine theory, research and practice. Thus, participants become equipped with the fundamentals of both knowledge and skills for the management of interdisciplinary teams at the European level.

Particular attention is paid to selected areas of project management (see Table II), which have a significant influence on a project's success.

TABLE II. LEARNING CONTENT FOR PROJECT MANAGEMENT

	Content
Introduction to EU's research funding philosophy	EU structure and EU programs
Societal challenges	Environment Stakeholder management Gender aspects Teambuilding Sustainability in engineering
Project objectives and finances	Objectives Tasks Business plan
Cooperation	Communication plan Dissemination plan Networking of networks
Quality	Quality management Risk management
Exploitation	Patenting Intellectual property Standardization on European level

The learning content also includes how to identify an excellent project consortium, how to get funds, and how to select research topics of relevance for Europe and beyond. Often these "hot" and promising topics are close to the societal challenges [8]. Students are taught how to write and evaluate a proposal, and how to calculate project costs. They learn how to get information on the results of previous projects. Many deliverables (i.e., documents bundling the results of European research projects) have the status "public", meaning that their content is available to everyone. Students are trained on how to identify and access this information.

Discussion and analysis of research results are practiced and promising applications are identified. The processes employed include:

- Comparison of products, processes or services created in the frame of the project with the pre-established solutions that already are available;
- Relevance and prospective impact of the yielded innovation (novelty in its sector and on the national and international level); and
- Technical and economic advantages in comparison with currently available solutions.

It is advisable to familiarize the students with the relevance of their projects for addressing societal challenges. Thus, they can classify their own project ideas according to their relevance for the society:

- Clean, secure, sustainable and efficient energy;
- Industrial renewal;
- Life, health and well-being;
- Food security and demographic challenges;

- Sustainable mobility and urban systems information and communication society;
- Innovative, inclusive and adaptive societies;
- Freedom and security of Europe, its citizens and its residents; and
- Resource management and adaptation to climate change.

V. CONCLUSION

Third level teaching institutions promote internationality in teaching, research and service. This strategy is essential for the success, development, profile and attractiveness of universities.

European research projects coordinated by lecturers of such institutions give an excellent opportunity to link classes with projects and to involve students who are both motivated and skilled. Tasks originate from research and/or management content of European Projects. Generally, legal aspects concerning intellectual property must be available in advance. The effort for the lecturers is high but the benefit is on both sides – for the students and also for the consortia/universities.

ACKNOWLEDGMENT

The mentioned European projects have received funding from the EU under grant agreements as given below: FP7 ASSEMIC n° 504826, FP7 SolarDesign n° 310220, H2020 CABRISS n° 641972, H2020 EMMC-CSA n° 723867.

REFERENCES

[1] Technische Universitaet Wien, Wien Austria, website, https://www.tuwien.at/en/studies/welcome/.

[2] N. Bednar, A. Caviasca, P. Sevela, N. Severino and N. Adamovic, "Modelling of flexible thin-film modules for building and product integrated photovoltaics," Solar Energy Materials and Solar Cells, Volume 181, pp. 38-45, July 2018, https://doi.org/10.1016/j.solmat.2017.12.035.

[3] S. K. Lascano, J. H. Núñez, I. E. Esparragoza, L. C. Schmidt and R. L. Nagel, "A case study approach for teaching students sustainability from a global perspective," IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, Volume 10, pp. 109-118, August 2015.

[4] C. Thomsen, "Sustainability - world commission on environment and development definition," in Encyclopedia of Corporate Social Responsibility, S. O. Idowu, N. Capaldi, L. Zu, A. D. Gupta, Eds. Springer-Verlag Berlin Heidelberg, 2013, https://doi.org/10.1007/978-3-642-28036-8.

[5] H2020 project CABRISS „Implementation of a circular economy based on recycling, reused and recovered indium, silicon and silver materials for photovoltaic and other applications," website, https://www.spire2030.eu/CABRISS.

[6] Directive 2012/19/EU of the European Parliament and of the Council of 4 July 2012 on waste electrical and electronic equipment (WEEE) Text with EEA relevance, Official Journal of the European Union L 197/38 , Document 32012L0019.

[7] W. Brenner, N. Bednar, P. Biermayr and N. Adamovic, "Standardization and life cycle cost assessment approach in circular economy for photovoltaic waste," in proceedings of 3rd International Conference on Smart and Sustainable Technologies SpliTech, Split 2018, pp. 1–6, ISBN: 978-1-5386-6296-0.

[8] European Commission, website, https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges.

[9] T.W. Druzovec, L. N. Zlatolas and N. Escudeiro, "Slovenian students on project/internships," in proceedings of the 24th Annual Conference of the European Association for Education in Electrical and Information Engineering EAEEIE 2013, Chania 2013, pp. 210-214, https://doi.org/10.1109/EAEEIE.2013.6576531.

Teamwork challenges and solution strategies of first-semester engineering students

A. Gorup, M. Grzunov, J. Petrović* and P. Pale*
* Department of Electronic Systems and Information Processing
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, HR-10000 Zagreb, Croatia
{ Andrija.Gorup , Matej.Grzunov , Juraj.Petrovic , Predrag.Pale }@FER.hr

Abstract **–** Teamwork is an important skill for today's engineers, who often work in international teams where decision making, problem solving, and all communication aspects need to be dealt with efficiently. It is important therefore to familiarize engineering students with challenges they will face in such situations and provide them with related practical experiences before they enter the labour market. And while engineering students typically have opportunities to work in project teams, those teams are often assembled by students themselves, unlike real project teams in which engineers cannot always choose their teammates**.** This paper explores dynamics of student teams formed in a way that maximizes group heterogeneity in terms of academic success, geographical origin, and gender, and thereby minimizes probability of students already knowing each other. In such conditions, students are assigned a task to solve in teams of five. Teams' work results and reflections collected through a questionnaire were used to identify typical challenges students face, and strategies they employ to solve them. Educational implications and improvements of this intervention are discussed in the paper.

Keywords – teamwork; groupwork; team communication; team formation

I. INTRODUCTION

Thanks to global migrations, the invention of the internet, and the process of globalisation, the world is more connected today than it ever was before. Most modern-day workplaces are diverse regarding employees' cultural and professional background and often require the employees to work as groups or teams, sometimes with complete strangers [1]-[3]. Since working in teams is considered to lead to higher efficiency [4], it is expected of engineers to work on projects, sometimes quite expansive and with many people they do not know, and this can cause conflicts to spark between co-workers. Conflict solving and efficient communication are, therefore, crucial skills for engineers entering the workforce to possess and that should be an integral part of higher education [3], [5]. Small task-oriented groups offer students an important insight into their later work in international teams as they learn the importance of teamwork [3], [4], [6], the benefits of which exist independent of the cultural similarity between students of the same university [7].

In higher education today, students' task-oriented teams are most commonly formed by letting the participants group themselves and choose their partners, a method unreminiscent of real-world project grouping where engineers seldom have the possibility of picking team members or co-workers. This research explores several aspects of teamwork challenges students face when working in teams with other students they typically haven't met before. The aim of the study is to explore which conflicts arise during work on the project, how the problems are dealt with, and how these relate to students' relationships after the project is completed.

The rest of this paper is organized as follows. The grouping methods commonly used in engineering education today are summarized in section II. The method of grouping used in this research, as well as its context and research questions and methods are described in section III. Research results are described and discussed in section IV, and conclusions are presented in section V.

II. COMMON GROUPING TECHNIQUES

There are several factors to consider when forming a group, most important of them being the size of a group and the method of allocating group members. Studies of task-oriented student groups show that the optimal number of members in a group is around four, and that going over the five to six member number can reduce the efficiency of the entire group by a noticeable amount and yield poorer results [6], [8]. The successfulness of five member groups is the result of the following factors. It is easier for such groups to adjure meetings, each member feels more involved in the work of the team and it creates interdependence [4], [8], [9]. In particular the key to good all-around group performance is outcome interdependence and its absence in smaller groups can cause members to act as though the assigned work is part of an individual task [9]. The downside of small sized groups is the possibility that students may be overwhelmed if one or more members cannot participate in the group's work. [8]

The method of allocating students in a group refers to the manner in which students are designated partners with whom they work on a given project. Generally there are four commonly used team member allocation methods [8]: self-selection, random appointment, selective appointment and task appointment.

- Self-selection is a group forming method where students select their partners themselves. This is generally perceived favourably among students and is the easiest to implement. Its drawbacks

include some students struggling finding partners and may be disadvantaged as a result. Moreover, it is highly unrepresentative of team formation in the workplace students are being prepared for.

- Random appointment entails a computer based random selection of partners or using a less sophisticated method of grouping based on relative in-class position. This way of partner assignment does not require substantial effort and it replicates to an extent real-world grouping scenarios. This allows people to potentially work with people with whom cooperation was otherwise highly unlikely. This and the former method are beneficial when a large number of students are to be split into small groups.
- Task appointment refers to students being offered task related topics which the students may choose and in doing so they indirectly choose their partners. Creating teams this way ensures all members have a positive perception of the topic which will raise optimism and cooperation with the downside being the possibility of uneven topic selection, where some topics will experience great popularity and other ones general avoidance.
- Selective appointment consists of the shared criteria method where students are grouped based on a common goal or interest, for example their stated grade ambitions. An example of this approach is CATME tool [10], which supports representative team formation based on students' survey responses. Specific criteria or algorithms are used to create representative groups where differently skilled students balance each other out. The former has the advantage of every category of student will be pleased with the grouping process as they will feel their expectations are being met by every other member, although segregation of this type can result in homogeneous student groups based on ability where lower ability students usually see little benefit [6], [9], [11]. Conversely, the latter attempts to shape heterogeneous teams and thus positively impacts student self-esteem and perception while presenting a group scenario somewhat similar to modern work environments. It does however take a lot of effort to implement and can be hindered by a majority of students having the same skill set.

It is important to note that no method by itself is conclusively best and each one has applications in different circumstances, and it is ultimately up to the organiser to decide on which to use.

III. MATERIALS AND METHODS

A. Team formation and assignment

This research was conducted within a first semester *Communication skills for engineers* course at the University of Zagreb, Faculty of electrical engineering and computing. The course is focused on most relevant communication skills for engineers, specifically knowledge communication, knowledge presentation, formal communication, cultural and individual differences, and conflict resolving. Within the course, one of the assignments in to prepare an elevator pitch – a 60 seconds presentation about what and how students would change at their institution. Students' teams were formed from five (when necessary, four) members through selective appointment with the goal of making teams as diverse as possible regarding their place of origin, their previous educational achievements, and gender. Student teams were announced using an email sent to every member of every group and containing a description of the task and a list of email addresses of other team members. The team members should prepare the pitch together, yet only one team member would be *ad-hoc* selected by the course lecturer to deliver the pitch in front of the whole auditorium of approximately 200 students. The pitch was graded by the lecturer (65%), and by the auditorium (35%), based on the pitch idea's viability, importance, and presentation quality. The course credit received for the pitch was finally distributed between team members by themselves.

B. Research questions

The objective of this research was to investigate students' perception, behaviours, and outcomes of the assignment, and find support for, or discard, the following hypotheses:

1. Students dislike being grouped with unfamiliar people since they know it brings them out of their comfort zone and requires them to solve conflicts within the group.
2. Conflicts that typically occur in groups include dismissing good ideas by majority decision, uneven motivation and effort of team members, and credit distribution among team members.

Aside from the two hypotheses, additional points of interest were:

- How successful students think they were in solving conflicts in their group?
- Do students plan to stay in contact with their team members after the end of the projects?
- How do students distribute project credit among themselves?
- How do students perceive their individual responsibility in the project team?

These hypotheses and points of interest were investigated using an 18-item survey conducted online after the pitch presentations. Most items in the survey came in pairs, where the first item was a multiple-choice question and the second allowed a student to type in a custom answer that could further explain their choice or define a new one when none of the offered answers was appropriate.

IV. RESULTS AND DISCUSSION

In total, 400 out of 670 first year engineering students (81.97% of those who participated in this course activity) completed the survey (76.25% M, 23.75% F).

A. Time efforts and distribution of work

The initial survey items investigated the amount of time students had invested in the assignment. On average, students spent 4 hours 40 minutes (with a standard deviation of 16 minutes) or 4 hours (median) working on the assignment. Students who rated their time efforts on the task as appropriate (59.90%) had spent on average 3 hours and 58 minutes (with a standard deviation of 17 minutes) or about 4 hours (median) working on it. Those time frames are in line with the expected assignment preparation time of around 5 hours. The assignment workload was planned to be relatively low to facilitate students succeeding in it, even if a student would have to prepare it on his own due to unfavourable circumstances like group conflicts.

The next point of interest in this research was to assess whether workload within groups was approximately evenly distributed. This was assessed by calculating standard deviation (SD) of the group members' self-reported amounts of time spent on working on the assignment. The distribution of SD values for all groups is graphically displayed in the Fig. 1 and suggests that for a majority of groups the time invested in the project was approximately equal for all group members. More than two third of groups had SD less or equal to 2 (approximately 50% of the mean value).

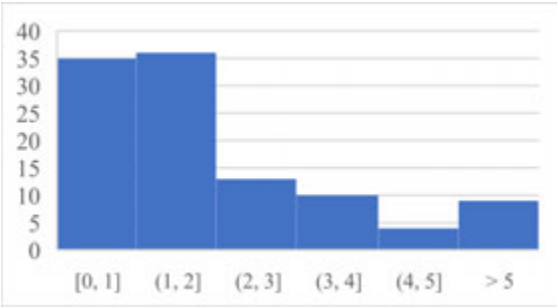


Figure 1. Standard deviation (x-axis) of the amount of time different members of a group had spent working on the task

The same research question was investigated, additionally, from students' perspective by asking students to assess how much their group members have been working on the assignment. In response to this survey item, 57.29% of students said their team members put the same effort as they did in the assignment, 36.20% said some of their team members put higher, and some lower effort than them, while 6.51% of students said all of their team members had put in a lower effort than them. Those results are similar to results reported in Fig. 1., which means either that students were equally motivated for the task, or that they successfully managed work distribution within a group.

The fact that all team members had dedicated an approximately even amount of time to working on the assignment does not imply that time was well used. Therefore, the third point of interest in the research was whether students have managed to utilize the potential of their group. For example distribute the tasks to use time more efficiently, or to benefit from different perspectives and ideas of group members. Of course, it is difficult to directly measure whether students have benefited from working with others, and not alone. In a survey item labelled "Do you feel you would have been equally or more successful if you had done the assignment by yourself?" questioning students' opinion on the topic 19.54% of students reported they would be more successful as they had been if they did the assignment on their own. A further 29.19% reported they would have been less successful had they been working on their own, and the rest (51.27%) believe they would be equally successful on their own. This assignment was not as demanding that students would depend on each other, but it is important to see that some of them have been successful in recognising or using some of the benefits of working in a team.

B. Grouping method

As described in the previous section, groups for this assignment were formed in a way to maximize the likelihood of students not being acquainted with their team members. In the part of the survey related to students' stress they were given the following statement: "Groups were formed in a way to maximize the likelihood of you not being acquainted with your fellow group members. How much stress did the fact you would have to work with strangers cause you?" Multiple answers were possible in that survey item and 65.91% of students reported that they experienced no stress linked to working with strangers. Students who did feel stressed reported as their main reason the uncertainty about team members' reliability (54.66%), uncertainty about team members' motivation (32.67%) or discomfort with teamwork in general (12.67%). In addition, some students reported they experienced stress for at least two of the aforementioned reasons (11.14%) and a small number experienced stress for all three of the stated reasons (1.36%).

C. Outcomes and grading

Pitch outcomes were graded by one of the course lecturers and by other students in the audience. The credit awarded by the lecturer had an average of 65.89% (SD=0.15), while the credit awarded by the audience had an average of 76.00% (SD=0.06). A moderate statistically significant correlation between students' and the lecturer's grades was found (Pearson's $r=0.43$, $p<0.001$). No correlation was found between group's average or total working time and the lecturer's grade.

In order for the group members to get their final assignment credit, the credit awarded to the group by the lecturer was multiplied with the number of group members

subtracted by one, and every group member was then able to grade all other group members by distributing that amount of credit between them. Final credit for every group member was calculated as the average credit awarded to him by other team members. This procedure was anonymous and other group members would only get their average score based on all their group members' grades. When asked about how had they distributed the credit among group members, 44.90% of students reported having done that based on an agreement between group members, while 55.10% of them reported dividing the credit based on their own opinion about group members' contributions.

D. *Within-group conflicts*

Within-group conflicts were another point of interest in this research. It was assumed that conflicts could arise mostly related to choosing the pitch topic, differences in team members' motivation, and distribution of credit between group members. Based on the reported results, however, it seems that very few conflicts emerged. Students almost unanimously reported that no conflicts occurred related to pitch topic selection (92.31%), uneven motivation of team members (95.35%) or credit distribution within groups (98.71%). Overall, 90.57% of students stated no conflicts arose during the preparation of the group project. Students were also encouraged to explain any other source of conflict that emerged during the assignment or to comment on that topic. Alongside with topic selection, uneven motivation and credit distribution, the only other sources of conflict were the inability to organize group meetings, mentioned by five students (1.29%), and the unwillingness of some students to communicate with the rest of their group, mentioned by four students (1.03%). Regarding the credit distribution, over half (55.10%) of students said there was no agreement involved and they handled the distribution at their own discretion. This is probably why only 1.29% of students mentioned a conflict occurring in this regard. When it comes to topic selection, out of the 30 students that stated a conflict occurred in this regard, 17 (56.67%) said an idea which they thought was great got outvoted by others in the group, while 13 (43.33%) persuaded others into choosing the topic they thought was the best. A high chance of multiple good ideas developing within the same group combined with a high impact a group's topic has on the quality of their outcome, and therefore their grades, is most likely why it is stated as the most common conflict cause. 18 students mentioned a motivation gap related conflict. A total of 11 (61.11%) of these reports said the conflict was solved with a mutual agreement, while seven (38.89%) said the conflict was not solved with an agreement. Regarding the motivation gap, one student also stated that *"... this is a natural phenomenon which we will have to get used to if we are to work in group environments."* Lastly, organising meetings did not seem to be a big problem in this task due to the relatively small amount of effort needed for this task as well as the possibility of online meetings.

The results suggest not many conflicts emerged during students' working on the assignment, and even if they did, students resolved them successfully.

E. *Follow-up*

In two questions related to pitch follow-up, students were asked to identify any skill that they believe would significantly help them in getting most out of teamwork or that they think they should improve. Students were given the possibility of choosing more than one answer and on average each student selected 1.66 of the offered communication skills. Percentages of students for each of the offered skills is displayed in the Fig. 2, with most students (44.00%) reported wanting to improve their creativity in problem solving.

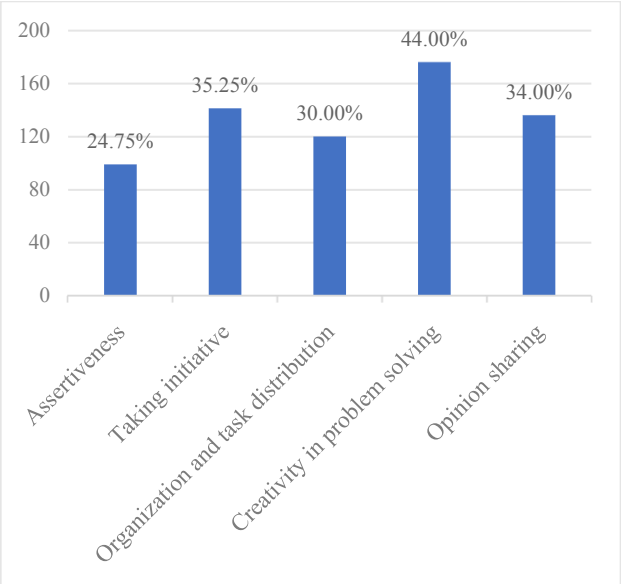


Figure 2. Skills that would help them solve the task better

Finally, 56.15% students reported they would remain in touch with at least one of their team members and 43.85% saying they wouldn't. While this was not the goal of the assignment, it is reported as a potential measure for the success of group communication and efforts. A more demanding task that would require students to work together longer would probably increase the probability of them establishing relationships that would last beyond the formal task requirements.

V. DISCUSSION

The obtained results suggest students mostly did not experience significant issues with completing the task and did not experience stress due to its context. Slightly more than 90% of students stated no conflicts arose during the preparation of the pitch. Still, the low time-intensity of the

task, combined with a relatively small impact this project has on students' grade, could be the reason for a vast majority of students not reporting any conflicts. It is much more probable someone will firmly advocate an idea when higher risks are involved, and that might lead to a conflict regarding idea dismissal. Having an unmotivated group member is far more strenuous in such a scenario. However, students should be taught not to be afraid of these situations and how to handle them with low stakes projects. The data gathered in this study show that student's fear and stress caused by being grouped with strangers is mostly irrational. This can be seen in the disparity between the stress caused by a thought that other group members will not be as motivated or reliable as they should be. Even though this stress was experienced by approximately one third of students, less than five percent of students reported a motivation gap, and less than two percent reported an unfair or uneven point distribution. These statistics hint at the aforementioned stress being irrational and, therefore, subjecting students to more of these low time-intensity projects could help in reducing the stress and help students adapt faster to their future workplaces. It is also important to note that even if a conflict does occur during group projects, it can be helpful in preparing students for a potential conflict in the future, as well as give them some practice in solving conflicts.

VI. CONCLUSION

The research described in this paper was conducted to assess the impact of an elevator pitch task on computer and electrical engineering students within a first-semester communication skills course. The results of the research suggest that a vast majority of randomly formed groups experienced no specific conflict and that about two thirds of students experienced no stress related to working in them. Even though these results are achieved in a safe environment with a low cost of failure (low assignment credit), these results are important for adjusting educational practices in order to help students obtain these skills which will be irreplaceable once they enter the labour market and start working in teams with other professionals. Further research could prove useful in helping to understand how risks involved in a group project impact the occurrence of conflict within group projects and how they impact students' relations.

REFERENCES

[1] J.D. Herbsleb, and D. Moitra, "Global software development", IEEE Softw., vol. 18, pp. 16-20. Mar/Apr 2001, doi: 10.1109/52.914732.

[2] N. DiTomaso, and G.F. Faris, "Diversity in the high-tech workplace: Diversity and performance in R&D", IEEE Spectr., vol. 29, pp. 21-24, Jun. 1992, doi: 10.1109/6.254039.

[3] E. Pfaff, and P. Huddleston "Does it matter if I hate teamwork? What impacts student attitudes toward teamwork", J. Market Educ., vol. 25, pp. 37-45, Apr. 2003, doi: 10.1177/0273475302250571.

[4] Michael A. West, and Joanne Richardson, "Teamwork", The Blackwell Encyclopedia of Sociology, Edited by George Ritzer

[5] M. R. Hovde, "Preparing students for global engineering workplace communication", paper presented at the American Society for Engineering Education Conf., Indiana Sectional Conference DeKalb, IL, 2005.

[6] Yiping Lou et al. "Within-class grouping: a meta-analysis", Rev. Educ. Res., vol. 66, no. 4, pp. 423-458, Winter 1996, doi: 10.2307/1170650.

[7] K. Jablokow, M. Myers, "Managing cognitive and cultural diversity in global IT teams", 2010 5th IEEE Int. Conf. on Global Software Engineering, Princeton, NJ, 2010, pp 77-86, doi: 10.1109/ICGSE.2010.17.

[8] "Forming effective groups", part of "Enhancing Experiences of Group Work", Univ. Technol. Sydney, Inst. Media & Learn., Accessed: Nov. 10, 2019. [Online]. Available: https://www.uts.edu.au/research-and-teaching/learning-and-teaching/enhancing/enhancing-experiences-group-work/forming-effective-groups

[9] E.G. Cohen, "Restructuring the classroom: conditions for productive small groups", Rev. Educ. Res., vol. 64, pp. 1-35, Mar. 1994, doi: 10.3102/00346543064001001.

[10] F. Jahanbakhsh, W.-T. Fu, K. Karahalios, D. Marinov, and B. Bailey, "You Want Me to Work with Who? Stakeholder Perceptions of Automated Team Formation in Project-based Courses," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, Colorado, USA, May 2017, pp. 3201–3212, doi: 10.1145/3025453.3026011.

[11] Z. Wang, "Effects of heterogeneous and homogeneous grouping on student learning", M.S. thesis, Univ. North Carolina at Chapel Hill, Chapel Hill, USA, 2013, doi: 10.17615/j3k8-fz95. [Online]. Available: https://cdr.lib.unc.edu/concern/dissertations/9z903079v

Differences in the Students' Achievements between Traditional and Project-based Learning of Basic Engineering Competencies: A Quasi-experimental Study

D. Purković*, M. Prihoda **
* University of Rijeka/Study of Polytechnics, Rijeka, Croatia
** Primary school Krk, Krk, Croatia
damir@uniri.hr

***Abstract* - Project-based learning (*PjBL*) can greatly affect the output competencies of technology and engineering students. However, *PjBL* is still not central to the curriculum of this education. This paper provides an example of *PjBL* in engineering education and provides research findings on differences in student achievement between project-based teaching versus traditionally organized. The *PUD-BJ* project learning model was used for implementation, and the research was conducted on a sample *(N = 64)* of undergraduate students of the Polytechnics of the University of Rijeka. Given the limitations, the research was conducted as a quasi-experimental intervention study. Common, traditional methods and instruments were used to evaluate achievements. The results show that the dominated achievement of students in traditional teaching is similar to project-based learning, but with significant individual differences. While in traditional teaching there are no students whose achievements are significantly higher than dominated, in project-based learning some students show exceptional achievements. The obtained results show the main advantages of project-based learning: it gives the opportunity to students who can and want more for exceptional achievements, motivates them for further research, contributes to self-awareness and development of their own learning paths.**

Keywords* - *engineering competencies; engineering education; higher education; project-based learning; technology education

I. INTRODUCTIONS

Engineering, as well as the education of engineers, is inevitably linked to problem solving, product design and construction, and their manufacturing and use. Therefore, this kind of education also needs modern learning strategies and approaches, including project-based learning (*PjBL*). Project-based learning is emerging as a popular strategy for maximizing self-study skills and promoting students' intellectual development while learning the basics of the subject [1]. Project learning is based on constructivist principles that emphasize the importance of a specific teaching context [2] in which students are actively involved in the learning process, and achieve their goals through social interactions and sharing of knowledge and understanding [3]. Such learning should be central to the curriculum of teaching, with questions or problematic tasks guide the student to the main concepts and principles of the teaching area, include student research that requires exceptional mental effort, allow for a high level of student autonomy and responsibility, and provide students with the necessary sense of authenticity with which education and the "real world" are linked [4]. Project-based learning, as a rule, begins with a specific assignment that requires the completion of one or more tasks that lead to the creation of the final product, design, model, device, or computer simulation, and ends with a report and presentation of what has been done [5]. Therefore, it represents a desirable and even a key component of engineering education programs [6].

When it comes to Engineering Competencies, research suggests that engineering teachers should focus on developing "general engineering competencies" rather than separately developing general and engineering (technical) competences [7]. Namely, the changes brought about by the professional engineering context also affect the requirements of engineers and engineering education and shift the focus from internal engineering activities to consulting firms, globalization, rapid technological changes and the development of technical specializations, the critical relationship of society, and environmental issues [7], [8], [9]. These changes highlight the gap between competencies developed during engineering education and competencies required for engineering work. This requires a change from the current focus on the theory and analysis of abstract problems according to communication, teamwork, management, problem solving, practical engineering and context [7]. [10]. In this regard, researches indicate significant deficiencies in the development of communication, social and leadership skills [11], [12], [13], insufficient development of emotional intelligence [14], employer dissatisfaction with the practical application of the theory and business skills of engineers [10], [15], [16] but also poor fundamental and engineering knowledge [17]. These gaps confirm research that demonstrates that there is no significant correlation between students' academic achievement scores and their actual work performance [7], [18] and that mental resourcefulness, entrepreneurialism, and interpersonal skills are correlated with performance at work [19]. Researches also proves that students can earn

high grades without demonstrating the competencies required for engineering practice and vice versa [7]. Given the view that engineering education should be aligned with engineering work, the challenge of engineering education is to prepare students for their first job at the same time, but also for long-term career development [7], [10], [11], which means that engineering education must provide initial competences for engineering work, but also developmental competences for successful career development. These are specific but also generic competencies that can be defined as knowledge, skills, attitudes and values that are important in all areas of engineering and facilitate the success of engineers as individuals and their contribution as engineers in good society [7], [20]. These competences include technical and non-technical competencies required in all engineering disciplines related to communication, ability to work in teams, developing professional attitudes, engineering business skills, problem solving, managing your own learning, critical thinking, creativity (inventiveness) and practical engineering skills [7], [10]. In doing so, communication skills, teamwork, self-management, professionalism and inventiveness are recognized as important for every engineer. This in no way diminishes the importance of engineering (technical) theories, but rather emphasizes the necessity of integrating them into meaningful concepts that allow students to apply them and really understand them. For this reason, complex project activities certainly play an important role in the development of engineering competencies, as evidenced by initiatives such as CDIO [21] and Engineers Without Borders [7].

Although many studies point to the importance of project-based learning, they often find it lacking in the certainty of a cause-and-effect relationship between *PjBL* and positive learning outcomes [22]. Most studies lack the random distribution of participants in control and experimental groups [22], which is difficult to achieve when it comes to higher engineering education. Namely, engineering students themselves already form a group that is not part of the wider population, and the conditions of project learning are difficult to fully control. Despite the objections, the research findings indicate many contributions to achievements, such as increased levels of student engagement, increased interest in content, stronger development of problem-solving strategies, better management of their own learning and greater depth of learning, better integration with the industrial environment, and transfer of knowledge and skills to new situations [4], [22], [23], [24].

II. THE PROBLEM AND AIM OF THE RESEARCH

Despite the many positive findings, the major differences between project-based and traditional teaching in students' specific achievement are often not clearly distinguished. The main problem addressed by this research focuses on differences in student achievement that accomplish certain learning activities in the traditional way and those with which this teaching is performed as project-based learning (*PjBL*). Therefore, the aim of this research is to determine as accurately as possible the differences between the achievements of students with whom certain learning and teaching, with the same goals and learning outcomes, are realized as traditional and those with which the project approach is implemented. The main aim of this research may be clearer if the research question is asked: *What are the main differences in student achievements between traditional and project-based learning in the implementation of an engineering material processing practicum, as an isolated segment of the engineering education?* In this research, it is an engineering practicum of material processing where students acquire a part of basic engineering competencies, among them are generic (teamwork, communication, problem solving, inventiveness), as well as practical engineering competences (product design, documenting, production of digital materials and contents, solving technical problems, application of technology and tools for making prototypes, and acquiring procedural knowledge related thereto).

III. METHODS, PROCEDURE AND SAMPLE

A. Methods

The research was conducted as a quasi-experimental design that was repeated twice over four academic years, with two experimental and two control groups. Given that the groups were neither uniform nor randomly selected, pre-screening of students' experiences in each group was conducted to determine their uniformity. This phase (some kind of pre-test) included examining students' experience with engineering and technical means, experience in technical drawing and design, and experience in mechanical engineering, electrical engineering, programming, and the use of ICT and e-learning environments. These areas were selected as important for the realization of projects in engineering, but also for the development of basic engineering competencies related to the specifics of this study. Pre-screening was conducted during the first week of class, and students' experience were evaluated on a scale: 1 - poor, 2 - sufficient, 3 - good, 4 - very good, and 5 - excellent. A t-test for independent samples compared the previous experience between the experimental and control groups of students and found that there were no statistically significant differences between groups (Tables 1 and 2). Subsequent analysis of the students' previous experience revealed that for most students, this is their first direct experience with

TABLE 1. T-TEST OF STUDENTS' PREVIOUS EXPERIENCE BETWEEN GROUPS A AND C

Areas of student experience	*t*	*df*	*P*
Experience with technology and technological equipment	0,591	27	0,560
Mechanical engineering experience	-0,920	27	0,366
Technology design and drawing experience	-1,955	27	0,061
Experience in the field of applied sciences and mathematics	-0,218	27	0,829
Electrical engineering experience	-1,359	27	0,185
Programming and program languages experience	-0,428	27	0,672
ICT and e-learning environments experience	-1,531	27	0,137

materials and using a tools, machines, and *hands-on* activities.

TABLE 2. T-TEST OF STUDENTS' PREVIOUS EXPERIENCE BETWEEN GROUPS B AND D

Areas of student experience	*t*	*df*	*p*
Experience with technology and technological equipment	-0,785	33	0,438
Mechanical engineering experience	-0,623	33	0,537
Technology design and drawing experience	-0,968	33	0,340
Experience in the field of applied sciences and mathematics	-0,405	33	0,688
Electrical engineering experience	-0,474	33	0,639
Programming and program languages experience	-0,683	33	0,499
ICT and e-learning environments experience	-0,040	33	0,968

B. Research procedure

The first part of the practicum (four exercises) was performed in the same way with each group, in the traditional way to adapt the students to the environment and rules of work. This part is realised as a set of standard exercises for creating a given object by processing materials (paper, wood, plastic and metal) with specific partial problem (engineering) tasks. This was followed by student preparation of additional documentation (technical drawings, reports, solutions). With the control groups, this type of work continued throughout the duration of the practicum, while with the experimental groups, further teaching was conducted as project-based learning. The planning and organization of project-based learning was carried out according to the *PUD-BJ model - from idea to product* [25], which consists of three segments: teaching

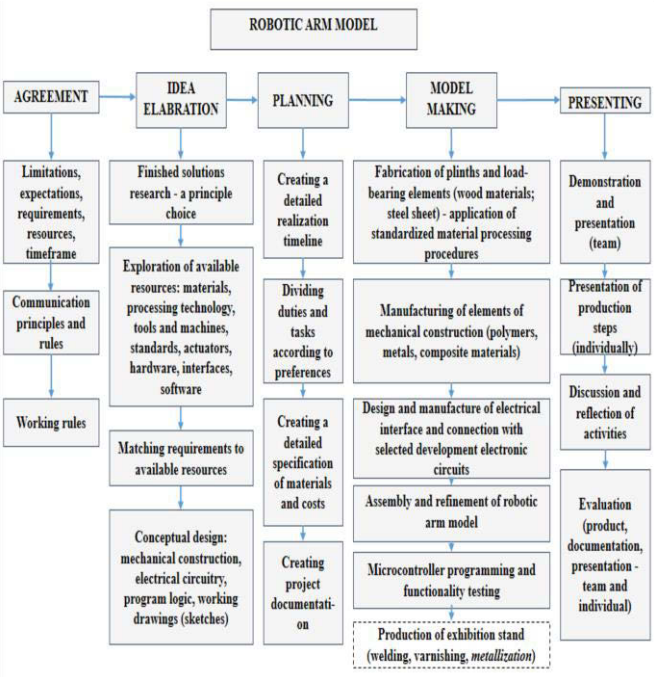


Figure 1. Macro-plan according *PUD-BJ* model

objectives, macro-plan (Figure 1), and micro-plan. An example of developing project-based learning for experimental student groups is presented here, in which expected learning outcomes are integrated into the development of a complex product - a robotic arm model.

The learning objectives were some of the more elaborate goals envisaged in the curriculum of the practicum, broadened by more complex outcomes from multiple fields of engineering. The macro-plan (Figure 1) represents the very roughly structured stages and major activities that students are expected to experience during the project. A micro-plan is a more detailed elaboration of a macro-plan that provides for each stage *what*, *how*, *where*, *when*, *who*, *with what* and *why* something is done. Although project-based learning cannot be fully predicted, a teacher necessarily needs this guide to be able to monitor and moderate student activities. The idea research and development stage took about 6 hours, the planning stage 3 to 6 hours, and the model development stage 18 to 21 hours, and each team had 3 hours to present their own work. To help with student research, problem solving, exchange of solutions and documents, a specially designed knowledge base (KB), integrated into the existing learning management system (LMS) environment, and was used. The features of this knowledge base and knowledge management system (KMS) included simplified visualizations, basic and / or simulated abstractions, segmented problem situations and tasks that lead student to success, ability for basic interactivity and collaboration, and clear individual and group expectations from students [23]. After completing the practicum and collecting data, a t-test was also used to later compare achievement between control and experimental groups.

C. Sample

The research was conducted on a sample *(N = 64)* of undergraduate students of the Polytechnic, University of Rijeka, during four academic years. The control group consisted of group A *(N_A = 15, 2014*) and group B *(N_B = 18, 2015),* and experimental group C *(N_C = 14, 2016*) and D *(N_D = 17, 2017).* These are students of the fifth semester of study, who have acquired theoretical backgrounds in sciences, technical drawing and design, mechanical engineering, electrical engineering, computer science and energy. The sample of students consisted of all students of each year of the study. The classes with experimental groups C and D were conducted according to the project-based learning model described above. The topics of the projects were chosen by the students in agreement with the teacher, but also according to the needs and limitations of the partners and conditions. Teaching with control groups A and B was conducted in the traditional way, through prepared exercises to solve partial engineering problems and to apply material processing procedures for product making. Project activities were carried out in teams of 7 to 10 students, while traditional classes were conducted as individual or tandem work.

D. Data collecting

Since the aim of the paper is focused on the differences in student achievements between project and traditional learning, data on student achievements were

collected in the same way as in traditional teaching. This method was chosen because it has been used for many years to evaluate student achievements. For this reason, common instruments were used to evaluate a student's individual achievements. The elements of the evaluation were: the quality of the project documentation (technical drawings, cost sheets, plans, programs, reports); quality of digital content produced and use of ICT (original multimedia materials, texts, spreadsheets, learning and sharing materials, diagrams, charts, program codes), engagement (initiative, organization, quality of interaction with the teacher and other students, taking responsibility and risks, receptiveness, participation in discussions, timeliness), product quality and performance (precision of work, technical correctness, aesthetic appearance, appropriateness of choice of tools and procedures of material processing, compliance of material processing regime), and knowledge and understanding expressed during presentation (terminology, knowledge technologies, presentations, answering questions). In traditional teaching, each student creates their own technical documentations, digital materials, product and presents their work. In project-based learning and teaching, the workload of each student is generally higher than in the traditional one, but the individual results are evaluated here, because otherwise they cannot be compared. Thus, each student, in accordance with the responsibilities of the team plan, creates certain documents, digital materials, make part of a complex product and presents what her / he did individually. This means that the complete project documentation or the quality of the final product as a whole was not evaluated here, but only individual achievements. For each element, students were awarded points ranging from 0 to 100 points during and after the practicum, which were later converted into grades. The grades were awarded in accordance with the current university grading rules, with a grading scale of 1 to 5, with 1 being inadequate, 2 sufficient, 3 good, and 4 very good and 5 excellent. The scores thus obtained for each evaluation element were used for further statistical processing and to determine differences in achievements between groups.

IV. RESULTS

The results of descriptive statistics (Table 3) shows better average scores (M) for the experimental groups than the control one. Nevertheless, the dominant achievement values of all groups are still most often within the average results (*Mod = 3*). Better average scores are more pronounced for engagement than other achievements. Variations in some average scores are greater for control than for experimental groups. Variations are highest for engagement estimates for group A (*CV = 45,42*) and for knowledge and understanding in group B (*CV = 37,71*). Insufficiently expressed differences in average grades by groups and relatively high coefficients of variation are not sufficient to draw conclusions about the success of one or the other approach, despite the pronounced and more uniform achievements of students from the experimental groups. For this reason, a statistical comparison of the results between control and experimental groups was performed. The comparison was performed by t-test for independent samples between groups A and C and groups B and D. Comparison of the results of achievement estimates between control group A and experimental group C (Table 4) shows that the differences were statistically significant only for students' engagement, at the level of statistical significance $p < 0,05$. This result clearly indicates that students' engagement is better in the project-based learning group.

TABLE 3. RESULTS OF DESCRIPTIVE STATISTIC: MEAN (M), DOMINANT VALUE (MOD), STANDARD DEVIATION (SD), COEFFICIENT OF VARIATION (CV)

Group	*Achievements*	*N*	*M*	*Mod*	*SD*	*CV*
A *(contr.)*	Documents	15	2,87	2	0,92	31,94
	ICT and digital materials	15	3,07	3	0,70	22,95
	Engagement	15	2,87	3	1,30	45,42
	Products and performance	15	3,07	3	0,80	26,05
	Knowledge and understanding	15	3,00	3	0,80	26,04
B *(contr.)*	Documents	18	2,72	3	0,83	30,36
	ICT and digital materials	18	2,61	3	0,61	23,27
	Engagement	18	2,77	3	0,94	33,94
	Products and performance	18	3,05	3	0,73	23,74
	Knowledge and understanding	18	2,40	3	0,92	37,71
C *(exp.)*	Documents	14	3,43	3	1,08	31,77
	ICT and digital materials	14	3,57	3	1,09	30,50
	Engagement	14	3,86	4	1,03	26,62
	Products and performance	14	3,21	3	0,97	30,33
	Knowledge and understanding	14	3,21	3	1,12	34,89
D *(exp.)*	Documents	17	3,23	3	0,97	29,98
	ICT and digital materials	17	3,11	3	0,86	27,50
	Engagement	17	3,71	5	1,21	32,72
	Products and performance	17	3,35	3	0,60	18,08
	Knowledge and understanding	17	3,26	3	1,09	33,73

TABLE 4. RESULTS OF THE T-TEST OF COMPARISON OF ACHIEVEMENTS BETWEEN GROUPS A AND C

Achievements	*t*	*df*	*p*
Documentation	-1,508	27	0,143279
ICT and digital materials	-1,493	27	0,147153
Engagement	-2,263	27	0,031873*
Products and performance	-0,447	27	0,658195
Knowledge and understanding	-0,410	27	0,684718

Comparison of the results of achievement estimates between the second control group B and the second experimental group D (Table 5) also reveals statistically significant differences. Here, the differences are statistically significant for student engagement but also and for knowledge and understanding, at a level of statistical significance of $p < 0,05$. Such results reaffirm the importance of project-based learning for student engagement, and indicate an impact on knowledge and understanding.

TABLE 5. RESULTS OF THE T-TEST OF COMPARISON OF ACHIEVEMENTS BETWEEN GROUPS B AND D

Achievements	*t*	*df*	*p*
Documentation	-1,688	33	0,100933
ICT and digital materials	-2,026	33	0,050959
Engagement	-2,536	33	0,016127*
Products and performance	-1,312	33	0,198628
Knowledge and understanding	-2,321	33	0,026623*

Further analysis of individual student achievements reveals certain regularities that are not clear from group analyses. Specifically, individual student achievement (Figures 2 and 3) shows that in both experimental groups there are students whose average achievement is excellent, while in the control groups there are no such students. This finding clearly supports the thesis that project-based learning is a challenge and an opportunity for some students to achieve higher achievement, which is still required today by engineers.

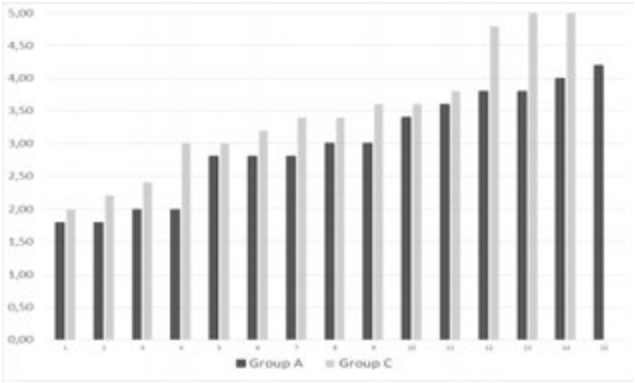


Figure 2. Individual achievements – group A and C

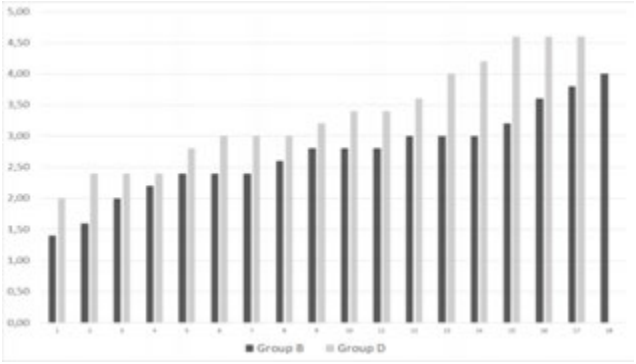


Figure 3. Individual achievements – group B and D

V. DISCUSSION

Aggregate research findings support student achievement during and after implementation of project-based learning, despite the dominance of intermediate grades. The results show statistically significant differences in achievements for student engagement and knowledge and understanding. Student engagement is assessed here on the basis of several important characteristics, such as: initiative, organization, quality of interaction with the teacher and other students, assuming responsibilities and risks, being receptive, participating in discussions, and fulfilling commitments in a timely manner. From the student achievement of these elements and the importance of generic engineering competencies described in the introductory section, it shows the importance of project-based learning for their development. This simply means that important generic competences such as communication skills, teamwork ability, managing one's own learning, and professionalism and inventiveness are much better developing during project-based learning than in traditional teaching and learning. Considering that, these competences are important for successful adaptation in the world of work and for further career development of engineers, the significant role of project and similar modern learning strategies in today's education of engineers is evident. This view is also confirmed by previous research [23], which noted that a group of students' experiences gained through project-based learning were the impetus and starting point for further research and development and were quickly integrated into the world of work.

Although the results of the comparison of assessments between control and experimental groups show the least significance and level for achievement related to the quality of the final product and performance, it should be emphasized that these estimates are somewhat more consistent across the experimental groups. In other words, here comes to the fore peer learning and collaborative work as important features of project-based learning. After all, the realization of the product (prototype) and the concrete use of tools, machines and equipment are just a learning path here, which gives students a sense and meaning of the activity, while the real learning goal is focused on the procedural level of knowledge and general competences they acquire. Therefore, this finding is not so important from the perspective of final grades, but more from the point of view of uniform development, teamwork, communication, inventiveness and problem-solving skills.

An analysis of the individual achievements of students from each group shows that students in traditional teaching do not achieve the excellence required of engineers today. The highest actual individual achievement in the control groups does not exceed a grade 4, although their tasks in traditional teaching are generally much simpler than the tasks of students in project-based learning. In contrast, some individuals in the experimental groups achieve excellent results. This finding reinforces the thesis that students who want and need challenges, and whose abilities enable them, get and use the opportunity for self-realization during project-based learning. For engineering education, this is very important, as the job of an engineer is focused on inventively and creatively solving problems and challenges in the “real world”. Of course, the prerequisite for such achievements is the acceptance of such activities by students, which means that the choice possibilities (here of products) that students will design and realize is a necessary segment of project-based learning. In other words, students should not be forced into what and how they will work, but must be carefully guided in developing the idea and discreetly

moderating their activities while realizing project-based learning. This segment highlights the importance and role of teachers in this process, as well as the complexity of planning, preparing, organizing and implementing project-based learning.

VI. CONCLUSION

The results of the research show some differences in student achievement between traditional and project-based learning and teaching, with higher average student achievements in project-based learning. The differences are particularly noticeable for the achievements related to student engagement, which supports the connection of project-based learning with the development of generic engineering competencies: communication, teamwork, entrepreneurship. The most important differences were observed by analysing individual student achievements, where excellent results being observed only in project-learning groups. This finding indicates that project learning is a challenge and provides an opportunity for those students who want and can do more for their own self-realization. In other words, project-based learning of technology and engineering can help develop the competencies required of engineers today and enable them to adapt faster and more successfully to the "world of work" and further career development far better than today's, largely traditional, engineering education. Here, one can also confirm the thesis that engineering (technical) competences cannot be acquired separately from general and humanistic ones, but must be acquired in an integrated way through simulated engineering practice, such as project-based learning.

Despite these findings, it should be noted that this research was conducted within engineering practicum whose curriculum is of a practical nature. This is somewhat limiting in terms of research results. Namely, the groups observed were not uniform, did not consider any external (environmental) factors or distractors that may have affected students' achievements. In addition, evaluation instruments are intended for traditional teaching and are not adapted to project-based learning. At the same time, some elements of evaluation, such as engagement and knowledge, are largely subject to the subjective evaluation of the teacher. Therefore, future research into the project approach should cover the broader context and overall curriculum of engineering education, not only to gain insight into achievements, but also to find acceptable implementation solutions for future engineering education.

REFERENCES

[1] G. Lipika, A. Pooja, S. Rajnish, C. Madhu, Project based Learning - an Enhanced Approach for Learning in Engineering, Proceedings of the International Conference on Transformations in Engineering Education ICTIEE 2014, Natarajan, R. (Ed.), Springer, 2015 (ISBN 978-81-322-1931-6).

[2] D. Purković, J. Bezjak, Kontekstualni pristup učenju i poučavanju u nastavi temeljnog tehničkog odgoja i obrazovanja, Školski vjesnik, 64(2015),1, pp. 131-152.

[3] D. Kokotsaki, V. Menzies, A. Wiggins, Project-based learning: a review of the literature. Improving schools, 19 (3), 2016, pp. 267-277.

[4] J. W. Thomas, A review of research on project-based learning. San Rafael, CA: Autodesk Foundation, 2000.

[5] M. J. Prince, R. M. Felder, Inductive Teaching and Learning Methods, Definitions and Research Bases, Journal of Engineering Education, 95(2), 2006, pp. 123-138.

[6] A. A. Frost, Synthesis of Knowledge Management Failure Factors, 2014, http://www.knowledge-management-tools.net/failure.html (14.12.2016.)

[7] S. A. Male, Generic Engineering Competences: A Review and Modelling Approach. Education research and Perspectives, 37(1), 2010, pp. 25-51.

[8] F. S. Becker, Globalization, curricula reform and the consequences for engineers working in an international company. European Journal of Engineering Education, 31(3), 2006, pp. 261-272.

[9] P. D. Galloway, The 21st-Century Engineer: A Proposal for Engineering Education Reform. Reston, Virginia: ASCE Press, 2008..

[10] M. Abdulwahed, M. O. Hasna, Engineering and Technology Talent for Innovation and Knowledge-Based Economies: Competencies, Leadership, and a Roadmap for Implementation, Springer International Publishing AG, 2017.

[11] C. Bodmer, A. Leu, L. Mira, H. Rutter, SPINE: Successful Practices in International Engineering Education: Engineers Shape our Future IngCH, 2002, http://www.ruetter-soceco.ch/wp-content/uploads/2016/07/spinereport.pdf (15.09.2019.)

[12] A. T. Kirkpatrick, AC 2011-2804: Vision 2030 creating the future of mechanical engineering education, 2011, https://peer.asee.org/vision-2030-creating-the-future-of-mechanical - engineering- education.pdf (12.10.2019.)

[13] C. S. Nair, A. Patil, P. Mertova, Re-engineering graduate skills - a case study. European Journal of Engineering Education, 34(2), 2009, pp. 131 - 139.

[14] G. Scott, K. W. Yates, Using successful graduates to improve the quality of undergraduate engineering programmes. European Journal of Engineering Education, 27(4), 2002, pp. 363-378.

[15] N. Spinks, N. Silburn, D. Birchall, Educating Engineers for the 21st Century: The Industry View. London: The Royal Academy of Engineering, 2006.

[16] S. A. Male, M. B. Bush, E. S. Chapman, Perceptions of Competency Deficiencies in Engineering Graduates. Australasian Journal of Engineering Education, 16, 2010, pp. 55-68.

[17] A. Johnston, R. King, Addressing the supply and quality of engineering graduates for the new century. Technical Report, 2008, DOI: 10.13140/RG.2.1.2378.4166. (15.10.2019.)

[18] R. Harvey, N. Lemon, Personal factors, education and the effectiveness of professional engineers. European Journal of Engineering Education, 19(3), 1994, pp. 351.

[19] C. L. Newport, D. G. Elms, Effective engineers. International Journal of Engineering Education, 13(5), 1997, pp. 325-332.

[20] OECD, Definition and Selection of Competencies (DeSeCo) Theoretical and Conceptual Foundations: Strategy Paper: Organisation for Economic Co-operation and Development, 2002.

[21] E. F. Crawley, J. Malmqvist, S. Ostund, D. R. Brodeur, K. Edstrom, Rethinking Engineering Education: The CDIO Approach. Springer, Cham, Switzerland, 2014.

[22] D. Kokotsaki, V. Menzies, A. Wiggins, Project-based learning: a review of the literature. Improving schools, 19 (3), 2016, pp. 267-277.

[23] D. Purković, Identifiction of optimal features of the knowledge base in project-based edngineering education-Qualitative analysis of application in engineering practicum. MIPRO 2018 41st International Convention Proceedings, Skala, K. (Ed.), 2018, pp. 656-663.

[24] S. M. Gomez Puente, Design-based learning: exploring an educational approach for engineering education. Eindhoven: Technische Univeriteit Eindhoven, 2014, https://doi.org/10.6.100/IR771111.

[25] J. Bezjak, Project learning of model PUD – BJ- from idea to the product. Klagenfurt: LVM, 2009.

Computer-based question and exam evaluation in summative knowledge assessment

S. Tomić *, V. Paunović*, and I. Bosnić*
* University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
{ sinisa.tomic , vlatka.paunovic, ivana.bosnic }@fer.hr

***Abstract* - This paper describes a system for automatic multiple-choice pen-and-paper exam generation, which includes evaluation of students' answers and evaluation of the quality of exams and questions. During the design phase examiner defines the number of questions, area for each question, and the number of variations of the exam. In the evaluation phase, the system automatically evaluates the students' answers. In the verification phase, the system uses Classical test theory (CTT) concepts to provide feedback to the examiner highlighting all questions identified as too easy or too difficult for students. When using the KR-20 formula to verify more than 100 exams and 5,000 students, results show that the reliability coefficients are higher than 80% on average, which is considered very high.**

***Keywords – e-Learning,* Classical Test Theory, Item Response Theory, automated exam evaluation, automated hybrid exam verification**

I. INTRODUCTION

The increased number of computers and the Internet's availability also increased its usage in the educational environment, which was exceptionally noticeable during the COVID-19 outbreak, where many educational institutions offered their courses online.

Today Learning management systems (LMS) are used not only for teaching but also as part of the examination process [1]. Students' knowledge can be examined using only the computers but also in the hybrid environment where the computer generates and evaluates the exam, but students write the test using the pen-and-paper method [2]. This approach may be more appropriate in scenarios where many students should be examined simultaneously at the same place, and when not enough computers are available.

This work describes a new system for automatic hybrid examination using the pen-and-paper method. The system, based on the examiner's input and question pool, creates one or more versions of the exam prepared to be printed and handed out to students. During the evaluation phase, the system automatically evaluates all answers. In the last - verification phase, the system analyses the answers and generates a detailed report regarding the confidence level of the exam and questions.

The system then calculates the exam confidence level using the Kuder and Richardson Formula 20 (KR-20), which is still in use today [3]. It calculates the confidence level for the exam based only on correct and incorrect answers that students provided. KR-20 is part of the Classical test theory (CTT) that Charles Spearman introduced in 1904 and is still in use today. This theory defines the result of the examination as the sum of correct and incorrect answers [4].

Based on the CTT, the system will highlight to the examiner any questions identified as too easy or too difficult for students. The described process supports the improvement of questions and may increase the quality of the question pool.

This paper analyses the proposed system for automatic hybrid exam generation and evaluation as well as results when using the feedback to the examiner on questions. It also answers the following research questions within the educational context:

RQ1: How to increase the quality of questions in a high-volume question pool?

RQ2: How to automatically create and evaluate exams in a hybrid evaluation environment?

The rest of the paper is structured as follows. Section II presents the related work, focusing on the use of learning management systems in the examination. Section III gives a theoretical background of the Classical Test Theory and Item Response Theory used in this paper. Section IV describes the introduced system. Section V shows the results of the evaluation of the exams and questions. The last section concludes this paper with a discussion and provides future work suggestions.

II. Related work

Learning management systems (LMS) are an integral part of today's learning when using computers in learning or examination. Today, many such systems are available as described by Kasim et al. [5] The authors analyze several modern systems such as Moodle, Sakai, ATutor, Blackboard, SuccessFactor, and SumTotal. They conclude their analysis with Moodle as the recommended system due to the high number of features and more than 80 milion users.

With such many features and users, it may seem that LMSs enable all types of examinations. The work by Drašković et al. [6] shows that even that is not enough and that the examination process, as well as question types, should be modified to be able to create an exam.

Authors such as Sierra et al. [7] analyze LMS security in which students submit their software as an answer to the question. They create a new LMS divided into two parts: one for question generation and the second one, which analyzes and evaluates submissions. They establish communication between these two parts using the *Learning Tools Interoperability* – LTI [8].

Kiraly et al., in their work [9], also suggest ignoring existing LMS's to create new question types not already available in existing systems. The same approach uses Siddiqi et al. in their work [10] while creating the new system just for the evaluation of short essay types of questions. Their work, as well as work of other authors, is then analyzed by Patil et al. [11], who conclude that all existing software is still not good enough to evaluate short essay questions properly.

The way the examiner would like to be able to ask a question is the problem for all listed authors. Without the possibility to ask a question in an intended way may be one of the reasons that not many authors analyzed the questions that they managed to ask. LMSs today, such as Moodle, provide much statistical information about the exam to the examiner. However, this information is not easy to understand for the examiners. It is based just on statistics with no analysis using the methods from *Classical Test Theory* (CTT) or even more advanced methods such as *Item Response Theory (IRT)*.

Azevedo et al. are one of the authors that use the CTT to examine their questions pool [12]. They used external software for the analysis of questions from the Moodle question pool. At the end of the examination, they left only 973 questions out of more than 2,500 in the question pool.

III. Theoretical background

There are two approaches to the assessment of student's knowledge: *summative* and *formative* [13]. The questions used in summative assessment can be verified using: the *Classical test theory* and the *Item response theory*.

A. Summative and formative assessment

Summative assessment assesses the knowledge or skill that the student possesses at the time of evaluation. Examples of summative assessments are multiple-choice questions, short answer, matching, true-false questions. At the end of the summative assessment, the student receives feedback about the results, which may include correct answers or steps needed to solve the given task. This evaluation may be performed several times during the semester, and final grades may be the result of several such assessments.

Formative assessment is a more complex type of evaluation. It is the result of continuous observation of student's work, behavior, and attitude. The student's knowledge can then be evaluated on the expected retention duration and not just based on factual knowledge. The main goal of this type of assessment is to provide the student with guidance and to include feedback on improving the knowledge.

Dixon et al. [14] evaluate both types of assessments. They also define types of formative assessment as: planned and spontaneous. Planned formative assessment occurs when the student is solving the task by themselves. Spontaneous is the one in which the examiner receives feedback about knowledge of the student in real-time by:

a) observing the student's behaviour during the examination;
b) asking the student to further explain the answer by giving examples of the described concept;
c) using the short quiz during the lecture.

In their book, Parke et al. [15] describe the distinction between summative and formative assessment. They propose to use the formative assessment to find out how successful students are and how successful is the teacher or examiner.

Wieman et al. [16] give another example of collaborative learning, after which the formative assessment is used. They divide the examination into two parts. During the first one, students have a written exam. In the second one, they group students into groups of three during where the goal of each group is to comment on the questions and answers from the first part. Each team must agree on their final answer for the first part. The final grade is the result of the score in the first and the second part. The first part, in

which they worked independently, has more influence on the final grade.

B. Classical Test Theory (CTT)

While grading student's work is an integral part of the examination, it is also essential to be able to verify if the questions asked during the examination were adequately selected. In 1904 Charles Spearman introduced the Classical Test Theory [17], which is in use even today. It enables the examiner to anticipate results of examination for which it uses the parameters such as the ability of students and difficulty of the task.

CTT is defined as (1) [4]:

$$X = T + E \tag{1}$$

X represents the observed score of the examination, T is the true score, and E represents the error. Both T and E are not observable directly.

To calculate the reliability coefficient of examination following formula can be used as in (2):

$$P^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} \tag{2}$$

It is the ratio of the variance of the correct result and the variance of the response.

Most prominent formula to establish the reliability of examination is the Cronbach's alpha or coefficient alpha [18] as in (3):

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} \sigma_{Y_i}^2}{\sigma_x^2}\right) \tag{3}$$

In which K is the number of questions, and σ_i^2 represents the variance of X_i and σ_x^2 variance of X.

As shown in the formula, the number of question K has a great impact on the result. With the increase in the number of questions, the α increases too.

Cronbach does not define the exact value of α. However, there are suggestions made by Nunnally [19]. He states that 0.7 or more is valid during the early research phase; 0.8 during the applied research phase and is 0.9 or higher when making an important decision. When the decision is crucial, the value should be 0.95 or higher; when the number is one, it may denote that the number of questions is too large.

A different version of the formula is the Kuder and Richardson Formula 20 (KR-20) [3], which is in use still today. When examination uses grading using the value 1 for correct and 0 for an incorrect answer, it can also use the KR-20 to calculate its reliability. Formula KR-20 is defined as (4):

$$KR_{20} = \frac{k}{k-1} * \left(1 - \frac{\sum_{i=1}^{K}(pq)_i}{V_y}\right) \tag{4}$$

C. Item Response Theory (IRT)

All described methods can be used to evaluate the complete exam. They cannot be used to evaluate the reliability or quality of a single question in it, which some authors consider an issue [20]. Evaluation of a single question in an examination can be performed using the Item Response Theory (IRT). This theory evaluates each question, or item, as the probability of examinee with appropriate abilities to answer it. There are several validation methods out of which the four parameters model (4-PL) is the most complex one and can be used in multiple-choice questions. This model presumes the student's ability, the difficulty of the question, and the probability that a student who does not know the correct answer will still try to guess it, and that student who knows the answer will still not choose it.

IV. Application of CTT and IRT in Student's Automated Examination Process

As described in this work, several authors struggle with existing software to create a proper exam [6]–[9]. One approach to the problem is to have the software create an exam familiar to the students. Creating an existing type of exam is the approach of the prototype system designed at the University of Zagreb, Faculty of Electrical Engineering and Computing.

The prototype system, written in PHP, as a web application, enables semi-automatic written pen-and-paper test generation with little to none intervention by examiner. The exam is generated from the question pool when the examiner selects the number of questions from each pool. The system then, based on the frequency of previously asked questions, creates a new exam. The examiner can choose to replace the question, but it cannot choose the exact one. The exam has several variations by shuffling questions and answers. Analysis of answers is performed using the KR-20 for the exam. The software analyses the questions in the exam using the ideas based on IRT.

This software prototype generated more than 100 exams. It examined more than 5,000 students using the semi-automatic exam generation Students wrote the exams using the pen-and-paper. They were not aware that the system semi-automatically generated and validated the exam.

The system consists of four main parts, as described in Fig. 1:

1.) Question pool with grouped questions
2.) Examination configuration
3.) Examination generation algorithm
4.) Evaluation and verification algorithm

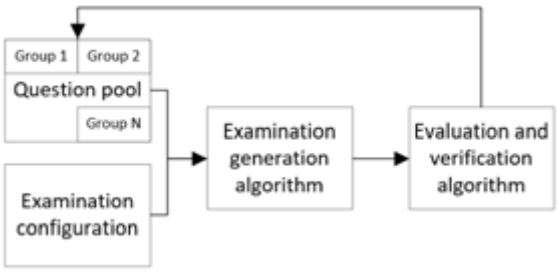


Figure 1 – Structure of the software

When the algorithm generates a new exam, it can be downloaded by the examiner as a list of PDF files: question sheet, answer sheet, and sheet with correct answers. After the students turn in their answer sheets, they are retrieved using the OMR, performed by commercial software Remark, which outputs students' answers as a CSV file, which we used as input for the software. These answers are then further analyzed in the software, and the analysis is stored in the question pool, improving its quality. The result of this process is two reports: exam report, and question report, both available as an HTML page.

A. Exam report

The Exam report, as shown in Fig. 2, is the result of the analysis of the complete exam. The software will first convert all these variations into the base question and answer list, if the students wrote the exam that had several instanced generated and the order of questions and answers, were shuffled.



Figure 2 – Example of exam analysis

Next, it will perform a statistical analysis of the exam. It produces statistical data about the exam. In the example from Fig. 2, the report shows the maximum number of points that students achieved (26 out of 30), the average number of points (16.05), the number of students attending the exam (128). The report also includes the score distribution chart for the exam.

In this scenario, all question scoring rules are 1 point for correct and 0 for an incorrect answer, which enables the analysis of the exam using the CTT, more explicitly using the KR-20 formula. This formula verifies the validity of the exam with the reliability of 0.84189, which is considered very high.

The CTT-based analysis shows the validity of the exam as a whole, but it does not give additional information about questions presented to the students.

B. Question report

The system generates an additional Question report to improve the question pool. It uses the ideas from IRT to give an analysis of each question in the exam. Fig. 3 shows the analysis of the first several questions. Each row represents one question. The correct answer is marked with green color and listed in the first column. The number of correct, incorrect, and unanswered questions is given in the third, fourth, and fifth columns.

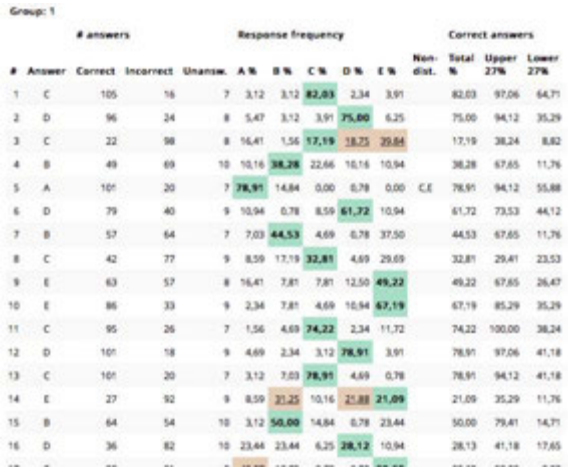


Figure 3 – Example of exam analysis

The central part of this report is the frequency analysis. It displays the percentage of answers for each question. It is essential feedback for the examiner, and the system highlights the questions that need further examination. There are three questions in this example that the examiner needs to review further.

In this example, third and 14th questions can be considered the difficult ones. The reason is that these questions are having more invalid answers selected than the correct one. There can be several reasons for this. It can be the result of the invalid question or invalid responses that are too similar to the valid one. It is also possible that this is the result of students not understanding the presented materials. The examiner

should reevaluate these two questions to make them more understandable.

Another example is the fifth question, which has two non-distractors. The non-distractor is the answer that no student has selected. In the case of this question, the student has to choose among just three answers since two of five suggested are too implausible and thus disregarded. To resolve this issue, the examiner should reevaluate and change the suggested answers.

It is essential to eliminate these types of questions in an automatically generated exam. If there are two groups of students in which group receives only questions like the third one and the other group gets only questions like the fifth one, they would not have the exam of the same difficulty.

V. ANALYSIS

Evaluation of the exams that system semi-automatically generates uses two parameters: the number of students attending the exam and reliability of the exam using the KR-20 formula. We selected only exams with more than 50 students attending for evaluation to have an adequate sample size. We then calculated KR-20 reliability for these exams.

As shown in Table 1, the reliability of exams remained similar regardless of the number of students attending them. For all exams, the reliability was higher than 0.7, and most of them had it between 0.8 and 0.9, which shows that automatically generating an exam, using the question pool, is possible.

TABLE I. DETAILED RESULTS FOR THE LAST 10 EXAMS

#	# Stud	KR-20	Max	Avg	Med	Std dev
1	101	89.38%	28	17.32	18	6.65
2	60	79.12%	24	16.32	16	4.83
3	123	83.88%	26	18.26	20	5.49
4	128	84.19%	26	16.05	17	5.62
5	56	91.55%	26	15.00	17	7.30
6	61	81.21%	24	16.41	17	5.22
7	81	82.12%	25	16.32	17	5.40
8	80	77.14%	24	14.80	15	4.87
9	118	85.96%	26	16.53	17	6.02
10	105	88.41%	30	19.94	21	6.19

All questions in these exams were analyzed previously using the CTT methods, as shown before. All invalid questions were then corrected as much as possible. Exams generated had a high-volume question pool so that each question occurs only once in three years, making it very hard for students to know all questions and answers before the exam.

VI. DISCUSSION

A. The summary of findings

After implementing and evaluating the software, we propose the following answers to our research questions:

RQ1: How to increase the quality of questions in a high-volume question pool?

This research shows that it is possible to increase the quality of questions in a high-volume question pool using the appropriate feedback to the examiner. While KR-20 gives overall information about the quality of the exam it does not provide enough data about the questions themselves. The proposed system uses ideas based on IRT to highlight to the examiner only the questions that should be revised.

This approach can significantly reduce the amount of additional question verification. In our example, only three out of 20 questions required additional validation.

RQ2: How to automatically create and evaluate exams in a hybrid evaluation environment?

This research shows that it is possible to evaluate exams in a hybrid evaluation environment when the questions asked are multiple-choice questions created using a higher volume question pool. We use two exam evaluation methods: KR-20 to verify the reliability of the exam and methods from the IRT to verify the validity of each question included in the exam. When we checked the reliability of the semi-automatically generated exams using the KR-20, they showed, on average, 80% reliability, which is considered very high.

B. Limitation

One of the limitations of this research is that it was performed on only one course and may show different results for other environments. The exam used multiple-choice questions before introducing this system, and the scoring was +1 for correct and 0 for an incorrect answer. The KR-20 may not be as appropriate for reliability testing when using the different grading or kind of questions.

C. Suggestion for further research

Further research will focus on the inclusion of multiple IRT methods in the question analysis, mainly using the 4-PL model to improve the high-volume question pool further and include new questions in the exam.

VII. CONCLUSION

The proposed system for automatic hybrid exam generation and evaluation showed promising results when verified using the KR-20 formula. The students

who attended these exams were not aware that these tests were semi-automatically generated and evaluated, had no issues with the exam, or included questions, thus proving that it is possible to generate valid exams.

Verified, high volume, question pool is the prerequisite for such exam generation. Using the positive feedback to the examiner by highlighting only the questions that students had issues with, such as too easy or too difficult questions, we showed that it is possible to increase the quality of the high-volume question pool highly. Changing less than 5% of the questions in the pool in our case resulted in the exam reliability higher than 80% on average, which is considered very high.

REFERENCES

[1] M. Ueno and Y. Miyazawa, "IRT-Based Adaptive Hints to Scaffold Learning in Programming," *IEEE Trans. Learn. Technol.*, vol. 11, no. 4, pp. 415–428, Oct. 2018, doi: 10.1109/TLT.2017.2741960.

[2] Y. Kataoka, A. H. Thamrin, J. Murai, and K. Kataoka, "Effective use of learning management system for large-scale Japanese language education," in *Proceedings of the 10th International Conference on Education Technology and Computers - ICETC '18*, 2018, pp. 49–56, doi: 10.1145/3290511.3290564.

[3] G. F. Kuder and M. W. Richardson, "The theory of the estimation of test reliability," *Psychometrika*, 1937, doi: 10.1007/BF02288391.

[4] R. Maclean *et al.*, Eds., *Applied Rasch Measurement: A Book of Exemplars*. Dordrecht: Springer Netherlands, 2005.

[5] N. N. Mohd Kasim and F. Khalid, "Choosing the Right Learning Management System (LMS) for the Higher Education Institution Context: A Systematic Review," *Int. J. Emerg. Technol. Learn.*, vol. 11, no. 06, p. 55, Jun. 2016, doi: 10.3991/ijet.v11i06.5644.

[6] D. Draskovic, M. Misic, and Z. Stanisavljevic, "Transition from traditional to LMS supported examining: A case study in computer engineering," *Comput. Appl. Eng. Educ.*, vol. 24, no. 5, pp. 775–786, Sep. 2016, doi: 10.1002/cae.21750.

[7] A. J. Sierra, Á. Martín-Rodríguez, T. Ariza, J. Muñoz-Calle, and F. J. Fernández-Jiménez, "LTI for Interoperating e-Assessment Tools with LMS," in *Advances in Intelligent Systems and Computing*, 2016, pp. 173–181.

[8] Educause, "Learning Tools Interoperability," 2013.

[9] S. Király, T. Balla, and G. Kusper, "Code School: Learn to code by practicing," in *CEUR Workshop Proceedings*, 2016.

[10] R. Siddiqi, C. J. Harrison, and R. Siddiqi, "Improving teaching and learning through automated short-answer marking," *IEEE Trans. Learn. Technol.*, 2010, doi: 10.1109/TLT.2010.4.

[11] R. G. Patil and S. Z. Ali, "Approaches For Automation In Assisting Evaluator For Grading Of Answer Scripts: A Survey," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 2018, pp. 1–6, doi: 10.1109/CCAA.2018.8777664.

[12] J. M. Azevedo, E. P. Oliveira, and P. D. Beites, "Using Learning Analytics to evaluate the quality of multiple-choice questions," *Int. J. Inf. Learn. Technol.*, vol. 36, no. 4, pp. 322–341, Aug. 2019, doi: 10.1108/IJILT-02-2019-0023.

[13] B. B. Frey, "Standards for Educational and Psychological Testing," in *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*, 2455 Teller Road, Thousand Oaks, California 91320: SAGE Publications, Inc., 2018.

[14] D. D. Dixson and F. C. Worrell, "Formative and Summative Assessment in the Classroom," *Theory Pract.*, 2016, doi: 10.1080/00405841.2016.1148989.

[15] J. Parkes and D. Zimmaro, "Formative and Summative Assessments," in *The College Classroom Assessment Compendium*, Routledge, 2017, pp. 89–91.

[16] C. E. Wieman, G. W. Rieger, and C. E. Heiner, "Physics Exams that Promote Collaborative Learning," *Phys. Teach.*, vol. 52, no. 1, pp. 51–53, Jan. 2014, doi: 10.1119/1.4849159.

[17] C. Spearman, "'General Intelligence,' Objectively Determined and Measured," *Am. J. Psychol.*, 1904, doi: 10.2307/1412107.

[18] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, 1951, doi: 10.1007/BF02310555.

[19] J. Nunnally and I. Bernstein, "Psychometric Theory, 3rd edn, 1994," *McGraw-Hill, New York*, 1994.

[20] M. Stenbeck, R. K. Hambleton, H. Swaminathan, and H. J. Rogers, "Fundamentals of Item Response Theory.," *Contemp. Sociol.*, 1992, doi: 10.2307/2075521.

Students' perception of summative peer review grading

L. Zrnić, L. Korov, J. Petrović* and P. Pale*
* Department of Electronic Systems and Information Processing
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, HR-10000 Zagreb, Croatia
{ Leon.Zrnic , Luka.Korov , Juraj.Petrovic , Predrag.Pale }@FER.hr

***Abstract* - Peer review is the process of evaluating scientific, academic, or professional work by others working in the same field. It has been used and researched in formal education for dozens of years. Still, the results of its application vary greatly depending on the usage context. This paper describes the context, students' perception, and outcomes of using peer review in summative assessments in a first-semester communication skills course for engineering students. Students' opinions were investigated using two surveys and analyzed together with the credit they achieved in peer review assignments. The obtained results describe students' responses to peer review and indicate specific aspects of peer review that need to be taken into consideration when implementing it in higher education.**

Keywords – peer review; peer assessment; summative assessment; scalability

I. INTRODUCTION

Peer review, or peer-assessment, is „the process of someone reading, checking, and giving his or her opinion about something that has been written by another scientist or expert working in the same subject area, or a piece of work in which this is done" [Cambridge dictionary; peer review]. A similar definition by Topping [1] states that peer-assessment is "an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status". In the last few decades, peer review has found a plethora of uses in higher education [2] and is widely used and researched today since it offers a scalable method of knowledge assessment.

This paper reports the results of using peer review in an mandatory course in the first semester on the Faculty of Electrical Engineering and Computing at the University of Zagreb. This course has been using peer review for several years as a way to grade students' work with the goal of creating scalability in grading, but also as a way of achieving course learning outcomes. Peer review was introduced to the course to help students learn to critically assess others' work, to gain insight into the way that other students in their generation solve similar problems. Students' opinions and perception of grades that they were given as well as the success of students in integrating feedback that they received were investigated with a questionnaire that was carried out twice during the semester.

The rest of this paper is organized as follows. In chapter II related research is identified and summarized, as well as the goals of the research described in this paper. In chapter III the materials and methods are listed and explained. Results are presented and discussed in chapter IV, and chapter V contains conclusions and suggestions for future work.

II. RELATED RESEARCH

Peer review is today mostly implemented in higher education [3], [4]. A possible reason for that might be the maturity and competence which higher-education students can offer. Peer review can be used as formative assessment, meaning its purpose is just to provide feedback to students on their assignments, but this research is based on summative assessment, the form of peer review that contributes to final grade [5]. Regardless of the differences between peer reviewing forms, peer review is undoubtingly a powerful tool for educational purposes, offering flexibility, reliability, and scalability [6], [7].

Even though there are dozens of different areas college and university students can specialize in, peer review researches' results are often similar. For example, peer assessment study between undergraduate students in process metallurgy discipline [8] concluded that factors like low level of trust among peers, anxiety over year marks and discomfort due to perceived incompetency when compared to their peers play a significant role in peer review's efficiency and results. Those same factors were identified in a study about peer review in language proficiency among first-year undergraduates on nonlanguage programs [9].

Since peer review outcomes are so dependent on educational and socio-cultural settings, the answer to the question of how to incentivize high-quality peer reviews is still open [10] and its answers can't be generalized across different settings. Hence, the aim of this research is to collect and interpret the data related to peer review procedures in a first-semester communication skills university course at a technical faculty. That data, describing students' perception of peer review, especially grading in peer review, could be helpful to adjust the peer reviewing practice. Using those adjustments, peer review could still be used to achieve scalability in grading and target learning outcomes in a manner that students find fair and helpful.

In this research, students' perspective, their opinions, and positions on peer review were therefore investigated with the aim of finding support for against the following hypotheses which could have a strong impact on students' participation in peer review:

- H1: Students overestimate the quality of their own work and perceive the grade that they received through peer review as unfairly low
- H2: Students who feel they received unexpectedly low or unexpectedly high assignment credit will blame it on their peers' subjectiveness in peer reviewing
- H3: Students will mostly ignore their peer's feedback since it doesn't come from a course teacher or another intellectual authority
- H4: On average, students will tend to give a higher assignment credit than they think a submitted assignment really deserves

III. MATERIALS AND METHODS

A. Course context

In this research, peer review was used as a summative assessment method for a total of six assignments which were part of the mandatory *Communication skills for engineers* course at the University of Zagreb, Faculty of Electrical Engineering and Computing. It is a humanistic course, and one such course is obligatory in every semester of the computer engineering or electrical engineering study program. The course corresponds to 4 ECTS (*European Credit Transfer System*) points, which means students are expected to invest a total of between 100 and 120 hours of work in the course. Around 650 students enroll in the course every year.

B. Peer review procedures

Peer review assessments have been used in this course for several years already, every year with slight changes with the goals of making the process simpler and grades more valid and reliable. The six peer review assignments are divided into homework assignments (five assignments together worth 19.5% of total course credit), and one final course assignment (worth 30% of total course credit). Each of those assignments' credit was divided in ratio of approximately 65% of the assignment credit awarded for the submitted work, and approximately 35% of the assignment credit. Students could receive a maximum of 12.8% of the overall course credit for reviewing their peers' assignments. Participation in the peer review was, however, optional and students who chose not to participate in peer reviewing could have still achieved the highest grade or a maximum of 97.2% of the course credit by completing other optional assignments instead.

The peer review assignments used in the course include writing/capturing: a formal email message on an individualized scenario, a MS PowerPoint (PPT) presentation including narration, a personal narrative *curriculum vitae* (CV) as well as a motivational letter for a job application, a technically valid photograph and a video sequence. The final course assignment was to create a two-minute video presentation of a topic a student finds interesting and would like to work on during his/her studies.

The peer review process was anonymous, meaning a reviewer didn't know whose assignment he was reviewing and vice versa. This was not true only for the case of the final course assignment where the name of the author was written in the introductory part of the assignment. Submitting a work for peer review was a precondition for reviewing others' work and the review criteria were predefined by course lecturers and they were made available to students before creating each work. The same criteria were used for reviewing submitted assignments.

Some criteria were binary (for example, weather a PowerPoint presentation contains narration) while other were measured on a scale with three to five values. Naturally, some reviewers had a more or a less strict stance on some aspects of the works they reviewed, which would lead to a grade difference with other reviewers. Therefore, reviewers could leave comments that would defend their stance or perhaps try and leave helpful suggestions for the student whose work was assessed.

Since every peer review activity contributes to the final course percentage, several control mechanisms were implemented to ensure validity of the grades students would get. About 5-10% randomly selected assignments submitted for peer review were assessed by course lecturers themselves, making students sure that the grades they got were objective. Additionally, about 10% of submissions that were previously graded were included in the assignments sent to students for review, being a baseline for their grading standards. To help ensure validity, some properties of submitted works were additionally assessed automatically using custom software support. For example, the existence of subtitles in a video file, or narration in a PowerPoint presentation. Every submitted assignment was assigned to five students for review. If at least four students submitted a review for that assignment, the final score would be calculated by discarding the highest and the lowest reviewer score and calculating the average of the remaining scores. If less than four reviews were submitted for an assignment, the assignment would be graded directly by course lecturers.

C. Questionnare

To learn how students perceive peer review and their success in it, a questionnaire was given to the students at their midterm exams and before the final exams. The questionnaire was not anonymous, yet students were assured that they would get submission credit regardless of their answers. In the questionnaire students were first asked to remember or to check the grades that they received for each of the peer review assignments, and then to answer questions regarding the difference between the assignment credit they expected and the assignment credit they got. The scale was from -2 to 2, with -2 corresponding to *My grade is at least 10% lower than I thought it would be*, 0 corresponding to *My grade is within a 5% difference of*

what I thought it would be and 2 corresponding to *My grade is at least 10% higher than I thought it would be*. If students answered anything except 0, they were offered a question in which they were given a chance to explain why they believe this had happened. Afterwards, students answered whether they believed they were objective in their reviews and also if they find the comments they received from their reviewers helpful.

IV. RESULTS AND DISCUSION

Distribution of the answers regarding students' difference between expected and obtained assignment credit is shown in TABLE I. According to the data, it seems about two thirds of students received the assignment credit they were expecting (*My grade is within a 5% difference of what I thought it would be.*). Even though it is possible that some students picked this answer knowing that in that case they won't need to write additional comments or opinions, it seems that a majority of students did not feel they got an unfair amount of assignment credit from peer review and/or that they understand reasons for the grades that they got. Even though this does not imply that the assignment credit they got in peer review would be the same as the assignment credit they would get from the course lecturer, it is a potential descriptor of their dissatisfaction with the fact that they have been graded by their peers.

TABLE I. DISTRIBUTION OF ANSWERS

Assignment	Difference between received and expected assignment credit				
	-2	-1	0	1	2
CV (N=453)	12.14%	11.70%	64.02%	6.18%	5.96%
PPT (N=496)	4.64%	8.47%	67.54%	9.88%	9.48%
Mail (N=447)	15.21%	14.32%	60.40%	7.38%	2.68%
Photo (N=465)	8.82%	11.83%	62.80%	10.54%	6.02%
Video (N=446)	8.07%	11.8%	58.97%	12.56%	8.52%
Project (N=426)	8.69%	9.39%	64.32%	8.69%	8.92%

Interestingly, the percentage of students who think they received a lower or higher credit than they should have received varies significantly. In one case, almost 20% students receiving a higher assignment credit than they expected for the PPT peer review task, while, in another case, almost 30% of students receiving a lower assignment credit than they expected for the mail peer review task. A possible explanation is that either the expected assignment properties were not clearly communicated, or that students feel more confident about some assignments than others, leading them to overestimate their work. Those data do not support the H1 hypothesis, as it seems around two thirds of students are typically satisfied with their peers' grading.

The answers presented in the TABLE I were used to calculate Pearson's correlation coefficients between the assignment credit students were given and their expectations. Correlation coefficients were calculated for every peer review task this way, but they were calculated separately using only answers from students that were graded by course lecturers and using only answers of students who were graded only by their peers. Correlation coefficients calculated from questionnaire responses are shown in the TABLE II. Calculated correlation coefficients indicate that there is a weak correlation between the students' assignment credit and their expectation, except for PPT and project assignments when it was not graded by the course lecturer. Reason for the exception is probably combination of freedom that PPT and project tasks offered and subjectivism of the reviewers. As indicated by the results in TABLE II, it seems there is a small correlation between students' expectations and the awarded assignment credit when they know that they have been graded by the lecturer. This correlation ranges from small positive to small negative correlation in case students are graded by their peers, dependent on the peer review assignment.

TABLE II. CORRELATION COEFFICIENTS

Assignment	Not graded by course professor	Graded by course professor
CV	0.215	0.218
PPT	0.027	0.191
mail	-0.153	0.148
photo	-0.143	0.189
video	0.149	0.140
project	0.044	0.279

The comments that some students took time to write in the survey were read and categorized. Those who said they got a higher or a lower grade then they expected were further asked why they think that was the case. The answers to that question are shown in the TABLE III.

Based on the results in TABLE III, it seems most students blame themselves when asked about the difference between the assignment credit they got and the assignment credit they expected. Both if they got a lower or higher assignment credit than expected, being too self-critical or not self-critical, and making a mistake in preparing the assignment are most common explanations selected by between 49% and 63% of students. A slightly lower percentage of students, between 32% and 48%, blames the difference between the awarded and expected assignment credit on their reviewers. The reviewers, according to students, are too subjective or incompetent to assess their work. This result partially supports the H2 hypothesis. It seems at around one third of students could believe they have been badly graded by their reviewers. This is an important result and it means additional mechanisms must be implemented to enable students to get another evaluation, possibly by the course lecturer in that case.

TABLE III. REASONS FOR MISJUDGING THE GRADE

Explanation	When getting a higher grade than expected	When getting a lower grade than expected
I am too self-critical.	55.21%	/
I am not self-critical enough.	/	15.63%
I made a mistake doing the assignment.	8.33%	33.75%
Reviewers are too subjective.	26.04%	30.00%
Reviewers are not competent enough to grade others.	6.25%	17.50%
I do not know.	4.17%	3.12%

Students were also asked in the survey how successful they think they are in integrating feedback to their submitted assignments that other students provide in peer review. This feedback can be helpful, but if it comes from another student who misunderstood the assignment it can also be confusing. It can also cause a student to feel that he or she got a lower grade because of an incompetent reviewer. In the answer to the survey question related to integrating peer reviewers' feedback, 5.86% of students said that they find it difficult to distinguish between the comments of their peer reviewers that they should or should not take into consideration to improve their work. Some students said they ignore the feedback they receive (3.35%), while the majority said they can mostly (51.76%) or completely (39.03%) distinguish between the comments that they should or should not take into consideration to improve their work. This result suggests the comments that students receive through peer review are helpful and it provides support against the H3 hypothesis. Based on this result, it seems that peer reviewers' comments and feedback should be available with the peer review assignment credit even though it is on the student to recognize which of them are valid and valuable feedback.

The last two question in the questionnaire investigated the objectiveness of students' reviewing and grading. In the first question, explaining how students grade their peers, 92.81% of students said they grade their peers as objectively as they can, 6.37% said they tend to give slightly higher grades than the submitted assignment deserves, and only one student said he tends to give lower grades than he thinks the submitted assignment deserves. Students who said they tend to give their peers slightly better grades than they think their assignments deserve were further asked to explain their choices with textual comments. The most common explanations were that they do not feel comfortable to express their real opinion, they hope they will get a better grade by others in return, they feel good about that, or because they value the perceived effort more than its results. It seems a large majority of students attempts to grade peer review assignments as objectively as they can. This result does not support H4 hypothesis. Still, social desirability bias could have affected students to select this answer, so this result needs to be considered carefully.

V. CONCLUSIONS

In this paper, the data related to students' perception and behavior in summative peer review assessments is reported and analyzed. While educational practices should be based on educational goals, teaching methodologies based on scientific evidence, and well-developed curricula, students' opinions are still important in this process. They can help to adjust educational practices to make them more stimulating and enjoyable, or at least less unenjoyable.

Overall, looking at the data gathered and analyzed in this paper, it seems that peer review is an efficient and scalable way of grading students' assignments. Apart from its efficiency and scalability, it also teaches student to critical think about both their and others' work. Still, the key findings of this research suggest students are generally satisfied with peer grading and a lot of them will first blame themselves when they are not. This is important, since the possibility that students express general dissatisfaction with being graded by their peers could hinder the usage of peer review in the course. Still, about one third of students had in some instance objections to their graders, and mechanisms should be implemented to help resolve the related issues. Also, it seems comments that students receive through peer review are useful for a majority of students, and, based on self-reported results, a great majority of students try to be as objective as possible in peer review grading.

The authors hope those results will be of help to other researchers and practitioners. In their future work the authors hope to focus on the comparison between assignment credit as awarded by students in peer review, and awarded by course lecturers, as well as mechanisms and procedures that can be used to ensure their similarity.

REFERENCES

[1] K. Topping, "Peer assessment between students in colleges and universities," Review of Educational Research, vol. 68, no. 3, pp. 249 , 1998.

[2] C. Morgan, and M. O'Reilly, "Assessing open and distance learners," Psychology Press, 1999.

[3] N. Falchikov and J. Goldfinch, "Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks", Review of Educational Research, vol. 70, no. 3, pp. 287-322, 2000.

[4] Zulfaneti, S. Edriati, and Mukhni, "Enhancing students' critical thinking skills through critical thinking assessment in calculus course," J. Phys.: Conf. Ser. 948 012031, 2018.

[5] W. Harlen, and M. James, "Assessment and learning: differences and relationships between formative and summative assessment," Assessment in education: Principles, policy & practice 4 (3): 365–379, 1997.

[6] R. Arruabarrena , A. Sánchez, J. M. Blanco, J. A. Vadillo, and I. Usandizaga, "Integration of good practices of active methodologies with the reuse of student-generated content," International Journal of Educational Technology in Higher Education, 2019.

[7] C. Kulkarni, K. Pang Wei, Huyle, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer, „Peer and self assessment in

massive online classes,“ ACM Trans. Comput.-Hum. Interact. 20, 6, no. 33, December 2013.

[8] E. Matinde, “Students’ perceptions on reciprocal peer tutorial assessment in an undergraduate course in process metallurgy,” Educ. Sci., 9, 27, January 2019.

[9] W. Cheng, and M. Warren, “Peer assessment of language proficiency,” Language testing 22 (1): 93–121, 2005.

[10] A. A. Hope, and C. L. Munro, “Criticism and judgment: a critical look at scientific peer review,” American Journal of Critical Care, vol. 28, no. 4, July 2019.

Assessing students' SQL knowledge and skills in gamification manner

G. Tuparov*, and D. Keremedchiev*
* New Bulgarian University / Department of Informatics, Sofia, Bulgaria
gtuparov@nbu.bg

Abstract - **The interest of educators in the game based education and gamification in university educational process grows in the last years. The paper discusses some approaches towards assessing students' SQL knowledge and skills and the application of some techniques to prevent cheating during the assessment process. In addition, assessment tools integrated in e-learning environments to test students' SQL skills are analyzed. An approach to students' self-assessment in gamification manner is presented. To support this approach a system for automated generation of the test questions in QTI format for e-learning environments is presented.**

Keywords - component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Contemporary society is more and more influenced by Information and communication technologies (ICT), and this is a stable trend. Nowadays many students want to obtain knowledge and skills in ICT as a primary or complementary part of their academic study, and in this way to have a successful professional career. Moreover, according to ACM Curricula Recommendations [1], the comprehensive database course including SQL is required in computer science, software engineering, and information systems majors. In other majors with some influence of ICT, a database course with basic knowledge of SQL is proposed as elective or some topics in this field are included in other courses. Unfortunately, some students simply follow this ICT trend, instead of being deeply involved in the field, and university lecturers are challenged to teach effectively such kind of students.

Furthermore, the ICT influence in everyday live changes the personal habits in communication, perceptions and learning preferences of nowadays students, as they are deeply involved in virtual worlds of gaming and Internet. They prefer to be online and access lectures and other learning materials in the same way they access social networks or play games through their mobile devices.

ICT innovations rapidly influence higher educational institutions. A great number of universities provide e-learning services for students in different modes – from simple support for face-to-face classes, through blended learning mode to pure distance learning programs. Game-based learning and gamification have also become popular at university level education. Unfortunately, the development of interesting serious games for university students is not so easy and one very popular alternative is to implement some game elements in teaching, i.e. apply gamification, preferably through e-learning environment.

The aim of this paper is to present an approach for gamified self-assessment of students' SQL knowledge and skills developed with respect to new students' learning profile. The paper is organized as follows.

II. RELATED STUDIES

A. Gamification in higher education

The gamification is "an example of a modern teaching method that can be used in any sphere of human activity, and is applicable to each age group" [2]. In the last 5 years, studies and papers have related to the problems of gamification in higher education growth in scientific database such as SCOPUS, Science Direct, ACM and IEE. In Table I. some statistical data are presented. The data was collected on 26.01.2020.

The documents with key words "gamification + higher education" are about 10% -11% of all documents with a key word gamification. For the last five years the

TABLE I. DOCUMENTS FOUND IN SCIENTIFIC DATABASE WITH SPECIFIC KEY WORDS

Database	SCOPUS		Science Direct		ACM		IEEE	
Key words	All	2015-2020	All	2015-2020	All	2015-2020	All	2015-2020
gamification	6070	5195	1965	1812	2526	1950	1098	904
gamification + higher education	627	569	253	231	174	143	42	37
gamification + university education	18	14	21	20	15	12	3	2
gamification + university education +SQL	10	10	1	1	3	3	0	0
gamification + higher education+ SQL	0	0	4	4	10	10	0	0
teach SQL	10	7	4	3	13	9	1	1
teach SQL + gamification	1	1*	1	1	0	0	0	0

The study is funded by Bulgarian National Scientific Research Fund, Contract DN-05/10, 2016

documents on the topic "gamification + higher education" increase up to about 90% of all documents on the topic.

This topic sets challenges in front of university lecturers and students [3]. It is an innovative approach for education at all educational levels [4].

The studies are directed to the implementation of the gamification approach in different educational areas: computer science, medical and nursing education, entrepreneurship, teacher education etc. In the field of computer science the focus of the studies is on programming and software engineering. A few papers discuss problems with gamification in Database courses, topics related to SQL (Figure 1 and Figure 2). Less than 0,67% of all papers related to gamification in the mentioned databases discuss gamification in higher education Database courses, topic SQL.

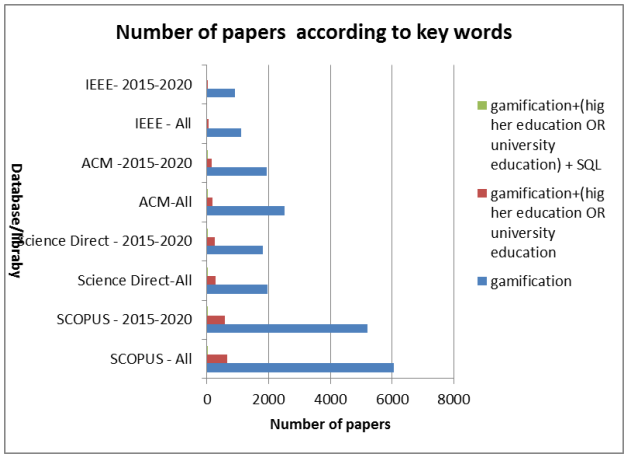


Figure 1. Number of papers related to key words "gamification", "gamification+(higher education OR university education)", "gamification+(higher education OR university education) + SQL" in SCOPUS, ACM digital library, Science Direct, IEEE Xplore

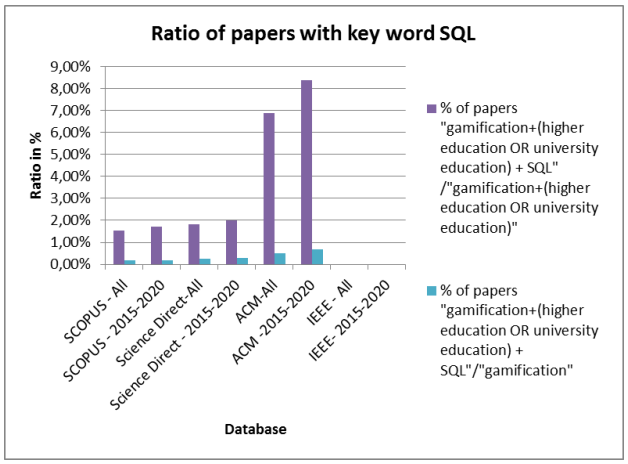


Figure 2. Ratios of number of papers with key words "gamification+(higher education OR university education) + SQL" to key words "gamification", "gamification+(higher education OR university education)"

B. Assessment and gamification in e-learning environments

Assessment is a driving engine in the educational process. In pedagogical literature, two main types of assessment are identified – criterion-referenced and normative-referenced assessment. According to Gronlund [5], criterion-referenced assessment measures students' individual achievements with respect to the set of learning objectives. On the other hand, normative-referenced assessment measures individual achievements against the achievements of other students. In e-learning environments, the use of badges or points combined with a conditional moving to the next level/module based on the awarded badges or points is an example of applying gamification with criterion-referenced assessment. The badges or points together with a leaderboard implement a normative assessment.

In [6] the authors present an approach for the evaluation of students' software projects via gamified peer assessment using badges and leaderboard. They have developed a Moodle plug-in for this purpose. The authors state that gamified peer assessment enhances students' understanding about the positive and negative aspects of their projects.

III. ASSESSMENT OF STUDENTS' SQL KNOWLEDGE AND SKILLS IN E-LEARNING ENVIRONMENT

The traditional database course consists of at least four major topics - data modelling, database models (by default a relational model), data normalization, and SQL as a language for database creating, maintaining and querying. The assessment of the knowledge and skills by default is based on solving tasks with open (textual or graphical) or short answer, multiple choice and Yes/No questions, filling gaps in template, and development of an individual or group project. Assessing open answer question and projects is a very complex task and is not suitable for automated assessment [7].

On the other hand, multiple choice and Yes/No questions are very suitable for automatic assessment and are implemented in any e-learning environments with testing capabilities. To prevent cheating during the assessment procedures, a question set used in each student's quiz or exam is selected randomly from a question pool. Moreover, shuffling of the questions and their answers is used too. Despite these techniques, to be successful against cheating, the question pool has to be big enough.

Four approaches suitable for on-line assessment of students' SQL problem-solving skills can be identified [7]:

- Creation of a SQL query from scratch in order to receive a specified result from a database with a given schema;
- Filling the missed part/s in a template of SQL query in order to receive a specified result from a database with a given schema and optionally state;

- Selecting the appropriate SQL query from the list of queries to receive a specified result from a database with a given schema and state.
- Evaluation of the SQL query result (number of rows returned) from a database with a given schema and state.

An interesting solution for on-line creation of SQL query from scratch as a part of the self-training and/or formal assessment procedure is presented in [8]. The authors state that this style of self-training and testing is close to the way professional developers create SQL to solve real tasks - starting from scratch and continuously refining their query up to the moment it becomes syntactically and semantically correct. In addition, the authors have developed an AsseSQL on-line tool for self-training and examination.

The AsseSQL tool provides on-line environment for creation of a SQL query and its execution against a training database in a particular state. During the process of creation, students can run the SQL query they have created as many times as needed, and receive feedback about its syntactical and semantical correctness. In the initial version of the tool, the semantical correctness was checked by comparing the desired result with the result from the execution of the SQL query against a current database state. In this way, any SQL query that returns the desired result is accepted as semantically correct, even though the result could be correct only in this database state. To avoid this weakness, in the next version of the tool, the query has been executed against two database states. In the end, the query is considered correct, when it can be executed without errors and returns the desired results with two different database states. [9]

Other tools also provide similar functionalities - SQLator [10], testSQL [11], and SQL Tester [12]. All of these tools are able to assess single table queries (SELECT-FROM-WHERE) optionally with GROUP BY and HAVING clauses, two tables join and/or simple subqueries. They don't have any or have very limited gamification capabilities and are not compatible with the standards for exchanging learning content or information among e-learning environments like LTI [13], xAPI [14], and CMI5 [15]. Being unique and very specific, they are not widely used in university teaching in SQL.

HackerRank [16] provides a good environment for self-training in programming languages and in SQL. Students can register and start creating SQL queries to solve problems and to obtain points for every correct solution. According to their personal achievement, students can move through the levels, receive awards (stars) and be ranked in a leaderboard. The lecturers may create a contest for the students as an element of self-training or assessment. Unfortunately, HackerRank cannot exchange the learning content and information about the students' achievements with other e-learning environments and in this way to be a part of the e-learning ecosystem.

The creation of SQL queries from scratch is used in some on-line games for SQL learning such as SchemaVerse [17], SQL Island [18], and GalaXQL [19]. These sites can be used only as supporting tools. Moreover the tasks cannot be changed or tuned according to the learning goals of the particular syllabus.

The second and third assessment approaches - filling the missed part/s in template of SQL query or selecting the appropriate SQL query from the list of queries in order to receive a specified result from a database with a given schema and state, can be easy implemented in e-learning environments. Their implementation can be done by using multiple choice questions and short answer questions. In addition, these approaches are widely implemented in many web sites for self-training and self-testing such as w3schools.com [20], sqlquiz.com [21], etc.

The approach for assessing the students' SQL skills by evaluation of the SQL query result as a number of rows returned from a database with a given schema and state can be implemented easily in the e-learning environment in the form of short answer questions. Because the answer is a number, any SQL query – from a very simple to a very complex one can be used in the assessment process. However, this assessment approach is not popular in the on-line systems for self-training and self- assessment, probably because it needs a database schema with a state provided. [7]

IV. GAMIFICATION IN E-LEARNING ENVIRONMENTS

Gamification features in e-learning environments can be classified in three types, according to the way they are implemented:

A. Core functionalities

These features are delivered with basic distribution of the e-learning system. By default, the basic gamification features included are: badges and points mark the individual achievement and progress of every student; levels to control the access to the learning content which guarantee the students' access to the next topic after passing the cut-off level of knowledge and/or skills; ranking (scoreboard) to compare the individual achievement of the student with those of other students.

B. System specific extensions

System specific extensions – practically every e-learning environment has its own mechanism to extend the core functionality by adding modules/plugins developed in specific for the particular environment way. Instead of the options to develop a special gamification extension for a particular purpose, many environments have a library of free or commercial extensions for gamification, such as scoreboards, leaderboards etc. For example, Moodle's extension module Level up! [22] enhances core gamification capabilities. Features such as leaderboard are used to display the ranking of the students; interface is used to display the current level and progress towards the next level and control is provided over content regarding the reached levels. This module has free and commercial versions.

C. *Standard-based extensions*

Standard-based extensions are extensions developed with respect to SCORM [23], LTI [13], xAPI [14], and CMI5 [15] standards. SCORM based extensions (named SCORM packages) can be "played" in all SCORM-compatible e-learning environments. The other three types of standard extensions can be used as externally provided services. An external provider can be a system developed for this purpose. In case of LTI resources the external provider can be another e-learning system. For example, badges in Canvas [24] are implemented as external LTI resource provided by Badgr [25].

V. APPROACH FOR SELF-ASSESSMENT IN GAMIFICATION MANNER

Teaching SQL nowadays is not an easy process. As was stated before, not all students are deeply involved into the ICT area. They prefer to learn in the same manner as they act in social networks or in game environment. In this way, traditional approaches on class training for solving SQL problems have become not so effective. Students want to have access to learning materials whenever and wherever they want to learn and train.

SQL knowledge and skills need to be built sequentially because each topic is based on the knowledge and skills obtained in the previous one. According to this, the lecturer has to be sure that the students have successfully completed the basic requirements for each topic. This can be done by a test after each of the SQL topics. To increase the students' achievements and avoid discouraging in case of bad test results, a set of tasks for on-line self-training and self-testing could be provided as a complementary part of the learning process. However, additional motivation is needed to involve the student in this extra training. Just informing the students that there is some useful training is not enough to motivate them to participate in it. The authors think that gamification can add the motivation needed as well as some fun in the self-training and self-assessment. This will help to build the students' SQL problem-solving skills more effectively.

Our approach for adding gamification in the self-training and self-assessment processes includes a scenario for gamification and SQL question generation system. We have decided to use evaluation of the SQL query result (number of rows returned) from a database with a given schema and state, as tasks for self-training and self-evaluation (Figure 3).

The students have to solve a randomly generated test for each SQL topic starting from the very first one. Questions from the current topic are included in the test. The test can be generated many times, each time selecting random questions, and the students can solve it as many times as they want. To obtain access to the next topic the student has to solve the test with more than 50% correct answers at least once. This proves that the students have obtained basic problem-solving skills for this topic and will be able to solve the problems in the next one as well. To gamify self-training and testing, badges are used for achievement of 50%, 70% and 90% correct answers. For each topic, a particular Bronze, Silver or Golden badge can be awarded according to the corresponding level of

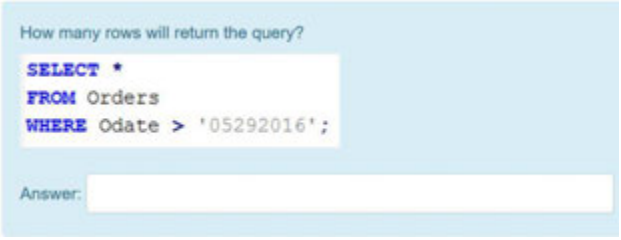


Figure 3. Evaluation of SQL query result test question, implemented in Moodle

achievements listed above. A badge leaderboard is used as an element of gamification as well. As extra stimulus, the students can take part in a pre-exam procedure instead of a final exam for final grading if they have collected silver or gold badges for every topic. If a student has collected all golden badges, extra points to the final grading can be awarded.

The Moodle e-learning environment is used for implementation of the scenario. As Moodle allows points and other rewards to be awarded automatically only after passing one cut-off per activity (there is only one level of passing), badges are collected manually by the students. A corresponding link becomes active when the student reaches 50, 70, or 90% of correct answers. Special bronze, silver and golden badges are prepared for each topic. After obtaining all silver badges, the student is awarded with a pre-exam badge and with an extra-point badge if all golden badges are collected. A Badge Ladder plug-in module is used as a leaderboard, where the students can compare their current performance with the other students anonymously.

As self-training is out of the class activity and gives some advance in grading, we have to use question pools big enough to prevent collective works and re-using of results. For this purpose, a system with automatic evaluation of SQL queries and generation of test questions in an appropriate format is developed. The system allows to evaluate sets of queries using the same database schema but in different states and to sequentially generate sets of test questions. The different versions of the question pool at different states for the same database schema allow the use of the database schema used for self-training in formal assessment procedures but in different states. The schema will be familiar to the students. The question pool can include the same questions but results will be different according to the database state used. In addition, the SQL

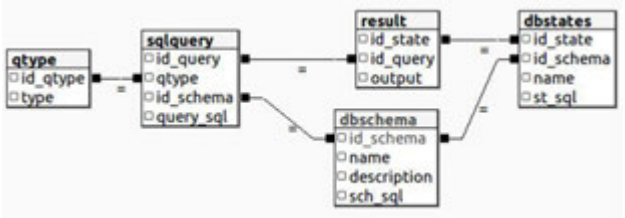


Figure 4. Database schema used to store database schemas, states and query sets used for automatic generation of question sets.

query code is presented as an image in order to avoid direct execution in an external system in case of leaks in a database state. During the formal assessment procedures, a database schema and state are given on paper.

VI. WEB-BASED SYSTEM FOR AUTOMATIC GENERATION OF SQL SHORT-ANSWER QUESTIONS

The first version of the web-based system for automatic evaluation of SQL queries and generation of test questions [7] has the following capabilities:

- Storing and managing relational database schemas and states;
- Storing SQL query sets related to the stored database schemas;
- Automatic execution of stored query sets using related database schema in different states;
- Converting the SQL query code to an image file before including it into a generated question.
- Generation of the short-answer questions in Moodle XML format for uploading in Moodle or other compatible e-learning systems.

In fact, the first version of the system has some limitations. They are expressed in the technology stack used for implementation as there is only one output format. To fix these drawbacks, we fully reengineer the system keeping only the database schema (Figure 4). Instead of Python and PostgreSQL, we decided to use PHP with PHP Data Object extension [26] to provide a data-access abstraction layer which allows different relational DBMS to be used for an automatic execution. Thus, the new system can be deployed in a wide range of execution environments – from desktops to clouds. In addition, the generation of the short-answer questions in IMS QTI 2.1 [27] format is added as a new functionality. This allows a wide range of open source and commercial e-learning environments to be used to implement our gamification approach.

VII. SOME RESULTS FROM PRACTICAL IMPLEMENTATION

It was planned to implement the presented approach during the spring semester of 2020 academic year as a part of a face-to-face SQL training course. The training tests were organized in four topics – Single table queries with WHERE clause, Single table queries with GROUP By and HAVING clauses, Multiple tables queries with JOIN, and Multiple tables queries – Subqueries/Sets.

Due to COVID-19 crisis, only three weeks of the semester were on-site and the other 12 were completely on-line, including all formal assessment procedures. We were not able to distribute a paper part of the proposed approach (database schema in some state) so we decided to include in each test question a part of the database schema in the state needed for solution. We have tested this new question view (Figure 5) during a self-assessment training. About half of the 32 students enrolled in the course took part in this training.

The results from a self-assessment training attempts are presented in Table II. In general, the students' feedback is positive – students are satisfied to train with assessment materials close to the formal exam and to compare their results anonymously with other students.

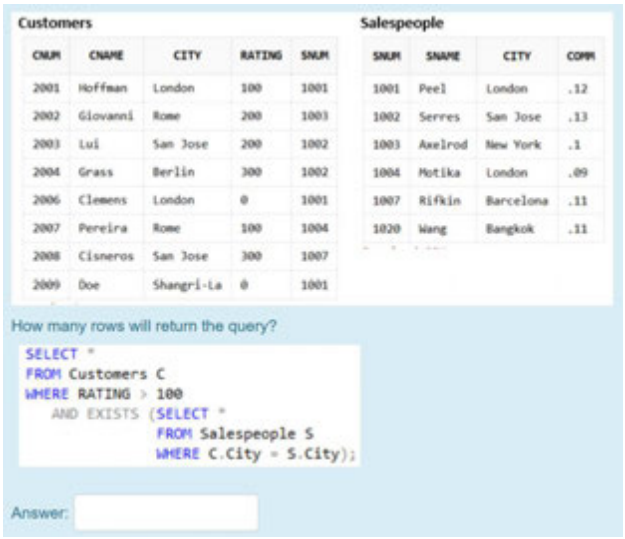


Figure 5. Test question with a part of the database schema in the state needed for solution, implemented in Moodle

TABLE II. SELF-ASSESSMENT REZULTS

SQL Topic	Attenders	Attempts	Bronze badge		Silver Badge		Golden Badge	
			Claimed	Achieved	Claimed	Achieved	Claimed	Achieved
Single table queries with WHERE clause	17	65	14	17	15	16	12	14
Single table queries with GROUP By and HAVING clauses	17	94	12	15	11	12	7	8
Multiple tables queries with JOIN	14	90	10	13	9	11	4	6
Multiple tables queries - Subqueries + Sets	11	42	7	10	7	9	4	6

Part of the students did not claim to obtain badges although they achieved appropriate results.

CONLUSION

This paper discusses the approaches towards the assessment of the students' SQL knowledge and skills and the application of useful techniques to prevent cheating during the assessment and self-assessment processes. The assessment tools integrated in e-learning environments to test students' SQL skills are analyzed. An approach to students' self-assessment in gamification manner is presented. To support this approach a system for automated generation of the test questions in Moodle XML and QTI formats for e-learning environments is presented.

The problems discussed can increase the students' motivation and achievements in database courses especially in SQL problem solving skills. Our future

works are directed to testing the approach with a larger number of students.

ACKNOWLEDGMENT

The study is funded by Bulgarian National Scientific Research Fund, Contract DN-05/10, 2016

REFERENCES

[1] ACM Curricula Recommendations, URL https://www.acm.org/education/curricula-recommendations

[2] Stoyanova, M., Tuparova, D., Samardzhiev, K.Impact of motivation, gamification and learning style on students' interest in maths classes – a study in 11 high school grade, Advances in Intelligent Systems and Computing, 716, pp. 133-142., 2018, DOI: 10.1007/978-3-319-73204-6_17, Springer Verlag, ISSN: 21945357, ISBN: 9783319732039

[3] Mora, A., Riera, D., Gonzalez, C., & Arnedo-Moreno, J. (2017). Gamification: a systematic review of design frameworks. Journal of Computing in Higher Education, 29(3), 516–548. doi: https://doi.org/10.1007/s12528-017-9150-4.

[4] Terzieva, T., Golev, A., & Stavrev, C.,Serious games – innovative tool for education. In proc. Innovative software tools and technologies with application in matematics, informatics, and education. 2017, pp. 104-107, University of Plovdiv "Paisii Hilendarski"

[5] N. E. Grounlund, Measurement and Evaluation in Teaching, 5th ed., Macmillan Publishing Company, 1985

[6] S. Simionescu, D. Šuníková, & Z. Kubincová, 2017, "Gamification of peer assessment in learning management system", 2017 18th International Carpathian Control Conference, ICCC 2017, pp. 571.

[7] G.Tuparov, M. Stanchev, Assessing students' SQL knowledge and skills in an e-learning environment, Computer Science and Education in Computer Science, vol. 1, 2019, pp. 77-79

[8] J. C. Prior, and R. Lister, "The Backwash Effect on SQL Skills Grading", ITiCSE '04, June 28-30, 2004, Leeds, United Kingdom.

[9] J. C. Prior, "AsseSQL: an Online, Browser-based, SQL Skills Assessment Tool, ITICSE'14, June 21–25, 2014, Uppsala, Sweden

[10] S. Sadiq, M. Orlowska, W. Sadiq and J. Lin. SQLator—an online SQL learning work-bench. InProceedings of ITiCSE'04, pages 223–227, Leeds, UK, June 2004.

[11] Joshua License. 2017. testSQL: Learn SQL the Interactive Way. In Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education - ITiCSE '17 . ACM Press, New York, New York, USA, 376–376

[12] A. Kleerekoper, A. Schofield, SQL Tester: An Online SQL Assessment Tool and Its Impact, TiCSE'18, July 2–4, 2018, Larnaca, Cyprus

[13] IMS Global Learning Consortium, Learning Tools Interoperability Core Specification, URL http://www.imsglobal.org/spec/lti/v1p3/, last visited on 25.01.2020

[14] Advanced Distributed Learning Initiative, xAPI Technical Specifications, URL https://adlnet.gov/projects/xapi-technical-specifications/ , last visited on 25.01.2020

[15] AICC, CMI5 Specification, URL https://aicc.github.io/CMI-5_Spec_Current/, last visited on 26.01.2020

[16] HackerRank, URL www.hackerrank.com, last visited on 25.01.2020

[17] SchemaVerse, URL https://schemaverse.com/, last visited on 25.01.2020

[18] SQL Island, URL https://sql-island.informatik.uni-kl.de/, last visited on 25.01.2020

[19] GalaXQL, URL http://sol.gfxile.net/g3/#, last visited on 25.01.2020

[20] w3schools.com, URL https://www.w3schools.com/quiztest/quiztest.asp?qtest=SQL, last visited on 25.01.2020

[21] SQL Quiz, URL http://www.sqlquiz.com/sql-quiz.php, last visited on 26.01.2020

[22] Level Up!,URL https://levelup.plus/, last visited on 26.01.2020

[23] ADL SCORM, https://scorm.com/scorm-explained/technical-scorm/, last visited on 26.01.2020

[24] Canvas Learning platform, URL https://www.instructure.com/canvas/en-gb, last visited on 26.01.2020

[25] Badgr, URL https://info.badgr.com/, last visited on 26.01.2020

[26] PHP DO, https://www.php.net/manual/en/intro.pdo.php, last visited on 25.01.2020

[27] IMS Question & Test Interoperability (QTI) Specification, URL http://www.imsglobal.org/question/index.html, last visited on 25.01.2020

Changing the Assessment Process in Mathematics for Students in Engineering

P. V. Georgieva* and E. P. Nikolova**
Informatics Department in FCSE, Burgas Free University, Burgas, Bulgaria
* penka.georgieva@bfu.bg, ** enikolva@bfu.bg

***Abstract* - Mathematical subjects studied in the Faculty of Computer Science and Engineering at Burgas Free University are focused on acquiring fundamental knowledge and developing essential skills required for future engineers. This paper looks into changing the process of assessment of students' achievements in mathematics from closed-book exams to diversification of tasks and expected outcomes. Three aspects of assessing the mathematical knowledge in engineering programs are considered: the importance of identifying the entry level, blended learning and creating students' portfolios with MatLab. The experience of the authors in using the tools of blended learning and the use of a software environment are presented. Some emerging problems in the educational process that need to be faced are outlined.**

Keywords - mathematical knowledge assessment; blended learning; students' portfolios

I. INTRODUCTION

The key goals of studying mathematical subjects in engineering bachelor's and master's degrees were unchanged for the last century and they are: acquiring fundamental knowledge and developing essential skills that are required for future engineers. Achieving these goals today is impossible by following the old-fashioned methods of admission, training and learning such as:

• for enrolling in engineering studies, applicants have to pass a difficult mathematical entrance exam with the presumption that strong mathematical knowledge somehow predetermined the high potential of engineering skills;

• "standard" lectures: the professor "reads" the lecture, and the students "listen" and eventually write down notes in almost complete silence;

• "standard" mid-exams: the lecturer offers several problems to be solved and the students are expected to write down the solutions on paper, with the requisition that all the necessary knowledge must be pre-memorized;

• each of the math courses ends with an exam procedure, in which firstly problems are solved as described above, and only those students who have shown the knowledge and skills to solve more than 50% of the problems correctly, pass to the second stage, in which they have to write on 2 or 3 topics from the exam questionnaire (again pre-memorized) i.e. the students sat closed-book exams.

But what was going on in the period between the lectures and the exam was not very important and/or interesting for the lecturer. What is more, after the mathematics exam, the future engineer was free to forget the lemmas, theorems and propositions, as well as their proofs.

Although the authors classify these approaches as old-fashioned, they have had produced generations of capable engineers, thanks to which engineering and technology have received such rapid development. Yet one fact has to be taken into account - in the previous decades, the engineering sciences remained steady and stable for a decade or two (in some areas).

Now the conditions, the requirements and the realities have changed substantially and each university is adjusting methodologically.

Firstly, the number of applicants for engineering majors has dropped significantly (Figure 1). For example, the decrease of engineering students in numbers for the last 10 academic year is almost 35% [1]. The number of students enrolled in engineering programs at BFU has been decreasing approximately by 40 % for the last 20 years. This leads to reduced competition at the entrance, where students with lower-than-average achievements enroll and begin to study in engineering programs.

Secondly, engineering knowledge itself is much more dynamic, rapidly evolving and thus requires different

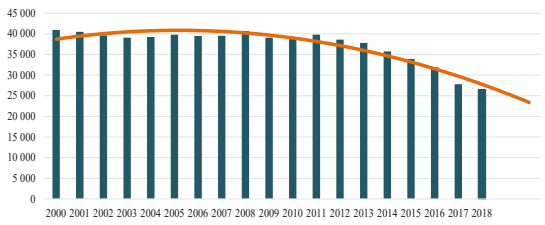


Figure 1. Number of students studying Engineering degrees in Bulgaria for the period 2000-2018

This research is funded by National Research Fund of Bulgaria under Contract No. KP-06-COST-8/ 06.08.2019 for providing national co-financing for the participation of Bulgarian teams in approved actions under the European program for cooperation in the field of research and technology COST under the project "Characteristics prediction and optimization of a photovoltaic system with artificial intelligence methods".

tools, technologies and skills to be acquired. For example, engineering students no longer draw by hand but use a wide variety of CAD/CAM software systems. BFU is not an exception and the discipline "Fundamentals of Design and AutoCAD" has been introduced for all engineering majors at FCSE since 2016. What is more, new specialties are emerging in the engineering field: robotics, IoT, and more. In response to current tendencies two new master degree programs have been introduced in FCSE: Artificial Intelligence and Robotics, and Electric Cars.

Thirdly, the students show little willingness to learn theoretical concepts. For example, the diagram in Figure 2 presents the results from the theoretical part of the linear algebra and analytical geometry exams (2015-2019). For this part of the examination procedure students are evaluated on a scale from 0 to 20 points. The chart shows that the percentage of students who received 15-17 points is below 17% (excluding academic year 2016/17), while the percentage of those within 18-20 points is below 12%. In addition, it is important to note that students' knowledge is not systematized in a coherent scientific system, but is somewhat granular, detached, and slightly chaotic. They have easy access to information and knowledge from all aspects of science and life. If they don't know something, they simply google it. If they can't do anything, they just watch a video (or two) and learn how to do it.

These facts also necessitate changes in learning, teaching and knowledge assessment. However, these changes in no way undermine the fundamental role that mathematical subjects play in the preparation and development of future engineers.

This article focuses on the transformation of the process of assessing the students' achievements in university mathematics from closed-book exams to diversification of tasks and expected outcomes.

Three aspects of assessing mathematical knowledge in engineering programs are considered: *the importance of identifying the entry-level; blended learning* and *creating students' portfolios with MatLab.*

The authors' experience in using blended learning tools and the use of a software environment is presented.

Some emerging problems in the educational process that need to be faced are outlined.

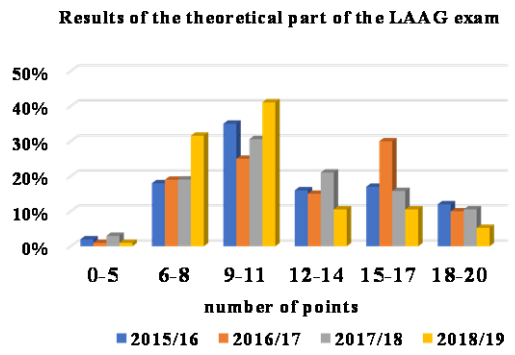


Figure 2. Results from the theoretical part ot the exam in Linear algebra and Analytic Geometry for the period 2015-2019

II. Changes in Learning Environment

Challenging changes in the learning environment are widely discussed in academic circles. E. g. prof. Sh. Alexander focused on the rising cost of education and employability for graduates and how technology might be used for a more sustainable future in education have been discussed in the plenary of *OEB 2019 Pushing Technological Boundaries*. [2]

This paper is focused on active learning, the use of software systems for computer mathematics and internet in the learning process.

A. Active Learning

The concept of active learning is based on shifting from memorization of knowledge and facts (surface learning) toward understanding (deep learning). [3]

Four broad categories of instructional approaches in active learning have been identified: (a) individual activities, (b) paired activities, (c) informal small groups, and (d) cooperative student projects. These methods encompass many activities such as concept mapping, brainstorming, collaborative writing, case-based instruction, cooperative learning, role-playing, simulation, project-based learning, and peer teaching. [4]

Project-oriented learning (POL) involves students in various projects, usually leading to products [5, 6]. However, this process aims at the learning effect rather than the product itself. POL didactic strategy portrays active learning as an educational paradigm that transforms direct experience into a tool for supporting and stimulating learning. Key aspects of POL are: working autonomy, practical relevance, learning of soft skills, cooperation between university and practice.

In this context, engineering students in Burgas Free University are involved in small projects within individual courses, progressing to a final year project course. The projects are combined with traditional teaching methods within the same course with focus on the application. Projects are undertaken throughout the length of the course and vary in duration from a few weeks up to a whole year. Although POL is closely related to specialized subjects, there are possibilities for its application both in mathematics as a tool for engineering calculations and in the study of pure mathematics. This is because POL, being directed to applications of knowledge through understanding, provokes strong interest in students. What is more, students receive both theoretical and technical knowledge and additionally soft skills such as working in a team, time and resource management, and last but not least - finding solutions, arguing and defending their opinion and position. [7]

B. Using Software Systems for Computer Mathematics

Many years of experience of the authors of this article show that lecturing on the whiteboard is an indispensable

methodological approach in teaching mathematics. The authors very much agree with and try to follow as much as possible Recommendation 4 of London Mathematical Society: *Despite the agreed importance of modern, computer-based teaching and learning, lectures delivered using clearly visible boards should continue to play an important role.* [8]

Nevertheless, the fact that easy access to powerful computers and software has led to significant changes in the teaching and learning of mathematics cannot be disregarded. Software packages that can be used as practical tools for the studying mathematics have been developed. MATLAB, MAPLE, and MATHEMATICA can support learning mathematics by performing complex symbolic calculations and transformations as well as for visualizing concepts and results. Statistical packages have also made a huge impact, allowing real-world data to be explored and analyzed. [8]

MATLAB is a software environment for performing various calculations (numeric and symbolic) and visualization of the obtained results. This software system consists of a kernel with built-in features that are dynamically complemented by toolboxes and additionally it has a built-in programming language that is capable of creating new applications. Key features of MATLAB are: accuracy in calculations, reliability of results, a large number of applications created in almost every field of science, the accessibility of applications (many are with open source code), vectorized operations, etc. These important advantages of MATLAB for engineering research are the prerequisite for its use in mathematical subjects at the Faculty of Computer Science and Engineering at Burgas Free University. [9], [10]

C. Internet in the Learning Process

Nowadays, using the internet for educational purposes plays an important role in both the teaching and learning processes. The benefits of incorporating internet in mathematics education are numerous. The Internet can be used as a methodological tool, e.g. video and animations are an effective way to illustrate geometric objects, vectors, etc. What is more, the students can contact their lecturers or/and other students almost continuously not only by emails but while working on assignments in web-based courses for assisted training by receiving comments and analytics. Internet resources help the students to explore new ideas and enrich their knowledge. Additionally, some online calculators visualize solutions step by step and thus are a useful tool in the learning process. Such are matrix calculators [11-13], derivative calculators [14-16], integral calculators [17-19], and others.

Changes in the learning environment necessitate changes in the forms of assessing knowledge and skills. The transition from traditional assessment forms to the diversification of the expected outcomes and how the lecturers in Mathematics at Burgas Free University proceed in achieving it are discussed in Section III.

III. From Closed-book Exams to Diversification of Outcomes

A. A Brief Overview of Assessment Practices

A preliminary clarification on the authors' opinion is needed at this point. The authors of this paper believe that assessment is a part of the learning curve and its purpose is to inform and thus to improve both learning and teaching. The focus is on developing students' mathematical knowledge as well as their ability to evaluate themselves and improve.

The main types of knowledge assessment are as follows [20]:

(1) diagnostic assessment - aims at assessing students' knowledge and skills before the start of the course;

(2) formative assessment - focused on students' performance during the course and is done regularly throughout the semester;

(3) summative assessment - measures students' achievement at the end of the course;

(4) norm-referenced assessment - compares a student's performance against other students;

(5) criterion-referenced assessment - measures a student's performance against a goal, specific objective, or standard;

(6) interim/benchmark assessment - evaluates the overall student's achievements at the end of a grading period.

Assessment methods are the techniques and instruments for gathering information about the attained level of knowledge and skills. Subject learning outcomes include completing an action (describe, analyze, evaluate, solve, design or create) in the content of the scientific area, according to the context or professional environment or practice in which the student can apply the skills and knowledge. Here are eight important ways to measure students' achievements in a university subject.

1. Tests - can be objective or subjective; multiple-choice or free-response, true-false answers, matching answers, filling blank space; written or oral; national or locally generated.

2. Surveys - written or compiled through interviews. Students can be surveyed in courses (about the courses), as they graduate (about the major), or as they change majors (about their reasons for changing).

3. Evaluation reports - an individual or group is evaluated through a checklist of skills and abilities.

4. Portfolios - collections of student work, usually compiled for individual students under faculty supervision following a standard departmental protocol. The contents may be sorted into categories and by types, such as homework, formal written papers, or examinations. The work collected in a student's portfolio should reflect the student's progress through the major.

5. Essays and reports – focused on writing skills in mathematics as well as knowledge of the subject matter.

Essays and reports can be a part of courses and should be candidates for inclusion in portfolios.

6. Oral presentations - demonstrate speaking ability, confidence, and knowledge of the subject matter.

7. Practical assignment – students are involved in project, research, discussion, modelling.

8. Dialogue with students – students' attitudes, expectations, and opinions can be a valuable element in the assessing process. Some methods for collecting data are student evaluations of courses, interviews held by faculty members or administrators, advising interactions, seminars, student journals, and informal interactions. [21]

Student cooperation and involvement are essential to most assessment methods.

Various alternative assessment methods have been developed over the last decade [22], for example:

- research skills in mathematics - the assessment of students' writing skills, oral presentations and peer-reviewing of paper and/or report drafts;
- designing test questions by the students - students are allowed to create own question and answer based on the course materials;
- presentation of a group work in the assessment - as part of the assessment the students present chapters from a book on group presentations;
- group project which consists of two presentations of the solution of an open-ended problem set by the lecturer and an individual written report;
- weekly online quizzes;
- presentations of applications of pure mathematics - the assessment consists of creating a poster which illustrates an application of math's theory to real-world problems and presenting it;
- continuous assessment in a history of mathematics module which is accomplished through essay writing, peer-assessed posters for mini-projects and the solution of a mathematical question with the appropriate historical tools;
- the use of an algorithmic e-assessment for summative and formative assessments;
- mathematical modelling project developed using a mathematical software package;
- a written group project with individual presentations of the same project;
- presentations and quick quizzes on basic material from the course;
- Moore-method tasks- students have to find alone answers to given problems without the help of supporting material and present their solutions in class.

B. Motivation for changing the assessment procedure

The starting point for designing the assessment procedure is to identify the area of knowledge, the appropriate depth and difficulty as well as the soft skills that the course will aim at. During the mathematical training of future engineers at BFU, the lecturers' intent is the students to acquire additional competencies such as communication skills, creative thinking, team working, ability to deal with other people, decision making, problem-solving, personal character traits, etc. [23]

Closed-book exam is a traditional assessment method in which students rely entirely on their memory to solve problems, answer questions or write on a topic. It was exclusively used for assessing the mathematical knowledge of the engineering students at BFU in the 20th century. The exam consisted of solving problems and for those who met the criteria for a positive result, a written theoretical examination (writing definitions, theorems and proofs on predefined topics) was held.

Later, in some mathematical courses, the closed-book exams were substituted with tests. Over the years this assessment method proved to be unable to evaluate students' mathematical knowledge and skills neither as a system nor in detail. Gradually the tests were replaced with mixed assessments.

Since 2000, the evaluation of the academic achievements of students at BFU has been regulated by the BFU Students' Knowledge and Skills Assessment System. One of the basic principles of the system is the systematic nature of the students' efforts throughout the learning process through the use of effective credit accumulation mechanisms. The point scoring system requires ongoing monitoring of the knowledge acquired during the semester. Current monitoring is realized by two or three midterm tests, individual exercises in the electronic training platform and individual/team coursework.

Key principals for diversifying the tasks that form the final mark are:

- ***formative assessment*** – the point-based mark system is introduced in many universities and thus the assessment is based on an ongoing assessment by monitoring students' progress and providing feedback on how to improve their learning process and a summary assessment that assesses how well the students have performed against the assessment criteria selected;
- ***assigning weights to different activities*** – allows the students to learn gradually and make decisions on when to complete which tasks;
- ***feedback*** - students receive feedback on their achievement throughout the academic year and thus are encouraged to work steadily and continuously and thus to improve their performance;
- ***creativity*** - students are provided with the opportunity to be creative in new ways of studying and to develop a range of transferable skills they need to successfully enter the labour market. On the other hand, the diversified assessment provides lecturers with an opportunity for creativity in the design and development of assessment tasks.

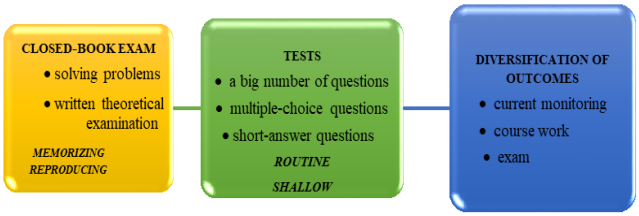


Figure 4. From closed-book exams to diversification of tasks

The final mark is complex and includes points from the current monitoring, completing the course work and from the exam with 20% gained from current monitoring, 10% - course work and portfolios, and 60% - from the final examination procedure. (Fig. 4)

IV. Experience, Results and Emerging Problems in Assessing the Mathematical Knowledge in Engineering Programs

Assessment of students' mathematical skills at BSU begins with an entry-level exam and continues in each of the mathematical disciplines during the training period. In addition to these skills, a number of other transferable skills are assessed through individual and group coursework. Web-based courses provide an opportunity for assessment during the semester. Development of skills for using software packages in solving problems, as well as the evaluation of the achieved results is a main goal in mathematical courses.

A. Identifying the Entry-Level

The entry requirements for engineering degrees reflect the fact that there is a substantial mathematical component. This is the key motivator for conducting an exam at the beginning of the first semester for determining the entry-level in mathematics for the last 10 years. The exam covers topics from the upper secondary education level required for university courses. Based on the results, students who did not perform well (less than 60% of the test points) are required to attend an optional mathematics course. At the end of this course, they take the exam again and after passing it they can proceed with higher mathematics subjects.

The results from high school math also show a decrease. Indeed, while in 1995-1999 80% of the students showed marks from high school diplomas in the interval [5.00, 6.00], in 2000-2018 80% of the students had marks in the interval [4.00, 5.25]. Another trend observed in BFU is skipping the first exam date and attending the exam later.

B. Blended Learning

Blended learning methods use the best of traditional classroom learning and e-learning. The role of assessment in the blended learning is to create engaging assignments and to monitor student progress.

During their education, the students of Burgas Free University may use web-based courses for assisted aining. BFU currently is offering 291 BA courses and 78 MA courses delivered via blended learning ethodology (https://moodle.bfu.bg/). The courses are eveloped in accordance with the framework approved at niversity level:

- *General information about the course*: subject and objectives of the course, for whom it is intended, what students will know after passing the course, what are the duties of all involved in the educational process, what is the duration of the course.
- *Assessment and evaluation methods, procedures, and tools;*
- *Schedule of the consultations.*
- *Course material;*
- *Teaching aids* (references to books or developed guides, examples)
- *Additional support materials* - software products; a list of internet resources related to the material, and more.

The ways in which BFU mathematics professors incorporate blended learning principles to the classroom are:

- *digital revision tasks*- the digital training platform is used to set up individual exercises for students and to monitor the progress of each student throughout the semester, as well as to identify the weaknesses;
- *digital assessments for feedback*- each of the courses provides students with self-assessment tests;
- *virtual consultations*- BFU mathematics professors use the digital learning platform for counselling the students.

C. Students Portfolios with MATLAB

Student's portfolio is a collection of research papers, reports, tests, exams, case studies, video, essays, journals, self-evaluations, exercises, etc. in which the students present examples of their best work. The portfolio is collected throughout a program or subject and is assessed by faculty members.

Introducing this additional opportunity for scoring points is a result of the lecturer's drive to improve students' programming skills, to develop students' mathematical modelling skills, to show them ways to use their mathematical knowledge in various subjects, and to develop students' employability skills.

A variant of the students' portfolio is the *focused portfolio*, which concentrates on a specific aspect of a mathematics course. In mathematical subjects at BFU, this aspect is the use of mathematical software package MATLAB. The authors of this paper, being lecturers, expect that students work on the assignments during the tutorials under the guidance of the tutor and/or individually as homework. They are provided with a template for presenting their assignments. The work, collected in the student's portfolio, reflects his/her progress through the main stages of the course and is

reviewed 2 (for Calculus) or 3 times (for Multivariable Calculus) per semester. The emphases are on differentiation, integration and graphing. There is one task involving the transition from problem-solving to proving. An additional practical assignment for developing a mathematical model and present it as a group project is included as optional. As an illustration three typical assignments for including in the portfolios are shown below.

1) Draw the area D: $x=y$; $x=3-2y^2$, calculate the double integral $\iint(y^2-x)dxdy$ and interpret the result. (use Matlab)

2) Define the integral creteria for series convergence. Use it to the convergence of ∑1/(n(ln(n))^(1/2)), n≥2. Using Matlab, visualize the first 150 partial sums of the series.*

3) Let $i(t)=I_m*sin(\omega t-\varphi)$ (sinusoidal current) and $u(t)=U_m*sin(\omega t)$ (sinusoidal voltage), where I_m and U_m are the amplitudes of the current and voltage, ω is the angular frequency and φ is the phase difference between voltage and current. Find the average power. Graph i(t) and u(t) for $I_m=2$, $U_m=3$, $T=\pi$ and $\varphi=\pi/2$.(use Matlab)

Several additional documents are included in the portfolio. These include an entry profile, an entry questionnaire, counseling forms, and advisor notes. The entry profile contains information about the student prior to enrolling at the BFU, such as the mathematics course grades from high school. It also contains the entrance questionnaire, the placement scores from the entry-level examinations in mathematics at BFU and the advisor notesDuring

In the current academic year, the authors plan to add student voice as an important addition to the portfolio. The students will be invited to choose the types of tasks to be included in their portfolios - homework, essays, exam papers, written samples, independent project reports, and may further open the door to students' attitudes and feelings (e.g. self-awareness of their path – "Where you started from, where you want to be, and where you are now"). Our experience shows that the process of portfolio creation itself is essential in the education of each student because it is focused on the individual performance, not on the accumulation of every possible information.

In our opinion, some emerging problems in engineering education are to be faced in the next decade. Three of them are outlined below. First, effective adaptation for the first-year students who show fewer academic skills and learning habits. Second, the almost unlimited volume of mathematical applications decreases the motivation of the students. And third, it is not always certain that students complete their assignments themselves.

V. Conlusion

This article presents the experience of mathematics lecturers at the FCSE of BFU regarding the new trends in the assessment of mathematical knowledge in engineering programs using blended learning tools and software environments. Attention is drawn to formative assessment in blended learning, which actively engages students in the learning process, builds peer and self-assessment skills and provides an opportunity to evaluate both academic achievement and transferable skills.

The authors of this paper are planning to implement some elements of learning analytics and using their experience in point-based assessment system and blended learning, to focus on new effective approaches is learning and assessment process.

References

[1] https://infostat.nsi.bg/infostat/pages/module.jsf?x_2=164

[2] OEB 2019, Pushing the Technological Boundaries, www.youtube.com/watch?v=KbYAezm7SiA&feature=youtu.be

[3] R. Ritchhart, M. Church, and K. Morrison, "Making thinking visible: How to promote engagement, understanding, and independence for all learners," CA: Jossey-Bas, 2011.

[4] A. Roehl, S. Reddy and G. Shannon, "The flipped classroom: an opportunity to engage millennial students through active learning strategies," JFCS, Vol. 105 (2), 2013, pp. 44-49.

[5] St. Bell, Project-Based Learning for the 21st Century: Skills for the Future, The Clearing House: A Journal of Educational Strategies, Volume 83, Issue 2, 2010, pp. 39-43.

[6] A. Kolmos, "Problem-Based and Project-Based Learning," in Skovsmose O., Valero P., (eds) University Science and Mathematics Education in Transition, Springer, Boston, MA, 2009.

[7] S. Mollova, M. Zhekov, A. Kostadinov, and P. Georgieva, "Laboratory model for research on computer cluster systems," Proceedings MIPRO, pp. 1595-1600, 2018

[8] "Mathematics Degrees, their Teaching and Assessment," London Mathematical Society, 19 Nov 2009, https://futurelms. wordpress.com/2010/04/08/teaching-position-statement/

[9] P. V. Georgieva, "Systems for computer Mathematics and MOOCs," Computer Sciences and Communications, Vol. 7 (1), 2019, pp. 125-131

[10] P. V. Georgieva, "Applying MUPAD in university Mathematics courses," BFU ANNUAL, Vol. XXX (1), 2014, pp. 236-245

[11] https://matrixcalc.org/bg/

[12] https://www.symbolab.com/solver/matrix-calculator

[13] https://www.emathhelp.net/calculators/linear-algebra/matrix-calculator/

[14] https://www.derivative-calculator.net/

[15] https://www.symbolab.com/solver/derivative-calculator

[16] https://www.mathpapa.com/derivative-calculator/

[17] https://www.integral-calculator.com/

[18] https://www.symbolab.com/solver/integral-calculator

[19] https://www.wolframalpha.com/calculators/integral-calculator/

[20] B. V. Prasanthi and V. Ж. Inti, "Classroom assessment methods and tools: a review," International Journal of Research and Analytical Reviews, vol. 6, issue 2, 2019, pp. 94-97.

[21] B. Gold, S. Z. Keith, and W. A. Marrion, Ed., "Assessment practices in undergraduate mathematics," MAA Notes No 49, 1999

[22] Mapping university mathematics assessment practices, University of East Anglia, Edited by Paola Iannone, Adrian Simpson, 2012, https://www.birmingham.ac.uk/Documents/college-eps/college /stem/mapping-universities-mathematics-assessment-practices.pdf

[23] P. V. Georgieva and E. P. Nikolova, "Enhancing communication competences through Mathematics in Engineering curriculum," 42nd MIPRO, 2019, pp. 1451-1456.

Design of Rubrics for Student Outcomes in 2019-2020 ABET Criteria

B. Pejcinovic*
* Portland State University, Electrical and Computer Engineering dept., Portland, OR, USA
pejcinb@pdx.edu

***Abstract* – ABET is the main accreditation body for engineering programs in the United States and they have recently introduced a new set of Student Outcomes. This set was reduced from 11 to 7 items by combining several outcomes into one and adding some new ones. In our electrical and computer engineering programs we decided to design a set of seven general rubrics, one for each ABET outcome. These rubrics could then be used unaltered if course content fits them, or they can be adjusted to fit a particular course. To use a common description for rubrics, we wanted to keep the Performance Criteria the same but can adjust the Performance Indicators to suit a particular course. Six rubrics are presented in detail. We also share some initial observation in practical implementation of these in course and program assessment. They have helped us identify a problem in our sophomore cornerstone class related to the quality of student designs. Similarly, in our senior capstone-related class we identified a problem with defining and understanding ethical dilemmas.**

Keywords – rubrics, assessment, evaluation, accreditation.

I. INTRODUCTION

Curriculum design of engineering programs is an ongoing process that has yielded many different approaches, each with its own strengths and weaknesses. The process itself is driven by many different factors, such as accreditation, available resources, government policies, and faculty beliefs. There have also been many calls for reforming, or even revolutionizing, engineering education, as well as increasing the number of engineering graduates, see e.g. [1] and [2]. Once new programs are designed and implemented, they must be assessed and evaluated.

Program assessment and evaluation are largely driven by various accreditation bodies. In the US, ABET is the main accreditation body for engineering and engineering technology programs. Accreditation requirements and procedures have exerted great influence on the structure and content of engineering programs. In the late 1990-s and early 2000-s, ABET started what is known as EC2000 criteria along with outcomes-based assessment and evaluation.

Even though the original intent was to allow flexibility in designing program outcomes, many programs (including ours) decided to keep the structure and requirements that ABET listed. These were the so-called "a through k" student learning outcomes (SLO), as discussed in, e.g., [4]. These were to be tied to program educational objectives (PEO). Both SLOs and PEOs were supposed to be assessed and evaluated periodically. For example, we evaluated SLOs every year and PEO-s every third year.

Over time, however, it became clear that it was very hard to do proper assessment and evaluation of PEO-s, due to the requirement that they be evaluated three to five years after students graduate. In our experience, this assessment and evaluation turned out to be a logistical nightmare and it rarely produced actionable feedback. On the other hand, ABET's insistence on continuous improvement approach to curriculum design and assessment has led to positive results because it shifted the emphasis from topics coverage to actual student learning, as discussed in [5] and [6].

Implementing their own requirement of "continuous improvement" ABET embarked on revising EC2000 criteria resulting in the new criteria listed in [3]. There are eight criteria used in program accreditation by ABET (quotations in italics are from ABET website [3]):

1. Students is concerned with evaluation, monitoring transfers, and graduation requirements.
2. Program Educational Objectives are "*broad statements that describe what graduates are expected to attain within a few years after graduation.*" In the past, programs were expected to assess and evaluate this criterion by, e.g., contacting graduates 3-5 years after graduation. At present, however, programs only have to document it and review it periodically.
3. Student Outcomes describe "*what students are expected to know and be able to do by the time of graduation. These relate to the knowledge, skills, and behaviors that students acquire as they progress through the program."* This is where the bulk of assessment and evaluation of student learning is addressed. The results are used as inputs to Continuous Improvement.
4. Continuous Improvement describes what processes should be in place to ensure that the results of assessment and evaluation are used to improve a given program.
5. Curriculum – ABET does not require specific courses for each engineering program but it does give some requirements regarding e.g. number of credits to be devoted to math and science courses. It is in here that ABET declares that "culminating

major engineering design experience" must be part of curriculum.

6. Faculty, Facilities and Institutional Support are the remaining requirements but do not directly deal with students and their learning.

Programs undergoing accreditation evaluation tend to be most carefully examined on criteria 3 and 4. These also cover areas that faculty are most directly involved in through collection of assessment data and its evaluation. Faculty involvement and buy-in is a critical component of successful preparation for accreditation.

In the sections below, we first introduce rubrics and follow that up with a discussion of how our faculty used rubrics in assessment and evaluation of ABET Criterion 3. Lastly, we provide some concluding remarks.

II. BRIEF INTRODUCTION TO RUBRICS

Many benefits are claimed for rubrics, such as clarifying course or program goals, improving feedback, communicating expectations, etc. When used in courses, rubrics can be helpful in scoring course assignments [7]. Our intent is to use them for assessment and evaluation of student outcomes at the program or departmental level. Given that outcomes are defined at the program level, the rubrics will be more general than what one may use in a course.

The easiest way to explain the construction of rubrics is given in Table 1. As an example, we use "Systematically develops, compares and ranks design alternatives to arrive at a final solution" as one criterion in evaluating how well students apply engineering design. How well they perform is described by a performance indicator and the level at which student is performing. In this example, student who is performing at the "Proficient" level would "develop several good alternatives" but a student at the "Beginning" level would "Consider only one design." Criteria, therefore, describe our general expectations from student work, and indicators give specific ways that students can demonstrate how well they have mastered them.

Table 1. Outline of a generic rubric in a table format.

Criteria	**Performance level 3 (e.g. Proficient)**	**Performance level 1 (e.g. Beginning)**
Criterion A. (e.g., Identifies ….)	Performance **indicator 1** at level 3	Performance **indicator 1** at level 1
Criterion B. (e.g. Develops …)	Performance **indicator 2** at level 3	Performance **indicator 2** at level 1

We standardized on four levels of performance: Exemplary (4), Proficient (3), Developing (2), and Beginning (1). There are many other ways that these can be segmented and labeled but this set seemed to explicitly capture how we approach evaluation of student performance. Note that the performance indicators in this work are given only for the Proficient and Beginning levels to save space. In general, we consider that graduating students should be at the Proficient level. Exemplary level would typically be attained by a smaller fraction of all graduating students. Transition from Beginning to Developing to Proficient can be used to gauge student development within a course or, more likely, within curriculum. Therefore, it may be quite acceptable in introductory courses to have a lot of students at the Developing or even Beginning levels.

Finally, rubrics that will be presented below are fairly general because they are meant to be used for program assessment. The expectation, however, is that individual instructors will tailor them to suit a particular course where they are being used. The initial development was largely based on author's experience with cornerstone and capstone courses where they can be readily applied.

III. ABET CRITERION 3. STUDENT OUTCOMES

Student outcomes are broken into seven individual criteria, as discussed below. For each we quote ABET description followed by a discussion of how each is addressed by a rubric. Where possible, we describe our initial experiences and future plans.

A. Problem solving and analysis (outcome 1)

On the surface, this outcome seems to be the easiest to explain to engineering faculty because they believe that most of what they teach directly addresses this outcome:

> *"an ability to identify, formulate, and solve complex engineering problems by applying principles of engineering, science, and mathematics" [3]*

We have identified the following criteria to help us assess it. Note that each criterion listed below has several performance indicators, but they are not given here.

A. Identifies problems with a quantifiable solution that can be approached systematically.

B. Selects appropriate methods for solving the problem.

C. Formulates the problem according to chosen solution method and identifies key issues/variables

D. Recognizes the need for multiple solutions

E. Analyzes alternative solutions to an engineering problem

F. Justifies a solution to an engineering problem

This criterion is perhaps the hardest to express in a general rubric format because it varies widely across courses. There are several issues with devising appropriate criteria as well as the use of the rubric:

1. Faculty are used to assigning problems that can be solved in a relatively short time, for example, during a test. Somewhat longer problems may be assigned as homework, but this brings up issues related to copying from other students and from resources on the internet.
2. This results in problems that are not complex enough to be directly applicable to this program criterion.
3. It is only rarely that students are asked to consider multiple solutions and select the most appropriate ones. Sometimes this is implicit in the problem

itself but in order to evaluate it this needs to be an explicit requirement.

4. Justifying a solution is not normally required so long as students reach correct solution. To assess this properly, this must be required in the problem statement.

What these issues indicate is not that this is impossible to do, but that it requires careful consideration and explicit planning for assessing this outcome. Different criteria can be assessed and evaluated in different problems, but they need to be taken together to evaluate the overall student performance.

In the first year of use we have found the rubric to be valuable in formulating proper assessment questions in individual courses. Once faculty are familiar with individual criteria then it becomes easier to identify possible topic areas and assignments that can be modified so that they provide good assessment data. However, this is still an ongoing process and one that requires constant communication with the faculty who are tasked with this assessment and evaluation.

B. Engineering design (outcome 2)

"Design" shows up in the titles of many engineering courses. However, design means many different things and is dependent on the context. Engineering design is process oriented with the goal of devising a system, component, or process. Typically, it will have a specific targeted need (goal) and it involves a decision-making process. More complete definition and details can be found in [3]. Given this definition, ABET identifies the following outcome:

> *"an ability to apply engineering design to produce solutions that meet specified needs with consideration of public health, safety, and welfare, as well as global, cultural, social, environmental, and economic factors"*

Our initial rubric addressing this outcome is given in Table 2. Note that only a subset of performance indicators is given, which suffices to illustrate what is going on and what we are looking for in students' work.

One of the curricular issues related to this outcome is over-reliance of programs on demonstrating student attainment of it through assessment and evaluation of capstone projects. While this is an obvious time and place to perform this evaluation, waiting until the senior year makes it impossible to track student development, identify causes of observed issues, and implementing improvements. This is why we have added a cornerstone course in the sophomore year [8] as a follow-on to three freshman courses with significant, but somewhat basic, project component [15]. This course covers similar grounds as our capstone course but in a simplified form [10].

We have applied this rubric in the evaluation of designs produced by student teams in the cornerstone course taught in Winter 2019 term. There were 57 students in 15 teams. One area of concern was students' inattention to formalizing requirements and constraints in their designs. Anecdotally, students like to dig into technical challenges but consider formal requirements an unnecessary burden. This has lead to changes how we teach and what we require from student projects. One change will be related to requiring that students justify their solutions by describing the design process and which alternatives they have considered. So far, we are observing student performance to be somewhere between Developing and Proficient, which is appropriate for a sophomore-level course.

Table 2. Criteria and a sampling of Performance Indicators for outcome 2 - Engineering Design

Criteria	Proficient (3)	Beginning (1)
A. Identifies and follows a logical and orderly design procedure.	Needs minimal help in identifying the procedure, understanding steps and staying on track.	No discernable effort made to identify or follow a procedure. Haphazard approach taken.
B. Systematically develops, compares and ranks design alternatives to arrive at a final solution.	Several good alternatives developed	Only considers one design option.
	Systematically compares and ranks alternatives	No evidence of systematic comparison or ranking
C. Creates a final solution that satisfies all requirements and constraints identified in formulating the design problem.	Requirements and constraints are appropriate and well justified	Identification of requirements and constraints in formulating the problem is missing or inadequate.
D. Justifies design decisions using analyses based on appropriate engineering and/or scientific principles.	Applies principles correctly for major design decisions	No analysis of design decisions performed.
E. Considers, where appropriate, factors: public health; safety and welfare; global, cultural, social, environmental, and economic	Produces a specific list of constraints resulting from these factors	No listing of constraints provided
F. Supports the design process with appropriate engineering documentation and references.	All supporting documentation is created. Minor improvements in quality are possible	Significant pieces of supporting documentation are missing or of a poor quality

C. Communication (outcome 3)

Ability to communicate has been identified not only as a very important ability for engineering students, but also as one in which they are persistently underperforming [9]. The new ABET outcome

> *"an ability to communicate effectively with a range of audiences"*

was expanded from the old version by the addition of "range of audiences." Given that students are asked to demonstrate this ability in a variety of contexts and audiences, our rubric is fairly extensive so that we report only on the simplified version, as given below. The rubric focuses on two main areas: written reports and oral presentations.

- Written report rubric criteria include: Content, Organization, Vocabulary and Grammar, and Mechanics (formatting, spelling, proofreading)
- Presentation rubric criteria include: Organization, Visual Aids, Technical Content, Posture and Tone, Handling of Questions, and Effective Use of Time

This rubric can be applied in many project-based courses as they typically require written and/or oral communication. Overall, we have found that written communication is more challenging for students than presentation one. It is also clear from our experience and from research that writing cannot be handled in isolation [11]. Instead, it must be taught and practiced across curriculum. One frequent issue is that faculty find it difficult to evaluate writing and we hope that this rubric will help alleviate some of these concerns.

D. Ethics and professional responsibilities (outcome 4)

Three outcomes from the previous ABET list of outcomes are now combined into one:

> *"an ability to recognize ethical and professional responsibilities in engineering situations and make informed judgments, which must consider the impact of engineering solutions in global, economic, environmental, and societal contexts"*

Assessment of this outcome is difficult. For the time being, we have focused on the first portion dealing with ethics proper by defining appropriate criteria, as given in Table 3. The second part dealing with "informed judgment" is interspersed in performance indicators, such as "Understands personal, professional, and wider social consequences of violations of codes." In the future, we will attempt to address it more directly.

At this time, we have one 2-credit class that directly addresses this outcome, but we do not require students to take classes in ethics from other departments within our University. However, given the complexity of this new outcome this will be something to consider in our future curriculum development. Our initial evaluations indicate that students can easily identify ethical dilemmas related to safety, but they have difficulties in formulating more ambiguous dilemmas and identifying stakeholders. These results will lead to placing more emphasis on these items during class discussions and in assignments.

E. Teamwork and project management (outcome 5)

Teamwork was listed in previous versions of ABET student outcomes but the current description is expanded to include what we interpret to be project management:

> *"an ability to function effectively on a team whose members together provide leadership, create a collaborative and inclusive environment, establish goals, plan tasks, and meet objectives"*

In our newly developed teamwork and project management rubric, we use three performance criteria: Project planning, Project implementation, and Team functioning, as given in Table 4 . The same table lists a selection of performance indicators to illustrate the kinds of behaviors and artifacts that we use in the evaluation.

Table 3. Criteria and Performance Indicators for outcome 4 - Ethics.

Criteria	Proficient (3)	Beginning (1)
A. Recognition of ethical dilemma	Clearly identifying the type of ethical dilemma	Not recognizing a problem at all
	Focused, clear, and detailed framing of a dilemma, not irrelevant digressions	Unfocused, unclear, not coherent description of a dilemma
	Clearly identifies stakeholders (those affected by the dilemma)	Stakeholders not identified
B. Providing relevant information for the case at hand	Produces well crafted arguments based on new information; justifies assumptions and brings information from their own experiences	Copies arguments from sources but does not critically examine them. Ignores pertinent facts or uses misinformation.
	Identifies critical issues & components of the new knowledge	Gets lost in unimportant details
C. Analysis and fairness	Provides clear arguments	Arguments are not clear or coherent
	Considers multiple points of view in their analysis; global view with perspectives from e.g. employers, professions, and society	Takes only one point of view when analyzing a dilemma
D. Ethics in professions	Familiar with the value and importance of professional codes	Cannot identify even the basic components of professional codes
	Understands personal, professional, and wider social consequences of violations of codes	Unaware of consequences of violations of codes of ethics

Table 4. Criteria and Performance Indicators for outcome 5 - teamwork and project management.

Criteria	Proficient (3)	Beginning (3)
A. Project planning (documentation, timeline, decomposition, requirements, specifications)	Project proposal provides full graphical and textual documentation for timeline, functional decomposition, requirements, specifications	Superficial planning document
B. Project implementation (planning, tasks, deadlines)	Team meets regularly	Team meets irregularly
	Tasks are specific, detailed, have responsibilities assigned with deadlines, and define when a task is done (who, when, what)	Intermediate tasks are too broad, lack responsible person, deadline, and definition of "done"
	Activities follow a plan	Activities driven by external deadlines - most activity just prior to some event
C. Team functioning (structure, communication, spirit)	Member roles and responsibilities are clear and effectively executed	Member roles and responsibilities are not clear
	Team contract is well written and team members stick to it	Team does not produce a team contract

We have used this rubric in evaluating teamwork and project management in our cornerstone courses where we use Scrum-like project management [8]. This enables continuous monitoring and feedback to students, something that was missing from the more traditional ways of organizing project-based courses [16]. So far, teams are having more problems in criteria A. and B., especially when it comes to implementing their plans in a complex academic environment. We also use peer-evaluation software CATME [14] to evaluate teamwork and are in the process of correlating it with the results from our rubric.

F. Experimentation and data (outcome 6)

It is hard to imagine engineers not doing some form of experimentation and this criterion describes it:

> *"an ability to develop and conduct appropriate experimentation, analyze and interpret data, and use engineering judgment to draw conclusions"*

The second part dealing with engineering judgment is newly added. While there are many courses with associated labs, note that this criterion requires a development of experiments, which is more complex problem than simply conducting an experiment. Similarly, students have to be placed in situations where they have to exercise judgment instead of following instructions. This rubric is currently undergoing testing and revisions.

G. Life-long learning (outcome 7)

This outcome used to require only "recognition of the need for life-long" learning, but the new outcome

> *"an ability to acquire and apply new knowledge as needed, using appropriate learning strategies."*

clearly requires that students demonstrate that they can do this. In the past, some form of survey may have been sufficient, but we now have to consider different approaches. Our initial rubric for it is given in Table 5.

Table 5. Criteria and Performance Indicators for outcome 7 - life long learning.

Criteria	Proficient (3)	Beginning (1)
A. Acquire new information	Examines all the widely known sources, e.g. internet and library	It cannot be determined which sources are used
	uses only reliable sources	does not examine reliability of sources or uses unreliable ones
	information used is current or appropriate mix of old and current	there is no way to discern how current the used information is
B. Apply new knowledge	Produces well crafted arguments based on new information	Copies arguments from sources but does not critically examine them.
	Identifies critical issues & components of the new knowledge	Gets lost in unimportant details
C. Learning strategies	Determined from survey [13]	

In the development of this rubric we tried to capture the active part, i.e. students doing things that require acquisition of new knowledge, followed by appropriate use of such knowledge. This was inspired by work done in [12], which seeks to evaluate whether information presented on, e.g., a website, is credible and valid. Their CRAP Test looks at four major areas: currency, reliability, authority and purpose. We now require students apply CRAP test to any presentation that involves research, such as analyzing a recent engineering ethics case.

In addition, we are trying to determine if students are deploying appropriate learning strategies. We are currently in the process of collecting data by using a published survey [13] and by using our rubrics in Table 5. Depending on our findings, we may deploy interventions addressing students' metacognition, i.e. improving their understanding of their own learning.

IV. Conclusions and Future Work

Program outcomes assessment is a critical component of accreditation. Recent changes in ABET criteria prompted us to reconsider how we perform student outcomes assessment. One critical part of our new approach to assessment is the development of general rubrics for each ABET outcome. Each rubric has several performance criteria and each criterion has several performance indicators. Not all criteria will be used in every course, but they can be adjusted to specific course requirements.

Initial implementation was done during 2018-19 academic year and we are currently examining possible improvements to rubrics and how we use them. Even in this initial form rubrics have been useful in identifying some specific issues at course and program level, such as placing different emphasis during ethics instruction, and clarifying the role of requirements and constraints in the engineering design process.

Faculty need training in application of rubrics and this is an on-going process. We hope that utilization of rubrics across courses in our programs will result in better uniformity and reliability of our assessment. We also hope that other programs will find this approach useful and are looking forward to exchanging information on successes as well as difficulties in implementing it.

References

[1] National Academy of Sciences, "Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future," National Academy of Sciences, Washington DC, 2007.

[2] NSF call for proposals, "IUSE/Professional Formation of Engineers: Revolutionizing Engineering Departments (IUSE/PFE: RED)," https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505105

[3] ABET Criteria for Accrediting Engineering Programs, [Online]. Available: https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2019-2020/, [Accessed Feb. 3, 2020].

[4] R. M. Felder and R. Brent, "Designing and Teaching Courses to Satisfy the ABET Engineering Criteria," Journal of Engineering Education, vol. 92, no. 1, pp. 7–25, Jan. 2003, doi: 10.1002/j.2168-9830.2003.tb00734.x.

[5] J. F. Volkwein, L. R. Lattuca, P. T. Terenzini, L. C. Strauss, and J. Sukhbaatar, "Engineering Change: A Study of the Impact of EC2000," International Journal Of Engineering Education, vol. 20, no. 3, pp. 318–328, 2004.

[6] L. R. Lattuca, P. T. Terenzini, and J. F. Volkwein, "Engineering Change: A Study of the Impact of EC2000 - summary," 2006. [Online]. Available: https://www.abet.org/wp-content/uploads/2015/04/EngineeringChange-executive-summary.pdf. [Accessed: 08-Mar-2020].

[7] D.D. Stevens and A.J. Levi, "Introduction to Rubrics," 2nd ed., Sterling, VA: Stylus Publishing, 2013. ISBN-13: 978-1579225889

[8] B. Pejcinovic, "Introducing Cornerstone Courses into Electrical and Computer Engineering Curriculum," in 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2019, pp. 1427–1432, doi: 10.23919/MIPRO.2019.8756847.

[9] P. Rogers and R. J. Freuler, "The 'T-Shaped' Engineer," presented at the 2015 ASEE Annual Conference & Exposition, 2015, pp. 26.1507.1-26.1507.18, Accessed: 04-Apr-2019. [Online]. Available: https://peer.asee.org/the-t-shaped-engineer. DOI 10.18260/p.24844

[10] M. Faust, A. Greenberg, and B. Pejcinovic, "Redesign of senior capstone program in electrical and computer engineering and its assessment," in *2012 Frontiers in Education Conference Proceedings*, 2012, DOI: 10.1109/FIE.2012.6462390.

[11] [1]S. Conrad, T. J. Pfeiffer, and T. Szymoniak, "Preparing Students for Writing in Civil Engineering Practice," presented at the ASEE Annual Conference, 2012.

[12] M. Beestrum and K. Orenic, "The CRAP test," [Online]. Available: http://loex2008collaborate.pbworks.com/w/page/18686701/The%20CRAP%20Test, [Accessed: 24-Mar-2020]

[13] J. C. Chen, S. M. Lord, and K. J. Mcgaughey, "Assessing Engineering Students as Lifelong Learners," International Journal of Engineering Education, vol. 30, no. 5, pp. 1063–1071, Oct. 2014.

[14] M.W. Ohland, M.L. Loughry, D.J. Woehr, C.J. Finelli, L.G. Bullard, R.M. Felder, R.A. Layton, H.R. Pomeranz, and D.G. Schmucker, "The comprehensive assessment of team member effectiveness: Development of a behaviorally anchored rating scale for self and peer evaluation," *Academy of Management Learning & Education*, vo. 11, no. 4, pp. 609-630, 2012.

[15] P. Wong, M. Holtzman, B. Pejcinovic, and M. Chrzanowska-Jeske, "Redesign of Freshman Electrical Engineering Courses for Improved Motivation and Early Introduction of Design," in *ASEE Annual Conference and Exhibition*, Vancouver, Canada, 2011.

[16] B. Pejcinovic, P. Wong, and R. Bass, "Adapting Scrum Project Management to ECE Courses," presented at the 2019 ASEE Annual Conference & Exposition, Tampa, Florida, 2019.

Electronic Learning Experience Setup:
Power Electronics and Electrical Drive Education

P.J. van Duijsen*, D.C. Zuidervliet ** and J.B. Woudstra**
* Simulation Research, The Netherlands
** The Hague University of Applied Science, Delft, The Netherlands
info@dc-lab.org

***Abstract* - A method to teach and instruct the workings of power electronics and electrical drives is presented. The combination and interaction between oral lectures and practical laboratory assignments is presented, as well as the required tools. The structure of the course should consist of individual sections of increasing level regarding subject and details. For each subject the required knowledge level can be examined and a laboratory test has to be passed. Design tools and simulation/animation tools are used throughout the course. The laboratory assignments should follow the theory and simulation/animation is used to show the similarity and differences between theory and practice. A typical power electronics course is given as example and laboratory assignments using a dedicated educational trainer are given.**

Keywords – Education Power Electronics, DC grid, Laboratory, Design, Simulation, Animation, Hardware Trainer, Droop Control, Prosumers

I. INTRODUCTION

Teaching power electronics is usually done using textbooks and laboratory assignments. Simulations to enhance the understanding of the working of converters is used for over three decades[1] and animations to visualize the converter operation are also well known[2]. Remote laboratories became feasible with the widespread use of fast internet connections for students[3][4][5]. However all these "gadgets" are an add-on on the curriculum. Important is also how the structure of the curriculum and content is organized in order to teach students from basics up to a professional level. In this paper attention is given how the contents of power electronics courses are developed where there is a strong interaction between the oral lectures and the practical assignments in a laboratory. Especially the step wise building up of the level[6][7] of knowledge during the oral lectures and the practical assignments and the interactions between them will be highlighted in this paper by a number of examples.

To encourage students, special attention is given to the practical assignments on a hardware trainer, which is constructed in the same way as an industrial application[8][9][10]. It allows multiple assignments and includes many measurement connections for important signals. The combination of oral lectures, exercises with online verification, educational simulations/animations and practical assignments on a newly developed hardware trainer are the subject of this paper.

TABLE I. TYPICAL APPLIANCES AND POWER LEVEL

AC/ DC	Application & Power Level		
	Application	***Mobile***	***Grid***
DC-DC	Adapters[(a)] for mobile phone and computers are the most common seen devices	Low	Low
AC-DC	Powering various appliances, Heating, Cooking	Medium	Medium
DC-AC	Drive Applications; Domestic appliances, Fans, Cooling, Heating.	Low	Medium
AC& DC	Solar Power MMP and Grid connected inverters	Low	Medium
DC-AC	Inverters in Electric Mobility	Medium	
DC-AC	Inverters in Industrial Drives Traction & Wind Energy	High	High

a. AC-DC as well as DC-DC

The Laboratory assignment is built around the theory from the lectures and has to be practiced using simulation, online design tools (self-assessment) and measurements. For the Laboratory assignment a new universal trainer containing four half-bridges with shoot-through protection, adjustable dead-time and current measurement is developed. Digital control is an add-on using an Arduino or TI Launchpad.

The hardware trainer allows students to perform experiments with safe voltage levels below 60 volts and under 200 Watts. Applications range from battery charging, solar panel maximum power tracker, DC – DC conversion to DC – AC conversion for electrical drive speed and position control. The curriculum covers both electrical and mechatronic students on Bachelor level[11]. The main objective in developing a new power electronics course is to combine theory together with a design phase, component selection and calculations, prototype building, measurements and in a final stage the self-reflection of comparing the original design idea with the final obtained results.

II. WHY AND WHAT?

Before starting the discussion about typical content in a power electronics course, it is wise to first consider what is expected from the student to learn and why this is important[12][13]. The importance of the subject is something that has to be justified when building a curriculum for Bachelor and Master students.

A. *Why?*

Looking at the current discussion on climate change and the energy transition, taking place from fossil fuel towards green energy, it is easy to justify a course on power electronics. In table I some of the main applications that are of interest in our daily life are summarized. Pointing at the importance of each application will justify the existence of a power electronics course in the overall curriculum.

B. *What?*

What is it a student has to learn? It starts with an oral lecture where the basics of the theory should be explained. Once there is a thorough basic understanding of power electronics and the waveforms and operation is clear to the student, the component parameters can be calculated. Instead of showing only formulas, now the student is forced to use the formulas and to understand where the formulas are coming from. Based on the knowledge of calculating component parameters, the student can make a first design. Design tools and simulations are helpful in understanding the influence of parameters. Depending on the results from the simulation, the student can do a redesign. Building a prototype in the laboratory is of tremendous value to a student to get a grip on the difference between design, simulation and final measurement results. If building of a prototype would take too much time, a pre-build design can be used and the student can do measurements. The student now has to explain the difference between the calculated and simulation result and the measurements on the prototype.

III. COURSE STRUCTURE

A course should contain theory and laboratory assignments. Each topic should be assessed during the oral lecture and, if possible, in the laboratory.

A. *Theory*

Every oral lecture or self-assignment needs written text to explain the theory. Various other media can accompany it, as long as there is a story that can be told to the student explaining the matter he has to understand. Although webinars and other video tutorials are helpful, a written transcript helps the student to go through the material at his own pace. Learning is practicing, the student should have enough worked-out short problems or questions and answers to rehearse and practice the topics. Let us take the example of a buck converter. First the students learn the input-output relations and next they have to understand how to set the inductor current ripple by choosing an inductor value. By just explaining the voltage levels during switching, the voltage over the inductor can be derived from which with known operating frequency the inductance value can be obtained as

$$0 < t < T_{on} \Leftrightarrow L\frac{di_L}{dt} = U_{in} - 0 \qquad (1)$$

$$T_{on} < t < T_{off} \Leftrightarrow L\frac{di_L}{dt} = U_{in} - U_{out} \qquad (2)$$

In either case the student should be able to calculate the value of L for continuous conduction mode. To verify the result he could use a simulation, but a design tool might be faster to use. Figure 1. shows how a design tool[14] is used to verify the peak inductor current when a fixed inductor value of 150uH is selected.

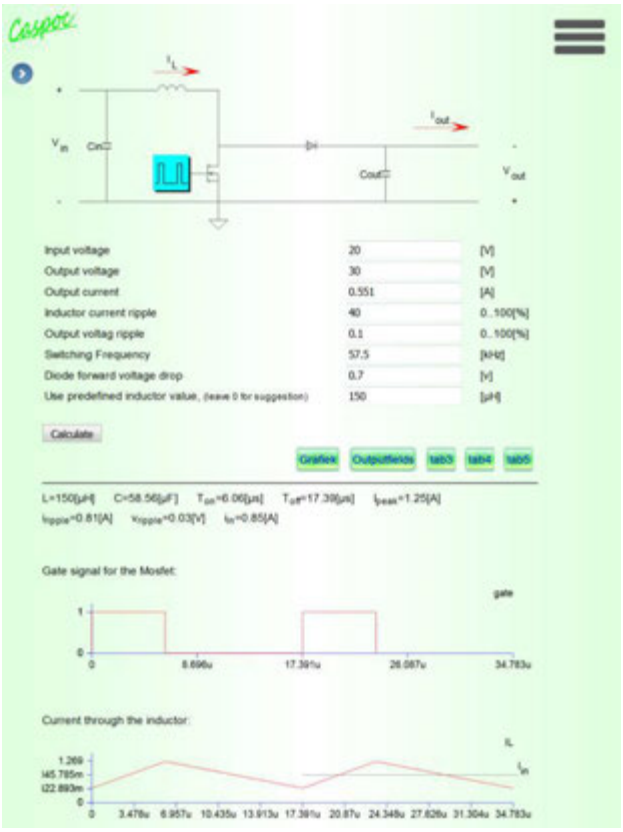


Figure 1. Design tool to prove the inductance value according to an assumed inductor current ripple.

By choosing a zero value for the inductor, the tool will suggest an inductor value to achieve the desired current ripple. This inductor current ripple value is relative to the average inductor current and is default set to 40%.

Using (1) and (2), the student can see the relation between the inductor current ripple and the value of the inductor and the value of the switching frequency. Using the design tool, he can directly see the influence graphically. Not only does it give a numerical insight, but also the waveform is directly available and shows that by decreasing the inductor value, the converter will eventually go from continuous mode into discontinuous mode.

By using the design tool, the student can get a better feeling of the relations between the component parameters and the current and voltage waveforms. Since the waveforms are presented in their actual voltage and current scale, the student directly will get an impression of typical component values. He will get more insight in the approximate size of the components for a typical operating frequency. This will be of help when going to the next level in the course, where he has to test the operation of the converter in a simulation to verify the waveforms.

The results from the design tool are directly applicable in a simulation/animation of the circuit[14]. Using the values like the operating frequency, inductor and output capacitor, he can directly set up a simulation/animation of the circuit and will very probably get nearly the same waveforms as presented by the online tool. This save the student time for trial and error to choose component values, run a simulation and based on the waveform, make changes to the component values.

The simulation/animation is primarily needed to show the operation of the circuit and especially the current paths when, for example in a buck converter, the switch is closed and when the freewheeling diode is conducting.

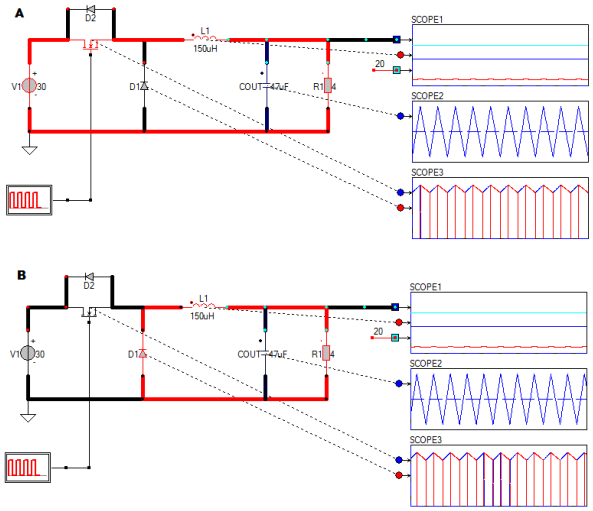


Figure 2. The current path when the Mosfet is turned-on(A) and when the freewheeling diode is conducting(B).

In Figure 2. The current path in the circuit is colored according the current level during the animation. In this way the student sees the current path periodically changing between Mosfet and diode and will see that the freewheeling always follows the on-time of the Mosfet.

By decreasing the inductor value in the animation of Figure 2, the inductor current will eventually get discontinuous. In the animation the student will see that in that case the current path through the freewheeling diode will gradually change from red(maximum current) to black(no current). If the animation was recorded using the Freeze and Go-Back function [1], the student can step through the animation and see the position in time on the actual waveform, the current path in the circuit as well as the value of each current, voltage and control signal in the animation. Stepping through the animation, the student can observe the changing of currents and voltages and how this affects the operation of the circuit. Using the Freeze and Go-Back feature he can trace signals and relate them to the switching operation in detail.

B. Laboratory

A laboratory should start with insight in the basic operation of the circuit that is going to be build or examined. Simulation is required here to quickly look at the operation of all sections and components and to give the student a basic understanding on the size of components and the to be expected current and voltage ranges in his circuit. Building a circuit in practice is an experience, students should go through in order to understand that in practice things behave different from what is expected from theory. Especially the influence of parasitic elements and thermal behavior becomes clear during a laboratory exercise. Most importantly at this stage, the students will see that the measured voltage and current waveforms differ from what is expected from theory.

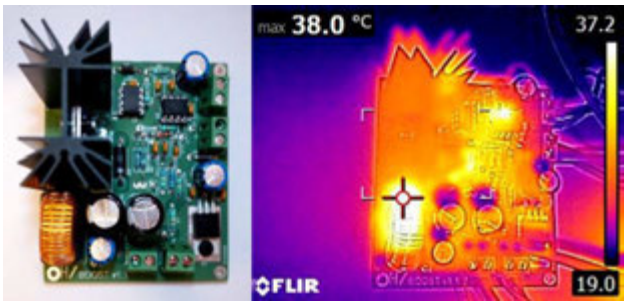


Figure 3. On the left the assembled boost converter and on the right a thermal image indicating severe losses in the inductor .

Figure 3. shows an assembled current mode controlled Boost converter and the thermal image when operating. The students can observe the losses in the inductor core material and this will give raise to thinking about core selection.

Through hole components are preferred during the first laboratory experiments, as they are easier to handle for inexperienced students, but also allows easier probing on the PCB during testing. The control IC is mounted using a socket that allows different circuit operations when the IC is not mounted. For example, the duty cycle for the boost converter can be injected using an external signal generator when the current mode control IC is removed. For newer components, an IC Socket Adapter to use SOP or similar SMD packages on a through hole PCB can be used.

The assembly of the PCB should be done step-wise and each step should be accompanied with measurements to verify the operation. For the boost converter in figure 3, the power circuit consisting of the inductor, Mosfet with gate driver and output capacitance is build first and the duty cycle is set using an external gate driver. The students can measure the current and voltage waveforms for varying switching frequency and duty cycle. Needles to say that an ohmic load and a power supply with maximum current protection have to be available at this stage.

Reporting of measurement results can be practiced during the laboratory assignment. It will encourage students to document the complete set up and wiring of several experiments in order to be able to reproduce measurement results.

IV. KNOWLEDGE LEVEL AND INTERACTION

The theory and practical assignments have to be organized in parallel during the course. The student will gain knowledge during the lectures and during the assignments and should preferably have theoretical knowledge about a typical topic before entering the

practical assignment. Figure 4. shows a typical gain in knowledge and the interaction between the lectures and the laboratory assignments for a power electronics course on switched mode power supplies.

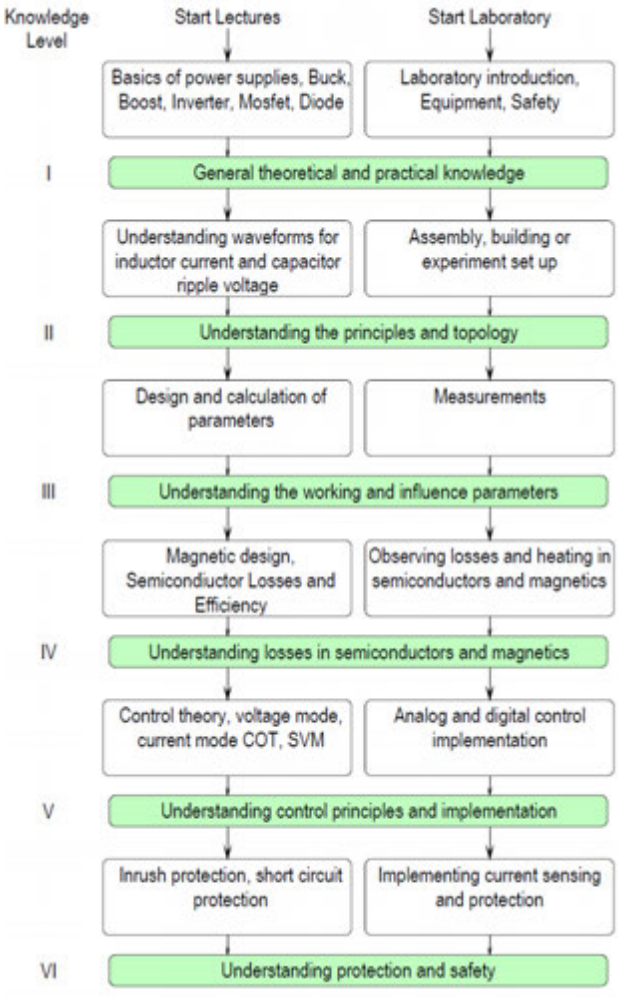


Figure 4. Knowledge level and interaction between theory and laboratory assignments for a typical power electronics course on switched mode power supplies.
(COT = Constant-On-Time, SVM = Space Vector Modulation)

The knowledge level in Figure 4. is displayed on the left side next to the achieved knowledge in the shaded box. Examination in the form of a written or oral examination or delivered measurement report can be used to follow the progress of each student individually. Continuing with the next knowledge level should be based on the fulfillment of the previous level. The lecture notes and tools are required in each stage. For example the design tool from Figure 1. is used during the knowledge level II and III for understanding the principle operation, but also for design and component values. Animation as shown in Figure 2. is typically used at knowledge level I where the basics of operation are explained. Specific design tools, for example for calculating the thermal impedance of heat sinks are again applied at knowledge level IV where the students observe the heating of components such as the losses in inductors as shown in Figure 3.

V. APPLICATIONS

The knowledge gained during the power electronics courses is important for other courses. Gained theoretical knowledge on inverter operation and modulation is of importance when starting a course on electric motor drives. Without prior knowledge on the operation of inverters a course on variable speed drives would need extra hours explaining the working of pulse width modulation and the control of currents through the motor windings. Assembled converters can be applied in other laboratory assignments, especially in electric motor control assignments or as interface to components such as solar panels and batteries. In this way the gained knowledge from the power electronics course as well as the laboratory assignments form a basis for other courses like mechatronics, electric drives, AC power systems and DC grid power systems like DC grid Droop Control and Solar Power applications.

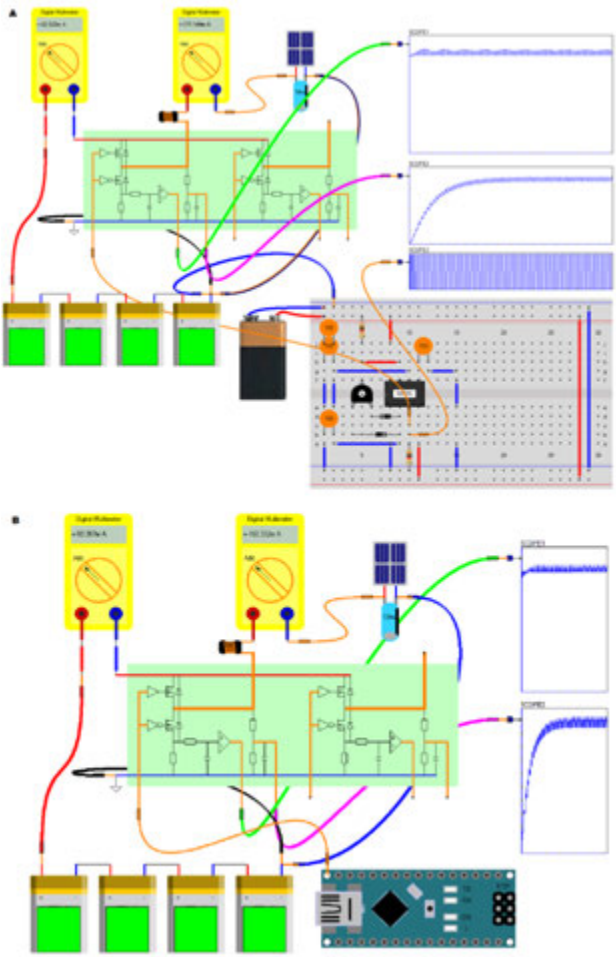


Figure 5. Charging Li-Ion batteries from a solar panel using and analog control(A) or digital control(B) and the Universal Four Leg inverter[10].

A typical application is the use of a boost converter to load a solar panel[10]. Figure 5 shows a set up where a pre-assembled inverter is used as a boost converter for loading a solar panel and charging 4 Li-Ion batteries.

VI. EXAMPLES

Typical examples during a power electronics course are DCDC and DCAC converters. There are many different types of possible examples that could be used during the course. At the power electronics courses at the THUAS there are two main converters that are assembled/measured during the practical laboratory assignments. The Boost converter is build during the lectures on DCDC converters and an inverter circuit is used in practical laboratory assignments to study the various modulation techniques used in DCAC applications[9] and DC grid droop control.

A. DCDC Boost Converter

After preliminary design, students have to assemble a boost on a PCB, see Figure 3. They start with a basic power circuit and the aim is to teach the students the influence of switching frequency and duty cycle in a Boost converter. A signal generator is used and controls the gate of the Mosfet in the boost converter. Students have to perform experiments for various switching frequencies and duty cycle and report current and voltage waveforms. During the oral lectures they have learned the basics of operation along with the waveforms. Using the design tool from Figure 1. and the animation and simulation from Figure 6. Here they are faced with the fact that from theory and during simulation it is easy to obtain the waveform for inductor current and voltage across the inductor. However, to achieve the same in a measurement, they have to improvise and measure the converter input current as well as the use of a differential probe for measuring the voltage across a component.

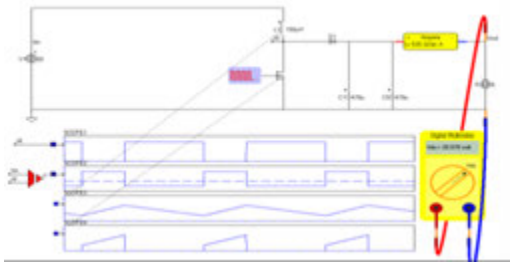


Figure 6. Understand the basic relation between the voltage across the inductor (V(Vin)-V(J8)) and the inductor current.

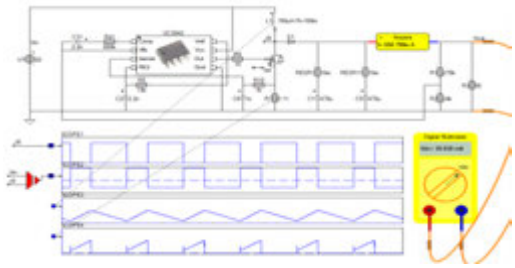


Figure 7. Closing the control loop using an industry-standard current-mode control IC.

From theory the current mode control can be implemented in a simulation and the students can start with designing the main components required to get a stable output voltage. Once they have built the Boost converter in the laboratory and start measuring, they have to explain the differences between the simulation results and the measurement results.

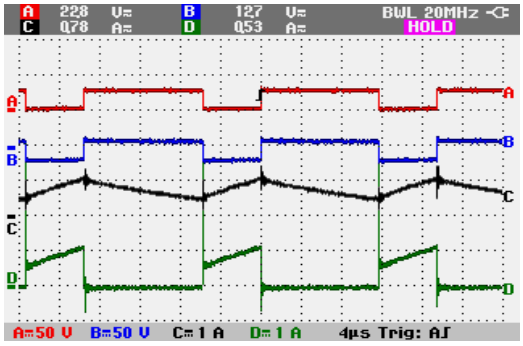


Figure 8. (A)Voltage over the Mosfet, (B)differential voltage across the inductor, (C)inductor current measured over shunt at the input, (D)current measured as voltage over the sense resistor R2.

From the result from Figure 8. the students have to explain the origin of the spikes on the current measurement. By adding parasitic components in the simulation they will step by step identify the most important parasitic components in the circuit, Figure 9.

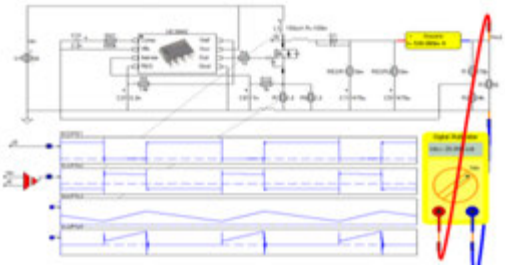


Figure 9: Parasitic components added to the simulation

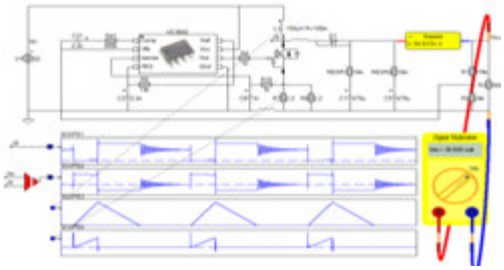


Figure 10. Ringing on the Mosfet's drain voltage because of parasitic loop inductance during discontinuous operation.

The ringing on the node where Mosfet, inductor and diode intersect is visible in the simulation. Measurements like in Figure 11. should confirm this.

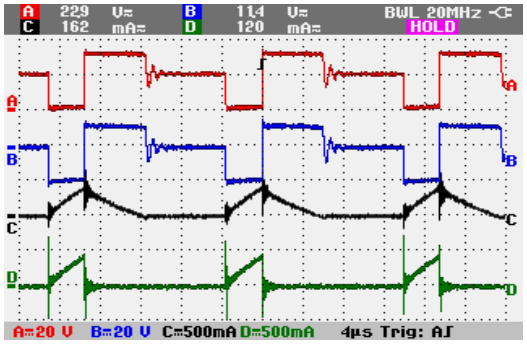


Figure 11. Discontinuous current causes ringing across the Mosfet. From top to bottom, voltage over the Mosfet, differential voltage across the inductor, inductor current measured over shunt at the input, current measured as voltage over the sense resistor R2.

B. DCAC inverter

The DCAC converter is already pre-assembled, as the PCB contains many SMD components that require too much time from the students. Also the goal of he laboratory assignments is to get a basic understanding on the working of the inverter and the modulation principles.

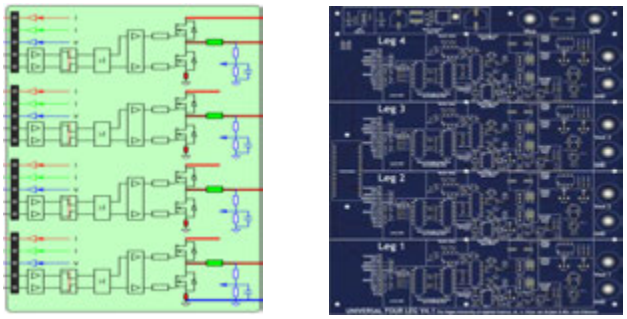


Figure 12. CASPOC Simulation model and hardware of the Universal Four Leg (U4L).

The students will use a pre-assembled Universal Four Leg[10]. In Figure 12 we see the simulation model of the U4L with the PCB of the hardware of the U4L. An additional analog control will be added for three phase modulation that can be connected to the U4L, as shown in Figure 13. The students will have to solder this analog circuit first before they can test it in combination with the U4L. This circuit can be used to modulate the input of three legs of the U4L for controlling an inductive load like a three phase motor.

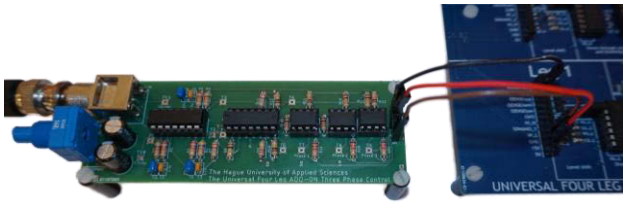


Figure 13. On the left the hardware for three phase modulation with one phase connected to Leg 1 of the U4L to control an inductive load.

The students can feed in their own triangular high frequency carrier wave and measure the output frequency of the three phase oscillator. The output signals of this circuit combined with the U4L can be seen in Figure 14. The three phase output signal should confirm the simulated results.

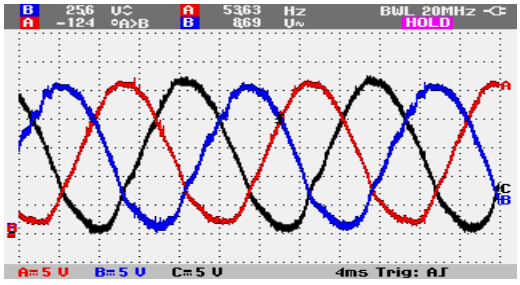


Figure 14. Three phase output currents in an inductive load created using the Universal Four Leg with analog control.

C. DC Grid Droop Control

The U4L is also used to study DC Grid Droop Control. The outputs of the U4L can regulate a bidirectional current flow, depending on the DC voltage level on their output terminal. Using the U4L students can model DC Grid prosumers like batteries, solar MPP controllers, non-linear loads and grid connections[11].

ACKNOWLEDGMENT

Thanks to Casper Grootes for providing the three phase analog control circuit and output current measurements.

VII. CONLUSION

Combining oral lectures including written lecture notes, design tools, simulation and animations tools with laboratory assignments enriches power electronics courses. The structure and knowledge levels in the course contents are given which are helpful in the progress of the students and allow intermediate examination and or passing of laboratory assignments. Dedicated educational trainers like the U4L enrich the laboratory assignments and can be applied to other courses like mechatronics and electric drives as well.

REFERENCES

[1] P. Bauer, and P.J. van Duijsen, Challenges and Advances in Simulation. IEEE 36th Conference on Power Electronics Specialists, 2005. doi:10.1109/pesc.2005.1581755

[2] P. Bauer & J.W. Kolar (2003) Teaching Power Electronics in the 21st Century, EPE Journal, 13:4, 43-50, DOI: 10.1080/09398368.2003.11463546

[3] P. Bauer and V. Fedák, e-Learning in Electrical (Power) Engineering, from Idea to Reality: State of the Art and Challenges. Article, January 2006.

[4] D. Lascu, M.Lascu and M. I. Băbăiță, E-Learning Practical Teaching of Resonant AC Circuits. Journal of Electrical and Electronics Engineering, May 2009.

[5] D. Lascu, M. I. Băbăiţă, P. Bauer and M. Lascu, Distance Education in Soft-Switching Inverters. Article in Journal of power electronics · November 2010.

[6] F. Blaabjerg, A Power Electronics and Drives Curriculum with Project-Oriented and Problem-Based Learning: A Dynamic Teaching Approach for the Future. Journal of Power Electronics, Vol. 2, No. 4, October 2002.

[7] Zhang, Z., Hansen, C. T., & Andersen, M. A. E. (2016). Teaching Power Electronics with a Design-Oriented and Project-Based Learning Method at the Technical University of Denmark. I E E E Transactions on Education, 59(1), 32-38. https://doi.org/10.1109/TE.2015.2426674

[8] J.B. Woudstra, P.J. van Duijsen, P. van Willigenburg and P.M. Witte, DC Educational Development Improving Understanding of DC to DC / DC to AC Conversion. 46th SEFI Annual Conference, 2018 Copenhage. ISBN: 978-2-87352-016-8

[9] P.J. van Duijsen, J.B. Woudstra and P. van Willigenburg, Educational setup for Power Electronics and IoT. IEEE 19th International Conference on Research and Education in Mechatronics, 2018. doi: 10.1109/REM.2018.8421802

[10] P.J. van Duijsen, D.C. Zuidervliet and C.F.M. Dirksen, Enhancing Laboratory Learning Experience A New Experimental Setup For Power Electronics And Electrical Drive Education. 47th SEFI Annual Conference, 2019 Budapest. ISBN: 978-2-87352-018-2

[11] DC-LAB Researchgroup, www.dc-lab.org.

[12] V. Fernao Pires, A. J. Pires and O. P. Dias, Self-Learning as a Tool for Teaching Power Electronics. 5th IEEE ICELIE, November 2011. DOI: 10.1109/ICELIE.2011.6130035

[13] E.A. McShane, M. Trivedi and K. Shenai, An Improved Approach to Application-Specific Power Electronics Education—Curriculum Development. IEEE TRANSACTIONS ON EDUCATION, VOL. 44, NO. 3, AUGUST 2001

[14] CASPOC Simulation Research, www.caspoc.com.

Study of Switching Forward Single-ended DC/DC Converter in the Course "Power Supplies"

S. Zaharieva*, I. Stoev *, A. Borodzheva** and S. Stoyanov***
* University of Ruse "Angel Kanchev"/ Electronics, Ruse, Bulgaria
** University of Ruse "Angel Kanchev"/ Telecommunications, Ruse, Bulgaria
*** Technical University of Varna/ Dobrudza Technological College, Dobrich, Bulgaria
szaharieva@uni-ruse.bg, istoev@uni-ruse.bg, aborodzhieva@uni-ruse.bg, svilen.stoyanov@tu-varna.bg

Abstract – **The proposed paper presents active learning methods using a laboratory module "Switching Forward Single-ended DC/DC Converter" during a laboratory exercise in the course "Power supplies" at the University of Ruse „Angel Kanchev". The „Switching Forward Single-ended DC/DC Converter" laboratory module enables students to solidify their knowledge, skills and experience in switching power supplies. During the class, students record basic timing diagrams of input and output voltages, input current, oscillator voltage, control pulses for the power switch, pulse width modulation, and the positive impact of protective circuits. From practical experiments, students receive results for load characteristics with and without feedback from the switching forward single-ended DC/DC converter.**

Keywords - active learning methods, DC/DC converter, load characteristics, protective circuits, PWM

I. INTRODUCTION

Nowadays, businesses require their staff to be adequate to the fast-paced "Industry 4.0" standard. In this regard, in the training of students in engineering specialties, it is necessary for them to acquire knowledge and experience through which they can successfully solve practical problems. An important aspect of an effective educational process is the application of different methods of active learning. The application of a research approach in the laboratory-assisted learning process develops communication, collaboration, problem-solving skills, adaptive and critical thinking, creativity and innovation.

The aim of the paper is to present active learning methods for studying switching forward straight single ended DC/DC converter used in the course "Power Supplies" in the specialty "Electronics" at the University of Ruse "Angel Kanchev".

The switching forward single-ended DC/DC converter laboratory module enables students to solidify their knowledge, skills, and hands-on experience in switching power supplies.

During the classes, students record basic timing diagrams of input and output voltages, input current, oscillator voltage, control pulses for the power switch, pulse width modulation (PWM) and protection circuits.

From practical experiments, students receive results for load characteristics with and without feedback from the switching forward single-ended DC/DC converter [1, 4, 6, 7].

The article presents results from Project 2020 - FNI - FEEA - 02, "Research and development of specialized electronic industrial process control systems based on statistical analysis" with financial support of the National Science Fund of University of Ruse.

II. TRAINING TECHNOLOGY IN THE COURSE POWER SUPPLIES

The fast development of electronics has a strong stimulating effect on development of power supplies, creating new requirements. As a result, the course "Power Supplies" is developing extremely dynamically, both in terms of electronic components and materials, and in terms of the principles of operation. The development of uninterruptible power supply systems creates new technical problems and complicates old ones. The technological process carried out in the production of new semiconductor devices and magnetic materials has made it possible to apply new control principles, such as pulse-width modulation and to reduce the volume and weight of the power supplies.

The course "Power Supplies" is a mandatory fundamental course in the bachelor's program in Electronics at the University of Ruse "Angel Kanchev". The course enables third-year students to master the principles of operation and ways of designing basic circuits in power supplies, such as transformer, rectifier, multiplier, filter, DC voltage and continuous current, as well as switching converters.

Attention is paid to modern methods for design and operation of a wide range of rectifier circuits and devices. The methods for synthesis and analysis of different types of rectifier circuits are considered.

The course is directly related to the disciplines "Analog Circuits", "Signals and Systems", "Electromechanical Devices", "Conversion Technology", etc.

The practical exercises enhance the knowledge from the lecture course and thus acquire valuable practical skills for working with basic corrective and stabilizing schematics of power supplies.

As a result of the training in the discipline "Power Supplies" students acquire:

- Extended and in-depth knowledge of design and construction of power supplies;
- Applying logical thinking and showing innovation and creativity in solving specific tasks related to the design and development of electronic equipment;
- Skills to design and maintain electronic equipment for power supplies of various electronic equipment.

The training in the discipline is carried out through lectures, practical exercises and controlled extracurricular work, including the development of a course project.

From the lecture material students get acquainted theoretically with the main issues, methods and tools for operation of rectifiers. The practical exercises are conducted with the participation of the students and the topics are directly related to the studied lecture material. The students are formed as separate working groups of four and perform measurements on laboratory modules (Fig. 1).

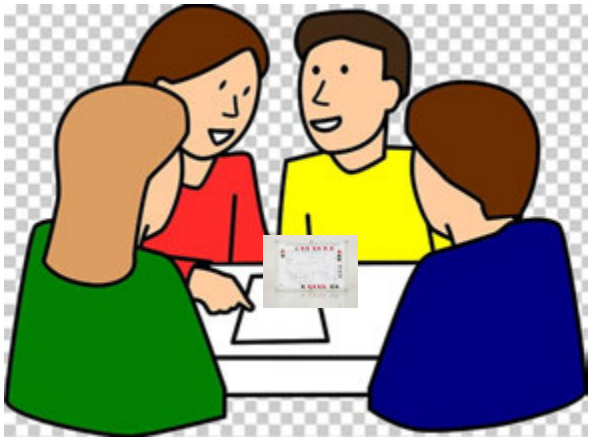


Figure 1. Working groups performing measurements on laboratory models

III. SWITCHING FORWARD SINGLE-ENDED DC/DC CONVERTER

A. Advantages and disadvantages

Switching voltage regulators operate on the principle of PWM (pulse width modulation), as a result of that, they have the following advantages [2, 3, 5]:

- High efficiency up to 78% due to the key nature of the operation of the control element;
- Low sensitivity to temperature changes of the environment;
- Small dimensions and mass.

The disadvantages are:

- Lower coefficient of stabilization;
- Finds application for smaller power consumption;
- The switching mode requires the use of input and output filters, which reduces the dynamics of the regulation and prevents the introduction of radio-interference in the main power supply network.

B. Structural and principle electrical schematic of the developed laboratory module

Fig. 2 shows the structural diagram of the switching straight single-ended DC/DC converter and Fig. 3 shows the principle electrical schematic of the developed laboratory module.

Description of the individual blocks in Fig. 3 of the switching straight single-ended DC/DC converter:

- Input DC voltage – external power supply (12V, 1A);
- Input filter (C8) – filters interference from external power supply;
- Transformer (TR) – with primary, demagnetizing and secondary coils with ratio of number of turns w1 = w2 = w3 (1:1:1);
- Rectifier (D3) and LC filter (L1, C10) – for smoothing the output voltage ripples;
- Linear stabilizer (LM7809) – power source for the control block;
- Control block (TDA1060) – an integrated circuit that controls the power switch (VT1) and elements by which its operating parameters are set – operating frequency (R4, C4), maximum fill;
- Maximum fill factor, feedback depth, soft-start time after switching off (R1, R2, C3);
- Power switch (VT1) – MOS transistor;
- Overvoltage protection (C6, D2, R8) – protection the power switch;
- Demagnetizing circuit (demagnetizing coil, VD1) – demagnetizing the transformer core;
- Current protection (R7, R6, C1) – protection when the source current of the transistor exceeds the set values;
- Negative feedback (R10, R9, C7, R3, C2) – maintains the constant value of the output voltage. Part of it, through a divider, is fed to the input of the control block;
- Output Load RL – necessary variable load for setting of the desired output current of the DC/DC converter required for the study.

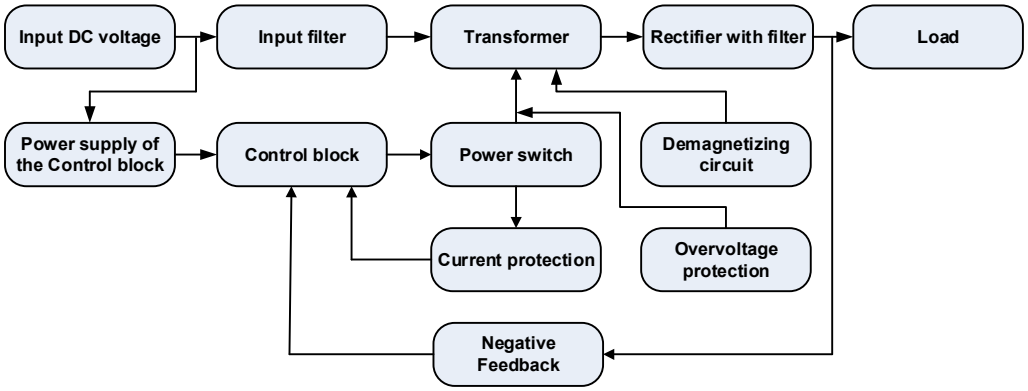


Figure 2. Structural diagram of the switching straight single-ended DC/DC converter

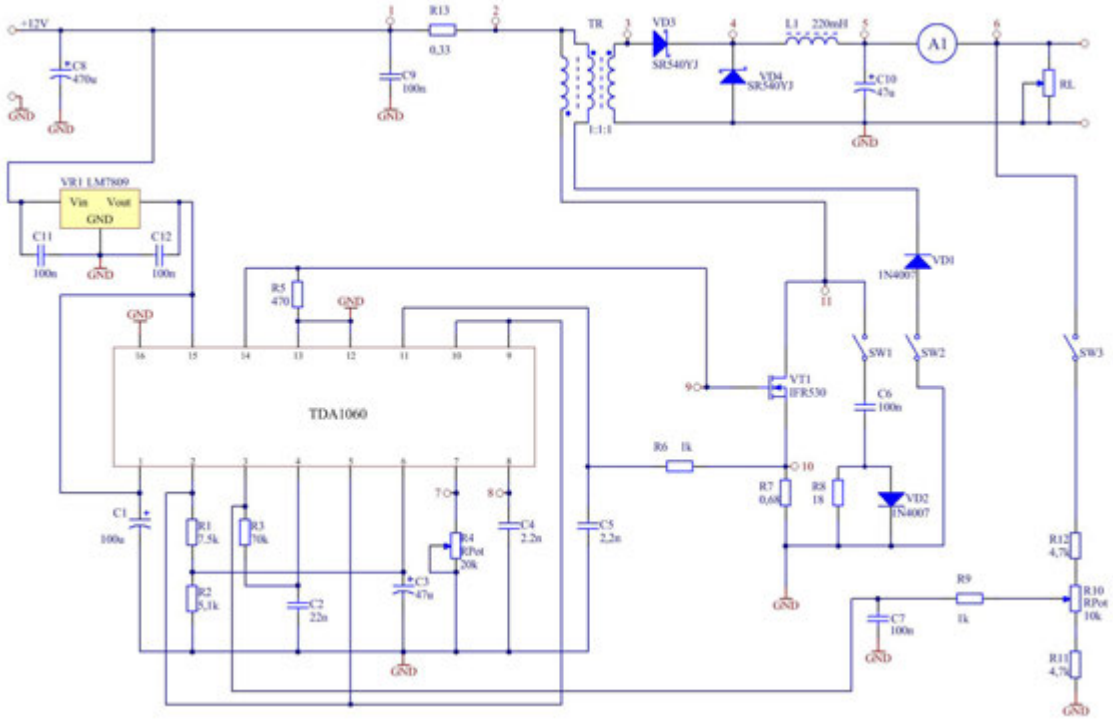


Figure 3. Principle electrical schematic of the switching straight single-ended DC/DC converter - laboratory module

IV. METHODOLOGY AND RESULTS

A. *Methodology of assigned tasks*

During laboratory exercises, students are assigned tasks to the following methodology:

Task 1: Record the time-diagrams of the input current, such as a voltage drop on the sensing resistor R13 (1, 2);

Task 2: Record the voltage diagrams of the sawtooth wave generator of the control block (7.0 and 8.0);

Task 3: Record the time-diagrams of the output of the control block (9.0);

Task 4: Record the voltage U_{DS} time-diagrams of the power switch VT1 with demagnetizing circuit and overcurrent protection (11.0). Measurements should be made with switches SW1 and SW2 in ON-state;

Task 5: Record the time-diagrams of the transformer secondary coil (3,0);

Task 6: Record the time-diagrams of the voltage of the reverse diode VD4 (4,0);

Task 7: Record the time-diagrams of the voltage of the inductor L1 of the output filter (4,5);

Task 8: Record the voltage U_{DS} time-diagrams of the VT1 power switch with disconnected demagnetizing

circuit and overvoltage protection (11.0). The measurement should be performed with switches SW1 and SW2 in state OFF;

Task 9: Record the voltage U_{DS} time-diagrams of the VT1 power switch with connected demagnetizing circuit and disconnected overvoltage protection (11.0). The measurement should be performed with switches SW1 in state ON and SW2 in state OFF.

Task 10: Record the voltage U_{DS} time-diagrams of the VT1 power switch with disconnected demagnetizing circuit and connected overvoltage protection (11.0). The measurement should be performed with switches SW1 in state OFF and SW2 in state ON.

Task 11: Take down the external characteristic U0 = f(I0) with disconnected and connected negative voltage feedback circuit using the SW3 switch. The ammeter connected between control points 5 and 6 measures the current consumed by the load. The output voltage is measured using a voltmeter connected between control points 6 and GND. Fill in TABLE I with the measured values.

TABLE I. U0 = F(I0), WITH AND WITHOUT NEGATIVE FEEDBACK

I_0, A	0,35	0,6	0,9	1,2	1,5
U_{0woNFB}, V					
$U_{0withNFB}$, V					

Note: For each task draw conclusions from the study and make an analysis of the results obtained.

B. Results from tasks

Results from Task 1:

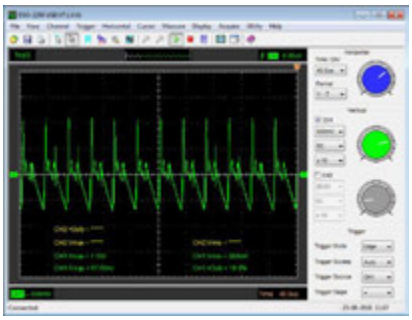


Figure 4. Voltage waveform on signal resistor R13

During the first time interval, when the power switch VT1 is released, the input DC voltage Uin is applied to the transformer. Current flows through the signal resistor R13, the transformer primary coil and the VT1 power switch. The VD1 and VD4 diodes are closed and VD3 is opened. Energy is transferred from the input source to the load.

From the voltage drop on R13 we could judge the magnitude and the rate of increase of the magnetizing current of the transformer iμ.

Results from Task 2:

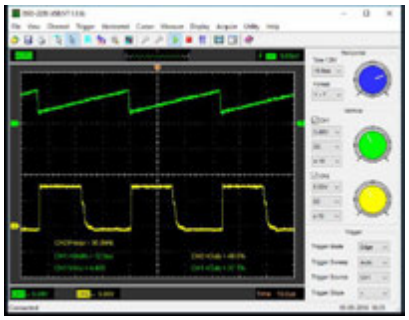


Figure 5. Voltage waveform of sawtooth wave generator of control block

The voltage on the capacitor C4 is of a saw-tooth shape and the voltage on the resistor R4 is of a pulse shape. These elements determine the frequency of operation of the integrated sawtooth wave generator in the control block.

Results from Task 3:

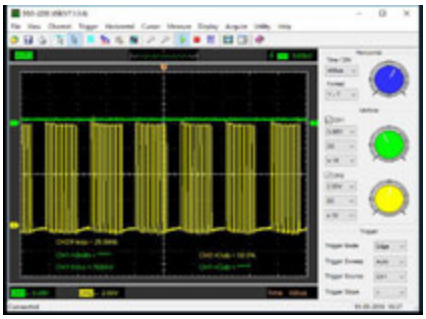


Figure 6. Waveform of output voltage of the control block

The control block generates packets of pulses and feeds them to the power switch VT1. Changing the pulse duration (Pulse Width Modulation) changes the value of the output constant voltage of the converter. Negative feedback supplies some of the output voltage to the control block. It adjusts the pulse duration to stabilize the value of the output voltage to the load.

Results from Task 4:

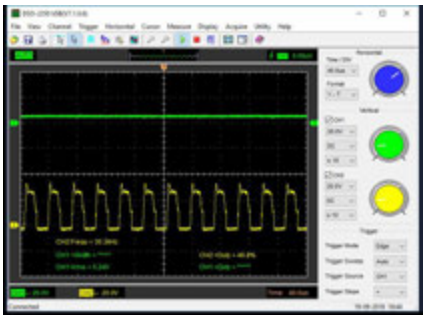


Figure 7. U_{DS} voltage waveform of the power switch VT1 with connected demagnetizing circuit and overvoltage protection

The pulses at the output of the power switch have twice the amplitude of the input DC voltage.

Results from Task 5:

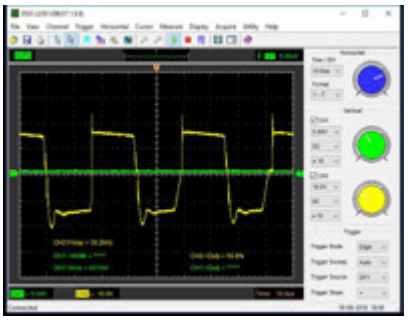


Figure 8. Voltage waveform of the secondary coil of the transformer

The voltage of the secondary coil is bipolar rectangular pulses with an amplitude corresponding to the rectified DC voltage of the converter. Positive pulses open the VD3 rectifier diode and negative impulses close it.

Results from Task 6:

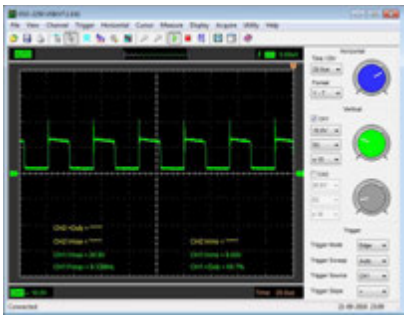


Figure 9. Voltage waveform of the reverse diode VD4

When the VD4 diode is closed, current flows from the transformer secondary coil, the rectifier diode VD3, and through the inductor L1 to the load.

When the VD4 diode is opened, the stored energy in the L1 inductor supplies current to the load.

Results from Task 7:

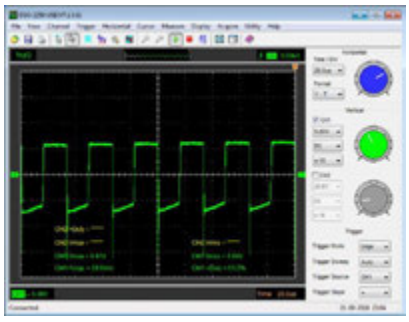


Figure 10. Waveform of the output filter inductor L1

The top of the waveform shows the time during which pulsed current from the rectifier diode VD3 flows through the inductor L1 to the load. As a result of this current, the core of the L1 inductor is magnetized. The lower part of the oscilloscope shows the energy stored in the L1 inductor, which is emitted into the load. The pulses have a duration equal to the VD4 diode opening time.

Results from Task 8:

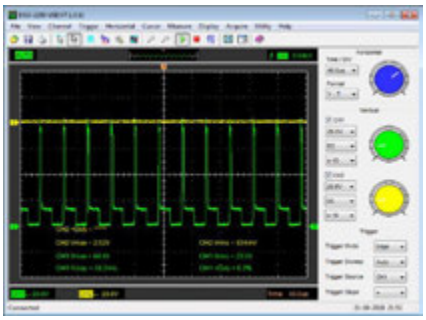


Figure 11. Waveform of U_{DS} voltage of the power switch VT1 with disconnected demagnetizing circuit and overvoltage protection

The oscilloscope shows peaks due to the self-induction of the voltage in the primary coil.

Results from Task 9:



Figure 12. Waveform of U_{DS} voltage time diagrams of the power switch VT1 with connected demagnetizing circuit and disconnected overvoltage protection

The oscilloscope shows peaks of 2Uin amplitude due to the dilution of the energy loaded into the primary coil with a duration equal to the time of the release of diode VD1. The transformer magnetizing current now flows through the demagnetizing coil and decreases by a slope equal to the increment slope shown in point 1 and point 2 waveforms. After the transformer is demagnetized, the voltage U_{DS} becomes equal to the input Uin (small edge in the middle).

Results from Task 10:

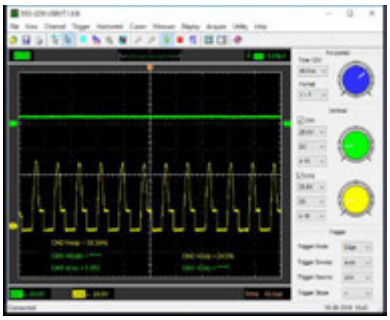


Figure 13. Waveform of U_{DS} voltage time diagrams of the power switch VT1 with connected overvoltage protection and disconnected demagnetizing circuit

The overcurrent protection provides an alternative route to the transformer demagnetizing current with closed

power switch VT1. Charging the capacitor C6 limits the rate of increase of the voltage U_{DS}. After the transformer has been demagnetized, the U_{DS} voltage equals the input Uin.

Results from Task 11:

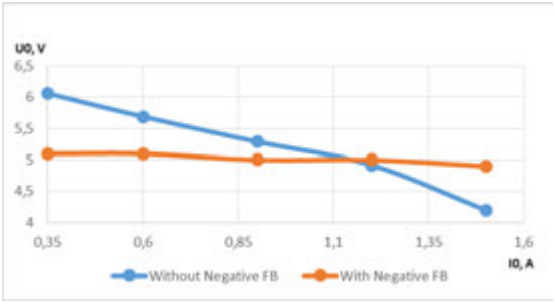


Figure 14. External characteristic of the converter without and with negative voltage feedback

The graph shows that the output voltage of the converter without negative feedback decreases as the load current increases. With negative feedback turned on, the value of the converter output voltage is stabilized.

V. CONLUSION

The tasks assigned in the research of the developed laboratory model aim to give students a thorough knowledge of the principle of operation of impulse straight switching forward single-ended DC/DC converter and its advantages.

Studying the negative feedback, the students find that it stabilizes the value of the output voltage of the converter and improves its parameters many times.

During the 2019/2020 academic year, the methodology described in the paper was used for the first time, with 34 students from the University of Ruse and its branches in Vidin and Silistra participating in the practical exercises. From the analysis of the results: 22 – excellent, 9 – very good and 3 – good, it can be concluded that the students showed interest, participated actively in the exercises, which increased their knowledge, skills and competences.

The practical experience gained in the study of switching power supplies would be useful in their future professional activities as engineers.

REFERENCES

[1] Cheng, Kuo-Hsiang & Hsu, Chun & Lin, Chih-Min & Lee, Tsu-Tian & Li, Chunshien. (2007). „Fuzzy–Neural Sliding-Mode Control for DC–DC Converters Using Asymmetric Gaussian Membership Functions“, IEEE Transactions on Industrial Electronic. 54, pp. 1528-1536. 10.1109/TIE.2007.894717.

[2] M. Zehendner, M. Ulmann, “Power Topologies Handbook” 2017, Texas Instruments.

[3] S. Roberts, “DC/DC Book of Knowledge, Practical tips for the User”, 2015q, pp. 38-40.

[4] Wang, Rong & Chen, Heng & Lei, Tengfei & Shi, Yun. (2014). “Research on Forward DC-DC Converter with High Power”. Advanced Materials Research. 978. pp. 63-66. 10.4028/www.scientific.net/AMR.978.63.

[5] V. Valchev, D. Judov, “Power Supply Devices”, Book, 2008.

[6] Yang, W., Qiu, L. J., & Wang, Y. X. (2014). Single-Ended Forward Switching Power Supply Design Based on UCC2800 PWM Controller. Applied Mechanics and Materials, 492, pp. 489–492.

[7] Do-Hyun Kim, Hyun-Woo Kim, Joung-Hu Park, and Hee-Jong Jeon (2015) “High Efficiency High-Step-up Single-ended DC–DC Converter with Small Output Voltage Ripple”, Journal of Power Electronics, Vol. 15, No. 6, pp. 1468-1479.

Designing an Interactive Multimedia Application for the Course "Communication Circuits"

A.N. Borodzhieva
University of Ruse "Angel Kanchev"/Department of Telecommunications, Ruse, Bulgaria
aborodzhieva@uni-ruse.bg

Abstract – **The paper describes a multimedia interactive application elaborated in English and Bulgarian, that will be utilized for teaching the course "Communication Circuits" of Bachelors in the specialty "Internet and Mobile Communications" in the University of Ruse, and mostly for Erasmus students (in English) attending the course. With the help of the application based on varied patterns of exercises students will be able to learn the basic concepts and issues related to the subject, including resonant circuits, electrical filters and modulations. It is planned to be multilingual, incorporating English, Spanish, French, Russian and Bulgarian.**

Keywords – Communication Circuits, Active Learning, Interactive and Multimedia Learning.

I. INTRODUCTION

Today, many universities in the world are striving to evolve more efficacious and attractive teaching and learning styles, supported by the rising of information and communication technologies (ICT). Innovations in education have incited the design of interactive multimedia technologies, such as "computer-based instruction and assessment", intelligent and integrated learning systems, etc. On its basis, the knowledge transfer needs to be resilient to the learners making the connections between "learners and learning resources", pursuing "higher levels of interaction between learners in order to increase the importance of independent learning".

Converting the "traditional teacher-oriented learning style" into "innovative student-oriented learning style" based on learners' interactions with other, improves the quality of education, stimulates students to show their knowledge they have learned during classes, builds communication and collaborative skills in students while working in teams, helps sharing experience between the students while interacting in the process [1].

The use of interactive multimedia systems in education helps students to better master the material in the course, increasing their effectiveness, respectively their grades, their motivation to learn and their skills for creative and logical thinking. The advantages of interactive multimedia systems are undoubtedly: a more interesting way of presenting information, the ability to learn anytime and anywhere, reducing teaching time and improving the motivation and quality of student learning [2].

The term "interactivity" is notably applied when talking about multimedia learning, but its definition and its advantages are not clearly specified. Some researchers have made attempts to define the term "interactivity" and the factors that influence it when designing interactive systems, as well as to present an interactive model involving "the user, learning environment and a system of connections and concepts" [3].

The paper describes a multimedia interactive application elaborated in English and Bulgarian. Using it, students may learn in an attractive manner the basic concepts studied in the course "Communication Circuits", including resonant circuits (serial, parallel, and more complicated), electrical filters (LC filters of type "K" and "m", passive and active RC filters) and modulations (amplitude, phase and pulse modulations) making various patterns of exercises ("learning by doing" and by playing). The application will be utilized in the course "Communication Circuits" for the educational qualification degree "Bachelor" in the specialty "Internet and Mobile Communications" in the University of Ruse, and mainly for Erasmus students attending the course. The application is planned to be multilingual, incorporating Bulgarian (designated for Bulgarian students), Spanish (the second most widely spoken language in the world), French and Russian (in addition to English).

II. A MULTIMEDIA INTERACTIVE APPLICATION FOR THE COURSE "COMMUNICATION CIRCUITS"

A. Interactive Multimedia

Computer, network and multimedia technologies have opened different channels for students to improve their knowledge and the competence of the scientific field they are involved in. Using multimedia applications, teachers can present the information in an innovative manner and motivate the students to learn quickly [3]. Interactive multimedia is equipped with user-operated controller tools and users can choose what they want for the next process, like in interactive games and CD applications (Fig. 1).

The most important feature of interactive multimedia is the students not only pay attention to media or object,

This paper is supported by the National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)", financed by the Ministry of Education and Science of Bulgaria. The work presented in this paper is completed as partial fulfilment of Project 2020 – FEEA – 03 "Design and development of a multifunctional robot for implementation and evaluation of autonomous navigation algorithms", financed under the Scientific and Research Fund of the University of Ruse "Angel Kanchev".

but also required to interact during learning. Interactive multimedia combines all media elements such as text, graphic, audio, video, visual and sound effects, and interactivity (Fig. 1) [3].

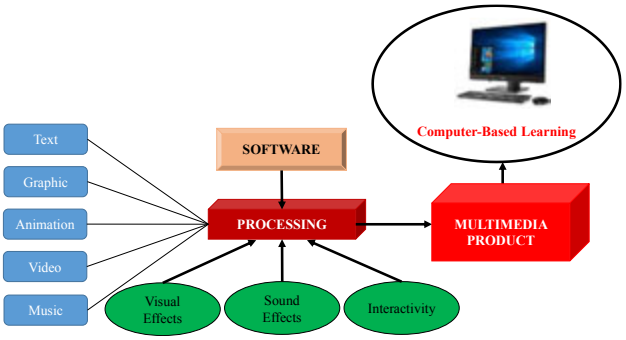


Figure 1. Process of development of learning multimedia based on computer

B. *Activities in LearningApps*

LearningApps is a Web 2.0 application, supporting the processes of learning and teaching based on small interactive modules used for self-studying or directly in learning materials. Reusable building elements, known as Apps, are collected and made available to everyone. The Apps do not include specific framework or learning scenario and they are not suitable as complete lessons or tasks, but they can be embedded in an appropriate teaching scenario [4]. The most important activities in LearningApps are explained in [4, 5] and presented in Fig. 2.

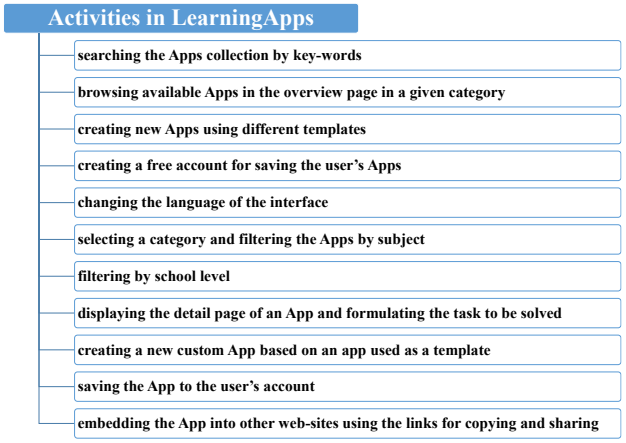


Figure 2. Activities in LearningApps

C. *Active Learning Using Project-Based Learning and Team Work in the Course "Communication Circuits"*

For the first time, this learning approach was applied in the training of two Erasmus students: M. I. Mazola Ortega (Spain) and J. Wysowski (Poland), who developed an application (in Spanish and English) for training in the discipline "Pulse and Digital Devices" and presented it at the conference of the University of Ruse, Fig. 3 [5]. For studying the topics in digital electronics testing exercises were developed using the templates "Matching Pairs", "Hangman", "Group assignment", "Multiple-Choice Quiz", "Pairing Game", and "Horse race", with details for creating the App based on the template "Matching Pairs" are presented in [5].



Figure 3. Presentation of the application developed by Erasmus students at the conference of the University of Ruse

For the purposes of the course "Communication Circuits" studied by bachelors (3rd year) in the specialty "Internet and Mobile Communications" and mainly for Erasmus students attending the course in the University of Ruse interactive multimedia applications have been created in English and Bulgarian during this academic year (2019-2020). The applications, covering varied types of testing exercises, allow students to learn in an attractive and fascinating manner the basic concepts studied in the course, including resonant circuits, electrical filters and modulations. The structure and the contents of the course is presented in Fig. 4.

LECTURES		
№	Topic	Classes
1.	Basic theory of communication systems and types of communication services. Series and parallel resonant circuits – resonance phenomena, frequency characteristics.	2
2.	Parallel resonant circuit with series connection of the load resistance of one of the reactive elements ($RL \| C$ and $RC \| L$) – resonance phenomena, frequency characteristics.	2
3.	Sectioned resonant circuits with capacitive and inductive divider. Transformer with one and two tuned oscillating circuits – resonance phenomena, frequency characteristics.	2
4.	Electrical filters. Types of filters depending on the component basis used. Approximations. Passive electrical filters. Analysis and design of reactive chain filters.	2
5.	Chain LC filters of type "K" and type "m" – serial and parallel derivatives.	2
6.	Passive RC filters. Crystal and ceramic filters.	2
7.	Active filters. Cascading implementations of active filters. Second-order active filters with low and mediate quality factor – sections with one operational amplifier with single-loop and multi-loop negative feedback.	2
8.	Second-order active filters with low and mediate quality factor – sections with one voltage controlled voltage source, biquadratic sections with one operational amplifier.	2
9.	Second-order active filters with high quality factor – filter sections with two operational amplifiers using the model with generalized impedance converter.	2
10.	Second-order active filters with high quality factor – universal biquadratic sections with three and four operational amplifiers. Active filters with switchable capacitors (SC filters).	2
11.	Amplitude and phase correctors – principles of amplitude and phase correction, transfer functions of amplitude and phase correctors of the first and second order, realization of passive and active amplitude-correcting and phase-correcting sections. Attenuating four-poles.	2
12.	Amplitude modulation – basic dependences, analysis, spectrum, power. Double side-band (DSB) and single side-band (SSB) amplitude modulations with suppressed carrier component.	2
13.	Angular modulation – frequency and phase modulations, basic dependences, frequency spectrum and power of angular-modulated signal. Comparison between frequency and phase modulations.	2
14.	Influence of amplitude-modulated and angular-modulated signals on resonant circuits.	2
15.	Pulse modulation – general information. Types of pulse modulations, features. Special types of modulations – pulse-code modulation (PCM), differential PCM and delta-modulation.	2
	Total	30

Figure 4. Structure and contents of the course "Communication Circuits" for 3rd year bachelors

At the beginning of the semester, students get acquainted with the idea of project-based learning and its advantages (Fig. 5), which argues the need for the introduction of team and project work during the classes. This proved to be extremely useful, especially this academic year (2019-2020), in the context of Covid-19, when the transition to distance learning was necessary.

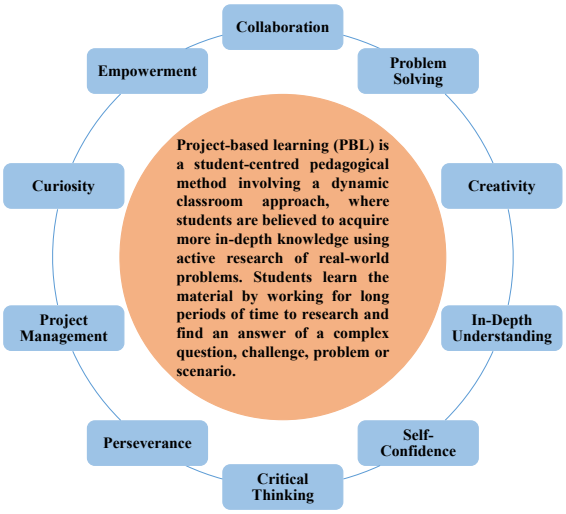


Figure 5. Project-based learning and its benefits

The students are assigned, as a task for their project activity, the development of a multimedia interactive application, choosing the appropriate environment for the purpose. At the beginning, an example of such a platform, LearningApps, is presented to the students, as the teacher introduces them to the various opportunities provided by the platform [4, 5, 6] and varied types of testing exercises developed for the purposes of different courses such as "Pulse and Digital Devises" [5], "Communication Circuits" (an example, presented in the paper), "Coding in Telecommunication Systems" [6]. There is no restriction on the means used for creating the applications. For example, students with good computer knowledge, skills and abilities can develop their own such applications, Web- or Windows-based. However, the practice has shown that in the presence of instructions for working with this platform, and mostly due to the presence of an interface in Bulgarian, all students chose this platform to create their application.

The application of active learning methods using project-based learning and team work in the course "Communication Circuits" is illustrated visually in Fig. 6. An example distribution of activities over time is shown in Fig. 7, in the form of a Gantt chart. First, students are divided into two groups: conditionally called Bulgarian and English. The English group includes students with relatively good knowledge and use of English. Teams are formed on a voluntary basis (students who know each other well and can work in a team), and the team includes one student from the Bulgarian group and one student from the English group. Each team is given a specific topic from the subject matter studied in the course "Communication Circuits" (Week 1, Fig. 7).

In the working process, the students in the team should choose the environment (the way) they will use to perform their task, look for adequate references (mostly textbooks, recommended by the teacher [7-13]) on the topic, select appropriate templates for the tests, choose appropriate questions and answers and the relevant photo materials (pictures, videos, etc.) that will be used in the construction of the tests, etc. Depending on the type of literature used (Bulgarian or English) the students must translate the materials into the other language (English or Bulgarian). This allows the construction of a bilingual multimedia interactive application in Bulgarian and English, which has the same structure and contents in both languages.

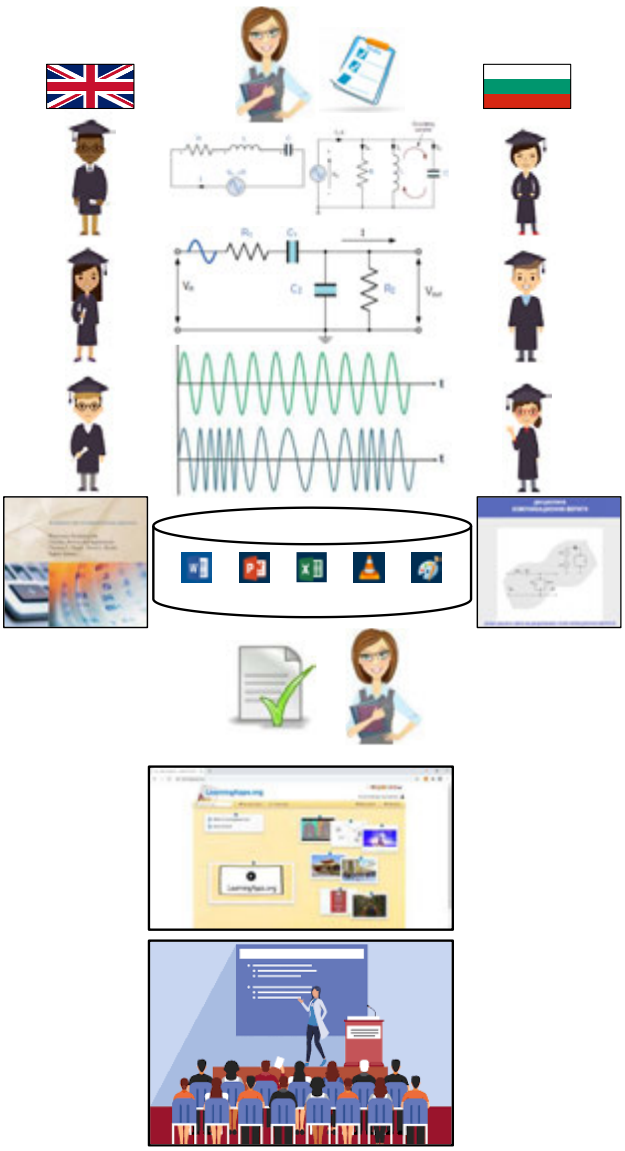


Figure 6. Applying the project-based learning and team work in the educational process

Students must prepare in advance (using different software products for this, such as MS Word, MS Excel, MS PowerPoint, Paint, etc.) in a Word document the materials that will be used to create the tests and present them to the teacher for checking, so that errors not to be made in the development of the testing exercises.

TASKS (PROJECT-BASED LEARNING)	WEEKS												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Forming teams and getting the assignment													
2. Selection of an environment to create the application													
3. Searching for references													
4. Preparation of questions and answers (documents)													
5. Verification of the documents by the teacher													
6. Creating tests in the selected environment													
7. Presentation of the development to colleagues													

Figure 7. Distribution of activities over time using Gantt chart

For completing the execution of their task, students must describe the steps in the application creation process citing the references used and attach screenshots (Fig. 8 and Fig. 9) to be able to follow these steps, as well as attach photos from the application testing (creating the pairs, filling in the blanks, etc., depending on the templates they have chosen), Fig. 10, explaining them. At the end of the semester, students from each team present their application to all their colleagues, providing the links to which they are published so that the tests can be used by their colleagues in the discipline to prepare for the upcoming exam. One of the students or more develop a Web-site with links of all the materials created by the students for their using when preparing for the final exam in the course (Fig. 11).

Below a brief description of creating such an application is given, as it is expected to be performed by students (by analogy with the development of user documentation accompanying a software product).

For building a testing exercise on the topic "Modulations", the template "Matching Pairs" is selected. The steps in the process of creating are given in Fig. 8 and they are as follows:

1) The title of the App is defined – in this case, the title is "Modulations" (Fig. 8, part 1); 2) A short description of the task for the App, shown on start up, is given – in this case, "Make the pairs: image + text to speech (audio) OR text + video!" is specified (Fig. 8, part 2). 3) The pairs are specified (Fig. 8, part 3A and part 3B), by entering the two media which belong together and they can be a mix of text, images, audio or video – here, different media are chosen. 4) Up to three additional elements that will be displayed but do not belong to the solution might be added, but in this case this option is not used (Fig. 8, part 4). 5) An option for making matching pairs disappear might be selected, where matching pairs will be automatically checked and then disappear. In this case, it is not checked, and matching pairs will not disappear without this option (Fig. 8, part 5). 6) A choice how the cards to be displayed with two options "Cards one above the other" or "Cards alongside one another" is available – here, the first option (by default) is chosen (Fig. 8, part 6). 7) Feedback is provided (Fig. 8, part 7) – here, the text "Great, you've found the solution!" is specified and it will be displayed when the solution is found. 8) Providing some hints how to solve the App, that can be viewed by the user via a small icon in the upper left corner is also possible, in this case it is left blank, no hints are supposed to be given to the student, because the answers are very easy. All these steps are described in the platform and it is not difficult for students to explain them briefly in their explanations.



Figure 8. Steps in creating the application – describing each step in the process

When defining the pairs in the template, for attracting users' (respectively students') attention, the elements used for the pairs are "text to speech" (Fig. 9 a) (audio), and video (Fig. 9 b). The video clips used in the application are from YouTube [15, 16, 17]. The choice of images and videos might be from Internet, giving URL address or from PC or the user's own.

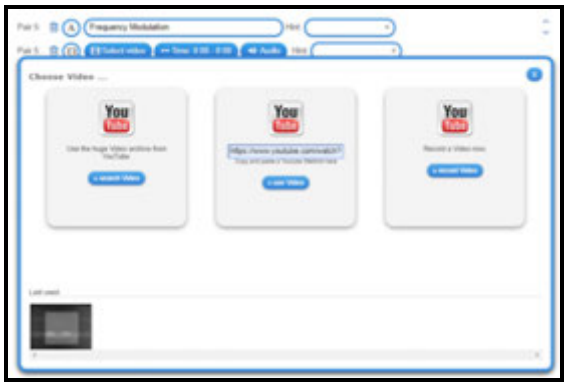
To their solution, students add in a separate folder all the graphic images, audio and video components, they have created for their application, as well as links to the used graphic, audio and video elements [15, 16, 17], if students use them in advance, due to copyright protection.

The initial state (initial view) of the exercise and the formulation of the task are given in Fig. 10 a. When clicking on "Audio", a window with an option for listening to the audio, appears (Fig. 10 b). When clicking on "Video", a window with an option for watching the video appears (Fig. 10 c). The final state where all the answers are right (marked in green) and the corresponding greeting message "Great, you've found the solution!" are

shown in Fig. 10 d. The situation with a wrong choice, marked in red, is presented in Fig. 10 e.



a)



b)

Figure 9. Creating the App based on the template "Matching Pairs" using audio and video components

In the academic 2019-2020 year, an experiment for applying active learning methods including project-based learning and team work was conducted with the students-bachelors of the specialty "Internet and Mobile Communications" in the course "Communication Circuits". The students had to implement such interactive and multimedia applications. All of them chose the platform LearningApps due to: 1) its interface in languages, including Bulgarian; 2) detailed in for working with the platform; 3) many developed with this platform for studying courses in their specialty.

The students did brilliantly their assignments interest in the platform, experimenting with templates and options of the templates and different testing exercises to their colleagues. templates were "Matching Pairs", "Group assi "Free-text input", "Matching Pairs on Images", " Choice Quiz", "Cloze text", "Crossword", "Wo "Guess the word" as the template "Matching Pai out to be the most preferred by the students.

One of the teams suggested as an example a to a problem, as the individual stages of the solu shown as images and the user who solves the problem must arrange them in the correct sequence, choosing the appropriate "Number line" template.



a)



b)



c)





e)

Figure 10. States of the testing exercise built on the template "Matching Pairs" on the topic of modulations: a) initial (formulation of the task); b) listening to audio; c) watching video; d) final (right choice and greeting message); e) final (wrong choice)

According to students, some of the templates did not seem particularly suitable for building learning applications for students at a university, such as "The Millionaire Game" (known for entertainments on the television) and "Horse race". The templates "Simple order", "App Matrix", "Audio/Video with notices", "Group-Puzzle", "Where is what?", "Pairing Game", "Guess", "Tools", "Voting", "Chat", "Calendar", "Notebook", "Pinboard" were not selected by the students during this academic year. Some of the students said that they could not understand their functioning because the examples were not in languages known to them, and they did not have enough time to make experiments.

Students demonstrated great interest in the material studied and shared that they learned a lot about the construction of multimedia interactive applications, as well as concepts in the terminology used in the discipline "Communication Circuits" using an attractive way of learning by doing.

One of the students interested in web design, suggested to create a Web-site with links of all the materials created by the students for their using when preparing for the final exam in the course (Fig. 11).

According to the Gantt chart (Fig. 7) the students from each team had to present their texting exercises to all their colleagues, but due to the Covid-19 crisis, the presentation could not take place in person, so the solution was using the system BigBlueButton [18], Fig. 12, preferred for synchronous distant learning during this semester by most of the teachers and students at the University of Ruse "Angel Kanchev" or Microsoft Team [19], the other alternative in the university.

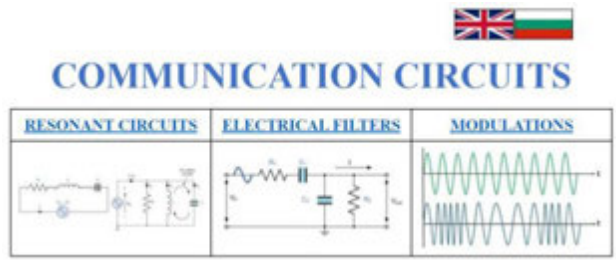


Figure 11. Index page of a Web-site developed by the student with all the links of the tests during the academic year 2019-2020

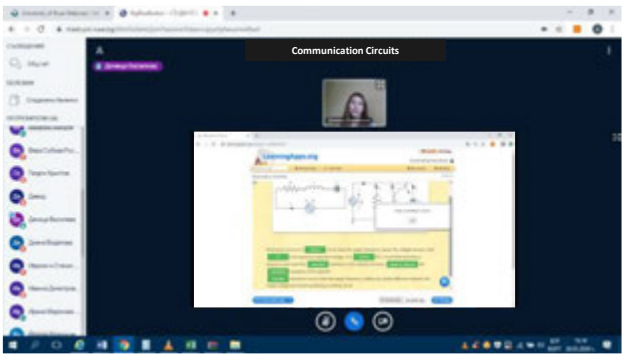


Figure 12. Presenting the multimedia interactive applications using the system BigBlueButton

III. CONCLUSION

The paper demonstrates a tool for designing interactive and multimedia training systems applied for attracting students' attention to the material studied in the course "Communication Circuits" in our university. The applications allow students to learn basic definitions based on texts, pictures, audio, etc. studied in the course. Various types of testing exercises and their creating are presented in the paper. The application is intended to be extended, covering more topics, more types of testing exercises, and more languages, for example Bulgarian and Spanish.

REFERENCES

[1] L. Fui-Theng, M. Neo, "Interactive Multimedia Learning: Innovating Classroom Education in a Malaysian University." Turkish Online Journal of Educational Technology 13(2): pp. 99-110, 2014.

[2] W. Wiana, S. Barliana, A. Riyanto, "The Effectiveness of Using Interactive Multimedia Based on Motion Graphic in Concept Mastering Enhancement and Fashion Designing Skill in Digital Format", International Journal of Emerging Technologies in Learning 13(02): 4, 2018.

[3] S. Heidig, R. N. Schwartz, J. L. Plass, "Interactivity in Multimedia Learning: An Integrated Model". Computers in Human Behavior 26(5): pp. 1024-1033, 2010.

[4] https://learningapps.org (Last visited: July 2020)

[5] M. I. Mazola Ortega, J. Wysowski, A. Borodzhieva, "Designing an Interactive Multimedia Bilingual Application for the Course "Pulse and Digital Devices"". 58th Science Conference of Ruse University, Bulgaria, 2019, Proceedings, Volume 58, book 3.2. Communication Systems & Networks, Ruse, 24-26 October 2019, pp. 37-44, ISSN 1311-3321.

[6] A. Borodzhieva, "Development and Description of an Interactive Multimedia Application Applied in the Course "Coding in Telecommunication Systems"". 16th International Scientific Conference "eLearning and Software for Education" (eLSE 2020), 23-24 April 2020, Bucharest, Romania, Proceedings, Volume 2, pp. 548 – 565, doi: 10.12753/2066-026X-20-158.

[7] A. Manukova, A. Borodzhieva, "Communication Circuits – Exercise Guide". Ruse, University of Ruse, 104 pages, 2002.

[8] R. E. Ziemer, W. H. Tranter, "Principles of Communications: Systems, Modulation, and Noise". Seventh Edition, John Wiley & Sons, Inc., pp. 746, 2015.

[9] S. Haykin, M. Moher, "Communication Systems". Fifth edition, International Student Version, Wiley, 2009.

[10] L. W. Couch, "Digital and Analog Communication Systems". Sixth edition, Prentice-Hall, 2001.

[11] Th. L. Floyd, D. L. Buchla, "Electronics Fundamentals: Circuits, Devices and Applications". Eighth Edition, Pearson New International Edition, Pearson Education Limited, pp. 1065, 2014.

[12] St. Winder, "Analog and Digital Filter Design". Second edition. Newnes, Elsevier Science (USA), pp. 458, 2002.

[13] W.-K. Chen, et al, "The Circuits and Filters Handbook". Editor-in-Chief: Wai-Kai Chen, CRC Press LLC, pp. 2881, 2003.

[14] Modulation. https://searchnetworking.techtarget.com/definition/modulation

[15] Amplitude Modulation Tutorial and AM Radio Transmitter Circuit. https://www.youtube.com/watch?v=3I_e7gIyfQg

[16] Frequency Modulation – Part I - Basic Principles. https://www.youtube.com/watch?v=gfz1FbIOMbs

[17] What is Pulse Code Modulation (PCM)? https://www.youtube.com/watch?v=YJmUkNTBa8s

[18] BigBlueButton, https://bigbluebutton.org/ (July 2020)

[19] Microsoft Teams, https://www.microsoft.com/bg-bg/microsoft-365/microsoft-teams/group-chat-software (July 2020)

Integrated laboratory complex

R. Simionov*, S. Mollova* and R. Dolchinkov*
* Burgas Free University/Faculty of Computer Science and Engineering, Burgas, Bulgaria
radoslav.simionov@gmail.com, mollova@bfu.bg, rado@bfu.bg

***Abstract* - This report examines the project-oriented training of students in Computer Science and Energetics through research and implementation of an Integrated Laboratory Complex in Burgas Free University. Through the construction of an internal local area network, three different premises are physically connected: the building - two laboratories and an adjacent roof space with solar panels for electricity. Information received from sensors is recorded and stored on the network server; processed in a Raspberry Pi based computer cluster; simulations are performed to predict the characteristics of the PV system and the amount of energy produced. The proposed model of an integrated laboratory complex aims at 1) engaging students in research activities of an interdisciplinary nature and 2) solving further problems related to the production of green energy for use in the university building.**

Keywords* - *project-based training, photovoltaic systems, computer clusters, computer networks

I. INTRODUCTION

The quality of the skills and knowledge base of engineering students gets significantly improved when their theoretical knowledge is used for the realization of specific practical tasks. One way to achieve this is to involve students in the development of different projects, ie. project-based learning (PBL) [1], [2]. The projects may be linked to a single discipline studied in a particular field. However, in the modern development of science, technics and technologies, there is often a need for knowledge that is both focused and interdisciplinary.

The offered integrated laboratory complex is built with the joint efforts of students and professors from several different specialties at Burgas Free University- students from the Bachelor's Program in "Computer Systems and Technologies"," Computer Networks","Power Supply and Equipment", "Communication and Renewable Energy", as well as students from the Master's Programs in "Integrated Computer Systems and Complexes" and "Engineering and operation of energy systems". The laboratory complex also plays a significant role in research of the teaching staff [3-7].

As part of active student learning, PDL focuses on three areas:

1) Construction, operation and research of parallel computer structures;

2) Construction of systems for collection, storage and analysis of data from photovoltaic plants;

3) Design, constraction and setup of an internal LAN network .

No. KP-06-COST-8, August 6, 2019, "Prediction of Characteristics and Optimization of a PV System by Artificial Intelligence Tools"

II. CONCEPTUAL MODEL OF AN INTEGRATED LABORATORY COMPLEX AT BFU

The complex connects a variety of components in a network. An internal LAN is being created that physically connects three different rooms in the building - Lab 31, Lab 314 and the roof space that houses the solar panels for electricity.

A. Description of the components - Laboratory 31

Two computer clusters are included in this laboratory. One is made up of 4 desktops and the other of 4 Raspberry Pi single board microcomputers.

Cluster technology [8] allows improvement of the two main features of a computer system: performance [9] and reliability (fault tolerance). Reliability is achieved by having a surplus of machines, which in practice is redundant, which in turn makes the system fault-tolerant [10].

1) PC based computer cluster

Hardware platform

In the manufacture of cluster machines provided the conditions:

- accommodate four motherboards and all necessary peripherals
- provide adequate space for fans and airflow for efficient cooling of machines
- provide easy replacement if any of the devices does not work

The corresponding serviceable components have been selected for the construction thus formed. Due to the fact that already used computers are used, each RAM board and each hard drive were tested. Motherboards that use AMD processors with integrated video have been selected.

Software platform

Separate computers are connected to one local network using a switch (Fig. 1). Since the equipment used is outdated, must be selected the appropriate operating system to support it. It was chosen Cent OS, which is free, with a number of flexible applications, and provides the ability to manually adjust as necessary.

The "Sudoers file" has been configured to provide administrator privileges.

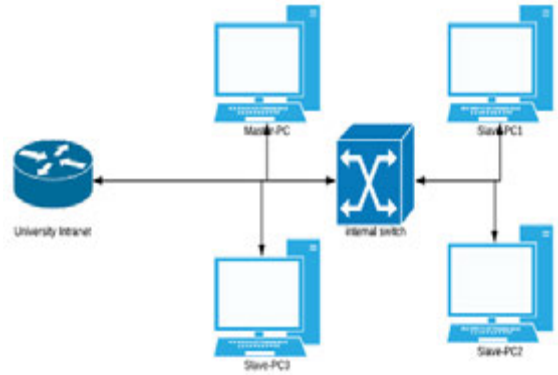


Figure 1. LAN of PC computers

In order to be able to add other computers to this cluster or replace damaged nodes, a program is included that checks for new computers on the local network and adds them to those that the head node recognizes.

The task of communication between the cluster machine and an external terminal is solved by SSH (SecureSHell), a network protocol that allows encrypted data transmission.

2) Cluster based on Raspberry Pi microcomputers

The following components are used in the construction of the computer cluster [11] :

- 4 x Raspberry Pi 3;
- 4 x Class 10 SD cards;
- 10/100 Mbit Awning S108 Switch;
- 6 ports x 2.4A Tecknet U601 14.4A Power Supply;
- 4 x CAT 5E UTP cables;
- 4 x USB power cords.

To implement the entire configuration, the components of the cluster are connected in a single module. Added peripherals - keyboard, mouse and monitor, which allows settings and control without an additional computer external to the system (Fig. 2).

The operating system Raspbian Jessie is used, which is a Linux operating system based on Debian. It is specially optimized for Raspberry hardware, which has an ARM architecture [12].

The Message Passing Interface (MPI) standard for libraries was used (Fig. 3)

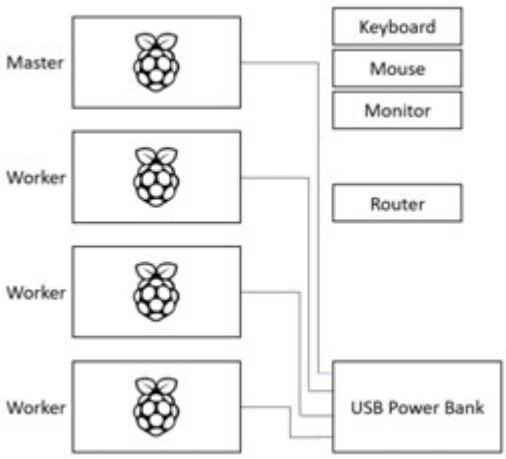


Figure 2. Compact Raspberry Pi module

Python is selected as the programming language. The platform is preferred for working with single-board computers. Additionally, an MPI4PY library is installed, which allows the use of a multiprocessor configuration.

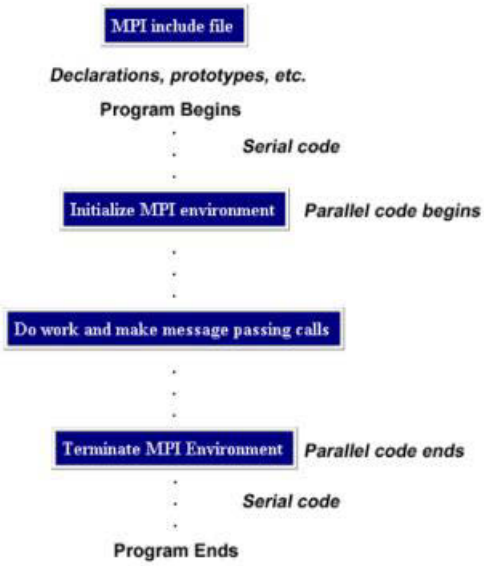


Figure 3. Message Passing Interface

Python is selected as the programming language. The platform is preferred for working with single-board computers. Additionally, an MPI4PY library is installed, which allows the use of a multiprocessor configuration.

B. Description of the components - Laboratory 314

In this laboratory are located components of the specialized system for collection, storage and analysis of data from the photovoltaic plant, taking into account the readings from the meteorological station [13], [14], [15].

1) Data logger Fronius - serves to display information from sensors located on the roof. It consists

of two parts - a module for connecting sensors "Fronius Sensor Box" and a module for storing data "Fronius Datalogger Web" (Fig.4)[16]. *The connection between the data storage module and the sensor module is serial, type RS- 232. The module transmits data in real time and archives data every 5 minutes. The memory capacity is 16 MB (4096 days);*

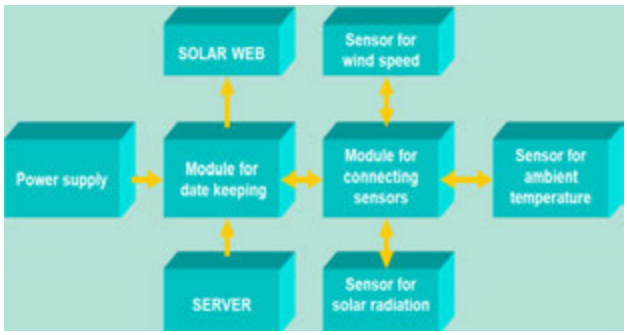


Figure 4. Monitoring and assessing the energy potential of solar energy

2) HP ProLiant ML370 Generation 5 server for storing and analyzing data from photovoltaic panels;

3) Solar inverter SolarMax 15MT2 to convert the DC input voltage coming from the photovoltaic panels to the AC output voltage

4) Router - HP Procurve Switch 5304XL. Represents a 24-port Rack mount 3U router;

5) Switch- 3com 1100 - 12 port Rack mount 2U.

C. Description of the components - Roof space

There are two main components on the roof:

1) Photovoltaic panels - Sunmodule SW 145 mono;
2) Development environment – Arduino.

The components of the two laboratories and the roof space described in this way are connected in a single complex, which is shown in Fig. 5.

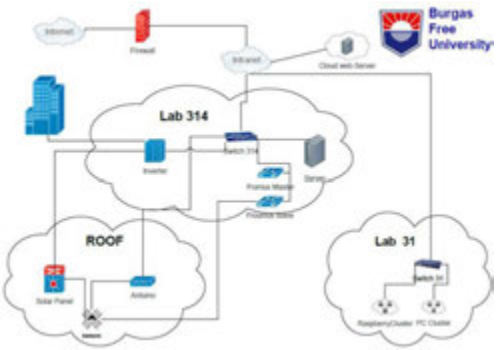


Figure 5. Conceptual model of Integrated laboratory Complex

D. Designing communication between individual components in an integrated laboratory complex

For the design of the network component of the integrated laboratory complex, a software product for modeling computer networks Boson NetSim 11 is used. It describes the settings that are performed for each component - a PC-based cluster, a Raspberry Pi-based cluster, Arduino, a Sunmodule SW 145 mono solar panel, a registrar Fronius Data, HP ProLiant ML370 Server. Communication between them is carried out using Cisco Catalyst 3750 switches.

1) Network topology

From the list of possible devices that can be used, three switches C3750V2-24TS-S (24 port Cisco 3750) are selected: INTRANET_BFU - the main, SWITCH_314 - the switch in Lab. 314 and SWITCH_31 - the switch in the Lab. 31.

Add all the components that are in Lab. 31 and Lab. 314., giving them the following names:

Solar inverter – INVERTER

Arduino – ARDUINO

Data logger Fronius – FRONIUS

Server HP ProLiant ML37 – SERVER

PC cluster – PC_CLUSTER

Raspberry Pi cluster – RASPBERRY_CLUSTER

Connections are made between the switches and the end devices. The following connection ports are set:

INTRANET_BFU FastEthernet1/0/1 to SWITCH_314FastEthernet1/0/1

INTRANET_BFU FastEthernet1/0/2 to SWITCH_31FastEthernet1/0/1

SWITCH_314 FastEthernet1/0/2 to INVERTER ETHERNET0

SWITCH_314 FastEthernet1/0/3 to ARDUINO ETHERNET0

SWITCH_314 FastEthernet1/0/4 to FRONIUS ETHERNET0

SWITCH_314 FastEthernet1/0/5 to SERVER ETHERNET0

SWITCH_31 FastEthernet1/0/2 to PC_CLUSTER ETHERNET0

SWITCH_31FastEthernet1/0/3 to RASPBERRY_CLUSTER ETHERNET

The topology of the constructed network is shown in fig. 6.

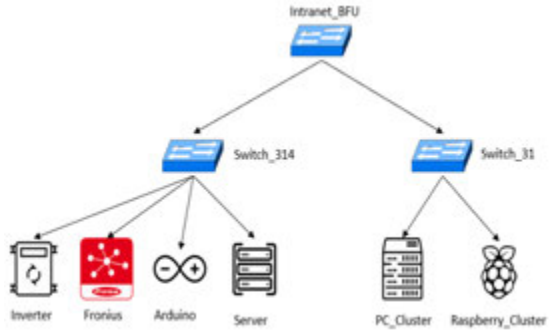


Figure 6. Connection between laboratory 31 and laboratory 314

2) Network setup and testing.

A simulation is started in order to configure all devices in the network.

In the main switch, two Vlan are set up, Vlan31 for Lab.31 and Vlan314 for Lab. 314. Thus, a connection is established between the laboratories through the existing network in BFU, but no one other than these devices will have access to each other.

Set the switch name BFU_INTRANET, the two Vlan ports FastEthernet 1/0/1 and FastEthernet 1/0/2. Vlan0031 has a network IP address of 10.0.0.1 and a mask of 255.255.255.0, and Vlan0314 has a network IP address of 172.16.0.1 and a mask of 255.255.255.0.

The following are the settings for SWITCH_314 and SWITCH_31.

The ping command and the IP address of the other device establish that there is a connection between the devices in Lab.314. The procedure is repeated for laboratory 31.

Once there is a connection between the devices in each laboratory, tests for connection between laboratory 31 and laboratory 314 are performed separately.

III. EXPERIMENTS CONDUCTED AND RESULTS OBTAINED DURING THE OPERATION OF THE LABORATORY COMPLEX

A. Experimental studies from the laboratory No. 31

Both clusters in the laboratory are developed exclusively by students studying at BFU.

The cluster, which is based on old desktops already used in practice, is assembled, set up and tested by students in the Bachelor of Computer Systems and Technology program. The leading idea in development is the analogy with the creation of RAID. In this case, cheap used machines are used instead of cheap disks. However, the implemented parallel structure is not only cheap, but with the surplus obtained, it also becomes fault tolerant.

The structure thus obtained enables students to study and explore the important performance and resilience characteristics of a scalable system [17].

More computers can be added and the relationship between their number and cluster performance explored. It is possible to connect with another, Raspberry Pi based cluster and examine the behavior of the system thus obtained

The Raspberry Pi 3 microcomputer cluster was developed by students enrolled in a Master's program in Integrated Computer Systems and Complexes. The process of creating this cluster, as well as the experiments conducted to investigate its performance and fault tolerance, are described in [18], [19]. An energy efficiency study was conducted for the same cluster [20] . With additional hardware upgrades, where different sensors are connected, a part of the emission monitoring system has been made.

Both developments are used in laboratory exercises in the disciplines of Computer Architecture and Failover Computer Systems.

B. Experimental results from laboratory tests 314

This laboratory has implemented a system for analysis and visualization of data from a meteorological station at BFU. It includes the following components [21],[22]:

- Inverters;
- Laboratory server;
- Metering systems for measuring energy output;
- Alternative electricity generation systems.

The data obtained from the system is used for the extracted electricity from the photovoltaic plant of the BFU. The specialized software developed, based on neural networks, makes a prediction of the performance of the panels when receiving meteorological forecast data [17].

C. Experimental results from the joint work of laboratory 31, laboratory 314 and the roof space

Fig. 7 and Fig.8 Show a part of the obtained results, presenting a study of the dependences of the electricity production upon change of different external factors, for two periods of time. From these, the students can obtain baseline information to analyze the performance of photovoltaic panels with appropriate changes in solar radiation levels, wind speed, temperature, and other meteorological characteristics.

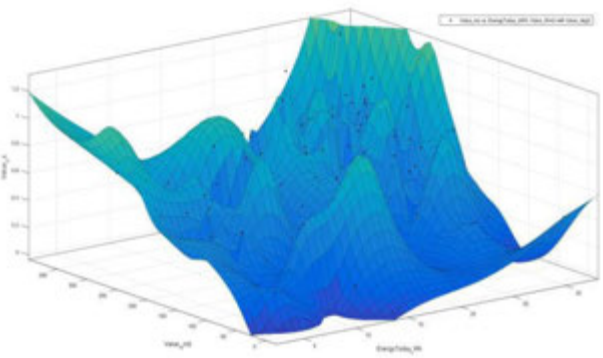


Figure 7. Electricity production upon change of different external factors for period of time 1

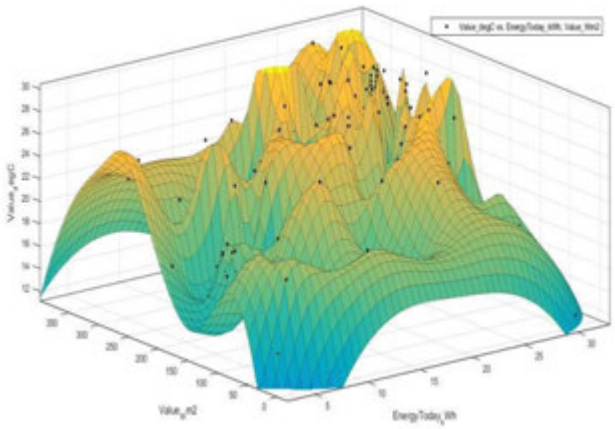


Figure 8. Electricity production upon change of different external factors for period of time 2

Show a part of the obtained results, presenting a study of the dependences of the electricity production upon change of different external factors, for two periods of time. From these, the students can obtain baseline information to analyze the performance of photovoltaic panels with appropriate changes in solar radiation levels, wind speed, temperature, and other meteorological characteristics.

D. Acquisition of practical skills by students

The achieved educational skills and the results for the students, who participated in the development of the project, can be indicated as:

- 12 students presented 17 papers at Student Scientific Conferences, which were then published in the respective proceedings;
- Five joint publications of students and lecturers were made in the Yearbook of BFU;
- The students presented the results of their participation in the work on the Integrated Laboratory Complex in the "Open Day" of the University;
- Five diploma theses of the students from the master's program "Integrated Computer Systems and Complexes" and "Engineering and Operation of Energy Systems" have been developed and defended on topics related to solving various tasks of the project [23],[24].

IV. CONLUSION

The goals set for the team related to PBL have been fulfilled:

Construction, operation and research of parallel computer structures- two clusters were created (PC and Raspberry Pi3 based) and three students from the master's degree "Integrated Computer Systems and Complexes" and six students from the bachelor's degree "Computer Systems and Technologies" took an active part. Computer clusters are used in the education of students in the discipline "Computer Architectures" from the bachelor's degree, and in the master's degree in the disciplines "High-performance computer systems" and "Fault-tolerant computer systems".

Construction of systems for collection, storage and analysis of data from photovoltaic plants – a server is programmed for storing and using the data from the sensors of PV panels located on the roof of the university building. This part of the study involved four students from the bachelor's program "Power Supply and Furnishings" and "Communication and Electronics of Renewable Energy Sources", three students from the master's program "Engineering and Operation of Energy Systems" and one PhD student in Energy. Students work on research on the reliability and efficiency of panels produced by different technologies, as well as defective ones related to the aging of materials.

Design, construction and setup of an internal LAN network - two students with a bachelor's degree in "Computer Systems and Technology" and two students with a master's degree in "Integrated Computer Systems and Complexes"were involved. The realization of the network allowed students to put into practice their knowledge acquired in the discipline "Computer Networks".

The work on the project connected students and teachers from different subjects, in a small team combining their theoretical knowledge and practical skills for solving an interdisciplinary task.

The established laboratory complex will allow to conduct practical classes for some of the technical specialties of the Center for Informatics and Technical Sciences at Burgas Free University, related to the specialties "Computer Systems and Technologies", "Computer Networks", "Power Supply and Electrical Equipment" programs and "Integrated computer systems complexes", "Communication and electrification of renewable energy sources" and "Engineering and operation of energy systems" from the master's programs.

Therefore, the educational goals increasing the motivation and creative activity of the students through their inclusion in specific practical developments, have been achieved.

The work on the Integrated Laboratory Complex, a result of the joint efforts of teachers and students at BSU, practically realizes PBL. The students managed to apply their theoretical knowledge in specific practical tasks, to create a creative work atmosphere, to build proper communication with each other. All this leads to their growth as future engineers, as well as to the development of teamwork skills.

The future goals:

1) To increase the number of computers connected in parallel in the clusters and to study the energy efficiency of scaling the system;

2) to do research and implement a system for monitoring a particular energy object based on a computer cluster.

ACKNOWLEDGMENT

This paper is funded by Scientific Research Fund, agreement No. KP-06-COST-8, August 6, 2019, "Prediction of Characteristics and Optimization of a PV System by Artificial Intelligence Tools" for providing national co-financing for participation of Bulgarian collectives in approved actions under the European program for cooperation in the field of research and technology COST.

REFERENCES

[1] Burdewick, "Aspects of methodology and education psychology In project-oriented studies," International Workshop on Project Oriented Learning, Groningen, 2003

[2] J. Mills, "Engineering education-is problem based or project-besed learning the answer?", Australian journal of engioneering education, AAEE Inc., vol. 3 (2), pp. 2-16, 2003

[3] Georgieva, P., I. Popchev, S. Stoyanov, A Multi-Step Procedure for Asset Allocation in Case of limited Resources- Cybernetics and Information Tchnologies Vol. 15, No 3, 2015, 41-51, Print ISSN: 1311-9702, Oline ISSN: 1314-4081, DOI:10.1515/cait-2015-0040

[4] P. V. Georgieva, Parameters of GFSSAM: Coding the Parameters of a Hybrid Genetic Fuzzy Systems- Proceeding of the Sevennth International Conference on Telecommunications and Remote Sensing, (ICTRS '18), 85-92, Barcelona, Spain, ACM New York, NY, USA 2018, ISBN:978-1-4503-6580-2, DOI:10.1145/3278161.3278174, 2018

[5] Popchev, I., P. Georgieva. A fuzzy approach to solving multicriteria investment problems. In: Innovative Techniques in Instruction Technology, E-Learning, E-Assessment, and Education (Ed. by Magued Iskander), Springer, Science+Business Media B.V., New York, 2008, 427-431, ISBN: 978-1-4020-8738-7. Library of Congress Control Number 2008931889.

[6] Georgieva P., I. Popchev. Fuzzy Q-measure Model for Managing Financial Investments. – Comptes Rendus de l'Academie bulgare des Sciences, vol. 66, 2013, №5, 651-658. ISSN 1310-1331. DOI: 10.7546/CR-2013-66-5-13101.

[7] Georgieva, P. Genetic Fuzzy System for Financial Management. - Cybernetics and Information Technologies, Vol. 18, 2018, No. 2, 20-35. Print ISSN: 1311-9702; Online ISSN: 1314-4081. DOI:10.2478/cait-2018-0025.

[8] Hoffmann K., Meyer A., Parallel Algorithms and Cluster Computing, „Springer", 2006

[9] Jeff Layton, Monitoring HPC Systems, ADMIN Magazine, 21/2014

[10] http://searchexchange.techtarget.com/definition/cluster

[11] Pajankar A., Kakkar A., Raspberry Pi By Example, „Packt", 2016

[12] Pajankar A., Raspberry Pi Supercomputing and Scientific Programming, "Apress", 2017

[13] P. V. Georgieva, Dolchinkov R., Fuzzy models for managing a micro grid PV system, 42nd International Convention on Informationand Communication Technology, Electronics and Microelectronics, MIPRO 2019, Proceedings

[14] Dolchinkov R., Teaching methods in computer design of technological systems, SEVILLE, SPAIN, 6TH International conference of education, Reserch and Inovation, ISBN 978-84-616-3849-9, p. 5785-5795, 2013

[15] Ginko Georgiev, Silviya Letskovska, Kamen Seymenliyski, Laboratory Equipment for Energy Accumulation from Renewable Sources, ICEST 2016, International Scientific Conference On Information, Communication And Energy Systems And Technologies, 28-30 June 2016, Ohrid, Macedonia

[16] S.A. Letskovska, K.D. Seymenliyski, E.D. Zaerov and R.R. Simionov. Educational technologies in energy efficiency training International Conference, Economic Aspects of Energy Efficiency of the EU Countries, Workshop on Energy Efficiency in Industrial and Buildings Sectors, 8-11 07. 2019 Nizhny Novgorod, Russian Federation, p. 27-35, University named after R.E. Alekseeva. - N. Novgorod, 2020 . - 246 p., ISBN 978-5-502-01284-3

[17] Georgieva P. V., E. P. Nikolova, Enhancing Communication Competences through Mathematics in Engineering Curriculum. 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2019

[18] Mollova,S., Kostadinov, A., Georgieva, P., Fault-tolerance of a Laboratory Computer Cluster "XX International Symposium on Electrical Apparatus and Technologies, 2018, Bourgas, Bulgaria

[19] Mollova S., M. Zhekov, A. Kostadinov, P. Georgieva, Laboratory Model for Research on Computer Cluster Systems. 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, DOI: 10.23919/MIPRO.2018.8400250

[20] Mollova, S., Simionov, R., Seymenliyski, K., A study of the energy efficiency of a computer cluster, ICTRS'18, Proceedings of the Seventh International Conference on Telecommunications and Remote Sensing, pp. 51-54, Barcelona, Spain, 2018; ACM New York, NY, USA ©2018, ISBN 978-1-4503-6580-2, DOI: 10.1145/3278161.3278174

[21] Silvija Letskovska Kamen Seymenliyski and Pavlik Rahnev, Laboratory Equipment for Hydrogen Energy Education, ICEST 2014, XLIX Iinternational scientific conference on information, communication and energy systems and technologies, Serbia, Nis, June 25 - 27, 2014, Proceedings of Papers, Volume 1, ISBN: 978-86-6125-108-5, p. 315-320

[22] Ginko Georgiev, Silvija Letskovska, Kamen Seimenliyski and Pavlik Rahnev, Creating Laboratory Models for Auto Backup Power, ICEST 2017: 52TH International Scientific Conference On Information, Communication And Energy Systems And Technologies, Niš, Serbia, June 28-30, 2017 Proceedings of Papers, p. 362-366, SSN:2603-3259

[23] Kamen Seymenliyski, Silvija Letskovska, Renewable Energy Sources and Pricing of Electrical Power, Journal of Energy and Power Engineering (ISSN1934-8975) and online (ISSN 1934-7367) by David Publishing Company, US, 2014, Volume 8, Number 5, May 2014(Serial Number 78), p. 896-902

[24] Silviya Letskovska, Kamen Seymenliyski, Ginko Georgiev and Pawlik Rahnev, Laboratory Classes for saved Emissions of Greenhouse Gases, ICEST 2017, International Scientific Conference On Information, Communication And Energy Systems And Technologies, 28-30 June 2016, Nis, Serbia

Elastic collisions visualization using OpenCV object motion tracking

M. Hajba*, E. Ciriković*, M. Pecimotika*
*Virovitica College/Department of ICT, Virovitica, Croatia
marko.hajba@vsmti.hr, enes.cirikovic@vsmti.hr, mario.pecimotika@vsmti.hr

***Abstract* - Robotics is rapidly evolving scientific and engineering interdisciplinary field applied in numerous aspects of everyday life. Computer vision is a domain of robotics which primarily involves scientific knowledge of mathematics, electrical engineering and computing. Practical demonstration of computer vision use in the physical experiment of the elastic collisions is presented in this paper. Open source framework OpenCV on Raspberry Pi 4 platform was used for object motion tracking on specifically constructed elastic collision experimental setup. The experiment was used in course of Physics, Numerical Mathematics and Introduction to Robotics within our institution.**

Keywords – Elastic collisions, object tracking, visualization of the collision, OpenCV, Raspberry 4 Pi with Camera Module

I. INTRODUCTION

In physics, a collision is an interaction between two or more bodies where a part of the initial momenta and energies is exchanged between bodies involved in the collision. All collisions conserve momentum but what distinguishes them is conservation of kinetic energy. Thus, collisions can either be *elastic* (both the law of conservation of energy and the law of conservation of momentum are valid) or *inelastic* (only the law of conservation of momentum is valid). The degree to which a collision is (in)elastic is determined by the parameter called *coefficient of restitution*, i.e. the ratio of the final to initial relative velocity between two objects. In the case of perfectly elastic collisions, coefficient of restitution is one, which means that initial relative velocity and final relative velocity between the objects have the same value. Collision of two billiard balls is good approximation of elastic collision. An example of inelastic collision is a collision of two snowballs.

Computer vision is a field of computer science that works on enabling the computers to identify, see and process information in a similar way that humans do and provide an appropriate result [1]. In this paper practical use of computer vision in the education of future bachelors of Electrical Engineering and Computer Science at the Virovitica College is presented, providing computer-vision supported characterization of the elastic collision, as a practical demonstration of two fundamental laws of physics – the law of conservation of energy and the law of conservation of momentum.

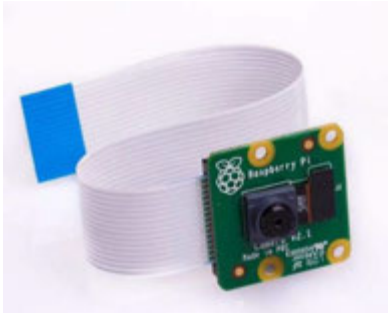


Figure 1. The Raspberry Pi Camera Module v2.1 [3].

OpenCV (*Open source Computer Vision library*) is a software library aimed at providing the necessary tools for solving computer-vision problems, written natively in C/C++ [2]. Today, it has C++, Java, MATLAB and Python interfaces supporting Android, Linux, MAC and Windows operating systems. OpenCV-Python is a Python wrapper for the original OpenCV C++ implementation that use NumPy, a library for numerical operations with a MATLAB-like styled syntax [3]. Python package `imutils`, a library providing the tools to make basic image processing functions such as translation, rotation, resizing, skeletonization, sorting contours and detecting edges, was also used.

The Raspberry Pi Camera Module v2.1 (Figure 1) with still resolution of 8 MP, connected with Raspberry Pi 4 is used to take videos, as well as still photographs, for time-lapse, slow-motion etc. It supports 1080p30, 720p60 and VGA90 video modes, as well as still capture [4]. The command line tool named `raspivid` was used to capture video.

II. EXPERIMENTAL SETUP

A. Construction of the experimental setup

The experimental setup (Figure 2) described below is entirely designed and constructed by the authors of this article. The experimental setup consists of two parallel wooden panels placed at an adjustable vertical distance with a maximum of 1 m. At the bottom, there is matte sleek black surface where collisions occur. Ball shooters are made of manual pinball plungers firmly attached to custom made wooden holders. Shooter holders are placed in such way that allows fine tuning of the initial coordinates and the angles. On the left and on the right edge hollow traps are installed where plastic balls are collected after collision. Painted ABS plastic balls with a diameter of 3.6 cm are used in this experiment.

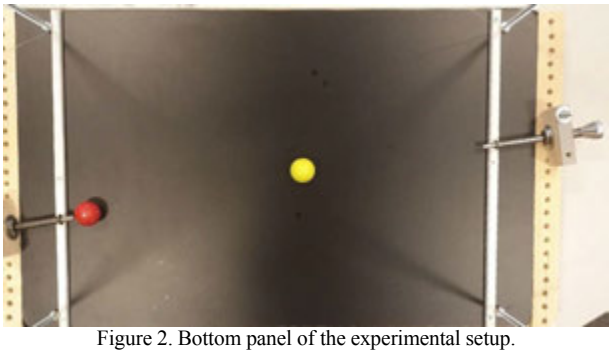


Figure 2. Bottom panel of the experimental setup.

At the bottom side of the top panel, the camera module along with the led illumination was installed, in order to improve the efficiency of the image capturing and to increase the capturing area (Figure 3). Capturing area is measured in centimeter to achieve realistic speeds and distances.

B. Central elastic collisions

As already mentioned in the introduction, in physics, a collision is an interaction between two or more bodies where a part of the initial momenta and energies is exchanged between bodies involved in the collision. Depending on energy and momentum conservation, collisions are characterized as elastic (both quantities are conserved) or inelastic (only momentum is conserved).

Collisions can also be typed as straight-line (central) collisions and non-central collisions (Figure 4). In central collision, the centers of gravity of the colliding bodies move along their connecting lines before and after the collision [5]. In non-central collisions, the centers of gravity of the colliding bodies move in different directions.

Consider two bodies of masses m_1 and m_2 moving in common along a straight path along the x-axis (i.e., central collision). During the collision, the total energy and the total momentum are conserved along the path. The vertical and horizontal component of velocity are, respectively:

$$v_{y,i} = v_i \sin\alpha = 0 \text{ m/s},$$

$$v_{x,i} = v_i \cos\alpha = v_i,$$

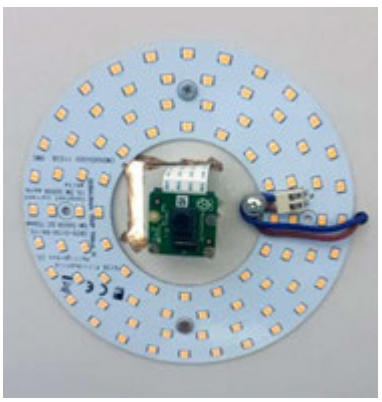


Figure 3. Camera Module v2.1 with the LED illumination.

where $\alpha = 0°$ is the angle between velocity vector and the horizontal axis, and $i = 1,\ 2$. Considering the above, the law of conservation of energy, with $v_{x,i}$ marked as v_i, yields:

$$\frac{1}{2}m_1 v_1^2 + \frac{1}{2}m_2 v_2^2 = \frac{1}{2}m_1 v_1'^2 + \frac{1}{2}m_2 v_2'^2, \quad (1)$$

where $\vec{v_1}, \vec{v_2}$ are velocities before collision, and $\vec{v_1'}$, $\vec{v_2'}$ are velocities after collision. The law of conservation of momentum yields:

$$m_1 v_1 + m_2 v_2 = m_1 v_1' + m_2 v_2'. \quad (2)$$

Combining (1) and (2) one can find the equations for velocities after the collision:

$$v_1' = \frac{m_1 - m_2}{m_1 + m_2} v_1 + \frac{2m_2}{m_1 + m_2} v_2, \quad (3)$$

$$v_2' = \frac{m_2 - m_1}{m_1 + m_2} v_2 + \frac{2m_1}{m_1 + m_2} v_1. \quad (4)$$

If $m_1 = m_2$ is inserted in (3) and (4), it yields:

$$v_1' = v_2, \quad (5)$$

$$v_2' = v_1, \quad (6)$$

i.e., in case of equal mass, the colliding bodies exchange their velocities. One should keep in mind that velocities in (5) and (6) have negative sign in the case of opposite traveling direction of masses.

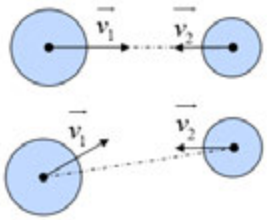


Figure 4. Elastic collisions. Top: Central collision. Bottom: Non-central collision.

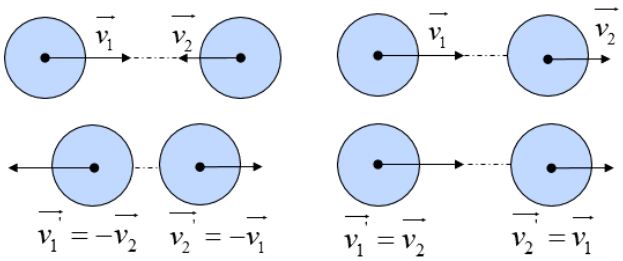


Figure 5. Elastic central collisions. Velocities exchange when masses traveling in opposite direction (on the left) and common direction (on the right).

C. Python script for collision detection

Script used to detect collisions and to calculate basic parameters was written in Python 3. Starting point is script explained on website [6]. In the script, single object motion tracking in real-time using Raspberry Pi and Camera Module is explained. OpenCV documentation and NumPy library knowledge was used to write the rest of the script. Script can process saved video file(s) or work real-time, depending on the used options during compilation. Yellow and red balls are chosen to be tracked, but this can be easily changed in the script. Ball colors are defined by RAL color matching system and converted in HSV color space using web converter. This is important as color masks are used to detect contours and its centers. Two separate windows are used, first one that streams captured video or works real-time and simultaneously is drawing object trajectories, and the second window where the visualization is shown. As soon as program detects two balls that should be tracked inside a single frame, it is calculating distance between centers of the balls. Moreover, coordinates of centers are saved during balls motion in an array, each center in an array respect to the tracked ball. This is important as it gives an opportunity to measure distance ball has travelled during single or multiple frames, simply by using Pythagorean theorem. If the distance is smaller than $2r+\varepsilon$, where r is the radius of the ball and ε is predefined measure for numerical error, on the visualization window message "COLLISION" is shown. One frame after the collision occurred (or more if it is defined), speeds and angles before and after the collision are calculated and shown on the visualization window.

Velocity (or speed –value is important, not direction, i.e. velocity is treated as a scalar), is calculated as "the distance traveled over the time". The formula "distance over the time" is used to describe a uniform straight-line motion. In our experiments, there is no guarantee that the motion of the balls is perfectly or completely uniform. However, if duration of the observed frame of time is very short, i.e. $\Delta t \to 0$, then this is a good approximation.

90 or 120 frames per second (FPS) videos are used which justify the given formula as time is small enough to consider it as a good derivative approximation of distance that ball had travelled. In order to calculate angles, standard idea of applying inverse tangent to distance in y-axis over distance in x-axis is used, respect to the proper frame before and after the collision occurred.

Number of FPS during the video capturing process is very important as capturing moment when collision occurred is extremely important in speed, angle and error calculation. Numerous tests have been done to achieve conclusions as Camera Module is inexpensive, but have limitations in FPS, focus speed, light sensitivity etc. On top of that, Raspberry Pi4 have its own limitation in terms of processing and memory speed, which should be considered with caution. This hardware limitations lead to limitations in experiments. The speed of balls should not be very large, camera should have enough time to obtain good focus and tracking becomes reliable, number of FPS should be adjusted accordingly (90 or 120 FPS for capturing video).

III. RESULTS

It was important to test many options, primarily number of FPS and resolution due to hardware limitations addressed in the previous section. Real-time visualization is hard to achieve due to the fact it works at around 20 FPS. It tracks object smoothly, but it is not reliable in collision detection. As a demonstration, let's take a situation where one ball is still, and second ball is travelling at a speed of 1 m/s. 20 FPS means that the moving ball will travel 1/20 m in one second, which is 5 cm. That is larger distance then a whole diameter of 3.6 cm. That complicates both the collision detection and parameters calculation just before and after the collision. Different options were tested in terms of FPS and video resolution. Central elastic collisions could be recorded at 120 FPS with a resolution of 600×100 pixels. In case of that, video area is narrowed, but this seems as a good choice as camera gets some time to adjust focus on the moving balls before collision occurs, and higher frame rates require smaller video resolution due to hardware limitations. In this setup, the radius of the ball takes 15 pixels, so a threshold distance of 35 pixels is chosen as a measure to detect collision. There is one more fact that should be pointed out. If the balls are slow, they could be near each other across multiple frames and collision would be detected multiple times, potentially resulting in a bad performance. Therefore, a control variable to prevent this to happen is used. One could compare distance between balls and take the smallest one inside collision detection distance range, but that is not perfect either. Message about collision and calculation would come later then supposed to in most cases, and there is a trap of taking data at the wrong time, after the collision already happened.

Fifty experiments of the central elastic collisions were performed, measuring previously mentioned parameters. The first type of experiment is when one of the balls is initially at rest, and the second type of experiment is when both balls have their respective initially velocities. Cases where the angle of deflection was $\leq 5°$ are considered, since $\cos(5°) \approx 1$ which satisfies central collision conditions. The script calculates the restitution coefficient in order to help us to analyze the physical background of a single event. The collision theory states that the restitution coefficient is equal to 1 in a case of perfectly elastic collision. However, it would be satisfactory if the restitution coefficient of 0.8 or better would be achieved. Indeed, the coefficient of restitution in the first experiment was in average 0.91, while the second type of experiment had an average of 0.79. During the experiments, it is noticed that previously mentioned problems could cause the restitution coefficient to be greater than 1. This could lead to a potentially better, but

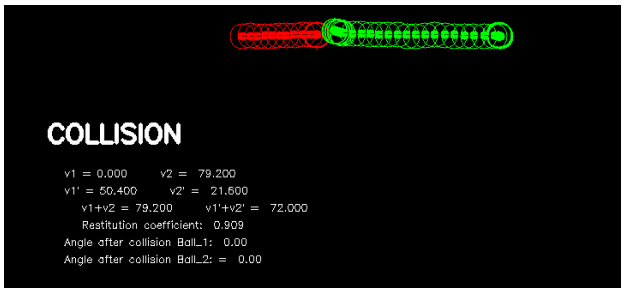


Figure 6. Example of the visualization window for the first type of experiment (the red ball is initially at rest), recorded after the collision.

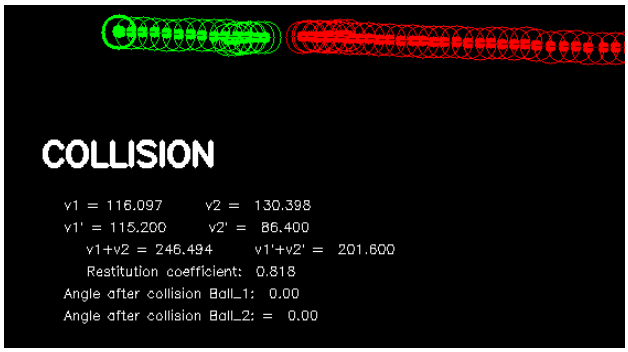


Figure 7. Example of the visualization window of the central elastic collision for the second type of experiment (both balls have their respective initial velocity), recorded after the collision.

false results in terms of average value of restitution coefficient. For example, if we consider two experiments with restitution coefficient 1.4 and 0.6, then average is 1, but the results are far from perfect. To ensure this is not an issue in our experiments, relative error is calculated as:

$$e_{rel} = \left| \frac{v_1 + v_2 - v_1' - v_2'}{v_1 + v_2} \right|. \quad (7)$$

In average, the relative error in percentage for the first type of experiments was 11%, while the second type of experiments yielded 25%, which is close to the results in terms of restitution coefficient. Second type of experiments yields higher error, which is expected since speeds are summing up which makes capturing the moment of collision less accurate.

IV. USE OF THE EXPERIMENT IN TEACHING ACTIVITIES

Raspberry Pi 4 with Camera Module is cost-effective, portable, but powerful enough to work on case studies in multiple areas with students. Collisions are important lecture inside Physics course. This experiment is used to demonstrate elastic collisions and can be used as an introductory example to illustrate uncertainties and types of error during experimental work. Moreover, this is good example to discuss difficulties someone may confront during experimental work. The experiment results are used in the course of Numerical Mathematics questioning different types of error that may be present while experiment is being performed. Also, significant digits, absolute and relative error is presented through the experiment. Raspberry Pi with Camera Module and OpenCV have a great potential to be used in course of Introduction to Robotics to study object motion tracking, recording and analyzing videos both real-time and postprocessing.

A. Physics

The course of Physics is taught during the first semester of the undergraduate professional study of the Electrical Engineering. The learning objectives are to develop basic knowledge in the core areas of physics including classical mechanics, electromagnetism, thermodynamics, atomic and nuclear physics, have the ability to apply the mathematics knowledge to problems in physics, have experience with modern laboratory methods, including experimental design and error analysis, basic scientific computing skills, and modern numerical techniques.

The third topic inside the course is Work, Power, Energy. Two hours of lectures and two hours of exercises are planned for the topic in the syllabus. Students are expected to acquire basic knowledge of the topic and among other things, should be able to:

- identify the mathematical quantities which effect the momentum and be able to calculate momentum,
- extend the third Newton's law in order to compare the momentum changes of two colliding objects,
- explain the meaning of the law of momentum conservation,
- explain the meaning of the law of energy conservation and to describe when the law does and does not apply to a collision,
- consider and describe different types of collision.

Learning outcomes are tested through class activities, homework assignments, written and oral exams, problem sets, written and oral laboratory and project reports, etc. Modern school is a creative and collaborative community created for the students, with the aim of developing divergent thinking, originality, and flexibility. It is a key element in the development of a society that encourages creative and modern thinking where new, divergent and original ideas and innovations are more than desirable. In modern society, the real goal of education is to teach students how to learn. In physics, this is mainly accomplished by the laboratory exercises.

Some research articles show that problem-oriented learning encourages the development of many skills in students, such as the ability of critical and logical thinking, scientific reasoning, experimental skills, and verbal expression (e.g. [7]). If the students' success is compared in the written and oral examinations, there is noticeable progress of students who studied the problem of the law of conservation of momentum through experiments in comparison to a generation before, which dealt with this problem from a theoretical point of view.

Students should start their discussion with the statement of the purpose of the lab experiment, providing

a brief theoretical explanation about expected results in different cases, perform the experiment measuring initial and final velocities giving the characterization of the collision and conclusions about whether mentioned laws are conserved or not, and submit their experimental results with error in written form. Discussion if the theoretical results fit within the experimental results is desirable.

B. Numerical Mathematics

The course of Numerical Mathematics is taught during the second semester of the undergraduate professional study of the Computer Science. The learning objectives are to develop knowledge about basic ideas and methods of numerical mathematics, to analyze the real-life problems and create corresponding mathematical models and give a review of the results.

The first topic inside the course is Errors. Two hours of lectures and two hours of exercises are planned for the topic in the syllabus. Students are expected to acquire basic knowledge of the topic and among other things, should be able to:

- classify types of errors and give examples,
- define and calculate absolute and relative error.

Learning outcomes are tested through class activities, homework assignments, written and oral exams, etc. In the course of Numerical Mathematics, this experiment is used as an example of different types of error during the auditory exercises performed in two groups of 15 students. The experiment is briefly described to the students, and then a link to the topic is given. Students should notice that the program code is executed on the computer, thus possible rounding errors and machine precision errors may be present. Furthermore, due to the imperfection of the measuring devices (Camera Module, latency, …), errors due to the inaccuracy of the input data may also be present. The method used to detect the moment of the collision is determined with a tolerance of several pixels, thus presence of a method error can be expected. Some physical phenomena such as air resistance and friction are neglected, which yields a model error. Error analysis is inevitable for making decisions about the validity of an experiment. The accuracy of the obtained approximation can be determined by a comparison of the obtained results with the expected measures. For this purpose, the absolute and relative errors should be calculated.

Example I: Let the velocities of the balls before and after the elastic collision are, respectively, 0 cm/s, 79.2 cm/s, and 50.4 cm/s, 21.6 cm/s. Calculate absolute and relative error of relative speeds before and after the collision if the balls were moving in opposite directions before the collision.

Initial relative velocity is calculated as:

$$x = v_1 + v_2 = 79.2 \text{ cm/s}.$$

Final relative velocity is calculated as:

$$\hat{x} = v_1' + v_2' = 72.0 \text{ cm/s}.$$

An absolute error is calculated as:

$$\Delta x = | x - \hat{x} | = 7.20 \text{ cm/s}.$$

A relative error in percentage is calculated as:

$$E_{rel} = \frac{| x - \hat{x} |}{x} \cdot 100\% = 9.09\% \cdot$$

Example II: Let the velocities of the balls before and after the elastic collision are, respectively, 0 cm/s, 139.5 cm/s, and 85.40 cm/s, 61.30 cm/s. Calculate absolute and relative error of relative speeds before and after the collision if the balls were moving in opposite directions before the collision. Compare the result with Example I and answer questions below.

- Do the theoretical results fit within the error of the experimental results?
- What is the discrepancy between the experimental and theoretical results? What is the relative error in the results?
- The experimental results are considered to be precise if the relative uncertainty is less than 0.1 (depending on the used experimental setup). Are the results accurate and/or precise?
- What are the reasons for errors in the experiments and how can they be reduced?

An absolute error is:

$$\Delta x = | x - \hat{x} | = 7.20 \text{ cm/s}.$$

A relative error in percentage is:

$$E_{rel} = \frac{| x - \hat{x} |}{x} \cdot 100\% = 5.16\% .$$

The relative error is smaller in Example II than it was in Example I. Students should conclude that relative error is a better indicator of error accuracy because it also considers the size of the data. The authors conducted a final student satisfaction survey on the Loomen platform. The average grade of the course increased by half compared to the grade year before. This yields the conclusion that students prefer to work with everyday problems and have a better understanding of the topic when it is presented and processed on the problem compared to the generic examples. A stronger hypothesis requires long-term surveys.

C. Introduction to Robotics

The course of Introduction to Robotics is taught during the sixth semester of the undergraduate professional study of the Electrical Engineering. The learning objectives are to train students to understand robotic systems and their application in everyday life, basic laws in robotics, kinematics and dynamics of robots, the structure of programming languages for robot control, computer vision and basic forms of artificial intelligence.

The last topic inside the course is Computer vision and image recognition algorithms. Two hours of lectures and two hours of exercises are planned for the topic in the syllabus. Students are expected to acquire basic knowledge of the topic and among other things, should be able to explain edge detection, object recognition, and robotic (computer) vision.

The authors would like to point out that this course has not yet been performed, given the situation that the undergraduate study of electrical engineering has been available for two academic years only. The authors hold that the presented experiment has good potential as an introduction to the topic of computer vision and image recognition algorithms. The above topic could cover the basics of OpenCV and examples that demonstrate the use of Python and the specified packages/methods in the field of robotics and/or computer vision.

D. Future work

Our plan is to use this project as role model to some new projects, but also to investigate an optimal frame rate, resolution, experiment with non-central and non-elastic collisions, and to involve students in our researches. There are some other physical experiments we would also like to focus on, e.g. the ticker-tape timer. The ticker-tape timer (Figure 6) is a device that tracks the object movement in one dimension. The timer is connected to the electrical power source (220 V, DC). Vibrating bar of the timer vibrates with the frequency of the electrical network (50 Hz), while the pin of the bar leaves a carbon paper trail on the tape. As the object moves, it drags the tape through the ticker leaving a trail of dots which represents the motion of the object as it changes its position over time. Since the spaces between the dots are equally distanced in time (only 1/50 of the second, since the frequency is 50 Hz), measuring the spatial distance of the dots, one can find the velocity and the acceleration of an object, and perform the analysis of body motion. An equal distance between dots indicates uniform motion. An increasing distance between dots indicates that the object was speeding up. A decreasing distance between dots means the object was slowing down over the time. The analysis of a ticker tape can help to characterize the motion of a body and to understand the specifics of different types of motion. Thus, the next step is to find a way to analyze the motion of an object using OpenCV framework, modeled on the traditional ticker-tape experiment, which may modernize the process of teaching physics and bring a field of kinematics to the students. Future upgrades may also include linear electromagnetic plunge system controlled by own-made software which should bring parameterized ball collisions and improve overall experiment precision.

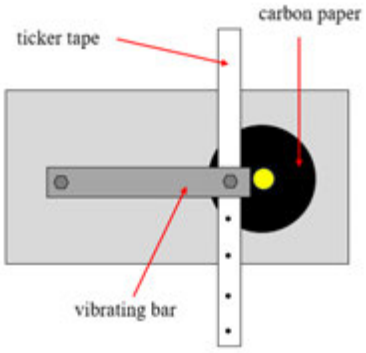


Figure 8. Ticker-tape timer. The strip is drawn to the bottom of page.

V. CONCLUSION

Raspberry Pi 4 with Camera Module was used to detect central elastic collisions of two plastic balls. Two types of experiment were performed. In order to detect collisions, an open source software library called OpenCV (*Open source Computer Vision library*) is used. It was noticed that real-time visualization is hard to achieve due to small FPS. Central elastic collisions could be recorded in 120 FPS with a resolution of 600 ×100 pixels. In average, the relative error for the first type of experiment was 11%, while the second type of experiment yielded 25%. This experiment and its results were used in the courses of Physics and Numerical Mathematics to achieve some of the core learning outcomes, and it has a great potential to be used in the course of Introduction to Robotics. It also brings numerous ideas for student projects.

REFERENCES

[1] "Computer vision ". Available: https://www.techopedia.com/definition/32309/computer-vision [Feb. 25, 2020]

[2] Kari Pulli, Anatoly Baksheev, Kirill Kornyakov, and Victor Eruhimov, "Realtime Computer Vision with OpenCV," *Queue* vol. 10, no. 4, pp. 40–56, April 2012.

[3] "OpenCV ". Available: https://docs.opencv.org/ [Jan. 21, 2020]

[4] "Raspberry Pi". Available: https://www.raspberrypi.org/ [Jan. 22, 2020]

[5] W. Benenson, Handbook of physics. New York, NY: Springer, 2001.

[6] "OpenCV track object movement". Available: https://www.pyimagesearch.com/2015/09/21/opencv-track-object-movement/ [Dec. 1, 2019]

[7] A. Letina, "Efikasnost istraživački usmjerene nastave Prirode i društva u razvoju prirodoznanstvene kompetencije učenika," Croatian Journal of Education, vol. 18, no. 3, pp. 665–696., October2015

Educational computer games and Gamification at the higher education – students' points of view

D. Tuparova*, G. Tuparov ** and D. Orozova***
* South-West University "Neofit Rilski", Department of Informatics, Blagoevgrad, Bulgaria
** New Bulgarian University, Department of Informatics, Sofia, Bulgaria
***Burgas Free University, Department of Informatics, Burgas, Bulgaria
e-mail address: ddureva@swu.bg

Abstract: Educational computer games (ECG) and gamification, step by step take place in educational process at all educational levels – from kindergarten to university level. And if in preschool and primary school age, games are a key method of learning, then in higher education games are seen as means of entertainment first and foremost. In this article, we present the results of a study regarding university students' perceptions and preferences for educational computer games and gamification. The study was conducted among 347 students in various professional fields at Bulgarian universities. The main research questions are related to the kind of end user devices used for playing computer games; type of game elements that are most preferred by the university students; appropriateness of ECG different learning situations; etc. The answers to these questions help to design and study the usability of educational computer games applicable to university education taking into account the professional area of the students.

Keywords **– gamification; educational computer game; higer education; student's perception**

I. INTRODUCTION

Educational computer games (ECG) and gamification, step by step take place in educational process at all educational levels – from kindergarten to university level. And if in preschool and primary school age, games are a key method of learning, then in higher education, games are seen as means of entertainment first and foremost. In this article, we present the results of a study regarding university students' perceptions and preferences for educational computer games and gamification.

Most of the studies related to students' and educators' perception of ECG and gamification are directed to primary and secondary schools. However, during the last five years the number of studies directed towards development and implementation of ECG and gamification at higher education level has rapidly increased. With the key search statement – 'student AND (perception OR attitude OR acceptance) AND game' we found in SCOPUS database around 2600 documents and 1282 of them have been published during the last five years. Approximately 390 of all documents, published during the last five years include key phrase "higher education" (27.01.2020). For educators and ECG developers it is important to analyze students' requirements and preferences towards the game elements, the type and utility of ECG.

The main purpose of this paper is to present the results of a survey directed towards identifying of university students' perceptions and point of view about ECG and game elements application in higher education.

The study was conducted among 347 students in various professional fields at two Bulgarian universities.

The main research questions are:

- What kind of end user devices do they use for computer games?
- Which game elements are most preferred and which are not preferred by the university students?
- In what learning situations are educational computer games appropriate?
- To what extent are games or game elements used in the educational process?
- Is there a difference in the attitudes and preferences of students in different majors?
- Is there a difference in preferences of students who use computer games for fun and those who do not?

The answers to these questions help to design and research the usability of educational computer games applicable to university education.

II. BRIEF LITERATURE REVIEW

A. Basic concepts

1) Serious games

A variety of definitions and viewpoints about serious game exists [1]. One core definition is proposed by Zyda [2]: the serious game is described as a "mental contest, played with computer in accordance with specific rules that uses entertainment to further government or corporate training, education, health, public policy, and strategic communication objectives".

2) Educational computer game (ECG)/ didactical computer game.

This study is funded by Bulgarian National Scientific Fund, Project contract DN-05/10, 2016

Under this term we consider serious computer games, applicable in educational subject-domain. The ECG has clearly specified educational objectives and outcomes. It is meant for training, learning or assessment of competences and set of learning objectives [3].

3) Gamification

In the work of Kapp [4] gamification is defined as "using game-based mechanics, aesthetics and game thinking to engage people, motivate action, promote learning and solve problems." The authors of [5] point out that gamification is: "The use of game elements and game thinking in non-game contexts." Gamification is a concept for introducing game mechanics into a variety of processes and areas such as business, marketing and education. The aims of gamification are to motivate, engage and attract users [6]. Gamification is popular in different areas of education [7], it is an innovative tool for education[8], and is extensevely used in education [9]. Mora et al [10] consider gamification as "use of game-like properties in learning scenarios".

4) Game elements

In literature different frameworks for description of game elements and game characteristics are defined. According to [11] more than 50 taxonomies of game elements exist, but not more than 10 are related to education. Some popular taxonomies are "Periodical Table of Gamification" developed by Marczewski [12] with 52 game elements; frameworks proposed and discussed in [10], [13] consists of 22 elements and characteristics grouped in 5 groups – knowledge, logic, psychology, measurement, interaction.

We focus our study on some game elements and characteristics such as: graphical design, logic and plot; audio support; possibility to play computer game on different end user devices; many gamers to play at the same time; possibility to play online; awards.

5) ECG and Gamification in higher education

Games and gamification are challenges for higher education [10]. The studies related to the application of ECG and gamification in higher education could be classified in the following groups:

- studies about process of development of ECG and gamification tools;
- analysis of the achievements and results from the implementation of developed ECG/serious game or learning scenarious with usage of gamification in some subject-domain as nursing or medical education, teacher training, engineering, software engineering and programming etc;
- analysis of the attitudes of stakeholders – lecturers and students, towards already used ECG or gamification in the university subject;
- mapping studies about gamification, seriuos games or ECG.

In Bulgaria several studies about attitudes and perception of different stakeholders towards gamification and computer games in education have been done. These studies are focused on the students and teachers at primary and secondary school [3], [14] and specialists that work with pupils with special educational needs [15], [16].

III. METHODOLOGY

A. Target groups

About 400 students from two Bulgarian universities have been invited to fill online questionnaires and a total of 347 students responded to the invitation. The profile of the respondents is presented in the tables I. and II. Students' areas of study are grouped according to Bulgarian Higher Education law act professional fields.

About 79% of the respondents indicate that they play entertainment computer games.

B. Research questions

The main research questions analyzed in the paper are:

RQ1. What kind of end user devices do they use for computer games?

RQ2. Which game elements are most preferred and which are not preferred by the university students?

RQ3. In what learning situations are educational computer games appropriate?

RQ4. To what extent are games or game elements used in the educational process?

RQ5. Is there a difference in the attitudes and preferences of students in different majors?

RQ6. Is there a difference in preferences of students who use computer games for fun and those who do not?

C. Used methods

We used a survey with 53 items, two of them with an open answer. The items are grouped in 8 groups: Profile of the students; Used devices for game playing, Important element and characteristics of games; Educational activities in which ECG have been used; Educational activities in which game elements have been used; Used

TABLE I. DISTRIBUTION OF THE RESPONDENTS REGARDING AREA OF STUDY

Area of study	Number	%
Pedagogy	36	10,37%
Humanity	27	7,78%
Social, Economics, Law	105	30,26%
Computer science, Math, Natural Sciences	137	39,48%
Technical science	39	11,24%
Art	2	0,58%
National security	1	0,29%

TABLE II. DISTRIBUTION OF THE RESPONDENTS REGARDING HIGHER EDUCATION LEVELS

Higher education levels	Number	%
1-2 year/bachelor degree	207	59,65%
3-4 year/bachelor degree	124	35,73%
Master or PhD degree	9	2,59%
Post diploma qualification	7	2,02%

types of ECG; Utility of ECG at university level of education; Utility of usage of ECG and game elements via smartphones; General attitude about usage of ECG and game elements at higher education.

The used types of items are: Dichotomous and 5 level Likert Scale type of items.

For data analysis, the following statistical methods are applied: Cronbach's alpha for reliability evaluation, Descriptive statistics, Non-parametric Methods – due to ordinal data. The analysis is conducted with MS Excel, SPSS and Orange application software.

IV. FINDINGS AND RESULTS

A. Reliability of the survey instrument

The coefficient Cronbach's alpha – α for the 50 Likert Types of items is 0,932. In subgroups coefficients Cronbach's alpha are presented in Table III.

B. Analysis of research questions

1) RQ1. What kind of end user devices do they use for computer games?

Students prefer to play games on smartphones (70%) and laptops (57,1%). (Table IV).

Regarding the educational level we consider data for 1-2 year/bachelor, 3-4 year/bachelor and joined data for Master students and Post-diploma qualification students.

To find out whether the year of study is a factor on the usage of end user devices we applied χ^2 test for independent samples. There is no statistically significant difference regarding the use of PC, laptop, smartphone, and tablet in the three groups mentioned above. (Table V). A statistically significant difference is observed regarding gaming consoles.

In the group of 1-2 year of Bachelor students about 28% use gaming console, 16% in the group of 3-4 year, and 19% in group of master and post-diploma students use gaming console. A statistically significant difference exists between the group of 1-2 years students and the group of 3-4 year students: $p = 0{,}013619 < 0{,}05$.

TABLE III. CRONBACH'S ALPHA COEFICENTS FOR SUBGROUPS OF ITEMS

Subgroup	N	α
Used devices for game plying	5	0.602
Important element and characteristics of games	7	0.789
Educational activities in which ECG have been used	4	0.935
Educational activities in which game elements have been used	4	0.925
Used types of ECG	8	0.880
Utility of ECG at university level of education	10	0.945
Utility of usage of ECG and game elements via smartphones	4	0.872
General attitude about usage of ECG and game elements at higher education	4	0.897

TABLE IV. PREFERRED ENDUSERS DEVICES FOR GAME PLAYING

End user device	Yes	No
Personal computer	38,60%	61,40%
Laptop	**57,10%**	**42,90%**
Smartphone	**70,00%**	**30,00%**
Tablet	22,20%	77,80%
Gaming Console	23,30%	76,70%

TABLE V. PREFERRED ENDUSERS DEVICES FOR GAME PLAYING BY YEAR OF STUDY

End user device	1-2 year	3-4 year	Master and Post-diploma qualification	p, χ^2 test
Personal computer	86	41	7	0,28087
Laptop	129	61	8	0,055197
Smartphone	151	79	13	0,125003
Tablet	48	27	2	0,605916
Gaming consoles	**58**	**20**	**3**	**0,042327**

2) RQ2. Which game elements and characteristics are most preferred and which are not preferred by the university students?

The most preferred game elements and characteristics are logic and game plot, and graphical design. In contrast awards in the game are less preferred. (Table VI.)

3) RQ3. In what learning situations are educational computer games appropriate?

According to the respondents ECG are appropriate in all educational situations, mentioned in the survey. (Table VII)

Regarding the utility of ECG at university educational level across groups of study field, there is a statistically significant difference in opinion of the students about knowledge acquisition (Fig. 1). The students from the field of Pedagogy evaluate the utility of the ECG for knowledge acquisition at university level as highest (Median = 5).

TABLE VI. PREFERRED GAME ELEMENTS AND CHARACTERISTICS

Game elements/ characteristics	Mean	Median	Mode	Std. Deviation	Variance
graphical design	**4,18**	**5,00**	**5**	**1,094**	**1,197**
logic and plot	**4,40**	**5,00**	**5**	**1,005**	**1,010**
audio	3,40	4,00	3	1,307	1,709
possibility to play computer game on different end user devices	3,48	4,00	5	1,469	2,158
many gamers to play at the same time	3,65	4,00	5	1,243	1,544
possibility to play online	3,83	4,00	5	1,302	1,696
awards	*3,12*	*3,00*	*3*	*1,387*	*1,923*

TABLE VII. UTILITY OF ECG IN DIFFERENT EDUCATIONAL SITUATIONS

Utility of ECG at university educational level for:	Mean	Median	Mode	Std. Deviation	Variance
knowledge acquisition	3,57	4,00	5	1,387	1,925
skills acquisition	3,93	4,00	5	1,189	1,415
creativity development	3,91	4,00	5	1,196	1,432
decision making skills development	3,84	4,00	5	1,271	1,616
development of trial and error skills	3,79	4,00	5	1,285	1,651
assessment of own ideas	3,77	4,00	5	1,249	1,559
development motivation	3,72	4,00	5	1,324	1,754
understanding that failure is a way of learning	3,64	4,00	5	1,313	1,725
development of critical thinking and problem solving	3,72	4,00	5	1,299	1,687
development of communication skills	3,73	4,00	5	1,316	1,731

4) RQ4. To what extent are games or game elements used in the educational process?

About 35% of the students report that ECG have not been available to them for mastering of knowledge, for 41% of the students ECG have not been available for self-learning, for 42% of the students ECG have not been available for assessment and about 46% do not have access to ECG for self-assessment.

Regarding the usage of game elements (GE) at the university education level, about 41% of the students indicated that GE have not been available to them for knowledge mastering, for 40% of the students GE have not been available for self-learning, for about 41% of the students GE have not been used for assessment and for 46% of respondents GE have not been applicable for self-assessment.

There is a statistically significant difference regarding the used ECG and GE for knowledge mastering and assessment by the area of study. Due to the small number in Art (2 respondents) and National Security (1 respondent) these data are excluded from the analysis. To compare distributions of students' answers we applied Median test. Students from study fields in Pedagogy and Social science, Economics and Law indicated higher level of usage of ECG for knowledge mastering and assessment activities at the university (Median is 4 for both groups). (Figure 2.)

5) RQ5. Is there a difference in the attitudes and preferences of students in different majors?

Regarding the preferences of students in different majors to the game elements and characteristics, we can mention that there is only a statistically significant difference for the use of audio in ECG. The Independent Samples Kruskal-Wallis test was applied. (Table VIII)

The results indicate that there is no statistically significant difference regarding the importance of audio for students in fields of Pedagogy; Humanity; Social

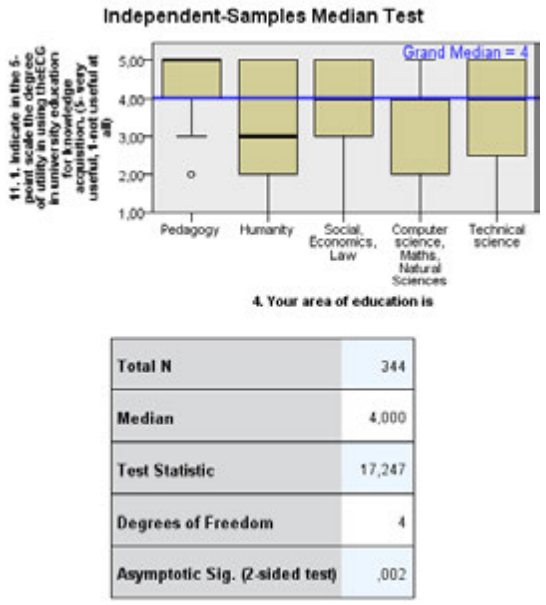


Figure 1. Median test for utility of the ECG in knwoledge mastering across study field

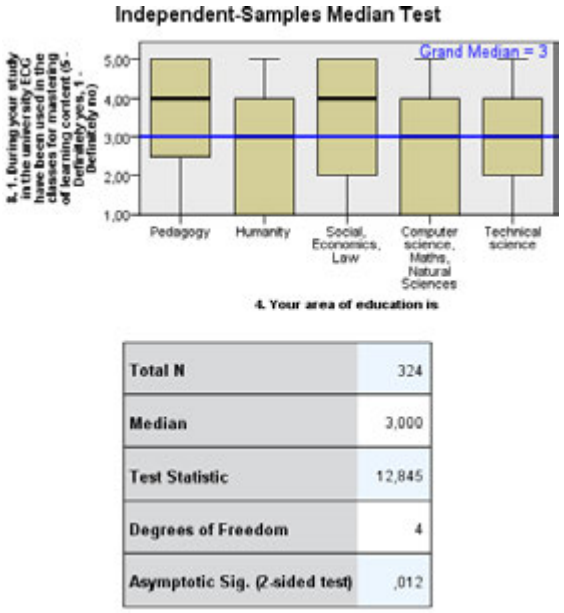


Figure 2. Median test for usage of the ECG in knwoledge mastering by area of study

TABLE VIII. INDEPENDENT SAMPLES KRUSKAL-WALLIS TEST FOR PREFERRED GAME ELEMENTS AND CHARACTERISTICS BY MAJORS

Test Statistics[a,b]			
Importance of Game elements/ characteristics	Chi-Square	df	Asymp. Sig.
graphical design	3,151	4	,533
logic and plot	9,257	4	,055
audio	**10,016**	**4**	**,040**
possibility to play computer game on different end user devices	9,206	4	,056
many gamers to play at the same time	1,456	4	,834
possibility to play online	3,798	4	,434
awards	2,676	4	,613
a. Kruskal Wallis Test			
b. Grouping Variable: 4. Your area of education is			

science, Economy and Law; Computer science, Mathematics and Natural Sciences. P-value obtained by Independent samples Kruskal-Wallis Test is 0.222, $\chi 2$=4.391, df=3. Totally different is the opinion of students from Technical field of study. About 69% of them responded that audio in the games is important or very important.

Statistically significant differences were not observed in the general perception of ECG and game elements by students from different fields of study. (Table IX)

6) RQ6 Is there a difference in the preferences of students who use computer games for fun and those who do not?

Statistically significant differences exist in the opinion of students who use and those who do not use computer games for fun regarding game elements and characteristics – graphical design, logic and game plot, audio, possibility many gamers to play at the same time, possibility to play online. For students who play games for fun the most important characteristics are logic and game plot, and for those who do not the most important are logic and game plot and graphical design. In the latter group, there is no statistically significant difference between the proportions of 63% (graphical design) and 72% (game plot and logic). The $\chi 2$ test for proportions comparison is applied and p=0.2573.

Less important for both groups are awards. Furthermore, for the students who do not play computer games for fun audio in the game is less important. (Table X)

Statistically significant differences exist in the perception of ECG and game elements in higher education between students who play and who do not play CG for entertainment. The students who play CG for entertainment demonstrate more positive attitude regarding the usefulness and suitability of ECG and game elements. (Table XI).

TABLE IX. INDEPENDENT SAMPLES KRUSCAL-WALLIS TEST FOR GENERAL PERCEPTION OF GAME ELEMENTS AND CHARACTERISTICS BY SPECIALTIES

Test Statistics[a,b]			
General acceptance of ECG and game elements - Definitely yes, Yes	Chi-Square	df	Asymp. Sig.
You would like to use an ECG in your training	5,272	4	,261
You would like to use game elements in your training	5,436	4	,245
ECG are suitable for university education	8,072	4	,089
Game elements are suitable for university education	4,257	4	,372
a. Kruskal-Wallis Test			
b. Grouping Variable: 4. Your area of education is			

TABLE X. PREFERRED ELEMENTS AND CHARACTERSISTICS OF ENTERTAINMENT GAMES NG BY STUDENTS WHO PLAY AND DO NOT PLAY GAMES FOR FUN

Important and very important is	Play	Do not play	p-value, $\chi 2$
graphical design	**81%**	**63%**	**0,000**
logic and plot of the CG	**88%**	**72%**	**0,000**
audio	**53%**	**38%**	**0,002**
possibility to play at different end user devices	56,0 %	54,9%	0,480
possibility many gamers to play at the same time	**62%**	**49%**	**0,018**
possibility to play online	**70%**	**58%**	**0,026**
awards	41%	39%	0,105

TABLE XI. PERCEPTION OF ECG AND GAME ELEMENTS AT HIGHER EDUCATION.

General acceptance of ECG and game elements - Definitely yes, Yes	**Play**	**Do not play**	**p-value, $\chi 2$**
You would like to use an ECG in your training	66%	49%	**0,000**
You would like to use game elements in your training	74%	59%	**0,001**
ECG are suitable for university education	70%	48%	**0,008**
Game elements are suitable for university education	70%	52%	**0,012**

V. CONCLUSION

In this paper, we analyzed some of the findings and results obtained from a survey among 347 students from different fields of studies at two Bulgarian universities. The results indicate that:

- students in all five fields of studies prefer to play games on smartphone and laptop. They wish to be mobile;
- the most preferred game elements and characteristics are logic and game plot, and graphical design;
- awards and audio belong to the less preferred game elements and characteristics, but for the students from the Technical field of study, audio in the game is more important.

These findings give some directions for educators and developers of ECG for higher education, e.g. ECG has to provide mobile or responsive interface; ECG has to offer interesting game plot, logic and attractive design; ECG for subjects in the engineering field has to provide audio with possibility to switch on/off.

Future work will be directed to study differences in university students' opinion with respect to gender.

ACKNOWLEDGMENT

The study is funded by Bulgarian National Scientific Fund, Project contract DN-05/10, 2016.

REFERENCES

[1] Djaouti, D., Alvarez, J., & Jessel, J.-P. (2011). Origins of Serious Games. In M. O. Ma, *Seriuos Games and Edutainment Applications*. London: Springer - London.

[2] Zyda, M. (2005). From visual simulation to virtual reality to games. *IEEE Computer, 38*(9), 25–32.

[3] Tuparova, D., Tuparov, G., Veleva, V., & Nikolova, E. (2018). Educational computer games and gamification in informatics and information technology education - Teachers' points of view. *41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018* (pp. 766-771). Opatija: Institute of Electrical and Electronics Engineers Inc. doi:10.23919/MIPRO.2018.8400142

[4] Kapp, K. (2012). *The Gamification of Learning and Instruction: Game-based Methods and Strategies for Training and Education.* New York: Pfeiffe.

[5] Werbach, K., & Hunter, D. (2012). For the win: How game thinking can revolutionize your business. *Philadelphia: Wharton Digital Press*.

[6] Stoyanova , M., Tuparova, D., & Samardzhiev, K. (2017). Gamification in 11th Grade Mathematics Lessons – One Possible Interactive Approach. In M. Auer , D. Guralnick , & J. Uhomoibhi (Eds.), *Interactive Collaborative Learning. ICL 2016. Advances in Intelligent System.*

[7] Kasurinen, J., & Knutas, A. (2018). Publication trends in gamification: A systematic mapping study. *Computer Science Review, 27*, 33-44.

[8] Terzieva, T., Golev, A., & Stavrev, C.,Serious games – innovative tool for education. In proc. Innovative software tools and technologies with application in matematics, informatics, and education. 2017, pp. 104-107, University of Plovdiv "Paisii Hilendarski"

[9] Dichev, C., & Dicheva, D. (2017). Gamifying education: What is known, what is believed and what remains uncertain: A critical review. International Journal of Educational Technology in Higher Education.

[10] Mora, A., Riera, D., Gonzalez, C., & Arnedo-Moreno, J. (2017). Gamification: a systematic review of design frameworks. *Journal of Computing in Higher Education, 29*(3), 516–548. doi: https://doi.org/10.1007/s12528-017-9150-4

[11] Toda, A., Klock, A. S., & Oliveira, W. (2019). Analysing gamification elements in educational environments using an existing Gamification taxonomy. . *Smart Learning Environment 6, 16* . doi:https://doi.org/10.1186/s40561-019-0106-1

[12] Marczewski, A. (2015). Even ninja monkeys like to play: Gamification, game thinking and motivational design. *Gamified UK*, 65-80.

[13] Toda , A. M., do Carmo , R. M., da Silva, , A. P., Bittencourt, I. I., & Isotani, S. (2019). An approach for planning and deploying gamification concepts with social networks within educational contexts. *International Journal of Information Management, 46*, 294-303.

[14] Paunova-Hubenova , E., Terzieva, V., Boneva, Y., & Dimitrov, S. (2018). Trends in the Application of Educational Games in Bulgaria in the Last Five Years. *National Conference on "Education and Research in the Information Society"*. Plovdiv.

[15] Stankova, M., Ivanova, V., & Kamenski, T. (2018). Use of Educational Computer Games in the Initial Assessment and Therapy of Children with Special Educational Needs in Bulgaria. *TEM Journal, 7*(3), 488-494. doi:10.18421/TEM73-03

[16] Tuparova, D, Stankova, M., Kamenski, T., Todorova, E., & Tuparov, G. (2019). Attitudes toward the use of educational computer games in the initial assessment and terapy of children with special educational needs. *Pedagogy-Pedagogika, 91*(5), 725-735.

Using Moodle e-learning platform in Mechanical Engineering lectures

P.Tomšič *, I. Demšar ** and T. Finkšt***
Faculty of Mechanical Engineering, University of Ljubljana, Ljubljana, Slovenia
* pavel.tomsic@lecad.fs.uni-lj.si
** ivan.demsar@lecad.fs.uni-lj.si
*** tomaz.finkst@lecad.fs.uni-lj.si

***Abstract* – With implementation of European Union's wide reforms of the educational system in the last decade, the academic lectures and increased complexity of the overlapping curriculum has become increasingly demanding. Usage of the online e-learning platforms in Higher Education has become an important part of the flexibility in modern educational system. There are many different e-learning platforms, both open source and commercially available. In the Laboratory LECAD (laboratory for Computer Aided Design) we have decided to use open source software Moodle for handling practical exercises of different academic lectures. The e-learning platform has been in use since 2015 and gradually adapted for many lectures, including traditional approach of weekly practical lectures as well as project-based lecture management. The paper describes impact of the implementation of Moodle on the curriculum, its advantages as well as its weaknesses. Each academic year more than 600 students were enrolled in Moodle. The advantages offered through online management tools have made lectures more attractive and accessible and has simplified keeping the topics being up to date. Having long time experience in conducting courses with usage of e-learning in the field of Mechanical Engineering it is hard to imagine e-learning as a full supplement of the traditional approach in educational system.**

Keywords – e-learning; digitalization, LMS, higher education in Slovenia, Moodle, Mechanical Engineering lectures.

I. INTRODUCTION

Constant implementation of new technology, digitalization and emergence of internet have changed the teaching and learning process in Higher Education (HE) to their core [1, 2]. Books, prints and other literature are mostly replaced with digital documents that offer more content and interactivity (videos, surveys etc.). Rapid expiration of knowledge and training combining with technological advancement has led to a necessity of just-in-time information delivery, greater flexibility of courses and rapid change of the presented topics. Due to higher complexity of curriculum and associated need for cost-effective teaching methods the classical forms of teaching and learning practices needed to be transformed.

In recent years the number of courses offered fully or partially (hybrid or blending) online by HE institutes has dramatically increased. Everyone seems to be interested in online available topics, a move that corresponds with EU effort and financing for digitalization, internationalization and more accessible educational system for everyone.

For HE institutes to offer such courses it is important to have an effective technology platform to support the online class environment and activities. Nowadays many technologies are used to support online education, particularly Learning Management Systems (LMS), which serve as the e-learning platforms. The LMS can be used for online education as well as a supplement for face-to-face courses. Popular LMS systems include Blackboard, Moodle, Sakai, Desire2Learn, Canvas, and eCollege.

In the LECAD Laboratory we have been using open source LMS software Moodle (Modular Object-Oriented Dynamic Learning Environment) since 2015. It is a popular e-learning solution used in schools, universities, workplaces and other sectors. Advantages of using such a platform include enrolling of blended learning, distance education, flipped classroom and other e-learning education activities. During the first year we have tested it on traditional learning courses and through the years we have adopted its usage also for project based courses.

The paper is organized in six sections. After introduction, e-learning background is briefly described. Implementation of e-learning platform Moodle for management of academic courses in laboratory LECAD is presented. In the fourth chapter statistics from different courses, classical weekly practical lectures and project based through years are analyzed. Teaching assistants, responsible for these courses, have also made SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis according to their gained experience gathered through the years. In conclusion final remarks on e-learning software inclusion are presented.

II. ABOUT E-LEARNING

A. Perspectives

One form of modern distance or remote learning presents e-learning. It allows users to be involved in learning independent from place and time. It has become an important part of the modernization and digitalization of traditional education, carried out by HE throughout EU. E-learning activities can be divided into two groups:

a) e-learning in a broad sense, and

b) narrower e-learning.

The first view of treating e-learning is with perspective towards any education principle that has a technological component (e-learning in a broad sense) [7]. The perspective defines any ICT (Information and Communications Technology) supported education, in form of either CDs, e-books, PowerPoint slides, PDFs, etc. Such form of education is a part of traditional principles and is used at all stages of educational systems. The goal of ICT inclusion is to enrich the learning process with additional sources. Through this perspective, e-learning could be referred to partly technology enhanced education [3].

In the second view we narrow the perspective upon e-learning. Marc Rosenberg et al. [4] defined e-learning as the usage of online technologies in various solutions to enhance knowledge or enhance educational activities. The essential difference between technology enhanced learning and the narrower concept of e-learning (referred to also as integrated e-learning) is conceptual. In integrated e-education, technological support is not only partial, but is fully integrated in all elements of the educational process. Thus it is involved in pedagogical and administrative support as well as distribution of learning material. A concept that enables the learning process to be carried out with the physical separation of the teacher and the participant(s). Blended learning can also be used for interaction between partially technology-enabled educations and integrated e-learning.

In Figure 1 educational principles are divided into four groups [3]: traditional education, partial technology education, combined education and integrated e-education.

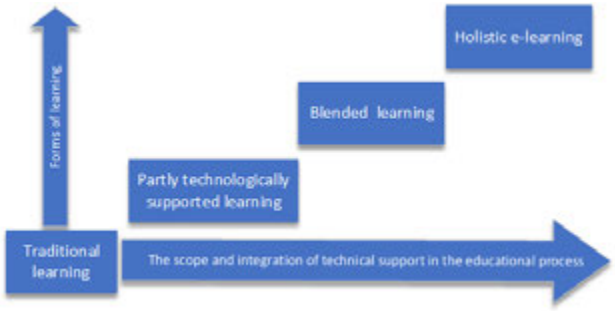


Figure 1. The scope and degree of integration of technological support in different forms of education

In the literature [5], e-learning can be defined as a workflow process between different participants: teachers/ mentors, students, e-education developers and others, especially administrative, technical and professional staff performing e-learning supporting activities.

Figure 2 shows which basic processes are carried out in an educational institution. There are three basic processes:

1) development of e-learning, which includes the distribution of e-learning,
2) implementation of e-learning,
3) evaluation of e-learning.

All three processes are carried out at the level of the individual course and at the level of the institution implementing the program.

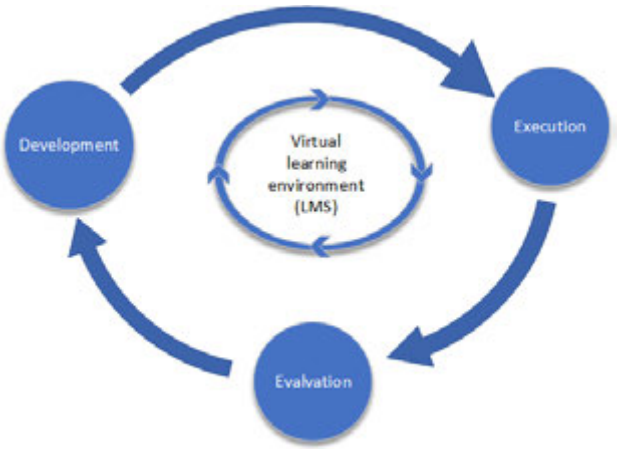


Figure 2. Interconnection of a basic e-learning processes.

B. *Benefits of e-learning*

General expectations from e-learning implementation are quite high. Its inclusion is expected to bring many benefits to participants and educational institutions, and more broadly, to education systems at the national level and internationally. E-learning is cost-effective and also saves time. By educating ourselves electronically, we reduce time spent in the office, classroom, etc., thus reducing travel and printing costs. The transition should increase productivity increase participant's satisfaction and focus. Digitalized education can be done anytime, anywhere.

However, traditional lectures are held only during office hours. E-learning allows students to process the topics outside the classroom and organize their own free time. In a virtual learning environment, student progress can be more efficiently assessed. Knowledge can be tested in real-time on a larger group simultaneously. Interactive materials can be more interesting than traditionally used materials. They can be enriched with pictures, animations, video presentations, quizzes, audio files, etc.

With online knowledge tests and home assignments, the students are also more motivated. The most significant advantages of e-learning from the perspective of education participants [3] are:

- increased flexibility in time, place, pace and content of education (just-in-time, just-in-place);
- increased interactivity and faster access to knowledge from various sources (synchronous and asynchronous forms of communication, online resources);
- the ability to adapt learning approaches to individual needs;
- transparency of educational conditions;
- development of new knowledge and competences.

C. *E-materials and the learning enviroment*

E-learning requires both electronic materials and suitable online learning environment. Most commonly, online classrooms are used as a learning environment.

A more specific definition for the e-material used in academics would state that is a digitalized material in various forms (audio, video, text, book, etc.) with a clear learning objective and in a LMS it presents a digital learning resource. The quality of e-materials, however, plays an important role in learning process. The higher the quality of the materials, the greater is expected the student's final knowledge to be.

If we want to create quality e-material, we must pay attention to the steps in preparing only the e-material (thus focusing our work). The preparation of the material is done in several steps: content preparation, design preparation, selection and inclusion of media in the program, and editing the LMS. This means the staff involved is required to have additional responsibilities and objectives compared with traditional learning approach.

The first step in the preparation of e-materials is the content preparation, which includes a detailed definition of the purpose and learning objectives of the program, the definition of basic content sections and their breakdown into narrower units and the choice of learning activities. When preparing the material, the design and graphic design of the text and web pages has to be defined, with addition of which media will be used. Activities related to content, design and media selection are often carried out in practice simultaneously, but the design and technological solutions must be content-based. The starting point and the learning objectives of the program must always remain in the forefront.

It has to be noted that although adoption of e-learning platforms throughout HE is extremely popular, there is little evidence to suggest that their implementation is actually improving students learning performance [6]. However, such approach encourages students to be creative and raises the satisfaction of the acquired knowledge [7].

III. IMPLEMENTATION OF E-LEARNING PLATFORM MOODLE

As Moodle is an open source software its purchase cost is null. Moodle requires a platform running Linux, Apache, MySQL and PHP- this set of applications is often referred to as a LAMP. We have installed it on our own laboratory server and are administrating it ourselves.

Main reasons why we have adapted e-learning platform are:

- Problems with management of high number of students,
- Need for constant updates of topics,
- Management the diversity of different courses,
- Need for defragmentation of students into small groups,
- Data management (in most of our courses students produce new data each week).

One of the challenges that we have faced presented the initial lack of discipline of the students enrolled in e-learning courses. Since detection of the issue we have successfully overcome the problem with implementation of weekly assessments.

Assessment of online activities can be challenging, as there is perception of a greater risk of cheating as in traditional face-to-face courses. We have overcome this issue by implementing a simple code for proof-checking the authorship as well as checking the dates and other specific properties of the electronic documents. Since its implementation the number of cheatings has dropped dramatically.

It seems as implementation of e-learning in Slovenia is behind in comparison to European countries [8] - 91% adaptability of EUA members in 2013 compared to 71% adaptability in Slovenia in 2017. From our perspective the adoption is problematic from operational side – there are issues to what extent LMSs could be efficiently used in order to cover specific needs of a course. Courses in different professional areas require different management tools- this gap is especially obvious when comparing Technical and Social Science domain. There has been some effort to establish LMS national wide in HE, in University and in individual Faculties but practically it is best that the teaching staff decides upon LMS integration according to specific needs.

Usage of e-learning systems has potential to further improve the quality of HE curriculum. Process mining techniques in e-learning system can be utilized to analyze learning paths in a specific domain [9] in order to discover e-learning process model. Such approach seems to be interesting, as it offers a way forward towards adaptive learning system. However, we have discovered that there is a great variety among different lectures, depending from knowledge level, academic year and the nature of thematic presentations (management). Therefore, we have not yet adapted such approach because of possible complexity issues and its non-universality threats.

IV. USAGE OF MOODLE IN MECHANICAL ENGINNERING

In various academic courses, that we are responsible for, we have been using Moodle since 2015. List of courses that are managed with Moodle:

a) Space Modelling and Representation; 1st academic year; master study; sl. Modeliranje prostora; weekly lectures

b) Space Modelling and Representation; 1st academic year; bachelor study; sl. Snovanje izdelkov in projektiranje; weekly lectures

c) Product Design and Development; 2nd academic year; master study; sl. Snovanje in razvoj izdelkov; weekly lectures

d) Power Engineering; 3rd academic year; bachelor study; sl. Delovni stroji; project-orientated course

e) Engineering Techniques; 1st academic year; master study; sl. Konstrukcijske tehnike; project-orientated course

f) Construction with polymers; 2nd academic year; master study; sl. Konstruiranje z nekovinskimi gradivi; mixed course

During the years before and after implementation of e-learning we had combined average of 600+ students enrolled in presented courses per year. The exact numbers are shown in Table I. Number of involved teaching assistants varied between courses, as they involve different number of students. Each year more than 8 teaching assistants were involved, some of them covering more than one course. During the years there has been some personnel changes, altogether 12 teaching assistants were involved. Management of Moodle and courses was done at two levels. Drafts for courses have been made and updated annually (macro management) and at each course one teaching assistant had the course managing role (micro management). Through such distribution knowledge transfer from senior towards new staff members was ensured, reconciling differences in opinions and updating course content was done on two levels.

TABLE I. NUMBER OF ENROLLED STUDENTS

Academic year	***2011/12***	***2012/13***	***2013/14***	***2014/15***
Nr. of students	758	699	772	661
Academic year	***2015/16***	***2016/17***	***2017/18***	***2018/19***
Nr. of students	659	621	601	535

It can be seen that in recent years our Faculty has lowered the number of enrolled students. This lowering in student numbers is especially visible in the courses, conducted in the first academic years. In the Figure 3 the highest numbers of students is constantly lowered. It seems that courses conducted in later years to be less or even none affected.

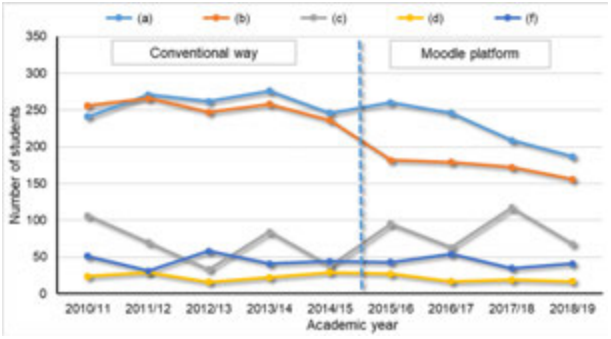


Figure 3. Graphical presentation of number of students troughout years.

We can compare the student average grades in the courses through academic years (Figure 4). Focus is on timeline before and after implementation of Moodle. Although students were more informed and in general more satisfied, unfortunately there seems to be no clear correlation regarding average grades.

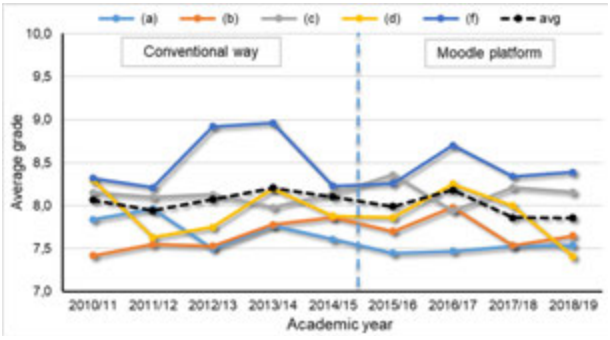


Figure 4. Graphical presentation of average grades troughout years.

The main benefits of e-learning implementation is change in the workflow. The teaching staff had significantly reduced time needed for administration, had better overview of student activity including grading and collaboration. The ease of publishing and changing the topics has led to topics improvement over the years.

As a possibility for a future upgrade to the courses we have tested implementation of educational videos to complement existing documents in academic year 2018/19 in the first course on the list (a). The videos in courses were divided into:

- Presentation of the theoretical background, and
- Presentation of practical usage.

Enrolled students at the course were separated into three groups. First group had an obligatory to watch the videos as a pre-assignment before each exercise and they did not have weakly assignments. The second group of students had weakly assignments and videos were not obligatory. The third group had a combination of video lectures and weakly assignments. During the duration of the course we had three testing of knowledge. After each testing, the group assignments have switched.

We have assumed that the outcome of video lectures introduction of will be self-learning and higher level of satisfaction among students. After the course completion we have prepared questionnaire for students. The group that started with video lectures, preferred them, where the group that was introduced to video lectures after first testing preferred the established topic presentation. With increasing complexity of the presented topics towards course completion majority of students preferred combination of video lectures and weakly assignments. Therefore, in the academic year 2019/20 we are conducting course as a combination of video lectures and weakly assignments. We will be further expanding the usage of video lectures to at least one more course.

Since this was a pilot project for us, we had to invest a lot of effort and time in planning and delivering of target video content.

V. SWOT ANALYSIS

One form of evaluation of implementation of e-learning platform for managing of several HE courses presents SWOT analysis (Strengths, Weaknesses, Opportunities and Threats). Through the past years of usage of Moodle platform the assistants, responsible for

the courses have gained operational experience and have evaluated its pros and cons (see Table II).

SWOT analysis was done on two levels. First was through questionnaire for students, involved at one of the courses after its completion. Second was through open discussion with senior teaching assistants, that conducted some of the courses over the years. Combining the two approaches, we have formed the SWOT Matrix, presented in the table below.

TABLE II. SWOT ANALYSIS OF MOODLE IMPLEMENTATION

Strengths	***Weaknesses***
• Reduced time needed for administration. • Fast and efficient teaching material distriburion. • Ease of updating and changing material for different user groups. • Provides a full range of activities. • Access to interactive materials. • Independence of learning from place and time. • Fast information transfer. • Lower costs: less need for mobility of students and education providers. • No need for printed materials.	• Lack of personal contact. • Unclear instructions for all students. • Limitations of uploaded material. • It requires a dedicated server for efficiency and cuntinuity. • Need for computer administration. • New responsibilities for lecturers.
Opportunities	***Threats***
• Positive effect of e-learning (quality, active learning). • Real-time tracking of progress. • Adding new interactive materials.	• Malicious codes that can compromise the information system, the network, and the privacy of the individual. • Students need to have equipment and internet access.

Using the Moodle e-learning platform over a number of years has revealed its strengths and weaknesses. The ease of implementation and change of the online materials in learning environment should be emphasized. No special knowledge of teaching staff is required for materials implementation. The GUI (Graphical User Interface) interaction for students presents no problems- it is easy to enlist in a course, to find downloadable material and vital information. Usage of a online classroom and e-materials is independent of place and time, which is definitely an advantage as the lectures schedule can be easily adapted in case of holidays, sickness etc. We have also noticed reduced costs by having materials online and lack of need for printing. Moodle also allows us to transfer information quickly. With information available 24/7 students are always up-to-date with latest activities and their progression.

In some classes we use the online platform also for student evaluation through weekly and monthly assessments. The platform is used for distributing material, as a database (students can upload files) as well as for setting the time limit for the assessments. Furthermore, it offers combining students in groups and distribution of material according to them. Since students have to process some of the material by themselves, their self-motivation is being raised. The lecturer has insight into each student's work and can keep up with his/hers progress. Having material always available for students is beneficial, but it also presents a risk. The instructions have to be clearly presented and having an open discussion about student's tasks is limited – problem with universality of tasks in the field of Mechanical Engineering have emerged. Oral communication has advantages over written messages. This issue was solved during physical exercises where each student had a chance to discuss about his/hers task.

Although usage of e-learning platform has many advantages and simplifies the academic process, also some disadvantages have occurred. Distance learning results in a lack of personal contact between students and the lecturers- this can lead to misinterpretation of online material. How much of the content can be processed by a student through a computer depends of the nature of the study (e.g. laboratory, experimental exercises cannot be done online). A new responsibility has occurred for the lecturer – he/she has to assess the amount and topics of learning material, gained through e-learning.

Using a LMS tool for conducting courses requires additional hardware requirements. To access and use the materials, students need a computer, the right software, and Internet access, things that can present an issue – we teach in the field of Mechanical Engineering, thus relative good equipment is needed. Furthermore, students were required to upload their working progress weakly. In addition, all exam outputs from practical exercises were also uploaded to Moodle, thus creating large amount of data after each academic year. With introduction of video lectures, file size limit of 100MB presented an issue. By raising the file size limit, download time can present an issue for some students with slow internet access. We have solved the issue by creating several short videos for each presented topic, each limited to 4 minutes. Altogether, conducting several courses with Moodle, enabling access for 600+ users and having large amount of data requires additional investments in server hardware and administration. The creation and implementation of the e-materials in Moodle represents additional work for involved teaching assistants. Threats also arise because Moodle is an open source system and can therefore be targeted by malware (viruses, worms, Trojans, etc.), which can damage the information system, corrupt data or endanger the privacy of individuals.

VI. CONCLUSION

Based on our experience gained with managing more than 2400 students through Moodle platform, we can clearly state that e-learning platform facilitates work for professors, teaching assistants and students. Unfortunately, by analyzing the average grades there seems to be no evidence of benefits towards improving students learning performance. The experience of a teaching assistants and hers/his ability of knowledge transfer presents the key factor of success.

As digitalization of HE is on the rise all over EU, also efficiency and quality of knowledge transfer by making e-learning a 100% learning tool is in question. An answer to

this question will present a key role in educational system of the future. In our opinion the path forward for conducting courses presents a hybrid approach – combination of partly online and partly in classrooms presented topics. With such approach we will be able to embrace the benefits of LMS while preserving the benefits of in-person communication.

REFERENCES

[1] Aristovnik, A. (2012). The impact of ICT on educational performance and its efficiency in selected EU and OECD countries: a non-parametric analysis. Turkish Online Journal of Educational Technology, 11(3), 144–152. Retrived from http://www.tojet.net/articles/v11i3/11314.pdf

[2] Aristovnik, A. (2014). Development of the information society and its impact on the education sector in the EU: efficiency at the regional (NUTS 2) level. Turkish Online Journal of Educational Technology, 13(2), 54–60. Retrived from http://www.tojet.net/articles/v13i2/1326.pdf

[3] L. Bregar, M. Zagmajster, M. Radovan: Osnove e-izobraževanja, priročnik, Andragoški center Slovenije, 2010 (in Slovene)

[4] P.E. Sanderson: E-Learning: strategies for delivering knowledge in the digital age, The Internet and Higher Education, Volume 5, Issue 2, Summer 2002, Pages 185-188 https://www.sciencedirect.com/science/article/pii/S1096751602000829?via%3Dihub

[5] V. Sulčič: E-izobraževanje v visokem šolstvu, Faculty of Management, University of Primorska, 2008 http://www.fm-kp.si/zalozba/ISBN/978-961-6573-90-0.htm

[6] Phillips, R., Maor, D., Cumming-Potvin, W., Roberts, P., Herrington, J., Preston, G., Moore, E. and Perry, L. (2011). Learning Analytics and Study Behaviour : A Pilot Study. In Proceedings ascilite 2011 (997-1007). Hobart, TZ: Australia

[7] M. Sertić, "Implementation of blended learning in Computer Science," 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 514-520. https://ieeexplore.ieee.org/document/8756916

[8] M. Puhek and L. Bregar, "Why e-Learning in Slovenia is left behind?," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2018, pp. 0749-0753. https://ieeexplore.ieee.org/document/8400139

[9] B. Blašković, F. Škopljanac-Mačina and I. Zakarija, "Discovering e-learning process models from counterexamples," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2018, pp. 0593-0598. https://ieeexplore.ieee.org/document/8400112

Creation of software platform for distance use of lab equipment and data in Virumaa College at Tallinn University of Technology

Oleg Shvets*, Kristina Murtazin **, Gunnar Piho***
Tallinn University of Technology, Information Systems Group,
Tallinn, Estonia
*oleg.shvets@taltech.ee
**kristina.murtazin@taltech.ee
***gunnar.piho@taltech.ee

***Abstract* - Modern e-learning technologies provide various opportunities for students and teachers. E-learning in engineering education has some specific facets, for instance, in classes where a task or experiment is associated with some specific equipment. Therefore, in engineering education, there are some domain-specific issues to solve: (a) the time of usage of equipment in laboratories is not always efficiently disposed, (b) data of experiments and practical work disappear or are saved on different devices and in different laboratories, (c) other educational institutions are not able to acquire or use special equipment or appliances, (d) main part of equipment can be inaccessible for those students who receive remote education, and (e) it is difficult to organise effective feedback between student and teacher. In the training platforms, it is also important to consider the need of the student to acquire practical skills in using laboratory equipment. The presented work proposes a prototype software platform that integrates the equipment, located in different laboratories of Virumaa College at Tallinn University of Technology, and enables students to use the equipment remotely.**

Keywords - e-learning, remote labs, remote students, personalised feedback, knowledge, competencies, IoT.

I. INTRODUCTION

Today, classes of electronics for students in technical specialisations and engineering, almost always need practical knowledge to use equipment or instruments in some laboratory or at home. Therefore, the laboratory instruments and equipment should be accessible for the students, preferably also remotely. Achieving this is sometimes challenging, especially when the number of students is increasing. In technical training, a combination of real-world laboratory practice with realistic virtualisation, that includes visualisation and interactive online intervention, helps students in the development of practical skills with a deep understanding of basic scientific and engineering concepts [1]. The Internet of Things (IoT) is here a promising tool to organise a practical work of students in engineering education.

IoT is a convenient way to use a variety of equipment. Its application is possible in various fields of science and technology. The relative simplicity of use and deployment provide unlimited possibilities in the application of IoT systems. Besides, IoT is a form of the Internet with the ability to collect, analyse and diffuse data, which further are processed to obtain information [2]. Quite often, e-learning is limited only by using virtual classes, video lectures and animations, online lessons, and distributing educational materials. However, IoT can change this practice.

Applying IoT in e-learning can probably help to transform education [3]. All the features of IoT systems make IoT quite universal for use in the learning processes, also in the various courses of higher technical education. Modern laboratories of engineering education institutions normally combine an amount of quite expensive devices that often duplicate each other. In those laboratories, one can meet a situation where laboratory equipment is idle because students work with them only during hours in the classroom. Virumaa College at Tallinn University of Technology is not an exception. It would be much more efficient to use some of these devices, stands and modules in this idle time by other students in the remote access mode. Modern software together with IoT provides this opportunity [16].

The goal of our work is to create a learning environment. This environment integrates the equipment, educational materials, data, cloud storage with output data for analysis and visualisation, as well as storage of student competence and the possibility of personalised feedback between student and teacher.

Following are the stages we followed so fare when developing such an environment:

- A concept development that included an analysis of relevant published papers over the past seven years.
- A prototype creation based on the developed concept, available equipment and needed data.
- Implementation of the prototype for obtaining laboratory data and for visualising these data.
- Identification of needs for future research and development directions.

In creating a prototype for a learning environment, the requirement is that student can select devices from different laboratories. The data, obtained from experiments with

these devices, are saved and can be visualised and analysed remotely. This aspect is especially important for those students who use laboratory equipment remotely. We use Wonderware (https://www.wonderware.com/industrial-information-management/online-insight) Online Insight for connecting the devices. Usually, this software is used in enterprises to automate industrial processes, but in our case, we use it to organise our processes in an educational laboratory.

II. METHODOLOGY

A. Literature Review

Using IoT systems in engineering education is a rapidly developing and changing field. Therefore, it is important to have an idea of what has been published in this area recently. A state-of-the-art literature review is a good way to understand this. To conduct the literature review, we formulated the following research questions (RQ) [17], [18]:

- RQ1 - What forms of integration of IoT and e-learning are in use now?
- RQ2 - What are the main components of educational software platforms that are using IoT?

IEEE Explore Digital Library was chosen as the main indexed database for finding relevant papers. IEEE Explore was chosen because especially the papers from electronic engineering domain are relevant. For the search of published papers, the simple query ((("All Metadata":IoT) AND "All Metadata": e-learning)) was selected. Only articles from 2012 up to 2019 we included in the selection. Based on this selection, we obtained 59 publications.

The next step was the analysis of the selected publications in order to find answers for the RQ1 and RQ2.

Related to the RQ1, the following features of IoT application in the learning processes were identified:

- IoT systems are affordable tools for organising of practical experiments [4], [11].
- IoT devices are used as a tools of monitoring and data collection, where these data are later used in learning environments for analysis [5].
- IoT devices and applications for these devices can be used for detecting and personalising the learning experiences [6].
- The IoT technology enables emulate physical situation, sensing, and analysis and visualisation of the environment [7].
- IoT-based learning systems can increase students motivation to complete learning courses [8] and generally increase the effectiveness of learning [9].
- Smart classes and classrooms created based on IoT systems, contribute to creating effective feedback between teacher and student in real-time [10].

Related to the RQ2, it was important to determine which modules generally exist in software and hardware platforms where IoT systems and their derivatives are integrated into the learning environments. The main components of such platforms include the following ([12], [13], [14], [15]) :

- A connectivity and normalisation component that guarantees accurate data transfer and interaction with all devices.
- A device management component that ensures the connection and uninterrupted operation of all the connected IoT components.
- A scalable database component that allows the storage of data received from the IoT components.
- A processing and action management component that allows, for instance, the event triggers and other smart actions based on specific data received from the sensors.
- An analytics component for clustering, analysis and deep machine learning right up to the predictive analytics.
- A visualisation component for a convenient interface to represent analytics results that allow building various graphs and diagrams. It is also necessary to provide the ability to display the data received from the IoT in a real-time.
- Some additional components, which, for instance, allows IoT developers to create and test prototypes of learning environments using test data.
- External interfaces to offer the possibility to transfer and exchange data between third-party systems.

Based on the conducted literature review, one can conclude that the integration of IoT and e-learning can improve the learning quality. It also indicates what components should be considered when creating prototypes of such learning platforms. As a minimum, such prototype should contain parts as (a) device management, (b) database, (c) processing and action management, (d) analytics, and (e) external interfaces.

B. Connection of devices

It should be noted that in the educational systems integrating IoT and e-learning, both the conventional and industrial controllers are in use. Before deciding, what is relevant for our college, we went through and analysed some important constraints.

In various laboratories of Virumaa College at Tallinn University of Technology, we have different types of equipment with various types of controllers. Some examples of these types of equipment are illustrated in Figures 1, 2 and 3. Furthermore, in the future, the set of devices can change.

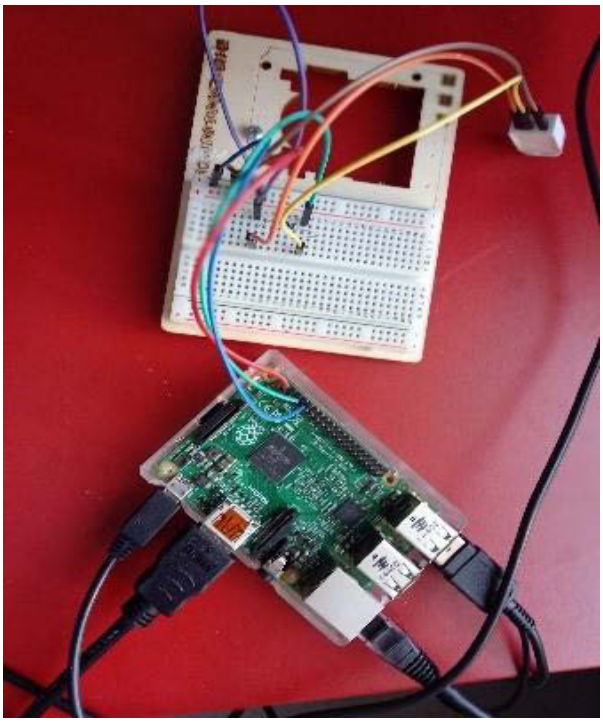


Figure 1: One of the controllers from the college laboratories.

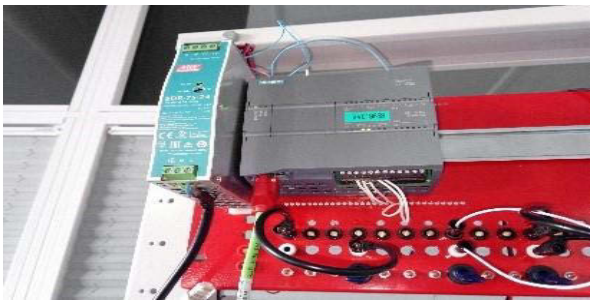


Figure 2: Example of the mount from the laboratory of automated technology with its programmable industrial controller.

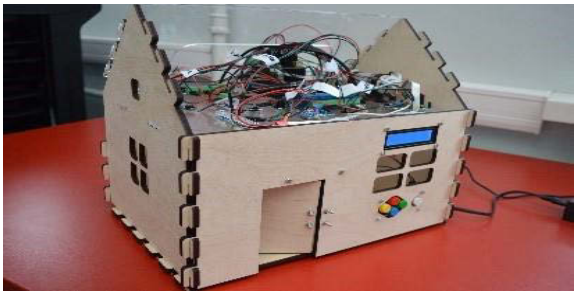


Figure 3: Prototype model of «Smart House» which uses a controller and some data for experiments.

Besides, we plan to use this learning environment in various classes of the college. In some of these classes, the students learn the principles of programming of controllers, including IoT systems. In other classes, the students undergo practical training at enterprises where they study the work of different industrial controllers.

After considering these various constraints, our choice was to use a universal industrial SCADA system from Wonderware. Wonderware (www.wonderware.com) HMI/SCADA software provides the reliable connection of devices based on controllers, and Wonderware Historian can be used as a data source.

III. RESULTS

The main result of the work done was the creation of a prototype software platform where we used the knowledge, that we obtained from the conducted literature review. Figure 4 illustrates the proposed schema by showing the interaction of all components of the developed system.

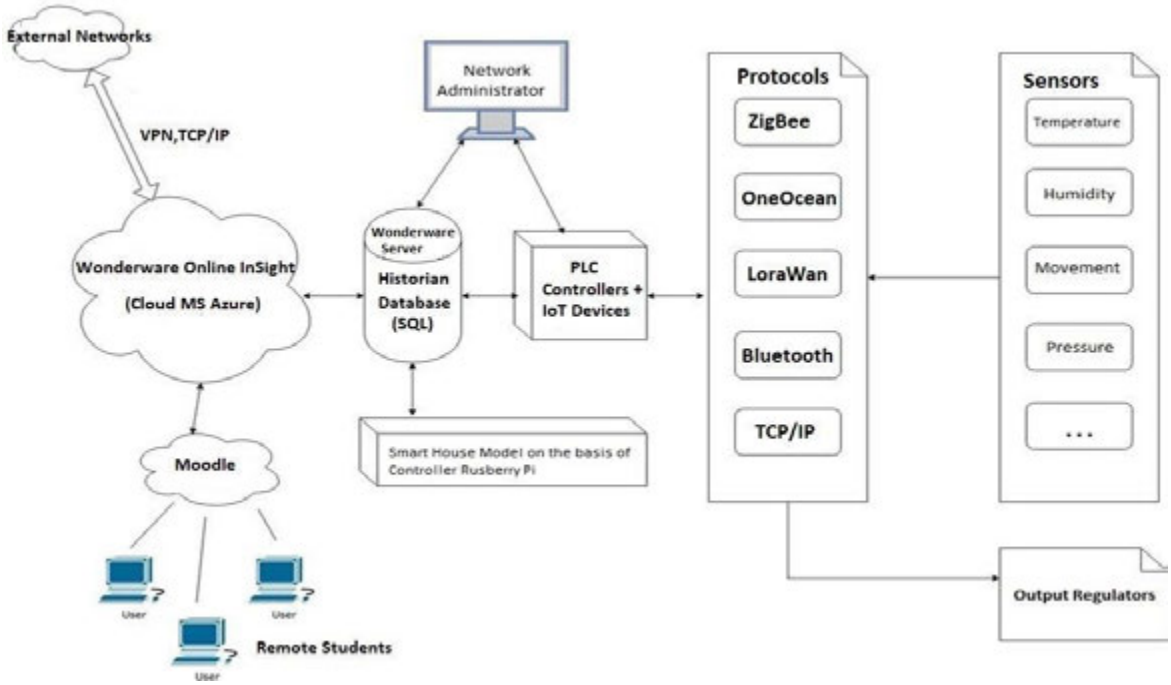


Figure 4:The conceptual schema of the proposed learning environment

The sensors module includes sensors that are installed in various laboratories of the college and measure such data as temperature, humidity, movement, pressure, etc. This data is collected by various controllers that are connected directly or through the PLC module to the Wonderware Server. Wonderware Server includes its own database. To ensure that this database works properly, the SQL module was installed.

Data, received from various sensors and other devices can be presented to students for analysis in various forms, for example, in the form of tables and graphs. It is possible to analyse the data obtained both from the local network of the college and remotely. For remote analysis, visualisation tools in the cloud storage are connected to Wonderware Online InSight. Online InSight uses MS Azure cloud. The system can also be used to visualise data (e.g. temperature and humidity) that are collected from different sensors. Data visualisation is exemplified in Figure 5.

Because the Moodle learning environment is an official learning environment for all the learning courses in TalTech, hence our software platform for distance use of lab equipment and data is also connected with Moodle. All is straight forward. In the Moodle learning environment, the official learning course is created. Students are registering to this course and gain access to the training materials, class assignments and schedules. Performing the assignments, students must make the analysis of schedules and draw conclusions. Graphs are the objects that visualize data coming in from various sensors in online mode. To date, related to the software platform for distance use of lab equipment and data, the Moodle is used only as an environment into where the data in the form of graphs are sent from the Wonderware Online Inside cloud storage. We used the Frames toolkit and placed graphics as HTML objects on one of the pages of the course in Moodle. This is illustrated in Figure 5.

IV. CONCLUSIONS

Based on the results of our work, we can conclude that the concept of a learning environment system was developed. This system combines the equipment, educational materials, data, and cloud storage for later analysis and visualisation of data. In order to develop the concept, the systematic literature review was conducted, and the needs of the learning processes of Virumaa College at Tallinn Technical University were considered. Especially we focussed in supporting the practical laboratory experiments. Visualised data can be transferred to the Moodle learning environment in the form of frames. Currently, during the educational process of remote students, some of the attached devices and Moodle environment for work with educational materials and for registration of students are used.

It should be noted that the main purpose of creating of an above-described software environment is to use it in the educational process. Students will have the opportunity not only to study the operation of equipment located in different laboratories of the educational institution, but also to analyse the results of this work. The proposed learning platform allows not only to transfer learning content, but also to connect many devices, including IoT, to unified laboratory experiments system. Data received from devices can be preserved both on the college server and in the provided cloud storage. Data can be visualised and analysed. It is possible to organise remote access for students and teachers to the created e-learning platform.

We can single out the following general directions of future research and development in improving the prototype environment: (a) attachment of more equipment to the project, (b) creation of own cloud-storage to collect experiment data, (c) creation of own web-interface for more flexible control and experiment data analysis, (d) analysis of feedback and knowledge obtained by students

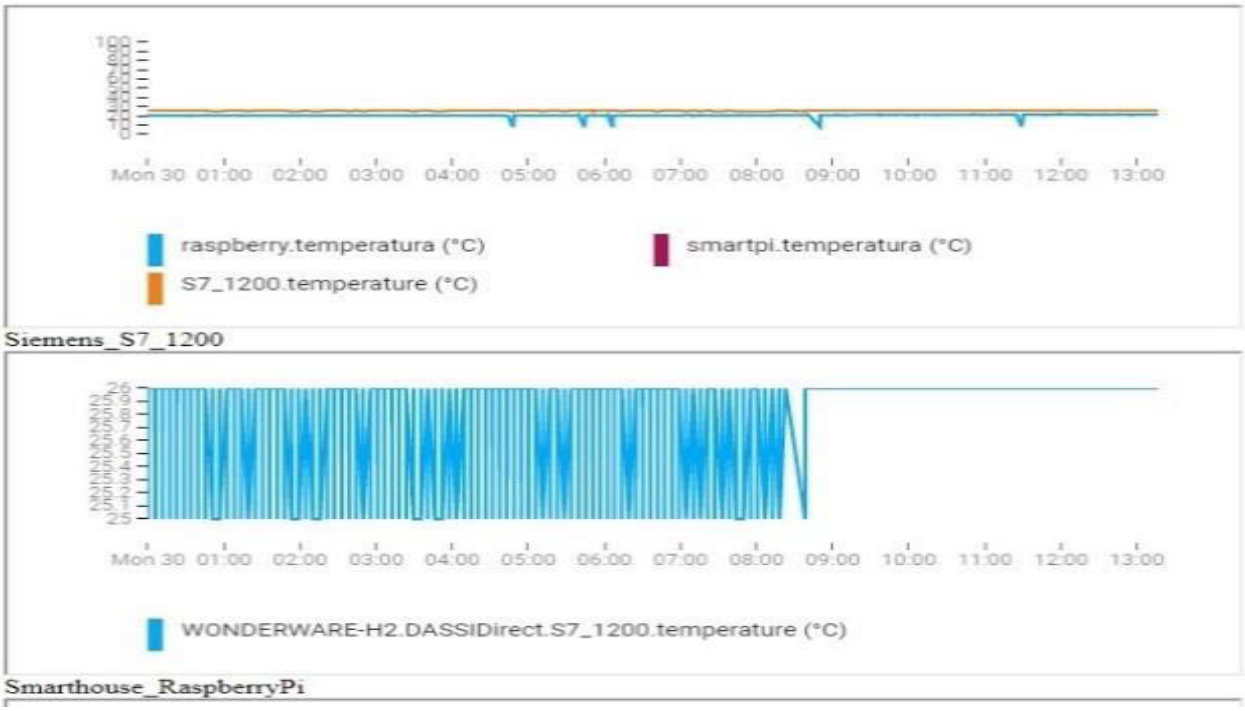


Figure 5: Visualised of data from sensors measuring temperature and humidity

REFERENCES

[1] Y. E. Cherner, J. Uhomoibhi, G. Mullett, M. M. Kuklja, C. Mkude, L. Fweja, H. Wang, "Implementation of Interactive and Adjustable Cloud-based e-Learning Tools for 21st Century Engineering Education: Challenges and Prospects," 2019 IEEE World Conference on Engineering Education (EDUNINE), March 2019.

[2] H.Y. AbuMansour, H. Elayyan, "IoT theme for smart datamining-based environment to unify distributed learning management systems," 2018 9th International Conference on Information and Communication Systems (ICICS), pp. 212-217, April 2018.

[3] Z. AjazMoharkan, T. Choudhury, S. C. Gupta , G. Raj, "Internet of Things and its applications in E-learning," 2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT), pp. 1-5, February 2017.

[4] Y. Ban, K. Okamura, K. Kaneko, "Effectiveness of Experiential Learning for Keeping Knowledge Retention in IoT Security Education," 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 699-704, 9-13 July 2017.

[5] W. Weihu, L. Yangshen, G. Yahui, J. Pengpeng, "An Educational Platform for Promoting Awareness of Lake Environmental Protection with Live Monitoring Technology," 2016 International Conference on Educational Innovation through Technology (EITT), pp. 238-241, 22-24 September 2016.

[6] Y. K. Ever, A. V. Rajan, "The Role of 5G Networks in the Field of Medical Sciences Education," 2018 IEEE 43rd Conference on Local Computer Networks Workshops (LCN Workshops), pp. 59-63, 1-4 October 2018.

[7] Y. Hayashi, R. Nakamura, O. Hasegawa, Y. Kitano, Y. Kiyoki, "A Cyber-Physical Learning Environment Design for Study Trajectory Creation," 2019 International Electronics Symposium (IES), pp. 76-81, 27-28 September 2019.

[8] M. Yamada, T. Oda, K. Matsuo, L. Barolli, "Design of an IoT-Based E-learning Testbed," 2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 720-724, 23-25 March 2016.

[9] R. Yakoubovsky, V. Sarian, "IoT in Effective Distance Learning Process," 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS), 24-26 June 2019.

[10] S. Mahmood, S. Palaniappan, R. Hasan, K.U. Sarker, A. Abass, P. M. Rajegowda, "Raspberry PI and role of IoT in Education," 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), 15-16 January 2019.

[11] A. Maiti, T. Byrne, A.A. Kist, "Teaching Internet of Things in a Collaborative Laboratory Environment, " 2019 5th Experiment International Conference (exp.at'19), pp. 193-198, 12-14 June 2019.

[12] A. A. Zaharov, O. V. Nissenbaum, K. Y. Ponomarov, A. V. Shirokih, "Use of Open-Source Internet of Things Platform in Education Projects," 2018 Global Smart Industry Conference (GloSIC), 13-15 November 2018.

[13] A. M. Njeru, M. S. Omar, S. Yi, "IoTs for capturing and mastering massive data online learning courses," 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), pp. 91-94, 24-26 May 2017.

[14] S. Shapsough, M. Hassan, S. E. Shapsough, I. A. Zualkernan, "IoT Technologies to Enhance Precision and Response Time of Mobile-Based Educational Assessments," 2016 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 202-205, 15-17 December 2016.

[15] M. Y. Idris, D. Stiawan, N. M. Habibullah, A. H. Fikri, M. R. A. Rahim, M. Dasuki, "IoT smart device for e-leaming content sharing on hybrid cloud environment," 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 19-21 September 2017.

[16] N. Fujii, N. Koike, "IoT Remote Group Experiments in the Cyber Laboratory: A FPGA-based Remote Laboratory in the Hybrid Cloud," 2017 International Conference on Cyberworlds (CW), pp.162-165, 20-22 September 2017.

[17] B. Kitchenham, R. Pretorius, D. Budgen, O. Pearl Brereton, M. Turner, M. Niazi, S. Linkman "Systematic literature reviews in software engineering – A tertiary study", Information and Software Technology, vol. 52, August 2010, pp. 792-805.

[18] C. L. Goues, C. Jaspan, I. Ozkaya, M. Shaw and K. T. Stolee "Bridging the Gap: From Research to Practical Advice", IEEE Software, vol. 35, September/October 2018, no. 5, pp. 50-57.

FPGA Design of Boolean Functions Using a Cascade of Decoders and Logic Gates

A.N. Borodzhieva*, I.I. Stoev**, I.D. Tsvetkova*, S.L. Zaharieva** and V.A. Mutkov**
*University of Ruse "Angel Kanchev"/Department of Telecommunications, Ruse, Bulgaria
**University of Ruse "Angel Kanchev"/Department of Electronics, Ruse, Bulgaria
{aborodzhieva, istoev, itsvetkova, szaharieva, vmutkov}@uni-ruse.bg

***Abstract* – The paper presents active and interactive methods used for improving knowledge that students will get while implementing different Boolean functions with a cascade of full decoders and OR logic gates based on their schematic view and their testing on the laboratory board built with devices of Spartan-6 FPGA Family. The courses "Digital Electronics" and "Pulse and Digital Devices" introduce the students-bachelors of the electrical specialties in our University to the main problems of digital electronics.**

Keywords – Boolean functions, combinational circuits, decoders, digital electronics, FPGAs, logic gates, Xilinx.

I. INTRODUCTION

Digital electronics is a part of the modern equipment like washing machines, dryers, dishwashers, toasters, air conditioners, game consoles, etc. The fundamentals and implementation of digital electronics units are basic for understanding the principle of operation and design of computers and communications, military and security equipment, consumer and industrial electronics, embedded systems, and so on [1].

At the University of Ruse one of the main courses is "Digital Electronics" in the curriculum of electrical and computer engineering programs. The course covers various topics in the field of synthesis and analysis of logic devices and it is based on two components – lectures and laboratory exercises. For the purposes of the effective educational process for electrical and computer engineers, a variety of active and interactive learning methods are used during the lab exercises together with the elements of the traditional presentation-based lectures. For faster and longer-lasting learning of the material being studied, active learning techniques such as case-based and problem-based learning, one-minute papers and "think-pair-share", decision-making activities, etc. are used [2, 3, 4]. Modelling and simulating, as well as FPGA designing give the students opportunities for finding a solution for different practical tasks and better understanding the material. Critical thinking, creativity, collaboration, research and problem solving, innovations and communication skills are used and developed in the laboratory exercises [2, 3, 4].

First, different active and interactive learning methods used in the training process are considered in the paper. Next, active learning in FPGA design of Boolean functions using a cascade of decoders is described.

II. ACTIVE AND INTERACTIVE LEARNING METHODS USED IN THE TRAINING PROCESS

Usually, students are asked to do some activities outside of school. In primary and secondary education, these activities are often referred to as home-based activities (homework) and are done by students at home. In higher education, these extracurricular activities are often called assignments (tasks), and students typically perform them at home, in libraries, or at other university facilities. Based on these features, education can be generalized as "lessons at classroom and homework at home". This model, known as a teaching-based model, has been used for decades at each educational level and it is often associated with passive student learning (Fig. 1).

Flip teaching presents a total change of perspective, seeking the active participation of students in the learning process. It performs an interchange of the two main activities of the traditional model: lectures and homework. The main change is that lessons are delivered at home and homework – in the classroom [5].

The active learning method consists of a group of tools allowing students to "actively" participate in the course by engaging in activities proposed to improve their skills rather than attempting to absorb unilaterally transmitted information.

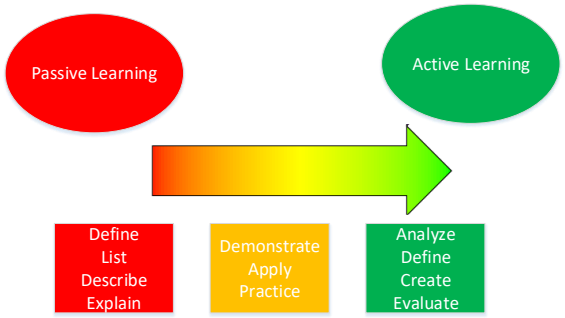


Figure 1. Comparison of passive and active learning

This publication is developed within the frame of Project "Forest Monitoring System for Early Fire Detection and Assessment in the Balkan-Med Area" (SFEDA), project code 2263, financed under the transnational Cooperation Programme INTERREG V-B "Balkan-Mediterranean 2014-2020", co-funded by the European Union and the National Funds of the participating countries and by the National Scientific Program ICTinSES, financed by the Ministry of Education and Science of Bulgaria.

The active learning method (Fig. 1) requires students to implement a variety of activities to generate full awareness of what they are doing as content processing enables them to learn. In essence, it is a student-centered approach in which the student determines his or her pace and style of learning and finds support in his or her peers, and where teachers play the role of "guiding and motivating" the student in the learning process. The active learning approach can be related to a variety of techniques, ranging from traditional problem-solving to case studies, learning from doing, learning through doing, discussion activities, or peer training. Modern technologies play an important role in the learning process, but they are not essential, so problems based on simulations and games are included along with the use of mobile devices and ultimately classrooms equipped with modern high technologies. Definitely, active learning arouses more student interest than traditional lectures, providing flexibility for course development, and encourages a tendency to take on individual and group challenges. In engineering education, experimental-based learning and the group approach offer activities where motivation and creativity are essential. Students are provided with tools to build their own knowledge, modified dynamically, along with the capacity to assess the difficulty of the problem and to determine the criterion to be used to solve it [6].

Continuous assessment throughout the course is considered motivating and maintains regular training habits. Therefore, student assessment requires rethinking. Rather than just an end-of-course exam, project work, exercises and course participation are also an essential part of the assessment. A joint (collaborative) assessment of the project work is used to motivate cooperation, but exams are evaluated individually to prevent free riding. The visualization using numerical software (for example, Scilab, MATLAB, Logisim, Multisim, etc.) aims to focus the learning on understanding the material. Design exercises and practical measurements are used to familiarize students with laboratory equipment and to motivate them studying the theory [7].

Some of the interactive teaching styles and the most effective ways to engage the students' interests are: 1) brainstorming; 2) think-pair-share; 3) buzz session; 4) incident process; 5) question and answer sessions [8].

The interactive brainstorming is performed in group sessions. It is useful for generating creative ideas and it helps students learn together in a group. Different types of interactive brainstorming are presented in Fig. 2 [8].

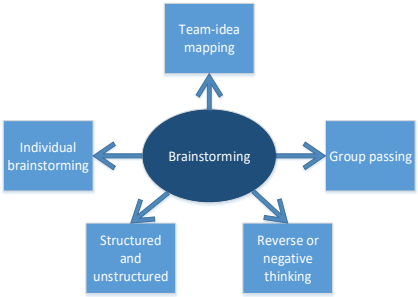


Figure 2. Brainstorming

In the process of think, pair, and share, the lecturer formulates a problem or a question, then groups students in pairs and gives each pair time enough to make a conclusion on the problem, and permits each participant to define the conclusion in voice, for example one student explains a concept while the other student evaluates what is being learned. Applying variations of the process, the students are engaged in the learning process, communicating, and retaining more information [8].

During the buzz sessions, students participate in session groups focused on a single topic. Every student in each group contributes ideas. The teacher encourages discussion and collaboration among the students in each group. During the buzz sessions, students should learn from one another's experiences [8].

The incident process is a teaching style involving case study formats. The focus of this process is on learning how to solve real problems involving real people, providing small groups of students with details from real situations and then ask them to find out a workable solution. This teaching style prepares the students for life outside the classroom [8].

During the Q&A sessions, each time when the topic is introduced, before the formal lecture, the teacher asks the students to write the thematic questions on cards. After collecting the cards, the teacher mixes them up, reads them and answers the questions generated by students [8].

III. Active Learning in FPGA Design of Boolean Functions Using a Cascade of Decoders

A. *Applied active and interactive methods during implementation of logical functions with decoders*

Based on the flip teaching, realizing an interchange of the two main activities of the traditional model (lectures and homework), during the classes in the courses "Digital Electronics" and "Pulse and Digital Devices" lessons are delivered at home and homework – in the classroom. Before the lesson, students should get acquainted with the theoretical issues discussed in detail in [9, 10] (or in other sources in Internet), as well as with the possibilities of Logisim [11] for studying digital circuits (Fig. 3) – for example, definition of decoder (Fig. 3 A), the truth table (Fig. 3 B), the output functions (Fig. 3 C) and the block diagram (Fig. 3 D) of a complete decoder with two inputs and four outputs. Students need to test the circuit (Fig. 3 D) for all four input combinations and the results observed are shown in Fig. 3 E, where the active input and output signals are 1s. For this result, students need to use the software product Logisim [11].

It is also necessary for students to get acquainted in advance with the solved two examples – Problem 1 and Problem 2 (Fig. 4), which allow the implementation of a Boolean function with a decoder and additional logic gates (Problem 1) or by cascading decoders and additional logic gates (Problem 2).

Students learning the courses "Digital Electronics" and "Pulse and Digital Devices" can use individual brainstorming or think, pair, and share activities to solve these problems and test the logic circuit shown in Fig. 3 E.

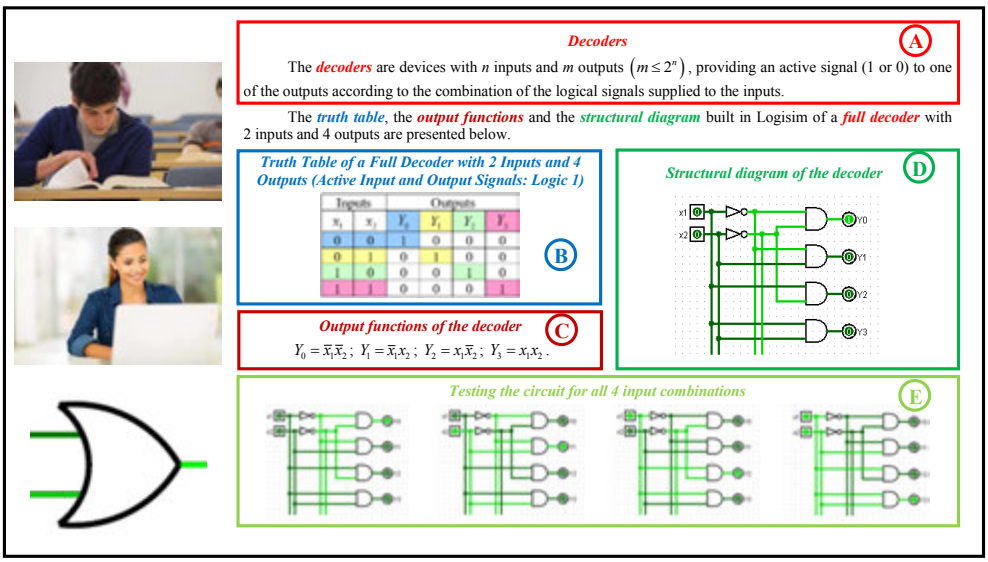


Figure 3. Flip teaching (an interchange of lectures and homework) during the classes in the courses "Digital Electronics" and "Pulse and Digital Devices"

Problem 1. Using a three-input decoder and additional logic gates, construct a structural diagram that implements the function $f(x_1,x_2,x_3) = \vee(0,1,3,7)$ if the decoder has: a) inverse outputs (integrated circuit SN74LS138, with three enabling inputs $\overline{E}_1, \overline{E}_2, E_3$); b) normal outputs. Justify the type of the additional logic gate by describing the operation of the circuit for the input combinations with numbers 3 and 6.

Depending on the decoder type (normal or inverse outputs) and the way of describing the function (Perfect Disjunctive Normal Form, PDNF, or Perfect Conjunctive Normal Form, PCNF), the chosen additional logic gate is given in the table below.

Decoder with: / Function given in:	normal outputs	inverse outputs
PDNF $(\vee)$	OR	NAND
PCNF $(\wedge)$	NOR	AND

Problem 2. Using two-input decoders with normal outputs and enabling inputs and additional logic gates, construct a structural circuit that implements the Boolean function $f(x_1,x_2,x_3,x_4) = \vee(1,5,6,8,12,13,15)$.

Solution: In the table below the minterms *m* corresponding to the input combinations for which the function has a value of 1 are given. Each minterm is divided into two parts: the first one (m') includes the variables x_1 and x_2, and the second one (m'') – x_3 and x_4. Each part can be obtained at the output of a decoder with two inputs. The products of the first two arguments can be obtained at the outputs of a two-input decoder and can be used as enabling signals to the next stage composed of two-input decoders whose information inputs are connected to the other two variables. In this way, the cascade of the decoders is realized and shown in the figure below. The justification of the type of the additional logic gate is based on the results presented in the table above.

m_{No}	MINTERMS – $m = m'.m''$
m_1	$m_1 = m_0'.m_1'' = \overline{x}_1.\overline{x}_2 \ .\ \overline{x}_3.x_4$
m_5	$m_5 = m_1'.m_1'' = \overline{x}_1.x_2 \ .\ \overline{x}_3.x_4$
m_6	$m_6 = m_1'.m_2'' = \overline{x}_1.x_2 \ .\ x_3.\overline{x}_4$
m_8	$m_8 = m_2'.m_0'' = x_1.\overline{x}_2 \ .\ \overline{x}_3.\overline{x}_4$
m_{12}	$m_{12} = m_3'.m_0'' = x_1.x_2 \ .\ \overline{x}_3.\overline{x}_4$
m_{13}	$m_{13} = m_3'.m_1'' = x_1.x_2 \ .\ \overline{x}_3.x_4$
m_{15}	$m_{15} = m_3'.m_3'' = x_1.x_2 \ .\ x_3.x_4$

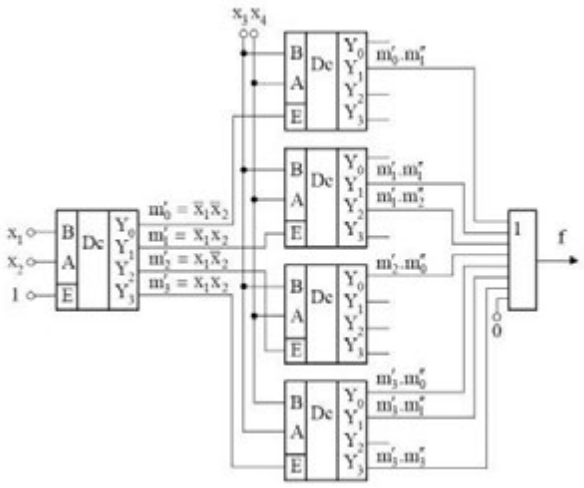


Figure 4. Problem 1 and Problem 2 – formulation and solution

Problem 3. Investigate the operation of a two-input decoder with normal outputs and an enabling input and determine the function implemented by cascading such decoders and with an additional logic gate (OR). Fill in the truth table of the device after programming FPGA with the file *casc_Dc_WG_xx.bit*, where xx = 01...30. The device realizes the function of four variables by a cascade of two-input decoders with normal outputs and an additional logic gate (OR) (Problem 2). After testing with the laboratory board and completing the truth table, build the schematic of the device used for creating the bit file in the environment of ISE Project Navigator. If you have some free time during the practical classes, create the bit file in ISE Project Navigator by yourself.

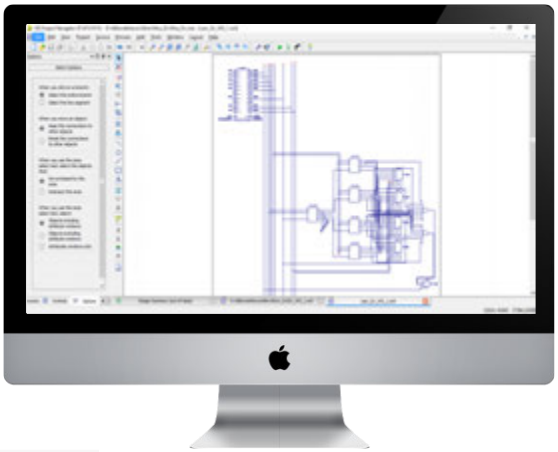




Figure 5. Problem 3 – formulation and solution

B. *Applied active and interactive methods during implementation of logical functions with decoders*

Students from the second year study some basic disciplines, where they learn essential skills. During the classes in the courses "Digital Electronics" and "Pulse and Digital Devices", students have to realize a function of 3 arguments with a three-input full decoder (Problem 1) and a function of 4 arguments by cascading two-input full decoders (Problem 2) in three ways: 1) manually; 2) by constructing and testing the circuit in Logisim; 3) by constructing and testing (or only testing) the circuit in the environment of ISE Project Navigator [12]. Here students need to use learning from doing, learning through doing, discussion activities, or peer training. A collaborative (joint) assessment of the project work is used to motivate cooperation, but exams are evaluated individually to prevent free riding. The visualization using software (for example, in this case, Logisim and ISE Project Navigator) aims to focus the learning on understanding the material.

During the classes, students get formulated tasks like those shown in Fig. 4 and Fig. 5. For every student, there is a personal assignment. They need to solve them and present a report with solutions like those shown in Fig. 4 and Fig. 5. The solution of the first problem is presented in details in [9].

Students can solve alone the first and the second assignment. To solve these problems students can use individual brainstorming or team idea mapping. In addition, they need to provide evidence from their work presenting a report consisting of step-by-step solutions.

The third assignment gives the students an opportunity to combine them in pairs, so using the activity think, pair and share they can get the solution. Also, this problem is aimed to provoke them to use learning through doing by programming. Students are divided in groups and they can help each other using discussion activities, or peer training. The generated bit file allows the exploration of two similar devices. After programming FPGA with the bit file, two students can work simultaneously on the lab-board. In the last part of this problem, students use a lab-board to test their solutions. An example of a screenshot from the software used for a part of the solution of a student, on the screen of the computer, is given in Fig. 5. A photo taken when testing with the lab-board is shown in Fig. 5 (the students around the board).

In engineering education, experimentally based learning and the group approach offer activities in which motivation and creativity are essential. These are basic skills that modern students must possess in order to be competitive. Students are given the opportunity to build their own knowledge that can be dynamically modified, along with defining the criteria to be used to solve the problems.

Students enrolled in the courses "Digital Electronics" and "Pulse and Digital Devices" have the opportunity to learn some essential skills, which will make them successful engineers. Lecturers try to improve knowledge

of the students using active and interactive methods and make them competitive in the future.

IV. CONCLUSION

The lectures in the courses "Digital Electronics" and "Pulse and Digital Devices" at the University of Ruse are really interested in improving the engineering education by applying active learning and interactive methods for attracting the students' attention, presented in the paper and promoting them within our university community.

On the basis of configuring the logic blocks in FPGAs complex Boolean functions are implemented. In the course "Digital Electronics" and "Pulse and Digital Devices" students study how to build a circuit using logic gates (for example, only 2-input NAND gates) or some combinational units (for example, decoders and multiplexers). One of the topics covered in the courses is the design of devices implementing Boolean functions using decoders or a cascade of decoders and logic gates, presented in the paper.

The laboratory exercises help the students to develop design experience and problem-solving skills, important for the engineering profession. They gain skills for analyzing, synthesizing and evaluating design information and problems. Therefore, it is very important for the lectures in the universities to develop effective teaching pedagogical methods to involve students in active learning for the laboratory classes in digital electronics.

REFERENCES

[1] A. K. Maini. "Digital Electronics: Principles, Devices and Applications", DOI 10.1002/9780470510520, 2007, research-gate.net/publication/290729475_Digital_Electronics_Principles_Devices_and_Applications/citations

[2] B. Balamuralithara, P. C. Woods, "Virtual laboratories in engineering education: The simulation lab and remote lab." Computer Applications in Engineering Education, pp.108-118, 2009.

[3] J. Pang, "Active Learning in the Introduction to Digital Logic Design Laboratory Course". ASEE Zone III Conference 2015. https://www.asee.org/documents/zones/zone3/2015/Active-Learning-in-the-Introduction-to-Digital-Logic-Design-Laboratory-Course.pdf.

[4] S. Senthamarai, "Interactive teaching strategies", Journal of Applied and Advanced Research, 3(Suppl. 1), 2018, pp. S36-S38.

[5] A. Fidalgo-Blanco, M. Martinez-Nuñez, O. Borrás-Gene, J. J. Sanchez-Medina. "Micro flip teaching – An innovative model to promote the active involvement of students." Computers in Human Behavior, Volume 72, July 2017, pp. 713-723.

[6] L. J. Cymberknop, R. L. Armentano. "Active Learning Approach to Enhance Engineering Education in Argentina: A Case of Study in Signals and Systems." Creative Education, 9, 1444-1456, 2018.

[7] A. Lehtovuori, M. Honkala, H. Kettunen, and J. Leppävirta. "Promoting Active Learning in Electrical Engineering Basic Studies". http://dx.doi.org/10.3991/ijep.v3iS3.2653.

[8] S. Senthamarai. "Interactive teaching strategies". Journal of Applied and Advanced Research, 2018: 3 (Suppl. 1), S36–S38 https://dx.doi.org/10.21839/jaar.2018.v3S1.166, ISSN 2519-9412, Phoenix Research Publishers, 2018.

[9] A. Borodzhieva. "Pulse and Digital Circuits. Laboratory Practice with Modern Circuit Solutions – FPGA (part 1). Ruse: Academic Publishing House "University of Ruse", 2019 (in Bulgarian).

[10] Combinational circuits using Decoder. https://www.geeksforgeeks.org/combinational-circuits-using-decoder/

[11] Logisim, http://www.cburch.com/logisim/

[12] Xilinx documentation, https://www.xilinx.com/content/xilinx/en/downloadNav/design-tools/v2012_4---14_7.htm

Interactive approach to Digital Logic

M. Špoljarić*, M. Hajba* i M. Pecimotika*
* Virovitica College/Department of ICT, Virovitica, Croatia
marijana.spoljaric@vsmti.hr, marko.hajba@vsmti.hr, mario.pecimotika@vsmti.hr

***Abstract* - Digital systems are designed to store, process, and communicate information in digital form, widely used in different spheres both of everyday life and science, including education, process control, digital instruments, computer science, robotics, telecommunications etc. In order to understand how digital systems are operated, basic knowledge of digital circuits and their logical function is required. Digital Logic is the topic shared between three courses taught at the Virovitica College (Mathematics I, Digital electronics, Computer Architecture) as a part of the Undergraduate Professional Study of Electrical Engineering. In this paper interactive approach to the topic of Digital Logic is presented. This is achieved in few steps: students are introduced to the theoretical background of the topic and computational simulation of digital circuits are performed. Based on the theory and simulations, practical exercises at the Laboratory for Electrical Engineering have been done.**

Keywords - digital logic in courses, simulation and design of digital circuits

I. INTRODUCTION

Today's world is based on digital logic. Digital systems are designed to store, process, and communicate information in digital form, widely used in different spheres both of everyday life and science, including education, process control, digital instruments, computer science, robotics, telecommunications etc. In addition to its practical application, digital logic is often subject of learning in different fields of higher education.

Moreover, students who can reason rigorously with formal logic are more likely to succeed in the field of computer science and electrical engineering [1]. In order to design software and hardware systems, students must use propositional logic and Boolean algebra to write specifications, validate designs, test rigorously, and optimize safely. Logic and reasoning skills are usually taught from the earliest phases of education, starting from the primary school in computer science curricula at the most vivid level (e.g. switching light on and off), to spreading the knowledge in more sophisticated curricula (e.g. as part of mathematics, computer science or digital logic classes) which serve as a foundation for many of the courses that follow.

Traditional lectures are more often being replaced by alternative teaching methods (e-learning, visualization and simulation of the problems by laboratory exercises, application of computer software, information and communication technologies, etc.), mostly because of their student-suited approach to learning, discussions and problem-solving. As a result, educators are in constant search for innovative, interactive, interdisciplinary, and contemporary approaches to teaching.

Using an interactive approach, we will present how the knowledge of digital logic is gradually built up, from the theoretical point of view, applying the Boolean algebra to specific problems, which includes software visualization and simulation of the problems and physical assembly of the digital circuits in the laboratory.

II. DIGITAL LOGIC IN COURSES

During their formal education at the Virovitica College, electrical engineering students are introduced to the topic of digital logic in courses of Mathematics I, Computer Architecture, and Digital Electronics. Classes, learning objectives, and contents of all three curricula are designed in a way that students can upgrade their knowledge by working on a different, gradually more complex problems and phenomena, avoiding literal repetition of the same contents and learning outcomes.

A. Mathematics I

The course of Mathematics I is taught during the first semester of the undergraduate study. The learning objectives are to develop propositional logic and reasoning skills, to perform the theoretical analysis of the real-world problems and to design appropriate mathematical models.

The first topic to be discussed is Mathematical Logic. Two hours of lectures and two hours of exercises are planned for the topic in the syllabus. Students are expected to learn how to write a truth table for a given problem, state the laws and rules of the Boolean algebra, and, of course, to apply them to given problems. Binary logical operations such as logical conjunction, logical disjunction, implication and negation are introduced.

Learning outcomes in the topic:

- list the characteristics of basic logical operations,
- apply the basics of Boolean algebra on specific problems,
- write and analyze the truth table for a specific problem.

Learning outcomes are tested through homework assignments, written exams, and class activities. Due to the extensive content of the course of Mathematics I, and the fact that in the course a total of only 180 minutes of lectures and exercises is provided for this topic, simulations of circuits and laboratory exercises are

planned for the courses of Computer Architecture and Digital Electronics.

B. Computer Architecture

The course of Computer Architecture is taught during the second semester of the undergraduate study. The teaching objectives are to familiarize the student with the components of the computers and the different computer architectures.

The third topic inside of course discussed with students is Boolean Algebra and Logic Gates. In two hours of lectures and two hours of exercises, students are expected to learn how to draw, analyze and characterized logic circuit, and how to design them and simulate using computer software. Students are expected to find a Boolean function for a given problem in a sum of products (SOP) and a product of sums (POS) form, to perform the minimization of logical functions using Boolean algebra.

Learning outcomes in the topic:

- minimize a given logical function,
- draw a logic circuit for a given logical function.
- realize a given logic function using logic circuits simulation software.

The exercises of the logic circuits simulations are performed in the laboratory for computer science. For this purpose, computers in the laboratory have installed freeware software which are also available to the students, thus, they can install them on their own computers and use them for repetition, learning, homework assignments, etc. During the exercises, each student has a set of tasks that must be completed during the classes. The support and assistance of the teaching assistant are available to students anytime. Student is expected to submit his or her report in written form before the following exercises. Students reports are graded in such a way that all lab works, and homework carries 20% of the final grade. Students are expected to solve all assigned tasks at least to the extent that satisfies the basic assumptions of each learning outcome, but they are also encouraged to ask questions, make conclusions and assumptions about problems, and approach those problems in their own way.

C. Digital Electronics

The course of Digital Electronics is taught during the third semester of the undergraduate study. The teaching objectives are to familiarize the student with different digital circuit elements ranging from simple gates to arithmetic logic units, such as adders, subtractors etc.

During the semester, students are expected to expand their knowledge of digital design, solve new and more complex problems, learn how to find a solution for a specific problem and design a digital circuit, minimizing complex logic circuits with aim of save the space, time, money, etc. Students are taught how to design combinational, synchronous and asynchronous sequential circuits approaching the same problems in different ways finding optimal and effective solutions.

For the first time, students are about to create specific digital circuits using integrated circuits in dual in-line packages, other electronic components and breadboards. The idea is to build those circuits that students have already designed and simulated in the courses of Mathematics and Computer Architecture, and to compare the results and characterized different forms of the same circuits.

Learning outcomes in the topic:

- distinguish and compare the technological features of digital circuits,
- plan and assemble the circuit for the given logical function,
- analyze the assembled logic circuit,
- compare a circuit with its minimized form,
- choose the most optimal solution to the problem.

III. Practical Approaches to Digital Logic

The goal of using digital logic software is to train students to visually monitor and analyze logic circuits and to help them to design and simulate those circuits [2][3]. Simulations tools are freeware or cost-effective over the longer period of time, they are suitable for demonstration and explanation of the abstract matter, transforming it into interactive visual content, easing the understanding, allowing interaction between student and simulated equipment, providing them with feedback and improving their skills and knowledge. Many different studies prove that the use of simulations can improve learning, e.g. [4].

Some of the successful software for digital logic simulations are: Logic Works, Digital Logic Simulator [5], CircuitVerse online Simulator, Logicly Online Simulator, Digital Logic Design (DLD), etc. In this paper, we decided to introduce two simulation programs, although, one simulation software would be enough for the topic of digital logic. However, for some other topics, such as multiplexers, subtractors, (de)coders, etc., DLD would be more appropriate because a smaller amount of those digital elements is available in CircuitVerse simulator. On the other hand, the advantage of CircuitVerse is its online availability, so by signing up, the users can access their projects and simulations from anywhere.

A. Digital Logic Design

Digital Logic Design is freeware software aimed at providing the tools for building and simulating digital circuits. It provides different digital circuit elements ranging from simple gates to arithmetic logic units, such as full adder, half adder, subtractors etc. In this software, circuit can easily be converted into a reusable module. A module may be used to build more complex circuits like CPU. The circuit working can be analyzed by using output parts like LEDs, seven segment display, digital oscilloscope etc. provided by the software. The application can be used to design combinational, synchronous and asynchronous sequential circuits [6].

B. CircuitVerse Online Simulator

This online simulator is freeware and open source. Digital circuits can be designed and simulated using a graphical user interface. It is designed primarily for educational purposes, and users can test and learn from numerous examples that are already available in the simulator. It is powerful enough to design a CPU, but very easy to learn and work with [7].

C. Breadboarding and Prototyping Circuits

A breadboard (Figure 1) is a solderless device for temporary prototype with electronics and test circuit designs [8]. The layout of a typical breadboard is made up of two types of interconnected electric terminals, so called strips - terminal strips and bus strips - connecting pin to pin inside the breadboard (the arrangement of solderless breadboard internal connections is shown in Figure 1).

Bus strips are used to provide power to the electronic components. Typically, the rows intended for a supply voltage are marked in red, while the rows for ground are marked in blue. The board is powered by 5 V DC, which can be obtained from the power source. However, some breadboards only provide a single-row power distributions bus strip on each vertical side. The inputs to the logic sources are supply voltage (for logical 1) and ground (for logical 0). The circuit output is connected to the logic indicator, e.g. light emitting diode (LED).

On the other hand, terminal strips are used to hold electrical components such as capacitors, resistors, DIPs, etc. ICs (Figure 2) integrated in DIPs can be inserted to straddle the centerline of the board, while into the remaining free holes are inserted other electrical components needed to complete the circuit. In addition to the power source, breadboard and ICs, connection wires and an output indicator such as LEDs are required. Other components, such as resistors and capacitors, can also be required in certain cases. The operation of the ICs used in the experiments is explained by referring to diagrams of similar components introduced by many different references, e.g. [9][10]. Each experiment is presented informally, and the student is expected to design the circuit for given problem and formulate a procedure for verifying truth tables.

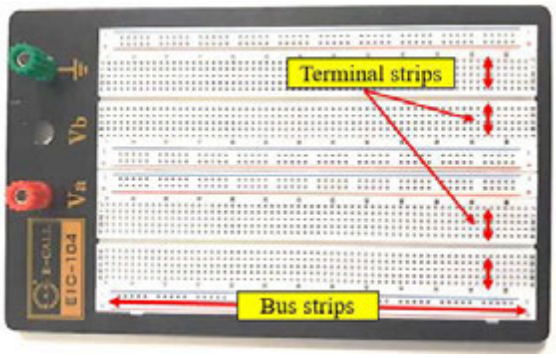


Figure 1. The arrangment of solderless breadboard internal connections.

TABLE I. THE TRUTH TABLE FOR EXAMPLE I.

A	B	$A+B$	$\overline{A+B}$	$\overline{\overline{A+B}}$	m_i	M_i
0	0	0	1	0	m_0	M_0
0	1	1	0	1	m_1	M_1
1	0	1	0	1	m_2	M_2
1	1	1	0	1	m_3	M_3

IV. EXAMPLES FROM PRACTICE

Through three courses, students are encouraged to apply theoretical knowledge to specific problems. To most students, digital logic is an abstract concept. Through simulation exercises and realization of logic circuits on the breadboard using real circuit elements, students are expected to improve their understanding of the digital logic.

Below, we present two examples discussed in the above-mentioned courses. Each example is divided into three sections so one can keep track of the progress and focus on the relevant steps.

A. Example I

a) Mathematics: For the given logical function write truth table and find its minimized form using Boolean algebra:

$$f(A,B)=\overline{\overline{A+B}} \tag{1}$$

In order to accomplish this task, students should know the truth tables for binary logical operations such as logical conjunction, logical disjunction and negation (Table I). It is also important to know the order in which logical operations are performed and how to apply Boolean algebra theorems to a specific problem. The exercises are usually designed to introduce students to the area of Boolean algebra (starting with simpler problems, then expanding their knowledge to a more complex level), learning how it can be used as a tool for problem solving. Implementing the rules of Boolean algebra, one can find minimized form (2) of given logical function.

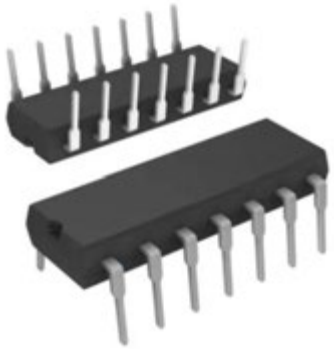


Figure 2. IC (or a chip) is a set of electronic circuits on small piece of semiconductor material [10].

$$\overline{\overline{A+B}} = \overline{\overline{A}\cdot\overline{B}} = \overline{\overline{A}} + \overline{\overline{B}} = A + B \qquad (2)$$

As shown in Table I, (1) and (2) have identical truth tables. It's because (1) can be reduced to (2) by applying Boolean algebra. With this approach to problems, students can find that every complex logic function can be simplified by using the appropriate rules, which later eases the physical implementation of logic circuits.

b) Computer Architecture: Express the given function as a sum of products (sum of minterms) and as a product of sums (product of maxterms):

$$f(A,B) = \overline{\overline{A+B}}\,.$$

Use the Digital Logic Design software to display the logic circuit for the given function.

$$f(A,B) = \overline{A}\cdot B + A\cdot\overline{B} + A\cdot B \qquad (3)$$

$$f(A,B) = A + B \qquad (4)$$

Now, comparing the results of (2) and (4), a student can determine that the POS is indeed a minimized form of some logical function. Function (3) expressed as the SOP can also be minimized by applying the appropriate rules of Boolean algebra, which may lead to further discussions on the topic.

Drawing a logic circuit for a given function and simulating its operation, the student can determine if the output values of each circuit are identical for the same input combinations. One can see in Figures 3 and 4 that the simulation results for both functions (i.e. circuits) are consistent with the results in the truth table, i.e. for both circuits, the LED will not light (the output is 0) only if both inputs are 0.

c) Digital electronics

For given logical function (1) write down the truth table and draw a circuit diagram. Connect the inputs to the logic sources and its output to the logic indicator. Apply various input combinations and observe output for each one. Verify the truth table for each input/output combination. Repeat the process for minimized form of the function. Write down your conclusions.

Students have access to most of the electronic components such as basic logic gates (AND, NOT, OR), gates as NOR, NOT, NAND, XOR, XNOR, flip-flops, counters, seven-segment displays, coders, decoders, etc.

Function (1) could be implemented by inverting NOR gate or using two NOT gates and one OR gate. It is up to the student how he/she will accomplish the given circuit using only the provided components. One of the possible solutions is shown in Figure 5.

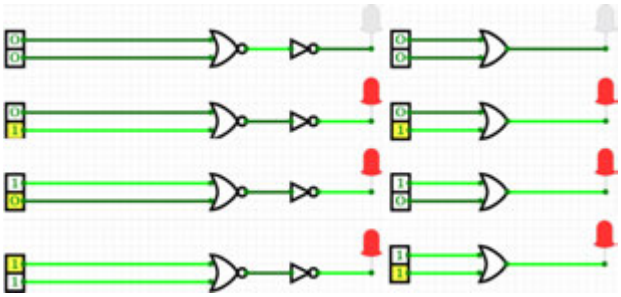


Figure 3. Logic diagram of (1) on the left, and (2) on the right, for input AB combinations 00, 01, 10, 11, respectively. Diagram was drawn with Circuit Verse simulator

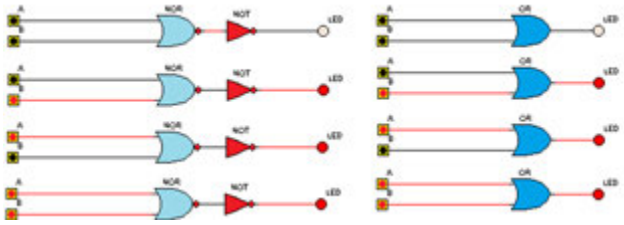


Figure 4. Logic diagram of (1) on the left, and (2) on the right, for input AB combinations 00, 01, 10, 11, respectively. Diagram was drawn with Digital Logic Design.

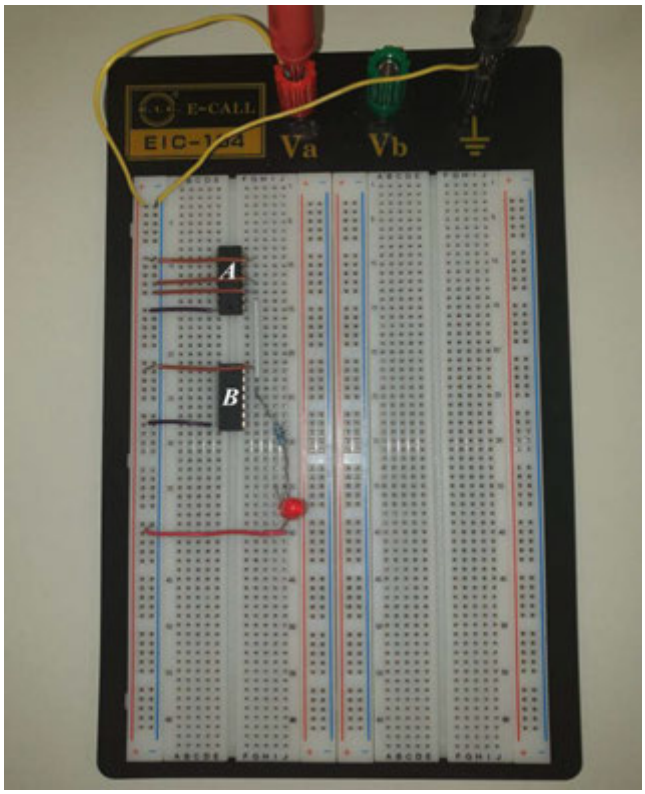


Figure 5. Breadboard prototype of function (1) using NOR gate (IC7402, labeled *A*) and NOT gate (IC7414, labeled *B*).

B. Example II

a) Mathematics: For the given function write truth table and find its minimized form using Boolean algebra theorems:

$$f(A,B)=\overline{\overline{A}\cdot\overline{B}+\overline{A}\cdot B}\ . \qquad (5)$$

In order to accomplish this task, students should know the truth tables for binary logical operations such as logical conjunction, logical disjunction and negation (Table II). It is also important to know the order in which logical operations are performed and how to apply Boolean algebra theorems to a specific problem (5).

$$\overline{\overline{A}\cdot\overline{B}+\overline{A}\cdot B}=\overline{\overline{A}\cdot\overline{B}}\cdot\overline{\overline{A}\cdot B}=\left(\overline{\overline{A}}+\overline{\overline{B}}\right)\cdot\left(\overline{\overline{A}}+\overline{B}\right)$$

$$=(A+B)\cdot\left(A+\overline{B}\right)$$

$$=A\cdot A+A\cdot\overline{B}+B\cdot A+B\cdot\overline{B}$$

$$=A+A\cdot\overline{B}+B\cdot A+0=A\cdot\left(1+\overline{B}+B\right)$$

$$=A\cdot(1+1)=A\cdot 1=A \qquad (6)$$

b) Computer Architecture: Express the given function as a sum of products (sum of minterms) and as a product of sums (product of maxterms):

$$f(A,B)=\overline{\overline{A}\cdot\overline{B}+\overline{A}\cdot B}\ .$$

Use the Digital Logic Design software to display the logic circuit for the given function and its minimized form.

Simulating (5) and (6), students can find that for both functions output always equals *A*, i.e. whenever is *A=1* LED diode will light. Obviously, the whole function is independent of input *B*, thus (5) can be minimized into (6). The output when *A=1*, *B=1* is shown in Figures 6 and 7.

c. Digital electronics

For (5) write down the truth table and draw a circuit diagram. Connect the inputs to the logic sources and its output to the logic indicator. Apply various input combinations and observe output for each one. Verify the truth table for each input/output combination. Repeat the process for minimized form of the function. Write down your conclusions.

Function (5) can be realized in many ways. For example, students may use an XOR gate or OR and NOT gates combined (Figure 8). At this point, students can perform a minimization of the given circuit. Using rules and theorems of Boolean algebra, (5) can be reduced to (8). By implementing both functions on the breadboard, students can verify if the given logic circuits are equivalent, i.e. if equivalent changes in input values for both circuits result in the same output values.

TABLE II. THE TRUTH TABLE FOR EXAMPLE II.

A	B	$\overline{A}$	$\overline{B}$	$\overline{A}\cdot\overline{B}$	$\overline{A}\cdot B$	$\overline{\overline{A}B+A\overline{B}}$	$\overline{\overline{A}\overline{B}+\overline{A}B}$	m_i	M_i
0	0	1	1	1	0	1	0	m_0	M_0
0	1	1	0	0	1	1	0	m_1	M_1
1	0	0	1	0	0	0	1	m_2	M_2
1	1	0	0	0	0	0	1	m_3	M_3

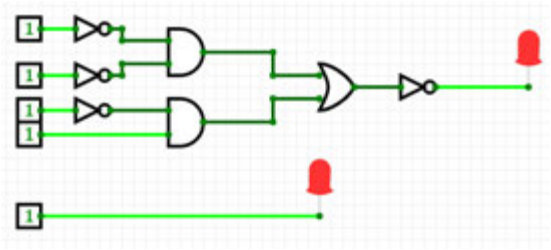


Figure 6. Logic diagram of (5) and (6), respectively, for input combination A=1, B=1. Diagram was drawn with Circuit Verse simulator.

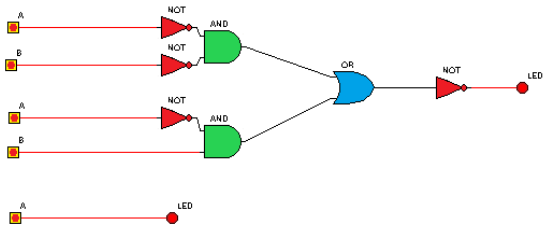


Figure 7. Logic diagram of (5) and (6), respectively, for input combination A=1, B=1. Diagram was drawn with Digital Logic Design.

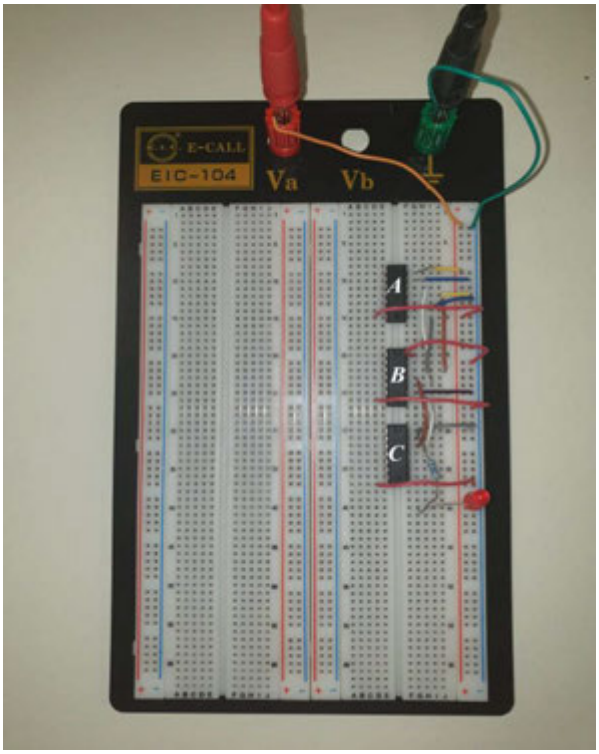


Figure 9. Breadboard prototype of function (5) using NOT gate (IC7414, labeled *A*), AND gate (IC7408, labeled *B*) and NOR gate (IC7402, labeled *C*).

V. Conclusion

In this paper interactive approach to Digital Logic was presented. We shared the topic within the three courses: Mathematics I, Computer Architecture and Digital Electronics. In the first two courses students are introduced to the foundations of the digital logic and Boolean algebra, which is then applied on more complex problems in the course of Digital Electronics. Problems are discussed both from the theoretical and practical point of the view. Students are encouraged to visualize problem with computer software and to realize the solution assembling the circuit on the breadboard.

In future, the aim is to discuss on the students' opinions about mentioned methods and how the presented methods affect the success of overcoming learning outcomes.

References

[1] https://doi.org/10.1145/2133797.2133800 (10th January 2020)

[2] C. Hacker, R. Sitte, „Interactive Teaching of Elementary Digital Logic Design With WinLogiLab“, IEEE Transactions on Education, Vol. 47, No. 2, pp. 196-203, 2004 (references)

[3] C. Hacker, R. Sitte, "A Computer-based Interactive Teaching Software for the Tracing of Logic Levels in a Digital Circuit", Global J. of Engng. Educ., Vol.6, No.1, Australia, pp. 85-90,2002. (references)

[4] Y. Zhu, T. Weng, C.-K. Cheng, "Enhancing Learning Effectiveness in Digital Design Courses Through the Use of Programmable Logic Boards", IEEE Transactions on education, vol. 52, no. 1, London, pp. 151-156, 2009. (references)

[5] V. Pavlović, Ž. Stanisavljević, B. Nikolić, J. Đorđević, "Digital Logic Simulator", Second Eastern European Regional Conference on the Engineering of Computer Based Systems, 2011. (references)

[6] https://sourceforge.net/projects/digitalcircuitdesign/ (10th January 2020)

[7] https://circuitverse.org/ (7th February 2020)

[8] http://wiring.org.co/learning/tutorials/breadboard/ (5th February 2020)

[9] https://www.instructables.com/lesson/Integrated-Circuits-1/ 5th February 2020)

[10] https://learn.sparkfun.com/tutorials/integrated-circuits/all (5th February 2020)

Analysis of secure data deletion and recovery with common digital forensic tools and procedures

Slaven Žulj*, Damir Delija*, Goran Sirovatka**

* VSITE, College for Information Technologies, Zagreb, Croatia

** TVZ, Zagreb University of Applied Sciences, Zagreb, Croatia

Slaven.Zulj@vsite.hr

Damir.Delija@vsite.hr

Goran.Sirovatka@tvz.hr

Abstract: **This paper presents how students practical's is developed and used for the important task forensic specialist have to do when using common digital forensic tools for data deletion and data recovery from various types of digital media and live systems. Digital forensic tools like EnCase, FTK imager, BlackLight, and open source tools are discussed in developed practical's scenarios. This paper shows how these tools can be used to train and enhance student understanding of the capabilities and limitations of digital forensic tools in uncommon digital forensic scenarios. Students' practicals encourage students to efficiently use digital forensic tools in the various professional scenarios that they will encounter**.

Keywords - secure data deletion, digital forensics, data storage media

I. INTRODUCTION

In the teaching of digital forensics, students must acquire practical and theoretical knowledge and a set of skills for their future careers. These goals are defined through learning outcomes in the subjects' curriculum ("Computer Forensics", "Digital Forensics"). Students have to get familiar with tools, common tasks, and procedures in which tools are used, which one usually cannot find in learning objectives, but it is almost equally important since it provides students with self-confidence and understanding of the "artistic" parts of digital forensics. Being able to use different tools and in different scenarios will give broader knowledge of what each tool can provide, flexibility, and understanding of what is the best tool for one of many different common scenarios. That experience and ability to choose the right tool in the real case will be achieved through using various tools even for the same tasks in developed students practical's.

II. DIGITAL FORENSIC TRAINING PRACTICAL TASKS

Digital forensic is defined as the process of preservation, identification, extraction, and documentation of digital evidence which can be used by a court of law [1]. It works with the key concept of digital evidence which is again defined as any probative information stored or transmitted in digital form that a party to a court case may use at a trial [2]. In a nutshell, students must have a good grasp of how to properly obtain digital evidence from storage devices with the available digital forensic tools. During the study, students have to get familiar both with concepts, tools, and procedures. In the VSITE [3] curriculum, there is the only introduction to digital forensic through one introductory subject, while the TVZ [4] program is more complex. The TVZ curriculum introduction is done through the "Digital Forensics" course, during the first semester, while more detailed knowledge is obtained through other courses in the following semesters, like "Computer Forensics", "Organization and Execution of Digital Forensics Investigation" and other elective courses.

During an introduction to digital forensics it is important to stress some key technical concepts like recovery of deleted data and proper artifact handling, which is again the basis of the crucial concepts:

- data deletions / secure data erasure / data sanitation and device destruction;
- device and data recovery and
- forensic storage media preparation.

These concepts are related to a set of tasks which are again part of any digital forensic work and have to be

accomplished through the standard procedures (SOP). Standard procedures again define how to properly use digital forensic tools in such tasks.

As an example, in the preparation phase of a digital forensic investigation, the storage media must be prepared and tested. This cover steps:

1. storage device erasure or scrubbing,
2. testing if scrubbing was successful,
3. re-initialization of storage media and
4. testing if the media is ready and reliable for use in an investigation.

This also covers secure data erasure, in the situation when duplicates of sensitive data have to be reliably destroyed or storage devices used in investigation sanitized. In such scenarios, it is again important to understand how the process works and how to test and record results of erasure and sanitation.

The same situation is with the process of data recovery, which is essential in cases where data was accidentally or purposely erased, it is important to understand how a device or data from it can be erased and how data can be recovered. It also requires a good understanding of what type of clues can indicate the type of erasure on a specific storage device, meaning students should be familiar with the internals of data erasure.

As it is mentioned for student's practical work, different tools provide flexibility and perspective on the process and actions, while also preventing students from knowing only one tool or process. For example, what we did when we developed students practical's is to do the same task with EnCase and FTK Imager is possible, and students have to understand how the tools work and what the tasks are to achieve correct results even if standard procedures are followed. Some important characteristics of digital forensic tools are listed in Table I. Tools like FTK imager, BlackLight and EnCase are commercial forensic tools, but with the important difference in the ability to do write-blocking or preventing changes on the digital storage media which is being investigated. Students must and will be aware during the practicals that they can easily compromise the investigation by accidental writing to the digital storage media with digital evidence. To avoid this, it is essential to have students practical's where students are solving the same set of tasks on the same media but with different tools. To stress write-blocking issues and to present the difference, separate physical write-blocker devices and software must be used, but not only on one type of operating system but all commonly used operating systems. Such an approach enhances a student's understanding of how the critical concept of write-blocking works and how it is presented in various operating systems and forensic platforms and this is the first of developed practical's.

The second practical's that we developed is in diving in the scenarios where data recovery is the goal, students must again use different tools and even a combination of tools to achieve better results. What we have found is that in the recovery test with EnCase it was not possible to recover erased disk partitions, because of a bug in EnCase. But students using EnCase as a framework and its feature were able to emulate hard disk from evidence image and were able to use the Partition recovery tool to recover erased partitions from the forensic image.

The same actions were done through FTK imager or full FTK. Again, in this practical, it was important for students to understand how the tools work and how the tools can be used in a given scenario, but in a forensically sound manner. Students need to remember that the level of standardization among digital forensic tool vendors is very low. For someone without a proper understanding of tools and evidence handling it will be natural after a forensic tool failure to try directly another purpose-designed tool like Partition recovery and possibly damage the evidence and that we this practical's succeeded to overcome. Students will with acquired knowledge and skills use Partition recovery in a forensically sound way and achieve the correct result. The most important that students will gain from the practical's is experience and possibility to understand how to use different, or even non-forensic tools, in the proper forensically acceptable way. Students have learned that such an approach and required expertise are essential in many specific situations where additional non-primary forensic tools have to be

TABLE I. LIST OF TOOLS

Tool	Type	Write-Block	OS	Data Recovery	Data Erasure
EnCase v8	forensic software, a commercial tool, academic license	Yes, soft	MS Windows	Yes	Yes
FTK Imager	forensic software, a commercial tool, free	No	MS Windows /Linux	Yes	No
FTK	forensic software, a commercial tool, academic license	No	MS Windows	Yes	No
Autopsy/Sleuthkit	forensic software, open-source	No	MS Windows /Linux	No	No
Blacklight	forensic software, a commercial tool, academic license	No	MS Windows	Yes	No
Degausser Prodevice ASM120	hardware magnetic media eraser, available only as a demo	No	Stand-alone device	No	Yes
Dd	OS software for disk erase/data copy	No	Linux	No	Yes
Eraser	Software for data erasure, open-source	No	MS windows	No	Yes
Partition recovery	Software for data erasure	No	MS windows	No	No

used in the digital forensic context. The commercial digital forensic tools can be used as framework tools which create a safe forensic environment. Such tools must be varied in use to show differences, but to keep the principle sound and to provide students with valuable experience in choosing a tool that is fit for the task, practitioner expertise and the environment.

Third practical deals with one of the most common tasks as detecting malware in digital evidence. In the malware search scenario, it is essential to use forensic tools to create a forensically safe environment in which ordinary antivirus or similar tools can search through digital evidence. Direct use of an antivirus tool can compromise digital evidence, while the use of a forensic framework tool prevents such development while enabling full antivirus functionality on the digital evidence.

TABLE II. LIST OF ALL TESTED DEVICES

Tested Device Label	Device type	Description
S1	SSD	Corsair Force F120GB3-BK
S2	SSD	Samsung PM830 MZ7PC128HAFU
H1	HDD	Hitachi HTS545025B9A300
H2	HDD	HGST HTS725050A7E630
D1	HDD	Western Digital WD800AAJS
D2	HDD	Maxtor DiamondMax 8S SATA150
D3	HDD	Seagate ST340014A

There are even more complex and sophisticated scenarios that depend on the proper understanding of forensic tool capability and digital evidence handling. It is possible to use the same knowledge and principles in live forensic tasks combined with remote forensic. In such situations, it is even more important to grasp the proper use of the forensic framework. Live forensic is a forensically safe procedure but not 100% forensically correct procedure since it works with a live machine that, by default, changes the state of potential digital evidence. The students need to understand how tools interact with a live machine and data on it to achieve proper results. In such scenarios, forensic tools like EnCase or FTK will be used to access remote machines through a servlet or an agent, presenting remote resources as part of a local forensic framework. This again allows operations like memory extraction and analysis, malware scans in memory and on the disks, registry analysis, data recovery, and other forensic tasks. Understanding how the tools work and interact with the evidence is key to proper usage of the tools without damaging the evidence and achieving proper results.

Different forensic tools must and are used in student practical work, to get an acceptable starting mindset and create a baseline for further professional practical work. To achieve this cooperation among educational institutions it is necessary to prepare materials and procedures and to share knowledge and prepare practicals. The graduation thesis that was done at VSITE, in cooperation with TVZ, and it was used as a base in developing those practical's. The idea was to do basic operations with available, well tested, forensic tools and then use these results as a source for exercises and practical work with other forensic tools available and develop the practical's for education and presented as practical's lab free for academia. The cooperation of academic institutions in developing practical Labs is another goal we achieved.

III. ANALYSIS OF THE SCENARIO TESTS

What was done? In the above-mentioned thesis, various scenarios were tested on a set of storage devices to create student workable scenarios. Devices used in the tests are listed in Table II, while the definition and results of the tests are presented in Table III. Digital forensic tools used in testing are presented in Table I.

Both magnetic disks and solid-state disks were subjected to different types of erasure and then recovery tests. Independent from the thesis work, some additional tests were later done through other, later acquired, forensic tools, but with the same goal and in same context.

TABLE III. RESULTS OF ENCASE TESTS ON DEVICES

Tested device label	Erasure method	Result of EnCase Forensic test
S1	Quick format	Device recognized as formatted, EnCase Forensic was able to recover files
S2	Quick format	Device recognized as formatted, EnCase Forensic was not able to recover files
H1	Quick format	Device recognized as formatted, EnCase Forensic was not able to recover files
H2	Full format	Device recognized as formatted, EnCase Forensic was not able to recover files
D1	Degauss	The device can not be activated
D2	Degauss	The device can not be activated
D3	Physical destruction	The device can not be activated

The purpose of the test was to define erase/recovery for a set of available storage devices, through unique common tools for erasing and recovery. Data erasure was done through format [5] (quick and full), device demagnetization, and physical destruction. In later phases of testing, not included in the graduation thesis, EnCase digital forensic tools feature „wipe drive" was used to erase data. For the EnCase remote forensic data removal was not used since EnCase academic license does not support this feature. If in the future appropriate EnCase enterprise license will be available, this feature will also be tested for student practical work.

Results of the test were as expected from well previously documented cases, the only discrepancy was a

problem with partition recovery using EnCase v8 caused by a bug in EnCase partition recovery module.

The data erasure/destruction test was done on VSITE, while data recovery was done on the TVZ since TVZ has commercial forensic products available. All products used are listed in Table I, with a short description of the capabilities in regards of write blocking and data recovery. Some of the tools are platform-independent, but most of them are based on one operating system. As this is an important characteristic when planning tests and a student's practical work, it is also listed in Table I.

Such tests can be done as part of student practical work. In local academic reality, there are exceptions. Destructive tests like degaussing and physical destruction can not be easily included since it destroys the storage device. The reason for this exception is a lack of required devices, it turns out that it is not easy to obtain permanent access to a degaussing device, while also gathering old hard drives was problematic due to some procedural legal obstacles. To partially overcome this situation, when some of the production devices fail due to hardware errors can then be used for physical destruction tests, in such a way that the device is dismantled, its magnetic surface damaged. The storage device is then re-assembled and finally tested with forensic tolls to conclude if data from the device can be recovered.

Storage media formatting and recovery tests from Table III are modified to be done not only with EnCase and FTK imager but also with all available forensic tools EnCase v8, FTK imager, FTK, Autopsy/Sleuthkit, Blacklight on different operating systems platforms in a form which students can replicate in a laboratory or at home. Such modifications provide flexibility for students and enhance their understanding of the mechanics of these tests and related procedures.

IV. CONCLUSION

Developed practical's work is of the utmost importance for students in the field of digital forensics. It is not easy to achieve satisfactory levels of individual practice for each student but it is necessary to get students through basic concepts of the digital forensic actions like data deletions/secure data erasure/data sanitation and device destruction, device and data recovery, forensic storage media preparation, and finally documenting results and procedures. Such practical's in the future can be derived from a simple test as it is presented, even based on another student's work like graduation thesis or seminar works. It can also be developed with more complex scenarios like live forensics and remote forensics, presenting the importance of the same basic tasks in different environments and with different equipment, devices, and tools. In academic reality, it is not easy to provide the necessary environment and set up for such work because of various limitations and logistics problems. Cooperation among practitioners, institutions, and teachers is necessary to ease or overcome such limitations and provide students with the best available tools and equipment .

REFERENCES

[1] K. Zatyko, "Commentary: Defining Digital Forensics," Forensic Magazine, 2007. https://www.forensicmag.com/article/2007/01/commentary-defining-digital-forensics.

[2] https://en.wikipedia.org/wiki/Digital_evidence

[3] https://www.vsite.hr

[4] https://www.tvz.hr

[5] Microsoft. Format. https://docs.microsoft.com/en-us/windows-server/administration/windows-commands/format.

MS Excel-Based Application for Implementing the Cryptographic Algorithm Shamir's Secret Sharing

A.N. Borodzhieva
University of Ruse "Angel Kanchev"/Department of Telecommunications, Ruse, Bulgaria
aborodzhieva@uni-ruse.bg

***Abstract* – The paper presents the usage of spreadsheets, and more specifically, MS Excel, to build an application that illustrates graphically the secret-sharing algorithm of Israeli cryptographer Shamir. The application helps students learn some aspects of cryptographic information security by presenting this algorithm extremely clearly and making easier to understand the cryptography theory, which is generally difficult because of its mathematical foundations. Numerous examples of Shamir's Secret Sharing obtained by the application with details about the embedded functions used are presented in the paper. The application is used in the training in the course "Telecommunication Security", included as compulsory in the curriculum of the specialty "Internet and Mobile Communications" for Bachelors at the University of Ruse.**

Keywords – Cryptography, Shamir's Secret Sharing, MS Excel, Active Learning.

I. INTRODUCTION

Cryptography is the theory and practice for studying techniques for secure communications in the presence of third parties, known as adversaries. Cryptography constructs and analyses protocols that prevent third parties or the public from reading private messages. Various aspects of information security, for example data integrity and confidentiality, authentication and non-repudiation are basic for modern cryptography. Modern cryptography exists at the intersection of the disciplines of mathematics, physics, computer and communication sciences, and electrical engineering. Some applications of modern cryptography include computer passwords, digital currencies, electronic commerce, chip-based payment cards, and military communications [1, 2, 3, 4].

One of the widely used algorithms in modern cryptography is Shamir's Secret Sharing, an algorithm created by Adi Shamir, an Israeli cryptographer and a co-inventor of RSA (Rivest-Shamir-Adleman) algorithm [1, 2, 3, 5]. Shamir's Secret Sharing is "a form of secret sharing, where a secret is divided into parts, giving each participant its own unique part. To reconstruct the original secret, a minimum number of parts is required. In the threshold scheme, this number is less than the total number of parts. Otherwise, all participants are needed to reconstruct the original secret" [3, 5].

The paper describes MS Excel-based application implementing and graphically illustrating the scheme of Shamir's Secret Sharing. Numerous examples obtained by the application are presented in the paper. Some details about the functions used for the implementation are given. The application is applied in the educational process in the course "Telecommunication Security", compulsory in the curriculum of the specialty "Internet and Mobile Communications" for Bachelors at the University of Ruse but it can be applied in other courses covering topics of cryptography and information protection, such as "Cryptography and Data Security" (for Bachelors in "Computer Systems and Technologies"), "Security in the Telecommunication Networks" (for Masters in "Internet and Multimedia Communications"), "Coding and Security in Telecommunication Networks" (for Masters in "Telecommunication Networks"), etc.

II. MS EXCEL-BASED APPLICATION FOR IMPLEMENTING THE CRYPTOGRAPHIC ALGORITHM SHAMIR'S SECRET SHARING

A. Shamir's secret sharing

Shamir's Secret Sharing is a cryptographic algorithm invented by Adi Shamir. Shamir's Secret Sharing is applied to safeguard the secret, for example the encryption keys, in a distributed manner. It is a type of secret sharing, in which the secret is divided into multiple pieces (parts), known as shares, giving each participant in the communications its own unique part. To reconstruct the secret, i.e. to unlock the secret using Shamir's secret sharing, a minimum number of shares, called threshold, is required. In the threshold scheme, this number is less than the total number of shares. Otherwise, all participants are needed to reconstruct the original secret [5, 6, 7, 8, 9, 10]. Various implementations of Shamir's Secret Sharing scheme are known, for example in Python [5], Ruby [8], C++ for different number of shares (1 – 10) and threshold (2 – 10) for sharing [11] and reconstructing [12], etc.

This paper is supported by the National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)", financed by the Ministry of Education and Science of Bulgaria.

The work presented in this paper is completed as partial fulfilment of Project 2020 – FEEA - 03 "Design and development of a multifunctional robot for implementation and evaluation of autonomous navigation algorithms" financed under the Scientific and Research Fund of the University of Ruse "Angel Kanchev".

A typical example for using Shamir's Secret Sharing is as follows: It is necessary for a company to make secure a password (for example a pass-code of their vault) and they could use a standard, for example AES (Advanced Encryption Standard), but if the keeper of the key is not available or dies, or is a crook, or if the key is compromised by a mischievous hacker, then not using secret sharing should be a great problem. In this case, Shamir's Secret Sharing can be applied for encrypting the pass-code and for generating a given number of shares, and some of these shares are distributed to each executive in the company (Fig. 1). They can unlock the vault, only if they combine their shares. The threshold can be suitably chosen according to the number of executives, so that the vault can be accessed by the authorized persons. If a share falls into the wrong persons, they would not be able to open the vault except the other executives collaborate them [5, 6, 7, 8, 9, 10].

B. *Mathematical definition*

In Shamir's Sharing Secret scheme, the goal is dividing the secret S (for example, the passcode for a safe or a vault) into n parts $S_1,\ldots,S_n$ so that: **1)** Knowing any k or more parts S_i leads to an easy calculation of S, i.e. the secret S can be restored from any combination of k parts. **2)** Knowing any $k-1$ or fewer parts S_i leads S to be completely undetermined, i.e. the secret S cannot be restored with fewer than k parts. This scheme is known as (k,n) threshold scheme. If $k=n$ then every part of the secret S is necessary for restoring the secret (Fig. 1).

The main idea of Shamir's threshold scheme is that an infinite number of polynomials of degree 2 might be drawn through 2 points, i.e. 2 points are enough for defining a line, 3 points – for a parabola, 4 points – for a cubic curve, etc., i.e. k points are necessary for defining a polynomial of degree $k-1$ (Fig. 2). The picture (Fig. 2) is only for illustration, because Shamir's scheme uses polynomials over a finite field that cannot be represented on a 2D plane [5].



Figure 1. Shamir's Secret Sharing scheme [9]

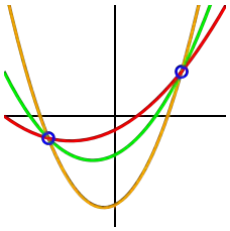


Figure 2. Main idea of Shamir's Secret Sharing scheme [5]

Assume a (k,n) threshold scheme is used for sharing the secret S, supposed to be an element in a finite field F of size P, where $0<k\le n<P$; $S<P$ and P is a prime. Then $k-1$ positive integers $a_1,\ldots,a_{k-1}$ where $a_i<P$ are chosen randomly and $a_0=S$. Then the polynomial of degree $k-1$ $f(x)=a_0+a_1x+a_2x^2+a_3x^3+\ldots+a_{k-1}x^{k-1}$ is built. Further, any n points of it are constructed, for example $i=1,\ldots,n$ to restore $(i,f(i))$. The prime P used for defining the finite field F and a point (a non-zero input of the polynomial and its corresponding output, both of them as integers) are given to each participant. The coefficients of the polynomial can be found using interpolation based on any subset of k of these pairs. The secret is the coefficient a_0.

C. *Example of Shamir's Secret Sharing and its implementation in MS Excel*

The principle of operation of Shamir's Secret Sharing is illustrated by the following example, where the calculations are realized based on integer arithmetic instead of finite field arithmetic. This example is not a real example of Shamir's scheme and does not envisage perfect secrecy, but it is appropriate for training purposes.

In the example it is supposed that the secret for sharing is the number 1325 $(S=1325)$ and it will be divided into 6 parts $(n=6)$, where any subset of 3 parts $(k=3)$ will be enough for reconstructing the secret. Randomly $k-1=2$ numbers, 197 and 65, are chosen $(a_0=1325; a_1=197; a_2=65)$, where the coefficient a_0 is the secret. Therefore, the polynomial for creating the secret shares (points) is:

$$f(x)=1325+197x+65x^2. \tag{1}$$

From the polynomial (1), six points $D_{x-1}=(x,f(x))$ are constructed:

$$\begin{aligned} &D_0=(1,1587); D_1=(2,1979); D_2=(3,2501);\\ &D_3=(4,3153); D_4=(5,3935); D_5=(6,4847). \end{aligned} \tag{2}$$

A different point (x and $f(x)$) is given to each participant. The points start from $(1,f(1))$ instead of $(0,f(0))$, because D_{x-1} is used instead of D_x and $f(0)$ is the secret.

This first part of the algorithm is implemented in MS Excel and it is presented in Fig. 3. The numbers $n=6$ and $k=3$ are fixed. In the case, $(3,6)$ threshold scheme is used. Then the user chooses the values of the secret S and two random numbers. In this case $S=1325$ and these two numbers are 197 and 65. Then the application fills in the values of the coefficients of the polynomial $a_0=S, a_1, a_2$ and the polynomial $f(x)=a_0+a_1x+a_2x^2$. When constructing the six points $D_{x-1}=(x,f(x))$, the first coordinates are fixed (from 1 to 6) and the second coordinates are calculated using the polynomial $f(x)=a_0+a_1x+a_2x^2$. The calculations are realized in MS Excel using the appropriate formulas as follows: *=$AN$12+$AN$13*AP20+$AN$14*AP20^2*. Here, *AP20*

is the relative address of the cell with the value of x and *$AN$12*, *$AN$13* and *$AN$14* are the absolute addresses of the cells where the coefficients of the polynomial are located. For reconstructing the secret, any 3 points will be enough, so that in the bottom of the first part of the applications three points are selected.

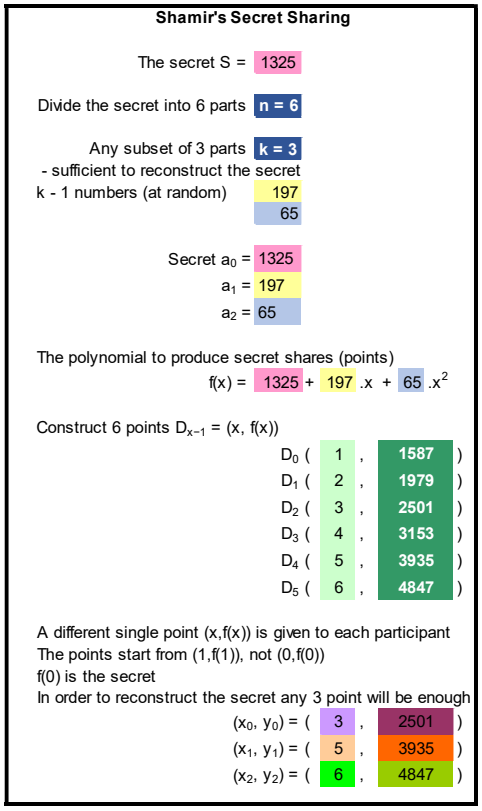


Figure 3. Example of Shamir's Secret Sharing and its implementation in MS Excel – part I

Reconstruction

In order to reconstruct the secret S any 3 points will be enough, for example, the points given in (3):

$$\begin{aligned}(x_0, y_0) &= (3, 2501);\\ (x_1, y_1) &= (5, 3935);\\ (x_2, y_2) &= (6, 4847).\end{aligned} \tag{3}$$

These three points are chosen in the bottom of the first part of the application (Fig. 3).

Then Lagrange basis polynomials are computed and they are presented with the expressions (4)…(6):

$$\ell_0 = \frac{x-x_1}{x_0-x_1}.\frac{x-x_2}{x_0-x_2} = \frac{x-5}{3-5}.\frac{x-6}{3-6} = \frac{1}{6}x^2 - \frac{11}{6}x + 5 \tag{4}$$

$$\ell_1 = \frac{x-x_0}{x_1-x_0}.\frac{x-x_2}{x_1-x_2} = \frac{x-3}{5-3}.\frac{x-6}{5-6} = -\frac{1}{2}x^2 + \frac{9}{2}x - 9 \tag{5}$$

$$\ell_2 = \frac{x-x_0}{x_2-x_0}.\frac{x-x_1}{x_2-x_1} = \frac{x-3}{6-3}.\frac{x-5}{6-5} = \frac{1}{3}x^2 - \frac{8}{3}x + 5 \tag{6}$$

Therefore, the polynomial will be given with the expression (7):

$$\begin{aligned} f(x) &= \sum_{j=0}^{2} y_j.\ell_j(x) = y_0.\ell_0(x) + y_1.\ell_1(x) + y_2.\ell_2(x) = \\ &= 2501\left(\frac{1}{6}x^2 - \frac{11}{6}x + 5\right) + 3935\left(-\frac{1}{2}x^2 + \frac{9}{2}x - 9\right) + \\ &+ 4847\left(\frac{1}{3}x^2 - \frac{8}{3}x + 5\right) = 1325 + 197x + 65x^2 \end{aligned} \tag{7}$$

As it was mentioned above, the secret is the free coefficient in this polynomial, which means that $S = 1325$ and the reconstruction is done.

The second part of the algorithm is also implemented in MS Excel and it is presented in Fig. 4 and Fig. 5. The application is designed with the aim to allow students to follow all the calculations and transformations according to formulas (4)…(7) (with all these details), which makes it extremely suitable for students in universities with poor mathematical background, as well as for students in schools who study the relevant material in mathematics.

Here, all the cells in colors are filled in automatically, these are the coordinates x and $f(x)$ of the three points (in the bottom of part I in Fig. 3). When Lagrange basis polynomials are computed, the formula (8) is used:

$$(x-x_i).(x-x_j) = x^2 - (x_i + x_j)x + x_i.x_j \tag{8}$$

When calculating Lagrange basis polynomials, for determining the denominators of fractions a formula of the form is repeatedly used: *=ABS((F37-H37)*(J37-L37))*, and the result is always obtained as a positive number.

When determining the sign of each fraction (positive or negative) and its correct visualizing, a loop of the type *IF* is repeatedly used:

=IF((F37-H37)(J37-L37)<0;"+";"-")*

In order to correctly write down the difference obtained (with a negative result), after the multiplication sign for its correct visualization, the obtained result should be enclosed in brackets, which is achieved by a formula of the form: *=IF((J37-L37)<0;CONCATENATE("(";(J37-L37);")");(J37-L37))*.

When calculating Lagrange basis polynomials, the following formula is also used: *=LCM(V36;V37)/V37*. The function *LCM* has the following arguments: *Number1*, *Number2*, ..., where *Number1* is required, but the subsequent numbers are optional. If the value is not an integer, it is truncated [13].

All of these formulas use absolute or relative addresses of cells that contain numerical values used to calculate Lagrange basis polynomials.

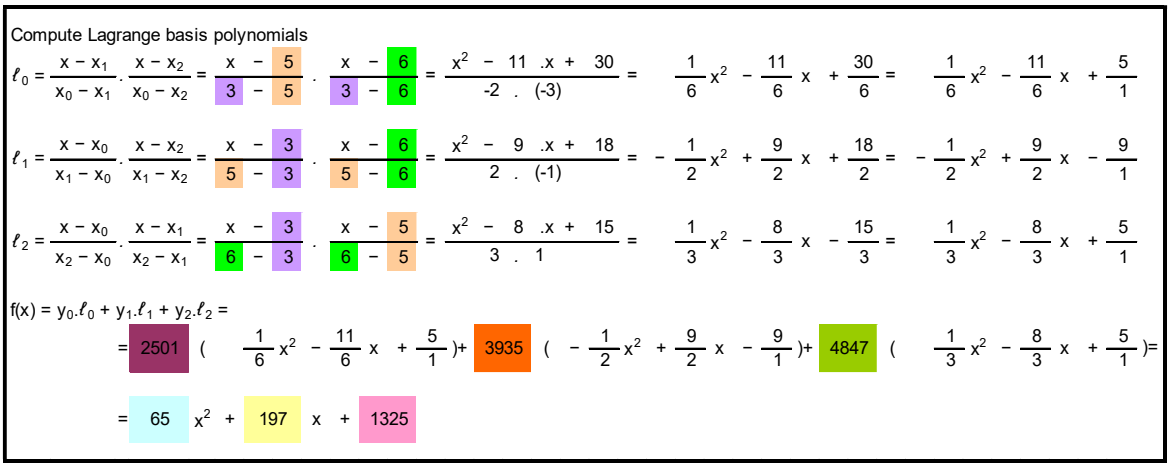


Figure 4. Example of Shamir's Secret Sharing and its implementation in MS Excel – part II (a view in MS Excel)

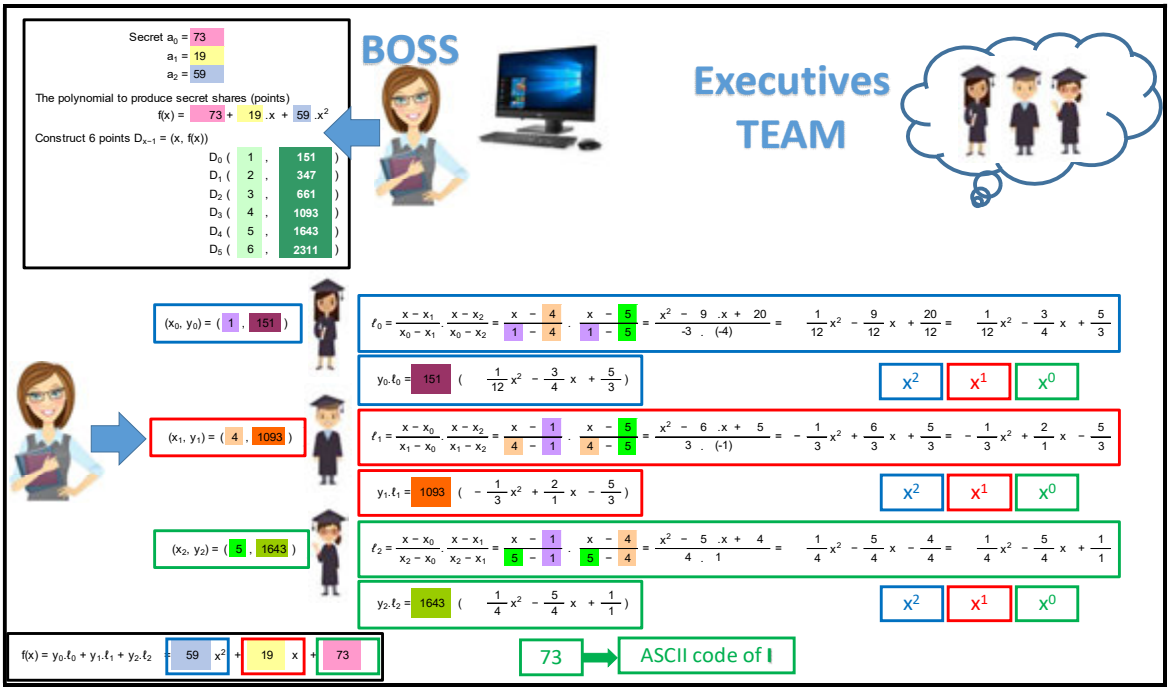


Figure 5. Activities of the lecturer (the boss) and the students (the executives) in Shamir's Secret Sharing scheme during the active learning with detailed calculations implemented in MS Excel (part II)

III. Active Learning in the Course "Telecommunication Security"

For better and lasting perception of the material being studied the active learning methods such as one-minute papers and think-pair-share, decision-making activities, case-based and problem-based learning, etc. are used [14]. Modelling and simulating give the students opportunities for solving diversified practical tasks and understanding better the material. The research approach using teamwork develops the critical and adaptive thinking, creativity and innovation, inquiry and problem solving, communication and collaboration skills [14, 15].

Using active learning methods during the practical classes, the lecturers see the students who have learnt the new lesson and the students who have not. Students' self-assessments incorporate students in evaluating their own work and learning progress and they can identify their own skill "holes" and see where their attention in learning to be focused, revise their work and track their own progress. Self-assessment helps students be involved and motivated and encourages responsibility for their learning [14, 15].

In order to facilitate the learning of the taught material, to attract the students' attention to the material, as well as to intrigue them on topics in the field of cryptography, active teaching methods are used during the training in the course "Telecommunication Security" at the University of Ruse [16]. The activities of the lecturer (the boss) and the students (the executives) in Shamir's Secret Sharing scheme during the active learning are presented in Fig. 5.

Before the class, the teacher devises some appropriate texts related to the field of cryptography. In this case, the text that will illustrate the topic being studied is SHAMIR'S SECRET SHARING (only uppercase letters are used, but a combination of lowercase and uppercase letters, such as Shamir's Secret Sharing, could be used). The teacher represents each of the symbols (including spaces and apostrophes) in the 7-bit ASCII code using their decimal equivalents [16]. In this case, this is the sequence of decimals: 83, 72, 65, 77, 73, 82, 39, 83, 32, 83, 69, 67, 82, 69, 84, 32, 83, 72, 65, 82, 73, 78, 71. Using the text of lowercase and uppercase letters, this sequence will be 83, 104, 97, 109, 105, 114, 39, 115, 32, 83, 101, 99, 114, 101, 116, 32, 83, 104, 97, 114, 105, 110, 103. In fact, these decimal numbers will represent the "secrets" that must be found. The determination of the decimal equivalent of the ASCII code of a symbol is performed with the built-in CODE function in MS Excel. In solving the problem by hand, the eight-bit EBCDIC code could also be used [16].

Then, for each symbol, the teacher thinks up two numbers, and forms a second-degree polynomial whose free term is the "secret" and the two fictitious numbers are the other two polynomial coefficients (before x^1 and x^2) used in the Shamir's Secret Sharing scheme, calculating the 6 points with their coordinates $(x, f(x))$, using the first part of the application developed in MS Excel (Fig. 3).

During the practical classes, the teacher (the boss) divides the students into groups of 3 students (teams, denoted as E in Table I, where the teams are numbered from 1 to 23) and gives each group three points with known coordinates $(x, f(x))$. Students (the executives) compile Lagrange basis polynomials, each student for his/her point $(x, f(x))$, using the knowledge of the other two students for their points $(x, f(x))$. Then the student multiplies the ordinate of his/her point $f(x)$ by the Lagrange basis polynomial he/she has composed. Finally, the students sum the coefficients before the equal degrees of x, for example, the first student calculates the coefficients before x^2, the second – before x^1, and the third – before x^0, i.e. the free term, which is also the "secret". In practice, it is sufficient to calculate only this last coefficient, but for the sake of completeness, each student calculates his or her own part. After calculating the free coefficients of the second-degree polynomial, i.e. after text detection, the student team transforms the decimal number representing the ASCII code into the symbol itself. Each team (that is numbered) shares their symbol and at the end of the class, the students discover the text, which is set as "secret".

For example, in Fig. 5, for team № 5, for the symbol "I" (in the word SHAMIR, with ASCII code $73_{(10)}$), the teacher thinks up two numbers: $a_1 = 19$ and $a_2 = 59$, and forms a second-degree polynomial whose free term is the "secret" $a_0 = 73$ and the two fictitious numbers are the other two polynomial coefficients (before x^1 and x^2) used in the Shamir's Secret Sharing scheme, i.e. the polynomial will be $f(x) = 73 + 19x + 59x^2$. Then the teacher, using the first part of the application developed in MS Excel (Fig. 3), calculates the 6 points with their coordinates $(x, f(x))$. In this case, they are as follows:

$D_0 = (1, 151)$; $D_1 = (2, 347)$; $D_2 = (3, 661)$; $D_3 = (4, 1093)$;

$D_4 = (5, 1643)$; $D_5 = (6, 2311)$.

TABLE I. RESULTS FOR IMPLEMENTING SHAMIR'S SECRET SHARING IN THE TRAINING PROCESS

			a_0							
E	**S**	**A**	x_0	x_1	x_2	a_1	a_2	y_0	y_1	y_2
1	S	83	1	2	3	11	51	145	309	575
2	H	72	4	5	6	13	53	972	1462	2058
3	A	65	1	3	4	15	55	135	605	1005
4	M	77	2	5	6	17	57	339	1587	2231
5	I	73	1	4	5	19	59	151	1093	1643
6	R	82	2	3	6	31	11	188	274	664
7	'	39	3	4	5	33	13	255	379	529
8	S	83	1	2	6	35	15	133	213	833
9		32	2	3	5	37	17	174	296	642
10	S	83	1	4	6	39	19	141	543	1001
11	E	69	3	5	6	51	91	1041	2599	3651
12	C	67	1	2	4	53	93	213	545	1767
13	R	82	2	3	4	55	95	572	1102	1822
14	E	69	1	5	6	57	71	197	2129	2967
15	T	84	3	4	6	59	73	918	1488	3066
16		32	1	2	5	71	51	154	378	1662
17	S	83	1	3	6	73	53	209	779	2429
18	H	72	2	4	5	75	55	442	1252	1822
19	A	65	1	3	5	77	57	199	809	1875
20	R	82	2	4	6	79	59	476	1342	2680
21	I	73	2	3	5	91	75	555	1021	2403
22	N	78	1	4	6	93	77	248	1682	3408
23	G	71	1	2	5	95	79	245	577	2521

Then the teacher gives the team № 5 three points, for example in Fig. 5: $(x_0, y_0) = (1, 151)$; $(x_1, y_1) = (4, 1093)$; $(x_2, y_2) = (5, 1643)$. Next, the students compile Lagrange basis polynomials, each student for his/her point, using the knowledge of the other two students for their points. More details about these calculations can be seen in Fig. 5, i.e. $\ell_0 = \frac{1}{12}x^2 - \frac{3}{4}x + \frac{5}{3}$ (student № 1), $\ell_1 = -\frac{1}{3}x^2 + \frac{2}{1}x - \frac{5}{3}$ (student № 2) and $\ell_2 = \frac{1}{4}x^2 - \frac{5}{4}x + \frac{1}{1}$ (student № 3) are the final results. Then the students multiply the ordinate of his/her point by the Lagrange basis polynomial he/she has composed: $y_0.\ell_0(x) = 151\left(\frac{1}{12}x^2 - \frac{3}{4}x + \frac{5}{3}\right)$ (for student № 1), $y_1.\ell_1(x) = 1093\left(-\frac{1}{3}x^2 + \frac{2}{1}x - \frac{5}{3}\right)$ (for student № 2), $y_2.\ell_2(x) = 1643\left(\frac{1}{4}x^2 - \frac{5}{4}x + \frac{1}{1}\right)$ (for student № 3).

Students who find it difficult to solve this task could use the second part of the application developed in MS Excel, which allows them to visualize in detail the necessary steps (Fig. 5).

Finally, the students sum the coefficients before the equal degrees of x, for example, the first student calculates the coefficient before x^2, the second – before x^1, and the third – before x^0, i.e. the free term, which is also the "secret". In this case, the results will be: $59x^2$ (for student № 1), $19x$ (for student № 2), 73 (for student № 3). After

calculating the free coefficient of the polynomial, the team transforms the decimal number 73 representing the ASCII code into the symbol "I". Each team that is numbered shares their symbol and at the end of the class, the students discover the text, which is set as "secret".

The results for all of the teams used in the training process are presented in Table I.

Before the classes, the teacher compiles such a table (as Table I), using the application – part I and part II. This table allows the teacher during the classes to check the results of students' calculations quickly, mainly in order not to errors accumulate, as well as finding students who successfully complete the task and those who fail.

Now, the application is not developed to allow a fixed number of points to be given to each operation, because the students solve their assignment manually, because sometimes using software tools does not lead to understand the principles of the processes under consideration.

In the academic 2018-2019 year, an experiment was conducted with 69 students (23 teams of 3 students). After teaching the theory about Shamir's Secret Sharing and making some exercises by the students, a self-assessment with the students was done. They were asked about their comments and their feelings on the lesson and the material studied. Each student was given three cards with the following meaning, representing how the student feels about the material studied in the class: a green card ("Easy! I get it! I can do this by myself!"), a yellow card ("OK! I need a little more help!"), and a red card ("Hard! I still need a lot of help") [16]. Only 9 of them (13 %) chose the red card, 23 of them (33 %) chose the yellow card and 37 of them (54 %) chose the green card. The students from the red and yellow groups shared that the application with the detailed step-by-step implementation of the calculations helped them to understand the principle of operation of the algorithm. However, all the students said that they had a lot of fun during the exercise and were most curious to find the "hidden text".

Using this methodology the teacher might see which students had difficulties in understanding the lesson, which students understood the new lesson and which students struggled a bit [16]. The cards divided students into groups according to their learning level. Using the self-assessment idea "networking for help" [16], students with green cards were paired up with students with red or yellow cards, students with yellow cards were paired up with students with yellow or green cards, and students with reds card were paired up with student with green cards. Finally students shared that this way of learning was interesting and helped them a lot in learning the new topic.

Self-assessment involves students in evaluating their own work and learning progress allowing them to: 1) identify their own knowledge and skill gaps; 2) see where to focus their attention in learning; 3) set realistic goals; 4) revise their work; 5) track their own progress. This process helps students to be involved and motivated and encourages self-reflection and responsibility for their learning [16].

IV. CONCLUSION

The practical exercises conducted in the course "Telecommunication Security" at the University of Ruse assist the students to build up skills for analyzing and synthesizing, as well as evaluating and problem-solving, substantial for future computer and communication engineers. Therefore, it is very considerable for lecturers at universities to evolve efficient teaching methods to ensnare students in active learning process during the practical exercises in this course for cryptography.

The topic could be presented in an accessible manner for pupils in the schools, without using the terminology required for students with the aim to attract future students to our university. Using groups of three or four students for reconstructing the original secret, known only to the teacher, makes the topic more attractive for the future students.

REFERENCES

[1] C. Paar, J. Pelzl, "Understanding Cryptography. A Textbook for Students and Practitioners", Springer-Verlag Berlin Heidelberg, 2010, http://swarm.cs.pub.ro/~mbarbulescu/cripto/Understanding%20Cryptography%20by%20Christof%20Paar%20.pdf.

[2] W. Stallings, "Cryptography and Network Security Principles and Practices", Fourth Edition, Prentice Hall, 2005, http://www.inf.ufsc.br/~bosco.sobral/ensino/ine5680/material-cripto-seg/2014-1/Stallings/Stallings_Cryptography_and_Network_Security.pdf.

[3] D. Boneh, V. Shoup, "A Graduate Course in Applied Cryptography", 2020, https://toc.cryptobook.us/book.pdf.

[4] Cryptography. https://en.wikipedia.org/wiki/Cryptography

[5] Shamir's Secret Sharing. https://en.wikipedia.org/wiki/Shamir%27s_Secret_Sharing

[6] A. Poyiatzis, "Shamir's Secret Sharing – A numeric example walkthrough", 2018, https://medium.com/@apogiatzis/shamirs-secret-sharing-a-numeric-example-walkthrough-a59b288c34c4.

[7] Understanding Shamir's Secret Sharing, https://medium.com/vault12/understanding-shamirs-secret-sharing-6a4bd27768c9.

[8] Divide and Manage Secret Data Securely With Shamir's Secret Sharing, http://kimh.github.io/blog/en/security/protect-your-secret-key-with-shamirs-secret-sharing/.

[9] S. Arora, "Taste of Cryptography: Secret Sharing and Secure Multiparty Computation", Advanced Algorithm Design, Princeton University, https://www.cs.princeton.edu/courses/archive/fall16/cos521/Lectures/lec21.pdf.

[10] J.-S. Coron, E. Prouff, and T. Roche, "On the Use of Shamir's Secret Sharing Against Side-Channel Analysis", https://www.ssi.gouv.fr/uploads/IMG/pdf/aesshamir_Coron_Prouff_Roche.pdf.

[11] Shamir Secret Sharing, https://asecuritysite.com/encryption/shamir

[12] Shamir Secret Sharing (Rebuild), https://asecuritysite.com/encryption/shamir_decode

[13] LCM function. https://support.office.com/en-us/article/lcm-function-7152b67a-8bb5-4075-ae5c-06ede5563c94

[14] C. J. Brame, "Active Learning, https://cft.vanderbilt.edu/guides-sub-pages/active-learning/.

[15] M. Bialik, Ch. Fadel, "Skills for the 21st Century: What Should Students Learn?", Center for Curriculum Redesign, 2015, https://curriculumredesign.org/wp-content/uploads/CCR-Skills_FINAL_June2015.pdf.

[16] Borodzhieva, A. Modeling of Cryptosystems Based on Linear Feedback Shift Registers Using Spreadsheets. 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2019, 20 – 24 May 2019, Opatija, Croatia, Engineering Education, Proceedings, pp. 1713 – 1718, ISBN: 1847-3946.

SSE

International Conference on

SOFTWARE AND SYSTEMS ENGINEERING

Steering Committee

Chairs: Tihana Galinac Grbac, Juraj Dobrila University of Pula, Croatia
Darko Huljenić, Ericsson Nikola Tesla d.d., Zagreb, Croatia

Program Committee

Members: Stipo Čelar, University of Split, Croatia
Andrej Grgurić, Ericsson Nikola Tesla d.d., Zagreb, Croatia
Igor Ljubi, Ministry of Public Administration, Croatia
Mladen Sokele, Zagreb University of Applied Science, Croatia
Nikola Tanković, Juraj Dobrila University of Pula, Croatia
Linda Vicković, University of Split, Croatia

PAPERS

SOFTWARE SYSTEM ARCHITECTURES

Database Integration Systems

J. Dončević, K. Fertalj
Faculty of Electrical Engineering and Computing, Department of Applied Computing, Zagreb, Croatia
juraj.doncevic@fer.hr kresimir.fertalj@fer.hr

***Abstract* – Database integration systems are not a novel concept in computing, but with frequenting application of NoSQL databases in real-world information systems the requirement of homogenizing structured and unstructured data is being expanded. In contrast to former applications on relational database systems, database integration systems now must also encompass schemaless data. The aim of this paper is to present a taxonomy, architectural alternatives and patterns of database integration systems, considering newer developments in the field of database systems.**

Keywords – database systems, integration systems, hybrid systems, heterogeneous systems, system architecture

I. Introduction

With the increase of interest for processing and storing big data it turned out that standard relational databases are not completely appropriate for such use-cases. In face of such deliberations, the idea of NoSQL systems started to take hold. Even though NoSQL database systems are a more recent invention than SQL (relational) systems, the superiority of NoSQL systems in terms of big data processing is becoming more apparent [1].

Despite the advantages of NoSQL systems, there is still an emphasized need for relational database management systems. The reason for this are the advantages of SQL systems regarding system security, data consistency and transaction mechanisms. It is apparent that most users are not interested in storing sensitive data in unstructured or semi-structured form.

Considering this, it is possible to conclude that SQL and NoSQL system are going to share the market for some time in the future [2]. It can also be expected that both types of systems are going to be used within the same business systems.

Also worth noting are NewSQL databases, providing ACID properties over large amounts of structured data that can be replicated or sharded[3]. They are systems of a relational nature, so in broader terms they can be viewed as relational systems.

Since the 1980s there has been an interest in systems that enable acquisition and joining of data from different relational systems[4]. It seemed that the field was well studied in the 1990s, but the resulting taxonomy was inconsistent and likely only adequate for use in specific cases – without the possibility of interchangeable or expanded use.

With the popularization of NoSQL databases new opportunities are appearing for expansion and systematization in the field of database integration systems. It is safe to assume a growing need for development of a technology that would enable connecting and integrating existing database systems at a higher level. The intention of this paper is to create a scientific and technological overview of the field of database integration systems, with the inclusion of NoSQL databases as possible data sources. This is achieved through the elaboration of a taxonomy and the presentation of key concepts and architectures.

II. Taxonomy

It is important to note that the taxonomy in this field is not completely defined. While examining past research it is expectable to stumble upon different ways of naming identical concepts. It is possible to illustrate a few examples. For instance, one author considers multi-database systems a category [5], while another author considers multi-database systems a subcategory [6]. One author names a defined concept as a federated system, a second author names it as a multi-database system [7], a third author considers them two different concepts [8]. This paper tries to reconcile the taxonomical differences in its used sources.

Database integration systems (or Multiple Sources Integration according to [9]) are database management systems with a task of integrating data of two or more databases (Figure 1.).

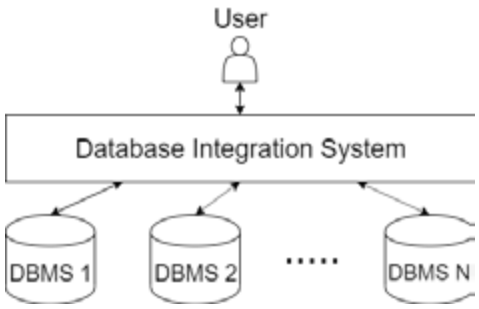


Figure 1. Conceptual schema of relationships between elements and users of a database integration system

Database integration system can be separated into two larger groups, depending on their mode of work and use: *virtualizing* and *materializing*.

Virtualizing systems store data in connected database systems, while materializing systems store data in a single data store. Virtualizing systems are used for unifying views on multiple connected databases. Materializing systems are used for unifying data in a single data store, as support for ETL (Extract Transform Load) processes during data warehouse creation.

Database integration systems basically integrate three database components: *schemas*, *queries* and *query results*.

Implementation variants of database integration systems are determined by three dimensions [5]: *heterogeneity*, *autonomy* and *distribution*.

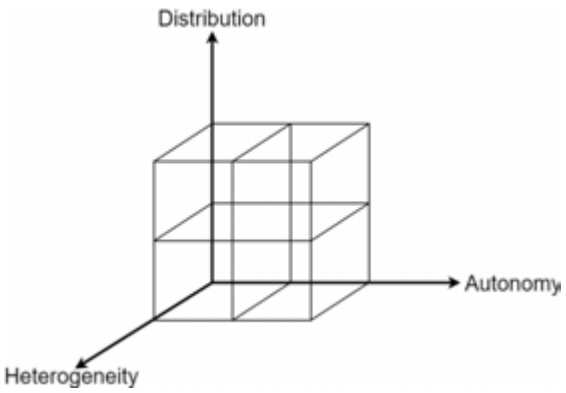


Figure 2. Implementation variants of database integration systems [5]

A. Heterogeneity

Database management systems that integrate multiple databases can be separated into two categories: homogeneous and heterogeneous [6]. The property of homogeneity and heterogeneity is determined by comparing the technological differences of connected database management systems. If an integration system supports only the connection of the same database management system products (e.g. only MySQL databases are supported), then the system is considered homogeneous. Vice versa, if the system supports the connection of different database management system products, then the system is considered heterogeneous.

There are five categories of heterogeneity that can be determined on a heterogeneous system: *technical* [6] [10], *model or structural* [7] [11], *semantic* [6], *syntactic* [10] and *paradigmatic*.

Technical heterogeneity implies the aforementioned condition for heterogeneity. If an integration system connects database systems of different technical implementations, then the system is considered technically heterogeneous.

Model or structural heterogeneity implies a difference in models of connected databases. For a system to be structurally heterogeneous it needs to meet the condition of connecting at least one database with a different model.

Semantic heterogeneity implies a structurally different display of semantically equal data in the connected databases. With the satisfaction of this category, schema conflicts are expected to appear [11].

Syntactic heterogeneity implies a difference in query languages of connected systems.

For the sake of a clearer system categorization, this paper introduces an additional heterogeneity category – paradigmatic heterogeneity. Paradigmatic heterogeneity implies the satisfaction of all the categories above but is used to indicate that a system is heterogeneous in terms of connecting SQL and NoSQL databases. It is important to note that this statement cannot be reversed – a system that satisfies all the categories above does not have to satisfy the paradigmatic heterogeneity category. A system that satisfies the category of paradigmatic heterogeneity is called a *hybrid database system*.

It is interesting to note that some authors [6] consider the criteria of heterogeneity could also encompass the operating systems on which the database systems are running.

B. Autonomy

In database integration systems autonomy is a measure of operation independency of connected systems. There are three categories of autonomy [6] [7]: *design autonomy*, *communication autonomy* and *execution autonomy*.

Design autonomy entails the autonomy of model design in connected database systems. An integration system satisfies this category if connecting and proper functioning of the connected system inside the integration system does not require changes to the connected system's data model. An example would be a relationship between two tables from separate connected systems – a foreign key attribute that references a remote table must be added to the referencing table.

Communication autonomy implies communication openness of connected database systems by their own accord with other remote systems not related to the integration system. An integration system will not satisfy this category if there is a standing condition by which the connected system is only allowed to communicate with the integration system.

Execution autonomy depends on the ability of connected systems to execute all tasks without outside interruptions. An integration system will not satisfy this category if it explicitly sets locks on elements of connected databases. Such behavior can cause deadlocks and disable further task execution on connected databases.

Three system behavior alternatives can be determined [5]: *tight integration*, *semi-autonomy* and *total isolation*.

The alternative of tight integration disables the connected systems' ability to autonomously control their data. User requests are sent and conducted over a single data manager that subordinates data managers in connected systems.

The alternative of semi-autonomy allows the connected systems to act autonomously, but they are still involved in a federated exchange of data. Each connected system gets to decide which part of the database is exposed to the integrated system, but the connected systems must be adapted for data exchange.

Total isolation implies that the connected systems act as stand-alone systems. Therefore, the connected systems don't know of each other's existence, nor do they know how to communicate with each other. In this case transaction management is delegated to the individual connected systems – this can make integrated control difficult.

C. Distribution

The dimension of distribution refers to the distribution of data in an integration system. According to [5] there are three abstract implementation alternatives: *undistributed system*, *client-server system* and *peer-to-peer system.*

The scope of this paper does not cover the field of undistributed systems, so they will not be considered.

The client-server alternative implies an architecture that consists of two software components: client and server. Clients contain software support for connecting to the server, while a server contains all functionalities regarding database integration. In this case servers are considered as containing two conceptual layers: the connected systems and the integration system. Considering such layering, the integration system can have multiple instances. Each integration system can then be connected to multiple different connected systems. Because of the need for alignment of various taxonomies, this paper is going to name systems of this type as *multi-database systems*. The conceptual diagram of this alternative is shown in Figure 3.

The peer-to-peer alternative (Figure 4.) doesn't strictly differentiate clients and servers, so every participant in the integration system acts as a component of the integration system. Each component of the integration system is connected to just one connected system and every connected system is connected to just one component of the integration system. Because of the need for alignment of various taxonomies, this paper is going to name systems of this type as *federated database systems*.

III. Domains

In the continuation of this paper terms local and global are used when referring to database components. On an abstract level, integration systems must take into consideration the existence of two database component domains: *local* and *global* (Figure 5.).

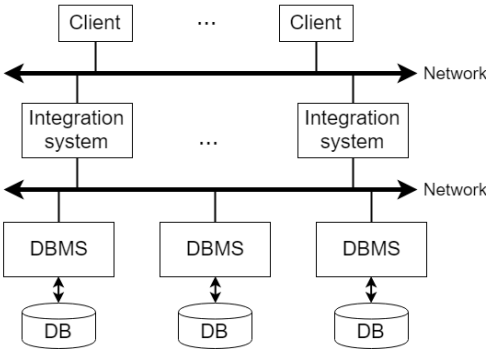


Figure 3. An example of a multi-database alternative (client-server)

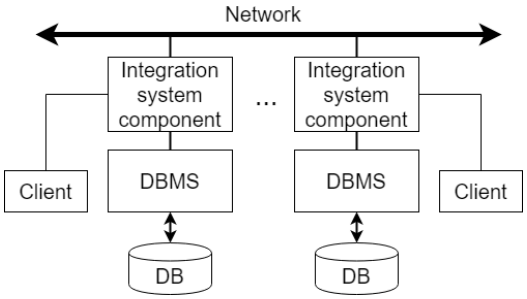


Figure 4. An example of a federated system alternative (peer-to-peer)

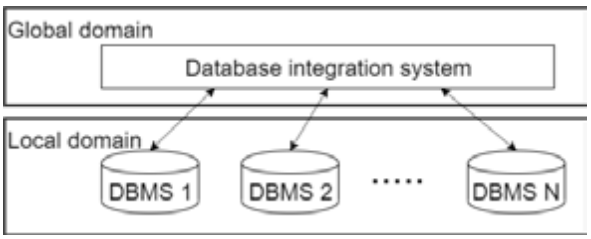


Figure 5. Domains in database integration systems

The local domain refers to components of the connected databases, while the global domain refers to components realized in the integration system. Each connected database system that is being integrated has its local schema, query language and query result form. Those local components are used for global component creation in a database integration system.

IV. System Architecture

Of the three dimensions of database integration systems, only the distribution dimension will have a significant impact on the overall system architecture. For that reason, integration systems can be divided into two groups: *multi-database systems* and *federated database systems*

The two groups are analogous with the abstract implementation alternatives obtained by breaking down the distribution dimension (Chapter II.C). It is interesting to observe the differences in internal components of these

two groups. The differences are caused by different communication and integration requirements.

There is also one more architectural approach – named *mediator-wrapper* [6] [9] [13]. This approach examines the architecture of an integration system as a division of components into two main functionalities of the system. The inner details of the components are not considered, but an interesting system componentization idea is introduced.

A. Multi-database system

A multi-database system implies the existence of a single integration system, that can be instanced multiple times to connect different local systems. The inner components of this system are referenced in different forms [2] [6] [10] [14] [15] to allow authors to better explain their set problem, but in general the ideas behind them remain the same. The architecture displayed in Figure 6. provides a general overview of the system's components and their coupling.

Users access the system exclusively through the access interface. The access interface can be implemented as web API [15], TCP server awaiting a JDBC/ODBC connection [2] or a web application.

The *global query parser* transforms the input of a user query into a form more adequate for machine processing. If a system uses a textual query defined by a language [16], [17], then it is required to perform lexical, syntactical and semantical analysis.

The *global query optimizer* at least uses predetermined heuristic rules for query optimization. Its purpose is to minimize the query cost and execution time. The optimizer works closely together with the query localizer. The *query localizer* segments the global query into multiple local queries whose merged results give the final global query result. The query localizer also conducts the procedure of translating local queries into the native query language of their source (local database).

The *query execution manager* receives the local queries and executes them over a set of connections to the local databases. To minimize the execution cost, the query execution manager should be able to execute the queries in parallel (asynchronously), as much as the query execution plan allows it. The query execution manager redirects the results, uniformed if possible, to the result globalizer.

The *result globalizer* realizes the optimal global query execution plan and finalizes the results. At the end of this procedure, the results must be uniformly formatted before forwarded to the access interface.

The *global schema manager* retrieves data about schemas of local databases and integrates them in a uniform format that is presented to the user. A globalized schema is passed to the user through the access interface. Depending on the system settings, the user might be presented with a limited view of some schema elements or even be provided with a view that omits some local data sources.

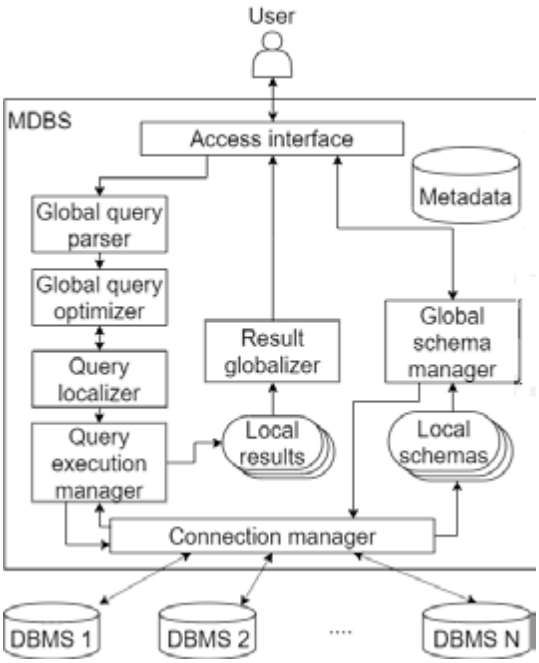


Figure 6. Components of a multi-database integration system

The *connection manager* is a component that concerns itself with the status of connections towards local databases. All system components communicate with the local databases exclusively through the connection manager.

Data concerning connection strings, connections, statuses of local databases, mappings of local data and system settings are stored in a *metadatabase*. The metadatabase is available to all system components. Although it is not shown in Figure 6., it is to be expected that all connections to the metadatabase will be achieved through a software interface within the system itself.

B. Federated database system

The federated database system has multiple instances running simultaneously. Each system instance of this architecture has a direct connection to only one local database. To achieve a complete system image, the system instances share among themselves data about their local databases. Federated database systems should follow the same ideas about system componentization as in multi-database systems. The principal difference in the two is the need for a federated system to exchange data among its instances. For valid data exchange it is required to use distributed transaction protocols that allow the creation of remote tasks and keep local databases in a consistent state [6] [13]. The conceptual architecture of a federated database system is show in Figure 7. The conceptual architecture has been adapted from the multi-database architecture (Figure 6.) but has been modified to fit a system with multiple instances according to [6] [13].

Users access the system through a *user access interface*. Ideally this interface should be the same as the access interface that would be implemented for a multi-database system.

An instance of a federated system also contains an interface for communicating with other instances – an *integration-communication interface*. System instances can exchange queries, results and metadata over this interface. This interface must be solely used by other system instances.

A federated system must have a global query parser, a global query optimizer and a query localizer. These components provide the same functionalities as in the case of a multi-database system.

Due to federated query execution in this implementation alternative, query management will also be run on remote system instances. To preserve consistency protocols like two-phase commit [6] [13] and two-phase locking) [5] can be used. One can conclude that a federated system will have to include a global and local query manager as system components.

The global query execution manager receives localized queries from the query localizer. Depending on the placement of data referenced in the query, the global query manager will send the localized queries to a local query execution manager in the same system instance or to a remote system instance over the integration-communication interface. The global query execution manager will receive remote results over the same interface or directly from the local query execution manager. The results are then forwarded to the result globalizer.

The result globalizer is similar to its counterpart in a multi-database system – it is used to merge and finalize results. In this case the difference is that the globalizer can receive results from both the local and global query execution manager. After the data has been processed, it is sent to the user through the user access interface.

The global schema manager is used to merge all the system instances' local schemas. Depending on the data distribution transparency, the user is presented with a schema as an integrated and homogeneous whole.

The metadatabase will in this case contain the locations (addresses) of remote system instances. Accordingly, it will contain system setting and metadata about global, local and remote schemas.

C. Mediator-wrapper

The *mediator-wrapper* is an architectural approach that can be implemented in database integration systems. The architecture breaks the system down into two main components: the mediator and wrapper. The *wrapper* is a software component that encapsulates the source data as an intercessor between the data source and the system middleware [19] [20]. The mediator is a software component that uses built-in knowledge of data sets to create information for upper system layers [20].

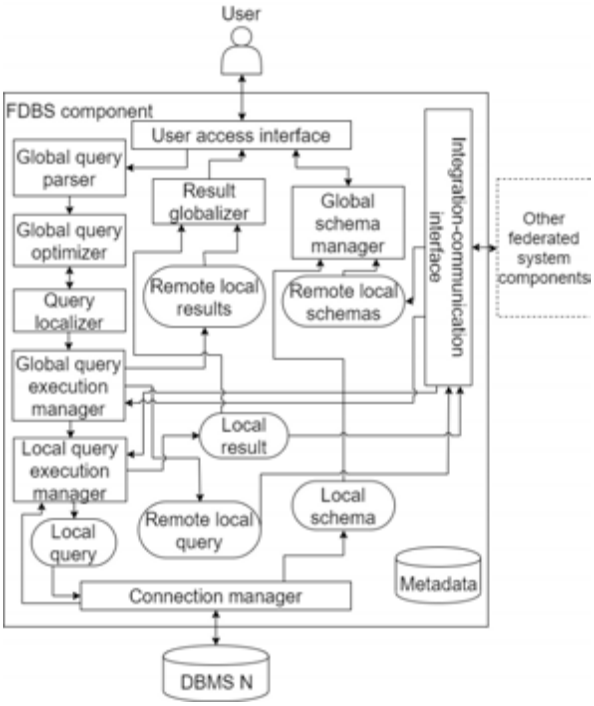


Figure 7. Components of a federated database integration system

In this architectural approach the mediator and wrapper exchange data between each other (Figure 8.). The wrapper retrieves data from its designated source and transforms it into a basic format that is intelligible by the mediator. The mediator receives data from multiple wrappers and tries to transform it into a uniform whole that is later sent to the upper layers. The mediator receives queries that it transforms and sends them on to wrappers with according data sources.

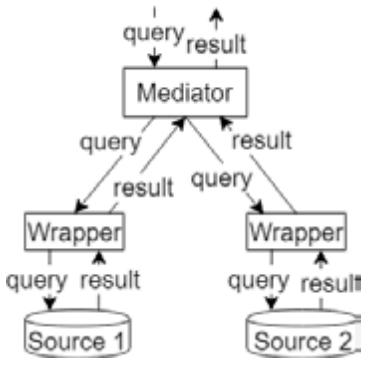


Figure 8. The relationship between components of the mediator-wrapper architecture [5]

An integration system is not limited to having just one mediator component (Figure 9). A system implementation is possible where there is a hierarchical connection between mediators. In that case, each hierarchical level can raise the data abstraction level. As an example, a system that connects multiple relational and non-relational databases can be observed. At a lower

hierarchical level, mediators can operate over paradigmatically similar data sources (mediators for graph databases, document databases, object-relational databases, relational databases, etc.). At a higher hierarchical level, the next possible categorical generalization can be implemented, e.g. mediators between relational and non-relational databases. At the highest level a mediator is used for complete data source integration, although that does not imply that only one mediator can be used. Mediators can be used for access control towards certain system components.

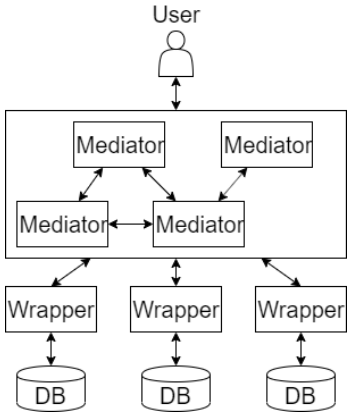


Figure 9. Mediator-wrapper architecture

V. Conclusion

Every endeavor should be founded on a set of agreed upon definitions of key field terms. In the case of database integration systems, key terms that appear in scientific sources are often overlapping and contradictory. This is the result of long-term uncoordinated scientific work, that was created in smaller work groups under the auspices of important industrial projects. In hopes of bettering the exchange of existing knowledge and future scientific breakthroughs, a revised taxonomy was introduced. In addition, the taxonomy allows extensions with terms related to integration of NoSQL systems.

The incompatibility of taxonomies has also reflected itself onto database integration system architectures. Therefore, it was important to present the architectural alternatives and approaches that the revised taxonomy entails.

The analysis of exact technologies that could be used for database integration are outside the scope of this paper. It is up to the system designer to decide whether to use an ad-hoc data integration system (e.g. Apache Spark, Apache Gobblin) or use a database integration system. While using an ad-hoc data integration system seems adequate at first glance, it can be argued that as system requirements change over time a database integration system might be more fitting – due to specialized process automatizations (e.g. query optimization).

References

[1] Z. Parker, S. Poe, and S. V. Vrbsky, "Comparing NoSQL MongoDB to an SQL DB," in *Proceedings of the 51st ACM Southeast Conference on - ACMSE '13*, Savannah, Georgia, 2013, p. 1, doi: 10.1145/2498328.2500047.

[2] R. Lawrence, "Integration and Virtualization of Relational SQL and NoSQL Systems Including MySQL and MongoDB," in *2014 International Conference on Computational Science and Computational Intelligence*, Mar. 2014, vol. 1, pp. 285–290, doi: 10.1109/CSCI.2014.56.

[3] A. Pavlo and M. Aslett, "What's Really New with NewSQL?," *SIGMOD Rec.*, vol. 45, no. 2, pp. 45–55, Sep. 2016, doi: 10.1145/3003665.3003674.

[4] M. T. Roth *et al.*, "The Garlic project," in *Proceedings of the 1996 ACM SIGMOD international conference on Management of data - SIGMOD '96*, Montreal, Quebec, Canada, 1996, p. 557, doi: 10.1145/233269.280363.

[5] M. T. Özsu and P. Valduriez, *Principles of distributed database systems*, 3rd ed. New York: Springer Science+Business Media, 2011.

[6] A. Sheth and J. Larson, "Federated Database-Systems for Managing Distributed, Heterogeneous, and Autonomous Databases," *Computing Surveys*, vol. 22, no. 3, pp. 183–236, Sep. 1990.

[7] Y. Breitbart, H. Garcia-Molina, and A. Silberschatz, "Overview of multidatabase transaction management," p. 59, 1992.

[8] S. Busse, R.-D. Kutsche, U. Leser, and H. Weber, "Federated Information Systems: Concepts, Terminology and Architectures," p. 41, 1999.

[9] C. Li and J. Gu, "An integration approach of hybrid databases based on SQL in cloud computing environment," *Software: Practice and Experience*, vol. 49, no. 3, pp. 401–422, Mar. 2019, doi: 10.1002/spe.2666.

[10] Á. Vathy-Fogarassy and T. Hugyák, "Uniform data access platform for SQL and NoSQL database systems," *Information Systems*, vol. 69, pp. 93–105, Sep. 2017, doi: 10.1016/j.is.2017.04.002.

[11] Chiang Lee and Chia-Jung Chen, "Query optimization in multidatabase systems considering schema conflicts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 9, no. 6, pp. 941–955, Dec. 1997, doi: 10.1109/69.649318.

[12] H. Garcia-Molina, J. D. Ullman, and J. Widom, "Database Systems: The Complete Book," p. 570.

[13] Y. Sun, J. Liu, L. Yang, X. Geng, and X. Cao, "A New Heterogeneous Database Integration Model Based on Multi-agent," in *2012 Fourth International Conference on Computational and Information Sciences*, Chongqing, China, Aug. 2012, pp. 421–424, doi: 10.1109/ICCIS.2012.25.

[14] G. Thomas *et al.*, "Heterogeneous Distributed Database-Systems for Production Use," *Computing Surveys*, vol. 22, no. 3, pp. 237–266, Sep. 1990.

[15] P. Atzeni, F. Bugiotti, and L. Rossi, "Uniform access to NoSQL systems," *Information Systems*, vol. 43, pp. 117–133, Jul. 2014, doi: 10.1016/j.is.2013.05.002.

[16] H. Valer, C. Sauer, and T. Härder, "XQuery processing over NoSQL stores," p. 6, 2013.

[17] K. W. Ong, Y. Papakonstantinou, and R. Vernoux, "The SQL++ Unifying Semi-structured Query Language, and an Expressiveness Benchmark of SQL-on-Hadoop, NoSQL and NewSQL Databases," p. 35, 2014.

[18] M. T. Roth, "Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources," 1997, pp. 266–275.

[19] Y. Papakonstantinou, H. Garcia-Molina, and J. Widom, "Object exchange across heterogeneous information sources," in *Proceedings of the Eleventh International Conference on Data Engineering*, Mar. 1995, pp. 251–260, doi: 10.1109/ICDE.1995.380386.

[20] G. Wiederhold, "Mediators in the architecture of future information systems," *Computer*, vol. 25, no. 3, pp. 38–49, Mar. 1992, doi: 10.1109/2.121508.

Microservice Performance Degradation Correlation

Marko Samardžić, Romeo Šajina, Nikola Tanković
Faculty of Informatics
Juraj Dobrila University of Pula
Pula, Croatia
{msamardzi, romeo.sajina, nikola.tankovic}@unipu.hr

Tihana Galinac Grbac
Juraj Dobrila University of Pula
Department of Engineering
Pula, Croatia
tgalinac@unipu.hr

***Abstract*—Microservice architecture offers many options to develop, deploy, and maintain applications using independent cloud services. Such a model is already widely accepted. However, the challenge of getting insights regarding the performance relationship between such services remains open. This paper introduces an approach to the investigation of performance degradation in microservice architectures. The algorithm inspects the correlation on multiple time series in order to make useful insights into the relationship between services as well as user behavior. In this research, we analyze collected run-time logs from six microservices in the retail domain used by several thousand SME customers.**

***Index Terms*—Microservice architecture analysis, Time series analysis, Microservice performance degradation, Time series anomaly detection, Time series correlation**

I. Introduction

Microservices architecture is an approach for developing a single application as a suite of small, autonomous services, which are easy to understand, deploy, and scale [3]. Despite the rapid growth of microservice architecture in the industry, there is still a lack of formal, systematic models that will evaluate the performance and their behavior. With architecture like these, it is often a challenge to develop a model that will intuitively describe user behavior of that system as well as how each service affects the other.

A web architect needs to assess the causal relationships between different services, but that is not an easy task. This relationship can emerge from the following patterns:

1) Service A directly invokes service B synchronously
2) Service A directly invokes service B asynchronously
3) A third party service (e.g., from frontend or external systems) bases its invocation on Service B by the results from service A, or sequentially regardless of results
4) An alive agent behavioral pattern or a certain use case determines the invocation of Service B from the results on Service A

Designing an accurate invocation or dependency model can be challenging, so we are basing this research on model-agnostic premises. Previous studies are based on postulating complex workload models which often consists of speculations about how user interact with the system. Besides, the model usually cannot take into account some unknown and unusual patterns of user interaction. The cloud environment often makes setting more complex to monitor. With additional rules regarding the elasticity and scaling in particular. If the predictive model cannot appropriately address these rules, it is often not useful, as we can see for their weak application in industry. Notably so in SME environment.

The significant amount of available run-time performance data demands fast online algorithm based on time series data. Our approach examines the relationship between multiple microservices from available workload data. Eventual correlation results could be useful for a Software architect to gain the intuition into the causal relationship between service performance and user behavior.

Our research offers several contributions:

1) a model-agnostic algorithm for detecting the correlation between performance degradation events in a microservice architecture
2) gave that the algorithm is non-intrusive and acts only on monitoring data, it can be applied to a wide range of different environments and architectures: from simple RPC services to fully-fledged microservice architecture and streaming analytics
3) it offers an easy to understand intuitive explanation on the correlation between degradations so one can easily interpret the results and assess further possible causations

The paper proceeds as follows: Section 2 presents the related research and background, where Section 3 outlines our approach. Section 4 presents the results of our method, together with the limitations. Section 5 concludes the paper.

II. Related work

The growing popularity of microservice architecture initiated the need for significant research effort to investigate the relationship between multiple services. Especially in perspective of root cause errors. Anomaly detection is proposed in [7] by examining monitoring data of individual components with a thresholding scheme. In [1], the RanCorr is presented, which computes anomaly ratings that are fully automated and consider contexts such as operation, component, and deployment, whereas we offer possible correlated relationships primarily for better insights to software architect. Some more recent studies, such as [8], introduce a FacGraph framework for anomaly investigation and root cause identification — the mining algorithm processed on the anomaly correlation graph that requires no domain knowledge. In [2], MonitorRank is proposed as an algorithm that finds the root causes of

anomalies in service-oriented architectures. When an event is detected, the MonitorRank provides a ranked order list of possible root causes for the monitor team to investigate. However, to investigate such a relationship, MonitorRank needs to carefully examine a graph for computation, which is often expensive in cloud environments.

Some studies have looked more of the intentions and strategies of migrating the monolith applications to microservice as well as comparing these two architectures in different circumstances. A qualitative study in [9] focuses on intentions, strategies, and challenges in the context of migrations to microservice architecture. The study conducted 16 in-depth interviews with software professionals from 10 companies. A comparison of the monolithic and microservices architecture has been made in [5]. The study warns about the negative impact on performance when moving to microservice. In [6], the infrastructure cost analysis is given for monolith and microservice architecture. The authors concluded that operational cost could be significantly lower with microservice architecture compared to monolithic.

III. Time series online approach

A. Dataset

For this work, we collected run-time logs from six microservices in the retail domain used by several thousand SME customers. We published the dataset at https://github.com/samardzicmarko/MicroservicePerformanceAnalysis.

Each dataset represents the timestamp of service invocation, together with the overall service time. Our datasets are synchronized and contain requests from 45 days of operation. Data is aggregated to one-minute frequency, which in our case, results with around 61000 data points.

Six services that were monitored for this study were:

1) **SaveInvoice** - this web service handles the user input in constructing and storing the retail invoices
2) **Sync** - this web service is handling the synchronization between user device data storage and online cloud databases
3) **Balance** - this web service computes the current cash-register balance based on the daily issued set of invoices
4) **Check** - a simple web service to assess the availability of stable connection with the cloud - this was included as a control web service, it should not show significant correlation with other web services
5) **Z-List** - a list of all the invoice headers to be included in the final daily report for a certain cash-register
6) **DataQuery** - a service to provide additional retail resources (like items and stock data) that are not cached offline

In table I we can see the mean service time value for each service as well as its standard deviation. Once again we observe service time for each incoming request.

Time series	Mean	Standard deviation
Balance	0.600892	0.660613
Check	0.316914	0.767031
Data Query	0.120254	0.858606
Save invoice	1.901311	1.714266
Sync	0.130768	0.443705
Z-list	2.125471	2.760975

TABLE I
DESCRIPTIVE STATSITIC FOR OBSERVED SERVICE TIME

B. Event detection

To investigate behavior for given time series, it is crucial to detect events that are interesting for our use case. We want to capture how variable service times are through time with the frequency of every minute. More specifically, events of our interest are the ones that we can think of as anomalies. For that purpose, we use Peak detection algorithm using smoothed z-scores [4]. The algorithm takes time series as input and has three parameters that need to be defined:

1) **Lag** - indicate the number of taken observations in the moving window. It determines how much data will be smoothed over.
2) **Threshold** - used to determine whether a data point is an anomaly or not. It is defined as the number of standard deviations away from the moving mean.
3) **Influence** - gives or ignores past calculations of signals on the algorithm detection threshold. If Influence is set to 0, past signals do not influence future detection, and threshold detection is calculated independently.

Each of the listed parameters needs to be addressed and tuned before running the algorithm. Keep in mind that our algorithm gives us the binary result; 1 if the data point is anomaly or 0 otherwise. To get the actual value of detected signals, we multiply each point in time series with its value so that we get the amplitude of the signal, finally producing $Z(ts)$. It is important to be aware that the z-score algorithm can return negative values that need to be addressed. The result of detected events is shown in the figure 2.

C. Time Series dependency analysis

Our proposed approach calculates correlation for two time series $Z(ts_1)$ and $Z(ts_2)$ at a time. Algorithm iterates over the positive values of $Z(ts_1)$. For each step t_i of iteration it captures values from $Z(ts_2)$ within the defined window. To define a window we need to introduce a parameter *window-size* (ws) that will determine a look in the future interval, i.e. how long after the event are we looking for an effect. Values captured from $Z(ts_2)$ will involve points $[t_i, t_{i+ws}]$ which are summed to produce a single value v_i. Finally we calculate the correlation with the folowing formula:

$$corr_{(ts1,ts2)} = \frac{\sum_{i=1}^{i} v_i}{\frac{i}{ws}} \quad (1)$$

where i is the number of iterations.

We repeat the same algorithm for all combinations of time series, which results in a causation matrix. An illustrative example of a calculated relationships is shown in figure 3.

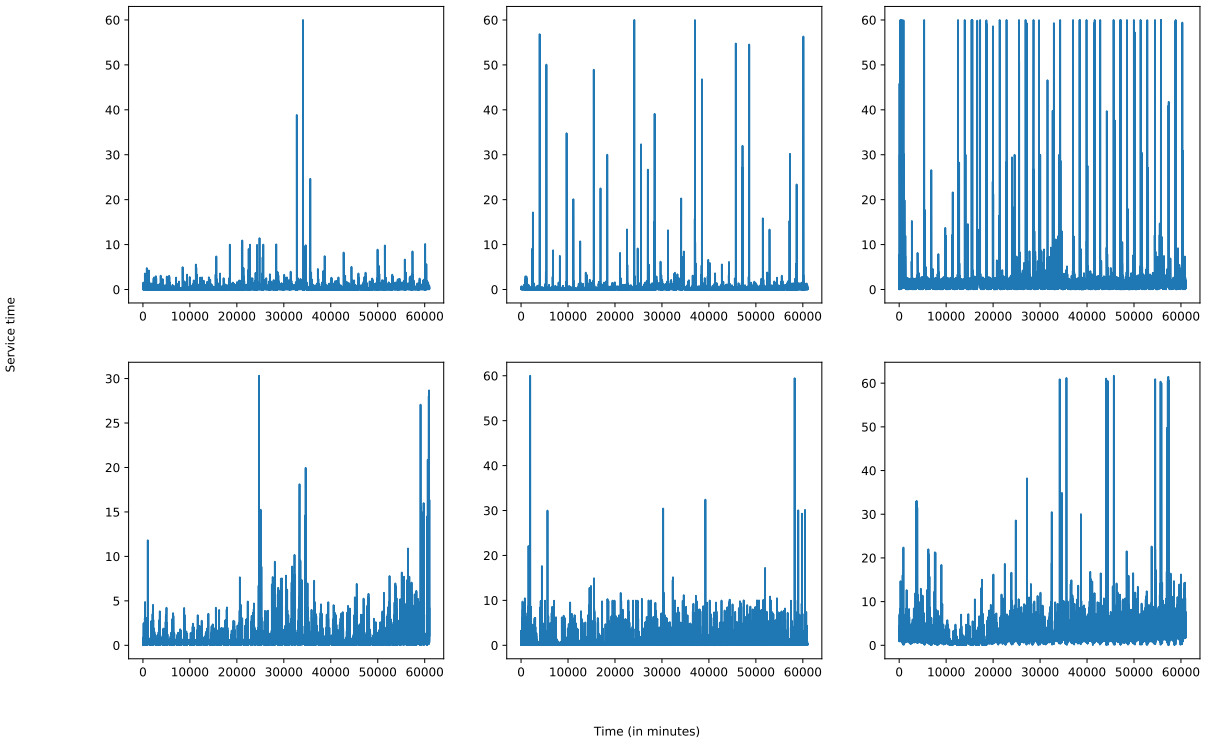


Fig. 1. Visualization of each time series

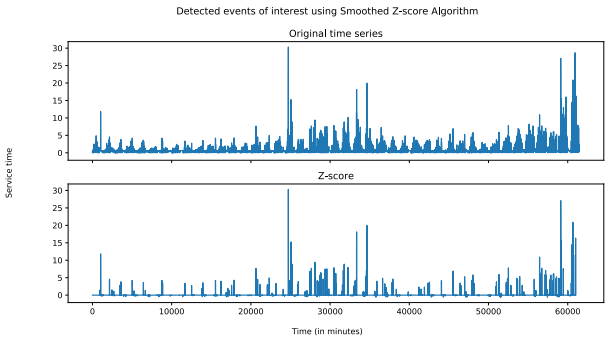


Fig. 2. Calculated Z-score algorithm on Balance dataset

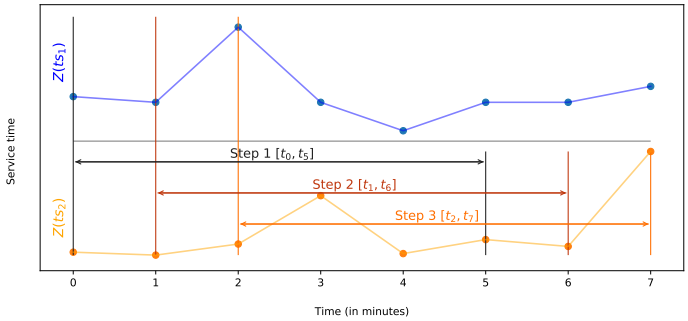


Fig. 3. Illustrative first few steps of correlation calculation algorithm. We iterate through positive values of $Z(ts_1)$ and capture corresponding values from $Z(ts_2)$ within the defined window.

IV. Results

Our algorithm generates the causation matrix for previously mentioned six observed services. It is important to be aware of the course in which we think about the dependency. In the following table, services placed in rows are affecting the anomalies of the ones in columns, as we can recall from the section III-C that there is a leader-follower relationship in the algorithm.

Results are represented with the causation matrix between each service which shows degradation effects across services.

Services	Balance	Check	Data Query	Save invoice	Sync	Z-list
Balance	0.053906	0.000816	0.000805	0.005503	0.000835	**0.011672**
Check	0.001015	0.027837	0.006520	0.003100	0.001127	0.007419
Data Query	0.001983	0.002087	0.023623	0.005603	0.002046	0.007638
Save invoice	0.001685	0.003471	0.000936	0.062899	0.001869	0.005490
Sync	0.001061	0.002732	0.004775	0.005698	0.011287	0.005326
Z-list	0.000690	0.000978	0.000782	0.001433	0.000590	0.083756

TABLE II
CALCULATED CAUSATION MATRIX BETWEEN SERIVCES

Intuitively we can see that one value stands out in the table for what we can think of as a significant. The Balance and Zlist services are showing a some kind of causal effect with anomaly events across the time window with the result of 0.011672. The significance of observed relation can be confirmed by evaluating its distance from the mean value. Before the calculation of the mean and standard deviation, we will remove the diagonal values, which results in the mean value of 0.0028 and a standard deviation of 0.0024. All but one values are contained within two standard deviations, while before mentioned relation between Balance and Zlist services shows a correlation that exceeds the third standard deviation.

A. Limitations and future work

Parameter optimization is worth the attention if we want to extend the presented approach. As we explained through the process of event detection, we define parameters such as threshold, influence, and lag. Further improvements of the Smoothed Z-score algorithm can be made with tuning those parameters based on given time series and introduction of the new parameters for the purpose of a better representation of events.

The window size can be further examined. Given the nature of microservices and their dynamic environment, the window size sometimes needs to be shorter. To extend our approach, we could use a weighted decay function. The different time windows for different microservices could also be considered for some systems. To aim for more concrete and insightful results, knowledge about the architecture that we are observing could be crucial. As mentioned in section I, there are several types of relationships between services, and not all can be treated equally. Our presented approach should be tested with some other algorithm besides Smoothed Z-score.

V. Conclusions

In this paper, we presented an approach for analyzing the performance degradation between multiple microservices. Two important steps stand out in our approach.

First is the extraction of relevant events that happened on multiple time series representing each service. Then follows the second step. On the extracted data, we iterate trough multiple causall combinations to examine how one service affects the other. The relationship is examinated for two time-series, with the two way leader-follower relationship. The causation matrix shows degradation effects between services. Based on the results, we can examine a possible relationship between services with a possible causal relationship in terms of performance degradation. In our case, we detected the relationship between Balance and Zlist services. Information like these can be useful to Software engineers for analysis of microservice architecture.

Acknowledgements

This work has been fully supported by Croatian Science Foundation under the project IP-2019-04-4216.

References

[1] Nina Marwede et al. “Automatic failure diagnosis support in distributed large-scale software systems based on timing behavior anomaly correlation”. In: *2009 13th European Conference on Software Maintenance and Reengineering*. IEEE. 2009, pp. 47–58.

[2] Myunghwan Kim, Roshan Sumbaly, and Sam Shah. “Root cause detection in a service-oriented architecture”. In: *ACM SIGMETRICS Performance Evaluation Review* 41.1 (2013), pp. 93–104.

[3] Martin Fowler and James Lewis. “Microservices a definition of this new architectural term”. In: *URL: http://martinfowler. com/articles/microservices. html* (2014), p. 22.

[4] J.P van Brakel. “Smoothed z-score algorithm”. In: *https://stackoverflow.com/questions/22583391/peak-signal-detection-in-realtime-timeseries-data* (2016).

[5] Takanori Ueda, Takuya Nakaike, and Moriyoshi Ohara. “Workload characterization for microservices”. In: *2016 IEEE international symposium on workload characterization (IISWC)*. IEEE. 2016, pp. 1–10.

[6] Mario Villamizar et al. “Infrastructure cost comparison of running web applications in the cloud using AWS lambda and monolithic and microservice architectures”. In: *2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*. IEEE. 2016, pp. 179–182.

[7] Janos Gertler. *Fault detection and diagnosis in engineering systems*. Routledge, 2017.

[8] Weilan Lin et al. “FacGraph: Frequent Anomaly Correlation Graph Mining for Root Cause Diagnose in Micro-Service Architecture”. In: *2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC)*. IEEE. 2018, pp. 1–8.

[9] Jonas Fritzsch et al. “Microservices Migration in Industry: Intentions, Strategies, and Challenges”. In: *2019 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. IEEE. 2019, pp. 481–490.

Robust and Scalable Online Code Execution System

Herman Zvonimir Došilović
University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
hermanz.dosilovic@gmail.com

Igor Mekterović
University of Zagreb
Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
igor.mekterovic@fer.hr

***Abstract*—In this paper, we present a novel, robust, scalable, and open-source online code execution system called Judge0. It features a modern modular architecture that can be deployed over an arbitrary number of computers and operating systems. We study its design, comment on the various challenges that arise in building such systems, compare it with other available online code execution systems and online judge systems, and finally comment on several scenarios how it can be used to build a wide range of applications varying from competitive programming platforms, educational and recruitment platforms, to online code editors. Though first presented now, Judge0 is in active use since October 2017 and has become a crucial part of several production systems.**

***Index Terms*—online code execution system, online judge system, untrusted code execution**

I. Introduction

Programming skills have been recognized as a core competency skill [1] and the demand for the educational services in this field surpasses the supply. Educational organizations that teach programming can extend the classical human code assessment of students' assignments with an automated assessment that relies on online judges (OJs) [2]. Online judges are systems designed for evaluation of the user's source code on a predefined set of test cases. There are many different categories of online judges which can be classified by use-case varying from competitive programming platforms, e-learning systems, recruitment platforms, etc. Studying the architecture design of many different online judges surveyed in [2] we have identified that their mutual component is a code execution engine (CEE) that uses some sandboxing mechanism for secure execution of the user's source code. Extracting a code execution engine into a standalone component and providing a feature-rich web API for it we were able to create a robust, scalable, and configurable online code execution system (OCES) that can be integrated into many different types of online judges. In section II we explore and define components of the OJ ecosystem. After that in section III we define OCES through functional and non-functional requirements and in section IV we study technical challenges in building an OCES and present modern architecture and execution pipeline of Judge0. Finally, in section V we give a quick overview of related work done in the field of online code execution systems.

II. Online Judge Ecosystem

Figure 1 shows the layered architecture of an online judge ecosystem. This ecosystem includes components commonly associated with OJ system such as platforms built on top of online judges and online compilers, as well as less addressed components that online judges rely on - sandboxes and code execution systems. Decomposing an online judge into separated layers allows us to better understand and define functional and non-functional requirements of each layer. In the remainder of this chapter, we define and comment on each layer to clarify how these layers fit together and how an online code execution system and online judges, in general, are built.

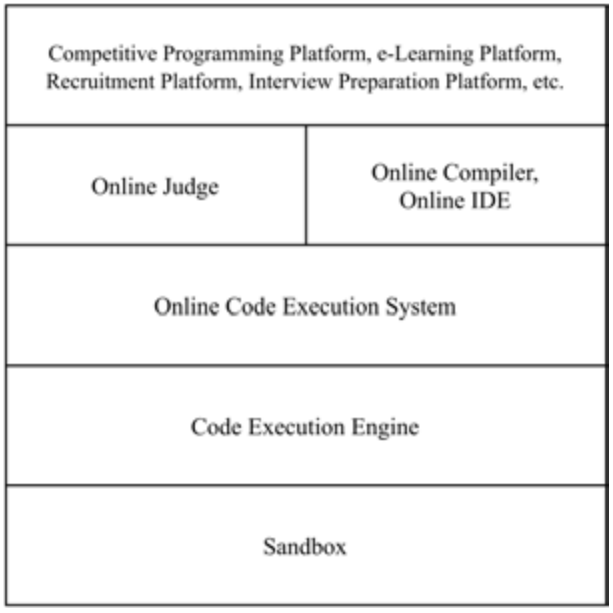


Fig. 1: OJ ecosystem architecture.

A. Sandbox

Sandbox is a security mechanism for separating running programs, often used to execute untrusted source code [3]. In the context of online judges, sandboxes are used for security measures and to enforce limits on resources [4] like memory or CPU. In competitive programming, every challenge has defined time and memory limits which should not be exceeded. Thus, with good test cases and resource limits, authors of

a challenge can check if the user solved a challenge with appropriate time and memory complexity. Many sandboxing techniques can be used for isolating untrusted source code with custom resource limitations such as LXC containers [5], Docker [6], virtual machines (VMs) [7] and others [3], [4]. In the context of online judges we expect that with a sandbox we can (1) initialize sandbox environment, (2) copy users code and other files into the sandbox, (3) compile and execute users code with custom resource limitations, (4) collect the output of a program and execution metadata, and (5) clean-up the sandbox environment.

B. Code Execution Engine

Code execution engine (CEE) is a component of an online judge that uses a sandbox to (1) compile and run arbitrary source code inside a sandbox and (2) to parse various metadata generated after compilation and execution. Building an online judge implies building a CEE, however, CEEs are rarely commented in the literature on online judges.

C. Online Code Execution System

Online code execution systems (OCESs) are (typically distributed) systems that provide a web API for compiling and executing arbitrary source code. They use CEE as a service for compiling and executing the given source code. Figure 2 shows the architecture of OCES. Up until now, OCESs have not received much attention in the literature in the context of online judges because online judges were developed for specific use-case for which OCESs were hidden inside the online judge architecture.

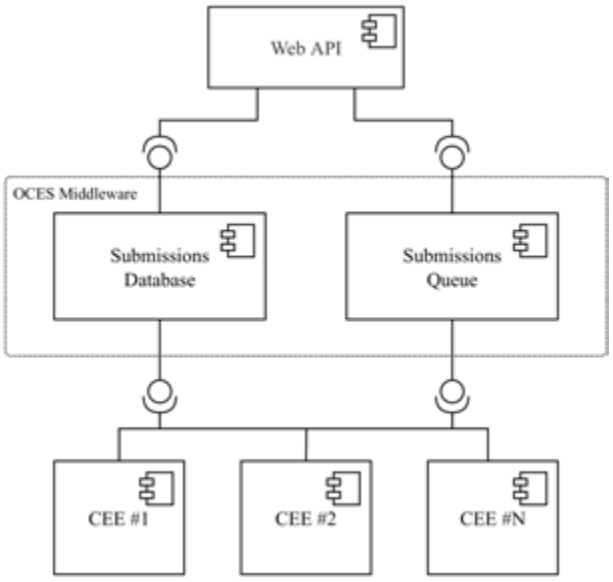


Fig. 2: Online code execution system architecture.

D. Online Compilers and IDEs

Online compilers (sometimes referred to as online code editors) like [8]–[11] are online platforms where users can write, compile and execute their source code in one or more programming languages. They usually allow users to write a single-file program, save their snippets, and share it with other users. Online IDEs like [12]–[14] are more advanced online platforms that allow users to develop bigger projects in a cloud. Some online platforms described in II-F have integrated online code editors so that users can write solutions in the browser without the need to install compiler and code editor on their computer. Detailed comparison of various online compilers and IDEs can be found in [2].

E. Online Judges

The definition of an online judge (OJ) given by [2] states that an online judge is an online service that performs either submission, assessment, or scoring in a cloud. In the submission phase, the user's code is compiled (if needed) and it is verified that the resulting binary can be successfully executed. In the assessment phase, the resulting binary is run on a given set of test cases and for each of the test cases it is verified that (1) the execution process completed without errors (2) resource limits have not been exceeded and (3) the obtained output complies to the rules described in the problem definition. In the last, scoring phase, the user's score is computed based on the results from the assessment phase. Following the above definition, figure 1 shows the architecture of an online judge and various online platforms that can be built on top of it. Online judges can be realized in many different ways that depend on the use-case, but the common functionality is problem management, assessment, and scoring of a user's solution.

F. Online Platforms as Online Judges

Different categories of online judges classified by use-case are described in [2]. Their common goal is the assessment of the user's solution of a given problem, but the user's score is subsequently used for different purposes. The most common types of online judges are competitive programming platforms such as [15]–[17] for organizing programming contests or for solving various programming challenges for practice. These challenges have well-defined input and output format and expect from users to implement the appropriate algorithm using appropriate data structures without exceeding given resource limitations such as time and memory limits. E-learning platforms can be also considered as online judges if they automatically assess students' solutions on a given programming challenge. They are often used in educational institutions and in the form of Massive Open Online Courses (MOOC) [2]. Interview preparation platforms such as [18]–[20] are another type of online judges that users can use to practice challenges, algorithms and data structures that appear in interview questions for software engineering positions. Recruitment platforms such as [20] are online judges that many companies use in their recruitment process to assess candidate skills with various programming challenges.

III. Online Code Execution System

In this paper, we propose that OCES should be distinguished from the OJs and treated a standalone entity - a "black-box"

component that can be used to build various applications atop of, OJ being one of them. OCES can be defined through functional and non-functional requirements.

A. Functional Requirements

Online code execution system **must** provide:

- a well-documented web API,
- support for compiling and running arbitrary source code for at least one programming language,
- sandboxed execution, with the predefined default resource limits of compiled binaries,
- support for specifying arbitrary input that will be given as standard input to the program and
- support for returning execution results that include at least the content of the standard output.

Additionally, online code execution system **could** provide support for:

- specifying custom compiler flags,
- specifying custom command-line arguments,
- specifying resource limits such as CPU time limit and memory limit,
- initializing a sandbox environment with multiple files other than source code,
- returning detailed execution results such as CPU time and memory usage, exit signal and content of a standard error,
- returning compilation warnings or errors if they exist,
- batched submissions,
- multiple inputs and
- authenticated requests

Of course, the more functional requirements it satisfies the more use-cases it can cover.

B. Non-Functional Requirements

Online code execution system **should** be:

- scalable,
- configurable,
- secure,
- easily deployable,
- efficient,
- resilient,
- extendable and
- robust.

The more non-functional requirements OCES satisfies it is easier for it to be integrated for specific use-case platforms.

IV. Overview of Judge0

Judge0 is a new and open-source [21] online code execution system that features scalable architecture and a new approach in building online judges (discussed in V). Development of Judge0 started in August 2016 with the goal of creating a new, open-source, robust, scalable, easy to use, and extendable OCES that can be easily integrated into various web applications. Since the academic year 2018./2019. Judge0 has been used by more than 3000 students from the University of Zagreb Faculty of Electrical Engineering and Computing (FER) as an OCES integrated into e-learning platform called Edgar [22]. It has now become an indispensable tool for several courses dealing with a large number of students, today, and in the foreseeable future.

Since October 2017 Judge0 has been publicly [23] available as a free OCES and till June 2020 it has executed more than 8 million programs submitted from all around the world. Most of Judge0 users integrate it into the online judge for specific use-case like e-learning or interview preparation platform, while others integrate it into online compilers like Judge0 IDE [8].

A. Requirements and Technical Challenges

There were many technical challenges that were considered and overcome while developing Judge0 and we study each in the following sections.

1) Scalability: Scalability of Judge0 can be achieved using three strategies. The first one is by scaling a number of CEEs run on a single computer. The number of CEEs determines the number of parallel submissions that can be executed simultaneously. Empirical research shows that the maximum number of CEEs that can be run depends on the computer resources and compilers/interpreters used for running submissions. However, determining the optimal number of CEEs with regards to the computer resources and compiler/interpreter type is still an open research question for us. Current empirical results show that the good choice for the number of CEEs run on a single computer is the same as the number of logical CPU cores. This kind of scaling, by increasing the resources and number of CEEs on a single machine, corresponds to the vertical scaling and is sure to hit a ceiling at a certain point. That is where the second, horizontal scaling strategy can be employed. The second scaling strategy assumes that all components of OCES shown in figure 2 are run on the same computer. Replicating OCES over many computers and abstracting them behind a load balancer achieves horizontal scaling. Note that the second scaling strategy can be used in the combination with the first one, i.e. first determine and run the maximum number of CEEs and then horizontally scale to an arbitrary number of nodes. This type of scaling is used at FER and has shown good results having done more than 210,000 submissions of C and Java code for more than 2000 students in the academic year 2019./2020. alone. The third scaling strategy is more complex but uses computer resources more sparingly. Since Judge0 features modular architecture where each module can be easily deployed as a Docker container, the idea is to scale only parts of the system that might present a bottleneck. For instance, by horizontally scaling just CEEs with web API and middleware being run on the same or different computers. This type of scaling is used for the public instance of Judge0 [23] where web API and middleware are run on one server and CEEs are horizontally scaled (in addition to the first strategy) on two servers. This type of scaling with servers having just 4GB of RAM and 2 vCPUs was able to run, at its peak, around 120,000 submissions per day.

2) Security and Sandboxing: Because Judge0 allows users to run arbitrary source code and binaries on a remote server,

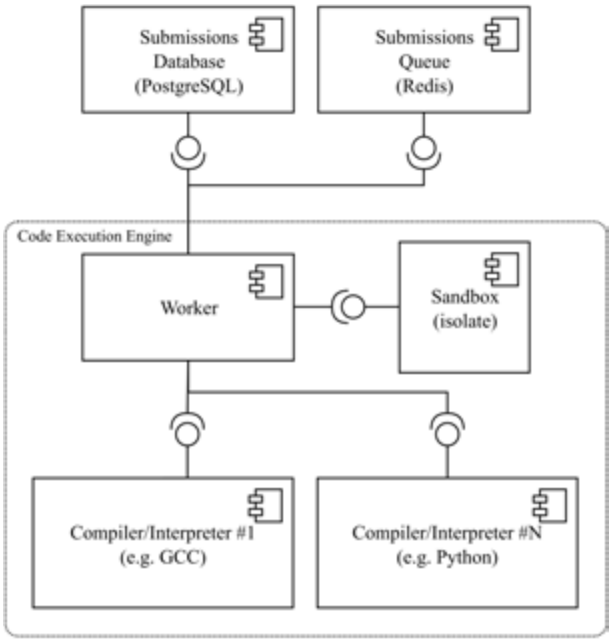


Fig. 3: Code execution engine architecture.

the attack surface is big and source code or binary for every submission is considered **untrusted** i.e. it must be compiled and run in a sandboxed environment. CEE of Judge0 uses *isolate* [4] (figure 3) for sandboxing compilation and execution. *Isolate* is a well-known and proven sandbox mechanism for popular competitive programming online judge called *CMS* [24], [25] which has been used on large competitive programming contests such as International Olympiad in Informatics (IOI) and Central European Olympiad in Informatics (CEOI). Another aspect of security is connected with wide configuration options for execution such as, for example, setting a custom maximum CPU time limit for the submission. Judge0 features rich configuration options (mostly inherited from *isolate*) for limiting resources (e.g. CPU) by setting the upper and lower bounds which enables Judge0 to support many different use-cases. Note that both compilation and execution should be sandboxed and resource-limited. To better understand this let us examine two code examples given with listings 1 and 2. Listing 1 shows a valid program written in C programming language that includes Linux pseudo-random number device driver which generates an infinite sequence of secure bits [26], resulting in infinite compilation time.

```
#include </dev/random>
int main() {
    return 0;
}
```

Listing 1: Valid C code with infinite compilation time.

The second example (listing 2) shows a denial of a service attack called *fork bomb* which generates many processes rapidly, thus exhausting the resources of the computer it is running on. As a result, any legitimate process can not start its tasks because the malicious processes are exhausting system resources [27].

```
#include <unistd.h>
int main() {
    while(1) {
        fork();
    }
    return 0;
}
```

Listing 2: Fork bomb in C programming language.

By using *isolate* as its sandbox environment Judge0 can safely compile and execute arbitrary source code and even binaries that have come from an unknown source. Moreover, parsing execution metadata from *isolate* CEE of Judge0 can save and report the status of every submission and the potential failure reason (for e.g. because of time limit exceeded, memory limit exceeded and etc.).

3) Robustness: Judge0 is designed to withstand any kind of malicious attempts with the following principle in mind: (malicious) submissions may fail, but the system will continue to function intact. This robustness in Judge0 is concerned with two aspects: (1) robustness to recovering from internal errors on program execution and (2) robustness to sending and receiving arbitrary data. The first one, robustness to recovering from internal errors, is achieved with separation of concerns in the CEE component. Malicious or faulty code will most likely cause an internal error in on of the following steps: (i) initializing the sandbox environment, (ii) compiling and executing source code in the sandbox, and (iii) clean-up of the sandbox environment. If CEE catches an undefined or unexpected behavior in any of the steps it will mark that submission with the status "Internal Error" accompanied with a status message - a reason for the internal error. The second aspect of robustness can be demonstrated with the program shown in listing 3 that outputs an invalid UTF-8 byte sequence. Even though this is a valid C program and it will run without any errors, the result of this program i.e. content of its standard output cannot be sent (serialized) as a raw sequence of bytes in a JSON format. Judge0 allows users to ask for the Base64 encoded data [28] in the JSON format. Respectively, Judge0 allows users to send Base64 encoded data in JSON format which allows them to avoid such problems.

```
#include <stdio.h>
int main() {
    printf("\xFE");
    return 0;
}
```

Listing 3: Program that outputs content that cannot be sent in a raw format with JSON.

4) Ease of use and deployment: Judge0 features simple JSON web API that has two main endpoints:

- create a new submission for one of the available languages and

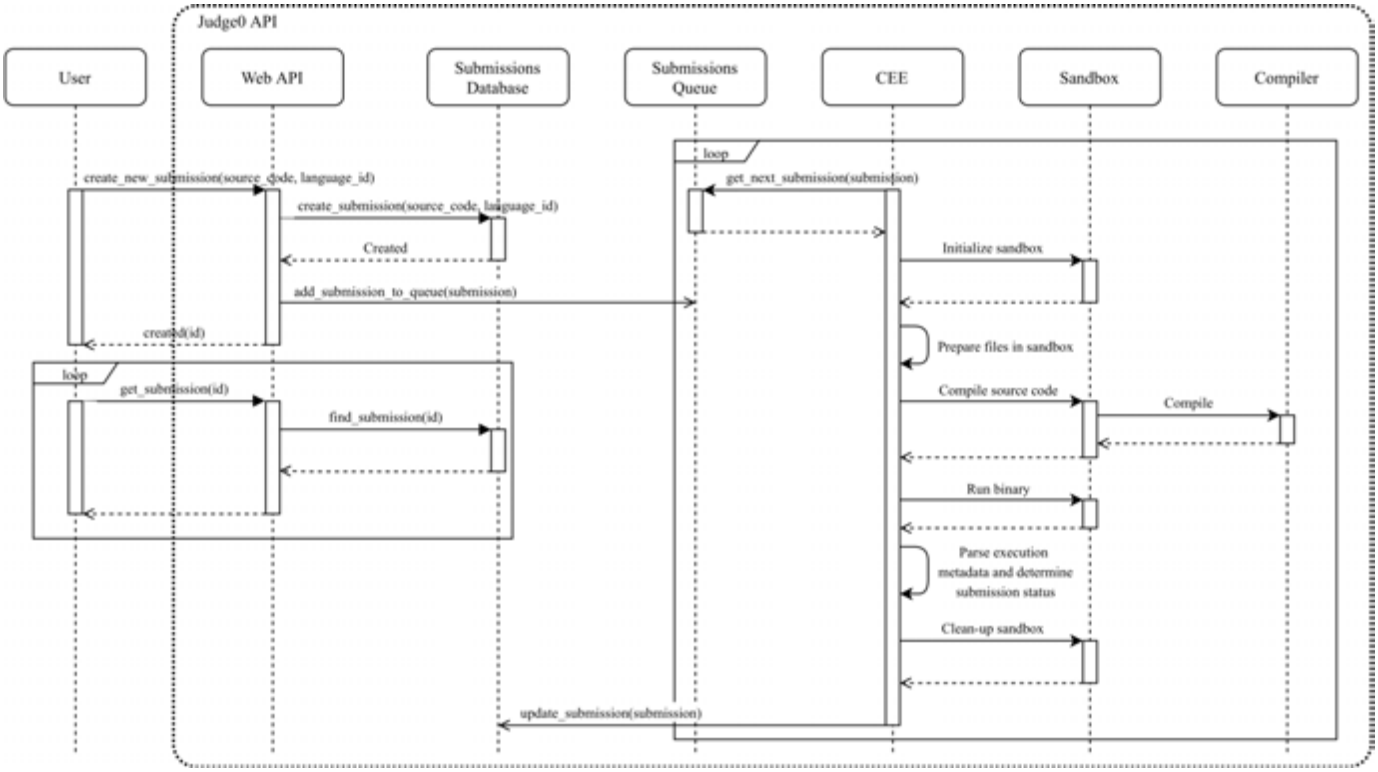


Fig. 4: Judge0 code execution sequence diagram.

- fetch submission status.

The first endpoint creates a new submission that must contain (1) source code and (2) unique identifier of a compiler or interpreter that needs to be used for compiling/interpreting given source code. Current version v1.9.0 of Judge0 supports 46 different compilers and interpreters but is not limited to those out-of-the-box languages: adding a new language/-compiler boils down to installing the compiler on the CEE machine and inserting one record (describing the newly added language) into the database. With the second endpoint, the user can fetch submission results i.e. get detailed information about execution results such as the content of standard output and standard error, CPU time, memory usage, etc. With just these two endpoints one can build simple e-learning online judge or online compiler like Judge0 IDE or [11].

Ease of deployment is achieved with Docker [6] that allows users to deploy Judge0 on any platform that has Docker Engine installed. Furthermore, because Judge0 depends on external services such as database (PostgreSQL) and submission queue (Redis) (figure 3) we use Docker Compose [29] for connecting these services which can then all be installed and run with a single command.

B. Architecture and Execution Pipeline

Judge0 features scalable architecture design as seen in figures 2 and 3. In-depth interaction between actors of the system can be seen in figure 4. First, the user creates a request to web API for creating a new submission providing at least source code and desired language. Web API then creates a new submission record in the database and pushes newly created submission to the submission queue. After pushing it to the queue, web API returns the submission unique identifier to the user. With the returned unique identifier user can request for submission results. Meanwhile, CEE continuously tries to pull the next submission from the queue. Upon successful pull it will initialize sandbox, prepare files in the sandbox, compile source code, run compiled binary, parse metadata after the execution and determine submission status, clean-up the sandbox environment and finally update submission in the database. User is expected to frequently poll submission from the web API until it receives the completed submission. We characterize this type of interaction between the user and web API as *async execution* because the user does not know when the submission is complete and, must *asynchronously* pull the submission from the web API. Judge0 also supports another type of interaction - *sync execution*, where the user gets submission result as a response instead of the submission unique identifier. Current empirical results show that *async execution* scales better when the incoming number of submissions is much higher than the number of submissions that can be executed by CEEs in the same time window. On the other hand, sync execution is a simpler paradigm for the client programs: a single request-response is easier to implement.

V. Related Work

A detailed survey on online judge systems and their applications has been done in [2] but not much attention has been given to online code execution systems because online judges

are typically developed for specific use-case for which OCESs are hidden inside the online judge architecture. In our research, we found two OCES that were developed with the competitive programming use-case in the main focus. The first one is *Sphere Engine* [30] which is a commercial and closed-source product that provides limited configuration options for each submission, whereas Judge0 exposes almost all sandboxing options *isolate* provides. Another OCES is *camisole* [31] that also uses *isolate* as a sandbox environment. It is open-source and free. However, they do not expose all configuration options that *isolate* provides, but only basic ones that are important for competitive programming. Deployment of *camisole* is done either by downloading a prepared VM image or by manual install. This type of deployment procedure is not easily scalable as the one we propose with Judge0. In conclusion, both of these systems are biased towards competitive programming while Judge0 presents an agnostic, general-purpose solution.

VI. Conclusion

With the rise of demand in computer science education, development of technology and increased connectedness, expansion of distance learning, and online courses the need for online code execution and assessment is considerable and growing. Online judges are typically built with specific use-case in mind such as for competitive programming, e-learning, recruitment, etc. and they typically have an online code execution system hidden inside their architecture which allows them to run users code in a secure and isolated environment. In such architectures, online code execution systems are strongly coupled and developed for a specific use-case. In this paper, we presented a novel online code execution system architecture defined through functional and non-functional requirements and designed as a standalone and multi-purpose entity in the architecture of online judges. Finally, the proposed system is implemented and presented here as Judge0, a generic, robust, configurable, and scalable online code execution system with feature-rich web API that can be used for many different use-cases. It is thoroughly tested and heavily used in many production systems and is available to the general public as open-source software.

References

[1] R. S. Lindberg, T. H. Laine, and L. Haaranen, "Gamifying programming education in k-12: A review of programming curricula in seven countries and programming games," *British Journal of Educational Technology*, vol. 50, no. 4, pp. 1979–1995, 2019.

[2] S. Wasik, M. Antczak, J. Badura, A. Laskowski, and T. Sternal, "A survey on online judge systems and their applications," *ACM Computing Surveys (CSUR)*, vol. 51, no. 1, pp. 1–34, 2018.

[3] C. Yi, S. Feng, and Z. Gong, "A comparison of sandbox technologies used in online judge systems," in *Applied Mechanics and Materials*, vol. 490. Trans Tech Publ, 2014, pp. 1201–1204.

[4] M. Mareš and B. Blackham, "A new contest sandbox." *Olympiads in Informatics*, vol. 6, 2012.

[5] W. Felter, A. Ferreira, R. Rajamony, and J. Rubio, "An updated performance comparison of virtual machines and linux containers," in *2015 IEEE international symposium on performance analysis of systems and software (ISPASS)*. IEEE, 2015, pp. 171–172.

[6] D. Merkel, "Docker: lightweight linux containers for consistent development and deployment," *Linux journal*, vol. 2014, no. 239, p. 2, 2014.

[7] J. E. Smith and R. Nair, "The architecture of virtual machines," *Computer*, vol. 38, no. 5, pp. 32–38, 2005.

[8] H. Z. Došilović, "Judge0 IDE - Free and open-source online code editor," accessed 2020-23-02. [Online]. Available: https://ide.judge0.com

[9] S. R. Labs, "Online Compiler and IDE ¿¿ C/C++, Java, PHP, Python, Perl and 70+ other compilers and interpreters - Ideone.com," accessed 2020-23-02. [Online]. Available: https://ideone.com

[10] melpon, "Wandbox - Social Compilation Service," accessed 2020-23-02. [Online]. Available: https://wandbox.org

[11] H. Z. Došilović, "Online compiler written in less than 200 lines of code." accessed 2020-06-30. [Online]. Available: https://online-compiler.dosilovic.com

[12] Codeanywhere, "Codeanywhere · Cross Platform Cloud IDE," accessed 2020-23-02. [Online]. Available: https://www.codeanywhere.com

[13] CodeSandbox, "CodeSandbox: Online IDE for Rapid Web Development," accessed 2020-23-02. [Online]. Available: https://codesandbox.io

[14] Repl.it, "Repl.it - The world's leading online coding platform," accessed 2020-23-02. [Online]. Available: https://repl.it

[15] M. Mirzayanov, "CodeForces," accessed 2020-23-02. [Online]. Available: https://codeforces.com

[16] S. R. Labs, "Sphere Online Judge (SPOJ)," accessed 2020-23-02. [Online]. Available: https://spoj.com

[17] Directi, "CodeChef - A Platform for Aspiring Programmers," accessed 2020-23-02. [Online]. Available: https://codechef.com

[18] AlgoExpert, "AlgoExpert | Ace the Coding Interviews," accessed 2020-23-02. [Online]. Available: https://www.algoexpert.io

[19] LeetCode, "LeetCode - The World's Leading Online Programming Learning Platform," accessed 2020-23-02. [Online]. Available: https://leetcode.com

[20] HackerRank, "HackerRank - Matching developers with great companies," accessed 2020-23-02. [Online]. Available: https://hackerrank.com

[21] GitHub, "GitHub - Judge0 source code," accessed 2020-26-02. [Online]. Available: https://github.com/judge0/api

[22] I. Mekterović, L. Brkić, B. Milašinović, and M. Baranović, "Building a comprehensive automated programming assessment system," *IEEE Access*, vol. 8, pp. 81 154–81 172, 2020.

[23] H. Z. Došilović, "Judge0 - Robust and scalable open-source online code execution system," accessed 2020-26-02. [Online]. Available: https://api.judge0.com

[24] S. Maggiolo and G. Mascellani, "Introducing cms: A contest management system." *Olympiads in Informatics*, vol. 6, 2012.

[25] S. MAGGIOLO, G. MASCELLANI, and L. WEHRSTEDT, "Cms: a growing grading system." *Olympiads in Informatics*, vol. 8, 2014.

[26] Z. Gutterman, B. Pinkas, and T. Reinman, "Analysis of the linux random number generator," in *2006 IEEE Symposium on Security and Privacy (S&P'06)*. IEEE, 2006, pp. 15–pp.

[27] G. Nakagawa and S. Oikawa, "Fork bomb attack mitigation by process resource quarantine," in *2016 Fourth International Symposium on Computing and Networking (CANDAR)*. IEEE, 2016, pp. 691–695.

[28] S. Josefsson *et al.*, "The base16, base32, and base64 data encodings," RFC 4648, October, Tech. Rep., 2006.

[29] Docker, "Overview of Docker Compose," accessed 2020-26-02. [Online]. Available: https://docs.docker.com/compose

[30] S. R. Labs, "Online Compilers and Programming Challenges APIs - Sphere Engine," accessed 2020-27-02. [Online]. Available: https://sphere-engine.com

[31] A. Prologin, "camisole," accessed 2020-27-02. [Online]. Available: https://camisole.prologin.org

Adapting CERIF for a national CRIS: A case study

D. Kremenjaš*, P. Udovičić * and O. Orel*
* University of Zagreb, University Computing Centre, Zagreb, Croatia
{davorin.kremenjas, petra.udovicic, ognjen.orel}@srce.hr

***Abstract* - A development and implementation of a national Current Research Information System (CRIS) in Croatia is on its way. During the design of the system, global trends in the area of research information management and exchange were considered. The initiative to share knowledge among a broad variety of stakeholders in the national research community resulted, among other, with CERIF case study.**

CERIF is a research information data model recommended by the EU and governed by euroCRIS. From a broad point of view, it meets the initial requirements for the system.

In the area of semantics and multilingualism CERIF is effective and valuable. Nevertheless, at the very beginning of implementation, functionality gaps emerged and it was necessary to make certain adjustments, especially regarding temporal aspects. With the intention to achieve the resolution of observed gaps within the model and to satisfy key stakeholders' concerns, architectural modifications and extensions of the model were made.

This paper includes the presentation of Croatian CRIS, CERIF and the case study explaining what is needed to adapt CERIF for a national research information management system and vice versa, all this while paying attention to the critical aspects.

Keywords* - *CRIS; research metadata; CERIF; information systems; temporal databases

I. INTRODUCTION

Research information management (RIM) is a field of work whose primary areas of concern are aggregation, curation and utilization of data regarding research activities [1]. In order to effectively perform these tasks, a new class of information systems emerged in the last decade, often referred to as Current Research Information Systems (CRISes) [2]. These systems are "current" in the way they hold information which is actual and will enable analyses that are presently important, in addition to all other historic information.

In this paper, we show varieties of CRISes and current state of affairs regarding research information in Croatia, including a case for creating a national CRIS. We will present some of the common data models used for RIM and consider one of them, Common European Research Information Format (CERIF) [3] in detail. A substantial amount of work deals with informational aspects of CERIF (e.g. [4,5,6]). Main contribution of our work is the implementation of CERIF in a huge research information system, while taking all the technical aspects into account. Out of several data models, CERIF was selected as a base for this system mainly because of its completeness, institutional support and flexibility.

We discuss the abilities and limitations of common relational databases to support such a model and consider changes in CERIF in order to achieve a level of practicality and ease of use for both system architects and developers.

This paper is organized as follows. In section II CRISes are discussed in more detail and a case for a national CRIS implementation in Croatia is presented. Section III shows common research information data models used today and gives more information about CERIF. In section IV challenges of implementing CERIF in a relational database are presented, along with some adaptations in order to make it more suitable for common relational databases, while the last chapter concludes the paper.

II. A NATIONAL CRIS IN CROATIA

The scope of CRISes varies. Some only contain information about scientific publications or projects, while some contain full-scale information about researches, institutions, publications, journals, projects, patents, events (such as conferences or congresses) and many more. Additionally, most CRISes are localized to a specific institution or a university, while the cases for a national CRIS are rare [7]. Some CRISes are implemented at a national level, as aggregators of information already present in institutional CRISes, like the ones in Finland [8] and the Netherlands [9]. There are also examples of a single-installation national CRISes, where all users directly connect to and use a central system, like the ones in Norway [10] and Slovenia [11].

A. Research Information in Croatia

The project Scientific and Technological Foresight (STF) is underway in Croatia. This project has several goals, one of which is creation of the national CRIS. In the project preparation, a political decision was made to create a new system instead of buying an existing one, and possibly adjusting it to specific Croatian needs. Currently, there are various "islands of informatization" regarding research information in Croatia. There are some systems which cover only a part of all information of the entire future CRIS - publications, equipment, projects, researchers, institutions [12,13,14]. These databases are

limited in scope and barely interconnected. On top of that, none of the CRISes should be isolated, so it is expected that this system is interoperable to other national and international systems and platforms (OpenAIRE [15], ORCID [16], etc.). Even though CRIS is not a data repository, it is a good practice to interlink CRIS with institutional repositories that usually hold full texts of the publications, research data etc.

B. Motivation for a Single-Installation CRIS

When considering the possible architectures of the national CRIS, the main dilemma is whether to create an aggregator of all the information available in existing systems and use it for analytic purposes, or to create a single-installation system, which will be used by all end-users. While the first option means that existing systems will continue to be used and most of all users should not notice any change, it also means additional applications should be made in order to cover the information which is currently not collected in any way (e.g. events, services, ...) and a huge part of interoperability features would need to be implemented in all those systems and applications. In this case, creating a single-installation national CRIS which will replace some existing systems and introduce many new features makes more sense, for following reasons:

- having a single source of truth for all the research information in the country
- simplicity of not dealing with interoperability between many smaller applications
- total cost of ownership (including maintenance, further features, etc.).

Croatian CRIS (CroRIS) will therefore be created and implemented as a single-installation system. At the beginning of the project, a planning phase was conducted, and a data model was chosen and the system itself is currently being developed. The whole system will be developed and implemented into production in stages, starting with official researchers and institutions registers, followed by the first version of CroRIS portal, etc. During each phase of implementation, a new set of users will be introduced to the system, along with import of relevant data from other systems and API subsystems, to provide the interoperability platform for other national systems. It is planned that the initial development and implementation will take more than two years.

III. Other Data Models for Research Information

As stated earlier, CERIF was selected as the basic data model for CroRIS. In this section, we give a short overview of some of the other actual data models for research information and state main shortcomings that prevented their further consideration.

A. Component MetaData Infrastructure

Component MetaData Infrastructure is a part of larger CLARIN (Common Language Resources and Technology Infrastructure) project which aims to create a research infrastructure that makes language resources and technology available and readily usable to scholars of all disciplines, in particular the humanities and social sciences [17].

CMDI is CLARIN project's framework for metadata reuse [18], using components and profiles as containers and main building blocks to be used later for operational systems implementation. It is a very high-level/conceptual framework similar to Dublin Core Metadata Initiative, currently having only early-stage prototypes with limited functionality.

B. Data Documentation Initiative

Data Documentation Initiative (DDI) [19] is an international, free XML based standard backed by high profile institutions such as MIT, UCLA and Eurostat [20]. Its main target is a limited set of sciences: social, behavioral, economic, and health. It has strong support for controlled vocabularies. Special care has been taken in the area of mapping to existing standards such as Dublin Core and MARC. It boasts a state-of-the-art documentation, but it has no direct mapping to relational data model.

C. B2FIND

B2FIND is a discovery service for harvesting existing/external repositories [21] rather than a data model to implement own research information repository/CRIS. It is supported by another well-established European research infrastructure provider entity – EUDAT – and their Collaborative Data Infrastructure [22] EU wide project (focusing on storage and network infrastructure specialized for data hosting) to support research activity. It has limited support for data modeling.

IV. CERIF in Detail

A. Origin and Basics

CERIF dates back to 1970s but only in 1987 was the first version developed and formally released [23]. Since its beginnings the model improved over time. Significant development of the model started with CERIF 1991. The model included only research projects, persons, organizations, publications, equipment, facilities and a few other entities. A need for a more contextual information was recognized later on, together with the functional dependencies problem.

First formally developed and released version of CERIF was developed by Research Database working group organized by EU. During the year 2000 custodianship of CERIF was handed over to euroCRIS, with the aim to achieve interchange, integration and standardization of European Research Area.

CERIF 2000 standard showed significant improvement in the area of describing the research domain. Main entities were interconnected through new attributes: role and date/time. Other mentionable improvements that came with this version were the implementation of multilingualism, storing contact information's in one entity and inclusion of a single classification system or controlled vocabulary.

More flexibility with capturing the semantics of relationships was achieved with the release of CERIF 2008 [24]. The model form as recognized today was built as a result of dividing entities into five groups:

- Core CERIF entities – Project, Person, OrganisationUnit, ResultPublication.
- 2nd Level CERIF entities – introduced with intention to capture the context of activity and interaction. Some of those are Service, Event, Facility, Citation...
- CERIF Relationships (LINK Entities) – those entities with the structure like Entity1Name_Entity2Name with goal to describe relationships among entities while considering temporal aspect.
- Language-dependent CERIF Entities – entities requiring representation in more than one language (Title, Abstract, ResearchInterest...)
- CERIF Semantic Layer – generically made with the intention to allow representation of virtually any kind of relationship within the model.

Later releases of CERIF made improvement on entities such as research infrastructure entities (Facility, Equipment, Service), together with the geographic bounding box in the same context. New entities were created (Medium entity), new attributes were added and semantic formalities were extended (vocabulary).

Attribute values in language-dependent entities of CERIF can be stored in any language. Values can be machine and/or human translated. Relations among entities are represented with a time range of validity (start date and end date for the relation).

CERIF used today contains 295 entities, 1829 attributes, 295 primary and 3 alternate keys as well as 638 relationships.

B. Semantics and Vocabularies

Standard relational data models often employ a relatively large number of domain-specific *type entities* which correspond with certain parts of controlled system vocabularies and can be visually represented as *leaves* in ER (entity-relationship) model, to be later referenced by fact-oriented data entities through referential integrity rules. Depending on the size of the overall relational model the number of such type entities can range from just a few to hundreds in larger systems.

A distinct feature of CERIF model dealing with such type entity proliferation is its *semantic layer* which minimizes the number of such hierarchies and employing a meta-layer for modeling types, categories, statuses, flags, varieties and similar concepts instead. Such concepts in CERIF are mapped to *classifications* which are further grouped into *classification schemes*. Each classification and classification scheme have several attributes helping to describe them. Any further type-based modeling becomes unnecessary as all main data model entities now reference this new meta-model comprised of just a handful of tables (as in an example of organizational structure depicted in Figure 1.) regardless of the size of the overall data model. What was previously modeled through a scheme structure now follows semantic rather than structural hierarchy. Any links between main entities also become classified through the above mechanism, allowing for virtually endless possibilities for expressing semantic relations between different entities.

This further enables easier handling and a lot more control over formal vocabularies used both within the system, but particularly for external systems data exchange where each term must be recognized and agreed upon between all the communicating parties. If deemed necessary at some point it could also become a key enabler of *semantic web*, a much-discussed concept for over a decade now but with very slow adoption worldwide.

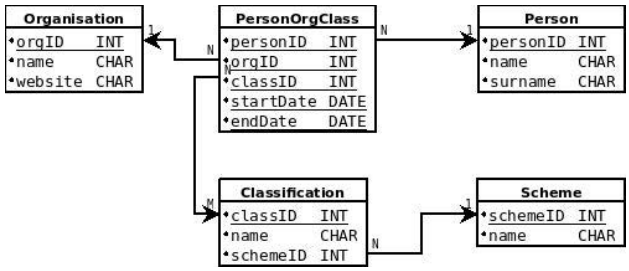


Figure 1. An organizational structure represented using Classification/Scheme

C. Why CERIF?

Among a multitude of research information targeted data models, CERIF stood out as the most appropriate for CroRIS project for a number of reasons:

- "open source": not controlled by any commercial entity; managed by euroCRIS
- strong institutional support: backed by the EU Commission
- standards based: SQL + XML
- simple: in comparison to more abstract/rich generic semantic frameworks such as ones listed above
- designed from ground up as a relational data model - our area of expertise
- extremely flexible: capable of modeling and meta-modeling just about any high-level concept using its semantic layer
- strong support for vocabulary formalization: key enabling point for EU-wide (potentially global) CRIS systems standardization and interoperability
- easy integration with external systems: OpenAIRE as a minimum, but a lot more to come
- adopted: the number of institutional, regional and national CRISes in the Directory of Research Information Systems (DRIS) [25] which claim to be CERIF compatible or use the software which is marked as CERIF compatible is 267.

Downsides:

- incomplete/inappropriate temporal aspect
- sporadically developed - several years of pause between development projects (current CERIF refactoring project started in January 2020 [26])
- no reference implementation/blueprint - some relational DB schema and XML docs but it is outdated.

Given these pros and cons of CERIF, it should be noted that, for CRIS purposes, CERIF is the most complete data model and the most suitable one.

V. CERIF Challenges and Adaptations

Relational databases still make the core of most of the big systems, in the face of recent popularity of other types of databases, namely NoSQL/NewSQL. For many reasons, a relational database will be a core of CroRIS as well.

Although CERIF provides an excellent starting point, it requires many changes and adaptations to be used efficiently and to cover all functional requirements. The below list is not exhaustive, but it still depicts a sheer number of changes introduced during the initial system design phase where a minority of planned features were covered.

A. Temporal Aspect

In CERIF, a special attention is given to the temporal information aspect. When it comes to temporal relational databases, CERIF as a data model focuses on valid time [27]. Therefore, valid time support in current relational database management systems (RDBMS) is of a special interest in our work.

Temporal support in relational databases is defined in the SQL:2011 standard [28], but current RDBMSs are yet to implement it in this regard. Many of RDBM's vendors support only system-time aspects. Implementing valid time in RDBMS means the database must support a "period" data type (a fixed interval in time), operators to deal with it [29] and temporal primary and foreign key constraints. Of these three, the most difficult is to support temporal primary and foreign keys because, in a typical relational database, these constrains are always checked only for equality, while in temporal regard the temporal part of keys must be checked using interval operators.

The most progress in implementing application-time temporal features has been made in PostgreSQL [30] and IBM DB2 [31]. Currently, PostgreSQL still lacks temporal foreign keys feature, while DB2 has temporal foreign keys included in the last version, but still with some minor restrictions.

CERIF strives to fully incorporate temporal data aspects from ground up, but falls very short of its intention, an area which proved to be the biggest obstacle to a real-world implementation in our case and probably an area of our biggest deviation from the proposed model. An example of this is presented in Figures 2a through 2c where a case of a subset of entity attributes are known to have their own temporal characteristics, but with no appropriate support in current CERIF to support it. A schema evolution is shown starting with Figure 2a. where only a Country entity has temporal aspect (supported by the current CERIF schema) through Figure 2c. where a country name has its own multilingual (supported by CERIF), but also independent temporal aspect (not supported).

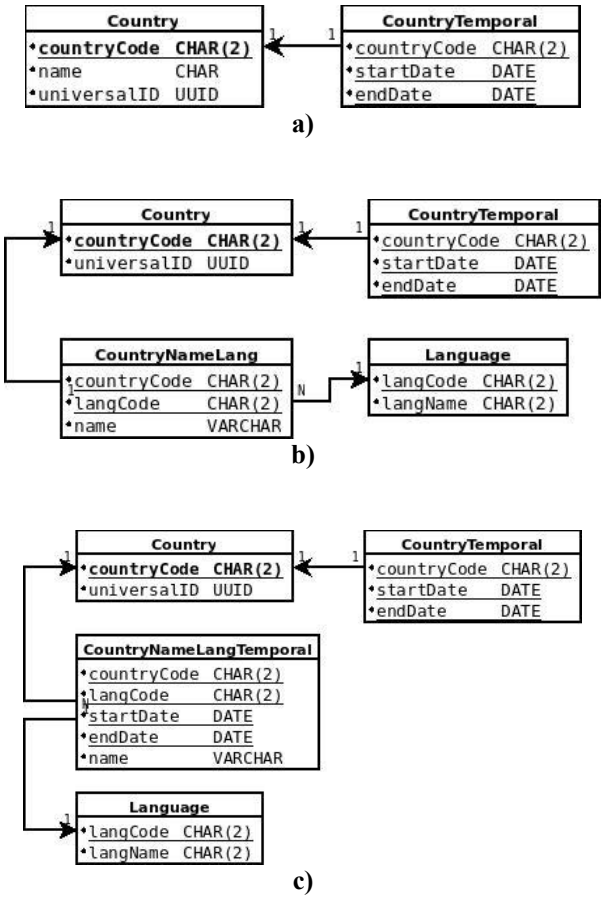


Figure 2. Multi Language Temporal Attribute within Temporal Entity

B. Changing UUIDs from CHAR (128) to CHAR (36)

A portion of the semantic layer is comprised of globally unique identifiers (UUID). An UUID is 128 bits long and can guarantee uniqueness across space and time [32]. Although well-known and easily implemented from a technical point of view, they pose a serious challenge from a data exchange perspective where disparate systems need to be fully in sync with their formal vocabularies and UUIDs representing terms of such vocabularies. As there is no European and especially no global standard to be followed in this respect within the area of research information, the only sensible approach was the adoption of the common denominator strategy, which in this case resulted in using somewhat limited OpenAIRE vocabularies.

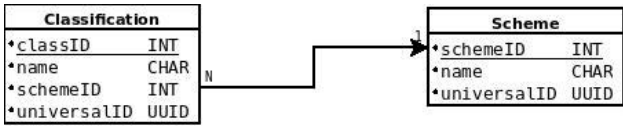


Figure 3. Classification/Scheme UUIDs

In CERIF it has been generically defined as CHAR(128) and assigned to all main entities (Figure 3.) presumably to store all bits as ones and zeros within a CHAR/string, but it is more efficient to use either a binary data type available from the underlying DBMS or a standard string-based representation which is 36

characters long (32 alphanumeric characters and 4 hyphens), e.g. 932f3658-d98c-21e2-b365-515746359182.

C. *Removed many NOT NULL constraints*

Numerous attributes within the existing CERIF model are defined as non-nullable, but for many of them there are no values defined in any publicly available standard. For example, sources for many classification definitions originally envisaged as being a part of CERIF controlled vocabularies are not available. As a consequence, all such attributes have been redefined as nullable.

D. *Descriptive attributes expanded from CHAR (127) to LVARCHAR*

A number of object definitions and descriptions did not fit into predefined templates and had to be expanded to 2k or even up to 16k variable character types.

E. *Switched from CHAR to VARCHAR for most attributes*

VARCHAR semantics proved to be a lot more flexible within the database layer and also more appropriate for object-oriented programming.

F. *Removed/Ignored unary *_Class tables where deemed unnecessary*

Most CERIF entities have a corresponding *_Class table (Figure 4.).

Our understanding of such unary classification tables is a re-affirmation of main object's existence plus their further classification and temporal limitations. In most cases such extra information was not necessary (in contrast to binary relationships where similar tables are always needed) and was removed.

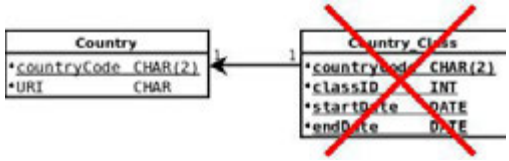


Figure 4. Country and its Classification Table

G. *Wrapped SELECT operations with views*

Due to underlying data model complexities as described above application read access to table data has been simplified and standardized using database views where possible.

H. *Wrapped INSERT/UPDATE/DELETE operations with stored procedures*

Data write access proved to be even more challenging as inserts/updates/deletes of one logical construct seen from the application point of view most often mapped to several database level objects. In this case database stored procedures proved to be of use, even though it is not the preferred approach in object-oriented programming world. For us they clearly demarcated responsibilities in each software code layer and provided a much simpler programming interface for application developers.

I. *Removed classification scheme ID from all primary and foreign key definitions*

Composite primary and foreign keys in the current CERIF version (1.6) proved to be unnecessary, adding complexity but not functionality and were removed (used simple keys instead). This is also a change scheduled for the next official CERIF version (1.7/2.0?).

J. *Removed UUID data type*

The research showed there is no standard, controlled vocabulary defining UUIDs. They were removed and primary and foreign keys were redefined as integers. This decision was taken with understanding once a widely adopted standard emerges, a key mapping to interface with the outside world will have to be defined.

K. *National Registers*

National specifics when dealing with scientific and research sources of information also had to be taken into account. Where CERIF did not provide an adequate data model to cater for local science and research information landscape, the necessary tables were added, modeling them to follow CERIF guidelines with a different table name prefix to be easily distinguishable.

L. *Added additional attributes into existing CERIF-defined tables*

A variation on the above comprised of expanding existing CERIF tables with new attributes where not all required information was already covered.

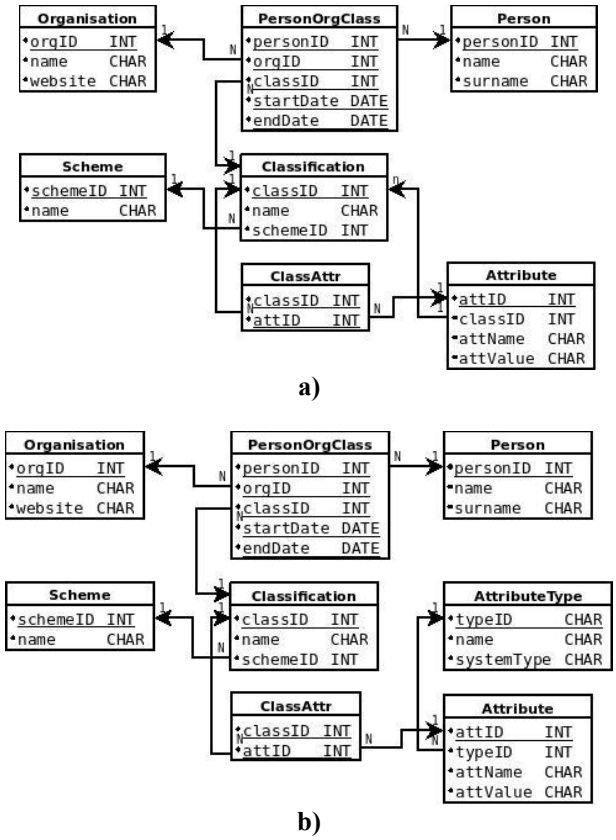


Figure 5. Organizational structure: a) no semantic layer for attribute type; b) using semantic layer for attribute types

M. Attribute Meta-modeling

Even the powerful semantic layer meta-modeling sometimes falls short when faced with more complex cases of abstract concepts. Non-trivial object classifications which themselves have extended attributes (thus evolving into first-order objects rather than just simple object classifications) frequently occur during modeling new or extending existing entities. Examples in Figure 5a. and 5b. solve such cases in a similar way: the latter using a "standard" approach; the former using CERIF-like semantic approach. This type of problems remains to be solved in our current model even though there is a general recognition of its existence and potential solutions for successful incorporation in the overall data scheme.

VI. Conclusion

In this paper we presented information systems focused on management and analyses of research information – CRISes. From our perspective of building a national CRIS, we presented some potential data models that could be used in such a system, emphasizing CERIF as the most developed one. Unlike the most of the previous work regarding CERIF, we focused on it from the technical point of view. As the main contribution of our work, we presented both positive and the negative sides of implementing it in the relational database and suggested a series of adaptations in order to make it more suitable, especially when implementing CRIS on a national level.

Our future work will include the continuation of CERIF adaptation regarding possibilities of relational databases and the needs of various stakeholders. We will focus on maintaining the most important ideas and concepts of CERIF, while at the same time enhancing it with new data in the national and international context, and giving a special attention to ease of use of temporal data and semantic layer.

References

[1] Bryant, R., Clements, A., Feltes, C., Groenewegen, D., Hoggard, S., Mercer, H., ... & Wright, J. (2017). Research information management: defining RIM and the library's role. OCLC Research.

[2] euroCRIS: Why does one need a CRIS? (https://www.eurocris.org/why-does-one-need-cris) [9.1.2020.]

[3] euroCRIS: Main features of CERIF (https://www.eurocris.org/cerif/main-features-cerif) [9.1.2020.]

[4] Jeffery, K., Houssos, N., Jörg, B., & Asserson, A. (2014). Research information management: the CERIF approach. International Journal of Metadata, Semantics and Ontologies, 1, 5-14.

[5] Ivanović, D., Surla, D., & Racković, M. (2011). A CERIF data model extension for evaluation and quantitative expression of scientific research results. Scientometrics, 86(1), 155-172.

[6] Macan, B. (2015). Model sustava informacija o znanstvenoj djelatnosti za hrvatsku akademsku zajednicu (Doctoral dissertation).

[7] euroCRIS: Directory of Research Information Systems (DRIS) (https://www.eurocris.org/eurocris-directory-research-information-systems-dris) [9.1.2020.]

[8] Puuska, H. M., Guns, R., Pölönen, J., Sivertsen, G., Mañana-Rodríguez, J., & Engels, T. (2018). Proof of concept of a European database for social sciences and humanities publications: description of the VIRTA-ENRESSH pilot.

[9] Jippes, A., Steinhoff, W., & Dijk, E. (2010, December). NARCIS: research information services on a national scale. In International Conference on Open Repositories: Proceedings.

[10] Karlstrøm, N. (2016). More than just a CRIS-How CRIStin is building a national infrastructure for Open Access.

[11] Korosec, A. (2014). SICRIS, V3. Organizacija znanja, 19(1), 22.

[12] Croatian Scientific Bibliography CROSBI (https://www.bib.irb.hr/) [9.1.2020]

[13] Šestar equipment database (https://sestar.irb.hr/en/) [9.1.2020]

[14] Croatian Science Foundation Projects database (https://www.hrzz.hr/default.aspx?id=1205) [9.1.2020]

[15] OpenAIRE (https://www.openaire.eu/) [9.1.2020]

[16] ORCID (https://orcid.org/) [9.1.2020]

[17] Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., & Koskenniemi, K. (2008). CLARIN: Common language resources and technology infrastructure. In 6th International Conference on Language Resources and Evaluation (LREC 2008).

[18] Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., & Windhouwer, M. (2012). Standardizing a component metadata infrastructure. In LREC 2012: 8th International Conference on Language Resources and Evaluation (pp. 1387-1390). European Language Resources Association (ELRA).

[19] Vardigan, M., Heus, P., & Thomas, W. (2008). Data documentation initiative: Toward a standard for the social sciences. International Journal of Digital Curation, 3(1).

[20] Gregory, A. (2011). The Data Documentation Initiative (DDI): An Introduction for National Statistical Institutes. Open Data Foundation.

[21] B2FIND (https://www.eudat.eu/services/b2find) [9.1.2020]

[22] Lecarpentier, D., Wittenburg, P., Elbers, W., Michelini, A., Kanso, R., Coveney, P., & Baxter, R. (2013). EUDAT: a new cross-disciplinary data infrastructure for science. International Journal of Digital Curation, 8(1), 279-287.

[23] Asserson, Anne, Keith G. Jeffery, and Andrei Lopatenko. "CERIF: past, present and future: an overview." (2002).

[24] Jörg, Brigitte. "CERIF: Common European Research Information Format-Insight into the CERIF 2008-1.1 Release." In CRIS. 2010.

[25] Directory of Research Information System (DRIS) (https://dspacecris.eurocris.org/cris/explore/dris) [26.5.2020]

[26] CERIF refactoring project (https://www.eurocris.org/cerif-refactoring-project-introduction) [26.5.2020]

[27] Jensen C.S., Snodgrass R.T. (2009) Valid Time. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA

[28] Kulkarni, K., & Michels, J. E. (2012). Temporal features in SQL: 2011. ACM Sigmod Record, 41(3), 34-43.

[29] Allen, J. F. (1983). Maintaining knowledge about temporal intervals. Communications of the ACM, 26(11), 832-843.

[30] PostgreSQL (https://www.postgresql.org/) [9.1.2020]

[31] IBM DB2 (https://www.ibm.com/analytics/db2) [9.1.2020]

[32] A Universally Unique IDentifier (UUID) URN Namespace (https://tools.ietf.org/html/rfc4122) [9.1.2020]

Knowledge-based system for data modelling based on verbalisation – an architecture proposal

S. Šuman * A. Jakupović * M. Kaluža*
* Business Department / Polytechnic of Rijeka, Rijeka, Croatia
ssuman@veleri.hr

Abstract - **Data modelling is a complex process that requires knowledge and experience of designers, where the quality of the data model significantly affects the quality of further phases in the development of the information system. The paper presents an architecture proposal for knowledge-based (KB) system that serves as a support for creating data models based on verbalisation. The system will use verbalized expert knowledge, which will be formalized using the methods of the theory of formal languages. Through the methodology for developing system architecture, system parts are described, and Purpose, Context, Inputs, Activities, Outputs, Effects for KB system are given through a variant of a logic model. Also, the functionalities which the verbalisation-based KB system for data modelling should have, are identified, along with all the potential beneficiaries. The paper presents the application of the theory of formal languages for translating business descriptions expressed in natural language into a data model expressed using the entity-relationship (ER) method. Example of the translation process is also given.**

Keywords - knowledge-based systems, translation, formal language, verbalisation, conceptual modelling, data models

I. INTRODUCTION

The purpose of this paper is to present an architecture proposal of the knowledge-based (KB) system for data modelling that will translate user specifications in controlled natural language (CNL) into a data model (using the entity-relationship method) presented in the formal language (verbalisation-based process). The research is carried out using the design science research (DSR) methodology [1], which prescribes the development of a conceptual model of the new system as the first phase.

The objectives to be achieved with this paper are to identify subsystems of the KB system and their interrelationships, identification of the system environment and the interaction of the system with the environment, as well as represent the basic direction of KB system development. This paper provides answers to the following questions: Q1. How to apply the theory of formal languages for translating a business description expressed in natural language into a data model expressed with ER method? Q2. What functionalities should such a KB system have? Q3. Who are its potential users?

What follows is the section Related Work which analyses existing KB systems for data modelling and highlights novelties of our approach the section Research Methodology describes the methodology for the development of the new KB system architecture, its purpose, activities and effects. The section Results and Discussion describes the developed architecture of the KB system. The Conclusion provides the answers to the research question, clarifies the limitations and describes further directions of the research.

II. RELATED WORK

The process of information system design through which informal and imprecisely defined requirements of end-users are translated into strictly defined database structures is knowledge-intensive, complex and difficult [2] and it is suitable as a problem domain for developing and applying the KB system because it is a non-algorithmic, non-trivial, and not completely deterministic problem [3]. KB system is a computer system that applies artificial intelligence methods and techniques for performing its tasks. It interacts with the user providing aid, training/learning and/or the resolution of problems. Its general structure mostly includes knowledge base inference engine, self-learning, explanation and reasoning and user interface [4].

Considering that conceptual modelling presents a critical phase in the development of information system and errors made in this stage are reflected in low quality of the database and incorrect query results [5], [6], KB systems that support this IS development phase seems convenient and useful. So far, many KB systems to support conceptual modelling in the development of the information system have been developed. KB systems' features for data modelling should enable it to intervene, propose design choice, identify errors in the conceptual model, advise users to optimize or correct the model so they could develop better data models [6].

In this paper, we synthesized the main conclusions from other authors' work (in total 14 systems or tools) [7-20]. Only the systems that use the ER method were observed. The detailed related work review is presented in [21] and [22]. The input type of all observed systems was natural or restricted natural language, Entity-Attribute-Relationship Cardinalities (EARC) extraction method was mostly performed by the system with user assistance, use of the case-based reasoning was implemented by four out of fourteen systems, targeted user type was mostly novice designers or students, output type was a textual ER model description or ER diagram.

Almost all the systems have a syntax analysis while semantic analysis was present only in some systems

usually with a series of rules or heuristics. Also, most of the systems have an output (data model) syntax analysis which is related to the successful ER method rules application.

In our new KB system, business description will be text-expressed according to a CNL, which is a subset of a natural language with restricted grammar and vocabulary. It acts as an intermediary between natural and formal languages reducing ambiguity and complexity of natural language and is suitable for knowledge representation and reasoning process [23–25].

A. New approach characteristics

This research aims to formalize the knowledge of human experts' reasoning process that is used during ER modelling. The sentential and lexical structures of business description sentences are observed and analysed to extract and formalize semantics necessary for the translation to the ER model. This approach differs from others in the following:

- development of EARC grammar that generates EARC sentences to enable the translation from one language to another,
- pattern identification from sentence parts of business descriptions that enable the creation of rules that for semantically precise translation in some EARC construct,
- additional rules based on specific lexical sentence parts that identify generalization, specialization and aggregation in the ER model, and
- specifically formed targeted questions that have an adjustable structure based on the identified missing construct – the construct that wasn't detected with automated translation.

III. RESEARCH METHODOLOGY

Building an ER data model following the user requirements is distinguished as the translation of one language into another. User requirements are viewed as text-expressed knowledge in controlled natural language that needs to be translated into text-expressed knowledge in data model formal language. Generally, the translation process consists of two phases: **lexical** and **syntactic** analysis of the input sentence (**recognition**) and generating output sentence (**translation**). The program for lexical and syntactic analysis is called the **recognizer**, while the program for translation is called the **translator**. The translator which uses recognizer in the syntactic analysis phase is called the transducer [26–30]. Fig.1 shows the model of the transducer for translating user specifications into a data model.

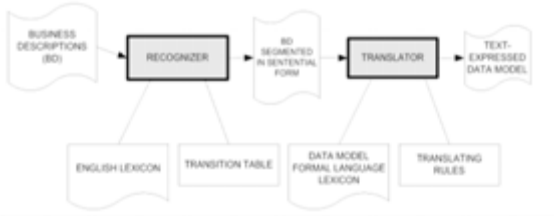


Figure 1. Business description to data model transducer

Based on lexicon and phrase-structure grammar (PSG) implemented with a transition table, recognizer performs lexical and syntactic analysis of the input sentences (business description in controlled natural language). The result is a set of BD sentences in a valid sentential form. A morpho-syntactic lexicon of a language contains word-forms, lemmas and morpho-syntactic descriptions, i.e. lexical items in the language and their linguistic properties.

The PSG was synthesized from [31], [32] and [33]and its productions are presented by symbols (Part-Of-Speech, POS tags) that are adapted to the format that is used in MULTEXT-East morpho-syntactic description (MSD) [34]. Further, the research collected 50 business descriptions, mainly examples from database textbooks and data modelling university exams which serve as BD text corpus. The current form of PSG for CNL is given in Table I.

TABLE I. PSG FOR BUSINESS DESCRIPTION EXPRESSED IN CNL

S	→	NP VP
NP	→	[D|NPs]N|Np|Pro|Ni
D	→	Dd|Di|Dg|Poss
Poss	→	NPs|Ds
NPs	→	([D|NPs]N|Np|Pg)['s|']
N	→	{AP}Nc{PP}
AP	→	{AP}([Rs|AdvP]Af)
AdvP	→	[Rs]Rm
PP	→	{PP}([PSpec]Sp(NP|PP))
PSpec	→	Rmp|McNc
Ni	→	{NP}[Cc]NP
Pro	→	Pp|Pg
VP	→	V[NP[NP|PP|AP[PP]]|AP[PP]|PP[PP]]
V	→	[Vo[Vo]Va[Rmp]|VmSp]Vm

Certain word classes of the lexicon are represented with the following MULTEXT-East lexicon MSD category:

TABLE II. WORD CLASSES AND CORRESPONDING MULTEXT-EAST LEXICON MSD CATEGORY

CNL PSG POS tag	POS category description
Af	qualificative adjective
Cc	coordinating conjunction
Dd	demonstrative determiner
Dg	general determiner
Di	indefinite determiner
Ds	possessive determiner
Mc	cardinal numeral
Nc	common noun
Np	proper noun
Pg	general pronoun
Pp	personal pronoun
Rm	modifier adverb
Rmp	positive modifier adverb
Rs	specifier adverb
Sp	preposition adposition
Va	auxiliary verb
Vm	main verb
Vo	modal verb

As it is shown in Fig. 1, the steps of the translation execution process are:

1. Obtain the sentential form - the Recognizer delivers business description in a sentential form which presents the input to Translator.

2. Execute the Translator - Translator runs the translation process as follows:

2.1. Apply the translation rules - Translator based on translation rules automatically translates parts of the BD sentential form to the word categories of the formal data model language.

2.2. Generate the question - if the data model sentential form would result incomplete (there are unmatched EARC word categories with BD sentential form parts), a specific question for every missing EARC part will be generated for the user.

3. Finalise - integration of the results of steps 2.1. and 2.2. - receiving a complete EARC sentence.

Using the induction method, a PSG for formal data model language is derived (PSG for EARC language that generates EARC, see Table III). EARC word categories are listed in Table IV. Every EARC sentence will have the following form:

S_{EARC} = $e_L((A_{pL}),(A_L))$ $(Cmin_L,Cmax_L)$ $r(Cmin_R, Cmax_R)e_R((A_{pR}),(A_R))$ of CNL

TABLE III. PSG FOR EARC LANGUAGE

S_{EARC}		Productions
A_{pL}	→	$a_{pL}\|a_{pL},A_{pL}$
A_L	→	$a_L\|a_L,A_L$
A_{pR}	→	$a_{pR}\|a_{pR},A_{pR}$
A_R	→	$a_R\|a_R,A_R$
$Cmin_L$	→	0|1
$Cmin_R$	→	0|1
$Cmax_L$	→	1 | n
$Cmax_R$	→	1 | n

TABLE IV. EARC WORD CATEGORIES

Label	Description
S_{EARC}	The form of every EARC sentence
e_L	Left entity type
e_R	Right entity type
$A_L=\{a_{L1},\ldots a_{Ln}\}$	Set of „non-primary key" left entity attributes
$A_{pL}=\{a_{pL1},..a_{pLn}\}$	Set of left entity primary key attributes
$A_R=\{a_{R1},\ldots a_{Rn}\}$	Set of „non-primary key" right entity attributes
$A_{pR}=\{a_{pR1},.a_{pRn}\}$	Set of right entity attributes that represent the primary key
r	Relationship type
$Cmin_L$	Minimum cardinality of left to the right entity (the minimum number of e_R that relates with one of the e_L)
$Cmin_R$	Minimum cardinality of right to the left entity (the minimum number of e_L that relates with one of the e_R)
$Cmax_L$	Maximum cardinality of left to the right entity (the maximum number of e_R that relates with one of the e_L)
$Cmax_R$	Maximum cardinality of right to the left entity (the maximum number of e_L that relates with one of the e_R)

As a demonstration of the translating method described in Fig. 1, the following CNL sentence is observed (example 1): **"A (really|very) adult person could be driving the city car."** The translation process for this particular case proceeds:

1. Obtain the sentential form (**SF**): "Di (Rmp) Af Nc Vo Va Vm Dd Nc Nc"

2. Execute the Translator

2.1. Apply the translation rules (Table V)

TABLE V. PART OF TRANSLATION RULES APPLICABLE TO THE OBSERVED EXAMPLE 1

Pattern	BD CNL POS tags (sentential form part) => EARC word category
*Nc*Vm*	Nc => e_L=person
* [Vo] [Va] Vm*	[Vo][Va] Vm => r =could be driving
If r contains Vo	$Cmin_R$ =0
Vm{Nc}Nc	{Nc}Nc => e_R =city car
*[Rmp|Rs]AfNc*Vm*	[Rmp|Rs]Af => a_{LV} =adult-=> a_{L1}=age

2.2. Question generation for missing EARC parts identification (still not implemented)

3. Finalise

The observed CNL sentence will be translated to the concrete EARC sentence following the **S_{EARC}** form:

person ((person_id), (name, surname, address, age**)) (0, n) could be driving (0,** n**) city car ((**car_id), (model, volume, age**))**.

The bold parts result from translation rules while normal text parts are expected to be derived from question generation. Fig. 2 shows the corresponding ERD for the above EARC sentence.

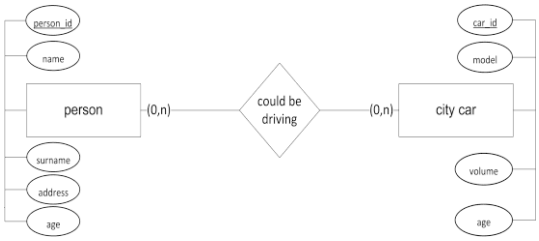


Figure 2. Corresponding ERD for the translated EARC sentence

Another example (example 2) of CNL sentence (from one of the business descriptions) translated to the EARC sentence is given in the following.

CNL: “Every employee is managed by one manager.”

SF: Dg Nc Va Vm Sp Mc Nc.

TABLE VI. TABLE 1. PART OF TRANSLATION RULES APPLICABLE TO THE OBSERVED EXAMPLE 2

Pattern	BD CNL POS tags (sentential form part) => EARC word category
Dg*Nc*Vm* and Dg =”every”	$Cmax_L$ =1
*Nc*Vm*	Nc => e_L =employee
* [Va] Vm*	[Va] Vm => r =is managed
*Vm*Mc*Nc & Mc=”one”	$Cmin_R$ =$Cmax_R$ =1

Vm{Nc}Nc	{Nc}Nc => manager
*[Rmp|Rs]AfNc*Vm*	[Rmp|Rs]Af => a_{LV}

EARC: **employee**((A_{pL}), (A_L)) (**1**, $Cmax_L$) **is managed (1, 1) manager** ((A_{pR}), (A_R))

The non-bold parts of the EARC sentence are expected to be derived from question generation.

In the development of the architecture for KB system, the following is taken into the account: potential users and their requirements, basic functionalities that every KB system should possess and functional requirements that emerged from the analysis of existing systems, tools and methods for converting BD into ER model. The following functional requirements for the new KB system are identified:

- syntax analysis
- semantic analysis
- insight into syntactically incorrect sentences - possibility for the system upgrade
- existence of question answering system that formulates the targeted questions for non-translated EARC constructs
- use of case-based reasoning that consists of the application of the same translation rules for the same or similar sentential form
- possibility of obtaining the list of all translation steps from the system input to output
- potential to upgrade, expand the knowledge base, the multipurpose and support for various input providers
- possibility to validate the system implementation performance based on the established evaluation method

IV. RESEARCH RESULTS AND DISCUSSION

To clearly describe the architecture of verbalisation based KB system for data modelling as well as its sub-systems and their information relationships, a variant of the basic Logic model [35] that describes the system through relationships between the elements (subsystems) will be used (Table VII)

Every description also includes a description of possible problems and their solutions, and ways in which to check whether the system/subsystem creates the desired result.

Fig. 3 shows the final version of the architecture proposal for the KB system for data modelling described in detail using the Logic model (Table VII).

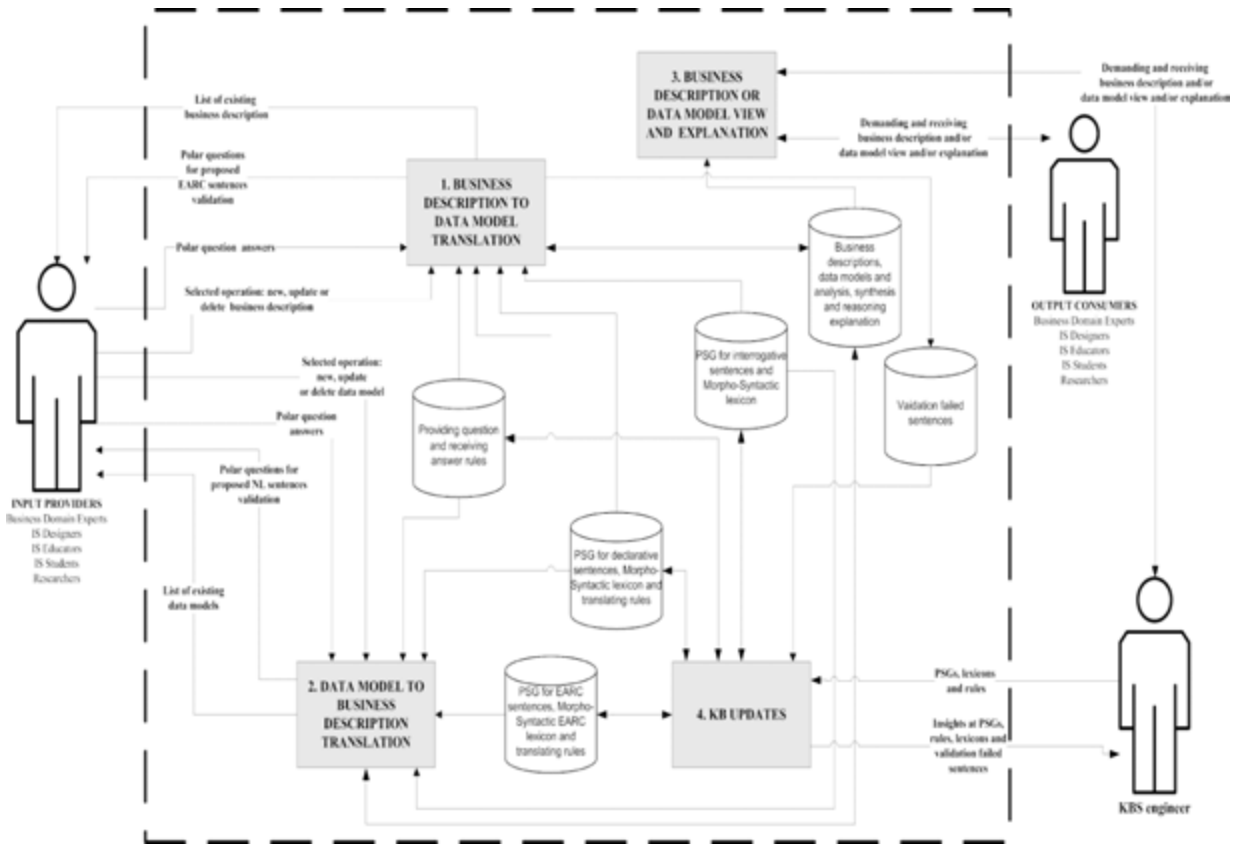


Figure 3. An architecture proposal for verbalisation based KB system for data modelling

TABLE VII. LOGIC MODEL OF VERBALISATION BASED KB SYSTEM FOR DATA MODELLING

PURPOSE

To support the process of:

- developing data models using natural language business description (for business domain experts and IS designers)
- education in data modelling domain (for IS educators and IS students)
- developing conceptual models using natural language (for researchers)

MOTIVATION

- to preserve the consistency of user requirements expressed in the natural language during its translation into a formal language of the data model
- to embed knowledge, experience and representation model of expert designers

CONTEXT

Three basic groups of users were identified:

Input providers - users who provide the KB system with necessary input (business domain experts, IS designers, IS educators, IS students and researchers)

Output consumers – users who use KB system outputs (business domain experts, IS designers, IS educators, IS students and researchers)

KBS engineer - a user who has access to KB system behaviour, and accordingly, the ability to update the system knowledge base

INPUTS

Inputs from **Input providers**:

- selected operation: create new, update or delete of business description or data model
- business and other domain-specific description (specification) written in non-formalized natural language
- answers to questions for validation of EARC sentences (for data model) and validation of NL sentences (for business description)

Inputs from **KBS engineer**:

- PSG and morpho-syntactic lexicons for declarative NL sentences, EARC sentences and interrogative NL sentences
- translation rules for NL to EARC sentences and vice versa
- rules for QA system

Inputs from **Output consumers**:

- demand for specification view and/or data model view and/or explanation

ACTIVITIES

Activities for inputs from **Input providers** (KB system creates a new, deletes or updates):

- existing business description and translates it to the data model
- existing data model and translates it to the business description
- uses QA system for missing or ambiguous part of the business description or data model, and validation of the proposed data model or proposed business description
- KB system uses a previously created data model for case-based reasoning about new data model or new business description from the given data model
- KB system does syntactic analysis and validation of NL Business description or data model described in a formal language
- KB system stores:
- each step in the process of translation from NL description to data model and vice versa
- each data model and corresponding NL specification
- every invalid sentence in NL description
- every invalid data model sentence

Activities for inputs from KBS engineer (KB system storing):

- PSG and morpho-syntactic lexicon for declarative and interrogative NL sentences and EARC sentences in the corresponding datastore
- translation rules and rules for the QA system in the corresponding datastore

Activities for inputs from Output consumers:

- based on demand, the KB system provides original NL specification, corresponding data model or an explanation of how it translated NL specification to the corresponding data model

OUTPUTS

Outputs for **Input providers**:

- list of stored business descriptions and data models
- explanation of invalid business description or data model
- specific questions for validation of proposed EARC sentences (for data model) and proposed NL sentences (for business description)

Outputs for **Output consumers**:

- NL specification view, corresponding data model view, translation steps explanation

Outputs for **KBS engineer** – Insights on:

- PSG and morpho-syntactic lexicon for declarative and interrogative NL sentences and EARC sentences
- translation rules and rules for the QA system
- business description sentences and data model sentences which failed validation
- steps of translation processes

EFFECTS

- creating a data model with the KB system decreases the problem of preserving consistency of user requirements expressed in the natural language during their translation into the formal language of the data model
- business domain experts can create a data model without IS design knowledge
- IS Designers are facilitated in the development of the data model
- Effectiveness and efficiency of IS education is increased by using the KB system
- researchers are facilitated in the development of the conceptual model from the domain of interest by applying the ER method for data modelling

V. CONCLUSION

This paper describes the architecture proposal for a KB system based on verbalisation for supporting data modelling using the ER method. We presented the purpose of the existence of computer support for the conceptual design of an information system. The novelties in our approach and possible solutions to problems in this area are also discussed.

The underlying method of applying a transducer from the theory of formal languages is described - the translation of business description textually expressed in natural language into a data model textually expressed in the formal language of the ER method. The structure of the entire system is also shown, with a detailed description of its purpose, inputs, activities, outputs and effects via the Logic model. Answers to the papers' research questions are Q1, Q2 and Q3 are:

A1. We have defined two artefacts necessary in the new KB system: PSG for CNL and PSG for EARC.

A2. We have identified functional requirements and the architecture of the KB system based on the analysis of the potential users and their requirements but also based on the analysis of existing systems, tools and methods.

A3. We have identified the following potential users: business domain experts, IS designers, IS educators, IS students and researchers, KBS engineers.

System limitations are reflected in the limited lexicon and limited grammar, as well as the use of one data modelling method. We plan to extend the corpus, lexicon and grammar with new practical examples that will result in knowledge base expansion. Future research activities include:

- the adaptation of the syntax analysis program module (finite deterministic automata) in Python,
- creation of controlled natural language corpus based on business descriptions and business dictionary terms for NLP,
- creation of special lexicon (a subset of MULTEXT - East) with extra word categories' tag that links a POS tag of a word to EARC tag,
- identification of various patterns in business description sentences that will determine translating rules,
- evaluation of the prototype, and
- extending the translation method to some other data modelling method, like UML.

REFERENCES

[1] A. Hevner and S. Chatterjee, Design Research in Information Systems, Theory and Practice, Integrated Series in Information Systems Volume 22. New York: Springer, 2010.

[2] T. J. Teorey, Database Modeling and Design: The Fundamental Principle, 3rd ed. St.Louis: Morgan Kaufmann, 1999.

[3] A. W. Lo and J. Choobineh, "Knowledge-Based Systems as Database Design Tools: A Comparative Study," in Intelligent Support Systems Technology: Knowledge Management, V. Sugumaran, Ed. IRM Press, 2002.

[4] P. S. Sajja and R. Akerkar, "Knowledge-Based Systems for Development," in Advanced Knowledge Based Systems: Model, Applications & Research, vol. 1, P. S. Sajja and R. Akerkar, Eds. Sudbury, MA, USA: Jones & Bartlett Publishers, 2010, pp. 1–11.

[5] J. Martin and J. Leben, Client/Server Databases: Enterprise Computing. Upper Saddle River, NJ: Prentice-Hall, 1995.

[6] C. Batini, S. Ceri, and S. B. Navathe, Conceptual Database Design An Entity-Relationship Approach. Redwood City, CA: The Benjamin/Cummings Publishing Company, 1992.

[7] F. Gomez, C. Segami, and C. Delaune, "A system for the semiautomatic generation of E-R models from natural language specifications," Data Knowl. Eng., vol. 29, pp. 57–81, 1999.

[8] S. Antony and D. Batra, "CODASYS: a consulting tool for novice database designers," ACM SIGMIS Database DATABASE Adv. Inf. Syst., vol. 33, no. 3, pp. 54–88, 2002.

[9] J. Choobineh and A. W. Lo, "CABSYDD: Case-Based System for Database Design," J. Manag. Inf. Syst., vol. 21, no. 3, pp. 281–314, 2004.

[10] P. Suraweera and A. Mitrovic, "An intelligent tutoring system for entity relationship modelling," Int. J. Artif. Intell. Educ., vol. 14, no. December 2004, pp. 375–417, 2004.

[11] N. Omar, P. Hanna, and P. Mc Kevitt, "Acquisition of Entity-Relationship Models from Natural Language Specifications using Heuristics," in Proceedings of the International Conference on 2005 Information Technology and Multimedia at UNITEN (ICIMU'05), Malaysia, 2005.

[12] N. Kim, S. Lee, and S. Moon, "Formalized Entity Extraction Methodology for Changeable Business Requirements," J. Inf. Sci. Eng., vol. 24, pp. 649–671, 2008.

[13] L. A. E. Al-Safadi, "Natural Language Processing for Conceptual Modeling," Int. J. Digit. Content Technol. its Appl. N 3, vol. 3, pp. 47–59, 2009.

[14] S. Lee, N. Kim, and S. Moon, "Context-adaptive approach for automated entity relationship modeling," J. Inf. Sci. Eng., vol. 26, no. 6, pp. 2229–2247, 2010.

[15] M. Shahbaz, S. Ahsan, M. Shaheen, and R. M. A. Nawab, "Automatic Generation of Extended ER Diagram Using Natural Language Processing," J. Am. Sci., vol. 7, no. 8, 2011.

[16] [Z. Kazi, L. Kazi, and B. Radulovic, "Analysis of data model correctness by using automated reasoning system," Tech. Technol. Educ. Manag., vol. 7, no. 3, pp. 1090–1100, 2012.

[17] O. Thonggoom, I. Song, and Y. An, "Semi-automatic Conceptual Data Modeling Using Entity and Relationship Instance Repositories," Er, pp. 219–232, 2011.

[18] M. I. Eid, "A Learning System For Entity Relationship Modeling," in PACIS, 2012.

[19] P. B. F. Njonko, S. Cardey, P. Greenfield, and W. El Abed, "RuleCNL: A controlled natural language for business rule specifications," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8625 LNAI, pp. 66–77, 2014.

[20] E. S. Btoush and M. M. Hammad, "Generating ER Diagrams from Requirement Specifications Based On Natural Language Processing," Int. J. Database Theory Appl., vol. 8, no. 2, pp. 61–70, 2015.

[21] S. Šuman, A. Jakupović, and F. Kuljanac Gržinić, "Knowledge-based systems for data modelling," Int. J. Enterp. Inf. Syst., vol. 12, no. 2, 2016.

[22] S. Šuman, A. Jakupović, and M. Pavlić, "Knowledge-Based Systems for Data Modelling: Review and Challenges," in Enterprise Information Systems and the Digitalization of Business Functions, M. Tavana, Ed. IGI Global, 2017, pp. 354–374.

[23] G. Hart, M. Johnson, and C. Dolbear, "Rabbit: Developing a Control Natural Language for Authoring Ontologies," in 5th European Semantic Web Conference (ESWC'08), 2008, pp. 348–360.

[24] S. Rolf, "Controlled Natural Languages for Knowledge Representation," in COLING '10, Proceedings of the 23rd International Conference on Computational Linguistics, 2010, pp. 1113–1121.

[25] N. E. Fuchs, K. Kaljurand, and T. Kuhn, "Attempto controlled english for knowledge representation," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5224 LNCS, pp. 104–124, 2008.

[26] A. V. Aho and J. D. Ullman, The Theory of Parsing, Translation and compiling. (Vol. 1.Parsing). NJ: Prentice-Hall, 1972.

[27] Z. Dovedan Han, Formalni jezici i prevodioci – sintaksna analiza i primjene. Zagreb: Element, 2012.

[28] Z. Dovedan Han, Formalni jezici i prevodioci – prevođenje i primjene. Zagreb: Element, 2013.

[29] A. Jakupović, M. Pavlić, and H. Z. Dovedan, "Formalisation method for the text expressed knowledge," Expert Syst. with Appl. 41, pp. 5308–5322, 2014.

[30] M. Pavlić, H. Z. Dovedan, and A. Jakupović, "Question Answering with a Conceptual Framework for Knowledge-Based System Development 'Node of Knowledge,'" Expert Syst. with Appl. 42, pp. 5264–5286, 2015.

[31] L. J. Brinton and D. M. Brinton, The Linguistic Structure of Modern English. Philadelphia: John Benjamins Publishing Company, 2010.

[32] M. Celce-Murcia and D. Larsen-Freeman, The Grammar Book: An ESL/EFL Teacher's Course. Boston: Heinle & Heinle, 1999.

[33] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, A comprehensive grammar of the English language. New York: Longman Inc, 1985.

[34] T. Erjavec, "MULTEXT-East - Morphosyntactic Specifications (Version 4)," 2010. [Online]. Available: http://nl.ijs.si/ME/V4/msd/html/msd-en.html. [Accessed: 20-Dec-2015].

[35] J. A. McLaughlin and G. B. Jordan, "Logic models: a tool for telling your programs performance story," Eval. Program Plann., vol. 22, no. 1, pp. 65–72, 1999.

QUALITY ASPECTS

Investigation of the accessibility of non-text content published on websites

Katja Kous, Saša Kuhar, Alen Rajšp, Boštjan Šumak
University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia
{katja.kous, sasa.kuhar, alen.rajsp, bostjan.sumak}@um.si

***Abstract* – Websites must be accessible for all users, regardless of their characteristics or needs. Although the many standards, guidelines and concepts for providing web accessibility exist, several websites still do not ensure a sufficient level of web accessibility. Existing studies report that many websites violate a large number of recommendations, and the most frequently identified violations are related to the use of alternative text for non-text elements (e.g. images). This paper focuses on the web accessibility evaluation of faculties' websites at the University of Maribor and the most frequent errors related to the guideline *Text Alternatives*. An automatic web accessibility evaluation was performed as well as a manual analysis of alternative text use for non-text elements. Based on the results, the recommendations for the improvement of alternative text use are presented.**

Keywords – evaluation; web accessibility; guidelines

I. INTRODUCTION

In recent years, the "accessibility" attribute has become an important segment of website quality characteristics. Websites must be accessible for all users, regardless of their characteristics and/or needs. Today, many web accessibility standards and guidelines provide definitions and recommendations on how to develop accessible websites [1]. Although many web standards exist, existing studies emphasize that several websites violate a large number of recommendations, and they are not consistent with guidelines presented in web accessibility standards. Many studies have reported that the most commonly identified violations are related to the use of alternative text for non-text elements (e.g. images) [2] [3] [4]. This means that non-text elements are not accessible for all web users, especially for people with disabilities. People with disabilities usually use assistive technologies to interact with the web (e.g. screen readers, text readers, speech input software, braille displays). Therefore, they use a different approach to interact with the web compared to people without limitations. For example (1) people who have difficulty perceiving visual content, who are blind, have low vision or have cognitive disabilities, use assistive technology that can read text aloud, present it visually, or convert it to braille; (2) people who are having problems with hearing or with understanding audio information use the assistive technology that can convert audio to text; (3) people who are deaf-blind read the text in braille [5]. All those assistive technologies cannot read the information presented with non-text content. For this very reason, all non-text content must be present with an alternative text. This allows assistive technologies to pass the presented information containing non-text elements to the user.

Our research, based on the automated accessibility evaluation of faculties' websites performed last year, showed that missing alternative text for non-text elements constituted the most frequently identified problem in both the years 2018 and 2019 [1]. This problem was found on 88% of web pages in 2018 and among 82% of web pages in 2019. On average, 35 identified problems per web page occurred in 2018 and 14 identified problems per web page occurred in the next year. Although the average number of occurrences decreased between these years, the number is still high. This means that people with disabilities cannot get information presented with non-text elements.

The findings presented above motivated us to investigate whether developers have increased their awareness of the accessible presentation of non-text elements and have started to follow the guideline Text Alternatives [6] [7], defined in the Web Content Accessibility Guidelines (WCAG) [8] or if the guideline Text Alternatives is still identified as the most violated one. Therefore, an automatic web accessibility evaluation was performed to identify the most frequently violated success criteria, while a manual analysis was used to identify the most common errors related to non-text elements, more specifically to images. Based on the results, the recommendations for the improvement of alternative text use are presented.

II. WEB CONTENT ACCESSIBILITY GUIDELINES

WCAG presents the most known guideline published by W3C, which explains how to create web content more accessible to both people with and without disabilities. The first version was released in 1999, while the second version was published in 2008. In 2012, WCAG 2.0 was approved as ISO standard ISO/IEC 40500 [9]. In 2018, the latest version, WCAG 2.1, was released.

WCAG 2.0 is divided into four main principles [1] [8]:

- *Perceivable* "means that the user must be able to perceive the information being presented" [8];
- *Operable* "means that the user must be able to operate the interface" [8];
- *Understandable* "means that the user must be able to understand the information as well as the operation of the user interface" [8];
- *Robust* "means that the user must be able to access the content as technologies advance" [8].

Each principle is supported by guidelines. Each guideline includes several *success criteria,* which present how suggestions can be achieved [1]. They are divided into three different conformance levels [1] [8]:

- *level A* with 25 success criteria presents the minimum level of conformance;
- *level AA* with 13 success criteria presents the medium level of conformance;
- *level AAA* with 23 success criteria presents the highest level of conformance.

A. *Guideline 1.1 Text Alternatives*

The first guideline defined in the perceivable principle is Guideline 1.1. Text Alternatives. The purpose of this guideline is to "provide text alternatives for any non-text content so that it can be changed into other forms people need" (e.g. large print, braille, speech, symbols or simpler language) [6]. The success criteria related to the mentioned guideline is the success criteria *1.1.1 Non-text Content*. Its purpose is "to make information conveyed by non-text content accessible through the use of an alternative text" [7]. The primary goal of alternative text is to make information accessible in such a way that it can be perceived with other sensory modality (e.g. auditory, visual or tactile) and satisfies user needs [7]. Alternative text is very important for people with different kinds of disabilities so that they can receive information otherwise presented with non-text content. For example, (1) people who have problems with hearing get an opportunity to display and read the alternative text; (2) people who have issues with seeing visual elements get an opportunity to listen to the alternative text using screen readers [7]. Besides, the information presented with text alternatives is more easily transform into both a simpler form and sign language [7].

Since our preliminary analysis showed that only visual elements are included in investigated faculties' web pages, this research only focuses on the alternative text for images. Most web developers know that every image must have an `alt` attribute following the HTML standard. Although they know that they need to provide alternative text for images, the success criteria *1.1.1 Non-text Content* presents the most frequently identified violations in several existing research. Because the alternative text often matters in personal interpretation, this success criteria is one of the most difficult to implement correctly [10]. This results in usually the missing, weak or even incorrect description of alternative text.

When the image includes a correct and appropriate alternative text, it can be marked as an accessible image. How to prepare the appropriate description to `alt` attributes usually depends on the purpose of the image. The images are divided into seven different groups based on their purpose [10]:

- Informative images include simple information, which can be presented with a short description in one sentence or phrase (e.g. icons (**Figure 1**), pictures, photos (**Figure 2**), and illustrations). The alternative text should present the main content or meaning of information presented in non-text elements. However, it should not be represented as a literal description of the image. A detailed literal description may be needed only in cases where the image content is all or part of the presented information. [11]



Koroška cesta 46, 2000 Maribor

Source: [12]

```
<img src="address.png" alt="Home address:">
Koroška cesta 46, 2000 Maribor, Slovenia
```

Figure 1: Example of an image used to label other information



Source: [11]

```
<img src="family.jpg"
alt="We're family-friendly.">
```

Figure 2: Example of an image conveying impression or emotion [11]

- Decorative images do not present supplementary information of the content on the website. Their primary purpose is to make the website more attractive and clear with graphical elements (e.g. the line that separates content (**Figure 3**)). For those types of images, the alternative text should be empty (`alt=""`), so that assistive technologies can skip them. [13]


```
<img src="line.png" alt="">
```

Figure 3: Example of an image used as part of the page design

- Functional images include information that presents actions (e.g. home link on a logo image (**Figure 4**), search button, shopping cart button (**Figure 5**)). The alternative text should describe the functionality presented by the image and should not include a direct image description. [14]



Source: [15]

```
<a href="http://digital-accessibility.eu/">
   <img src="Digital-Accessibility.png"
    alt="Digital Accessibility home">
</a>
```

Figure 4: Example of an image used alone as a linked logo



Source: [16]

```
<input type="image" src="shopping_cart.png" alt
="add to shopping cart">
```

Figure 5: Example of a stand-alone icon image that has a function

- Images with the text present the images that include the text, which should be read. In many cases, experts suggest avoiding text on images and replacing it with the original text. The size of the original text, background and text colors can be modified to suit the users' individual preferences without losing clarity. As such, the original text is much more customizable than images with text. In some cases where images with text are used, the alternative text must include the same text as presented in the image (**Figure 6**). [17]

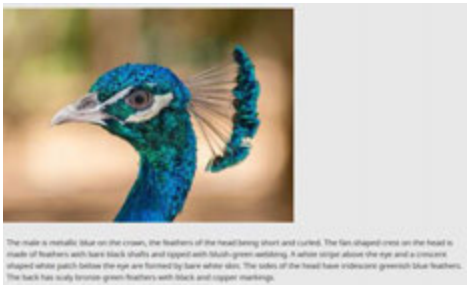


Source:[17]

```
<img src="access-city.png"
alt="Your access to the city.">
```

Figure 6: Example of styled text with decorative effect [17]

- Complex images usually include much information (e.g. graphs, diagrams, illustrations). The alternative text is presented in a two-part description: the first part is presented as a short image description and includes the location of the long description, while the second part is presented as a long image description with detailed information presented of the image. [18] [19] As an alternative to this approach, WAI-ARIA aria-described by attribute, as presented in **Figure 7**, can be used.



Source: [18]

```
<img src="peacock.jpg"
    alt="Male peacock head"
    aria-describedby="description">
[…]
<p id="description">
 The male is metallic blue on the crown, the
feathers of the head being short and curled. The
fan-shaped crest on the head is made of feathers
with bare black shafts and tipped with blush-
green webbing. A white stripe above the eye and a
crescent-shaped white patch below the eye are
formed by bare white skin. The sides of the head
have iridescent greenish blue feathers. The back
has scaly bronze-green feathers with black and
copper markings.
</p>
```

Figure 7: Example of a description containing textual information [18]

- Groups of images are images where groups of images represent one piece of information. The image, where the whole image is composed of several images (e.g. a collection of star icons presented in **Figure 8**) only needs the alternative text for one image that is considered as a whole, while the other images can have an empty `alt` attribute so that they can be skipped by assistive technology. The images presented with a collection of related images (e.g. "before and after images", images over the years) need alternative text for each image. The alternative text should describe each image individually, as well as its relationship within the collection of related images. [20]

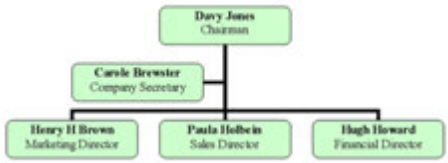


Source:[20]

```
<img src="star-full.jpg"alt="3.5 out of 5 stars">
<img src="star-full.jpg"  alt="">
<img src="star-full.jpg"  alt="">
<img src="star-half.jpg"  alt="">
<img src="star-empty.jpg" alt="">
```

Figure 8: Example of multiple images conveying a single piece of information [20]

- Image maps are images with multiple clickable areas (e.g. work breakdown structure in **Figure 9**). Each individually clickable area should include alternative text with a description of the purpose or destination of the link. [21]



Source: [21]

```
<img src="orgchart.png"
     alt="Board of directors and related staff: "
     usemap="#Map">
<map id="Map" name="Map">
    <area shape="rect"
          coords="176,14,323,58"
          alt="Davy Jones: Chairman">
    […]
    <area shape="rect"
          coords="6,138,155,182"
          alt="Harry H Brown: Marketing Director
               (reports to chairman)">
    […]
</map>
```

Figure 9: Example of an organizational chart [21]

Summarized - all images types mentioned above must have an alternative text, except decorative images or images in groups, which convey a single piece of information, where the `alt` attribute should be empty. The way of alternative text presentation depends on the image type. In addition to the automatic investigation of the use of alternative text, we also investigated if the image included the correct and appropriate description of alternative text according to their respective types.

III. RESEARCH

A. Research goal

The primary goals of this research were to investigate the accessibility of faculties' websites at the University of Maribor in 2019 and to verify whether the success criterion *1.1.1 Non-text content* was again identified as the most frequently violated one. Additionally, the research includes an investigation of the most common errors related to the mentioned success criterion and verification of correct and appropriate use of alternative text.

B. The investigated websites

Our research included 16 faculties' websites (94% of faculties' websites at the University of Maribor) where each of them has a unique structure and design. One website was excluded because it was not available during the evaluation period and the web accessibility evaluation could not be performed. According to existing research [1][4][22][23][24][25], the investigation focused on an entry web page (hereinafter web page), because it represents the first connection with users [1]. If the entry web page is not accessible, the users would have trouble accessing other pages within the website [1].

C. Data collection protocol

The evaluation of web accessibility was divided into two phases. The first phase included an evaluation of web accessibility with the automatic evaluation tool AChecher, which is the most commonly used open-source web accessibility evaluation tool [1] [26]. In this research, we checked the compliance levels of faculties' websites with WCAG 2.0 (conformance levels A, AA, and AAA). The second phase was based on a manual review of the violated success criteria *Non-text Content* and included an investigation and comparison of using the images and the alternative text for non-text content.

IV. RESULTS AND FINDINGS

In 2019, 677 errors were identified for all the investigated web pages. On average, one web page included 42 problems, of which 15 identified problems related to level A, 13 problems to level AA and 14 problems to level AAA of WCAG conformance.

As shown in Table 1, the maximum sum of the identified problems was 228 problems for the success criteria *Contrast* (level AAA), followed by 175 problems for the success criteria *Non-text Content* (A) and 173 problems for the success criteria *Resize text* (AA). The most frequently identified problem was the success criteria *Non-text Content* (A) which was found on 88% of web pages, followed by the success criteria *Resize text* (AA), which was identified among 63% of web pages, the success criteria *Headings and Labels* (AA) that was found on 56% of web pages and the success criteria *Contrast* (AAA), which was identified on 50% of web pages.

Table 1: The most frequently identified success criteria

Success criteria	SUM	Avg. number per web page	Web page [%]
1.1.1 Non-text Content (Level A)	175	11	88
1.3.1 Info and Relationships (Level A)	11	1	31
1.4.4 Resize text (Level AA)	173	11	63
1.4.6 Contrast (Enhanced) (AAA)	228	14	50
2.1.1 Keyboard (Level A)	22	1	6
2.4.2 Page Titled (Level A)	1	<1	6
2.4.4 Link Purpose (In Context) (Level A)	19	1	44
2.4.6 Headings and Labels (Level AA)	31	2	56
3.1.1 Language of Page (Level A)	4	<1	13
3.1.2 Language of Parts (Level AA)	1	<1	6
3.3.2 Labels or Instructions (Level A)	7	<1	31
4.1.1 Parsing (Level A)	5	<1	31

As expected, the most frequently identified problem was the success criteria *Non-text Content* (A). 88% of investigated web pages included, on average, 11 problems related to the success criteria *Non-text Content*. AChecker identified four different error types related to the success criteria *Non-text Content* with a different number of occurrences. As presented in Table 2, the most frequently identified error is Check 1, which requires that all `img` elements should have an `alt` attribute. On average, the error occured six times among 75% of the investigated webpages. The second most frequently identified error is Check 7, which requires that "`alt` text for all `img` elements used as source anchors is not empty when there is no other text in the anchor" [27]. This error was found among 56% of investigated webpages and on average, it occured five times per one web page. The third most frequently identified errors were Check 64 and Check 238. Check 64 requires that all `area` elements have an `alt` attribute, while Check 238 recommends that all `input` elements, except those with a `type` attribute value of "image", do not have an `alt` attribute. Both those errors were identified on 6% of investigated web pages and their occurrence averages less than 1 error per web page.

Table 2: The most frequently identified known problems

ID Check	Error	Sum of identified errors	Avg. number per web page	Web page [%]
64	Imagemap `area` element missing `alt` attribute.	2	<1	6%
1	`img` element missing `alt` attribute.	93	6	75%
7	Image used as anchor is missing valid Alt text.	79	5	56%
238	`input` element has `alt` attribute.	1	<1	6%

Based on the manual analysis of alternative text usage for non-text elements, we analyzed 256 images. On average, one web page included 16 images. Most of the investigated images could be classified as functional and informative images. The image map and decorative images were found only on one web page, while the other types were not identified on the investigated web pages.

A manual analysis revealed that only 24% (62) of the investigated images included the correct use of alternative text. This means that the `alt` element was used according to recommendations given for specific image types. The remaining images (76% of investigated images) included one of the identified errors related to the use of alternative text:

- 37% of investigated images did not contain an `alt` attribute (Check 1), and a further 1% of those 37% did not include `alt` attribute for the Image map (Check 64),
- 31% of them included an `alt` attribute, but it was empty, and did not describe the image information (Check 7) and
- 7% of the investigated images did not include the correct description of `alt` attributes according to their image type.

V. Discussion and Recommendations for Improved Presentation of Alternative Text

To summarize, 24% of the investigated images included the correct use of alternative text. This means that the alternative text for the informative images was presented with a short phrase or sentence that described the concept or information presented in the image, while the alternative text for the functional images was presented with a description of activities (action) that will be initiated (the purpose of the image) and did not include an image description.

According to the results, most images (37% of the investigated images) did not have an `alt` attribute. When images are missing the `alt` attribute, they cannot be accessed by people with disabilities. This presents significant issues for users who use assistive technologies. For example, a missing `alt` attribute for functional images generates significant issues for users who use screen readers because functional images present the functionality of the content [10]. Screen readers usually read the image file name, the image URL, or the URL for the link destination, which does not help users to understand what functionality will be activated when clicking on the image [10]. The missing `alt` attribute for an informative image also causes similar problems. In this case, the user cannot obtain the important information that is present on the image.

The next issue that was discovered was when the `alt` attribute was included, but was empty, and did not present the images' information (31% of investigated images). This empty `alt` attribute would be suitable only for decorative images. Nevertheless, since investigated images have been classified as functional or informative, they again pose a problem for recognizing the content or action meaning for persons with disabilities.

The last issue, found in 7% of the investigated images, was that images included the `alt` attribute, but that the description was defined incorrectly. This means that the description of the alternative text did not provide the correct information related to image type and image content.

To avoid the issues presented above, the functional images must include an alternative text with the description of the action that will be initiated, while an alternative text for an informative image should provide a short description with the same information as presented in the images.

To summarize, the first step to implementing an appropriate alternative text is to identify the image type. After that, the appropriate alternative text is easier to define if the rules for the specific image type are considered. Although the separated recommendations for specific image types exist, some general suggestions for defining an alternative text can be given. Therefore, the description of the `alt` attribute should [19]:

- be precise and equivalent to the information presented on the image;
- be understandable and short;
- be specific;
- not present unnecessary and redundant information;
- include the full title of the window, tab, heading, menu, or screen;
- not use the phrases " picture/image/diagram/graph of" to descript the content presented on images;
- not include line breaks, because the screen reader recognizes this as a pause.

A. Limitations and future work

This study has limitations that should be taken into consideration. In this research, only entry pages were included in the evaluation. If the research included the full web pages of each faculty, the number of identified problems might be different. Additionally, if a different automatic evaluation tool would be used, different problems might also be identified. A further limitation is related to the investigation the web accessibility of the specific domain. Therefore, the results cannot be generalized to other domains.

Our future work will be focused into the investigation of accessibility non-text content published on faculties' full web pages. Additionally, the next research will also be focused on the evaluation of websites in other domains (e.g. e-government, e-health). Based on the results, we will perform a comparative analysis between the different domains.

VI. Conclusion

This research focused on web accessibility, especially on the most frequently identified problems related to the

success criteria *Non-text Content*. The research was divided into two phases. The results of the first phase, which was based on automatic web accessibility evaluation, indicated that the success criterion *Non-text Content* is still identified as the most frequently violated one. The second phase was based on a manual review of the violated success criteria *Non-text Content*. It included an investigation of alternative text use for non-text content, more specifically for images. The analysis showed that only 24% of investigated images included the correct use of alternative text. The remaining images (76% of investigated images) included one of the identified errors related to the use of alternative text. In most cases, images were missing the `alt` attribute, or included the empty `alt` attribute. In some cases, the alternative text did not include the correct description of the `alt` attribute according to their image type. This means that all those images were not accessible for people with disabilities, especially for users who use assistive technologies to interact with the web.

ACKNOWLEDGEMENTS

The authors acknowledge financial support from the Slovenian Research Agency (research core funding No. P2-0057).

REFERENCES

[1] K. Kous, S. Kuhar, M. Pušnik, and B. Šumak, "Comparative analysis of faculties' websites accessibility based on an automatic evaluation," in 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2019, pp. 1498–1502.

[2] A. Silva, M. J. A. Goncalves, S. Teixeira, A. M. Silva, and T. Maia, "Accountability and Web Accessibility of the Portuguese official municipal Websites," in Vision 2020: Innovation Management, Development Sustainability, and Competitive Economic Growth, 2016, pp. 3941–3952.

[3] B. A. King and N. E. Youngblood, "E-government in Alabama: An analysis of county voting and election website content, usability, accessibility, and mobile readiness," Gov. Inf. Q., vol. 33, no. 4, pp. 715–726, 2016.

[4] C. Pribeanu, P. Marinescu, Ruxandra-Dora Fogarassy-Neszly, and M. Gheorghe-Moisii, "Web Accessibility in Romania: The Conformance of Municipal Web Sites to Web Content Accessibility Guidelines," Inform. Econ., vol. 16, no. 1, 2012.

[5] W3C WAI, "Understanding SC 1.1.1." [Online]. Available: https://www.w3.org/TR/UNDERSTANDING-WCAG20/text-equiv-all.html.

[6] WAI, "Understanding Guideline 1.1: Text Alternatives." [Online]. Available: https://www.w3.org/WAI/WCAG21/Understanding/text-alternatives.

[7] WAI, "Understanding Success Criterion 1.1.1: Non-text Content." [Online]. Available: https://www.w3.org/WAI/WCAG21/Understanding/non-text-content.

[8] W3C WAI, "Understanding WCAG 2.0," 2016. [Online]. Available: https://www.w3.org/TR/UNDERSTANDING-WCAG20/Overview.html#contents.

[9] International Organization for Standardization, "ISO/IEC 40500:2012 Information technology - W3C Web Content Accessibility Guidelines (WCAG) 2.0," International Organization for Standarization, 2012, 2012.

[10] WAI, "Images Concepts," 2019. [Online]. Available: https://www.w3.org/WAI/tutorials/images/.

[11] W3C WAI, "Informative Images," 2019. [Online]. Available: https://www.w3.org/WAI/tutorials/images/informative/.

[12] Flaticon, "Home Address free icon," 2020. [Online]. Available: https://www.flaticon.com/free-icon/home-address_1239525.

[13] W3C WAI, "Decorative Images," 2019. [Online]. Available: https://www.w3.org/WAI/tutorials/images/decorative/.

[14] W3C WAI, "Functional images," 2019. [Online]. Available: https://www.w3.org/WAI/tutorials/images/functional/.

[15] Certified Digital Accessibility Training, "Certified Digital Accessibility Training," 2020. [Online]. Available: http://digital-accessibility.eu/.

[16] Mobirise, "Cart, add, shopping Free Icon," 2020. [Online]. Available: https://icon-icons.com/icon/cart-add-shopping/99553.

[17] W3C WAI, "Images of Text," 2019. [Online]. Available: https://www.w3.org/WAI/tutorials/images/textual/.

[18] W3C WAI, "Complex Images," 2019. [Online]. Available: https://www.w3.org/WAI/tutorials/images/complex/.

[19] WebAIM, "Alternative Text," 2019. [Online]. Available: https://webaim.org/techniques/alttext/.

[20] W3C WAI, "Groups of Images," 2019. [Online]. Available: https://www.w3.org/WAI/tutorials/images/groups/.

[21] W3C WAI, "Image Maps." [Online]. Available: https://www.w3.org/WAI/tutorials/images/imagemap/.

[22] Y. Akgül and K. Vatansever, "Web Content Accessibility of Municipal Web Sites in Turkey," J. Adv. Inf. Technol., vol. 7, no. 1, 2016.

[23] J. Lazar et al., "A longitudinal study of state government homepage accessibility in Maryland and the role of web page templates for improving accessibility," Gov. Inf. Q., vol. 30, no. 3, pp. 289–299, 2013.

[24] Y. Shi, "The accessibility of Chinese local government Web sites: An exploratory study," Gov. Inf. Q., vol. 24, pp. 377–403, 2007.

[25] N. E. Youngblood and J. Mackiewicz, "A usability analysis of municipal government website home pages in Alabama," Gov. Inf. Q., vol. 29, no. 4, pp. 582–588, 2012.

[26] N. A. Karaim and Y. Inal, "Usability and accessibility evaluation of Libyan government websites," Univers. Access Inf. Soc., no. 0123456789, pp. 1–10, 2017.

[27] "Web Accessibility Checker - Suggestion." [Online]. Available: https://achecker.ca/checker/suggestion.php?id=7.

RESCCUE RAF App – Using Technology to Mitigate Climate Change Urban Impacts

P. Lopes[1], A. Oliveira[1], C. Pereira[2], R. S. Brito[2], M. A. Cardoso[2], R. Martins [1], M. David[3], J. Gomes[3], J. Pina[3]

[1] Information Technology in Water and Environment Research Group, Hydraulics and Environment Department, LNEC, Av. do Brasil, 1700-066 Lisbon, Portugal

[2] Urban Water Unit, Hydraulics and Environment Department, LNEC, Av. do Brasil, 1700-066 Lisbon, Portugal

[3] Laboratory of Instrumentation and Experimental Particle Physics, Av. Prof. Gama Pinto 2, 1649-003 Lisbon, Portugal

plopes@lnec.pt, aoliveira@lnec.pt, clpereira@lnec.pt, rsbrito@lnec.pt, macardoso@lnec.pt, rjmartins@lnec.pt, david@lip.pt, jorge@lip.pt, jpina@lip.pt,

Abstract - **Climate Change (CC) is nowadays one of the most important concerns for urban management. Given its scalability, technology can take an important role to help city managers identify and mitigate specific CC-related problems. Herein, an innovative and user-friendly web application is presented to empower city managers with a tool to cope with CC, giving them detailed information , accessible from anywhere. The usage of this app allows city managers to decide on optimized measures to address vulnerabilities to CC in their city. A methodology of city assessment that goes deep into city services processes, their characterization and interdependencies is behind the app, materialized by the Resilience Assessment Framework (RAF), identifying the specific areas and topics that have higher vulnerability to CC z d. T y' xp d within the app, at multiple levels, using intuitive and interactive chart and tabular information. Given its generic nature, the RESCCUE RAF App can be applied to other cities and take a major role in the urban resilience improvement to cope with climate change effects. Adaptation for other cities requirements can be developed g dm p f w g RAF' classification levels are possible.**

Keywords - Climate change; Resilience assessment; Urban areas;

I. INTRODUCTION

A. City resilience importance and short state-of-the-art on assessment methodologies

Urban areas are complex and dynamic systems, with interacting and interdependent strategic services and assets. Therefore, the management of these areas is complex, requiring involvement of a multiplicity of stakeholders. Potential effects of climate dynamics, such as intense precipitation events, tidal and surge effects or droughts, may significantly affect the strategic services, people, natural environment and economy in these areas. These challenges require an integrated and forward-looking approach to resilient and sustainable urban development [4]. The resilience concept has evolved along time and among disciplines [14], [16]. Herein, urban resilience refers to the ability of human settlements to withstand, recover quickly and adapt from any plausible hazards. Resilience to disruptive events not only refers to reducing risks and damage from disasters, but also the ability to quickly bounce back to a stable state. Besides addressing disaster risk reduction, resilience includes changes in circumstances [1], [14].

Since the cities are dynamic systems and considering the principle of continuous improvement [11], it is essential to undertake regularly the assessment of their resilience and to have the reliable information needed and tools to support this. Different tools and frameworks for resilience assessment have been developed in several fields of study e.g. ICLEI 2010, UN-Habitat CRPT 2013, Rockefeller&Arup 2014, World Bank 2015, UNISDR 2015, EPA 2017 among others [14], [19], [20], [17], [6].

A resilience assessment is necessary for identification of the real needs for resilience improvement, as well as the efficiency and effectiveness of planned or implemented measures. Therefore, assessing current and future resilience constitutes the basis for cities to know where they stand, to support decision on strategies, actions and measures to consider, planning in the long, medium and short terms and assessing progress. In this context, a Resilience Assessment Framework (RAF) was developed [4] within the H2020 RESCCUE project – Resilience to cope with climate change in urban areas [19].

B. Technology importance for city support and for resilience to CC assessment

Technology has been exponentially evolving on recent years, and its capabilities have become essential across innumerous areas allowing resolving tasks in a more efficient way or creating new processes that were not possible without it.

Cities are constantly evolving and have the tendency to increasingly grow and it's difficult to city managers to get detailed information on all the service processes of the cities. Given its size and complexity, cities can benefit

with the use of technology to gather detailed information to cope with CC and to facilitate communication among different involved stakeholders.

In this paper, a general description of the RESCCUE RAF methodology is presented and the overall architecture of the application, its components and user's profile is described. A brief presentation of the use of the application by the cities, in the scope of the RESCCUE project, is made.

C. *A brief overview of the RAF methodology*

The Resilience Assessment Framework (RAF) is an objective driven framework that was developed considering the RESCCUE's scope i.e. urban resilience to climate change (CC), with focus on the urban water cycle. The emphasis is on city, services and infrastructure resilience [3], [4]. Therefore, the RAF purposes are to facilitate an objective-driven resilience diagnosis of urban cities and services; to support decision on selection of resilience measures and the development of strategies to enhance resilience to climate change; to support the development of resilience action plans tailored to each city or to each service, and to track progress over time.

The RAF considers four resilience dimensions [13]: organizational, spatial, functional and physical. The first integrates governance relations and urban population involvement, at the city level. The second, also at the city level, refers to urban space and environment. The resilience of strategic services is assessed in the functional dimension, while the physical dimension focuses on the resilience of their infrastructure. The last two dimensions also allow to know the contribution of each service to city's resilience [3], [4]. The services within the RAF scope are water supply, wastewater, storm water and solid waste management, energy supply and mobility.

The RAF validation involved different stakeholders, representatives of research, city and urban services, allowing incorporating their concerns as well as their own context and reality. To ensure coherence, feasibility and effectiveness of the approach, this framework was applied to three cities Barcelona (Spain), Lisbon (Portugal) and Bristol (UK) by their cities and strategic services managers [3], [4].

The RAF has a tree structure (Fig. 1), where, for each dimension, resilience objectives are defined, translating the ambitions to be achieved in the medium-long term by the city and services. The functional and physical dimensions firstly unfold into sub dimensions (one for each service under assessment). For each objective, key criteria are specified, which express the different points of view through which the objectives are to be assessed. A set of metrics is associated to each criterion. Metrics are parameters or functions that allow undertaking the criterion assessment. By comparing the result of a metric with reference values, it is possible to assign a judgment to the responses, reflecting a resilience development level. The whole assessment translates the resilience maturity of the city or of the service [4] and allows to identify the main strengthens and weaknesses.

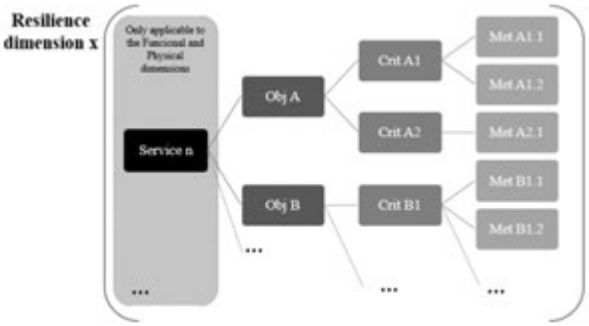


Figure 1. RAF structure (Cardoso and Brito, 2019)

This framework is complemented by the characterization of the city and the services context that frame the interpretation of the assessment results. The main characterization themes are geography, climate, population, economy and governance, built environment and infrastructure, type of customers, components and dimension of services and infrastructures.

II. RESCCUE RAF APP DESCRIPTION

A. *Concept and general description – integrating complexity into simple metrics*

Given its size and complexity, cities can benefit with the use of technology to gather detailed information to support coping with CC. To gather and process this information a web application was developed. RESCCUE RAF App relies on a web framework to facilitate the usage of the RAF methodology in a dynamic and easy-friendly way and able to producing instant results.

Developing a tool to provide such potentialities requires the identification of the necessary requirements. A list of the main requirements and a short explanation for each one is presented below:

1. Simplicity: the application should be the materialization of all the complexity of the RAF structure in a user-friendly web interface without compromising the detail of RAF. The intuitive menus within the application allow that anybody can use the app efficiently.

2. Fast and fluid: the user experience in the app should be fluid and different components of RAF should be dynamically highlighted for quick user understanding. Using client-side web technologies the transitions in the app are fast and fluid.

3. Flexible: depending on user knowledge about resilience assessment, the application should be flexible so the information required and given in the app is adjusted to the user knowledge level. There are three distinct levels of detail of information required to user, corresponding to a relevance level: essential, complementary and comprehensive. The user can alternate between levels at any time according to his knowledge or experience.

4. Customizable: advanced/authorized users should have the possibility of customize components of the application according to their needs and objectives. As administration interface was developed, so authorized users can easily tailor any of the components of the RAF structure to his needs.

5. Instant results: the application should provide instant results after data input from the user. The app uses resources from *Infraestrutura Nacional de Computação Distribuída* (INCD) as computation and storage.

6. Accessible: the access to the application should be made from anywhere using a large range of devices. Relying on a web framework, the app usage does not require any local installation on the user device. The only requirement is a web browser (*Google Chrome* or *Mozilla Firefox*) and an internet connection.

7. Confidentiality: all the data should be confidential; each user needs to be assured that his information in the app cannot be accessed by other users unless he wants to. The app access is made by user authentication ensuring the confidentiality of the data inserted.

The RESCCUE RAF App is integrated as an on-demand online service of the Portuguese Infrastructures Roadmap, under the *Infraestrutura Nacional de Computação Distribuída* (INCD) project. INCD provides the resources for the apps' data computation and storage, and assures its scalability to handle multiple, simultaneous user requests as well as database storage growth.

B. The overall architecture and components

The RESCCUE RAF App can be accessed at http://resccue.lnec.pt/ upon registration. This app has a typical Django web app architecture, containing the Frontend where the user interaction components are and the Backend containing the web framework itself and the application database (Fig. 2).

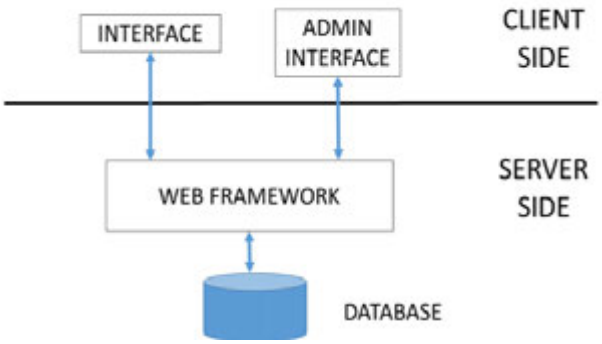


Figure 2. RESCCUE RAF App architecture

- Interface: The application main interface materializes several menus that present to the user RAF structure in a dynamic way. After authentication, the user can manage his previously created studies, insert new information or check results.

Each study takes the user through several categories of information required include among others the historic records of climate-related events and a description of city services and infrastructures. This information is processed on-the-fly and several indicators of the RAF are calculated, organized along resilience dimensions, objectives, criteria and metrics.

After the results generation, these can be explored by the user on the results section of the RESCCUE RAF App. The analysis can be made at different levels, ranging from an overall metric analysis (Fig. 3) to a more detailed analysis of a specific city criteria.

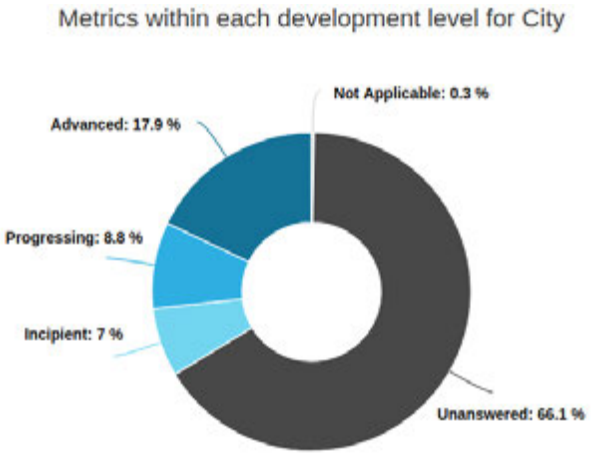


Figure 3. Donut chart with metric overall statistics

- Administration interface: interface where authorized users can customize the application.

There are three major user profiles in this application:

1. IT administrator: admin user with deep understanding of the deployment of the application and all its inherent

processes. Manages users, authentications, uptime of the application.

2. Staff user: users with deep understanding of the RAF and brief knowledge of how the application works. Can customize the RAF structure and the way that app uses to calculate the assessment level in each metric using administration interface.

3. User: only has access to the "main" interface of the app, can create new studies, fill metric with information regarding his studies and check results.

In the admin interface, staff users can customize all RAF components from dimension drilling down until metric. In Fig. 4, one of the menu to assign each service to respective dimension is shown.

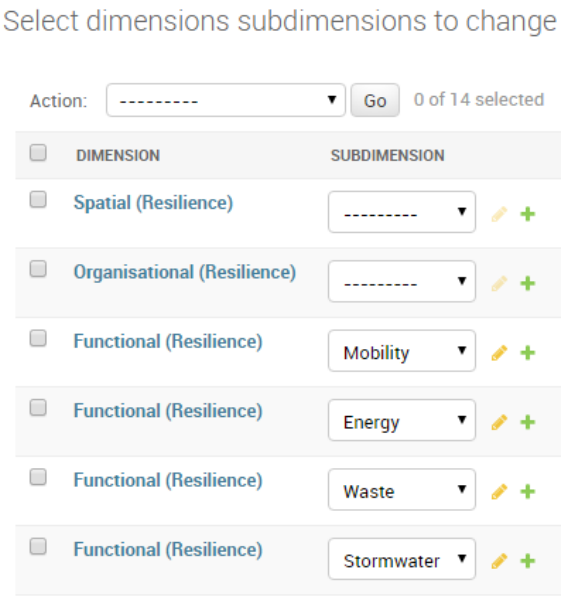


Figure 4. Administration menu of RAF dimensions and services management

Besides RAF structure customization, the methodology of calculation of the assessment to each metric can also be customized.

This customization capacity makes the RESCCUE RAF App, a tool that can be finely adapted to each city type, being it a coastal city with emphasis in water or a high dense populated city where waste management can be an important service to assess.

- Web framework: this component provides access to the application, through web pages hosting the major configurations. The Django Web Framework is used as the development framework of this component taking part of its properties for to a correct behavior of the application.

- Database: The storage component keeps the state of all information in the app. It is accessed by the Web framework to process all the requests made by interface.

III. DEALING WITH THE COASTAL DIMENSION AND FLOODING ISSUES

The city of Lisbon is located at the northern margin of the Tagus estuary, one of the largest estuaries in Europe, being therefore subject to inundation associated with its water levels rising [7]. Water level variability in this estuary is mostly driven by tides and storm surges, the Tagus morphological characteristics promoting tidal amplification [7], [10] and the risk for flooding of its margins [15]. The combination of extreme tidal levels and storm surge conditions has led to several inundation episodes, some of them with very severe consequences. Examples include the February 15th 1941, with high human casualties and property damages [12] and more recently the flood event associated with the passage of the Xynthia storm by the Portuguese coast [8][9]. At a more local scale, the impact of urban drainage has led to severe inundation episodes both at the city of Lisbon [5] and adjacent areas [2].

As this city's results obtained using the app are confidential, we focus here on how the app is prepared to deal with a coastal city and what are the relevant metrics, rather than presenting the real results for the use of the app for Lisbon.

The RESCCUE app can be used by a coastal city's managers to evaluate the city resilience globally and regarding the flooding hazard, since some metrics are hazard-specific. If a city or service intends, the RAF may be used to assess different hazards. In this case, a different RAF study needs to be created for each hazard. For instance, to assess flooding, heat waves, drought and combined sewer overflows, four different studies need to be considered, although they will only differ in the answers related to the CC scenarios metrics. It is possible then to compare and assess the resilience constraints associated to each hazard.

Additionally, for each hazard, the city may have different risk sources/variables associated. For example, flooding might be caused by rainfall or by sea level rise. When answering to the scenarios' metrics for flooding, if there are differences regarding the impacts or consequences that depend on the type of variable, then the answer should be done for the hazard/risk source that causes the most aggravated answer for the metric, and the variable should be specified as a comment to the result.

IV. CONCLUSIONS

Climate change adaptation in cities is a complex endeavor for the multitude of city profiles, adaptation status and services available. Information technologies, in particular based on web tools, can play a leading role in supporting city adaptation and promote a fast adoption of the most relevant options that enhance resilience. RESCCUE RAF App can be a massive step towards into city evolution to cope with CC. Through the

administrator profile, a tailored evaluation system can be built making the app suitable to any urban area. The app is anchored on a structured resilience assessment methodology that helps cities and services to identify the most critical aspects to be improved, supporting a resilience diagnosis and identification of areas for improvement, contributing to a resilient and sustainable urban development. This methodology and its Web implementation can support resilience planning, having a major impact on city managing decisions, and may facilitate communication among different stakeholders, essential requirements to build resilience in a city.

ACKNOWLEDGMENT

RESCCUE has received funding from European Commission by means of Horizon 2020, the EU Framework Programme for Research and Innovation, under Grant Agreement no. 700174. It has also been supported by the Lisboa2020 Operational Program through the INCD project (LISBOA-01-0145-FEDER-022153). This work made use of results produced with the support of the Portuguese National Grid Initiative; more information in https://wiki.ncg.ingrid.pt.

REFERENCES

[1] ARUP (2015). City Resilience Framework. 100 Resilient cities. The Rockefeller Foundation, ARUP.

[2] Cardoso, M. A., Almeida, M. C, Brito, R. S., Beceiro, P. (2015). Modelação integrada da drenagem urbana em zonas estuarinas: avaliação da inundação para cenários climatéricos. 10º Encontro Nacional de Entidades Gestoras de Água e Saneamento , ENEG 2015, Porto, 1-4 de dezembro.

[3] Cardoso, M.A., Almeida, M.C., Telhado, M.J., Morais, M., Brito, R.S. (2018). Assessing the contribution of climate change adaptation measures to build resilience in urban areas. Application to Lisbon. 8th International Conference on Building Resilience, ICBR, 14-16 November, Lisbon, Portugal.

[4] Cardoso, M. A. and Brito, R.S. (2019). Approach to develop a climate change resilience assessment framework. IWA Specialist Conference PI 2019, September 23- 27, 2019, Vancouver, BC, Canada.

[5] David, L.M., Cardoso, M.A., Matos, R.S. (2000). Urban flood control – Technical support to the diagnosis and rehabilitation of critical areas in the City of Lisbon. Flood Issues in Contemporary Water Management, J. Marsalek, W. Ed Watt, E. Zeman and F. Sieker (eds.), NATO Science Series, 2. Environmental Security, Vol.71, © 2000 Kluwer Academic Publishers, The Netherlands, p. 333-342.

[6] EPA (2017). Evaluating Urban Resilience to Climate Change: A Multi-Sector Approach. U.S. Environmental Protection Agency, Washington, D.C., EPA/600/R-16/365F, January 2017, www.epa.gov/research.

[7] Fortunato, A.B., Oliveira, A., and Baptista, A.M. (1999). On the effect of tidal flats on the hydrodynamics of the Tagus estuary. Oceanol. Acta, 22(1), 31-44.

[8] Fortunato, A.B., Freire, P., Bertin, X., Rodrigues, M., Ferreira, J., Liberato, M.L.2017. A numerical study of the February 15, 1941 storm in the Tagus estuary", Continental Shelf Research, 144: 50-64.DOI: 10.1016/j.csr.2017.06.023

[9] Freire, P., Tavares, A., Sá, L., Oliveira, A., Fortunato, A., Santos, P., Rilo, A., Gomes, J. L., Rogeiro, J., Pablo, R., Pinto, P.L. (2016). A local scale approach to estuarine flood risk management, Natural Hazards, doi:10.1007/s11069-016-2510-y

[10] Guerreiro, M., Fortunato, A.B., Freire, P., Rilo, A.R., Taborda, R., Freitas, M.C., Andrade, C., Silva, T., Rodrigues, M., Bertin, X., Azevedo, A. 2015. Evolution of the tidal dynamics of the Tagus estuary in the 21st century. Revista de Gestão Costeira Integrada, volume 15 (1), 65-80. doi:10.5894/rgci515

[11] ISO 9001:2015 – Quality management systems. International Organisation for Standardisation, Switzerland.

[12] Muir-Wood, R., 2011. The 1941 February 15th Windstorm in the Iberian Peninsula. Trébol, 56: 4-13

[13] Pagani, G., Fournière, H., Cardoso, M.A., Brito, R.S. (2018). Report with the resilience diagnosis for each city, D6.1 Project RESCCUE, 2018.

[14] Patel, R., Nosal, L. (2016). Defining the Resilient City. Working Paper 6. United Nations University Centre for Policy Research.

[15] Rilo, A., Freire, P., Tavares, A.O., Santos, P.P:, Sá, L., (2015). Historical flood events in the Tagus estuary: Contribution to risk assessment and management tools. In: Safety and Reliability of Complex Engineered Systems – Podofllini et al. (Eds) 2015, Taylor & Francis Group, London, ISBN 978-1-138-02879-1: 4281-4286.

[16] Sharifi, A. (2016). A critical review of selected tools for assessing community resilience. Ecological Indicators. 69 (2016), 629-647. 10.1016/j.ecolind.2016.05.023.

[17] Summers, J. K., Smith, L. M., Harwell, L. C., and Buck, K.D. (2017). Conceptualizing holistic community resilience to climate events: Foundation for a climate resilience screening index. GeoHealth, 1, 151–164, doi:10.1002/2016GH000047.

[18] UN-Habitat City Resilience Profiling Programme (2018). Guide to the City Resilience Profiling Tool. United Nations Human Settlements Programme (UN-Habitat). Available at: http://urbanresiliencehub.org/wp-content/uploads/2018/07/CRPT-Guide-18.07-Pages.pdf

[19] UNIDSR (2017a). Disaster resilience scorecard for cities. Preliminary level assessment. United Nations Office for Disaster Reduction. United Nations International Strategy for Disaster Reduction.

[20] UNIDSR (2017b). Disaster resilience scorecard for cities. Detailed level assessment. United Nations Office for Disaster Reduction. United Nations International Strategy for Disaster Reduction.

[21] Velasco, M., Russo , B., Martínez, M., Malgrat, P., Monjo, R., Djordjevic, S., Fontanals, I., Vela, S., Cardoso, M. A., Buskute, A. (2018). The RESCCUE Project - RESilience to cope with Climate -- Change in Urban arEas – a multisectorial approach focusing on water – The RESCCUE Project. Water 2018, 10, 1356. doi:10.3390/w10101356.

The advantage of using SWOT analysis for companies with implemented ITIL framework processes

Aida Granulo, Anel Tanović
Faculty of Electrical Engineering, Sarajevo, Bosnia and Herzegovina

***Abstract* - Each company aims to remain competitive in the market and provides the services that their clients seek, all in accordance with the cost-effectiveness and fulfillment of customer expectations. In order to do the same, companies are looking for the best practices that help in organizing their work and delivery of their services, as well as maintaining and determining the competitive advantage. In this paper an example of such practice is explained in the ITIL framework. The purpose of the SWOT analysis is to identify strategies that match the resources and capabilities of the company with the needs of the environment in which it competes. The purpose of this analysis is to use the company advantages, to explore its capabilities, to correct weaknesses, and to counteract the threats of the environment. The aim of the paper is to demonstrate the benefits of SWOT analysis with companies with the implemented ITIL framework, which is mainly reflected in simplicity, flexibility, low cost, but with tremendous efficiency, good estimates, and finding negativity and positivism in business.**

Keywords - ITIL framework, ITIL v3, SWOT analysis, Gap analysis,Balanced Scorecard

I. INTRODUCTION

This paper describes the benefits of SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis for companies that have implemented the ITIL (Information Technology Infrastructure Library) framework. The motivation for writing this paper is to analyze the market and provide guidance to companies, as well as to research the ITIL framework and its practical application.

Section II gives basic theoretical concepts of ITIL and SWOT analysis. Section III shows the implementation of ITIL in the company, and section IV deals with the application of SWOT analysis in the business environment of the company. Section V presents a comparison of the SWOT analysis to the other analyzes and highlights the benefits of the SWOT analysis. A conclusion is reached at the end of the paper in section VI.

II. THEORETICAL BACKGROUND

A. ITIL

ITIL is a 20-year-old commonly used framework for managing IT services and the ITIL structure is a wellspring of good practice in service management. ITIL is utilized by associations worldwide to set up and improve abilities in service management. This paper focuses on ITIL version 3. (V3) ITIL V3 is more attuned to today's market needs. The ITIL framework will help companies develop and improve their overall IT Service Management capabilities, increase organizational cohesion, optimize and demonstrate business value. The framework consists of best practices, procedures and processes to be followed for IT Service Management provision. ITIL provides a valuable body of information to achieve different standards and goals in service management. It was developed in the 1980s by CCTA (Central Computer and Telecommunications Agency, now Central Office of Government Commerce, OGC) at the instruction of the British government. The organization needed to develop a vendor-independent approach to make the most efficient and cost-effective use of IT resources from the Ministries and other UK public sector organizations [1].

The ITIL framework consists of 5 components of the service life cycle. Each component describes the content of the life cycle in detail, and those components cover the life cycle stages. Furthermore, each stage of the life cycle demonstrates how processes are used to improve service management in the organization. [1] It is very important to assign and define roles well within the life cycle. In order for the requirements and results of the process to be fulfilled, it is necessary to share tasks and roles well with individuals [2]. ITIL offers a set of roles: Service Owner, Process Owner, Service Manager, Product Manager [3].

The service lifecycle as mentioned above consists of 5 phases as follows:

1. Service Strategy
2. Service Design
3. Service Transition
4. Service Operation
5. Continual Service Improvement

In total, the ITIL framework consists of 5 phases, 26 processes [4].

Service Strategy phase represents the strategic approach to managing an organization's services and the core of he life cycle. This phase explains who the customers, services offered to meet customer expectations. It also explains IT capabilities and resources

needed to develop offers and requirements for successful execution [2], [4].

The Service Design phase offers guidelines for designing services based on customer requirements and guarantees that the services are designed to meet their expectations. In addition, the processes required to manage services are also part of this phase [2], [4].

The Service Transition phase enables the transition of services to the environment. This phase binds to change management, asset control, and configuration items (hardware, software, etc.) [2], [4].

Service Operation manages the day-to-day delivery of services, including optimizing efficiency and effectiveness. This phase provides the service on an ongoing basis, monitoring the daily overall behavior of the service [2], [4].

Continual Service Improvement is a phase that provides guidance for maintaining the created service, but also for its continued alignment with business needs. Offers a mechanism to measure and improve the level of service, technology, effectiveness and efficiency or processes used in overall service management [2], [4].

TABLE I. ITIL LIFECYCLE PROCESSES

Service Strategy	**Service Design**	**Service Transition**	**Service Operation**	**Conitnual Service Improvement**
Strategy management for IT services	Design coordination	Transition planning and support	Event management	Seven-step continual service improvement process
Financial management for IT services	Service catalog management	Change management	Incident management	
Demand management	Service level management	Service asset and configuration management	Problem management	
Service portfolio management	Capacity management	Release and deployment management	Request fulfillment	
Business relationship management	Availability management	Service validation and testing	Access management	
	IT service continuity management	Change evaluation		
	Information security management	Knowledge management		
	Supplier management			

B. SWOT analysis

SWOT is an acronym for Strength, Weakness, Opportunity, and Threats. It is often used to evaluate a business or a proposal. It is especially used when future progress is to be made. The process involves defining the strengths and opportunities that an organization has against the weaknesses and threats of the market [5].

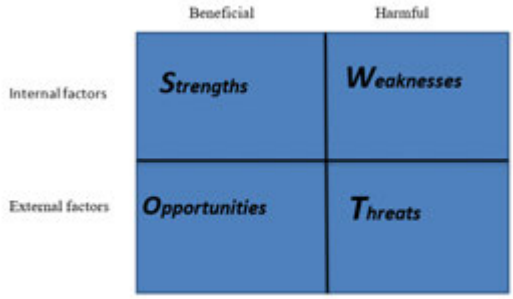


Figure 1. SWOT Analysis

SWOT analysis is a customizable and very popular tool, but it uses a lot of subjective opinions and decisions and it is good to use it as a guide. Also, there is no template and default analysis to be used, but each analysis is specific and depends on the goal to be achieved. Internal factors: strengths and benefits are those that can be measured and can be influenced in some way. But external factors: opportunities and threats are factors that cannot be influenced.

Strength- internal factors favorable to achieve the goal, factors that are re currently good and need to be maintained, but used as a foundation

Weaknesses - represent internal factors that are not favorable to the goal, bad practices, factors needed to be replaced or repaired

Opportunities - Represents external factors that are favorable for the goal, factors good for the future, needed to be prioritized, it is good to have and save them, optimize and use in improvements.

Threats - Represents external factors that are not good for the goal, are bad for the future and should be avoided or overcome [5].

Depending on the organization, but also the goal, the strengths and weaknesses could be changed. It is a dynamic process where decision making and thinking need to be involved. For the purpose of creating the SWOT analysis, meetings are organized which are attended by people of interest, with competent knowledge and skills. Sometimes, it is difficult to reach factors that fall under weaknesses, because people in an organization find it difficult to talk about the them. Participants should be encouraged to provide as much information as possible. The strength of SWOT analysis is that it can be used for different organizations, and the disadvantage is that it requires clear and smart thinking as well as good judgment, decision making and value creation [5].

There is no exact number of steps in the analysis process. In this paper, the process of SWOT analysis will be explained through four simple steps:

1. Select a target, collect and evaluate key

2. Classification of data into external and internal categories, respectively into 4 groups of factors

3. Creating a SWOT matrix

4. Using analysis to make strategies and decisions

SWOT analysis is a good defense mechanism for any company [6], [7].

C. GAP analysis

Gap analysis is a technique of comparing current conditions with the desired ones in order to identify the gap. Thus, it serves to find and analyze the gap between the current and desired state.

Gap = Preferred Performance - Current Performance

After identifying the gap, it is possible to move on to planning the solution. Steps in conducting the gap analysis include:

- Determine the purpose and topic of the analysis, as well as define organizational goals
- Determine your current position based on metrics or attributes
- Determine the desired position within a specific time frame
- Determine the gap between the two positions
- Create a report based on quantitative data collected, qualitative reasons why the data is below the benchmark, and identify the action elements needed to achieve the organization's goals [8]

D. Balanced Scorecard

In the early 1990s Robert Kaplan and David Norton, developed Balanced Scorecard technique and singled out the fact that companies base their business management on financial measures as the main business problem. However, Kaplan and Norton recognize the value of information in decision making and for this reason create the Balanced Scorecard technique. Balanced Scorecard is a management system that enables a company to set, monitor and achieve key business strategies and goals. It consists of 4 perspectives that relate to vision and strategy:

Financial, User, Internal business processes, Learning and development [9].

III. IMPLEMENTATION OF ITIL FRAMEWORK IN DISNEY COMPANY

The ITIL framework has been implemented in various companies around the world. Some of these companies are: Ohio State University, Yale University, AXA Insurance Group, Disney, Honda, Sallie Mae, Shell Oil, UnitedHealth Group, Visa, Australia Post, Bank of Canada, Bank of Montreal, Cebra, Harris Bank, British Airways, British Telecom, CERN, Müller Dairy, Telkomsel, PT Telekomunikasi Selular [10].

Implementing the ITIL framework in a business environment requires certain actions, knowledge and skills. Pink Elephant is a company that helps in implementing ITIL framework to other companies. Usually, the first step is to evaluate the benefits of the ITIL framework and the processes in which it can be applied. So, the processes most commonly applied to ITIL are problem management or change management, incident management, but also financial accounting, supply chains and capacity management. After that, the current situation in the company is documented and process assessments are made. Setting goals to create successful strategies is the next step. After selecting a project management methodology and implementing the ITIL framework, progress should be monitored, and goals met [11], [12].

The implementation of the ITIL framework brings changes in the company, and the changes affect technology, employees, human resources management. Implementation should improve business results, IT process efficiency, which often leads to reduced IT costs and service delivery [11], [12].

An example of a European company that has adopted ITIL is CERN (European Organization for Nuclear Research). Implementation of ITIL started with a service catalog describing the services available. CERN has embraced best practices from ISO 20000 and ITIL V3 to run a ServiceNow driven facilities management technology database that serves 25000 people on its campus. ISO/IEC 20000 provides a standardized and uniform framework for companies that are pursuing evaluation and validation of their service management capability. Although ISO/IEC 20000 is a standard to be followed and upheld, ITIL provides a valuable body of information to meet the standard [13].

The Disney company was officially formed in 1923 by the Walt Disney brothers and Roy O. Disney. Company is best known for its products in the movie studio, Walt Disney Studios, one of the largest and best-known studios in American cinema. But the company has 5 segments: Media Networks; Parks, Experiences and Products; Studio Entertainment; and Direct-to-Consumer and International [14].

In mid-2000, the company began adopting the ITIL framework. But in 2008, Glen Taylor, the company's vice president and chief information officer, brought ITIL adoption into the company. Taylor has previously been a proponent of this framework. He says that "ITIL aligns IT with business, and our goal was to become a more proactive IT organization. We were looking for an improved level of service. ITIL helped provide tools and metrics to define the value of IT services. " [15].

As the main goal was to meet customer expectations, for IT it meant accessibility, reliability and sustainability. For the company, this meant that it had to ensure that broad change did not result in incidents. It is also necessary to increase and ensure the management of releases and new capabilities as well as that outsourcing contracts are managed with professionalism. ITIL is the solution to all these challenges [15].

The implementation of the ITIL framework in this company has been carried out through three main steps: presenting ITIL to employees, training employees and selecting experts. Introducing the ITIL framework to employees means getting them familiar with the framework. The second step involved training employees. The training was attended by 250 employees and training sessions of 2 to 3 days are organized, as well as those specific to Disney and its use of ITIL, or its application. [15].

Finally, in the last step, 20 experts from across the segment were selected. They were selected for the ability to adopt and understand information provided by ITIL, a character that can persuade and influence people, work in a team and fight resistance, the ability to articulate a vision for the processes they manage or operate, and to reap ITIL benefits.

Tips from this company regarding the implementation of the ITIL framework relate to the importance of communication, which is crucial for them to succeed, utilize existing tools and documents. They also advise that you should not waste time on things that already exist and on the importance of uncomplicated and practical behavior [15].

IV. DISNEY COMPANY SWOT ANALYSIS

Below is a SWOT analysis of the Orlando resort (Disney Park and Resort). The SWOT analysis is applied to the process of continuous improvement.

Strengths

- The most visited entertainment center in the world
- Largest entertainment center
- Diversity in topics such as Magic kingdom, Hollywood studios
- Great benefit from the brand

Weaknesses

- The size of the center prevents visitors from visiting it in its entirety
- A large number of hotels and entertainment centers make centralized management and monitoring difficult

Opportunities

- The Disney brand name makes marketing campaigns easier
- The company has experience in managing resorts and entertainment venues and is extremely popular around the world
- As travel abroad is popular, it is possible to attract people from all over the world
- Many promotions and discounts attract customers
- The introduction of new attractions and facilities attracts many users

Threats

- High operating costs
- Local competitors are gaining in popularity
- Competitors are slowly taking market share if the company fails to create innovation constantly [16]

V. COMPARISON OF SWOT ANALYSIS TO THE OTHER ANALAYSIS

A. *SWOT vs. GAP analysis*

An example of the gap analysis is presented below:

TABLE II. GAP ANALYSIS

Objectives:	Achieved:
Family fun	Kids and teen parties
Honesty and Integrity	Money driven business
Friendly and imaginative	Friendly and imaginative
Creativity and innovation	Creativity without innovation
International brand	International brand

[17]

After the analysis, certain strategies for improving the business were elaborated. But the main conclusions are about improving and innovation. As competitors appear in the market every day with innovation, so does this company need to keep up with them. The company itself must work on the definition of innovation, but also on its realization. Also, the problem was how to innovate rather than spoil tradition, and it was necessary to create with these guidelines.

The gap analysis is quite cumbersome and provides less information than SWOT analysis does. It has been proven that SWOT analysis really gives a complete picture of the company in the market, but also a picture of its competitors so that the company can create better strategies. The gap analysis only identifies shortcomings that could be further analyzed by SWOT analysis and is more suitable for smaller processes. SWOT analysis can minimize deficiencies, increase power, eliminate threats and seize opportunities. SWOT analysis is used for long-term planning, while gap analysis corresponds to short-term goals. SWOT analysis assesses many aspects and competitors and, as such, is a comprehensive study, while gap analysis is focused on a single process.

Thus, gap analysis is there to identify the weaknesses and poor utilization of the potential that the company has, while SWOT analysis facilitates activity planning and decision making. Collaboration and visualization play an important role in presenting both analyzes visually.

The gap analysis is there to reveal gaps and weaknesses, while SWOT analysis is a key process for finding what needs to be done to reduce it. Some authors point out that the gap analysis is only an analysis or comparison of current strengths and weaknesses. As such, it provides only one piece of information that is opposed to it by the SWOT analysis. When used in the ITIL framework, both are put in the same context only when analyzing the company on the question "Where are we now?".

The biggest advantage of a SWOT analysis is gaining advantage over competitors emphasizing strengths and acting to correct weaknesses or address potential threats.

B. SWOT vs. Balanced Scorecard

An example of a Balanced Scorecard analysis for Disney is given [18].

TABLE III. DISNEY'S FINANCIAL PERSPECTIVE

FINANCIAL PERSPECTIVE		
Objective	Measurement	Target
Company assets	Increase in assets	Number of new assets submitted for each year
Work efficiency	Increase work efficiency	Obtain ISO certification
Annual revenue	Increase in annual revenue	Increase in revenue
New Investor	More Potential Investors	Increase the number of investors

TABLE IV. DISNEY'S CUSTOMER PERSPECTIVE

THE USER'S PERSPECTIVE		
Objective	Measurement	Target
Increasing stakeholder satisfaction	Meeting the expectations and needs of stakeholders	Holding all the expectations and needs of stakeholders
New users	Gaining more users	Constant increase
User loyalty	Keeping customer loyalty	Old user data saved
Improving customer service	Quality customer service	Achieving customer satisfaction

TABLE V. DISNEY'S INTERNAL PROCESS PERSPECTIVE

PERSPECTIVE ON INTERNAL PROCESSES		
Objective	Measurement	Target
Retention of assets and good employees	Extension of contract to good employees	Retention of all good employees
Strengthening the improvement of the management process	Performance evaluation	Magnifying enhancement graphics
Administration systems	Countable administration systems	Weekly reports on administration systems
Rapid Employee Progress	Changing employee positions to higher levels	Annual number of new employees

TABLE VI. LEARNING AND IMPROVEMENT PERSPECTIVE

A LEARNING AND IMPROVEMENT PERSPECTIVE		
Objective	Measurement	Target
Sustainable technical aspect	Meeting technical standards	Achieving ISO certification
Sustainable Human Resources Aspect	Meeting minimum standards	Obtaining certification
Developing employee skills and productivity	Employee skills and productivity	Getting certified
Healthy environment and relationships	Strengthening relationships	Involvement of all employees

This kind of analysis leads to the creation of a plan that can be acted upon. So, from a financial perspective, the action plan calls for spending money to buy new assets, standardizing every aspect of work, reducing unnecessary costs. The user perspective, as an action plan, requires creating the ability to propose changes anonymously and in an easy and accessible way (creating a suggestion box), creating a Customer Club, rewarding users.

The Internal Process Perspective proposes to increase pay, create reports for each transaction, evaluate employees. The perspective of learning and improvement suggests meeting the standards of work, organizing courses, but also shared vacations, trips, or team building.

As noted above, SWOT analysis is useful in broader planning, while the Balance Scorecard is a tool that helps achieve a specific goal. The main difference that needs to be pointed out regarding these two analyzes is their use. Namely, SWOT analysis is used to define a goal and Balance Scorecard is is a tool that helps achieve a specific goal.

Both methods are relatively inexpensive to design and implement, and they both provide valuable insight into the key aspect that will determine an organization's success. Although differences between these methods are often sought, they have been proven to be complementary and can be used together.

In the ITIL framework, Continual Service Improvement process more specific, Balanced Scorecards is used for metrics, measurements, or answers to the question "How do we know we have reached a certain goal or stage?" So, it refers to the part where monitoring, measuring success and acting are performed.

It was mentioned earlier that the SWOT analysis answers the question "Where are we now" and analyzes the current situation. SWOT analysis helps identify key strengths, weaknesses, opportunities and threats. Also, the analysis gives results in terms of strategic planning, such as the market, customer needs, et cetera. In this way, it enables the company to analyze the current situation and current positioning in relation to competition, market conditions.

SWOT analysis is also a step that precedes the creation of Balanced Scorecard analysis. In most cases, weaknesses, opportunities, and threats will provide the most insight into what the goals of the business of the company should be. All of this helps the company identify and implement improvement initiatives that try to identify the most important areas where performance is not at the desired level. It is important to emphasize that periodic review of deficiencies in order to check progress towards goals is crucial.

Also, the Balanced Scorecard is not able to decide on the strategies to measure and analyze. Since explaining an association's vision and system is the initial phase in leading a similar investigation, it is conceivable to characterize a lot of procedures to break down and center around utilizing SWOT examination. By connecting SWOT examination to Balanced Scorecard investigation,

an association can adjust its qualities against shortcomings and advance its market openings.

SWOT examination is useful here in light of the fact that it will assist us with nailing down the reasons why we need to embrace a best practice, what are our present difficulties to be tackled through this, what are the limitations that cutoff our capacity, what is at present conceivable and what benefits we will harvest on the off chance that we are successful. An overview of the comparison of analyzes is given in Table VII [19].

TABLE VII. COMPARISON OF ANALYZES

SWOT	GAP	Balanced Scorecard
Evaluate long-term opportunities	Address and identify short-term threats and needs	Balance between short-term and long term goals
Focuses on external factors and evaluates competitors	Used internally	No focus on competitors and external factors
Understanding relative market value in a measurable way	Reveals gaps	No focus on competitors and external factors
Comprehensive analysis	Focused on a single process	Focus on both high-level strategy and low-level measures
Helps with identifying a company's value	Helps to identify if a company is performing to its potential	Visual picture of the strategy; tool for accomplishing goals

VI. CONCLUSION

The aim of the paper is to show the advantages of SWOT analysis for companies with implemented ITIL framework. One of the benefits of SWOT analysis is identifying potential pitfalls while allowing the company to react on time and take the necessary actions to avoid them. The biggest advantage of SWOT analysis is gaining an edge over competitors by emphasizing strengths and acting to correct weaknesses or address potential threats.

SWOT analysis is comprehensive study and does not necessarily have to be focused on one process. Thus, it facilitates activity planning and decision making. Also, this analysis is a tool for defining a goal, but not for achieving it. This analysis can help investigate the impact of introducing new services and products, easily apply new business processes and functions. It is also useful in assessing the effectiveness of technologies, markets, and suppliers.

It is appropriate for companies that have an ITIL framework implemented. The framework clearly identifies and outlines the processes by which SWOT analysis helps facilitate decision making and pave the way for strategies.

ITIL provides better visibility of costs and assets, better business alignment, cost reduction, better risk management, stable support service environment, better knowledge of the process by employees. And when SWOT analysis for any service is added to these benefits, the change is clear and company will get an accurate analysis of the market and will be able to better match business with the market and with customer satisfaction.

A small number of companies in the region have implemented ITIL. Future work is based on the idea of creating a manual that would give clear guidelines for the implementation of ITIL in a small specific local business, as well as the use of SWOT analysis after its implementation with the aim of creation another case study.

VII. REFERENCES

[1] Valerie Arraj, "ITIL: the basics," Compliance Process Partners LLC, 2013. [Online]. Available: https://www.vanharen.net/Player/eKnowledge/itil_the_basics[100].pdf. [Accessed 23 7 2017].

[2] Liz Gallacher, Helen Morris, ITIL Foundation Study Guide, Chichester: John Wiley & Sons, 2012.

[3] "ITIL® v3 Foundation Study Guide," 2009.

[4] Alison Cartlidge, Ashley Hanna, Colin Rudd, Ivor Macfarlane, John Windebank, Stuart Rance, An Introductory Overview of ITIL v3, itSMF, 2007.

[5] FME Team, SWOT Analysis, free management e books, 2013.

[6] J. P. Harrison, Essentials of Strategic Planning in Healthcare, Health Administration Press, 2010.

[7] "Make SWOT Analysis in 4 Easy Steps," [Online]. Available: http://www.entrepreneurshipinabox.com/763/swot-analysis-4-steps/. [Accessed 10 8 2017].

[8] R. Kelly, "5 Simple Steps On How To Do A Gap Analysis," 8 11 2009. [Online]. Available: http://robdkelly.com/blog/getting-things-done/gap-analysis/. [Accessed 5 8 2017].

[9] P. R. Niven, BALANCED SCORECARD STEP-BY-STEP, Hoboken, New Jersey: John Wiley & Sons Inc., 2006.

[10] "Who's doing ITIL," [Online]. Available: http://itsmtransition.com/downloads/whos-implementing-itil/. [Accessed 21 7 2017].

[11] "How to implement ITIL," [Online]. Available: http://www.wikihow.com/Implement-ITIL. [Accessed 7 21 2017].

[12] PinkElephant, "Top 10 Things Managers Should Know About Implementing ITIL," 2005.

[13] C. Saran, "CERN uses ITIL to rethink facilities management," 29 May 2013. [Online]. Available: https://www.computerweekly.com/news/2240184788/Cern-uses-ITIL-to-rethink-facilities-management. [Accessed 24 May 2020].

[14] "The Walt Disney Company," [Online]. Available: https://thewaltdisneycompany.com/about/. [Accessed 5 8 2017].

[15] G. Taylor, "Disney's ITIL Journey," [Online]. Available: https://www.axelos.com/CMSPages/GetFile.aspx?guid=7eda3209-8364-4fd6-85d9-74a18b9c0bf3. [Accessed 2 8 2017].

[16] L. Peters, "Brand Strategy: Disney," 2012.

[17] L. Caroll, G. Dowling, L. Folkard, W. Hsuan Lie, P. Saville, B. Szechy , "Disney Land".

[18] Rama P. Renspandy, Randa Tio Alexi, "Balanced Scorecard".

[19] "SWOT vs Gap Analysis: Strategies to Analyze Your Business's Health," Lucidchart, [Online]. Available: https://www.lucidchart.com/blog/gap-swot-analysis. [Accessed 24 05 2019].

TECHNOLOGIES

Smart Contracts as a diploma anti-forgery system in higher education - a pilot project

D. Čeke [1,2] and S. Kunosić [3,4]
[1] Department of Software engineering, Faculty of Polytechnic, University of Zenica,
[2] IT Department, University of Tuzla, Bosnia and Herzegovina,
[3] Department of Physics, Faculty of Natural Sciences and Mathematics, University of Tuzla, Bosnia and Herzegovina,
[4] Faculty of Health Studies, University of Bihać, Bosnia and Herzegovina
denis.ceke@untz.ba, suad.kunosic@untz.ba

***Abstract* - The problem of counterfeiting diplomas in education with the advancement of digital technology is increasingly pronounced. The process of forging documents is almost always accompanied by reduced transparency in issuing the documents with no possibility to easily check the validity of the document. One of the currently very attractive and challenging technology in digitized society is the blockchain technology and all the sequential systems that have emerged based on it. One such system is the Ethereum platform, which uses blockchain technology and enables the creation of decentralized applications programmed to run on the Ethereum network. One of the Ethereum use is a Smart Contract, which allows applications to be executed online, completely autonomously without the influence of a third party, once a previously defined condition is satisfied. The objective of this research is to explore the possibility of using a Smart Contract in the process of the creation and issuance of diplomas at a higher education institution. At the end of the analysis, we provide an overview of the advantages and disadvantages of this procedure, as well as potential possibilities for its improvements. The possibilities of automation and the cost of such a process were also considered.**

Keywords – blockchain; Ethereum; Smart Contract; diploma; forgery

I. INTRODUCTION

Since the publication of the first article describing the blockchain technology by Satoshi Nakamoto in 2008 [1], this technology has been constantly evolving. In addition to cryptocurrency trading, for which this technology was created, the possibility of expanding the use of this technology has been constantly explored because of its several crucial advantages, such as the no need for the central authority in any type of application or transactions that run on this network and very good security of the network reflected in the inability to break through or compromise the data transmitted over it. Thus, on January 25, 2014, Vitalik Buterin [2] formally introduced the Ethereum platform [3], with the support of the blockchain technology to create smart contracts and transfer cryptocurrencies to other users by validating these contracts in the form of decentralized applications, thereby expanding capabilities of blockchain technology in areas other than paying for services, or sending and receiving cryptocurrencies through the blockchain network.

After a short period, decentralized applications, or popularly called *Dapps*, began their major development phase. The first major use was in the area of creating a special type of decentralized applications called s*mart contracts* (SC). The very name "s*mart*" implied that these applications had some form of "*smart*" management each time they were executed. According to [4], a smart contract is a self-checking, self-executing, and anti-fraud computer program. Smart contracts aim to negotiate negotiations between strangers by eliminating the need for a third party to audit the negotiation.

This type of application has expanded the classic use of blockchain technology in that way that now is possible to run different types of applications on a dedicated network, called the Ethereum network [3], decentralized, without intermediaries, distributed, with a high degree of security.

Some examples of SC usage can be found in the legal sphere, education [5][6], healthcare [7][8][9][10], for creating real estate registers [11] and in the are of IoT technology [12][13], supply chain systems [14][15] and for creation and storage of the various types of certificates that individuals can obtain in an education or training process [16].

Ethereum network uses the term transaction to describe the data packets that are carriers of information or messages that are transmitted between users. Each transaction has its basic elements, such as the message body for the recipient, the signature identifying the sender, the amount of Ether to transfer or to pay, and the data to be sent. In addition to this information, each transaction also contains two additional information, gas price (value paid in Ether by a unit of gas spent) and gas limit (maximum amount of gas to use) [17]. Every transaction that takes place on the Ethereum network has a price, that is, the cost of doing a transaction that is named "gas". The price of gas varies depending on the operation that needs to be done online. For example, the cost of publishing a smart contract on the Ethereum platform costs 32,000 gas, while storing 256-bit data value costs 20,000 gas. At the moment of writing this article (January 2020.), 1 ETH (Ethereum currency) had a value of $161.31 [18].

Each transaction has its unique identification number (network address) that is unique on the network and is

never repeated. In this way, the uniqueness and security of each data on the network are ensured because there is no possibility to overwrite the existing data in the Ethereum network with new ones, but with each transaction, a new address with data is created.

II. Related Work

As this is an area that is still under research and it is relatively new, there is a lack of relevant work that discusses some possible implementations of such systems in the field of education. Namely, most of the papers that talk about the applications of the Ethereum network are in other areas outside of the education, such as real estate, logistics, IoT and healthcare, mostly. Possible reasons for this situation probably lie in the fact that education is not an industry, that education does not produce a commercially usable product, and therefore the interest, at least at the moment, of the research community is more towards the commercial applications of this technology. Nevertheless, there are some proposals for the possible use of blockchain technology in the field of education, namely, in the process of awarding academic diplomas.

For example, the University of Nicosia [19] implemented blockchain technology-based solutions for the certificate registration process for students who completed the DFIN-511 *Introduction to Digital Currency* course. On the other hand, the Massachusetts Institute of Technology (MIT), in collaboration with a company called the Learning Machine [20], provides students with the opportunity to use the capabilities of the bitcoin blockchain network for academic credentials.

The Open University and the Knowledge Media Institute [21][22][23] also can place student graduation information on the blockchain network. Their approach has one additional possibility that it is for each student to create a so-called *ePortofolio* that includes information on all completed and required tasks that students need to complete to qualify for a diploma.

Another solution focused on integrity, proof of existence, and privacy of academic records is implemented through the system called *Disciplina* [25]. Unlike other solutions, in this case, the data is modeled in the form of transactions placed on a private blockchain. The public blockchain serves only as an additional module to house the data related to which the data is linked to the private blockchain.

Sony Corporation also uses a private blockchain to secure the data [26]. A detailed view of the implementation of the Sony solution can be found in Zhang's work [27].

Some of the currently active implementations have been developed by some of the universities or companies, and many of them base their examples on using the bitcoin blockchain network infrastructure to secure the digital credentials of the user and issuer of digital documents. However, since the bitcoin blockchain network is primarily intended for money transactions, it is suggested that it is not advisable to place information other than financial transaction data on it. It is even mentioned that over time, any online transactions that are not of a financial type will be removed because the network is not designed to host this kind of data [28] [29] [30]. For this reason, our approach to solving this problem is focused on creating Ethereum smart contracts, because, by their structure, implementation, and capabilities, they are currently the most optimal solution for the problem that is the topic of this paper.

The proposed solution presented in this paper is based on the simple development and implementation strategy with the main aim to reduce the quantity of data that needs to be stored on the Ethereum network. The main reason for this approach is motivated by the fact that a large amount of data requires a lot of network resources, financially and physically. The integration of the smart contract with the student registry application within the university information system was performed so that student's data is retrieved directly from the system. Once the students have fulfilled the requirements for obtaining a diploma, the system activates the option for the university administration personnel to publish the diploma to the Ethereum network.

III. Proposed System Design

Solidity is a specially created programming language for the development of Smart Contracts [31][32]. To facilitate the SC development process, integrated development environments (IDEs) have been created that are specifically designed for this purpose. Thus, to create, compile and test SCs programmers can use an online IDE called Remix IDE [33][34]. For the best practice it is recommended to test developed smart contracts in the local development environment because apart from the higher speed of development and testing, transactions are only paid with virtual money (gas) that has no real currency coverage. The local Ethereum network is implemented using the Ganache tool [35], which allows us to run a virtual Ethereum network on a local machine and test our SCs much faster and more efficiently. Before publishing smart contracts on the main Ethereum network, there is another additional testing option that allows us to get an even more realistic picture of how our smart contract will work in a real environment. We can publish SC to one of the free Ethereum test networks such as Ropsten [36], Kovan [37] or Rinkeby [38].

All of these individual elements for the development of *Dapps*, writing SC source code in the Solidity language, compiling smart contract, test it, and finally publish it to our local and/or Ethereum test network, have been integrated into a single framework, a Truffle suite [35].

To make our users have a simple user interface (UI) for the interaction with SC, a custom web application has been developed. This web application is specially designed for this purpose. It has a couple of modules. The first module is for authorization and authentication of users. An authorized user can create a digital document with data related to a student's graduation information and they can publish information on the Ethereum network. The second module is displayed after the user logs on to the system. The user receives information for all graduate students sorted by study programs. The third module is a

link between the web application and the *Dapp* system whose main task is to prepare data for SC, connect to the Ethereum network and to provide infrastructure for the SC publishing process.

Upon completion of the SC publication process on the Ethereum network, an address on the network containing information about students' diplomas is recorded for each one of them in the local system.

A web application was developed using the Laravel framework [39] which additionally uses javascript elements, such as Web3.js [40] [41] library so that it can interact with SC.

A. Data used for system testing

For system testing, data about graduated students of the Faculty of Electrical Engineering, all study programs, University of Tuzla were used. The test set consisted of data on 33 graduates who completed their studies in the academic year 2018/2019. Data were retrieved from the university information system in the JSON (*Java Script Object Notation*) format (Fig. 1).



Figure 1. Generated *JSON* file with students data from university information system (excerpt)

Students' data is loaded into a web application which then generates a list of students for who we need to create a smart contract or whose data will be placed in the university's decentralized application on the Ethereum network.

In the generated list of data about students in a web application, each student data set in the list has a *Publish* button that is active if the contract has not yet been published online, or inactive if the contract has already been published online. This option is made available by comparing data in a local JSON file, using unique student code, and data contained in the SC published on the Ethereum network. This ensures that there is no duplication of data on the network because there is no possibility to initiate an action twice to publish data to the network.

B. Smart Contract

A smart contract consists of two parts. The first part is a diploma registry (*DiplomaRegistry*) which houses information about all published contracts and acts as some type of array of data. The second part is the *Diploma* module itself, which is practically a data structure with its methods and attributes that forms the basis for the creation of a single contract. Each time a new contract is created, it is placed in a diploma registry that also has the methods that make it easier to retrieve information about published contracts. Source code excerpt of the *DiplomaRegistry* and *Diploma* contract is shown in Fig. 2 and Fig. 3, respectively.

```
pragma solidity ^0.5.0;

contract DiplomaRegistry {
  address[] public registeredDiplomas;
  event ContractCreated (address contractAddress);

  function createDiploma(uint _sifStud, string memory _imeStud,
  string memory _prezStud, string memory _datumDiplomiranja,
  string memory _nazivZvanje, string memory _nazSkGodina) public {

  address newDiploma = address (new Diploma(msg.sender, _sifStud,
  _imeStud, _prezStud, _datumDiplomiranja,  _nazivZvanje, _nazSkGodina));

    emit ContractCreated (newDiploma);
    registeredDiplomas.push (newDiploma);
  }

  function getDeployedDiplomas() public view returns (address[] memory) {
    return registeredDiplomas;
  }
}
```

Figure 2. Source code of *DiplomaRegistry* contract (excerpt)

The *DiplomaRegistry* contains several key elements that enable it to do its part of the job. First, an array of data is defined which is used to store data about all published diplomas. This array is an address data type that represents a special type of data created within the Solidity programming language and is solely intended to hold data about addresses on the Ethereum network.

Thereafter, an *event* type of function (event broadcast) is created after the diploma contract creation function (*Diploma*) is called and its purpose is to serve solely as an element of communication to the user environment where it is thus recorded that the SC publication event occurred. And finally, within *DiplomaRegistry* we have the *createDiploma()* function, which does the final part of the work, publishing diploma data to the Ethereum network.

Diploma contract, on the other hand, is a class object that smart contracts, *DiplomaRegistry*, will use as the template for instantiating future diploma contracts on the blockchain. Within the *Diploma* contract, global variables are defined, such as owner's address, and student's data, like students_id, first name, and last name, date of graduation, academic title and title of the academic year when the diploma was obtained.

C. The possibilities for the automation of the process

Process automation is implemented at the application level. Therefore, the data is not manually entered into the smart contract but is generated from the university's information system and exported as a JSON file.

The JSON file is then uploaded to the web application for publishing the contract, and based on its contents, a list

of students with their graduation data is generated. Within that list, an authorized system user (who is an authorized person on behalf of the university) cannot modify the generated data but only can click the *Publish* button and publish data to the blockchain network. In this way, the system is made to be error-prone (no errors can be made due to the absence of hand-filled forms) because there is no intermediary in transferring data from the information system to the blockchain network.

The rest of the work is done by the SC itself. The smart contract is called from the web user environment and after completing the publishing process, the list of all students whose graduation information should have been published online is reloaded, except that the *Publish* button is now disabled for those students for whom the publishing process has been successful. This way we have managed to make almost complete automation of the process, retrieving data from the information system and publishing to the Ethereum network.

```
contract Diploma {
  address public owner;
  uint sifStud;
  string imeStud;
  string prezStud;
  string datumDiplomiranja;
  string nazivZvanje;
  string nazSkGodina;

  modifier onlyOwner() {
    require (msg.sender == owner);
    _;
  }
  constructor (address _owner, uint _sifStud, string memory _imeStud,
  string memory _prezStud, string memory _datumDiplomiranja,
  string memory _nazivZvanje,string memory _nazSkGodina) public {
    owner = _owner;
    sifStud = _sifStud;
    imeStud = _imeStud;
    prezStud = _prezStud;
    datumDiplomiranja = _datumDiplomiranja;
    nazivZvanje = _nazivZvanje;
    nazSkGodina = _nazSkGodina;
  }
```

Figure 3. Source code of Diploma contract (excerpt)

D. *Advantages of using smart contracts*

There are numerous advantages of using Smart Contract and documents created in this way, which, among other things, become transparent, digitally protected from any form of manipulation or modification, without intermediaries, permanently stored on a distributed digital network. Transparency, security, and inability to change a contract and data once created (immutability) are some of the basic elements when it is necessary to determine the validity of official and public documents, like educational certificates, diplomas, notary contracts, etc.

E. *Disadvantages of using smart contracts*

Currently, some of the current shortcomings of SC use could include, among other things, the legal status of such applications and documents archived in such systems (uncertain legal status) in most countries, since there is no legal regulation that treats SCs and their legal strength. Development and implementation costs also present some of the disadvantages, since SCs cannot be created without knowledge of programming. If there are errors or security bugs in the application, their usage can make a big financial loss and jeopardize the very purpose of the existence of such applications.

IV. RESULTS

The process of creating, testing, publishing and using SC is time-consuming and financially burdensome. Each contract published on the Ethereum network has its own cost and execution time. After application testing, we have observed that the publishing time on the local Ganache Ethereum network, which is on the local computer, is very short and acceptable for testing smart contracts, and lasts for a few seconds with almost instant information feedback for the published contract.

When contracts were published on the *Ropsten* test network, the time of publication of contracts varied, from 58 to 126 seconds, although it was the same contract with the same content and the same data types as it was on local Ganache network.

The cost of publishing a contract is usually higher than the cost of using it because the contract is initially more programmatic in terms of content (it has variables, functions, constructors, etc.), and it is published only a small number of times (usually only once if everything is properly set up and programmed). The use and/or invocation of functions from the contract have a lower cost, and therefore the time of execution because essentially only one function is called.

Therefore, it is important to notice that using SC involves at least two steps: publication of SC on the Ethereum network and using an SC by calling its functions.

The cost of performing any activity on the Ethereum network (publishing contracts, interacting with contracts, transactions) is paid in a cryptocurrency called Ether. Every operation performed on the Ethereum network has its price, which is expressed in units called "gas". Gas is an abstract number that represents the relative complexity of operations. It is important to note that all transactions cost 21000 gas.

For example, if a transaction involves only money transfer without interacting with the contract, then the price of that transaction is 21000 gas. However, if there is also an interaction with the contract, then the price of that entire transaction is 21000 gas plus the additional amount of gas associated with the interaction with the contract.

The amount of gas is fixed for a particular operation, while the price of gas, *gas_price*, is dynamic and is determined by its market value. *Gas_price* is a value that represents the amount of Ether the user is willing to pay per gas. When a user initiates a transaction, he specifies gas_price in *Gwei* / Gas units (1 *Gwei* is equal to 0.000000001 ether). The total cost the user will pay is equal to the product *gas_price x gas_used*. The higher the amount of *gas_price* we are willing to pay, the faster the transaction will be completed. Separation of *Gwei* units from Ether allows maintaining an indirect link between

the current market value of cryptocurrency and the cost of a transaction.

At the time of writing this paper (January 2020), the median market price of Ethereum was 20 Gwei while the current USD / ETH exchange rate was $ 161 / ETH (United States Dollar - USD $).

The cost of publishing the contract (*transaction fee*) on the *Ropstein* test network was 0.001264819 Ether, which is equivalent to $ 0.204800. The average time it took to complete this transaction was about 40 seconds. The average estimated time to publish the SC online and the corresponding prices are shown in Table I.

The average cost of data storage (*transaction fee*) in the *Ropstein* test network was 0.000634878 Ether, which is equivalent to $ 0.103103. The average time it took to complete a transaction was about 20 seconds. The average estimated time for data to be published online and the corresponding prices are shown in Table II.

It is important to emphasize that time it takes for the diploma to be published on the network depends significantly on the network load because even if we offer a higher price for publishing our contract, it simply will not physically be possible to complete the transaction quickly due to the possible current network load.

The estimated total cost to publish 898 diplomas issued by the University of Tuzla to all students in the academic year 2018/2019, together with all costs, time and financial, is shown in Table III.

TABLE I. THE ESTIMATED PROCESSING TIMES FOR INITIAL CONTRACT PUBLISHING

	Transaction speed		
	Slow	***Average***	***Fast***
Time	~ 5 min	~2 min 18 sec	~30 sec
Price (ETH)	0.0012648	0.00015177	0.01011855
Price (USD$)	0.20	0.024	1.63
Type of interaction	Contract publishing	Contract publishing	Contract publishing

TABLE II. THE ESTIMATED PROCESSING TIMES FOR DATA STORAGE IN CONTRACT ON THE NETWORK

	Transaction speed		
	Slow	***Average***	***Fast***
Time	~ 7 min 6 sec	~3 min 6 sec	~30 sec
Price (ETH)	0.0006199	0.0007438	0.0049593
Price (USD$)	0.10	0.12	0.80
Type	Contract interaction	Contract interaction	Contract interaction

TABLE III. THE ESTIMATED TOTAL COST OF PUBLISHING DIPLOMAS FOR THE WHOLE UNIVERSITY FOR ONE ACADEMIC YEAR (2018/2019)

	Transaction speed		
	Slow	***Average***	***Fast***
Time	~ 7 min 6 sec	~3 min 6 sec	~30 sec
Price (ETH)	0.0006199	0.0007438	0.0049593
Price per unit (USD$)	0.10	0.12	0.80
Quantity	898	898	898
Total price (USD$)	89.8	107.76	718.4
Type of interaction	Contract interaction	Contract interaction	Contract interaction

V. CONCLUSION

Creating this kind of system is currently a rather complex task because of an underlining technology that is still quite under development. Required good knowledge of many technologies to develop a fully functional application is mandatory.

We have shown that for some applications development of SC may not be a very complex task. Primarily, SCs can serve as a simpler distributed database system and the complexity of the application itself must be maintained to some reasonable limits. Through the example application, we have seen that maintaining simplicity in the development of the SC leads to a decrease in the possibility of an error occurring in the application runtime.

A proposed solution presents one of the possible ways to enhance the safety, transparency, accuracy, and trust between involved and all interested parties, and to remove the need for intermediary's involvement. Generally, it is difficult to get exact data related to the number of forged diplomas, but in recent years, there has been a not so irrelevant number of scandals related to this topic. In this paper, one possible way to implement a diploma anti-forgery system is presented, one way to preserve document validity, as a legal public document, not taking into consideration its content. Therefore, in this context, other ways to check if some document is legally valid surely exists (with the support of software tools, contacting authorities that issued the document, etc.). Nevertheless, on the other hand, we think that the proposed solution is a promising, a very fast, secure, and generally, helpful way to discover if a diploma is fake or not. Blockchain technology definitely is not the only way to accomplish this task. This technology with supported tools provides secure, transparent, reliable, and without an intermediary's involvement very good framework to implement this kind of system.

To further develop this system, it is planned to create a fully automated system so that the students' graduation data, currently supplied through the generated JSON file, will be generated to the Ethereum network directly from the university's information system. In that way we will try to achieve functionality that when the student finishes his studies and when his status is concluded - he is graduated; data will be automatically transferred to the web application, which will then forward it to the smart contract on the Ethereum network to achieve complete

process automation. In this way, we will avoid one step in the current process and make the application much more efficient.

REFERENCES

[1] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 2008. [Online]. Available: https://bitcoin.org/bitcoin.pdf. [Accessed: 27-Jan-2020].

[2] V. Buterin, "Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform," *GitHub*, 2019. [Online]. Available: https://github.com/ethereum/wiki. [Accessed: 27-Jan-2020].

[3] "What is Ethereum?," *ethereum.org*. [Online]. Available: https://ethereum.org. [Accessed: 27-Jan-2020].

[4] B. K. Mohanta, S. S. Panda, and D. Jena, "An Overview of Smart Contract and Use Cases in Blockchain Technology," in *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2018, pp. 1–4, doi: 10.1109/ICCCNT.2018.8494045.

[5] G. Chen, B. Xu, M. Lu, and N.-S. Chen, "Exploring blockchain technology and its potential applications for education," *Smart Learn. Environ.*, vol. 5, no. 1, p. 1, Jan. 2018, doi: 10.1186/s40561-017-0050-x.

[6] A. Alammary, S. Alhazmi, M. Almasri, and S. Gillani, "Blockchain-Based Applications in Education: A Systematic Review," *Appl. Sci.*, vol. 9, no. 12, p. 2400, Jan. 2019, doi: 10.3390/app9122400.

[7] T. S. Kumar, S. Prabakaran, A. Sharma, and D. Vaidya, "Smart Blood Management and Tracking System," *Int. J. Recent Technol. Eng.*, vol. 8, no. 1, p. 6, 2019.

[8] H. L. Pham, T. H. Tran, and Y. Nakashima, "A Secure Remote Healthcare System for Hospital Using Blockchain Smart Contract," in *2018 IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–6, doi: 10.1109/GLOCOMW.2018.8644164.

[9] G. Leeming, J. Cunningham, and J. Ainsworth, "A Ledger of Me: Personalizing Healthcare Using Blockchain Technology," *Front. Med.*, vol. 6, Jul. 2019, doi: 10.3389/fmed.2019.00171.

[10] A. Shahnaz, U. Qamar, and A. Khalid, "Using Blockchain for Electronic Health Records," *IEEE Access*, vol. 7, pp. 147782–147795, 2019, doi: 10.1109/ACCESS.2019.2946373.

[11] S. M. Joshi and K. Rajeswari, "Efficient and Accurate Property Title Retrieval Using Ethereum Blockchain," in *Sustainable Communication Networks and Application*, Cham, 2020, pp. 424–438, doi: 10.1007/978-3-030-34515-0_45.

[12] Q. Xu, Z. He, Z. Li, and M. Xiao, "Building an Ethereum-Based Decentralized Smart Home System," in *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, 2018, pp. 1004–1009, doi: 10.1109/PADSW.2018.8644880.

[13] G. Papadodimas, G. Palaiokrasas, A. Litke, and T. Varvarigou, "Implementation of smart contracts for blockchain based IoT applications," in *2018 9th International Conference on the Network of the Future (NOF)*, 2018, pp. 60–67, doi: 10.1109/NOF.2018.8597718.

[14] R. C. Koirala, K. Dahal, and S. Matalonga, "Supply Chain using Smart Contract: A Blockchain enabled model with Traceability and Ownership Management," in *2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, 2019, pp. 538–544, doi: 10.1109/CONFLUENCE.2019.8776900.

[15] L. Augusto, R. Costa, J. Ferreira, and R. Jardim-Gonçalves, "An Application of Ethereum smart contracts and IoT to logistics," in *2019 International Young Engineers Forum (YEF-ECE)*, 2019, pp. 1–7, doi: 10.1109/YEF-ECE.2019.8740823.

[16] J.-C. Cheng, N.-Y. Lee, C. Chi, and Y.-H. Chen, "Blockchain and smart contract for digital certificate," in *2018 IEEE International Conference on Applied System Invention (ICASI)*, 2018, pp. 1046–1051, doi: 10.1109/ICASI.2018.8394455.

[17] C. Saraf and S. Sabadra, "Blockchain platforms: A compendium," in *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, 2018, pp. 1–6, doi: 10.1109/ICIRD.2018.8376323.

[18] "Ethereum Price Index — Real-time Ethereum (ETH) Price Charts," *CoinDesk*. [Online]. Available: https://www.coindesk.com/price/ethereum. [Accessed: 27-Jan-2020].

[19] "University of Nicosia. Academic certificates on the blockchain," *Institute For the Future*. [Online]. Available: https://www.unic.ac.cy/iff/blockchain-certificates/. [Accessed: 27-Jan-2020].

[20] E. Durant and A. Trachy, "Digital Diploma debuts at MIT," *MIT News*, 2017. [Online]. Available: http://news.mit.edu/2017/mit-debuts-secure-digital-diploma-using-bitcoin-blockchain-technology-1017. [Accessed: 27-Jan-2020].

[21] D. D.- KMi, "Open BlockChain," *Open BlockChain*, 11-Aug-2019. [Online]. Available: https://blockchain.open.ac.uk/. [Accessed: 27-Jan-2020].

[22] A. Mikroyannidis, J. Domingue, M. Bachler, and K. Quick, "A learner-centred approach for lifelong learning powered by the blockchain," in *EdMedia+ Innovate Learning*, 2018, pp. 1388–1393.

[23] P. Labs, "IPFS Powers the Distributed Web," *IPFS*. [Online]. Available: https://ipfs.io/. [Accessed: 27-Jan-2020].

[24] M. Sharples and J. Domingue, "The Blockchain and Kudos: A Distributed System for Educational Record, Reputation and Reward," in *Adaptive and Adaptable Learning*, Cham, 2016, pp. 490–496, doi: 10.1007/978-3-319-45153-4_48.

[25] K. Kuvshinov, J. Mostovoy, and I. Nikiforov, *Disciplina: Blockchain for Education*. 2018.

[26] "Sony Global Education Develops Technology Using Blockchain for Open Sharing of Academic Proficiency and Progress Records," *Sony Global - Sony Global Headquarters*, 22-Feb-2016. [Online]. Available: http://www.sony.net/SonyInfo/News/Press/201602/16-0222E/index.html. [Accessed: 27-Jan-2020].

[27] Z. Zhang, "Zhang Z. US patent application no. 2017043667a1." [Online]. Available: http://www.freepatentsonline.com/20170346637.pdf. [Accessed: 27-Jan-2020].

[28] "OP_RETURN - Bitcoin Wiki." [Online]. Available: https://en.bitcoin.it/wiki/OP_RETURN. [Accessed: 27-Jan-2020].

[29] L. M. Palma, M. A. G. Vigil, F. L. Pereira, and J. E. Martina, "Blockchain and smart contracts for higher education registry in Brazil," *Int. J. Netw. Manag.*, vol. 29, no. 3, p. e2061, 2019, doi: 10.1002/nem.2061.

[30] A. Kosba, A. Miller, E. Shi, Z. Wen, and C. Papamanthou, "Hawk: The Blockchain Model of Cryptography and Privacy-Preserving Smart Contracts," in *2016 IEEE Symposium on Security and Privacy (SP)*, 2016, pp. 839–858, doi: 10.1109/SP.2016.55.

[31] "Solidity — Solidity 0.6.2 documentation." [Online]. Available: https://solidity.readthedocs.io/en/v0.6.2/. [Accessed: 27-Jan-2020].

[32] "Solidity," *Wikipedia*. 17-Dec-2019.

[33] "Welcome to Remix documentation! — Remix, Ethereum-IDE 1 documentation." [Online]. Available: https://remix-ide.readthedocs.io/en/latest/. [Accessed: 27-Jan-2020].

[34] "Remix - Ethereum IDE." [Online]. Available: https://remix.ethereum.org/. [Accessed: 27-Jan-2020].

[35] "Sweet Tools for Smart Contracts," *Truffle Suite*. [Online]. Available: https://trufflesuite.com. [Accessed: 27-Jan-2020].

[36] etherscan.io, "TESTNET Ropsten (ETH) Blockchain Explorer," *Ethereum (ETH) Blockchain Explorer*. [Online]. Available: http://ropsten.etherscan.io/. [Accessed: 27-Jan-2020].

[37] etherscan.io, "TESTNET Kovan (KETH) Blockchain Explorer," *Ethereum (ETH) Blockchain Explorer*. [Online]. Available: http://kovan.etherscan.io/. [Accessed: 27-Jan-2020].

[38] etherscan.io, "TESTNET Rinkeby (ETH) Blockchain Explorer," *Ethereum (ETH) Blockchain Explorer*. [Online]. Available: http://rinkeby.etherscan.io/. [Accessed: 27-Jan-2020].

[39] "Laravel - The PHP Framework For Web Artisans." [Online]. Available: https://laravel.com/. [Accessed: 27-Jan-2020].

[40] "web3.js - Ethereum JavaScript API — web3.js 1.0.0 documentation." [Online]. Available: https://web3js.readthedocs.io/en/v1.2.4/. [Accessed: 27-Jan-2020].

[41] *ethereum/web3.js*. ethereum, 2020.

A Review on Generating Random Numbers in Decentralised Environments

Srđan Daniel Simić, Robert Šajina, Nikola Tanković, Darko Etinger
Faculty of Informatics
Juraj Dobrila University of Pula
Pula, Croatia
{ssimic, robert.sajina, nikola.tankovic, darko.etinger}@unipu.hr

***Abstract*—Recent advances in blockchain gained significant social attention, mainly due to substantial price fluctuations of Bitcoin and Ethereum cryptocurrencies. By its design, blockchain is an open, distributed ledger that can record transactions between two parties efficiently and in a verifiable and permanent way, providing solutions for many complex tasks without third party involvement. To achieve that, they employ a set of Byzantine Fault-tolerant consensus algorithms that require the implemented logic to be deterministic. The lacking source of randomness is a consequential limitation since many application domains, like games, lotteries, or random elections, require random sources. Given the Byzantine Fault-tolerance, generating random numbers should also be publicly-verifiable and tamper-resistant, but still hold the premises of being unpredictable.**

In this paper, we will provide an overview of the current research surrounding pseudo-random number generation on a decentralized network that satisfies those requirements.

***Index Terms*—Random number generation, Blockchain**

I. Introduction

Publicly-verified randomness is probably the most fundamental technical problem on a decentralized network. Its applications range from random elections, lottery games, to securing the network itself. Thus, a careful investigation into scalability and the security of the available solutions is crucial. The process of generating random numbers from an entropy source is called a Random Number Generator - RNG. The security of the system corresponds to the degree of the quality of the entropy source. There are two main requirements for the entropy source: **RQ1** - uniform distribution, **RQ2** - unpredictability of the result.

Random beacons like the NIST beacon [14] use the hardware entropy sources based on the different quantum-mechanical phenomena. These beacons are centrally governed and require an ongoing trust towards their publishers.

When the sources for random numbers are applied to the decentralized systems, additional requirements arise: **RQ3** - the characteristic of being publicly-verifiable and **RQ4** tamper-resistant which implies that the process of generating random number should not rely on any single actor. Those additional requirements led to many solutions of distributed RNGs - dRNGs that we cover in this review.

Prooving **RQ1** requires observing the process in the long run. A well known sufficient and agreed-upon approximation for achieving that requirement are cryptographic hashes [11]. Prooving the **RQ2** requirement is much harder because the traces of predictability are manifested by a statistically unlikely series of events that demonstrate the advantage for some actors. Generally, no one actor should be able to predict or extract any information in advance. The amount of information on future generated numbers should be the same for each actor.

RQ3 ensures that the generated random number should be verifiable by each actor - correctness. Because of a possible conflict of interest, actors must be able to verify if the received number is indeed the correct one and generated by a specific predetermined procedure.

We analysed the current state of the art and identified several unique ways in ensuring these requirements where most of them are based on a commit-and-reveal scheme and use public distributed ledgers.

II. Methodology

Studies included in this review are the most recent ones which tackled the problem of random number generation given restrictions of decentralised environment. The goal for this review is to enumerate the proposed techniques in decentralized random number generation and to answer the following question about each approach:

- Who are the actors in the proposed method?
- Which algorithmic techniques were used?
- What are the main characteristics of the algorithm in use?
- What is the most appropriate domain of use?

To identify relevant studies we queried Scopus database in March 2020 using **blockchain**,**random beacon**,**random number** keywords ,after which we iterated and converged on a relevant search query that captured all of the relevant aspects and used forward snowballing technique. After reading the abstracts, we selected the promising publications and found the final set of publications included in this short review.

III. Background

In this section, we give a brief overview of a set of methods that were applied in different stages of proposed approaches in order to satisfy the requirements **RQ1-4**.

Publicly verifiable secret sharing. Verifiable secret sharing scheme has the property that the validity of the shares distributed by the dealer can be verified by any party [17]. It is based on classic secret sharing scheme which distinguishes two protocol phases: *distribution* phase where a secret is

distributed (splitted) by a dealer among the participants, and a *reconstruction* phase in which the secret is recovered by fetching the shares, usually in a (t, n) scheme where at least t out of n participants can together reconstruct the secret. Verifiable secret sharing (VSS) resists malicious players, including the dealer itself. Public verification enables that any third party can verify that participants received correct shares.

Collective Signing - CoSi [18] is a scheme based on a variation of well-known Schnorr multi-signature algorithm [10]. Threshold model (t, n) is applied where at least t out of n participants are required in order to decipher the secret message. The secrets are usually shared via the aforementioned PVSS system so that an honest threshold of users can later recover them and form a publicly verifiable proof of their validity.

CoSi is used in [19] for *RandHound*, *RandShare* and *RandHerd* algorithms. RandHound uses a commit-then-reveal approach implemented via PVSS and CoSi to prevent client equivocation (explanation needed). The requirement is that at least f of $3f+1$ contributors are honest. RandHound is similar to RandShare, except that contributors are divided in smaller groups to enable scalability.

Homomorphic encryption (HE) is a kind of encryption scheme that allows a third party to perform certain computable functions on the encrypted data while preserving the features of the function and format of the encrypted data. An HE is primarily characterized by four operations : *KeyGen*, *Enc*, *Dec*, and *Eval*. While *KeyGen*, *Enc* and *Dec* are no different from their classical tasks in conventional encryption schemes, *Eval* is HE specific operation which takes ciphertexts as input, performs the function $f()$ over ciphertexts without seeing the messages and outputs evaluated ciphertexts. The most crucial point in HE is that the format of the ciphertexts after an evaluation process must be preserved in order to be decrypted correctly. To support an unlimited number of operations the size of the ciphertext should also be constant otherwise the increase in the ciphertext size will require more resources [1]

- Partially homomorphic encryption - allows only one type of operation with an unlimited number of times [1]
  - RSA- only homomorphic over multiplication
  - ElGamal- multiplicatively homomorphic
- Somewhat homomorphic encryption - allows some types of operations a limited number of times [1]
- Fully homomorphic encryption - allows an unlimited number of evaluation operations on the encrypted data and resulting output is within the ciphertext space [1]

Zero-knowledge proofs. ZK-proof [7] is a scheme in which a *prover* can prove to any interested party (*verifier*) that he knows a secret value x without revealing any information on the x, except to the fact that he knows the value. It comes in the interactive flavor where messages are exchanged between prover and verifier, or in non-interactive flavor where no such interaction is necessary, but requires a bit-string to be shared between prover and verifier that is **guaranteed** to be random.

Verifiable random function (VRF). VRF [12] is a pseudo-random function that given an input x and a secret key (SK) produces a pseudo-random number $f_{SK}(x) = y$ together with a non-interactive publicly verifiable proof π_y for y so that everyone with the knowledge of π_y and the public key (PK) can verify whether y was generated correctly, but still cannot recover SK. VRFs can be regarded as an asymmetric version of keyed hash functions. They are useful for mapping low entropy inputs ($x \in X$ where $H(X)$ is low) in order to provide preimage resistance: it is difficult to find x such that $h(x) = y$. Ordinary cryptography hashes and encryption are not preimage resistant.

BLS signature/scheme. BLS signature uses a group of *N* members that collectively hold private key *S*, with each member only holding a part of that key denoted s_i. Public key corresponding to *S* is marked as *P* and messages signed by *S* are marked as *SIG*. In the process of signing, each member signs a message with their key s_i resulting in sig_i. *SIG* is calculated only after receiving *k* signatures [15]. BLS scheme utilizes these properties by dividing nodes in the network into h groups. A member of each group produces and releases its share for a BLS signature of the message (e.g., a round number), sharing it with other members of the group. Each member of a given group waits until a threshold number *k* of BLS signature shares are available for that group and then forms the BLS signature SIG_i for this group. The first member who collected *k* shares announces or gossips these signatures with members of other groups. The next group is chosen based on SIG_1. This chosen group repeats the same process as the first group but uses SIG_1 as a signing message. Process of combining this groups signatures continues until a global BLS signature is available [15] [19] [16]. Only the last group can manipulate SIG_h if the group has more than *k* malicious members, although it is not predictable which group will do the SIG_h generation task [15].

IV. Commit and reveal approaches

These approaches are based on generating a random number from a series of contributions from multiple agents. Multiple agents are used to satisfy **RQ4**. The base assumption here is pretty realistic: at least one agent must be honest and not in malicious coordination with other agents. If that is the case, **RQ2** is also satisfied. The protocol is divided into time epochs where each epoch generates a single random number. Each epoch is further divided into two main phases: (1) the collection phase, where all the contributions are gathered, but hidden in a way that no one agent can gain insight from them, and (2) the reveal phase in which the agents have to reveal their contribution to others satisfying **RQ3**. There are two essential rules to adhere to: the second phase does not start until the first phase is finished, and all the contributions must be revealed in the second phase for the epoch to be considered valid.

A known challenge in the commit-reveal approach is that the last agent to reveal its commitment has an unfair advantage of knowing the final output before revealing their secret and thus can withhold revealing if it considers for that outcome

not to be favorable. This can be regarded as a form of Denial of Service attack, and there are separate agreements on techniques on how to enforce the final revealing. In this chapter, we will see different methods authors used to tackle that challenge.

A. Dealing with non-reveals

To address the flaw in a commit reveal scheme authors [20] proposed protocol where they provided means of hiding the randomness output from any subset of participants that are not abiding by the rules of the protocol while also maintaining the public verifiability which is necessary for randomness source. Their protocol has five steps where a new step can only be started when all the participants have finished the previous one: (1) Commitment, (2) Revealing, (3) Distribution, (4) Validation, (5) Reconstruction. To ensure that no participant has an unfair advantage in (2), (1) was designed so that each participant computes $y_i = g^{x_i}$, where $x_i \in Z_q$ is their random value and g is an element of a cyclic multiplicative group G, and together with another random value r_i with at least $log_2(||Z_q||)$ random bits calculates a secure cryptographic hash function $H_i = H(y_i, r_i)$ which is then recorded on a blockchain. Authors argue even though $y = g^x$ is publicly known, nobody can extract x unless he is capable of calculating the discrete logarithm in G, which is assumed to be a hard problem. The last step in (2) is extracting randomness R from blocks of the blockchain. Authors state that this step stops any attacker, who does not control at least k participants, from manipulating the block randomness in their favor. In step (3), participants distribute random value x from a uniform distribution using verifiable secret sharing protocol. In (4) each participant P_t decrypts and verifies $s_{i,t}$ of others P_i from (3) using predefined equation. If the equation results in non-equality, that will imply that P_i had been cheating. Finally in (5) all participants submit their decrypted $s_{i,t}$ and by solving the system of linear equations $s_{i,t}$ any party can calculate values of x_i and therefore $x = \sum_{i=1}^{n} x_i$. Public randomness r is finally achieved with randomness extraction function (secure cryptographic hash algorithm), which processes the value x concatenated with R, to ensure **RQ1**. Some of the possible security issues in this approach are 51% attack and having k corrupted participants.

B. Usage withing lottery-alike games

Authors in [9] addressed the issue of unfairness in lottery games and have proposed a blockchain-based smart contract for the fair and efficient lottery. Their approach is based on a commit-reveal scheme that can be described in 6 phases: (1) Deploy, (2) Register, (3) Purchase, (4) Publish, (5) Verify, (6) Award. Phase (1) starts with lottery center generating two pair of keys (LC_k and PL_k public and private) after which they deploy smart contract, which contains ticket value of users and specified $EndTime$, signed with their private LC_k while also providing their public LC_k to protect the confidentiality of communication. In (2) players generate their key pair PL_k for which they receive an address $PL_{address}$ from the blockchain platform. Players buy their tickets in phase (3). The ticket contains ticket value and identification code (v_i, s_i) chosen by the player. For ticket to be valid players have to send $Hash(v_i, s_i)$ to the blockchain before $EndTime_{buy}$, which will generate list of unverified transactions, after which they have to submit $LC_{P_k}(v_i, s_i)$ before $EndTime_{open}$ so that lottery center, using their private LC_k, can verify if the (v_i, s_i) matches the value submitted before $EndTime_{buy}$. Phase (4) is what distinguishes it from other research. To generate a winning number, authors have chosen to use Barycentric Lagrange Interpolation, which is an improvement of Lagrange Interpolation. Lottery center treats all $(Hash(v_i, s_i), v_i)$, submitted by the players, as points (x_i, y_i) that are used to get a polynomial $f(x)$ which goes through all of them. Authors argue that the coefficient of the polynomial is unpredictable and "truly random", therefore ensuring **RQ2** . If a player suspects that a winning number is false, he can verify it in (5) by validating $f(x_i) = y_i$, which provides answer to **RQ3**. The last step is (6) where the winner provides his $PL_{address}$ and (v_i, s_i) to the lottery center which verifies it and rewards the player accordingly. Authors conducted a security analysis of their approach that addressed several attacks on the system, along with solutions for them.

Authors in [5] addressed the problem of data manipulation by the game providers and designed algorithms to prevent human manipulation, which allows all the participants to be involved in the generation of random numbers. It has three actors: players, game providers (GP), and a blockchain. The algorithm can be summarized in a few steps:

- Process starts when GP creates a random number with a public-private key pair and records it on a blockchain
- Players generate and send their own random numbers to GP (**RQ02**)
- GP generates a random number according to a pre-announced algorithm by using players number, chain random number and its own random number
- GP uses this newly generated random number to start a game
- When the game is over GP announces its number and private key which players can use to verify if the random number is correct or if there was cheating involved satisfying **RQ03**

The algorithm was deployed on a consortium chain-based game platform with mixed Byzantine fault tolerance (MBFT) consensus algorithm. Within their experimental environment, the authors concluded that the latency between the player sending its random number and game provider sending response is approximately equal to the load time that a player has to wait to start a game. When there are more than 2000 players, there is a need for more than one consensus group because latency doubles with an increasing number of players.

C. Using beacons

Authors in [8] proposed a protocol used inside beacon that combines properties of blockchain and timed commitment, thus making the probability of adaptive attacks negligible

while also providing a quickly verifiable result. The main advantage of their protocol is in their verifiable timed commitment algorithm. By adapting the flaw of timed commitment method where adversaries can hinder the protocol with false commitments, authors offered a solution in which the commitment can be guaranteed to open correctly by verifying the commitment before accepting a block, including the one. Their protocol can be divided into five steps:

1) Commit
2) Verify commitment
3) Open commitment
4) Verify result
5) Generate public randomness

The evaluation was done on availability, unpredictability, impartiality, and verifiability of their protocol. Chain growth, common prefix and chain quality properties of blockchain are guaranteeing availability, unpredictability (**RQ2**) and impartiality, while verifiability (**RQ3**) has been ensured with verifiable timed commitment.

In [4], the authors propose an approach with participants who send their inputs to the beacon operator. Participants send their inputs combined with all data required for the computation named commitment, **RQ4**. Such a technique improves transparency by allowing any party to compute this randomness alongside the beacon operator, **RQ3**. To decrease the possibility of the operator manipulating the results by trying different commitments before releasing them, the operator must run a verifiable delay function, which ensures that the operator cannot try more than one commitment before running out of time, **RQ4**. The adversary can not bias the outcome as long as there is at least one honest party. As the authors state, there are two options for implementing this approach: (1) using a smart contract as a beacon operator, (2) beacon operator runs separately, but publishes some of its artifacts on the blockchain. While the first solution offers benefits in terms of decentralization, each computation executed on a smart contract consumes virtual currency, requiring many users to compensate for the high on-chain computation cost. Authors further break down the second option in four options, each providing different trade-offs of decentralization, security, and costs.

D. Using homomorphic encryption

Authors of [13] proposed a protocol composed of three main components: Requester, Core Layer (consists of many parties responsible for PRNG), and Public Distributed Ledger (PDL). Their protocol works in rounds where each round consists of a few stages. To initialize the process, Requester sends a $nonce$ together with his PK to the PDL for what he receives the ticket T, which is used to determined which parties, called contributors, are eligible to contribute to generating a random number. For each T, an instance of verifiable random function (VRF) had to be executed by Core Layer's parties, after which proof-of-eligibility (PoE) has to be sent to PDL that verifies all received PoE. Each contributor, which is a member of the eligible party, encodes his contribution, chosen value M_i, with Requester's PK, and sign it with their private key after which proof-of-contribution is generated. After all contributions have been made public, they are summed and sent to the Requester, which is possible because the protocol employs El-Gamal encryption. At the final stage, Requester uses his secret key to decrypt the tallied contribution and publishes decrypted value M along with a proof of proper decryption, **RQ3**. Security and complexity analysis has been conducted in which authors showed that their protocol could be implemented for practical use, after which they compared it with other possible solutions on dRNG market. The authors also noted the limitation of their protocol, which is an assumption that the Requester does not share his private key with anyone else, negating **RQ4**.

E. Game-theoretic approaches

In [2] authors proposed a commit-reveal approach based on a random bit generation game (RBG) and implemented their algorithm using a single-instance smart contract called Random Bit Generator Contract (RBGC). In the RBG game, each player is represented as a number and can be either an even-numbered or odd-numbered player. Even-numbered players can choose between playing either 0 or 2, while odd-numbered players can play either 1 or 3. A mini-game is simulated for each *(i, j)* pair where *i* and *j* are of different parity's. The winner is the player with its chosen number greater by one than that of its opponent. This way, players are incentivized to play uniformly at random to maximize their earnings, enforcing **RQ1**. The RBGC contract supports multiple requests for a random (in experiments 176 bits/s) with requesting clients specifying random bit generation due time. This approach produces random numbers as long as one of the participants is honest. The client who requested the first random bit would receive its bit only if all precipitants revealed its bits truthfully. The RGBC contract will fail to deliver random bit in cases where the deadline was to close (time available for generating the random bit is less than a predefined constant t_{min}), there were no participants or some participants failed to reveal its bit. In the latter case, the RGBC contract refunds the client with confiscated deposits. Finally, in the case of the successful reveal of all participants, a random bit is generated as XOR of all bits. Authors argue that generating a multi-bit random number can be achieved by asking participants to submit an N-bit number and running an independent RGB game for each bit. Looking at the implementation code provided by authors, we noticed that running multiple independent RBG games for each bit would be imposed on the participant in the form of transaction execution costs - *gas*.

V. Pure DLT approaches

One of the first approaches in the area of blockchain-based smart contracts used to solve the pseudo-random number generator problem was to use the information on the ledger itself as the source of entropy, hence the name of this section Pure Distributed Ledge Technology approaches (Pure DLT approaches). In the proof-of-work secured ledgers, the primary

source was the computed block *nonce* which comes from a high entropy uniform distribution computed with a delay for securing the ledger itself. Therefor using proof-of-work as primary source of entropy satisfy **RQ1**, **RQ2** and **RQ3**.

There are two main branches in this approach: (1) using the information from previous blocks and (2) using the information from yet unseen future blocks. In the cases where additional security is required, just using the block nonce is not sufficient, because of possible attacks on a blockchain network itself, so additional sources of entropy are required. In this chapter, we will explore how authors used data from blocks spiced with some additional data to produce a reliable pseudo-random number generator.

A. Previous blocks

In this subsection, we will review [3] in which authors ensured **RQ4** by utilizing data from previous blocks of a blockchain paired with cryptocurrencies price information to produce one-dimensional Cellular Automaton. To create a seed, which will be used to generate a one-dimensional CA, the authors described the formation of RngBlocks. These blocks V_{t_i} consists of sequence of bits of length k, timestamp t_i that denotes when the RngBlock was created and set of statistically independent evidence $E = \{e_1, e_2, ..., e_n\}$. That evidence is data extracted from previous blocks of a blockchain, and they must be represented in a binary form. Authors aligned the population of RngBlocks with the evolution of a blockchain, which is considered to be a continuous sequence of blocks, in a bounded time window $[L, A]$ of block stream S that contains $(V_{t_i}, block.hash_i)$ pairs. The generation of the genesis RngBlock $(V_{t_n}, block.hash_n)$, which is used as the initial seed for 1-d CA creation, is triggered by a self-executing smart contract, with a deployment date L, in a specific point in time A when all execution conditions are met. Configuration of proposed CA, which consists of a Boolean cellular array with $s \in \{0, 1\}$ states, is as follows : initial seed $(V_{t_n}, block.hash_n)$, length of the CA, which is the same as the RngBlock length, transition rules, which are specified in the form of a 3-bit rule table with each entry capturing every possible neighborhood state, and the time of execution q, which sets the finite number of evolution steps for CA. Authors defined q as a decimal form of the last z digits converted from a specific hexadecimal value that has been gathered from a block's hash which is determined by *blockhash(block.number - offset)* where *block.number* is the height of the current block and *offset* is related to Ethereum's price fluctuation normalized within a domain $[0 - 255]$. In each evolutionary step using the genesis RngBlock as the initial seed, their proposed methodology instantiates a CA, while the selection of transition rule, which contributes to the next evolution state of the CA, is set according to binary representation of a normalized integer value that resides within the domain [0-255] derived by $y = (255 * ms)/1000$ where ms is milliseconds collected from A. When the number of steps reaches q, a binary number of length k is produced as the final outcome of the pseudo-random generator process. The authors conducted an experimental evaluation using *Dieharder* and *NIST* statistical test suits that yielded positive results.

B. Future blocks

In [6], the authors defined their protocol with four participants: users, service system, random number generator, and blockchain. The random number generator is implemented as a smart contract and can execute two kinds of queries: Reg, which registers the received data on the blockchain, and $AskNonce$, which is used to obtain the nonce of the future block. Authors implemented their protocol inside a lottery game, and the process for obtaining the winning number can be described as a two-phase process: user and service system registration and lottery phase. During the user and service registration phase, each participant sends their seed and address to the random number generator smart contract G, to ensure **RQ4**. The registration phase must be executed at least once before the lottery phase can begin. During lottery phase G makes identification number sn and executes $Reg(sn, a, b)$, where a is user's and b system service address, after which $AskNonce$ would be executed to get nonce n of the next block. Last step is computing pseudo-random number α by calculating a hash function of sn, n, a, b , r and p, where r and p are user's and system service seed. The authors noted that their protocol has performance and scalability problems that will be addressed in their future work.

VI. Conclusion

Generating pseudo-random numbers on a decentralized network is a non-trivial task. Lack of entropy and transparent ways to secure the data so that no malicious party can manipulate or gain unwanted knowledge from it is still a major bottleneck in that process.

In this paper, we present some of the recent attempts in generating a pseudo-random number on a decentralized network, i.e., blockchain. Although, during our research, we encountered various approaches, based on volume of relevant research we concluded that at the moment, the most promising, and also the most prominent approach for solving this particular problem, is by using a commit reveal scheme.

References

[1] Abbas Acar et al. "A Survey on Homomorphic Encryption Schemes". In: *ACM Computing Surveys* 51.4 (2018), pp. 1–35. ISSN: 03600300. DOI: 10.1145/3214303.

[2] Krishnendu Chatterjee, Amir Kafshdar Goharshady, and Arash Pourdamghani. "Probabilistic smart contracts: Secure randomness on the blockchain". In: *ICBC 2019 - IEEE International Conference on Blockchain and Cryptocurrency* (2019), pp. 403–412. DOI: 10.1109/BLOC.2019.8751326. arXiv: 1902.07986.

[3] Klitos Christodoulou et al. "Randomblocks: A transparent, verifiable blockchain-based system for random numbers". In: *Journal of Cellular Automata* 14.5-6 (2019), pp. 335–349. ISSN: 15575977. DOI: 10.13140/RG.2.2.21366.96328.

[4] Samvid Dharanikot et al. “Breeding Unicorns: Developing Trustworthy and Scalable Randomness Beacons”. In: (2020), pp. 99–106. DOI: 10.1109/blockchain.2019.00022.

[5] Mingxiao Du et al. “A blockchain-based random number generation algorithm and the application in blockchain games”. In: *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics* 2019-Octob (2019), pp. 3498–3503. ISSN: 1062922X. DOI: 10.1109/SMC.2019.8914618.

[6] Yuto Ehara and Mitsuru Tada. “How to generate transparent random numbers using blockchain”. In: *Proceedings of 2018 International Symposium on Information Theory and Its Applications, ISITA 2018* C (2019), pp. 169–173. DOI: 10.23919/ISITA.2018.8664239.

[7] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. “Knowledge complexity of interactive proof systems”. In: *SIAM Journal on Computing* (1989). ISSN: 00975397. DOI: 10.1137/0218012.

[8] Yanxue Jia and Lei Fan. “Generate public randomness based on blockchain”. In: *Proceedings - 2018 IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovations, SmartWorld/UIC/ATC/ScalCom/CBDCo* (2018), pp. 263–270. DOI: 10.1109/SmartWorld.2018.00080.

[9] Jiasheng Li, Zijian Zhang, and Meng Li. “BanFEL: A Blockchain Based Smart Contract for Fair and Efficient Lottery Scheme”. In: *2019 IEEE Conference on Dependable and Secure Computing, DSC 2019 - Proceedings* (2019), pp. 1–8. DOI: 10.1109/DSC47296.2019.8937559.

[10] Gregory Maxwell et al. “Simple Schnorr multi-signatures with applications to Bitcoin”. In: *Designs, Codes, and Cryptography* 87.9 (2019), pp. 2139–2164. ISSN: 15737586. DOI: 10.1007/s10623-019-00608-x.

[11] Alfred J. Menezes, Scott A. Vanstone, and Paul C. Van Oorschot. *Handbook of Applied Cryptography*. 1st. USA: CRC Press, Inc., 1996. ISBN: 0849385237.

[12] Silvio Micali, Michael Rabin, and Salil Vadhan. “Verifiable random functions”. In: *Annual Symposium on Foundations of Computer Science - Proceedings* (1999), pp. 120–130. ISSN: 02725428. DOI: 10.1109/sffcs.1999.814584.

[13] Thanh Nguyen-Van et al. “Scalable Distributed Random Number Generation Based on Homomorphic Encryption”. In: October (2020), pp. 572–579. DOI: 10.1109/blockchain.2019.00083.

[14] NIST. *Interoperable Randomness Beacons*. 2019 (accessed January, 2020). URL: https://csrc.nist.gov/projects/interoperable-randomness-beacons.

[15] randao.org. “Randao: Verifiable Random Number Generation”. In: 2017. URL: https://www.randao.org/whitepaper/Randao_v0.85_en.pdf.

[16] Philipp Schindler et al. “HydRand: Practical Continuous Distributed Randomness.” In: *IACR Cryptology ePrint Archive* 2018 (2018), p. 319. URL: https://eprint.iacr.org/2018/319.pdf.

[17] Berry Schoenmakers. “A simple publicly verifiable secret sharing scheme and its application to electronic voting”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 1666.i (1999), pp. 148–164. ISSN: 16113349. DOI: 10.1007/3-540-48405-1_10.

[18] Ewa Syta et al. “Keeping Authorities "honest or Bust" with Decentralized Witness Cosigning”. In: *Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016*. 2016. ISBN: 9781509008247. DOI: 10.1109/SP.2016.38. arXiv: 1503.08768.

[19] Ewa Syta et al. “Scalable Bias-Resistant Distributed Randomness”. In: *Proceedings - IEEE Symposium on Security and Privacy* (2017), pp. 444–460. ISSN: 10816011. DOI: 10.1109/SP.2017.45.

[20] Habib Yajam et al. “Improvement on Bitcoin’s Verifiable Public Randomness with Semi-Trusted Delegates”. In: *9th International Symposium on Telecommunication: With Emphasis on Information and Communication Technology, IST 2018* (2019), pp. 53–57. DOI: 10.1109/ISTEL.2018.8661008.

Accessibility Standards and Their Implementation in Custom Data-Driven Maps

Irena Serna-Marjanović, Anel Tanovic, Ajla Cerimagic
Department of Computer Science
Faculty of Electrical Engineering, *University of Sarajevo*
Sarajevo, Bosnia and Herzegovina
irena.serna@etf.unsa.ba, atanovic@etf.unsa.ba, acerimagic@etf.unsa.ba

Abstract - This paper provides an overview of several standards and gives an example of their implementation in custom visualizations such as dynamic data-driven maps. A specially developed tool MapSpice for data visualization will be presented in this paper. This tool is developed and designed to fulfill accessibility standards and requirements of the web publishing industry. There are descriptions and examples of important features and functionalities in map visualizations that are developed using accessibility best practices. Following these guidelines and recommendations map content is accessible to a larger range of people with different disabilities, including blindness and low vision, photo-sensitivity, speech disabilities, deafness and hearing loss, limited movement, learning disabilities, cognitive limitations, and others.

The result of this paper will be an overview of accessibility standards and guidelines and their implementation in custom data-driven maps developed by MapSpice.

Keywords— Accessibility, ISO/IEC 40500:2012, ISO 9241-171:2008, WCAG 2.0, Web Content Accessibility.

I. INTRODUCTION

Presenting and publishing data in the digital era is extremely important, but also demanding. One of the high important requirements is content accessibility.

If publishing content is not accessible, then it becomes a social and legal problem. A lot of publishing agencies today set accessibility as one of the main features of their publishing content on the web.

Accessibility is a term that describes how certain product, application or device is available for a different type of users with a wide range of abilities and impairments. Accessibility today is not only a recommendation but a need and requirement. And in a lot of countries, it is also a legal obligation. For specific industries and business fields, accessibility should be taken into consideration from the beginning of the project as one of the most important project results. Map visualizations are also no exception.

Accessibility is also a great benefit for people without disabilities or with temporary disabilities such as broken arms, broken or lost glasses, people who have to work on small devices, old people, etc.

In this paper, several web accessibility standards will be described. Special focus will be on those sections and guidelines that are applicable for web content and data-driven maps. For these purposes, the authors examined and researched map visualizations developed by Social Explorer's tool, MapSpice [10]. Using assistive technology and source code insight these elements were recognized and described in the following sections. The main goal was to present the implementation of rules and given guidelines in order to recognize further accessibility improvements.

Accessibility standards and their main points are described in Section II.

Section III describes the platform for publishing creative and flexible data maps to the web – MapSpice. It enables a user to create custom data-driven maps for presenting and publishing data. This platform was developed by customers who need to present different data on maps. MapSpice follows rules and guidelines set up by agency for education publishing and assessment that is following the U.S. legislation.

As a part of this section, it will be presented how maps are created with MapSpice and what aspects of accessibility are considered.

II. ACCESSIBILITY STANDARDS

There is a variety of standards that give guidelines and principles of web accessibility and software quality. Also, a lot of organizations and companies give their recommendations and set their standards for web accessibility based on their products and customer needs. These rules and guidelines are just adjustments to existing standards and their principles.

These adjustments sometimes can be expensive in the manner of time and human resources and it is best practice to follow guides and principles given in several standards for web accessibility.

A. ISO 9241-171:2008

ISO 9241 provides guidance and specifications for the design of accessible software and other interactive systems. Accessibility is strongly related to the concept of usability and therefore increases effectiveness and efficiency for people having a wide variety of capabilities and impairments.

ISO 9241 consists of the 33 parts, under the general title: Ergonomic requirements for office work with visual display terminals. One of these parts is 171 with title Guidance on software accessibility.

Eleven clauses define controls and guidelines for accessibility adjustments [1].

Although in many cases organizations may define and set up their guidelines that depend on their user's needs, it is very difficult to consider all aspects that need to be met. Especially when developers are new to accessibility. It is always a good practice to follow the rules and guidelines of the known standards.

B. WCAG 2.1

Web Content Accessibility Guidelines (WCAG) 2.1 presents a set of technical documentation with recommendations for making Web content more accessible.

Twelve guidelines fall under four basic principles that were used to develop a platform for the implementation of data-driven maps that will be presented in Section III. These four principles were recognized and observed on these maps and data visualizations:

- Perceivable

The user interface and all data on the screen are visible and distinct. Users can access, recognize and understand content using assistive technology[1].

- Operable

There is a lot of users that are not able to use a mouse due to permanent or temporary impairment. These users rely on keyboard controls or other tools to navigate. In order to enable access for these users so they can interact with a website following is fulfilled:

  - All functionality and content are available using a keyboard.
  - Users have enough time to read and use the content on the web
  - Users can easily navigate, find content, and determine where they are on the map with the use of focus, element order, paged titles.

- Understandable

Web content on presented maps uses clear and concise language. The text is readable, understandable and easy to interpret. When a user navigates around the website and accesses different content, he is announced of any change and the result of these actions is obvious and predictable.

- Robust

Term robust means that specific website works across different operating systems, browsers, and on tablet/mobile devices. Data-driven maps that were a subject of this paper adjust automatically for different desktop, mobile and tablet views [10].

All of these layers (principles, guidelines, success criteria, and sufficient and advisory techniques) work together to provide guidance on how to make content more accessible. In order to meet the needs of different groups and different situations, three levels of conformance are defined: A (lowest), AA, and AAA (highest) [2][5].

The WCAG guidelines have also been adopted as International Standard ISO/IEC 40500:2012.

C. WAI-ARIA - Accessible Rich Internet Applications Suite

WAI-ARIA, the Accessible Rich Internet Applications Suite, is a specification that defines a set of rules and recommendations to make Web content and applications more accessible to people with disabilities.

To improve accessibility, WAI-ARIA provides developers with the option to add additional information about the structure of the document. It enables assistive technology to recognize and present to users the structure of the web pages.

This semantic information can be roles to describe the type of widget or structure of the web page, properties to describe the state widgets or live regions, etc.

Their combinations and usage on the page help users to understand and perceive the content of the web page using assistive technology.

WAI-ARIA defines a set of additional HTML attributes that better describe elements and gives them additional information. These additional attributes don't change the content of the page or affect its behavior. They should be considered from the perspective of assistive technology, such as screen reader which will recognize these tags and give info to the user about the page itself and elements that presents it.

There are three main features that element should contain:

- The role is the feature that defines the purpose of the specific element. Examples of these roles are 'button", "banner", "tab", "menu", "menu item", etc.
- Property is a description of an element, which can be used to give them extra meaning. For example, aria-decsribedby="Some description" attribute is used to indicate the IDs of the elements that describe the object. It is used to establish a relationship between widgets or groups and text that described them.
- The state is a special property that defines the current condition of a specific element. An example of it is aria-disabled="true". In this case, the screenreader will not notify the user about these elements as it is currently disabled [5][6].

[1]Assistive technology is any device, software, or equipment that helps people with disabilities or the elderly population so they can learn, communicate, and function better

III. MAPSPICE

MapSpice is a solution for creating and publishing data maps to the web. It was developed by Social Explorer[2] who is using it to create on demand data-driven maps for its customers [10].

It is a platform specially developed and designed to fulfill customers' needs to present data on maps. It includes a lot of interactive components and elements which provide a full experience of searching and exploring.

Wizards help to get users started with templates that can be customized to produce rich data maps quickly and easily. It also includes a software development kit and a code editor for more advanced users who want to customize their maps further.

MapSpice comes with professionally designed templates and components the user can use in visualizations. MapSpice uses the latest web technologies such as WebGL and vector-based map tiles to serve high volumes of maps quickly and consistently.

Also, maps are adjusted automatically for desktop, mobile and tablet views.

MapSpice as a tool for visual presentation of data on maps followed the guidance and recommendation of the publishing agency. Legal requirements are also considered when developing solutions that conform to a single shared standard that meets the needs of individuals, organizations, and governments internationally.

Implementation of accessibility standards can be seen in the following:

- Users can easily navigate, find content, and determine where are they on the map
- Every interactive component is accessible with a keyboard
- Text is readable and set up with recommended contrast

All of these features and functionalities are fully implemented on the map using MapSpice.

A. *Keyboard accessibility*

Every component of the map is designed to be possible to access its content using a mouse as well as keyboard navigation. All functionality of the map content is operable through a keyboard interface without requiring specific timings for individual keystrokes.

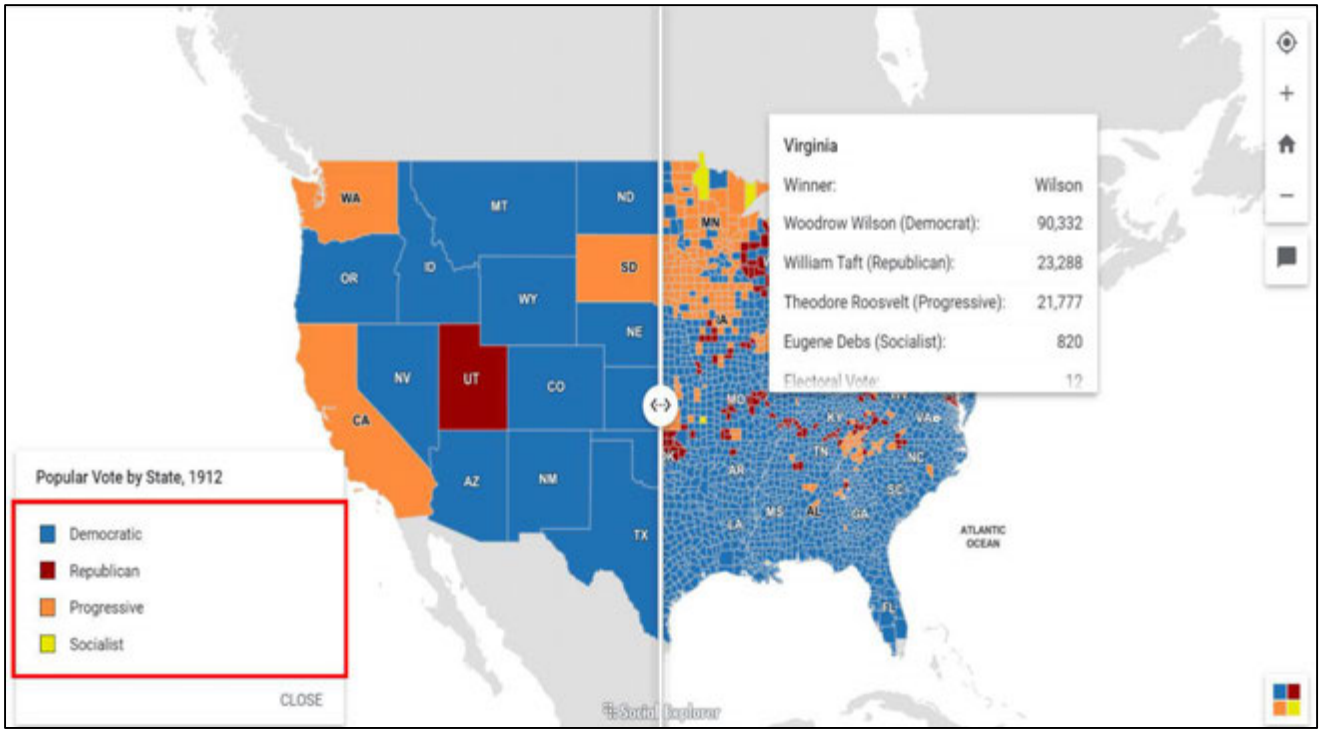


Figure 1. Data driven map created with MapSpice

[2] Social Explorer is a software company that offers several solutions for data visualization, mapping and geospatial data [9]

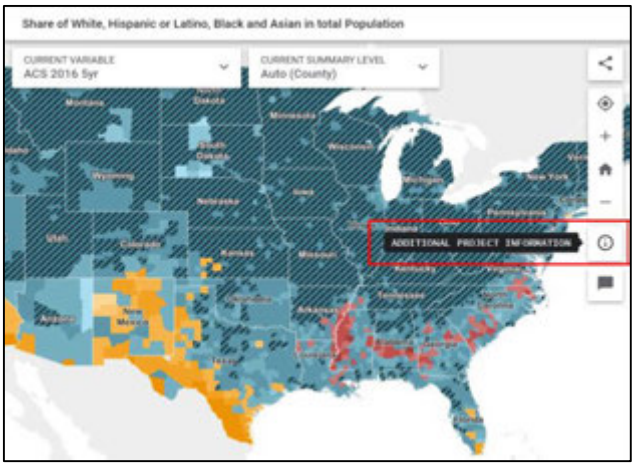


Figure 2. Focus indicator of a selected element

The keyboard focus does not get trapped in any part of the content. Components that have specific functionality can be accessed using event keycode for Enter (keycode 13). For every action taken on this map, the user is notified, and the focus is moved to the desired element. A visual focus indicator has been provided which ensures that the user in every moment knows where he is currently on the page.

When the map is loaded user is announced that content on the screen is a map graphic and the map title is red (when using assistive technology such as screen readers).

One of the great features of these maps is that the content of the map is also accessible using a keyboard. Only part of the map is focused and there is a visual element (annotation) with all data presented on the map. When a user is using a screen reader, focused areas and elements are being read. The user can easily navigate to another area of the map using up and down keyboard keys.

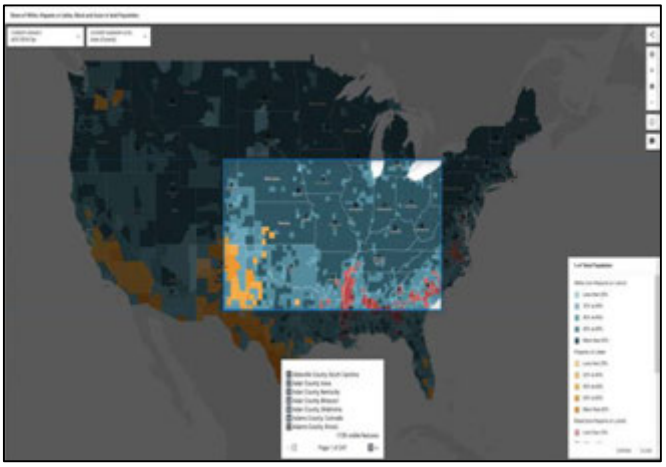


Figure 3. Accessible map content

B. *Tab index*

Every interactive element on the map has an attribute tab index that indicates that this element can be focused when using keyboard navigation:

- An element with a negative value of tab index (usually '-1') cannot be focused by default but using with Javascript or visually if needed.
- tabindex="0" means that the element will be focused, but the order is defined by the document's source order.
- A positive value of the tab index means that the order of focused elements is defined by the value of the number ('4' is focused before '5', but after '3') [5].

Elements on the map have either tab index '0' or '1' depending on what priority we want to give to the desired element. Usually, elements get focus from top to bottom, going from the left side to the right.

Some components are more complexed then others and may contain sub-items that are accessible after the user indicates that wants to operate within this component. Such components are menus and legend component.

For legend component sub-elements have tabindex=-1 and their index is programmatically changed when the user moves to the legend component and enters it in order to see change on the map for different color elements.

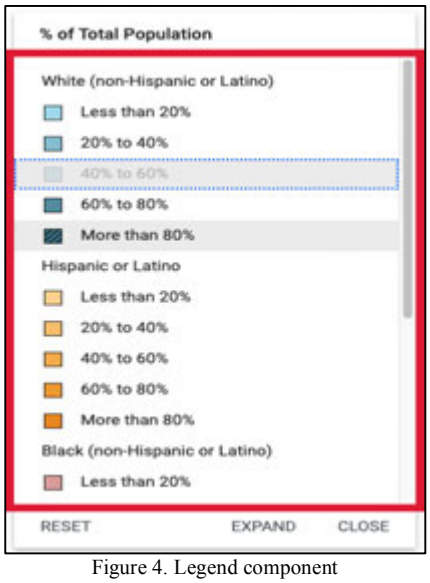


Figure 4. Legend component

C. *WAI-ARIA - Accessible Rich Internet Applications Suite*

Every map element has its aria-label attribute that defines a string that labels elements. In most cases, aria-label is used to replace an existing label with more precise information.

Also, every element has a role attribute that describes the role of an element. This role helps assistive technology (like screen readers) to recognize this element and to present it to a user in order to take appropriate action with it:

Explore

If an element isn't labeled and without a given role it wouldn't be recognized as an interactive element and the user wouldn't be able to make an interaction. Important information would be lost, and the purpose of this page would be defeated.

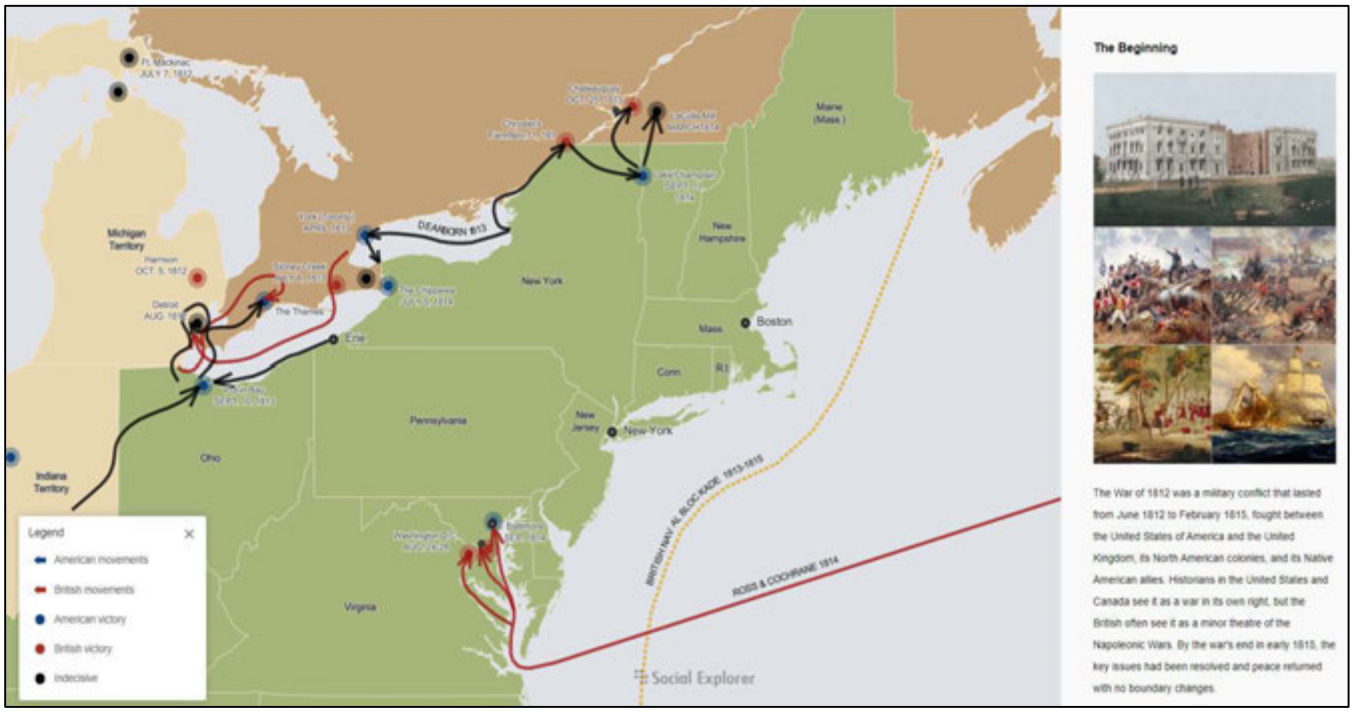


Figure 5 Use of color and contrast on the map

D. Contrast and color

Accessibility standards suggest using appropriate colors and sufficient color contrast between content and background in order to be visible and accessible to visually impaired users.

Best practices include the following segments:

- The color contrast ratio should be at least 4.5:1 for regular-sized text. A recommendation is to use complementary colors that provide enough contrast between content and the background for anyone with low vision or color deficiencies.
- Interactive elements should have other indicators that they are interactive elements, not only color (e.g. underline links). Color should not be used as the sole method of conveying content or distinguishing visual elements.
- Red/green color blindness is the most common and their combination should be avoided [6].

Meeting these requirements, especially contrast ratio, avoids cases when text is difficult to read due to insufficient contrast.

Specialized tools help designers and developers to check color contrast. One of these tools used on most of the maps is Color Brewer. This analyzer has WCAG 2.1 compliance indicators and a color blindness simulator.

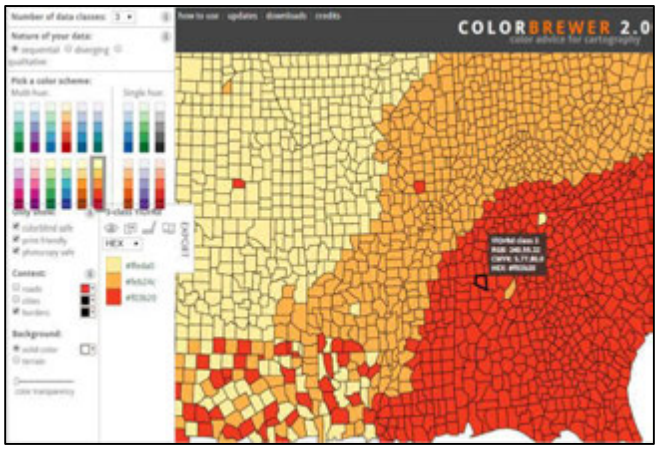


Figure 6. Color Brewer 2.0

Color palettes on the maps are created following standard guidelines using one of the available color schemes tools. Only suggested color palettes are used with an adequate color contrast ratio. Also, every time the color palette is changed it is tested again with a color contrast analyzer in order to check if the criteria of WGCAA are met.

Very important is to check all colors related to the text above colored areas. In order to ensure that every area on the map has sufficient color contrast, there is a complex background color saturation recognition. If the background is less saturated then the label above this area will be dark ad vice versa. In the example below are presented colored countries with their names. For countries with darker shaded area labels are white with shading under the text to make it more visible. Countries with lighter background have black text color with white shading.



Figure 7. Background and label color recognition

When color is used to provide information, this info is also provided with some other method. Usually, color on the map presents either area features (like rivers, mountains, terrain) or data (higher number is presented with a darker color, lower number with a lighter number). To ensure other methods for this info, every area has a tooltip with relevant info. This info usually contains the name of the area, coordinates, data presented and all other info that can gibe useful to the user.

To ensure that sufficient contrast is provided, the text is not placed over complex background areas, and where needed, outline and shading around letters are used.

E. Assistive technology

Users with some kind of impairment, especially visual, use a variety of assistive technology tools, such as screen readers, to access web content.

Using assistive technology is also very useful in accessibility testing that will help developers and QA engineers to understand the barriers some users face when navigating websites or maps.

At the point when the map is loaded, the screen reader will notify a user about content on the map. It will also read all additional descriptions and text as a user navigates the map.

To make digital content accessible a lot of rules have been set up. The screen reader will read first the title and map description and then, moving from left to right, from top to bottom all other relevant elements as a user moves from one element to another using the keyboard.

Alt-text is a written description of an element (image, photo, figure) that describes this element and its essential. The screen reader recognizes native elements on the visualization and reads aloud alt text and area labels to ensure that no information or functionality is lost. Every focused widget on the map should contain information about specific features, content and possible behaviors so the user can decide which action to take [8].

IV. CONCLUSION

Visualizations that present data on the map can be very demanding in the sense of accessibility. There are a lot of elements such as legend, title, descriptions, location pointer, zoom component, annotations. Also, there is a map itself with a lot of data that should be readable and understandable to all users. This map can consist of different colors, areas, elements such as rivers, mountains, cities, points of interest.

This work presents the implementation of accessibility standards in data-driven maps that may contain all of the listed elements. Using a specially developed platform for easy creation, MapSpice, users with and without any impairment can access and use elements and data on these maps. Every component and element of the map, as well as the data on the map, can be accessed and perceived using a mouse or keyboard. Using assistive technology will enable to dive deeper into the content of the map and additional descriptions of elements which is important for users that are not able to recognize it at first.

For future researchers, this paper can present a good example of accessibility standard best practice in data-driven maps, and web applications in general. Also, it can help them to understand how we can improve the accessibility of web visualizations.

Future work in this area is planned to explore the impact and significance of assistive technology in the development and testing of accessible web applications. Also, the role and impact of different web browsers from different vendors will be considered.

REFERENCES

[1] ISO, ISO 9241-171:2008, Ergonomics of human-system interaction, Part 171: Guidance on software accessibility

[2] Chris Felton, DWP, Best practice guide - A comprehensive guide for ICT accessibility standards v0.6, pp 1-16

[3] Raúl Miñóna, Lourdes Moreno, Paloma Martínez, Julio Abascal, An approach to the integration of accessibility requirements into a user interface development method, 3 Elsevier B.V, 2013

[4] W3C: W3C Web Standard Defines Accessibility for Next Generation Web (press release, 11 December 2008)

[5] "Web Content Accessibility Guidelines (WCAG) 2.0". W3C. Retrieved 17 December 2014.

[6] W3C: "Techniques for WCAG 2.0: Techniques and Failures for Web Content Accessibility Guidelines 2.0 ", W3C Working Group Note 7 October 2016

[7] Wan Abdul Rahim Wan Mohd Isa, Ahmad Iqbal Hakim Suhaimi, Nadhirah Ariffrn, Nurul Fatimah Ishak, Accessibility evaluation using Web Content Accessibility Guidelines (WCAG) 2.0, IEEE, Published in: 2016 4th International Conference on User Science and Engineering (i-USEr)

[8] S Sandhya ; K.A. Sumithra Devi, Accessibility evaluation of websites using screen reader, Published in 2011 7th International Conference on Next Generation Web Services Practices

[9] Social Explorer official web page www.socialexplorer.com

[10] MapSpice web page https://company.socialexplorer.com/mapspice

Improvements of Computer Assisted Virtual Environment (CAVE)

M. Fandáková*, K. Zábovská*/**, B. Bučko*/**, M. Zábovský*/**
* University Science Park of the University of Žilina, Žilina, Slovakia
** Faculty of Management Science and Informatics, University of Žilina, Žilina, Slovakia
miriam.fandakova@uniza.sk, katarina.zabovska@uniza.sk, boris.bucko@uniza.sk, michal.zabovsky@uniza.sk

***Abstract* - Virtual environment has become a common part of our everyday lives. From the dashboard projected into the car windshield to smartphones, gaming industry or even training courses. Pushing the limits of virtual reality further has long history and has escalated into the creation of virtual environment called CAVE. This paper presents creation of custom-made solution of computer assisted virtual environment – CAVE. Usage is supported with specific case studies and considers future challenges of this technology.**

Keywords - CAVE; virtual reality; virtual environment, simulation, visualization

I. INTRODUCTION

CAVE was first time designed in 1991 by Thomas DeFanti and Dan Sandin and then implemented by Carolina Cruz-Neira. The initial intend behind this idea was to develop a system which could erase virtual reality limitations such as low image resolution, isolation from the real world and inability to simultaneously share virtual experiences with multiple users in the same time [1].

Back then the word CAVE (if we do not consider natural underground chamber) did not have further meaning but as the development of similar technologies and concepts continued the CAVE has become abbreviation for various solutions like this, for example:

- Computer Assisted Virtual Environment
- Computer Automated Virtual Environment
- Computer-Aided Virtual Environment
- Cave for Automated Virtual Environment

Although the virtual reality cave is not the latest concept it surely brings new point of view when considered within the modern technologies.

II. USAGE

CAVE technology is used not only in a research field but also in the commercial sphere, where covers demand for high quality visualization platform, which could provide virtual surroundings with real time user interaction.

There are many CAVE systems. They differ in many parameters: size, audio system, projection system, tracking system. Generally, we can divide these systems into three categories:

1. **Professional solutions** - Many CAVE system solutions are built using a cluster or multi-core approach and the rendering is spread over many machines. Currently, it is possible to run rendering on a single PC that uses multiple GPUs at once or has an SLI architecture. For example, Nvidia Quadro Plex graphics card as an external graphics processing unit is designed for large-scale 3D visualizations and allows you to create a stereo graphic four-wall system CAVE with parallel walls with a large width, which is known feature of the CAVE system. [2].
2. **Low-cost solutions** - The HIVE (in Hanover) is a low-cost implementation of a small three-wall CAVE setup made at less cost. HIVE like systems are among today's less expensive approaches to build a multi-side virtual display with the usually low budget available for education [3]. Unfortunately, such systems usually are transitioned into very small environments with fewer walls and poorer quality when compared to other more expensive but also superior solutions [4].
3. **Home-made solutions** – There are versions of CAVE systems with only two walls. One of them is V-CAVE, (the name comes from two screens, which form the letter v) [5]. Some variants may also be installed using plastic frames. This solution can be easily assembled and creates portable version of the CAVE system.

III. CAVE ALTERNATIVES

The CAVE system is not the only one visualization solution within the modern technologies. Some of them are well known, such as:

- **Virtual reality (VR)** displays spatial models and scenes within the virtual environment. The aim is to deliver real world experience using VR glasses loaded with virtual content.

- **Augmented reality (AR)** overlays digitally created content on top of the real world and allows users to interact with the real world and digital elements in the same time. AR can be offered to users via various applications installed on a smartphone. These applications are able to visualize AR using smartphone´s video camera [6].
- **Mixed reality (MR)** connects real world with digital elements. In mixed reality users can interact and manipulate with both physical and virtual elements and environments, using next generation sensing and imaging technologies.

Extended reality (XR) is a term which covers all immersive technologies like VR, AR, MR and also those to be created in the future. If we consider CAVE as multi-device technology with various software and hardware elements, we can assume that CAVE also belongs to the Extended reality group.

IV. Our Solution

There are many new technologies and software applications which can be used to improve 3D environment and user interactions in it. Although there are several CAVE solutions, it is not ready-made product which can be directly bought in a market.

Our latest CAVE solution consists of 3 surrounding walls and 1 floor. These 4 surfaces provide enough space for desired visualizations.

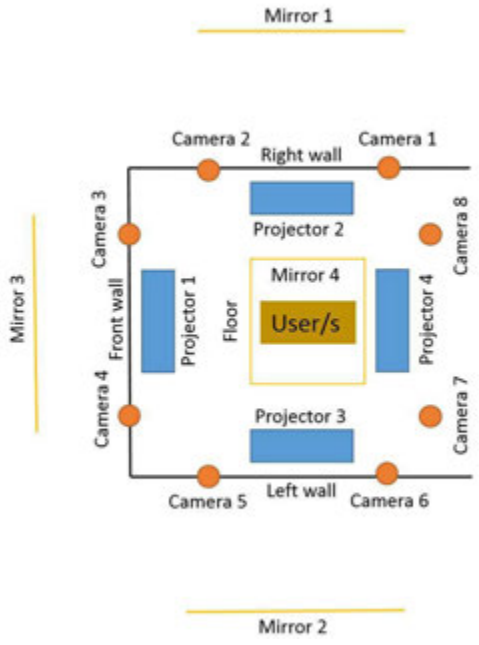


Figure 1 – CAVE top view layout

It is important to mention that we have considered creating CAVE with 4 surrounding walls and floor (cube concept) but we have then decided for 3 walls and floor variant because users could easily lose special orientation and get confused. Also, the closed 4 wall concept would require additional manipulation and calibration as the 1 wall must be detachable in order to allow users to get in/out of the CAVE. This situation could lead to mistakes.

Our first CAVE solution consisted only of three walls and gamepad for motion/user interaction. Later, after few years, within the scope of development of new technologies, we faced the challenge to upgrade our CAVE with floor and tracking system.

There are at least three types of tracking systems:

1. Camera and vision-based algorithm to track markers located on the user's head
2. Magnetic tracking system
3. Multiple viewpoints optical tracking system with active markers



Figure 2 – CAVE room

Camera with vison-based algorithm is progressive technology but, in some cases, lacks reliability and needs maintenance. Magnetic tracking system can run into issues associated with metallic conductivity. That is why we have decided for option number 3. Our CAVE consists of 3 walls, floor, 4 projectors, mirrors and optical tracking system. The top view layout is illustrated in the Figure 1. You can see final solution in the Figure 2.

The whole technology consists of following parts:

A. Hardware

- 4 walls - special acrylic walls which suppress reflectivity.
- Cluster of 4 PC, each PC for one wall, including synchronization module, separate output for left and right eye, one switch for all 4 PC to ensure outlook.
- 4 mirrors – special metal surfaces very sensitive, they cannot be touched, because of sensitive layer, even fingerprints would leave a trace and it would appear in the cave permanently.
- 4 active 3D HD projectors with remote control, 3 projectors with long lenses, one for the floor with short lens. There is a stereo synchronization at front projector. Basic principle of 3D active projection is separation

of the image for the left and right eye by transmitting two images at different times.

- Active 3D glasses with RF transmitter, connected to front projector, including tracking marks.
- Wireless gamepad with tracking marks.
- Control computer with RF transmitter for gamepad, software and 3 network adapters for: cluster, tracking software, global network.
- Optical tracking position system – 8 cameras with active IR flashes including image evaluation and data export to tracking software. There is a scheme of tracking system in the figure 2 below.

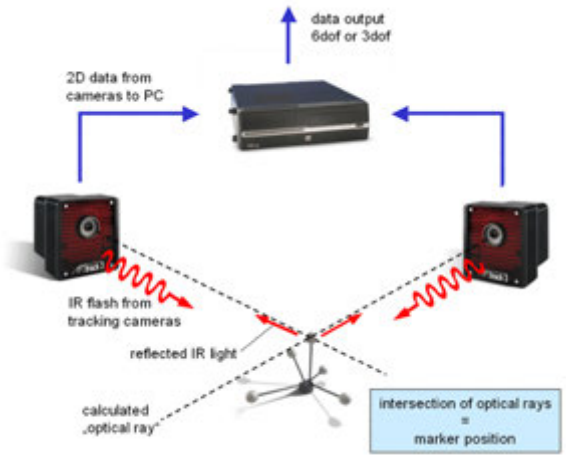


Figure 3 - IR optical tracking principle [7]

B. Software

- Cluster Launcher and calibration programs based at CAVELib - Calibration pattern for projector settings, functionality check of stereo 3D projection, shut down cluster of 4 PC.
- Tracking software - multi-camera rigid body tracking, 3 markers minimum per rigid body for 6 degrees of freedom can track over 300 rigid bodies in real time.
- Tracking communication – data from tracking system and gamepad are converted to data for visualization software.
- Projector remote control software – control panel enables to turn on and off all projectors, this is highly recommended, because manual switching can move projectors and visualization may be disturbed.
- Visualization software – all 3D models are loaded, constrains are defined, visualization scene is created and exported via plugin.
- Startup software – exported data from visualization software are synchronously loaded to projectors and walls.

V. CASE STUDIES

In past, visualized objects were created in 3D parametric or explicit CAD software. With development of 3D laser scanners, there is a possibility to create scanned point clouds of all components of 3D environment including buildings with equipment and all physical objects which can be scanned.

Laser scanning has significantly accelerated creation of 3D virtual models. Point clouds are with help of reverse engineering in 3D CAD applications processed into 3D models. Later, after conversion to corresponding formats, are these models uploaded to visualization software for scene creation. After the synchronization of a scene in the CAVE projection, the virtual environment is prepared for desired exploration.

There are many areas (architecture, automotive and different kinds of industry, research, medicine) in which CAVE projection can be used for visualization, exploration, analysis, optimization and simulation. Below are mentioned some examples and projects, we have been working on.

A. Reconfiguration and Optimization of Manufacturing Plant Workplaces

After creation of digital twin of manufacturing plant workplaces, there is a quick possibility to reconfigure workplaces and evaluate, if workplace corresponds to desired change. Mentioned use case is illustrated in the Figure 4. Edges between the walls are barely visible, therefore we have highlighted transitions between the front wall, left wall and the floor. This can be seen within the thumbnail in the Figure 4.

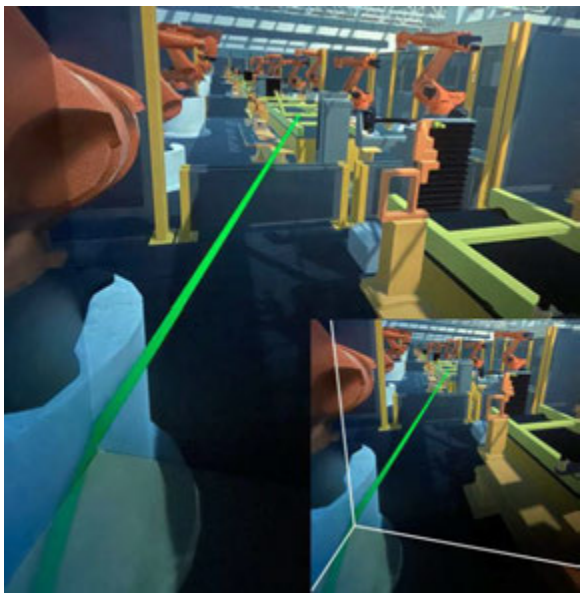


Figure 4 - Reconfiguration and optimization of workplaces

B. *Optimization of virtual buildings*

Exploration of newly generated 3D virtual prototypes of the buildings can among other things lead to shape correction of concrete stairs from bottom view to improve style, long before building of the foundation stone. Figure 5 represents virtual house simulation.

C. *Optimization of car interior*

Many car companies use these visualizations with average, tall, short, fat people to analyze their cone of view, reach cone and other ergonomic analyses. They can also compare these simulations to virtual simulations, created in 3D CAD software. Figure 6 refers to car interior visualization.



Figure 5 – Virtual house with visible IR cameras (blue circles) at the top

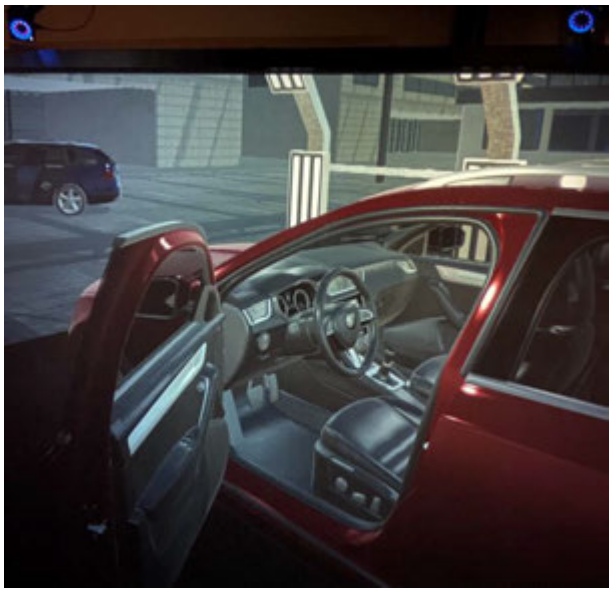


Figure 6 – Car interior with visible IR cameras (blue circles) at the top

D. *Impact of virtual environment to human body physical functions*

The virtual environment can induce various mental conditions in humans that are closely related to a change in physiological functions. These are for example adrenaline, panic, fear or courage. Some people have different health problems when using VR glasses including headaches, dizziness, increased heartbeat, etc. Many users are relieved of these problems when using CAVE, and in some cases they do not even occur. In the early stages of physiological function monitoring by non-contact sensors (surface EMG, rhythm and respiratory amplitude, skin surface temperature), CAVE appears to be a more acceptable alternative than VR systems. This study is under investigation from various perspectives. Figure 7 shows gaming mode in CAVE.

VI. FUTURE CHALLENGES

As mentioned above, CAVE is not ready solution, which can be simply bought. It is a complex system that allows various device variables.

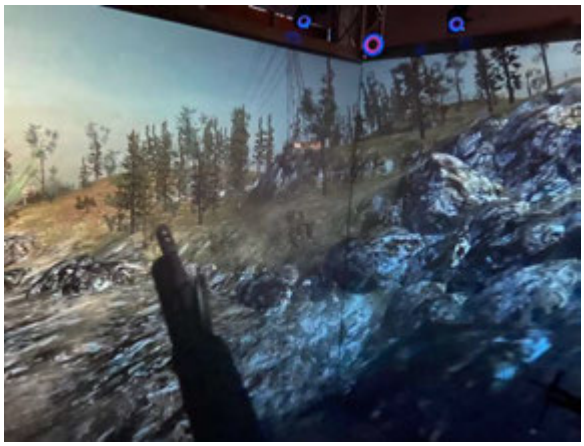


Figure 7 – Gaming mode with visible IR cameras (blue circles) at the top

With the development of technological solutions, it is possible to integrate ostensibly incompatible devices with the help of various programmed applications. It is important that the devices are open source.

1. **Upgrade of walls**
   Projectors and acrylic walls which suppress reflectivity can be replaced by LCD monitors and panels, top wall can be added.
   Advantages: better resolution of LCD´s, higher luminous intensity, the room does not have to be dimmed to the same level as when dealing with projectors, which greatly facilitates the work of the operator or discussion of work teams.
   Disadvantages: LCD screens and panels are connected through the edges and slats. Those can disturb perception of the virtual environment, final price is much higher.

2. **Replacement of gamepad**
   Remote control gamepad can be replaced with other kinds of remote input VR devices. There is a wide range of devices available and those

selected will be affected by the display device that has been chosen for the system. They can be divided into these groups:
Joysticks – basically held in both hands at the same time.
Hand-held wands – held in one hand.
Gloves – there is an option to use one glove or two, according to a request.

VII. CONLUSION

Virtual reality develops still very quickly, and its possibilities rapidly grows. Also, it is more available for wider range of users. This leads to high expectations.

Many companies are trying to develop devices that are compatible with multiple types of virtual reality. The next challenges will consider integration of the CAVE into an environment that allows users to accomplish a given tasks collaboratively, even when they are geographically dispersed. Using mentioned approach, the user in CAVE can share visual experience remotely with another elsewhere person. Shared visualization would be interpreted via a Head-Mounted-Display or 2D display.

ACKNOWLEDGMENT

"This publication was realized with support of Operational Program. Research and Innovation in frame of the project: ICTproducts for intelligent systems communication, code ITMS2014+ 313011T413, co-financed by the European Regional Development Fund".



REFERENCES

[1] Cruz-Neira, Carolina & Leigh, Jason & Papka, Michael & Barnes, Craig & Cohen, S.M. & Das, S. & Engelmann, R. & Hudson, R. & Roy, Trina & Siegel, L. & Vasilakis, C. & DeFanti, T.A. & Sandin, Daniel. (1993). Scientists in wonderland: A report on visualization applications in the CAVE virtual reality environment. 59 - 66. 10.1109/VRAIS.1993.378262. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Kageyama, Akira & Masada, Youhei. (2013). Applications and a Three-dimensional Desktop Environment for an Immersive Virtual Reality System. Journal of Physics Conference Series. 454. 10.1088/1742-6596/454/1/012077.

[3] Daniel Cliburn, John Krantz, "Towards an effective low-cost virtual reality display system for education", J. Comput. Small Coll 2008, pp. 147-153. (3) Siddhesh Manjrekar, Shubhrika Sandilya, Deesha Bhosale, Sravanthi Kanchi, Adwait Pitkar, Mayur Gondhalekar: CAVE: An Emerging Immersive Technology - A Review, 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, ISBN 978-1-4799-4923-6.

[4] Siddhesh Manjrekar, Shubhrika Sandilya, Deesha Bhosale, Sravanthi Kanchi, Adwait Pitkar, Mayur Gondhalekar: CAVE: An Emerging Immersive Technology -- A Review, 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, ISBN 978-1-4799-4923-6

[5] Jeffrey Jacobson, "Using caveut to build immersive displays with the unreal tournament engine and a pc cluster", I3D '03: Proceedings of the 2003 symposium on Interactive 3D graphics, pp. 221-222.

[6] Unity 3D, "What is AR, VR, MR, XR, 360?", https://unity3d.com/what-is-xr-glossary

[7] Figueroa Fernandez, Nadia Barbara (2012) 3D Registration for Verification of Humanoid Justin's Upper Body Kinematics. Master's. DLR-Interner Bericht. DLR-IB 572-2012/03, 133 S.

APPLICATIONS

The Order Batching Concept Implemented In Real Smart Warehouse

S. Delalić[1,2], E. Žunić[2,3], A. Alihodžić[1] and E. Selmanović[1]
[1] Faculty of Science, University of Sarajevo, Sarajevo, Bosnia and Herzegovina
[2] Info Studio d.o.o., Sarajevo, Bosnia and Herzegovina
[3] Faculty of Electrical Engineering, University of Sarajevo, Sarajevo, Bosnia and Herzegovina
Email: delalic.sead@pmf.unsa.ba, emir.zunic@infostudio.ba, adis.alihodzic@pmf.unsa.ba, eselmanovic@pmf.unsa.ba

***Abstract* - Distribution companies often store goods in large warehouses. Orders are collected and prepared for transport. Large-scale warehouses are often divided into sectors. Each worker collects a part of the order from the assigned sector. In that case, workers often pick small orders and the process is not optimal. Therefore, order batching is done, where one worker collects multiple orders at a time. In this paper, an innovative concept of orders batching in a warehouse with a 48-hour delivery based on a metaheuristic approach is described. The algorithm divides each order by sectors. An analysis of each part of the order is done and the possibility of batching based on the order content is checked. The order batching is based on the discrete Bat algorithm. The transport scheme and the order of loading goods into the truck are observed. In the order picking process, a number of standard constraints such as weight and item priorities are considered. The concept has been implemented and tested for 50 days of warehouse operation in one of the largest warehouses in Bosnia and Herzegovina. The algorithm is compared with the earlier approach of collecting orders in the warehouse, and significant progress has been observed in the number of kilometers traveled on a daily basis.**

Keywords - Order Batching; Order Picking; Smart Warehouse Concept; Large-Scale Warehouse; Bat algorithm

I. INTRODUCTION

The workflow of distribution companies consists of several significant processes. Companies order large quantities of goods from manufacturers and store them in the warehouse. Shopping centers, markets and other customers order goods from a distributor. The goods are collected from the warehouse, packed for transport, and delivered to customers.

The operation of the warehouse also consists of several processes, such as the process of recording and arranging goods at the entrance, the process of goods relocation within the warehouse, and the process of collecting goods after orders are received. When the orders are collected, items are controlled and placed on the final location for transportation. Each of the above processes in the warehouse can be optimized. Of all the above processes, the process of collecting orders takes the most time. This process is usually referred as the order picking and based on [1], the retrieval of goods takes about 55% of the operating costs.

The goods are stored in pallets in the warehouse. The pallet places are arranged on racks that can consist of multiple floors. The part of the warehouse where workers can collect goods without the use of forklifts is called the picking zone, while the other zone is called the stock zone. Between the racks, there are aisles. Workers usually walk through the aisles and collect the required items to complete the order. As the orders usually consist of a large number of items, workers use carts for easier transport of goods.

A large number of customers are served from the warehouse on a daily basis. Orders may contain items located at different locations in the warehouse. Therefore, the order picking process can be a tedious process for one worker. To simplify the process of collecting goods, various solutions are available. One of the most common solutions is the division of the warehouse layout into sectors. Sectors are formed by article type or based on physically separated rooms, such as temperature chambers.

Although the division into sectors has several advantages, such as reducing psychological pressure for the worker, and the fact that workers collect items in a smaller part rather than in a complete warehouse, division into sectors does not improve the efficiency in the order picking process. After division, it is common to find a small number of items in a sector. Workers then go through the sector to collect only several items, and these new orders are not balanced.

One of the most efficient approaches to order picking optimization is to merge multiple orders within a sector. In order for this process to be feasible, it is necessary to have a cart with separate boxes for collecting multiple orders. The rest of the process is no different from collecting orders that are not merged. The process of collecting multiple orders at once is called the order batching. One example of the order batching is given in Figure 1.



Figure 1. Order batching in the warehouse sector (a) first order (b) second order (c) batch

This paper describes the concept of order batching implemented for some of the largest warehouses in Bosnia and Herzegovina. Warehouses are divided into sectors and the process is adapted to the way warehouses operate, although the same process can be used in warehouses not divided by sectors. In this case, the entire warehouse can be seen as one sector. The warehouse has a 48-hour delivery system, where customers place their orders one day; workers collect orders the next day and transport them a day later. The algorithm for order batching is based on the Bat algorithm.

The paper consists of five sections. In the current section, a short introduction is given and the common processes in the warehouse are described. In the second section, a related work is described. In the third section, the implemented concept based on the Bat algorithm is described in details. In the fourth section, the results are given, such as the comparison between implemented algorithm and one of the most widely used practice approaches. In the final section, a conclusion and future work are observed and described.

II. Related Work

In the paper [2], the concept of smart warehouse and processes that can be optimized are presented. The observed warehouses are picker-to-parts systems, where the picker walks through the warehouse and collect orders. In larger warehouses, it can be a tedious job and orders are split up into smaller ones. There are multiple approaches to divide an order, such as solution based on Ant Colony Optimization presented in [3] or simply by dividing the order after the picker route is created. However, in practice, simple division into sectors is most commonly used and easiest to implement among pickers. The smart product positioning can increase the order picking efficiency [4].

One of the most important factors for optimizing warehouse operation is the arrangement of storage racks and aisles. In practice, standard and traditional layouts consist of racks that are lined up in parallel. The aisles are lined up and they pass through the middle of the warehouse. In paper [5], Flying-V and Fishbone layout are proven to have advantages over traditional layouts. In practice, some other non-standard layouts are encountered, and it is essential to deal with them when planning warehouse operations. In those layouts, racks are not always parallel and aligned to start from the same position.

The problem of optimal order picking can be reduced to the process of solving the Traveling Salesman Problem. This problem largely depends on calculating the distance in the warehouses. One approach for calculating distances in the warehouse of general layout is described in [6]. The algorithm is based on dynamic programming and the efficient use of this algorithm is necessary for the order batching process and the testing phase.

In [7], the order batching and zoning problems are defined. Many benefits of using mentioned approaches together are also described. The order batching problem is a well-researched area in the literature as significant cost savings can be achieved through process optimization. Many different approaches have been developed and described. As mentioned earlier, there are two standard batching types: static (off-line) and dynamic (on-line). When orders are not known in advance, the batching is called on-line. Benefits of the integrated routing and order batching are described in [8]. This paper focuses mainly on static warehouses with 48-hour delivery, where the orders to be picked are known in advance.

In the paper [9], a comprehensive description of the problem and a classification of the solution algorithms are given. The order batching in the NP-hard problem, so many heuristics are implemented. These solutions can be classified into constructive solution approaches and metaheuristics. Priority rule-based algorithms, seed algorithms, saving algorithms based on the Clarke and Wright algorithm and data mining approach belong to the first group of solutions. These approaches are mostly simple to implement, but at the same time make significant savings in the ordering process. The second group includes various metaheuristic approaches such as Local search, Tabu search or population-based approaches such as genetic algorithm (GA), which can lead to additional savings.

In the literature, many other metaheuristics are used for order batching. Iterated Local Search and Attribute-based Hill Climber are used in [10]. The last approach is based on the Tabu search. Mentioned approaches can improve picking process by 46% on average compared to the standard constructive heuristics. In the paper [11], and Ant Colony Optimization was implemented and compared to the Iterated Local Search approach. In the paper [12], the GA is used to solve the batching problem for different warehouse layouts. Similar approach is described in [13], where the grouping GA is implemented.

In the [14], a heuristic algorithm based on a general graph model is described and implemented in the warehouse for the large e-commerce business in Central Europe. In the paper [15], a simulated annealing algorithm combined with the A* algorithm is implemented. The algorithm produced results that are 1.2% far from the optimal. The algorithm was tested in a Finnish warehouse and this algorithm has made a saving of 5000 kilometers in distance traveled in 3 months (~16%). Simulated annealing and genetic algorithm were also used to solve the order batching problem in [16].

Many nature-inspired metaheuristics are developed. Algorithms often mimic swarm traits and the collaboration of individuals. The algorithms are based on directed search of the solution space, which can be significantly reduced. Some of the most famous algorithms are Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Artificial Bee Colony (ABC), Bat Algorithm (BA), Firefly Algorithm (FA), Elephant Herding Optimization (EHO)

and others. Some of these algorithms and their properties are described in detail in [17].

III. CASE STUDY

This paper describes the complete order batching concept implemented as part of the warehouse management system (WMS) in some of the biggest warehouses in Bosnia and Herzegovina. The complete workflow consists of several steps, as shown in Figure 2.

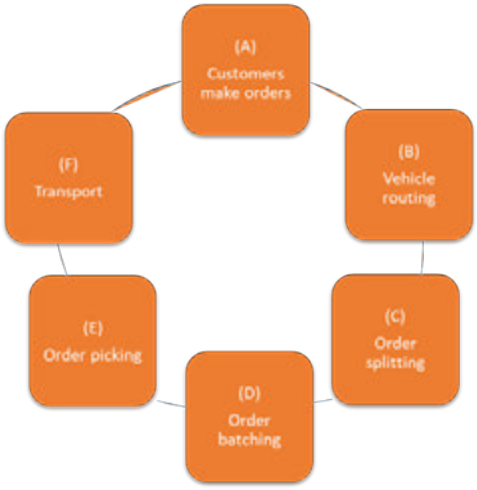


Figure 2. The order batching concept

A. Customers make orders

Distribution companies usually have customers who regularly place orders. Traveling salesmen visit customers [18] and collect orders or they are placed online. In any case, orders are saved in the WMS. The warehouses operate on the principles of 24-hour or 48-hour delivery. Observed warehouses operate on a 48-hour delivery principle, which means that orders are created on the first day, collected on the second, and transported on the third day. Therefore, orders are delivered to customers within 48 hours. The above practice allows off-line order batching principle to be used, as the orders that are collected are known in advance. The concept can also be implemented in 24-hour warehouses with minor modifications to work organization.

B. Vehicle routing

The biggest financial cost in the process of delivering is transporting the goods from the warehouse to the final locations of the customers. Therefore, the priority process that needs to be optimized is the process of routing transport vehicles, determining which vehicle should visit which customer, and in which order the customers are visited in the route. As part of this process, many constraints such as time-windows, vehicle capacity, and driver working hours must be respected. The ultimate goal is to create the most cost-effective routes. In the literature, the mentioned problem is called Vehicle Routing Problem (VRP). VRP is one of the most difficult problems of combinatorial optimization in the literature. In the smart WMS concept, the transport optimization is a separate application module that allows routes to be created by the transport manager and allows manual modification of the routes created. At the end of the process, the final driving routes are created and the preparation of goods for collection and transport begins.

C. Order splitting

When the orders are stored in the WMS and when the transportation routes are created, orders are divided according to the established principle. In practice, orders are most often divided by sectors. This concept implies division based on previously defined sectors.

For each article, the sectors has to be defined. In practice, this concept allows two ways to determine sector for each item. In the first case, articles are in fixed sectors such as “food”, “cosmetics” etc. In each order and in any moment, the article is in the same sector. In the second case, the sector can vary for article in different moments. In some warehouses, sectors are created based on the physical location of racks. This means that sometimes sectors can be determined only in the moment when the order splitting process begins. In this case, the location (pallet place) from which the item will be collected is determined. Then the item sector is taken to be the pallet place sector.

Each order is divided by sectors and smaller orders are created and labeled. As the process continues, the system only observes newly created orders.

D. Order batching

The process of order batching is the main process to be described in this paper. In this paper, the order batching problem is solved using the metaheuristic Bat algorithm (BA). To our knowledge, this is the first use of the Bat algorithm to solve the order batching problem in literature.

The Bat algorithm was developed in 2010 and it belongs to the group of nature-inspired metaheuristic algorithms for global optimization. Algorithm was developed by Xin-She Yang [19]. The algorithm was inspired by the behavior of microbats. Microbats have special echolocation behavior that can be presented through a series of rules implemented as part of the algorithm. In nature, bats send pulses into the environment with the primary purpose of hunting and navigating through space. After that, the bats listen to the echo and, based on the signals they receive, determine their location, location of obstacles and prey. The main inspiration for the bat algorithm is the ability for each individual to find the best area of the search space of all the bats in the swarm, and direct their movement toward that area.

Although the Bat algorithm is primarily intended to solve problems in the area of continuous optimization, many examples can be found in the literature where the algorithm is used to solve combinatorial optimization problems. In these cases, a modification of the original algorithm is required. In paper [20], an improved discrete Bat algorithm is used for symmetric and asymmetric Traveling salesman problem (TSP), which is one of the most important problems in the field of combinatorial optimization.

This paper is inspired by the use of the Bat algorithm to solve the Traveling salesman problem, and in a similar way,

the order batching problem is approached. The complete pseudocode for the implemented approach is given in Algorithm 1.

Algorithm 1 The modified BA for Order Batching

```
function runBatAlgorithm
begin
   Define the objective function f(x)
   Initialize the bat population X = x_1, x_2 ... x_n
   Find the best bat in population x_best

   for each bat x_i in the population X do
         Initialize the pulse rate r_i
         Initialize the velocity v_i
         Initialize the loudness A_i
   end

   repeat
    for each bat x_i in the population X do
     Generate new solution
     Generate the random number rand
     if rand > r_i then
      Generate solution around the best bat
     end

     if rand < A_i and f(x_i) < f(x_best) then
      Accept the new solution
      Increase r_i
      Reduce A_i
     end
end
```

As shown in the pseudocode, the implemented Bat algorithm has been modified from its original form. Each part of the algorithm should be tailored to the discrete version, and it must reflect the core of the problem being solved. Each bat represents one solution. The solution is represented by the order schedule in each batch. Each order belongs to exactly one batch.

First, the objective function has to be determined. The objective function is the function that has to be minimize in our implementation, and when the function is minimized, the solution is acceptable (and ideally optimal). The main goal is to decrease the walked distance for pickers. Because in the batching process for one vehicle, the number of locations to be visited can increase up to 1000 or more different locations in the sector, the distance calculation between each two location in the warehouse is not an easy and fast task. In addition, in every execution of the goal function, it would be necessary to solve the TSP problem for optimal order picking. In this paper, the objective function is calculated in a different way.

As part of the objective function, limitations are included. First, the number of boxes in a picker cart is limited. The picker is not able to collect more than a constant number of orders at the same time. The volume and mass of orders are limited, so the complete batch cannot exceed the parameters determined by the cart size and capacity. Because the total tour length is not calculated, the goal is set to visit the minimum number of racks in a batch, and at the same time to visit the minimum number of racks for all orders. For each rack, the goal is to minimize the distance between the farthest pallet locations. This way, a fast and easy-to-implement objective function is created, but the objective function that guarantees that the walked distance is reduced.

In the case of fewer locations, the goal function can be calculated as a solution to the TSP problem for a set of locations. In doing so, a simple 2-opt algorithm can be used to solve the TSP problem. In any case, this can be a challenging task because of the number of invocations of the goal function during the Bat algorithm.

When the objective function is determined, the population is initialized. The initial population was created by shuffling all orders, and shuffled orders were added to the batch until the batch capacity was filled. When one batch is filled, a new batch is created. This process continues until all orders are assigned to a batch.

The best bat in the population is determined as the bat with best objective function value. Pulse rate and loudness are determined by the standard formula used for the BA. At the same time, the velocity for each bat has to be determined differently. The velocity of a bat depends on the distance between the best bat, and the velocity in the previous iteration. Each order is placed in one batch in one population. Let a and b be the bats between which the distance is calculated. For each order, a corresponding batch in the bat a and b is checked. For each order in the observed batches, if they are not contained in both observed batches, the distance is increased by a constant value.

One of the steps is to generate new solution from the existing one. When generating the new solution, one of three methods with equal probability is used:

- two randomly chosen orders swap batches;
- one randomly selected order leaves the batch and creates a new one;
- one randomly selected order leaves the batch and joins another;

The process of increasing pulse rate and reducing loudness is implemented in a standard way by the formula given in [19] and [20].

At the end of the process of order picking, the best bat is selected and used as the final solution.

E. Order picking

When the orders are divided into smaller orders, and when the smaller orders are batched, the worker gets a list and order of the final locations to visit. For each location, the item information and associated order within the batch are listed. The worker then walks through the aisles and pick items into the appropriate box. Each box is labeled based on the original order and contains only items from the corresponding order.

When all locations are visited and when orders are picked, workers leave the collected orders at the designated loading points and start collecting the next free order.

As stated earlier, orders are collected for one vehicle for ease of control. When all orders for one vehicle are picked, the controller checks that all items have been correctly collected and placed in the appropriate lading location. The

TABLE I. RESULTS OBTAINED FROM REAL-WORLD WAREHOUSE

Observed property	*Value*
Number of days and orders	50 / 8200
Average number of orders per vehicle and day	28.38
Total tour length for APP1	2068.4 km
Total tour length for APP2	955.2 km
Total tour length for the Bat algorithm	802.3 km
Average BA route length improvement over APP1	**61.2%**
Average BA route length improvement over APP2	16.01%

cost of the controller does not increase the cost of warehouse operation, as the controller is obliged to control all outbound orders regardless of the earlier collection process.

F. Transport

The day of delivery begins with loading of prepared orders into previously specified vehicles. The vehicles follow the routes created and deliver orders to end customers. This completes the 48-hour concept, which consists of creating routes for routing vehicles, splitting orders, merging orders into batches, collecting batches based on created routes for vehicles and the delivery process.

IV. RESULTS

In this section, results of the previously described algorithm for the order batching and the complete concept are described. The concept described has been introduced gradually in the smart WMS. First, the order picking process was optimized and the algorithm for finding the optimal picking route was implemented. Then, the vehicle routing problem was introduced, and the complete process of transport optimization, including the tracking system and VRP solver improvements. When larger warehouses were introduced, few approaches for the order splitting process were implemented. The process of dividing orders into sectors is mainly used because it is easiest to implement and deploy among workers in warehouses. At the same time, other algorithms described in [2] are implemented in the complete concept of smart warehouse.

During the process, various and useful information about the processes in the warehouse is collected. Some of these are information about order details, order splitting and picking process details etc. The order batching algorithm was introduced as the last improvement of the complete warehouse management system.

The results of the concept implementation have been tested in one of the largest warehouses in Bosnia and Herzegovina. The observed warehouse serves several hundred customers daily. Once transport routes have been created, each vehicle delivers orders for up to 30 customers. Due to time constraints and limited capacity, the vehicle is not able to visit more customers. Once transport routes have been created, each vehicle ships orders for up to 30 customers. Due to time constraints and limited capacity, the vehicle is not able to visit more customers.

The observed warehouse is divided into six sectors. Each sector is created based on the type of items stored in it. After the division of orders into sectors, the number of orders increases. However, the number of orders cannot be increased for each sector. In the next step, the merging of orders for each vehicle and sector is approached.

Bat algorithm for the order batching process was tested on different vehicles and sectors through 50 days of warehouse operation. More than 280 vehicle routes and 8200 real orders were observed during the testing phase.

When implementing the bat algorithm, it is necessary to adjust the parameters. The population size is set to be 50. Initial loudness for each bat is set as a random real number between 0.7 and 1. The initial value for pulse rate was set as a random number between 0 and 0.4. Values α and γ where set to 0.98. The bat algorithm was run exactly 10 times for each instance. The maximum number of iteration was set to 30 000. The best result is further processed.

The algorithm assumes that four orders can fit in one cart at a time, and that the total mass of coupled orders is less than 100 kilograms. The volume is neglected due to the dimensions of the carts and the dimensions of standard orders in the warehouse.

The algorithm results were compared with two other approaches. In the first approach, the orders were not batched (APP1). In this case, the performance of the warehouse was compared before and after implementing the order batching algorithm. Orders were collected one by one. In the second approach, orders are sorted by creation time and batches are created by a simple rule: orders are added to the batch until the batch is full, in which case a new batch is created (APP2).

After the batches were created using all three approaches, an algorithm was used to find the optimal order picking route in the warehouse and the total length of the routes obtained were calculated. A comparison of the obtained values was made. For testing, an algorithm was implemented that optimize the order picking process. The route starts from the position of taking the order documentation and ends at the final place for the order picking process (place marked for loading in vehicles).

Complete results are shown in Table 1. As we can see, the first approach is less efficient than other two. This was to be expected, because the order batching has the main goal to reduce the walking route length. The length of routes obtained in the case of separate collection of each order can be reduced by more than 50% in the case of multiple orders being collected simultaneously.

In approximately 97% of tested cases, the Bat algorithm performed better than the APP2 algorithm. The total path length for collecting the tested orders was 16.01% shorter for the BA than the path length for collecting the same orders using the APP2 algorithm.

The Bat algorithm has shown stability in the solutions obtained. Deviations from the best results were rare for different runs. It has been observed that the algorithm often

finds the best solution when the number of iterations is greater than or equal to 20,000.

As part of the test, 11 different vehicles were used, including smaller transport vehicles and trucks. Of course, the amount of goods and the number of orders required to collect depends on the size of the vehicle. For the 50 days observed, the average number of shipment orders was 164.13.

In the case of non-batching orders, the total length of the picking routes to collect all orders in observed 50 days is about 2068 kilometers. In the case of APP2 approach, significant savings are achieved, with a total length of just over 955 kilometers. In the case of merging orders based on the Bat algorithm, additional savings of 153 kilometers over APP2 are achieved. Compared to the previously used APP1, it is possible to achieve savings of up to 60% using the Bat algorithm.

After a thorough analysis of the proposed routes based on the Bat algorithm, it was found that the routes are practically usable and that they give satisfactory order batches. Due to the number of combinations, it has been noted that the routes obtained are difficult to test even for experienced workers and managers.

V. CONLUSION

The paper describes the warehouse processes starting from the concept of accepting orders, through the concept of collecting orders, all the way to packaging and transport. The focus of the paper is the concept of batching orders implemented at picker-to-parts manual warehouses in Bosnia and Herzegovina. The paper offers an innovative approach to the warehouse operation process and optimization of the most challenging order picking process. The algorithm is based on the nature-inspired metaheuristic Bat algorithm. The results of the implemented process are satisfying and the concept gave good results in practice.

The problem of order batching is one of the problems whose improved results can produce significant savings in the speed of the warehouse operation process. Optimal work establishment can also improve employee satisfaction, further enhancing the importance of order batching.

Future research will focus on implementing and testing other nature-inspired metaheuristics to solve the order batching problem, such as firefly algorithm, elephant herding optimization etc. In addition, the objective function will be further modified so that the implemented bat algorithm can produce even better results. Other ways to optimize warehouse performance and improve workflow will be researched.

REFERENCES

[1] Bartholdi III, J. J., & Hackman, S. T. (2008). *Warehouse and distribution science*. Supply Chain and Logistics Institute, School of Industrial and Systems Engineering, Georgia Institute of Technology.

[2] Žunić, E., Delalić, S., Hodžić, K., Beširević, A., & Hindija, H. (2018, November). Smart Warehouse Management System Concept with Implementation. In *2018 14th Symposium on Neural Networks and Applications (NEUREL)* (pp. 1-5). IEEE.

[3] Žunic, E., Delalic, S., Tucakovic, Z., Hodžic, K., & Beširevic, A. (2019). Innovative Modular Approach based on Vehicle Routing Problem and Ant Colony Optimization for Order Splitting in Real Warehouses. *Annals of Computer Science and Information Systems*, *20*, 125-129.

[4] Zunic, E., Hasic, H., Hodzic, K., Delalic, S., & Besirevic, A. (2018, May). Predictive analysis based approach for optimal warehouse product positioning. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 0950-0954). IEEE.

[5] Pohl, L. M., Meller, R. D., & Gue, K. R. (2009). Optimizing fishbone aisles for dual-command operations in a warehouse. *Naval Research Logistics (NRL)*, *56*(5), 389-403.

[6] Žunić, E., Beširević, A., Delalić, S., Hodžić, K., & Hasić, H. (2018, May). A generic approach for order picking optimization process in different warehouse layouts. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1000-1005). IEEE.

[7] Yu, Mengfei M. *Enhancing warehouse performance by efficient order picking*. No. EPS-2008-139-LIS. 2008.

[8] Scholz, A., & Wäscher, G. (2017). Order Batching and Picker Routing in manual order picking systems: the benefits of integrated routing. *Central European Journal of Operations Research*, *25*(2), 491-520.

[9] Henn, S., Koch, S., & Wäscher, G. (2012). Order batching in order picking warehouses: a survey of solution approaches. In *Warehousing in the global supply chain* (pp. 105-137). Springer, London.

[10] Henn, Sebastian, and Verena Schmid. "Metaheuristics for order batching and sequencing in manual order picking systems." *Computers & Industrial Engineering* 66.2 (2013): 338-351.

[11] Henn, Sebastian, et al. "Metaheuristics for the order batching problem in manual order picking systems." *Business Research* 3.1 (2010): 82-105.

[12] Hsu, Chih-Ming, Kai-Ying Chen, and Mu-Chen Chen. "Batching orders in warehouses by minimizing travel distance with genetic algorithms." *Computers in Industry* 56.2 (2005): 169-178.

[13] Koch, Sören, and Gerhard Wäscher. "A grouping genetic algorithm for the Order Batching Problem in distribution warehouses." *Journal of Business Economics* 86.1-2 (2016): 131-153.

[14] Pferschy, Ulrich, and Joachim Schauer. "Order batching and routing in a non-standard warehouse." *Electronic Notes in Discrete Mathematics* 69 (2018): 125-132.

[15] Matusiak, M., de Koster, R., Kroon, L., & Saarinen, J. (2014). A fast simulated annealing method for batching precedence-constrained customer orders in a warehouse. *European Journal of Operational Research*, *236*(3), 968-977.

[16] Ardjmand, E., Bajgiran, O. S., & Youssef, E. (2019). Using list-based simulated annealing and genetic algorithm for order batching and picker routing in put wall based picking systems. *Applied Soft Computing*, *75*, 106-119.

[17] Yang, X. S. (2010). *Nature-inspired metaheuristic algorithms*. Luniver press.

[18] Zunic, E., Hasic, H., & Delalic, S. (2020, January). Strategic Development, Improvement and Integration of Representative Routing in Warehouse Management Information Systems. In *Proceedings of the 2020 the 3rd International Conference on Computers in Management and Business* (pp. 46-52).

[19] Yang, X. S. (2010). A new metaheuristic bat-inspired algorithm. In *Nature inspired cooperative strategies for optimization (NICSO 2010)* (pp. 65-74). Springer, Berlin, Heidelberg.

[20] Osaba, E., Yang, X. S., Diaz, F., Lopez-Garcia, P., & Carballedo, R. (2016). An improved discrete bat algorithm for symmetric and asymmetric traveling salesman problems. *Engineering Applications of Artificial Intelligence*, *48*, 59-71.

Integration of photogrammetry within laser scanning approach

P. Kudela*, M. Palčák*, K. Zábovská*/**, B. Bučko*/**
*University Science Park of the University of Žilina, Žilina, Slovakia
**Faculty of Management Science and Informatics, University of Žilina, Žilina, Slovakia
pavol.kudela@uniza.sk, michal.palcak@uniza.sk, katarina.zabovska@uniza.sk, boris.bucko@uniza.sk

***Abstract* - This article is focused on photogrammetry and its basic parameters for achieving the highest quality output and subsequent application in practice with 3D scanner interconnection. Photogrammetry along with 3D digital spatial scanning is nowadays in demand. Output is more than just a 3D model. It is the actual state of the photographed and scanned object, which is accurate to one thousandths of a millimeter. This procedure is used mainly for the reconstruction of historical buildings, statues, sculptures, but also for the construction of new projects. Photogrammetry consists of several methods and procedures that are directly influenced by software and hardware. The main fields of application are construction, urbanism, forensic science, medical industry, aerospace industry. Photogrammetry can be used in small as well as in complex projects to create visualization of changes created by human or natural activity.**

Keywords - photogrammetry, laser scanning, visualization, iris

I. INTRODUCTION

Photogrammetry [1] has long been used as a tool for collecting three-dimensional (3D) information of objects as well as texture information. The 3D coordinates of points on an object surface are determined on overlapping images with camera position and orientation information known as exterior orientation. The parameters of interior orientation that describe the principal point location and calibrated focal length of the camera have to be known. The exterior orientation can be determined if at least three control points are available in the overlapping image area. The 3D coordinate of all measured points in the overlapping image area can be determined by collinearity equation that defines the relationship between object and image coordinates [2]. With the classical approach, the overlapping images were recorded according to the stereographic principle with stereo metric or single metric cameras [3]. The 3D measurement of the object points in the stereo model was performed by a human operator after orientation of a stereo model using analog or analytic stereo plotter. These traditional methods are replaced with new methods in digital photogrammetry within significant advances in electronics, computers, and information technologies, such as digital cameras, parallel processing, and storage capacity [4].

Recent advances in terrestrial laser scanning technology allow us quick and efficient collection of 3D object coordinates automatically. Different techniques and principles are used to determine the 3D coordinates of objects with the terrestrial laser scanning systems. The range laser scanner measures and computes the distance between the laser and object surface by either the time-of-flight method (TOF) or comparison of the transmitted and received wave form of modulated signal. The range distance in the TOF principle is computed based on the measurement of the time of light Δt of the laser pulse between transmission and reception. In the phase comparison method, the transmitted laser beam is modulated by a harmonic wave, and the distance is calculated based on the phase difference between the transmitted and received wave [5], [6].

II. ADVANTAGES OF CONNECTING PHOTOGRAMMETRY WITH 3D SCAN

In the last decade, the evolution of electronics has moved forward rapidly. Computing devices are powerful and open up to new possibilities in photogrammetry and laser scanning approach. Increasing CPU and GPU performance has allowed the development of photogrammetric software that can transform photos of any object into a 3D model. These software products can also cooperate with laser scanners, which are very often used in photogrammetry work, such as buildings and ground infrastructure. The main advantage of photogrammetry is its high accuracy and time effectiveness. It is possible to obtain high-quality input data within a short time frame. Linking the 3D laser scan model to photogrammetry model delivers a high quality and accurate model. It is a connection of very high measurement accuracy with photography. The output thus obtained, which connects the scan outputs and photos, gives the opportunity to look at the flawless model, obtained in real time. Another advantage of photogrammetry is that the actual laser scan output, thus the finished 3D model, also has so-called dead spots. These are places where the scanner beams cannot reach at a certain angle, and thus the model loses not to quality but to aesthetics, which, however, can be supplemented by photogrammetry. It is necessary to detect dead spots and then to overlap unscanned areas from the laser scan using the taken photos. Then resulting 3D model meets all the demanding criteria for output quality requirements such as dimensional accuracy, color and model will thus become the digital twins of the real object.

III. 3D LASER SCANNING PRINCIPLE

3D laser scanning can be used in simple, or even in more complex projects in order to monitor changes created by human activities or by nature itself. Multiple scans can be taken over time and then compared, for example before and after natural disasters, urban changes, plants reconfigurations. Scanning is used to obtain future technical solutions from interior and exterior spatial arrangements. It is a creation of digital twins of real objects and is widely used in industry, architecture, health, aviation and other sectors. In addition to gaining a number of points, based on the principle of a rotating laser beam reflected from the mirror, photographs (camera is embedded inside laser scanner) are taken from the position of the laser scanner. These photos are used to integrate RGB color spectrum and also the device assigns this spectrum to a single point. In the resulting portrayal of the scanned object, this feature may be used for better orientation in the software environment of a user.

Generally, 3D scanning is a 3D data collection of selected surface points of a specific object. Mainly, the points of this object are placed in coordinates X, Y, Z. Actual laser scanning works on the principle of transmitting the laser beams that are reflected by a rotating mirror. In scanning, the laser scanner for larger distances is rotating at 360 degrees around the vertical axis and at the same time, it is transmitting laser pulses by using a rotating mirror. Thanks to this, the scanner covers almost the entire surroundings of the scanned object. There is the only one blind angle that arises during scanning, resulting from technical specifications and it is being generated directly by the scanner. Laser beams spread out and fall on the objects covered in the field of sight of the scanner. After that, they are being reflected from objects and returned back to the scanner, where the object will be again scanned by a sensor. Afterwards, a sensor will delineate point cloud, which has the shape of the scanned object surface, based on the time and the angle of beam return. The output of the scanning process is a partial point-cloud of the scanned object or its spatial part. After scanning, it is necessary to process and edit all scanned objects according to the project requirements, which means, a post processing of a scanned object must be accomplished.

IV. DIGITAL SLR CAMERA AS A PHOTOGRAMMETRY TOOL

Modern cameras, from phones to high-end DSLRs, are designed to make their own decisions by taking photos in AUTO mode. The drawback is that images taken in AUTO mode tend to look similar to one another, with a uniform depth of field and exposure. If there is need to move beyond the automatic camera settings, photographer needs to understand camera settings to obtain desired photographs. Then it is necessary to realize under what conditions it will be photographed. By conditions we mean exterior or interior. If the photo shoot will take place outdoors, the lighting conditions should be considered: sunny day without clouds, cloudy, dusk. In the interior to be considered: neon cold or warm light with or without illumination. There are many possibilities and variations and therefore, each situation requires different settings. You can ground in with manual settings for digital SLR in [7], [8]. However, the most important factor is ISO, iris and shutter [7]. Each digital SLR has automatic settings that will be appreciated especially by beginners. The next level in shooting is the semi-automatic mode, in which only one parameter needs to be set manually, the other settings are set automatically by the camera. However, the best outputs are achieved with manual adjustment. Below are listed basic needed settings for the digital SLR before taking a photo.

- Exposure time
  When shooting, it is necessary to constantly pay attention to the exposure time. The reason is simple. If the exposure time is too long, the motion non-sharpness will appear as a blur in the photos. Even when taking a stationary subject, it is important to note that even with image stabilization, the photographer cannot keep the camera still. Therefore, when the aperture decreases, it is necessary to choose a sufficiently short time for shooting, which ensures a sharp photo and possible movement of the camera will not affect the shooting results. Setting the ideal exposure time is very individual, we set the time of shooting to 1/160 s. At this exposure time we had no problem with blur. The number of photographs was in hundreds when photographing the sculpture shown in the fig. 1. The result is a photogrammetric scan of the photographed statue. White dots are the positions of the camera at the time of shooting a specific photograph within the set.
- The effect of focal length on the time factor
  We used a zoom lens in which we usually shot a focal length of approximately 80 mm. At time 1/160 it was not difficult to get sharp photos. From the point of view of photography, it is always necessary to realize with what large focus we shoot and what large lens we keep in our hands. Thus, to avoid blur, it is necessary to shoot a scene with a time equal to the inverse of the focal



Figure 1 – Photographed statue

length, therefore, when we were shooting at about 80 mm focus, the selected time should be at least 1/80 s. Because we used a camera body with a crop factor of 1.6, we multiplied the focal length of 80 mm by a crop factor of 1.6. The resulting value of 128 mm represents the minimum time shorter by 1/128 s.

- ISO sensitivity setting
  There was a situation where shooting with a short time was no longer realizable at our location, the lighting conditions were poor (low light) and the time to take a handheld was too long. In this case, do not be afraid to raise the ISO sensitivity. The maximum safe ISO value is ISO 3200. There is no problem with values up to ISO 6400 when setting a full frame shooting. In our case we used the ISO 3200 setting.

- Exposure best practices
  Exposure is an elementary part of each photoshoot session. It consists of a well-chosen combination of elements such as iris, time and sensitivity which form the so-called exposure triangle. The aim of these settings is to produce an image that is neither too light (overexposed) nor too dark (underexposed). A properly exposed photo should be as real as we normally perceive it.

- White balance
  During the photoshoot, the light changes constantly sometimes its color is golden but sometimes it is cool white or even blue. Human eyes can adapt to these colors automatically. However, the camera is a machine and relies only on white balance settings. Every photographer has to "set up" the camera which color tones to choose. Each camera has its own white balance functions. For example, on a sunny day, cloudy, taking pictures indoors or photo shooting in dark places. With these functions, you can set the compensation for too cold or too warm light. The best way to set the white balance correctly is to take a few shots with different settings to see which one exhibits the required properties. This ensures that the color properties we want to keep are not lost. In our case, we have set up the automatic function for cloudy weather.

- Iris
  Iris is a small mechanical device situated in the camera lens. It is an adjustable aperture used to control the amount of light coming through the lens onto the camera chip. When taking pictures of the statue, we had worse lighting conditions and therefore we tried to set the lowest aperture value. Exposure times were longer, so it was necessary to use the lowest possible iris setting of 2.8 f. It is important to mention that low iris number can cause very low sharpness level of the taken image.



Figure 2 – 3D laser scan with noise

The photo shoot itself took place in the park. It was a semi-cloudy day, so we tried to take pictures when the sun was behind the clouds and the sunrays intensity was not so high. The total number of photographs was over 400. We have used photo overlap approach. The essence of the overlap is that each subsequent photo must have common points with the previous photo. Otherwise, there would be a problem within the software postprocessing, and the photos would not create complete appearance of the photographed object. All photos were checked immediately during the photo shoot to make sure that they are all sharp enough with good light parameters.

V. DATA PROCESSING

The next step after the successful data gathering (photo shooting) is a separate scanning of the object. After the successful 3D laser scanner setup followed laser scanning procedure which involved moving the scanner to predefined locations around the statue. The scanning process according to weather conditions took approximately 2 hours. After a quick evaluation of the



Figure 3 – Noise-free 3D laser scan



Figure 4 – Statue detail from photogrammetry (left) and laser scanning (right)

laser receiver, we have assumed that the point cloud was of the required quality (enough points and little noise) and could be used for further processing. After combining the scans within the software tools, we have obtained a complete noise-free 3D image of the scanned object, shown on fig. 2. But the object (statue) still had to be cleaned up of small impurities captured during the scan (birds flying, pedestrian crossings etc.). Fig. 3 shows noise-free 3D laser scan.

VI. TEXTURING

The last step is texturing of the finished scan. It is a process in which photogrammetry is linked to a 3D scan. Fig. 4 shows statue details from photogrammetry (left) and laser scanning (right). Assuming that photogrammetry and laser scanning were performed precisely to thousands of a millimeter, the result would be quality output. Fig. 5 shows a finished 3D model enriched with texture that has a realistic visualization.

VII. CONLUSION

Photogrammetry in 3D virtualization is nowadays very actual. It significantly contributes to the quality of the resulting 3D model. Its implementation in cooperation with laser scanning can significantly reduce the cost and time required to carry out an object digitization project. Usage in conjunction with a 3D scanner ensures high-quality output that can be used in multiple applications. Thanks to the development of more powerful software and hardware, the usability of given technology in other sectors is increasing and this opens up new possibilities of of application. Developments over the past decade have shown that these two technologies are likely to be interconnected with other technology solutions in the near future, allowing for even more efficient and faster processing into a digital form of object display.



Figure 5 – Final noise-free 3D laser scan linked with photogrammetry

ACKNOWLEDGMENT

This publication was realized with support of Operational Program. Research and Innovation in frame of the project: ICTproducts for intelligent systems communication, code ITMS2014+ 313011T413, co-financed by the European Regional Development Fund.



REFERENCES

[1] James S. Aber, ... Johannes B. Ries, in Small-Format Aerial Photography, 2010, https://www.sciencedirect.com/topics/earth-and-planetary-sciences/photogrammetry.

[2] J.C. McGlone, Manual of Photogrammetry, (5th Edition), American Society for Photogrammetry and Remote Sensing, Bethesda, MD (2004).

[3] K. Kraus, Photogrammetry, Volume 2, Dümmler, Bonn (1993).

[4] NaciYastikli, Journal of Cultural Heritage, Volume 8, Issue 4, September–December 2007, pp. 423-427.

[5] W. Boehler, A. Marbs 3D scanning instruments, W. Boehler (Ed.), Proceedings of the CIPA WG 6 International Workshop on Scanning for Cultural Heritage Recording (September 1–2, 2002), pp. 9-12.

[6] M. Pieraccini, G. Guidi, C. Attzeni, 3D digitizing of cultural heritage, Journal of Cultural Heritage, 2 (2001), pp. 63-70.

[7] K. Srnec, M – manuálny expozičný režim, available: https://www.ephoto.sk/fotoskola/clanky/zaciname-s-fotografovanim/manualny-expozicny-rezim/.

[8] M. Marčiš, Automatizované fotogrametrické metódy v procese digitalizácie kultúrneho dedičstva, 2019, ISBN 978-80-227-4895-7.

Some elements for assessing the radiated heat in urban areas

Ivica Kožar*, Krešimir Peša*, Marijana Cuculić* and Neira Torić Malić*
* University of Rijeka Faculty of Civil Engineering, Rijeka, Croatia
ivica.kozar@gradri.uniri.hr

***Abstract* - Heat island describes increase in air temperature between a city and its surrounding area. The air temperature increase significantly endangers the quality and health of life in a city. Radiation plays a crucial role in heating of open surfaces in a town. Solar radiation heats the surface that later cools itself radiating heat during the night. The presented model calculates the amount of daily sun-light for arbitrary location and arbitrary day of year. This is the basis for calculation of shades from buildings and estimation of the heat radiated to the street. This is an input for the second part of the model: calculation of the street surface temperature and heat exchange with the surrounding air. Heat exchange problems involving radiation lead to a non-linear formulation, which can be simplified with careful parameterisation, i.e., contribution of radiation should be expressed through radiated heat instead of direct influence of distant body temperature. That approach requires introduction of a non-linear convective heat transfer coefficient between the pavement and the open air. In order to determine model parameters, small experimental field with several differently paved areas has been established near the Faculty. Some of the recorded data is presented in the paper.**

Key words - heat island; solar radiation; radiated heat; city model; heat transfer coeficient

I. INTRODUCTION

The term "heat island" describes a phenomenon where the air temperature inside a city is significantly higher then air temperature in the area surrounding a city. The exact temperature difference depends on many parameters, city size being one the most relevant. So, Los Angeles is approximately 6˚C hotter then the surrounding area, while Rijeka is only 2˚C hotter. Heat island effect can be very dangerous [9] and should be put under control. Assuming that air conditioning devices could solve the problem is completely wrong [10].

It is believed that the main cause for increased temperature comes from the sun heating the paved areas within the city. Understanding the mechanism of the heating is crucial for taking measures to reduce the heat island effect. Here, we are trying to develop a mathematical model that would put into relation all the relevant parameters and predict the pavement temperature based on the insolation data.

There are two main groups of parameters comprising the model: i) the amount of urban area exposed to direct sunlight and ii) the process of heat exchange between the surface and the surrounding air.

The amount of urban area exposed to direct sunlight is determined from city maps where streets (paved areas) and buildings are treated separately from green (and similar) areas. It is important to assign different convective heat transfer coefficient to different areas. Here, the amount of direct sunlight plays an important role. We have devised a procedure for determination of shade area around a tall building for more precise assessment of the area exposed to direct sunlight.

Regarding the process of heat exchange between the surface and the surrounding air, we have re-parameterized the relevant equations so that convective heat transfer coefficient could be introduced as a parameter. In [8] we have established that the non-linear convective heat transfer coefficient is sufficient for description of the observed phenomena. Namely, there are many parameters that influence radiated heat exchange [1] but the assumption is that, similar to the eddy diffusivity approach in [2], all the contributions could be introduced through one parameter [3]. Presented numerical experiments confirm that hypothesis, opening the way to further investigation of parameter determination from experimental results (like in [4]).

II. ESTIMATION OF RADIATED SOLAR HEAT

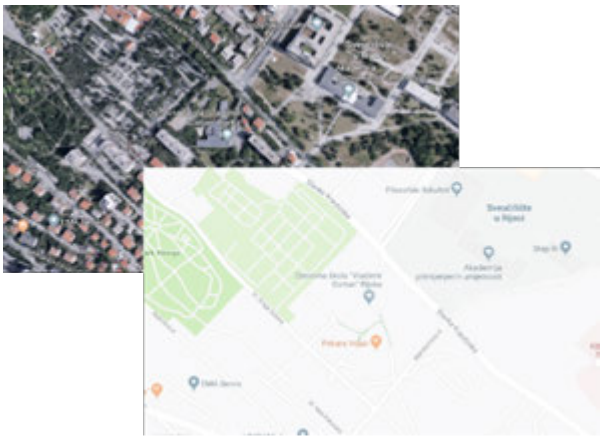


Figure 1. City area used for model parameters, satellite image for surface properties and schematic image for area calculation

A. *Terrain model*

Terrain model is starting point of our city model. From there, we get areas of various surfaces and streets comprising the city. At this point of development, we are taking into consideration only a small city area (up to 1 square kilometer). Figure 1 presents area from Google Earth map that is used in the example. From satellite photo, we assume terrain reflective properties (street, building, green area or trees) and from schematic image we calculate areas and coordinates.

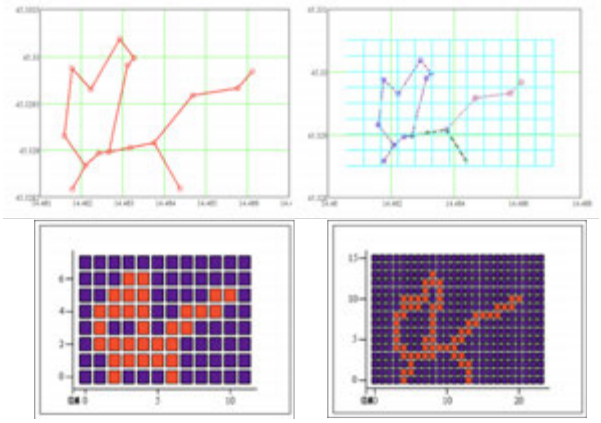


Figure 3. Digitalized streets and the coressponding city area, digitalization resolution is adjustable as needed

In the next step, streets are digitized and digitized surfaces are created, as shown in Figure 3. At this point, digitalization requires human to choose characteristic points that are latter connected into a polygon. Resolution of digitalized surfaces can be adjusted as needed (see Figure 3). For the future development, we are considering use of maps from the Croatian Geo-portal and possibly, automation of the digitalization.

B. *Experimental field*

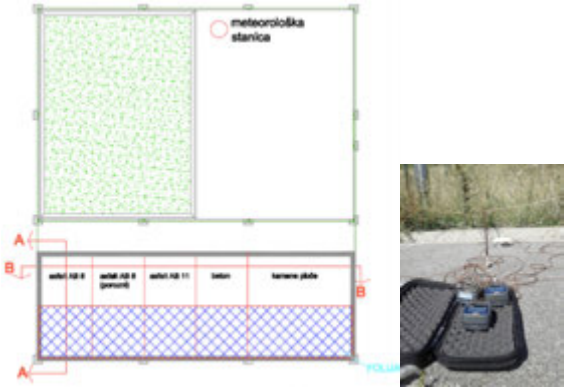


Figure 5. Test field "Kampus" and some measuring devices

In order to get insight into the process of heating and cooling of pavement in urban area measurements have been performed during the summer of 2013. Small area 3x9 m with different pavement materials has been set up at a location just outside of Faculty of Civil Engineering in Rijeka, Croatia. The test field is presented in Figure 5. Measurements comprise temperature at the pavement surface, at the pavement bottom where sub-grade material starts and at the depth of 20 cm. In addition, radiated heat flux has been measured using solar-meter and corrected for the reflexivity parameter (we assume R_{reflex}=0.5). Excerpt from the measurement data (about 3 consecutive days) is presented in Figure 2.

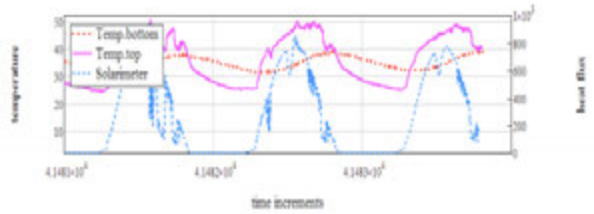


Figure 2. Experimental measurements of the pavement surface and bottom temperatures and radiated heat flux, presented as time series

C. *Sun position and shadow calculation*

The main source of information for Sun and Moon positions is [13]. Sun position is calculated for various seasons and times of day and shadow length and direction is determined for each corner of a building. Based on that calculation shadow area is estimated. Figure 4 gives schematic illustration of the calculation, where e_0 – Sun elevation angle, Φ - topo-centric azimuth angle, h – building height. During calculation we transfer coordinates between degrees and kilometers (and vice versa).

Calculation has to be performed for each building and its influence on the solar radiation on the street (surface) is

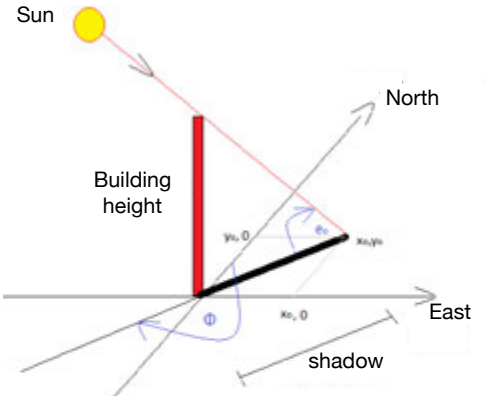


Figure 4. Schematic representation of shadow calculation

estimated. This is one of input variable for our heat exchange mathematical model.

III. MODEL OF HEAT EXCHANGE FOR OPEN SURFACES

A. *Mathematical model*

The work develops a model for heat exchange between the pavement and surrounding air and soil. The mathematical model is based on 1D non-stationary heat differential equation, see [5]

$$\rho c \frac{\partial T_{pav}}{\partial t} - \frac{\partial}{\partial z}\left(k(z)\frac{\partial T_{pav}}{\partial z}\right) = Q \quad (1)$$

Here T_{pav} is (through the depth) pavement temperature filed, t is time, $k(z)$ is diffusivity coefficient, z is pavement depth and Q is rate of heat (coming in or out), ρc is unit mass and heat capacity of the pavement. In our case, Q represents solar radiation (dividing it with the surface area, we obtain radiated heat flux q). The relationship between the pavement surface temperature and the radiated heat is obtained with the introduction of the convective heat transfer coefficient between the pavement and the open air. The equation reads

$$\alpha_{sol} q_{sol} + \epsilon_{IR}\sigma T_{sky}^4 = \\ = \bar{h}(T_{pav} - T_{air}) + \epsilon_{IR}\sigma T_{pav}^4 \quad (2)$$

The convective heat transfer coefficient h is related to the eddy diffusivity principle (see e.g., [2,3]) and the thin film coefficient concept (see e.g., [15]). The idea is that a constant cold be replaced with a functional value that incorporates the influence of some other parameter not appearing explicitly in the equation. In the case of eq.(2) it permits us to rewrite the heat flux from radiation in a form similar to the Newton heat exchange law, i.e., $q = h(T)\Delta T$, where $h(T)$ is convective heat transfer coefficient as a function of temperature and ΔT is the temperature difference (here $(T_{pav} - T_{air})$).

Here α_{solar} and ϵ_{IR} are solar absorptance and infrared emittance for the pavement, q_{solar} is solar radiation in [W/m^2] and $\bar{h}$ is convective heat transfer coefficient between the pavement and the open air. The sky temperature T_{sky} is assessed using empirical relation, as described in [6]

$$T_{sky} = \\ T_{air}\sqrt[4]{0.711 + 0.0056T_{DP} + 7.3\cdot 10^{-5}T_{DP}^2 + 0.013cos(\frac{2\pi t}{24})} \quad (3)$$

Q in eq.(1) is determined from eq.(2); it equals either the left or the right hand side of the eq.(2). We are using the left hand side for calculation of T_{pav} from eq.(1) since we are measuring q_{sol} and calculating T_{sky} from eq.(3). However, when calibrating for convective heat coefficient h we are using the right hand side.

Temperatures are in Kelvin and T_{DP} is dew point temperature, calculated as $T_{DP} = T_{air} - \frac{100 - RH}{5}$ with RH being relative humidity and t is time in hours after midnight [7]. All parameters are functions of time with their value changing during the day. Influence of the wind is taken through dependence of $\bar{h}$ on the wind speed, as in [8].

Data connecting h and various weather conditions could be found in the literature like [14] but due to local conditions it is better to estimate it locally. Due to the eddy diffusivity principle h would reflect the local conditions at the recording time.

The convective heat transfer coefficient h is also a function of position and has to be calibrated for each type of soil that we want to analyse. The key data in the inverse procedure required for assesment of h is change of the surface temperature in time and in various parts of the city it could be obtained with large-scale infrared shooting (e.g., from a plane). Data from larger area would lead to more precise temperature predictions.

The convective heat transfer coefficient h has to be calibrated for each type of soil that it to be analysed. The idea is that each city area has to be calibrated once and that value is then used in all subsequent analysis.

B. *Finite element discretization*

Mathematical model described with Eq.(1) has been discretized using the finite element method with a special finite element for heat flux towards the soil [9]. Special

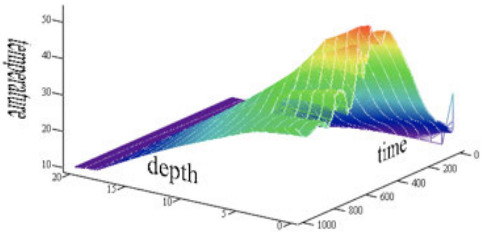


Figure 6. Temperature in space and time for non-calibrated convective heat transfer coefficient.

elements are needed in order to avoid measuring the temperature at a certain depth below the pavement. Infinite elements used could have boundary condition

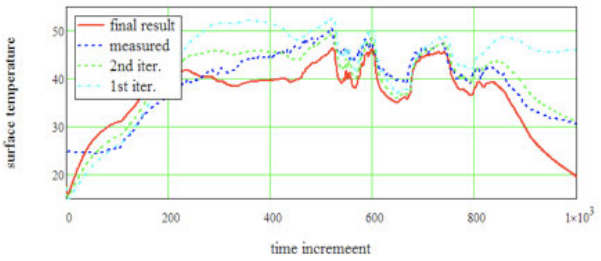


Figure 7. Temperature in space and time during calibration of the convective heat transfer coefficient.

expressed in terms of heat flux but that quantity is not measurable. Intermediate temperature measurement is used for determination of the bottom heat flux and is later used only for control purpose. Discretized model equation is as in [2] and in matrix notation reads

$$\mathbf{C}\frac{\Delta T}{\Delta t} + \mathbf{K}\mathbf{T} = \mathbf{Q} \quad (4)$$

where $\boldsymbol{C}$ and $\boldsymbol{K}$ are heat capacity and heat conductance matrices respectively and Δt is a time increment used in the simulation. Vector $\boldsymbol{Q}$ has only entry at the last element and that one involves convective heat transfer coefficient $\bar{h}$ between the pavement and the open air. Model is calibrated through measurements of q_{solar} on the site and validated through comparison with pavement temperatures. For the purpose of validation, temperatures were measured on the surface and bottom of the pavement for an extended period.

C. *Infinite 3D finite element*

In this work, 3D infinite finite element has been used to enable analysis of large areas of ground. Applying one-dimensional element as in [9] prevents inter-element communication in the surface direction (1D element extends along the depth). This difficulty is overcome by applying 3D infinite elements that extend infinitely into depth direction but are connected with adjacent nodes into a mesh of surface elements. The resulting mesh has only 2 nodes in depth and asymptotically approaches given deep earth temperature while allowing for arbitrary temperature distribution on the surface.

D. *Calibration of the model*

Experimental set-up has been mimicked in a numerical simulation so that numerical results could be compared with measurements. Twenty 1D finite elements and 1000 time increments have been used in the backward-Euler integration procedure. The results of simulations have been compared with measurements and inverse procedure has been applied (Levenberg-Marquardt method, as in [2]) to determine appropriate convective heat transfer coefficient.

E. *Numerical example*

Numerical example uses convective heat transfer coefficient determined in the calibration procedure. Experimental set-up has been mimicked in a numerical simulation so that numerical results could be compared with measurements. Twenty 3D finite elements and 1000 time increments have been used to reproduce results from experimental field measurements. The backward-Euler integration procedure used for solution of this non-linear matrix equation (4). Figure 6 presents, in space and time, the first solution with non-calibrated convective heat transfer coefficient, while Figure 7 presents comparison of measured values and those calculated in the numerical simulation. Good of fitness could be estimated using e.g., least square differences or Pearson correlation coefficient.

IV. CONCLUSION

This work confirms our hypothesis that non-linear (time and temperature dependent) convective heat transfer coefficient $\bar{h}$ between the pavement and the open air suffices for modeling the radiated heat exchange between the pavement and the surrounding air. Further development includes development of an inverse procedure for determination of $\bar{h}$, similar to [4]. In addition, factor analysis of $\bar{h}$ is planned, to obtain insight of possible factors contained in the heat transfer coefficient between the pavement and the open air.

REFERENCES

[1] J.G. Gui, P.E. Phelan, K.E. Kaloush, J.S. Golden. *Impact of Pavement Thermophysical Properties on Surface Temperatures*, Journal of Materials in Civil Engineering (19), p.683-690, 2007.

[2] D. Lozzi-Kožar, I. Kožar. *Estimation of the eddy thermal conductivity for Lake Botonega*, Engineering Review (37), p.322-334, 2017.

[3] Q.J. Slaiman, M.M. Abu-Khader, B.O. Hasan. *Prediction of heat transfer coefficient based on eddy diffusivity concept*, Chemical Engineering Research and Design (85), p.455-464, 2007.

[4] I. Kožar, N. Torić Malić, T. Rukavina. *Inverse model for pullout determination of steel fibers*, Coupled Systems Mechanics (*7*), p.197-209, 2018.

[5] J.H. Lienhard IV, J.H. Lienhard V. *A Heat Transfer Textbook.* Phlogiston Press, Cambridge Massachusetts, 2012.

[6] P. Berdahl, M. Martin. *The emissivity of clear skies*, Solar Energy (32), p.663-664, 1984.

[7] L. Adelard, F. Pignolet-Tardan, T. Mara, P. Lauret, F. Garde H. Boyer. *Sky temperature modelisation and application in building simulation*, Renewable Energy (15), p.418-430, 1998.

[8] A. Hermansson. *Mathematical model for paved surface summer and winter temperature*, Cold Regions Science and Technology (40), p.1-17, 2004.

[9] M.R. Motley, J.H. Prevost. *Simulation of transient heat conduction using one-dimensional mapped infinite elements*, Int. J. Numer. Meth. Engng. (83), p.598-610, 2010.

[10] I. Kožar, M. Cuculić, N. Torić Malić, "Establishing of parameters for model of radiated heating of urban pavement", in Proceedings of 4th International Conference on Multi-scale Computational Methods for Solids and Fluids (Eds.: A. Ibrahimbegovic et al.), p.184-187, 2019.

[11] WIRED magazine, "Urban Heat Islands Can Be Deadly, and They're Only Getting Hotter", 14 June 2017

[12] The Guardian, "The air conditioning trap: how cold air is heating the world", 29 August 2019

[13] The National Renewable Energy Laboratory (NREL), *Solar Position Algorithm*, https://www.nrel.gov/grid/solar-resource/renewable-resource-data.html

[14] Suresh Kumar, S.C. Mullick. Wind heat transfer coefficient in solar collectors in outdoor conditions, Solar Energy (84), p.956-963, 2010.

[15] Jiji L.M., *Heat Convection,* Springer – Verlag Berlin Heidelberg, 2006.

PROFESSIONAL PAPER IN CROATIAN

Unity – 3D i virtualna stvarnost

Bojan Fulanović , Vilim Šepl
Visoko učilište Algebra, Zagreb, Hrvatska
bojan.fulanovic@algebra.hr vilim.sepl@racunarstvo.hr

Sažetak **- Unity je cross-platform program za izradu i razvoj video igara. U ovom radu naglasak će biti na izradi 3D video igre u okruženju virtualne stvarnosti (VR) pomoću SteamVR plugina. Pomoću VR opreme i dobro napisanog softvera, čovjek je u mogućnosti fizički predočiti samog sebe u bilo kojem virtualnom svijetu. Program je danas kompatibilan s većim brojem vodećih VR setova, kao što su HTC Vive, Oculus Rift i PlayStation VR.**

Ključne riječi – Virtualna stvarnost (VR), Unity, 3D, SteamVR, HTC Vive

I. UVOD

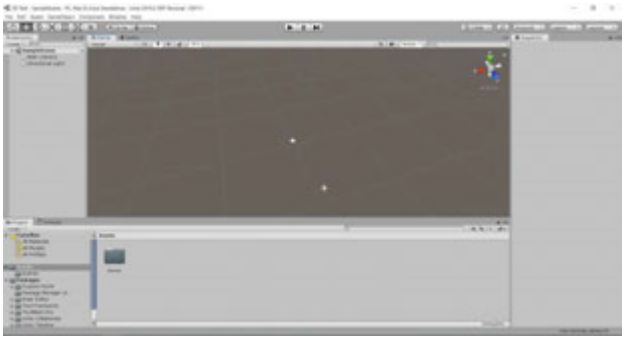
Program Unity razvija tvrtka Unity Technologies, a prvi puta je objavljen 2005. godine kao ekskluzivan softver za razvoj video igara na Mac OS-X operacijskom sustavu. Od 2018. Unity podržava više od 25 različitih platformi. Program se može koristiti za razvoj 2D, 3D, virtual reality i augmented reality (AR) video igara, kao i simulacija za podržane platforme[1].

Pokretanje Unityja znatno je pomoglo programerima započeti izrađivanje kvalitetnih aplikacija u novom, moćnom okruženju koje je svima lako dostupno. Unity je tako ubrzo postao omiljen program za razvoj različitih video igara te se logo programa pojavljuje na brojnim popularnim naslovima video igara poput „Cities: Skylines“, „Hearthstone“ i „Cuphead“. Od 2018. godine gotovo polovica video igara za mobilne uređaje te otprilike 60% VR/AR sadržaja nastalo je pomoću Unityja, a čak više od 2000 zaposlenika radi na njegovom razvoju i održavanju[2].

Kao glavni programski jezik za skriptiranje koristi se isključivo C#, no prije je osnovni razvojni jezik bio Boo, koji je uklonjen pojavom Unity 5 verzije programa[3]. Također se u prijašnjim verzijama programa koristila i inačica JavaScripta, prikladno nazvana UnityScript, koja je uklonjena u kolovozu 2017. nakon izlaska Unity 2017.1 u korist C#-a[4]. U ovom trenutku najnovija stablina verzija programa je 2019.2.

Od 2010. godine, korisnici programa mogu razvijati, prodavati, kupovati te preuzimati razne dodatke za izradu video igara preko Unity Asset Storea. Dodaci uključuju 2D i 3D oblike, animacije i razna druga okruženja i dodatke koje su razvili drugi programeri te ih dodali u trgovinu. Asset Store tako služi kao glavni izvor pomoćnih materijala za izradu bogatih igara[5].

Pojavom Unityja pojavile su se i brojne prednosti i mane samog programa. Možda i najbitnija karakteristika, ujedno i prednost programa, je njegova prenosivost te kompatibilnost sa širokim spektrom uređaja. Među ostalim prednostima je i jednostavnost razumijevanja i učenja programa pomoću službene dokumentacije, visoka kvaliteta zvučnih i vizualnih efekata, mogućnost debugginga i sl. Kad govorimo o nedostacima programa, najčešće spominjemo njegove niske sustavne performanse u odnosu na druge programe, djelomično zbog toga što koristi C# kao primarni programski jezik. Još je nedostatak i visoka cijena licence koja je potrebna ako se želi postići najbolja kvaliteta grafike, razvoja programa i performansi[6].



Slika 1: Prikaz glavnog Unity sučelja

Unity pruža podršku za nekolicinu različitih VR setova, poput Oculus Rifta, Gear VR-a ili Vivea. Svi objekti i okruženje dodano u igru mogu se direktno pregledati i testirati u VR načinu rada kojeg program omogućuje. Unity je tako postao zadani alat za razvoj VR aplikacija zbog svoje jednostavnosti i brzine izrade[7].

II. VIRTUALNA STVARNOST I SUSTAVI

Virtual Reality (VR) ili virtualna stvarnost je simulirani doživljaj koji može biti virtualno sličan ili pak posve različit stvarnom svijetu. VR tehnologija se, osim za zabavu, koristi i u edukaciji, npr. u medicini ili vojnoj obuci. Trenutni standardni sustavi za virtualnu stvarnost koriste posebne setove za glavu s ekranom, a najčešće i par ručnih upravljača koji omogućuju korisniku pregled, kretnju i interakciju s virtualnim okruženjem[8].

Razlika između VR-a i proširene stvarnosti (AR) je to što je VR okruženje, kako mu ime ukazuje, posve virtualno, tj. nestvarno, dok AR u postojeće, realno orkuženje ubacuje virtualne objekte. Oba oblika stvarnosti koriste različite senzore i algoritme kako bi odredili poziciju i orijentaciju korisnika, samo što VR koristi kameru postavljenu u simuliranom okruženju, a AR stvarnu kameru na nekom elektroničkom uređaju, npr. pametnom telefonu.



Slika 2: HTC Vive Pro set s ručnim upravljačima[9]

Najprepoznatljivija komponenta VR sustava definitivno je set za glavu s ekranom (head-mounted display). Tehnologija koja se koristi u takvim setovima glavna je poveznica između čovjeka i virtualnog okruženja te je ujedno i najveća razlika u odnosu na klasična, tradicionalna korisnička sučelja. HDM korisniku omogućuje vizualni, ali i auditivni doživljaj i pregled virtualnog okruženja. Uvjerljive virtualne stvarnosti zahtijevaju mnogo više od same realistične grafike. Vid i sluh su centar ljudskog osjetila za prostor, no ljudsko biće brže reagira na zvučne znakove. Stoga je za ostvarenje idealne virtualne stvarnosti izuzetno bitna preciznost i prirodnost zvuka u okruženju. Dok je audio-vizualne informacije vrlo lako replicirati, još uvijek se provodi aktivno istraživanje i razvoj virtualnih rješenja za ostala ljudska osjetila. Osim HMD-a postoje i ručni kontroleri koji omogućuju taktilnu interakciju s objektima u simulaciji. Ti kontroleri, iako su dostojna zamjena, nisu u potpunosti preslika korištenja fizičkih ruku u realnom okruženju. Također, postoje i podloge koje služe kao senzori za kretnju čovjeka kako bi se postiglo maksimalno realno 3D iskustvo u VR-u, umjesto da korisnik koristi kontrole na upravljačima kao izvor kretnje dok sjedi ili stoji na mjestu[10].

Unity pruža podršku te se može povezati s mnoštvom različitih VR sustava. Među najpoznatijima prepoznajemo:

A. Oculus Rift

Oculus je tvrtka koja se bavi tehnologijom, specifično proizvodnjom hardvera i softvera za razvoj VR-a. Osnovana je 2012., a iste godine objavljuje Rift, set za glavu dizajniran i namjenjen za igranje video igara. Također su pokrenuli kampanju kako bi programerima omogućili dostupnost VR setova. Kampanja se pokazala uspješnom te je tvrtka zaradila 10 puta više nego što je planirala. Tako su nastale prve dvije pred proizvodne inačice VR sustava namijenjene programerima:

- Oculus VR DK1 (Development Kit 1)
- Oculus VR DK2 (Development Kit 2)

Prvi proizvod namijenjen korisnicima izdan je u ožujku 2016. s novim dizajnom specijaliziranog VR ekrana, pozicijskim zvukom te tracking sustavom za praćenje položaja VR seta, upravljača i dijelova tijela. Od ožujka 2014. Oculus razvija tvrtka Facebook Technologies[11].

B. HTC Vive

Vive je VR set kojeg razvijaju tvrke HTC i Valve. Prvi je set koji pruža „room-scale“ iskustvo, što omogućava korisnicima da se slobodno kreću u 3D prostoru te pregledavaju okolinu u realnom mjerilu. Uređaj koristi tehnologiju „Lighhouse“ s različitim senzorima kako bi pratio pokrete glave i do jedne desetine stupnja. Vive je prvi VR HMD koji podržava SteamVR. SteamVR je platforma koja implementira OpenVR programsko sučelje kako bi se uspostavila veza između VR opreme i samog programa. Od pojave prve inačice Vive je ostao zaključan u snažnoj konkurenciji s Oculus Riftom koji je također uspješno plasiran na tržište[12].

	Oculus Rift	HTC Vive
Display	OLED	OLED
Resolution	2160x1200	2160x1200
Refresh rate	90Hz	90Hz
Platform	Oculus Home	SteamVR
Field of view	110 degrees	110 degrees
Tracking area	Standing/sitting	15x15ft
Built-in audio	Yes	Yes
Built-in mic	Yes	Yes
Controller	Xbox One controller	HTC Vive controller, any PC compatible controller
Sensors	Accelerometer, gyroscope, magnetometer, 360-degree positional tracking	Accelerometer, gyroscope, laser position sensors, front-facing camera
Connections	HDMI 1.3 minimum, 2x USB 3.0	HDMI 1.4 or DisplayPort 1.2 minimum, 1x USB 2.0, Bluetooth

Slika 3: Usporedba Oculus Rifta i HTC Vivea (2016)[13]

C. Gear VR

Gear VR je Samsungova inačica VR opreme koju tvrtka razvija u suradnji s Oculusom. Prvi set izdan je krajem 2015. godine. Za razliku od Oculus Rifta i HTC Vivea, da bi VR set pravilno funkcionirao potrebno je u njegovo kućište umetnuti Samsung Galaxy pametni telefon koji ima ulogu ekrana i procesora, dok se na samom kućištu nalaze kontrole kao što su tipke, pokretni kotačić i podloga osjetljiva na dodir te senzor neposredne blizine koji otkriva kada je VR set upaljen[14].

Glavni fokus Gear VR-a je pružiti najbolje moguće iskustvo pri VR mobilnim igrama i simulacija, ali pokazuju sve veći interes i za učinkovitu uporabu uređaja u znanosti i medicinskom obrazovanju[15].

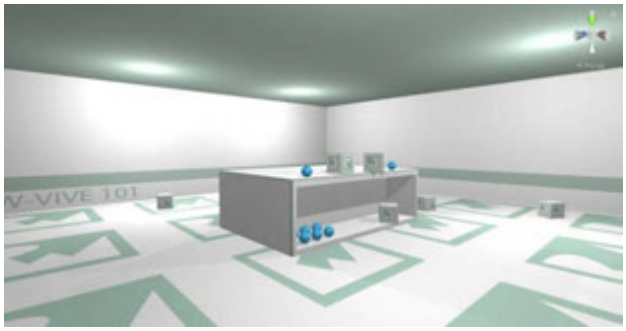
D. PlayStation VR

Za razliku od spomenutih VR sustava, PS VR namijenjen je isključivo konzolama, točnije Sonyjevim PlayStation 4 konzolama. Stoga PS VR od samog početka ima veliku prednost pred ostalim sustavima jer deseci milijuna obožavatelja igara u cijelom svijetu već posjeduju PS4 konzolu te ne moraju kupovati ili nadograđivati računala kako bi osposobili VR sustav za uporabu. Ta prednost omogućila je PS VR-u da postane najprodavaniji VR sustav na tržištu dostižući brojku od 4 milijuna prodanih jedinica od prvog izdanja sustava[10].

III. POVEZIVANJE VR-A S IGROM

Za uspješno povezivanje odabrani VR set najprije moramo konfigurirati da bi ga Unity mogao prepoznati. U ovom slučaju kao primjer koristit ćemo HTC Vive VR set s HMD-om i ručnim kontrolerima za interakciju. Nakon što uređaj povežemo s računalom te konfiguriramo postavke u programu HTC Vivea, spremni smo ga povezati s Unityjem.

U Unityju kreiramo novi testni 3D projekt. Nakon što se učita prazna scena projekta, najprije moramo dodati objekte koji će činiti naše okruženje u sceni. Objekte možemo uvesti preko Unity Asset Storea, iz nekog vanjskog izvora ili ih možemo napraviti po našoj želji. U ovom slučaju koristit ćemo gotove materijale uvezene iz nekog vanjskog izvora. Materijali sadrže komponente sobe, kao što su pod, zidovi, osvjetljenje, stol te interaktivne komponente poput lopti i kocaka. Pokrenemo li igru sada možemo primijetiti da se ništa ne događa, osim što vidimo sobu i njene objekte.



Slika 4: Prikaz scene s dodanim objektima

Kako bismo mogli upravljati interaktivnim objektima i kretati se u prostoru, nužno je dodati SteamVR SDK[16] da bi Vive znao komunicirati s programom. SteamVR služi kao library koji olakšava razvoj koda za sve glavne VR setove, među kojima je i HTC Vive, a može ga se preuzeti i uvesti u projekt putem Unity Asset Storea. Pri uvozu paketa SteamVR-a plugin otvara prozor u kojemu možemo odabrati postavke programa koje će unaprijediti i prilagoditi performanse i kompatibilnost s odabranim uređajem. Kada proces uvoza plugina završi, možemo zatvoriti Asset Store i pogledati što se sve nalazi u SteamVR folderu koji se stvorio u Project prozoru Unityja. Vidimo nekoliko ključnih komponenti koje su neophodne za rad VR seta u virtualnom okruženju aplikacije.

- [SteamVR] upravlja renderiranjem svih postavljenih VR kamera.
- [CameraRig] upravlja Vive HMD-om i ručnim upravljačima.

Obje komponente dodajem u Hierarchy te brišemo postojeću komponentu Main Camera da izbjegnemo miješanje s ugrađenom kamerom CameraRig komponente. Ako u ovom trenutku pokrenemo aplikaciju, pomicanjem ručnih upravljača možemo primijetiti da se pomiču i virtualni upravljači. Također možemo primijetiti i da HMD prati naše pokrete glave u stvarnom vremenu.

Zaključujemo da se trenutno možemo slobodno kretati po sobi i upravljati rukama, no još uvijek ne možemo utjecati na objekte. Kako bismo omogućili upravljanje objektima, moramo isprogramirati određene funkcionalnosti i podesiti kontrole te ih međusobno povezati da bi SteamVR znao pokrenuti određene akcije.

IV. KONFIGURACIJA AKCIJA

SteamVR omogućava prilagođenu konfiguraciju akcija, odnosno pridruživanje ulaza na upravljačima odgovarajućim akcijama. Pristupamo im pomoću SteamVR Input prozora koji se nalazi u Window meniju. U prozoru se nalaze sve zadane akcije podijeljene u setove akcija. Postoje 3 predefinirana seta akcija: default, platformer i buggy. Od njih je u ovom primjeru najbitniji default set u kojem se nalazi GrabGrip akcija. Ona će omogućiti interakciju s objektima u igri. Kako bi je prilagodili da obavlja funkciju po našoj želji, akciju GrabGrip preimenujemo u Grab.

Prije korištenja akciju trebamo povezati s kontrolama na upravljaču. Unity sadrži ugrađenu lokalnu web stranicu pomoću koje povezujemo željene akcije s odgovarajućim tipkama na upravljaču. Među ponuđenim tipkama trebala bi se nalaziti i Grip tipka koju povezujemo s Grab akcijom. Svaki put kada se na upravljaču pritisne Grip tipka, okida se Grab akcija. Zbog svojstva zrcaljenja dovljno je konfigurirati samo jedan upravljač koji će svoja podešena svojstva automatski preslikati na drugi upravljač.



Slika 5: Prikaz SteamVR Input prozora

Nakon konfiguracije akcija ponovo pokrećemo igru. Pritiskom i držanjem tipke Grip kada smo u blizini nekog pomičnog objekta, npr. lopte, primijećujemo kako smo ju uspješno uhvatili. Pustimo li tipku lopta ispada iz ruke te ponovo možemo odabrati i uhvatiti drugi objekt. Ovime završava proces osposobljavanja sustava za igranje igre. Sada se uspješno možemo kretati prostorijom i koristiti ručne upravljače za interakciju s objektima. Ostatak komponenata igre dodajemo i uređujemo po želji.

ZAKLJUČAK

Osim što Unity ima visoku reputaciju kao alat za izradu klasičnih 2D i 3D video igara za sve vrste platformi, razvio se i postao jedan od glavnih programa za razvoj aplikacija u VR okruženju. Razlog tome je njegova podrška za različite VR sustave koji se lako prilagode kreiranom 3D okruženju te se kontrole jednostavno konfiguriraju kroz konfiguracijske postavke. Jednostavnost povezivanja različitih VR sustava s Unityjem postignuta je dodavanjem implementacije API-a, u ovom slučaju SteamVR-a, koja služi kao sučelje između sklopovlja VR opreme i Unityja. Također, to sučelje dodaje sloj apstrakcije između VR opreme i programa što omogućava da željenu prilagođenu akciju povežemo s kojom god kontrolom upravljača želimo.

Unity na taj način u isto vrijeme postiže fleksibilnost zbog toga što podržava širok spektar platformi i VR sustava te jednostavnost korištenja u razvoju 3D aplikacija u VR okruženju. Zbog toga danas visoko kotira među programerima video igara.

LITERATURA

[1] Unity (game engine), Wikipedia https://en.wikipedia.org/wiki/Unity_(game_engine)

[2] Bonfiglio, Nahila (1. listopad, 2018). "DeepMind partners with gaming company for AI research"

[3] Unity Technologies, "What's new in Unity 5.0"

[4] Unity Technologies Blog, "UnityScript's long ride off into the sunset"

[5] Savov, Vlad (30. lipanj, 2017). https://www.theverge.com/2017/6/30/15899446/unity-cinemachine-unite-europe-2017-animation

[6] Potenza Global Solutions, http://www.potenzaglobalsolutions.com/blogs/5-rarely-known-advantages-and-disadvantages-of-unity-game-development

[7] Jerad Bitner, Lullabot (20. srpanj, 2017), https://www.lullabot.com/articles/11-tools-for-vr-developers

[8] Virtual Reality, Wikipedia https://en.wikipedia.org/wiki/Virtual_reality

[9] Vive https://www.vive.com/us/

[10] Joe Bardi, Marxent (26. Ožujak, 2019) https://www.marxentlabs.com/what-is-virtual-reality/

[11] Oculus VR, Wikipedia https://en.wikipedia.org/wiki/Oculus_VR#cite_note-Kotaku_FB-2

[12] Dante D'Orazio and Vlad Savov, The Verge (1. ožujak, 2015) https://www.theverge.com/2015/3/1/8127445/htc-vive-valve-vr-headset

[13] Ivy Bruce, Recovery Android (10. Kolovoz, 2016) "Which is Better: Oculus Rift vs HTC Vive"

[14] Samsung, "Samsung Gear VR - Specs"

[15] Christina Moro and Zane Štromberga, "Virtualisation devices for student learning"

[16] Kristopher Sandoval (2. Lipanj, 2016) "What is the Difference Between an API and an SDK?"

SP

International Conference on

MIPRO JUNIOR - STUDENT PAPERS

Steering Committee

Chairs: Nikola Filip Fijan, University of Zagreb, Croatia
Dubravko Sabolić, University of Zagreb, Croatia
Benjamin Kušen, University of Zagreb, Croatia

PAPERS

Development of Portable System for Determination of Aircraft Motion

S. Marijan, D. Franjković
Faculty of Transport and Traffic Sciences / Department of Aeronautics, Zagreb, Croatia
stjepanmarijan1@gmail.com , dfranjkovic@fpz.unizg.hr

***Summary* – Portable system for determination of aircraft motion is developed using Arduino 9 Axes Motion Shield sensor, Arduino Uno REV3 microcontroller board and notebook with suitable software. Some preliminary measurements of aircraft motion in turns were conducted on the Cessna 172R airplane and results were compared with theoretical predictions. In this paper are given: description of the Portable System and its components, working principles, characteristics of the Cessna 172R airplane, short review of theoretical background and presentation of some measurements results. Possibilities of further developmnet and application, as well as some constraints, are considered and commented. Conclusions are drawn at the end of the paper.**

Key words – aircraft motion; sensor; accelerometer; gyroscope; load factor

I. INTRODUCTION

The aim of this paper was to measure the amounts of load factors in horizontal turns and to compare them with theoretical assumptions. Measuring the amount of load factor in aeronautics is very important, because the load factor represents a global measure of the stress that acts on the structure of the aircraft. Excessive load factors must be avoided, because there is a possibility of exceeding the structural strength of the aircraft, and also of stalling the aircraft. While building an aircraft, we need to know for which amount of load factor we are designing it. Also, human beings have a limited ability to withstand a load factor significantly greater than 1, both positive and negative. Measurements were conducted on a Cessna 172R aircraft and were carried out only in horizontal turns. Measured data was subsequently analyzed and graphically represented.

The paper describes the horizontal turn and what the coordinated turn looks like. Basic forces that act on an aircraft in a turn are described and explained. Also, it is explained what is the aircraft load factor.

System for measuring the aircraft motion in a turn is also described, parts it consists of and its hardware and software. In addition, the calculations and the way the sensor is placed inside the aircraft is described.

Some of the results obtained from the measurement are presented and the way the measurement was conducted. Each turn was defined by a time duration so that it can be analyzed more easily.

The results of the measurement were graphically analyzed in such a way that the actual load factor values recorded by the measuring device were compared with the theoretically assumed load factors.

At the very end, a conclusion is drawn based on all the insights gained through the testing and the reasons for deviation from the theoretically assumed values within the paper are stated.

II. AIRCRAFT MOVEMENT IN HORIZONTAL TURN

Horizontal turn is a manoeuvre by which the pilot, by the coordinated use of all control surfaces, forces the aircraft's center of gravity to move along a circular trajectory in order to change the direction of flight. A coordinated turn is a turn in which the pilot maintains the amount of centripetal force equal to the amount of centrifugal force, that is, a turn in which the ball inside the inclinometer is maintained centered between two reference lines by the direction command. The pilot brings the aircraft into a coordinated flight using a simple rule „step on the ball“, that is, the pilot pushes the rudder with the leg on which side of inclinometer the ball is situated at the moment until he gets the ball to remain in the middle of two referent lines. If ailerons and rudder remain coordinated during the turn, the ball will remain centered between the two referent lines and there will be no slipping or skidding during the turn.[1] Considering the forces (Figure 1) acting on the aircraft in a horizontal turn, it is noticable that there are lift, drag, thrust, weight, centrifugal force and centripetal force.[2]

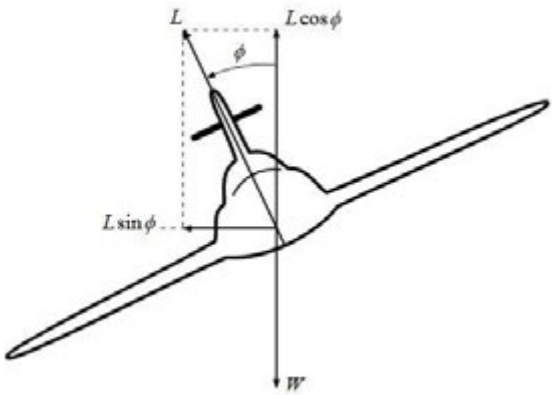


Figure 1 Forces in a horizontal turn

In a horizontal turn with constant speed, thrust is equal to drag:

$$T = D \quad (1)$$

T – thrust
D – drag

The weight of the aircraft is equal to vertical component of lift:

$$mg = L\cos\phi \quad (2)$$

m – mass of the aircraft
g – gravitational acceleration
L – lift
ϕ – aircraft bank angle in degrees

Horizontal component of lift, that is, centripetal force is equal to centrifugal force:

$$L\sin\phi = (mv_z^2/r) \quad (3)$$

v_z – velocity of the aircraft
r – turn radius

From figure 2 it is also possible to calculate the load factor of the aircraft as the ratio of total lift of an aircraft to the total aircraft weight by this formula:

$$n = L/(mg) \quad (4)$$

n – load factor

By including formula (2) into formula (4) an expression is obtained from which we can conclude that the load factor of an aircraft depends only on the bank angle of the aircraft (Figure 2):

$$n = 1/\cos\phi \quad (5)$$

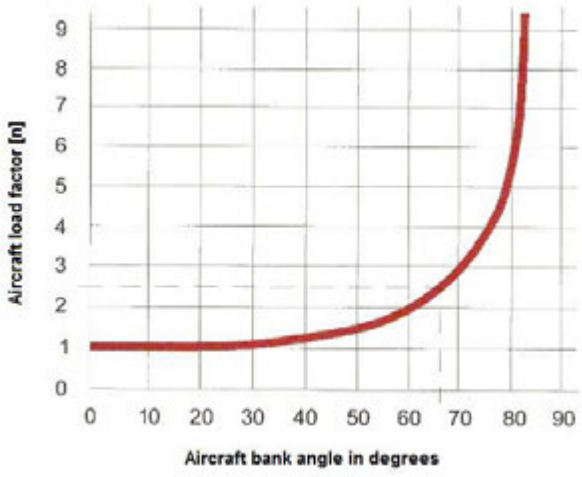


Figure 2 Relation between aircraft load factor [n] and aircraft bank angle

III. SYSTEM FOR DETERMINATION OF AIRCRAFT MOTION

System for determination of aircraft motion consists of: notebook, Microsoft Excel, PLX-DAQ software, Arduino programming console, Arduino Uno microcontroller board and absolute orientation sensor which consists of accelerometers, gyroscopes and magnetometer.

Arduino Uno (Figure 3) is a microcontroller board. It has 14 digital input/output pins, 6 analog inputs, a USB port, a 5 volts power jack, an ICSP header and a reset button.[3]



Figure 3 Arduino Uno Rev3 microcontroller board

The Arduino 9 Axes Motion Shield (Figure 4) is based on the BNO055 absolute orientation sensor from Bosch Sensortec GmbH which integrates a triaxial 14-bit accelerometer, a triaxial 16-bit gyroscope with a range of ±2000 degrees per second and a triaxial geomagnetic sensor with a 32-bit microcontroller running the BSX3.0 FusionLib software. The sensor gives data of three-dimensional acceleration, yaw rate and magnetic field strength data by using 3 perpendicular axes.[4]

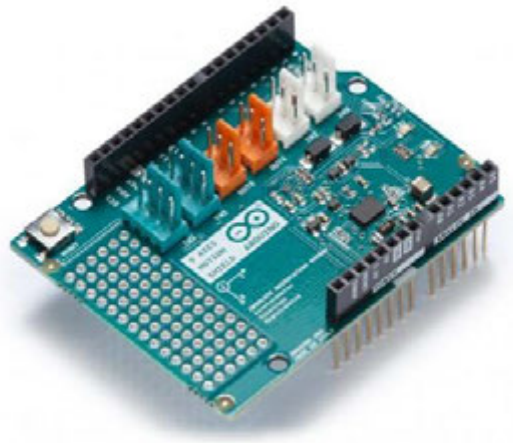


Figure 4 Arduino 9 axes motion shield

In the Arduino programming console there is a code for managing the microcontroller and the sensor for collecting data. The code also writes out the data collected to Arduino serial monitor in a readable form. (Figure 5)

Parallax Data Acquisition tool (PLX-DAQ) is a software add-in for Microsoft Excel which can parallelly process up to 26 channels of data from any microcontroller and arrange the numbers into columns as they arrive. [5] (Figure 6)

PLX-DAQ provides easy spreadsheet analysis of data collected in the field, laboratory analysis of sensors and real-time equipment monitoring.[5]

The only objection to the PLX-DAQ software is that older versions of Microsoft Excel such as version 2000-2003 must be used, as the software is not supported on newer versions of Microsoft Excel and on newer operating systems. In Microsoft Excel, we get a more transparent print of sensor data, which is then easier to analyze in various ways. All of these programs, microcontrollers and sensors are controlled by a notebook in real time.

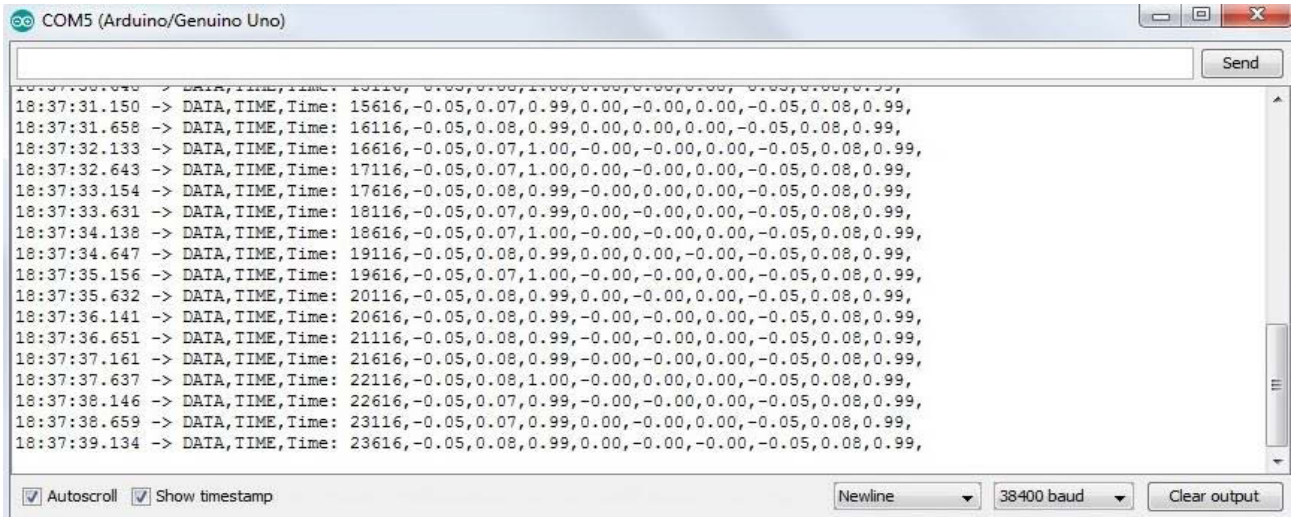


Figure 5 Real-time print of sensor data

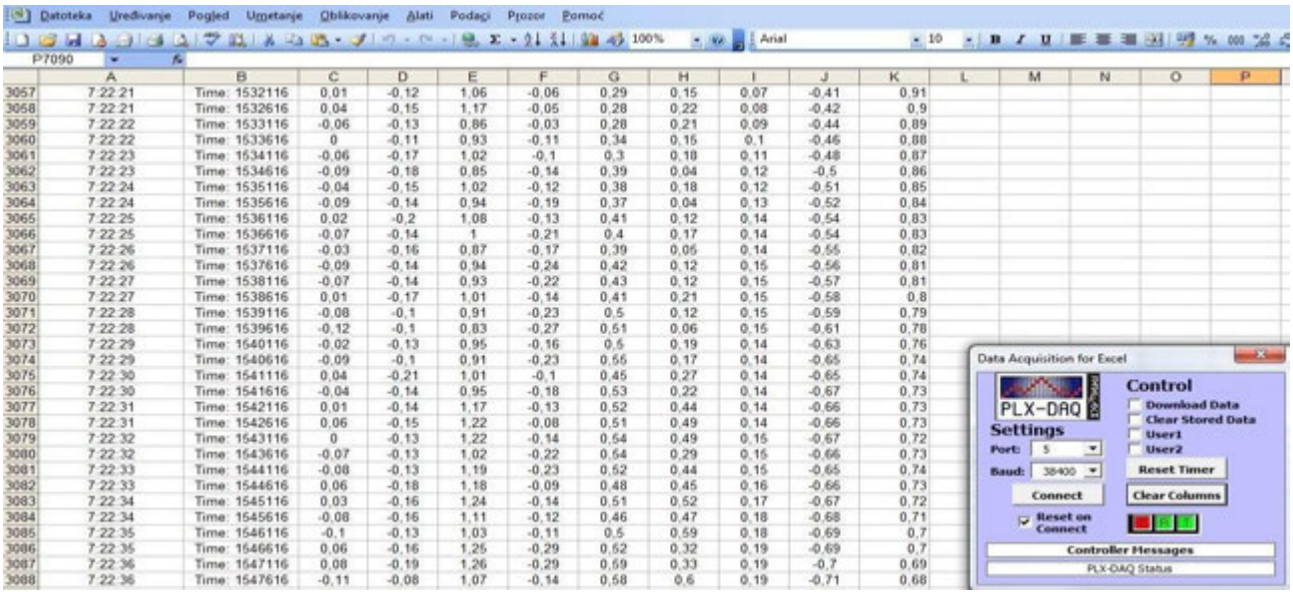


Figure 6 PLX-DAQ and Microsoft Excel 2003

IV. PLACEMENT OF SENSOR IN THE AIRCRAFT

The Cessna 172R aircraft (Figure 7) of license plate 9A-DAD was used for testing. It is an American-made aircraft made by the Cessna Aircraft Company. The aircraft is a single-engine, four-seater with high-wing. It can carry up to three passengers with one crew member. It is powered by a 160 horsepower Lycoming engine. The engine is four-cylinder atmospheric boxer piston engine that is air-cooled and fuel injected.[6]

The aircraft has a constant pitch propeller. The Cessna is one of the most widespread and most numerous aircraft in the world, and it is still produced today, even though the production began in 1955. The use of Cessna 172 aircraft is widespread. The aircrat is used for a variety of purposes. It is most widely used for training of new pilots where it has proven its outstanding handling and stability characteristics. Among the pilots it is known as a fault-tolerant aircraft. In addition of being used for training purposes, the aircraft is also used in agricultural aviation, parachuting, air taxi, and many private owners use it for various of private purposes.



Figure 7 Cessna 172R

Accelerometers are usually placed in the center of mass or as close as possible to the center of mass of the body or object whose acceleration we want to measure.

The sensor is located at the center of mass inside the aircraft and it is oriented in such a way that the *x*, *y*, and *z* axes of the sensors coincide with the *x*, *y*, and *z* axes of the aircraft.

Since the aircraft is in a uniform gravity field, the center of mass and center of gravity will be in the same location. Thus, the position for positioning the sensors along the aircraft's *x* axis was calculated by dividing the total momentum at take-off by the total mass of the aircraft at take-off, and thus obtaining the sensor distance from the aircraft reference line along the *x* axis.

$$x = M/m = 91.62 \cdot 1000/2262 =$$
$$40.5 \text{ in} = 102.87 \text{ cm} \quad (6)$$

The mass and balance calculation document from the aircraft's operating manual was used to calculate the center of gravity.

The placement of the sensor along the *y* axis was determined by the symmetry of the aircraft so that half the distance between the widths of the main wheels was taken. The distance between the main wheels is 2.54 meters. Half of the Cessna 172R cabin width can also be used to determine the position of the sensor along the *y* axis. The cabin width is 1.2 meters. Another way of determining the position of the sensor along the *y* axis is to use half the wingspan of the Cessna 172R aircraft. The wingspan of the Cessna 172R aircraft is 11 meters.[6] (Figure 8)

Determining the *x* and *y* axes yields the intersection of these two axes at one point, through which the *z* axis also passes. Regarding the placement of the sensor along the *z* axis, it was considered that the sensor should be positioned as close as possible to the floor of the aircraft regarding the center of gravity. And also for the sheer simplicity of attaching the sensor to the surface.

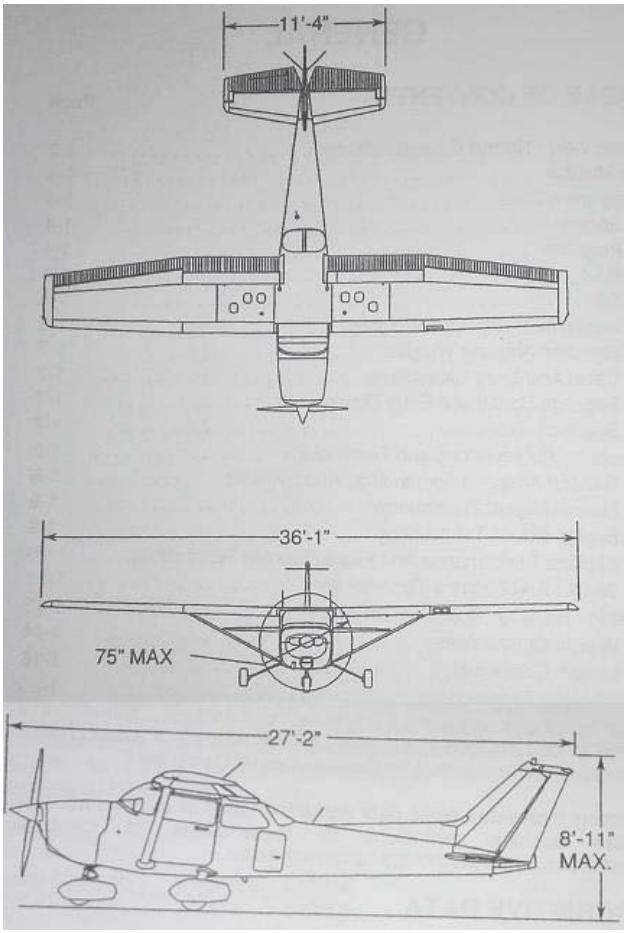


Figure 8 Basic aircraft dimensions of Cessna 172R

V. PRELIMINARY MEASUREMENTS

The measurements were conducted in such a way that the sensor was placed and attached to a predetermined position within the aircraft.

The test was conducted on an aircraft in flight at an altitude of 2,500 ft above sea level where the pilot was required to perform the desired manoeuvres. The requirements were based on certain speeds, bank angles and coordination, or uncoordination of a turn. During the test, the examiner would first wait for the pilot to stabilize and bring the aircraft into level flight at a specified speed, and then require the pilot to maintain the speed and bank angle of the aircraft constant throughout the entire turn. All the turns performed during the test were full 360 degrees turns and were peformed in the horizontal plane.

During the test, the examiner had in front of him a notebook with which he could monitor the system for measuring the load factors on the aircraft in real time. (Figure 9) In addition, he also had a sheet of paper with a table on it in which he noted whether there was a change of altitude in the turn, which turn was performed in relation to the left or right orientation and whether the turn was coordinated or not. The time at which the turn was performed was also marked, so that later, when processing the data, it would be easier to find the period in which each turn took place.



Figure 9 Real time system monitoring

VI. DATA ANALYSIS

For easier analysis of results, the load factors are plotted against time, and compared with the theoretical calculations also shown on the graphs. Theoretical calculations are in Table 1 and were obtained by the formula (5):

Table 1 Theoretically assumed load factors for certain aircraft bank angles

ϕ [˚]	10	13,5	15	20	30	45	60
n [g]	1,02	1,03	1,04	1,06	1,20	1,41	2,00

In the case of a coordinated turn (Figure 10), the load factor was a constant and coincided with the theoretical assumption. While the uncoordinated turn (Figure 11) shows an increase and decrease in the amount of load factor in relation to the theoretical assumption. These noticeable changes were caused by the increased amount of lift required to keep the aircraft in horizontal flight.

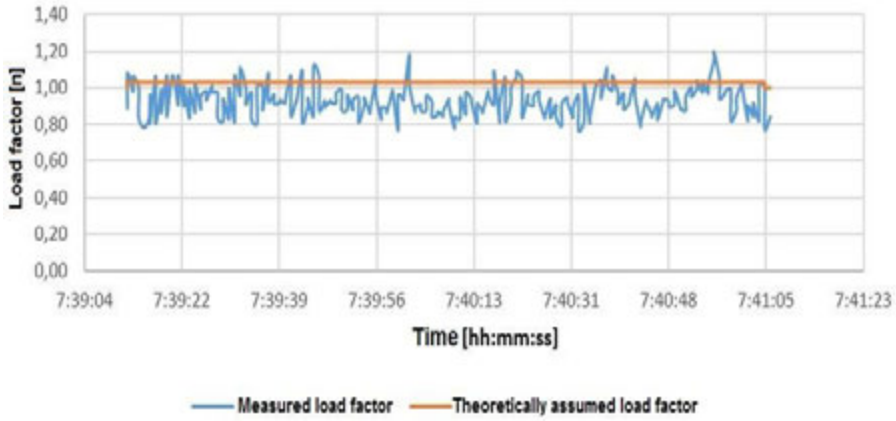


Figure 10 Left coordinated turn of standard bank angle (13.5˚) at speed of 90 knots

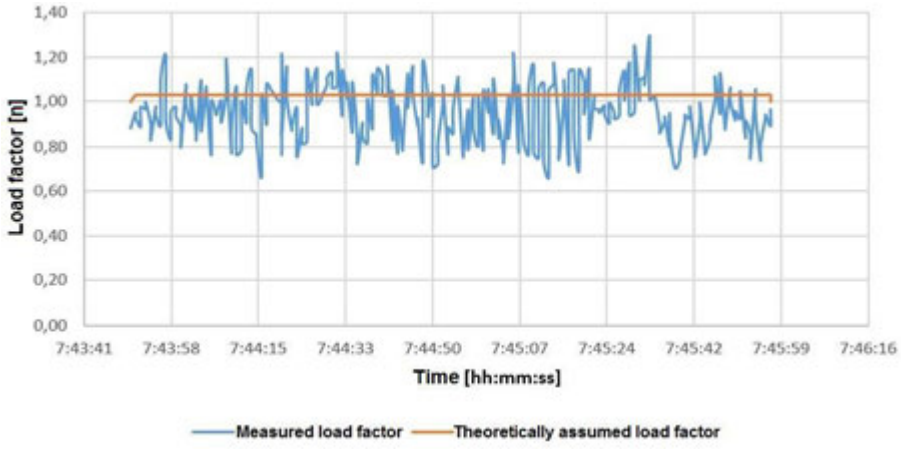


Figure 11 Left uncoordinated turn of standard bank angle (13.5˚) at speed of 90 knots

VII. FUTURE APPLICATION

The system is easy to operate. It is necessary to calibrate the device before conducting the measurements. Calibration is performed automatically. The sensor itself has a number of options such as a magnetic sensor that provides the ability to detect the magnetic field of the Earth, which allows the magnetic heading of the aircraft to be read. Also, the system can be fitted to only certain parts of the aircraft to measure the load factors for each component of the aircraft, which can be applied when testing new aircraft for load factors. Since the system measures acceleration on all three axes, it is possible to develop it into an inertial navigation system for general aviation, where the system, in conjunction with the computer and the appropriate software, would be able to calculate the current speed and position of the aircraft depending on the departure point of the aircraft. The sensor also allows tracking the amount of roll, yaw and pitch of the aircraft, which allows us to know, in addition to the load, at each moment, the current attitude of the aircraft relative to a reference line.

VIII. CONCLUSION

After conducting mesuarments on a Cessna 172R aircraft, which were performed only in horizontal flight the load factors measured by the sensors were compared with theoretical assumptions. The following conclusions of this paper are derived by analyzing and graphically presenting the obtained results and comparing them with theoretical assumptions. During the tests, sudden changes in the amplitude were recorded on the graphs. This amplitudes were caused by the aircraft vibration and they were also caused by sensor taking two samples of data in one second. By comparing the graphs obtained from the results and theoretical assumptions for different speeds and aircraft bank angles it is observed, and it is also confirmed by the theoretical assumptions, that the load factor depends only on the bank angle of the aircraft while the duration of the turn depends on the speed of the aircraft.

Also, with the increase in the aircraft bank angle during the turn, the duration of the turn is reduced because the radius of the turn is smaller. This confirms the theoretical assumption that the minimum possible radius of turn requires the maximum possible load factor which is obtained by increasing the bank angle of an aircraft.

During the test, in one of the cases a dynamic stall was observed, which manifested itself in a decrease in load factor, which confirms the theory of the V-n diagram that during the flight at certain speeds the aircraft would rather stall than there would be any structural damage to the aircraft.

Without keeping records during the test on which side of the inclinometer the ball has moved, it is not possible to determine whether it is a slipping or skidding turn based on the results obtained, only that there has been a lack of coordination.

The measurements carried out confirmed all theoretical assumptions that were made prior to the conduction of measurements.

LITERATURE

[1] Jeppesen, Oxford Aviation Training: Joint Aviation Authorities Airline Transport Pilot's Licence: Theoretical Knowledge Manual (080 Principles Of Flight), Frankfurt: Jeppesen GmbH, 2001.

[2] S. Janković, Mehanika leta zrakoplova, Zagreb: Fakultet strojarstva i brodogradnje, 2002.

[3] Arduino, »Arduino,« [Mrežno]. Available: https://store.arduino.cc/arduino-uno-rev3

[4] Arduino, »Arduino,« [Mrežno]. Available: https://store.arduino.cc/9-axis-motion-shield

[5] Parallax, »Parallax,« [Mrežno]. Available: https://www.parallax.com/downloads/plx-daq

[6] The Cessna Aircraft Company, Pilot's Operating Handbook and FAA Approved Airplane Flight Manual (Model 172R, Serial No. 17280174, Registration No. 9A-DAD), Wichita, Kansas: The Cessna Aircraft Company, 1996.

Small Piston Engine Aircraft Vibration Measurement and Analysis

F. Juretić (student), D. Gerhardinger, A. Domitrović, J. Ivošević
University of Zagreb, Faculty of Transport and Traffic Sciences, Zagreb, Croatia
fabjur116@gmail.com, dgerhardinger@fpz.unizg.hr, adomitrovic@fpz.unizg.hr, jivosevic@fpz.unizg.hr

***Abstract* - This paper is focused on vibrations of a small aircraft with piston engine in operation. The aircraft Cessna 172R is chosen for measurement and analysis of vibration. The paper describes sources of vibration, such as landing gear vibrations, powerplant and propeller vibration, which make a significant contribution to vibration of the whole system, aerodynamic vibration etc. Considering importance of vibration on an aircraft in operation, equipment for measuring vibration is assembled. Equipment, e.g., the measuring set consists of one Arduino UNO board, a tilt sensor SW-420 and a ADXL345 accelerometer. The measuring set is mounted on the aircraft, and data is collected during flight. The collected data is analyzed and shown in charts.**

***Keywords*: piston engine aircraft; vibration; Arduino; operation; instrument; accelerometer; sensors**

I. INTRODUCTION

During operations, piston engine aircraft are exposed to vibrations from various sources. Flying takes place in atmospheric conditions, which involve variable meteorological elements such as temperature, winds, precipitation, etc. Also, all aircraft are equipped with vibration-generating systems, which together can impair both comfort and flight safety.

Vibrations are an indispensable consequence of operation, on small piston-engine aircraft. Vibrations are not usually measured. Therefore, this paper focuses on in-flight vibration measurement and analysis on a Cessna 172R aircraft (hereinafter called C172R). This is accomplished through the installation of a vibration measurement kit. The paper presents and analyzes the measured data.

The first part of the paper describes a C172R on which vibration measurements were performed. The sources of vibration on the aircraft are described. A measuring set is described, it is assembled and fitted to the aircraft for vibration measurement purposes. Data measured during aircraft flight are shown in charts. The analysis of the data given in the charts makes it possible to identify the modes of operation of the aircraft with respect to vibrations. At the end of the paper, conclusions are made.

II. AIRCRAFT CHARACTERISTICS AND SOURCES OF VIBRATIONS

Vibration measurements were performed on the C172 R aircraft, callsign 9A-DAD, owned by the Faculty of Transport and Traffic Sciences, University of Zagreb (Figure 1). The aircraft is intended for flight training of the Faculty of Transport and Traffic Sciences students, at the undergraduate study of Aeronautics, civil pilot program.



Figure 1: Training aircraft Cessna 172R
(© *Faculty of Transport and Traffic Sciences, University of Zagreb*)

The C172R is a single-engine piston light aircraft, four-seater, metal structure, high-wing, with a conventional tail, and a tricycle fixed landing gear [1]. The propulsion system of C172R is a piston engine manufactured by Lycoming, model number IO-360. The engine has four-cylinders, in a boxer arrangement, right rotation with a capacity of 5899 cm^2 (360 $inch^3$) and 134 kW (160 HP) output power at 2400 RPM [2]. The technical and operational characteristics of the C172R are listed in Table 1.

TABLE 1. Technical and operational characteristics of C-172R [2]

Specifications	Values
Lenght:	8.28 m (27 ft 2 in)
Wingspan:	11.00 m (36 ft 1 in)
Height:	2.71 m (8 ft 11 in)
Aeroprofil:	NACA 2412
Wing Area:	16.17 sq m (174 sq ft)
Aspect Ratio:	7.32

Maximum Takeoff Mass:	1111 kg (2450 lb)
Usable Fuel Volume:	212 l (56 US gal)
Engine:	Lycoming IO-360-L2A
Power Output	134 kW (160 HP)
Propeller:	McCauley, 2 blade metal, fixed pitch
Cruise Speed (IAS):	226 km/h (122 kt)
Maximum limit speed: (IAS)	302 km/h (160 kt)
Service ceiling:	4114,8 m (13500 ft)
Maximum Occupants:	4 (1 pilot and 3 passengers)

Piston engine aircraft are exposed to landing gear vibrations, propulsion system vibrations and aerodynamic vibrations.

Landing gear vibrations occur due to aircraft movement on uneven ground, that is, on asphalt or grassy runways. Specific vibrations occur on the nose wheel for aircraft with a tricycle landing gear layout. For vibration damping, the nose wheel is made of two cylinders that are filled with compressed air or nitrogen and hydraulic fluid [3]. The shock absorber has an elastic, damping and dissipation component that protects the aircraft from landing impact. Vibrations occur around the vertical axis of the wheel in the form of a rotary displacement that results in the wheel having a sinusoidal trajectory, and thus, significant stresses that generate material fatigue over time. The problem is solved by installing a viscous torsion piston that dampens the rotational motions around the vertical axis, which is the case in the C172R [3].

The C172R propulsion system consists of a piston engine and a propeller. The vibrations of the propeller blades are caused by torsion impulses. The source of this impulses is engine crankshaft, which is most often directly connected to the propeller, so the same impulses occur on the propeller [3]. The piston engine contains movable parts which oscillate periodically and have corresponding accelerations in either straight or rotational motion. The vibration of the engine is transmitted through the engine mount to the aircraft structure, i.e. to the cabin of the aircraft. In order to minimize vibrations, damping elements are installed on the engine mount [3].

The aircraft moves through the air which is not perfectly still, nor are the movements of the aircraft produced by the pilot through the control surfaces uniform, so oscillations can be expected on the wing. At specific flow rates, which are relatively high, flutter can occur. The source of energy for these vibrations on the wing is an air current that, relative to the position of the wing, i.e. the angle between the free air flow and the chord line of the airfoil, creates a certain force on the airfoil, i.e. on the wing. For a given angle of attack, the wing produces a certain lift force, and with it, a certain intensity of momentum, that twists the wing. Under the influence of free air flow, there is a possibility of vibrations which can be dangerous to the wing structure [3].

III. MEASUREMENT SET FOR THE COLLECTION OF VIBRATION DATA DURING FLIGHT

Measurement set for the collection of vibration data during an aircraft flight, consists of the Arduino UNO platform and the associated sensor platform. The most important function of a measuring set is to transfer data from a sensor to the computer, in order to display data in tables or charts. Another function of the system is to signal the presence of vibration via light emitting diode, using the SW-420 sensor.

The following sections describe the parts of the measuring set and how to mount it on an aircraft.

A. Grove SW-420

The SW-420 is a multifunctional vibration sensor manufactured by Grove, also called a tilt or displacement sensor. This platform is very simplified in terms of vibration interpretation. Simplification is achieved in such a way that no specific vibration parameter such as amplitude, frequency or period is obtained during measurement. Only information is provided on whether or not vibration is present. [4].

The platform is used to mount to the vibration source together with the Arduino UNO platform with a tighter connection that will not dampen the vibration. Also, the mounting position of the measuring set on the aircraft is of great importance. The orientation of the sensor placement is not very important for measurement since the sensor measures vibrations omnidirectionally [4]. Furthermore, during the measurement, the system recognizes vibrations in such a way that the mass under the influence of inertial forces is removed from its initial position and breaks the circuit. Such information is interpreted as vibration and the system indicates that vibration is present. Vibration is presented in μs, which represents time of the mass displacement. System is programmed that it can register vibration period from 10 μs up to 3 minutes. Sensitivity threshold can be adjusted. This can be achieved by adjusting the potentiometer, which, due to the adjustment, will filter only vibrations of higher or lower interval between periods. [4]. The coupling of the peripheral platform SW-420 together with the associated LED to indicate the presence of vibration is shown in Figure 2.

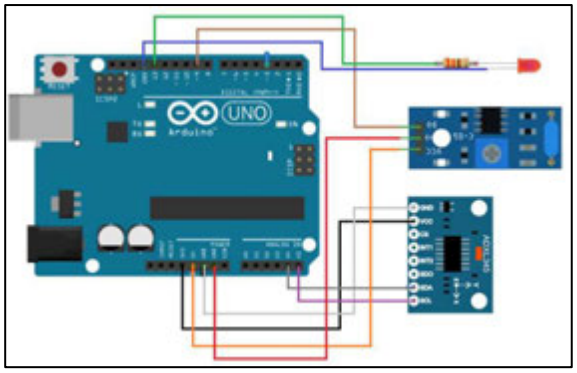


Figure 2. Connecting the Arduino UNO Platform to the SW-420 and ADXL345 Peripheral Platforms, [5]

The appearance of the total measuring set with built-in and coupled sensors and light emitting diode inside the housing is visible in Figure 3.

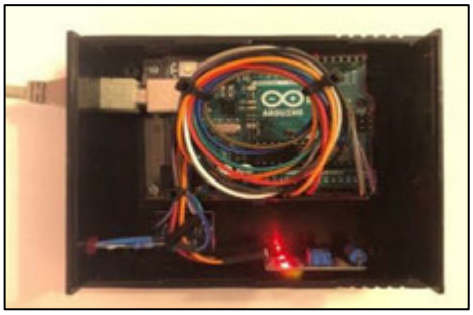


Figure 3. Measuring set

When vibration or displacement affects the system, the time of the impact of the inertial force on the mass is measured. The measured time is displayed in microseconds at the user-visible interface. To make the system's vibration easier to monitor graphically, numerical values are added to the time intervals and accurate recording times.

B. ADXL345 accelerometer

Accelerometers are used to measure the acceleration of individual systems. The application of these systems is broad: measuring angular acceleration and therefore position, applications in navigation with inertial navigation systems, and vibration measurement. In addition to the simplest mechanical accelerometers, Micro Electro Mechanical System (MEMS) are used on the peripheral platform of the ADXL345 measuring instrument [6].

The MEMS has an electromechanical principle of operation. The mass is placed between two metal plates which together with the mass make up two empty spaces between the mass and the plates. The system is housed in a vacuum that protects the system from variable atmospheric parameters for accurate measurement. As the mass moves, the gaps between the metal plates and the mass change in proportion to the acceleration. Ideally, this system is designed as two capacitors, which in this case represent the spaces between the mass and the plates, so the capacitor capacity also changes in proportion to the acceleration influenced by the system. The difference in capacities is proportional to the acceleration and so it is interpreted [7].

The possibility of connecting the ADXL345 peripheral platform to the Arduino UNO is visible in Figure 2.

The ADXL345 measures acceleration in three axes, so the installation of the system on the a/c must be designed to assume the vibration axis in which the vibration will be most intense. Its measurement resolution is 3.9 mg/LSB. [8].

The intensity of the vibrations that are assumed to be highest is expected in the transverse axis of the a/c due to vibrations resulting from the operation of the engine. Acceleration measured by the instrument will show an acceleration expressed in terms of standard gravity – g (9.81 ms^{-2}). Accordingly, the instrument should be positioned perpendicular to all three axes of the a/c. In this case, measurements without vibrations will show that in the vertical axis Z a reading of 1 is obtained, while the remaining two axes will give a reading of 0.

C. Mounting of measuring set on C-172R aircraft

The measuring set is designed to allow measurements to be made to confirm the vibrations and to make the data for further analysis. The sensors and Arduino UNO platform are built into a rigid plastic housing that allows vibration to be transmitted to the peripheral platforms where the sensors are located. The system has one displacement sensor and one accelerometer, so vibrations are measured for the entire a/c. The measuring set is mounted inside the cockpit to avoid the influence of atmospheric conditions (low temperature, high relative humidity, precipitation, etc.) that could affect the accuracy of the measurement. The chosen location for mounting the measuring set is the rear seat support structure. The structure consists of square-shaped metal elements directly connected to the supporting beams of the a/c structure. The location was chosen because it is directly connected by rigid connections to the fuselage structure, which ensures the vibration transmission of all elements of the a/c.

The orientation of the axis and the appearance of the installation of the measuring set can be seen in Figure 4.



Figure 4. Installation and orientation of measurement set on a/c

The measuring set is attached to the structure with a self-adhesive tape, which enables easy installation without additional refinements of the a/c structure. The housing, and therefore the sensors inside, are positioned vertically on all three axes of the aircraft. The X axis represents the longitudinal axis of the a/c, the Y axis represents the transverse axis of the a/c, while the Z axis represents the vertical axis of the a/c. Before installation on the a/c, the entire measuring set is calibrated.

IV PRESENTATION AND ANALYSIS OF DATA OBTAINED BY IN-FLIGHT MEASUREMENT

Vibration measurement is performed during IFR flight (*Instrument Flight Rules-IFR*) on an aircraft C-172R, registration 9A-DAD. Flight is conducted on 30th August 2019. Flight operation begun at Lučko airfield (LDZL) and was continued toward IFR waypoint KOTOR with intention of flying approach at Zagreb Airport – Franjo Tuđman (LDZA). During the entire flight, three approaches were made to runway 05, followed by a return and landing at Lučko Airport. The route is recorded by the GPS receiver and the part of the route that is relevant for the analysis of the set with the vertical profile is plotted in figure 5. The numbers on the figure 5. indicate the characteristic points associated with certain flight phases. The track of the flight and vertical profile in Figure 5. are extracted from the MapSource 2007 application used by Garmin's GPS76 GPS receiver.

Numbers in figure 5. indicate following:

1. Engine start, start of measurement in phase of taxiing
2. Start of measurement in phase of climb
3. Start of measurement in phase of cruise
4. Start of measurement in phase of descent
5. Start of measurement for landing.

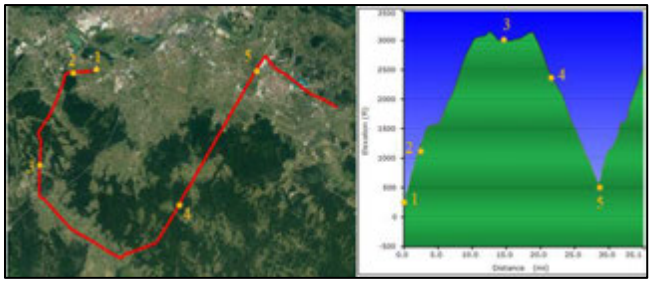


Figure 5. Track and vertical profile with significant points

The data collected from the measurements is presented graphically. Each chart shows a specific accelerometer axis, or oscillation time collected from sensors. Data collected from ADLX 345 accelerometer is expressed in g acceleration [m/s²]. Data collected from tilt sensor SW-420 is expressed in microseconds [µs] which present displacement of mass inside sensor. Each flight phase is associated with the description and the graphical representation. Also each phase of flight are associated with a number, which in Figure 5. shows the point on the route to which the graph relates.

A. *Engine Start*

The engine start involves cranking the DC starter after which engine continues with self-sustainable operation. Chart 1. shows the operation of the engine at the parking position on the apron at 1000 RPM. Higher intensity oscillations are seen when operating the DC starter and generally starting the engine, followed by the engine operation at 1000 RPM with smaller oscillations.

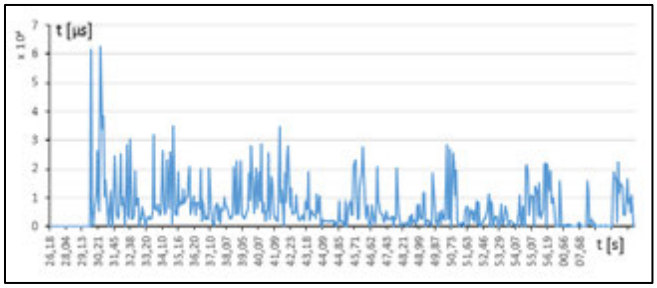


Chart 1. Data collected from SW-420 sensor during engine start

B. *Taxiing*

Taxiing is the phase of flight when the aircraft begins to move on the maneuvering area by force of its engine. Taxiing involves moving from a parking position, testing the braking systems, taxiing to the runway holding point. Taxiing in this case involves maneuvers on paved taxiways and taxiways with natural surface which is in this case grass. The difference of the data shown in terms of unevenness of

maneuvering surface is also evident in chart 2. Taxiing was performed at a speed of 10 km/h at 1200 RPM. Chart 2. shows the data collected through the Z axis of ADXL345 sensor.

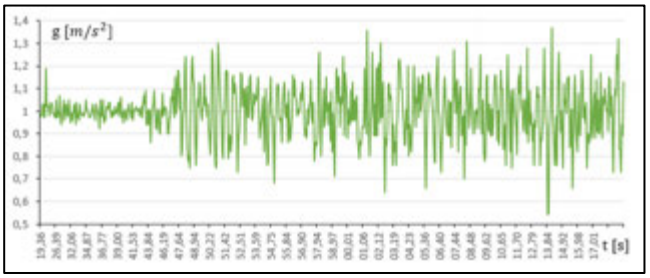


Chart 2. Data collected from ADXL345 sensor for Z axis during taxiing

C. Climb

The climb phase follows the take-off and contains several types of climb. For measuring purposes, the best rate of climb is considered, which in this case is performed on acceleration altitude, or 91 m (300 ft) above ground level (AGL). Data shown is measured during the climb from 213 m (700 ft) AMSL (above mean sea level) to 457 m (1500 ft) AMSL. Best rate of climb implies that the power setting is set to maximum power and the aircraft speed is 75 knots (V_y). In addition to the phase of flight, which has a significant contribution to the oscillation of the complete system, it is worth noting that the meteorological situation also has a significant contribution, especially when there are relatively high vertical airflows (convection) as it was during the flight on which the measurement was performed. In contrast to taxiing, there is decreased acceleration in oscillation measurements, especially due to rotation and initial climb, which implies that the aircraft no longer moves on uneven ground. Chart 3. shows the data collected through the X axis of ADXL345 sensor.

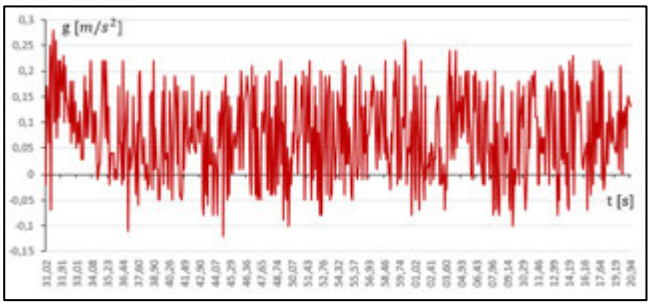


Chart 3. Data collected from ADXL345 sensor for X axis during climb

D. Cruise

The cruise phase occurs after the aircraft stops climb, or reaches the desired altitude of flight level. In general, cruise is characterized by a change in aircraft attitude (decrease in pitch) and a reduction in power relative to climb power. For the specific measurement case, cruising was performed at 915 m (3000 ft), maximum power is reduced to 61% which implies (at the time of atmospheric conditions) engine operates at 2100 RPM and cruising speed is 185 km/h (100 kt IAS). Cruise measurements show that the vibrations have lower intensity, primarily due to the engine's reduced power comparing to climb power. Chart 4. shows the data collected using the X axis of ADXL345 sensor.

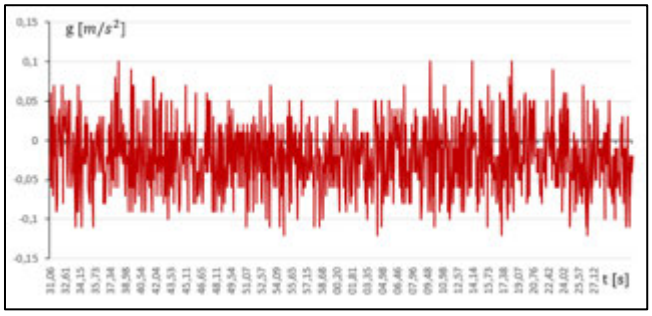


Chart 4. Data collected from ADXL345 sensor for X axis during cruise

E. Descent

The descent follows the cruise phase and its purpose is to reach the altitudes which are prescribed for specified approach. Such change in aircraft operation is characterized by a reduction in power, decrease in cruise speed, and change in aircraft attitude. Also the flaps are extended in order to reduce speed and obtain the desired approach and landing performance of an aircraft. In the particular case of measurement, the descent began at a height of 915 m (3000 ft). The descent was initiated for ILS approach which was followed by touch and go. The descent was continued with airspeed of 185 km/h (100 kt IAS) and vertical speed of 531 ft/min while engine was set to 1650 RPM. The measurement data shown in the graphs shows that oscillations have lower intensity in the descent than the cruise which is primarily caused due to a change in power setting to lower power. Chart 5. shows the data collected through the X-axis of ADXL345 sensor.

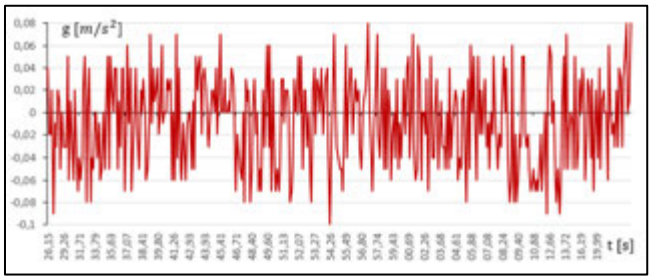


Chart 5. Data collected from ADXL345 sensor for X axis during descent

F. *Landing*

Landing is the phase of flight that follows at the end of the final approach and involves the aircraft touching-down at the runway surface and slowing down on the ground. In this specific case, the speed was reduced to 120 km/h (65 kt) at an altitude of 473 m (1550 ft AMSL) at outer marker-OM. The flaps had been gradually extended to 20°. Just before touch-down, the power setting was reduced to idle and the pitch is increased which is also called flare. Chart 6. shows landing data, where is visible distinctive moment of touch-down followed by an acceleration, as it was planned to perform a touch and go. Chart 6. shows the data collected through the SW-420 sensor.

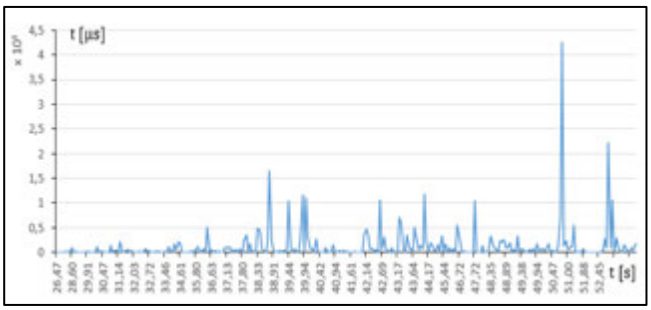


Chart 6. Data collected from SW-420 sensor during landing

V. CONCLUSION

Vibrations are a significant factor in the operation and maintenance of all aircraft. If vibrations act on certain elements of the structure in an unforeseen way, they significantly shorten the time that the structural elements can spend in operation before failure. Vibration monitoring on light aircraft is not regulated, although vibration is present. In-service vibration measurement allows vibrations to be monitored and thus to determine the state and remaining life of the aircraft system. Therefore, for the purpose of research, a measuring set was designed to measure vibrations based on the Arduino UNO platform with associated sensors. The measuring set was mounted on a small piston engine aircraft (C172R) and data from the measuring set were collected on a training flight. Data are graphically presented and analyzed by characteristic phases based on the analysis.

It is concluded that the installation of this simple measurement set provides basic information on aircraft vibrations, which can be subsequently analyzed, due to the fatigue of the aircraft structure. Furthermore, it is possible to observe the impact of vibration on the crew, that is, students and instructors, who are exposed to the vibration during the flight. It possible to parallelise the system, thus measuring vibrations at multiple positions within the aircraft. Also, the same measurement device can be utilised to monitor vibrations in real time, where the occurrence of amplified vibrations on aircraft would indicate that there is a potential failure and warn pilot.

Based on the research presented, it is concluded that vibrations are a significant factor in the operation and maintenance of aircraft. In this paper it is explained the type of equipment needed to measure vibrations on small aircraft. According to the collected data measured during aircraft operation, vibration analysis can be used to explain the influence of vibrations on the flight crew members as well as on the aircraft condition with respect to the number of cycles and flight hours that aircraft and crew are exposed to.

.

LITERATURE

[1] Textron Aviation, Cessna Aircraft, https://cessna.txtav.com/en. 10. 07. 2019.

[2] Cessna Aviation Company, Pilot's operating handbook Cessna Skyhawk, Wichita, Kansas, SAD: Cessna Aircraft Company, 1977.

[3] D. Hartog, Vibracije u mašinstvu, New York, Toronto, London: McGraw Hill Book Company, 1956.

[4] Seed studio, http://wiki.seeedstudio.com/Grove-Vibration_Sensor_SW-420/. 08. 02. 2019.

[5] Arduino.cc, https://www.arduino.cc/. 01. 08. 2019.

[6] Sparkfun, https://learn.sparkfun.com/tutorials/accelerometer-basics/all. 04. 08. 2019.

[7] M. Andrejašič, MEMS accelerometers, Ljubljana: Fakultet matematike i fizike, Sveučilište u Ljubljani, 2008.

[8] PCB Piezotronics, MTS System Corporation, https://www.pcb.com/resources/technical-information/mems-accelerometers. 05. 08. 2019.

Mazzilli Oscillator as a tool in education

Luka Matić, Saša Stojanović i Vladimir Šimović
Department of Electrical Engineering
Zagreb University of Applied Sciences
lmatic@tvz.hr, sstojanovic@tvz.hr, vsimovic@tvz.hr

***Abstract* - In an effort to visually show students the effects of Eddy currents (also called Foucault's currents), we have researched the possibility of using a Zero Voltage Switching circuit for the generation of such currents in a large, light and thin piece of metal. The Mazzilli oscillator provided with a much needed opportunity for such research and an interesting task for application of knowledge and skills acquired during studies.**

Keywords - Eddy currents; ZVS; Mazzilli; Oscillation

I. INTRODUCTION

As an integral part of studying and graduating from the Department of Electrical Engineering at the Zagreb University of Applied Sciences, students have to prepare a graduate thesis paper with an emphasis on practical implementation of acquired theoretical knowledge and practical skill sets during their education. Students choose their mentors based on the possible future direction of their further education or employment. The thesis is formulated in communication between a potential mentor and a student. This paper is a result of one such collaboration.

During the semester, as a part of student teacher evaluation process at the Zagreb University of Applied Sciences, students are asked to fill out a survey for each course and course teacher. The survey, aside from some standard questions about the quality of the course, that are answered with a numerical value from 1 to 5 (1 meaning a failing grade, 5 an excellent grade) also features a comment section.

Over the years, throughout the comment section and its contents, it has become evident that students have a hard time visualizing phenomena depicted in books with 2D images and described by math equations, and that better visualization would help their learning process. This anecdotal evidence is backed by research in this area [1][2][3][4][5].

Specifically, while students are subjected to experimental work in electrical engineering laboratories and simulated environments like Simulink and LabView, they have expressed a need for help in visualizing basic electricity and magnetism phenomena. So, as an interesting phenomenon that was hard to visualize, Eddy currents and their effects were chosen.

A. *Eddy currents*

Eddy currents (also known as Foucault's currents) are loops of electrical current induced in a conductor by a changing magnetic field. This effect is best described by Faraday's law of induction. As seen in Figure 1, Eddy currents flow in closed loops, in a plane perpendicular to the magnetic field. One of the side effects of thusly induced Eddy currents is heat. By achieving a large enough changing electromagnetic field we can produce large Eddy currents which produce a pronounced heating effect. This effect is used as a basis of and induction furnace, but we also use it to show the existence of currents in a conductor.

B. *Inspiration*

The idea for this project was inspired by a permanent exhibit piece at the Palais de la Decouverte in Paris. The piece shows an experiment where a 1kg aluminum plate is levitated above a large coil of wire that is being supplied

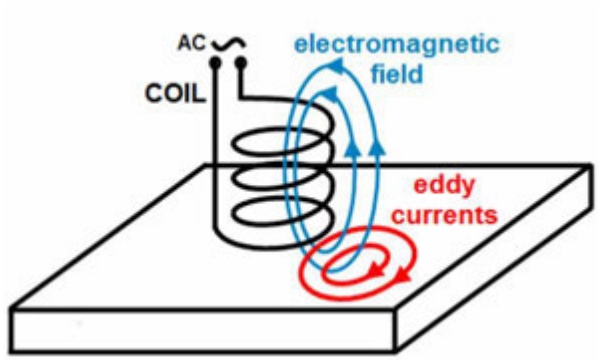


Figure 1. Eddy currents induced by an electromagnetic field generated by the coil



Figure 2. Exhibit piece at the Palais de la Decouverte in Paris

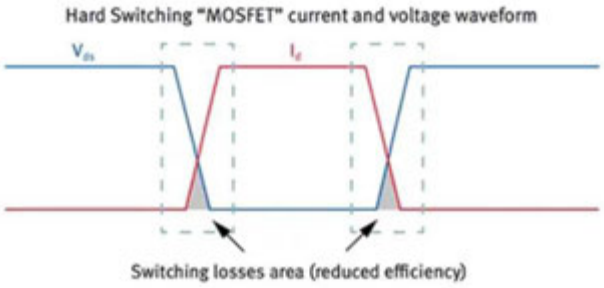


Figure 3. Loss occurence during voltage/current overlap in aswitching MOSFET

with 800 A of alternating current at 900 Hz as shown in Figure 2 [6].

In the exhibit piece, an alternating current is used to create a continually changing magnetic field that induces an alternating current in the plate, producing an opposing magnetic field, which is why the plate is able to levitate, as show in Figure 2. Levitation aside, the existence of the current is easily shown by putting a light bulb on the plate, which then glows. One thing to note is that the lamp in the Figure 2. has a metal oval beneath it, that acts as a coil which acts as a recipient of inducted voltage. We can also spray the plate with water, which would evaporate immediately, because of the heating effect of the induced current has on the plate itself. These are all interesting effects and phenomena that would hopefully help students understand the subject matter better.

After careful thought and research a Zero Voltage Switching (ZVS) circuit was selected for the generation of the needed current for achieving desirable effects that have been described in the paper. The circuit's name is the Mazzilli Zero Voltage Switching converter. To build it, we first had to build and test the Mazzilli oscillator.

II. THE DEVICE

A. *Zero Voltage Switching*

Most contemporary buck voltage regulators are plagued with high switching losses due to the simultaneous incidence of high-current and high-voltage stress imposed on the regulator's integrated Metal Oxide Semiconductor Field-Effect Transistor (MOSFET) switch during the on and off transition periods, as seen in Figure 3. The losses increase with an increase in switching frequency and input voltage which in turn limits maximum frequency operation and general efficiency.

One possible solution, for faster switching frequency at a higher input voltage is Zero Voltage Switching. Zero voltage switching can best be defined as conventional square wave power conversion during the switch's on-time with "resonant" switching transitions. For the most part, it can be considered as square wave power utilizing a constant off-time control which varies the conversion frequency, or on-time to maintain regulation of the output voltage [7].

B. *Principle of the device operation*

First and foremost, we have to note that the constructed device the so called Mazzilli oscillator, while we will be explaining the theory of the Mazzilli converter. The only difference is that the converter is un upgrade from the oscillator, where instead of a coil a transformer is used. It is the next iteration of this project.

The Mazzilli converter, as show on Figure 4 was formed as an idea by Vladimiro Mazzilli, as a derivative of the inductive-capacitance (L-C) MOS oscillator and Royer oscillator [8]. It self oscillates at the L-C oscillation frequency.

The key difference between the L-C MOS and Royer oscillators is the fact that the Royer one needs a resistor at the entry point and a tertiary primary winding for driving the switching transistors while the L-C MOS one uses the

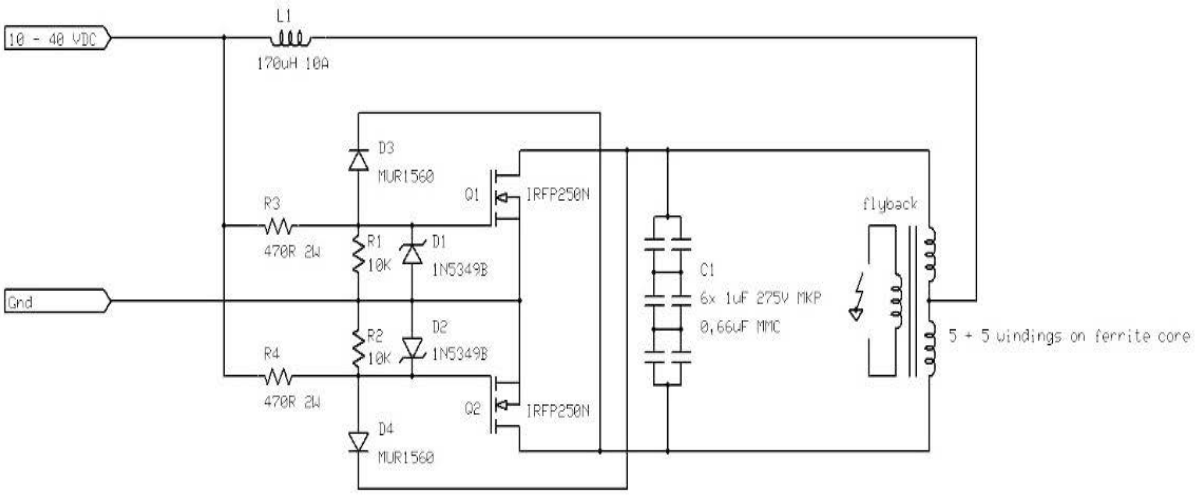


Figure 4. Mazzilli Converter Topology



Figure 5. MOSFET 2 voltage and current oscillogram

opposing drains. The Mazzilli converter uses a combination of the two: pulling power at the entry point while cross coupling the gates, through a diode, to the opposing drain, much like the L-C MOS oscillator.

The converter itself has four modes of operation and they will be best described using the measurements shown on Figures 5 and 7. The measurements are a representation of voltages and currents between two MOSFETs and their various pins, as shown on Figure 6 and in Table 1.

The first mode is when the drains of both MOSFETs are approaching zero and the cross connection ensures that MOSFET 1 is switched off and the MOSFET 2 is turned on. During this process, the resonant capacity is completely discharged and thus all the resonant energy is stored in the inductive component, which is just below saturation. We have used an air core coil. If a ferrite core coil was used, saturation becomes an issue in need of close monitoring.

The switching function monitors the L-C oscillation through the switches themselves, i.e. MOSFETs will not pass resonant current. After the MOSFET 1 is switched off and the MOSFET 2 is turned on, the resonant current begins to resonate, which can be seen in the rise and fall of the voltage U_{ds1}, transferring energy from the inductive element to the capacitive. This would be the second mode.

As U_{ds1} falls to zero via resonance, MOSFET 1 turns on and MOSFET 2 shuts off, switching to the third mode. As in the first mode, the capacity is completely discharged and the resonant current is in the inductive component but in the opposite direction.

The last, fourth mode happens when U_{ds2} resonates to zero, much like in the second mode. When the fourth mode is completed, the first mode restarts. As seen on Figures 5 and 7, the voltages and currents are almost identical between the two MOSFETs.

TABLE 1. FIGURE 5 AND 7 LEGEND

Channel	MOSFET	
	1	2
Yellow	U_{gs1} – gate to source voltage	U_{gs2} – gate to source voltage
Blue	I_{d1} – drain current	I_{d2} – gate to source voltage
Red	U_{ds2} – drain to source of MOSFET 2 voltage	U_{ds1} – drain to source of MOSFET 1 voltage
Green	U_{gs1} – gate to source voltage	U_{gs2} – gate to source voltage

C. *Final implementation*

When implementing the device, at first the IRF740 MOSFETs were used but their usage was abandoned because of the produced current of 22A was too high for them. In the end, the IRF540 MOSFETs were chosen.

By using two IRF540 MOSFETs, made by Vishay Siliconix, connected in the aforementioned way, we have

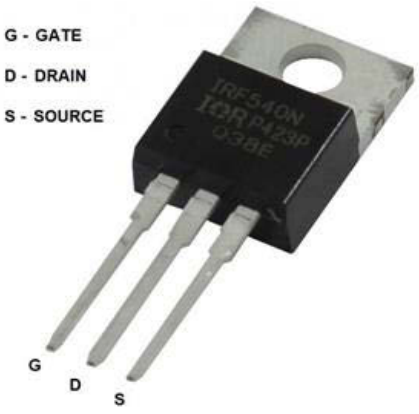


Figure 6. IRF540 MOSFET and it's pins

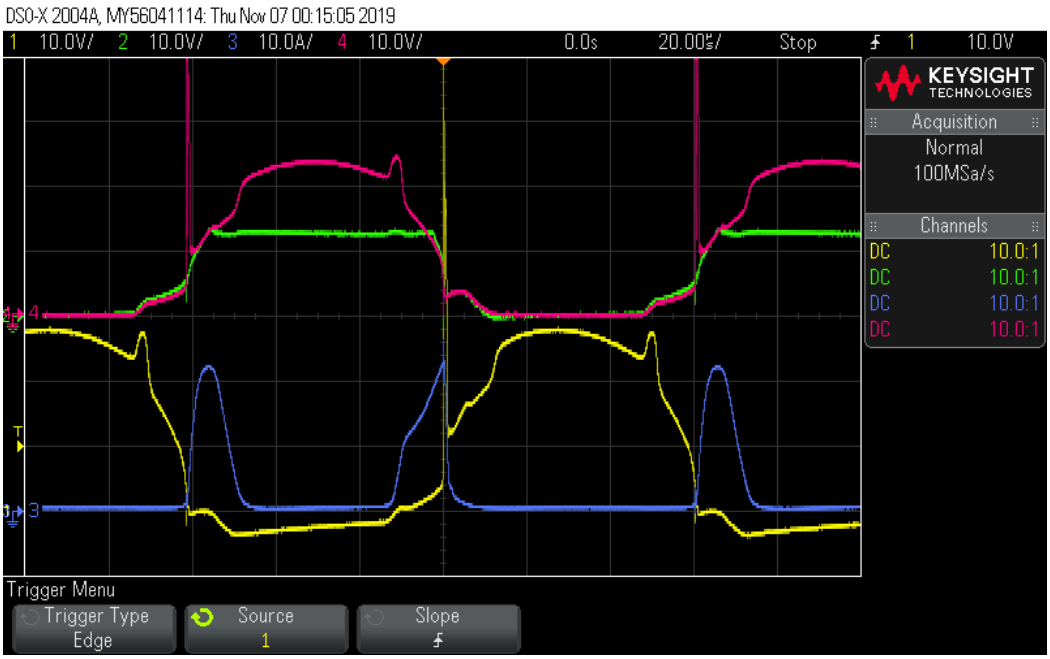


Figure 7. MOSFET 1 voltage and current oscillogram

achieved a device that resonates at a frequency of 8.33 kHz.

Said MOSFET was used because of its characteristics that are suitable for this type of oscillator. Specifically, because of the combination of features of fast switching and low switching resistance and a power dissipation of about 50 W. MOSFET heating is a big problem on device, so 50W of dissipation was a requirement for this circuit to ensure safe operation of the oscillator.

The heat sinks used were selected to dissipate as much power as possible and to draw as much thermal energy from the MOSFET as possible through direct connection without additional thermal insulators as seen in the final implementation that is shown on Figures 8 and 9.

The final product is able to produce a magnetic effect strong enough to induce Eddy currents in a thin plate of aluminum foil and thus repel it and showcase a warming effect. Unfortunately, the effect is not pronounced enough to be showcased in a classroom surrounding, so to exaggerate it even more, an increase in reactive power of the circuit is needed and that is one the planned improvements of the device.

D. Measuring equipment used

The measuring equipment used to aid in the construction of this device was the following:

- Keysight EDUX DSO-X 2004A oscilloscope
- Victor VC890C digital multimeter
- Keysight U1733C LCR meter
- Hall effect sensor

The EDUX DSO-X 2004A oscilloscope was instrumental because of its ability to monitor four channels at the same time which was crucial to obtain visual confirmation of the devices workings, as shown on Figures 5 and 7, among others.

The U1733C LCR meter was ideal for precise measurements of inductance, capacitance and resistance of specific elements of the device that produced the resonant frequency of the circuit.

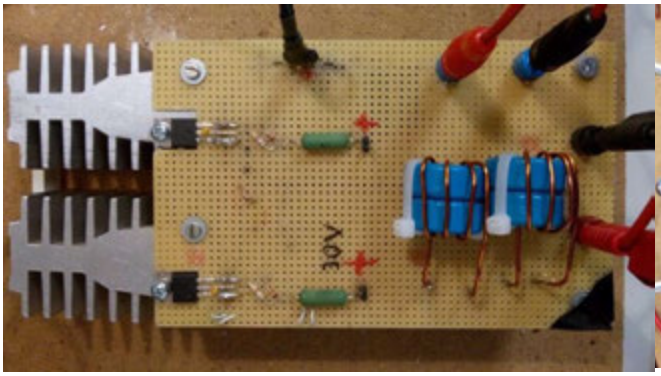


Figure 8. Final version of the device, top side

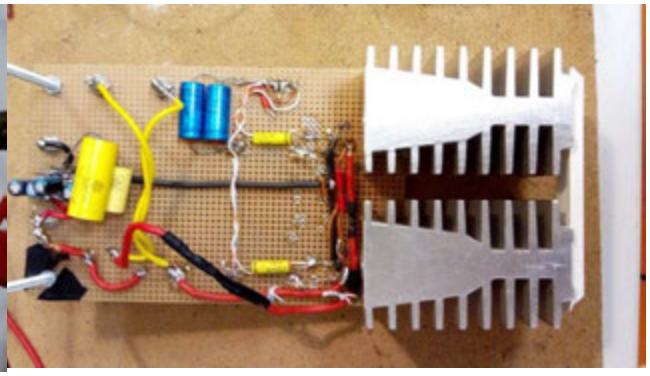


Figure 9. Final version of the device, bottom side

The VC890C multimeter was mostly used to check the soldering connections.

The Hall effect sensor was used to measure the magnitude of the magnetic field.

III. CONCLUSION

The work done on planning and construction of the Mazzilli oscillator has been highly instructive. A large amount of data was collected about the various impacts of the physical position and joining of specific components and its effects on possible oscillations and interferences.

With that valuable data and knowhow, it is safe to proceed to the next step, the construction of the Mazzilli converter and its usage in courses like "Electricity and Magnetism" and "Fundamentals of Electrical Engineering" at the Zagreb University of Applied Sciences.

The project has also provided an excellent opportunity for a student to apply and test lots of the theoretical knowledge he attained during his studies and it will continue to do so in the next phase of the project.

BIBLIOGRAPHY

[1] A. D. Ionita, A. Olteanu, "Support students' experimental work in electrical engineering with visual modeling", Revue Roumaine des Sciences Techniques Serie Electrotechnique et Energetique, vol. 59, 2014, pp. 107-116

[2] M. Balazinski, A. Przybylo, "Teaching manufacturing processes using computer animation", Journal of Manufacturing Systems, vol. 24, 2005, pp. 237-243

[3] F. L. Tan, S. C. Fok, "Development of engineering courseware for university undergraduate teaching using computer animation", Computer Applications in Engineering Education, vol. 3, 1995, pp. 121-131

[4] M. B. McGrath, J. R. Brown, "Visual learning for science and engineering", IEEE Computer Graphics and Applications, vol. 25, 2005, pp. 56-63

[5] M. E. Ismail, I. I. Mahazir, H. Othoman, M. H. Amiruddin, A. Ariffin, "The use of animation video in teaching to enhance the imagination and visualization of student in engineering drawing, IOP Conference Series: Materials, Science and Engineering, vol. 203

[6] "Levitating plate demonstration", Veritasium, 2012, https://www.youtube.com/watch?v=txmKr69jGBk

[7] B. Andreycak, "Zero Voltage Switching Resonant Power Conversion", UNITRODE Power Supply Design Seminar SEM-700, 1990

[8] J. D. Paolucci, "Novel Current-Fed Boundary-Mode Parallel-Resonant Push-Pull Converter", California Polytechnic State University, San Luis Obispo, MS thesis, June 2009

[9] S. L. McClusky, "High Voltage Resonant Self-Tracking Current-Fed Converter", California Polytechnic State University, San Luis Obispo, MS thesis, March 2010

An interactive punch power tracker for heavy bag training

A. Vinković, I. Linardić, D. Meštrović, J. Petrović* and P. Pale*
* Department of Electronic Systems and Information Processing
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, HR-10000 Zagreb, Croatia
{Ante.Vinkovic/Ivan.Linardic/Dario.Mestrovic/Juraj.Petrovic/Predrag.Pale}@FER.hr

***Abstract* - For a martial artist wanting to have a good insight into their progress it is necessary to regularly measure and evaluate their training performance. Unfortunately, most of today's performance tracking devices applicable in martial arts training are expensive and thereby available only to professional sportsmen.**

This paper describes the model and prototype implementation of a device which facilitates information gathering and automated feedback provision during or after a heavy bag training session. Its main features are measuring relative punch strength, frequency, and spatial distribution with a minimalist setup and affordable hardware components. Device prototype implementation is evaluated in authentic settings and shows promising results.

Keywords* - *punch tracker, heavy bag training, punch power sensor, training performance tracking

I. INTRODUCTION

For a martial artist wanting to have a good insight into their progress it is necessary to regularly measure and evaluate training performance. Unfortunately, most of today's performance tracking devices applicable in martial arts training are expensive and thereby available only to professional sportsmen.

In order to address those two issues, the concept and prototype of a device a device for use on a standard heavy bag with the aim of helping martial artists track their training progress were developed. This device was designed with several key principles in mind. First, to provide an efficient way to record physical hits to a heavy bag and to differentiate them from vibrations and other interference during training. Second, to measure user's relative punch power. Third, to design such a device to be resilient, yet as cheap as possible. This paper describes the efforts and results of an attempt to create such a device using accelerometers, piezoelectric film pressure sensors and an Arduino microcontroller board, as well as the methods of registering punches and measuring their strength while maximizing cost efficiency. Two approaches are described, one using both an accelerometer and piezoelectric film sensors to gather data, and one relying solely on use of an accelerometer.

The rest of this paper is organized as follows. In section II similar systems, their functions, advantages and disadvantages are described. Section III lists the requirements and describes the model of the system proposed by the authors. Prototype implementation is described in the section IV and conclusions and future work are described in section V.

II. RELATED WORK

While the idea of using electronic devices to track and improve performance of sportsmen is certainly not new, it seems that in this specific field there is plenty of space for improvement. Several projects and commercial solution related to the topic of this paper have been identified in online searches and are described in this section.

A. "Interactive punching bag" by Fight Cancer

This interactive punching bag was designed by a Dutch non-governmental organization [1]. The bag is supplemented with motion sensors (assumed to be accelerometers), impact sensors (assumed to be piezoelectric sensors), and LEDs that can withstand punches. Sensors measure precision, relative punch strength and the punch angle. The bag is rentable and intended for cancer awareness events, where it is used as a creative way to collect donations for cancer research based on punching sessions.

B. PADIPATA Smart interactive Punching bag IPB-XL

This interactive punching bag is manufactured by a Chinese company [2] and it is designed to have attributes of both sport equipment and of an entertainment device. It comes as a heavy bag on a stand featuring an interactive touchscreen module to choose training programs. While technical details of the implementation are not disclosed, the bag implements a mode where a part of the bag that starts glowing in color needs to be hit and punch power and reaction time are measured, suggesting at least two types of sensors are used again: sensors to locate hit area and sensors to measure punch power.

C. A system for automated scoring and performance analysis for amateur boxing

This paper [3] describes a system consisting of wearable punch sensors integrated into boxers' gloves, vests, and head guards, with wireless data transmission. Obtained signals are passed through a high pass filter and

classified as hits, blocked hits, or deflected hits. Two types of sensors were used and described in the paper: piezoelectric pressure sensors on thin plastic substrates, and accelerometers. Following from additional information about the system and authors' endeavors [4], polyvinylidene fluoride (PVDF) piezoelectric sensor were used, while accelerometers were placed on a boxer's wrists to ensure that a recorded hit was performed by the other boxer.

III. REQUIREMENTS AND SPECIFICATIONS

Following from the identified similar project s and systems, piezoelectric sensors and accelerometers were identified as most suitable for implementing a punch power sensor for a heavy bag. Due to its availability and low price, Arduino microprocessor was used to acquire data from sensors. Arduino mega 2560 was used in order to ensure enough current and pins for multiple connected sensors if necessary. is a single-board microcontroller. It has 54 digital input/output pins (of which 15 can be used as PWM outputs), 16 analog inputs, 4 hardware serial ports, a 16 MHz crystal oscillator, a USB connection, a power jack, an ICSP header, and a reset button. This microcontroller was used to handle all calculations and to interact with the sensors used.

Pressure sensor used in this project is the LDT0-028K piezoelectric film sensor. It is a flexible component composed of a 28-micrometer thick piezoelectric PVDF polymer film with screen printed Ag-ink electrodes, laminated to a 0.125-milimeter polyester substrate, and fitted with two crimped contacts. As the piezo film is displaced from the mechanical neutral axis, bending creates very high strain within the piezopolymer and therefore high voltages are generated. When the assembly is deflected by direct contact, the device acts as a flexible "switch".

The signals from the piezoelectric film sensors were originally meant to be used for both punch recognition and strength measurement. During testing it was concluded that the switch-like nature of these sensors could be used to signal whether contact with the bag has been made. This was concluded due to the sensors giving out too strong signals to be suited for deducing the relative strength of a punch. The piezoelectric sensor is able of providing over 70 volts when bent at a right angle, while the Arduino Mega board's analog inputs are capable of measuring signals ranging from ground to 5 volts, cutting the signal off otherwise. The analog inputs are then digitalized to a 10-bit resolution (i.e. 1024 different values, maximum being 1023). This fact made the calculation of punch strength using the LDT0-028K PVD impossible without further use of electrical components to decrease voltage output or to reduce it by mechanically limiting sensor flex.

The acceleration sensor used in this project is the ADXL337 accelerometer. It is a compact, 3mm × 3mm × 1.45mm, low power, three-axis accelerometer with signal conditioned voltage outputs capable of measuring acceleration with a minimum full-scale range of ±3 g. The output signals are analog voltages that are proportional to acceleration. It can measure the static acceleration of gravity in tilt-sensing applications, as well the dynamic acceleration resulting from motion, shock, or vibration. The accelerometer was chosen to be the better of those two solutions for optimizing the device's punch strength calculations.

IV. RESULTS AND PROTOTYPE IMPLEMENTATION

The first explored method of implementing a device for punch power sensing was by using both the piezoelectric film sensors and accelerometers. The piezoelectric film sensors are used for contact registration while accelerometers are used for measuring punch strength. One pin of the piezoelectric film sensor is connected directly to the Arduino mega microcontroller analog input pin, with the other being grounded. Sensors are then mounted to the bag using adhesive film. ADXL337's X, Y, and Z axis are connected to three analog inputs of the Arduino board. The voltage supply (Vcc) pin is connected to the 3.3 volt output and the ground pin is connected to Arduino's ground pin. The USB 2.0 cable was used as the power source, although it can be substituted with batteries or a regular voltage adapter can be used to achieve device mobility. Using adhesive film, the accelerometer was mounted to either the top or the bottom of the bag in order to minimise obstruction of the heavy bag during training. Testing indicated negligible sensor output differences depending on the position of the accelerometer in relation to the heavy bag.

Accounting for the sensitivity of the LDT0-028K piezoelectric film pressure sensor, being capable of outputting up to 70 volts when bent at the right angle, it was deduced that the best way to benefit from this characteristic would be to consider the piezoelectric film sensor as a switch with the task of registering hits. When the bag is hit, the resulting flex of the bag bends the piezoelectric film sensor mounted on its' surface, producing a voltage spike. This information travels to the Arduino board where it is compared to a preset trigger level. In the case that the signal input is stronger than the trigger level, the signal is marked as a hit.

During early testing it had been noticed that the piezoelectric film sensors give out two voltage spikes when hit. This is known to happen due to the sensor's elasticity. Making it rebound back towards its natural position, which in return flexes the sensor in the opposite direction, producing a negative voltage spike during the process. This problem was dealt with by implementing a so-called "enable trigger". This trigger is a Boolean value with the sole purpose of enabling the Arduino board to recognize piezoelectric sensor's signals. The hit is "enabled" for a read-out only when the previous hit signal decreases to a value lower than the enable trigger's manually set value. The enable trigger value is also designed to be altered during the testing process. Significant sensor output value differences were noted during testing, depending on the material on which the sensors were placed. These differences required the trigger values to be adapted accordingly.

While the device originally consisted of piezoelectric film sensors and accelerometers to register hits and measure their power respectively, an alternative of relying solely on an accelerometer was also considered. This idea consists of using the accelerometer for both detecting hits and measuring their relative intensity by sampling its output data in a way to recognize irregularities and acceleration spikes, all while ignoring noise and other interference. If the data is to be processed correctly, these input spikes would signal the Arduino board that a hit has been received. Amplitude of these signals would then be used to calculate a value which can then be sent back to the user as information on punch strength.

The method with which hits would be registered correctly is based on calculating the difference between the current and previous value within the *loop()* function of the Arduino program. The calculation is done for each of the sensor axis. This delta is compared to a static value chosen semi-arbitrarily. The value is then adjusted by hand in response to device sensitivity and false triggering.

During prototype testing of the accelerometer, data was collected from a series of different states during a simulated training session. Fig. 1 shows the accelerometer outputs from the Arduino board while the heavy bag is resting with Fig. 2 depicting the results of three consecutive punches without obstructing the bag's movement. Fig. 3 shows the bag being hit with a series of light taps and powerful punches while in a swinging motion, simulating conditions during a training session. Horizontal axis marks the elapsed time in milliseconds and the vertical axis shows acceleration intensity digitalized by the microcontroller on a scale from 0 to 1023. As seen in the figures, there are clear differences in signal impulse shape when comparing the hits with the heavy bag swinging freely.

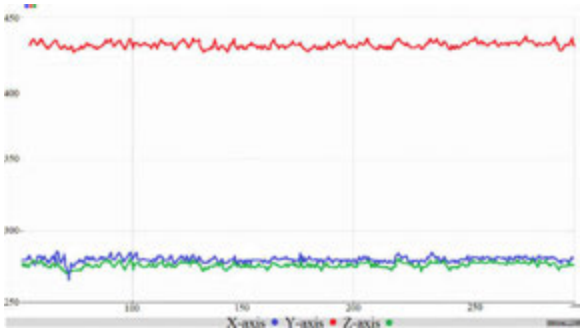


Figure 1. Accelerometer signal output during rest

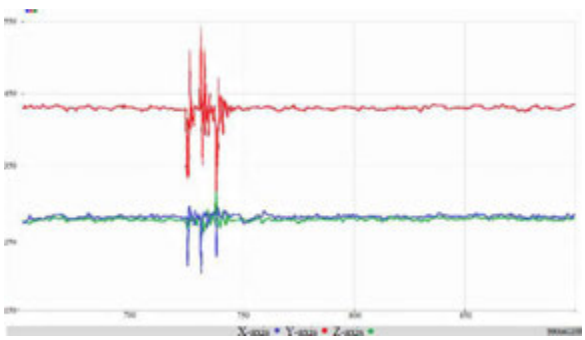


Figure 2. Accelerometer signal output during a three punch series

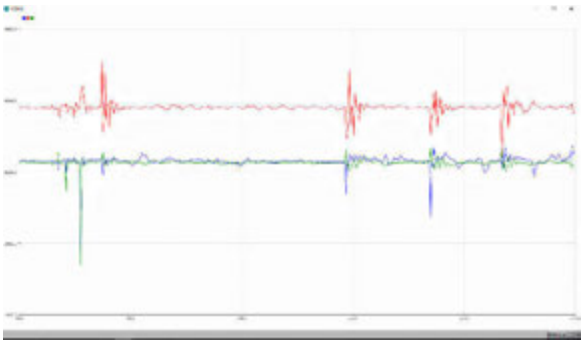


Figure 3. Accelerometer signal output during a short simulated training session

Punch strength can either be measured in relation to prior punches or it can be calculated by converting accelerometer data back into acceleration and converted into punch force by additionally measuring the mass of the punching bag used. Overall acceleration is calculated using signal outputs from the three axes of the accelerometer as $a = \sqrt{a_x^2 + a_y^2 + a_z^2}$, while the heavy bag mass is assumed to be around 40 kilograms. Scaling and differentiation of punch strength as well as the calculation of the strongest and the weakest punch is done by repeated use of method on subsequent punches. A counter for the overall number of punches landed by the user during the training session is included. Every time punch is registered the counter increments by one. The punch counter has been proven to work reliably with high intensity punch series, tested up to ten hits per second. Fig. 4 shows how active punch strength calculation values compare to the three axis acceleration outputs.

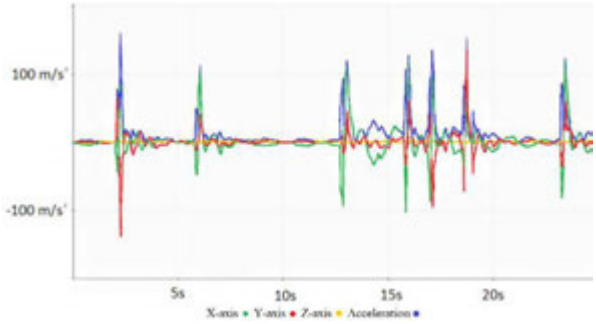


Figure 4. Accelerometer X, Y, and Z axis outputs are shown in green, red and yellow respectively. Calculated overall acceleration is drawn in blue

Since the proposed hardware implementations seem to be suitable for the device function and they were the main limitation in system design, the rest of the system is modelled as follows. Data acquired from an accelerometer will suffice for acquiring timestamps of hits, and their acceleration per axis. This data can be used to calculate last punch power, maximal punch power, or average punch power in the current session. This

information can be displayed on a 7-segment display for feedback during the workout session. A buzzer can be used for a different type of training session, where reaction time between the buzzer time and a punch is measured to provide feedback on reaction time. If using a 2.4GHz module, the system can be connected to a PC using wireless link, thereby record all data regarding a workout session, and evaluate improvements across sessions. This model was, due to a lack of time in the project, not yet implemented as part of the prototype.

The method of realizing the device using only an accelerometer, compared to adding piezoelectric sensors, would prove to greatly simplify how the device could be fitted onto a standard heavy bag by greatly reducing the risk of mechanical damage that is present with the use of piezoelectric sensors for hit registration. Another benefit of this approach is that the cost of assembling the device would reduce significantly, as the numerous piezoelectric film sensors cost more than the low number of accelerometers used. Accounting for the fact that device sensitivity, and therefore hit registration performance is partially determined by the density of the bag on which the piezoelectric film sensors are placed, a structured grid of these sensors would have to be assembled. The average price of a LDT0-028K piezoelectric film sensor is USD $3.50. The cost of assembling the mentioned grid would quickly exceed the price of a single ADXL337 accelerometer that is currently sold at around USD $10.00. It is also important to consider the additional effort that goes into assembling such a grid, as well as designing and building its' housing. All the connections to the sensor pins must be rigid enough to withstand large quantities of hits with varying power. An expected cost of additional components is USD $5.00 for the large 7-segment display, USD $3.00 for the wiring, USD $20.00 for the wireless modules. In total, this adds up to a total system component cost of USD $41.50.

V. CONCLUSIONS

The work described in this paper describes efforts in creating a cheap device for measuring punch power sensor and provide feedback to its user during or after a training session. Two designs were investigated based on the conclusions drawn from related projects: by using two different types of sensors, piezoelectric film pressure sensors for hit registration and an accelerometer for punch power calculation, or by using only an accelerometer for both functions. The latter method is cheaper, less complex in terms of assembly and equally as efficient. The design was shown to be feasible by testing it testing on a heavy bag. Accelerometer readings show clear difference between punches and other irregularities such as bag swing even in the most extreme simulated conditions. During limited testing opportunities, promising results were shown and further improvements for the model of the entire system are discussed in the paper.

REFERENCES

[1] The Interactive Punching Bag to Fight Cancer. www.interactivepunchbag.com. Accessed Dec. 13th 2019.

[2] PADIPATA, www.padipata.com. Accessed Dec. 13th 2019.

[3] A. Krajewski, R.J.N. Helmer and S.R. Lucas, *Signal processing for valid score determination in amateur boxing*, 5th Asia-Pacific Congress on Sports Technology (APCST), 2011

[4] A.G. Hahn, R.J.N. Helmer , T.Kelly , K.Partridge, A. Krajewski, I.
Blanchonette, J.Barker, H.Bruch , M.Brydon , N. Hooke and B.Andreass, *Development of an automated scoring system for amateur boxing*, 8th Conference of the International Sports Engineering Association (ISEA), 2010

[5] Buśko, K., Staniak, Z., Szark-Eckardt, M., Nikolaidis, P. T., Mazur-Różycka, J., Łach, P., Michalski, R., Gajewski, J., Górski, M., *Measuring the force of punches and kicks among combat sport athletes using a modified punching bag with an embedded accelerometer*, Acta of Bioengineering and Biomechanics Vol. 18, No. 1, 2016.

Simulink model of oxygen distribution in skeletal muscle

Amra Džuho
International Burch university
Department of Genetics and Bioengineering
Sarajevo, B&H
amra.dzuho@stu.ibu.edu.ba

Amina Aleta
International Burch university
Department of Genetics and Bioengineering
Sarajevo, B&H
amina.aleta@stu.ibu.edu.ba

Sabina Pandža
International Burch university
Department of Genetics and Bioengineering
Sarajevo, B&H
sabina.pandza@stu.ibu.edu.ba

Irma Ramić
International Burch university
Department of Genetics and Bioengineering
Sarajevo, B&H
irma.ramic@stu.ibu.edu.ba

Nurayim Mamatnazarov
International Burch university
Department of Genetics and Bioengineering
Sarajevo, B&H
nurayim.mamatnazarov@stu.ibu.edu.ba

Lemana Spahić
International Burch university
Department of Genetics and Bioengineering
Sarajevo, B&H
lemana.spahic@stu.ibu.edu.ba

***Abstract*—Cellular respiration is a pathway that uses energy from food molecules and oxygen for different processes of life such as movement. Oxygen, transported through mass transport from blood vessels to skeletal muscle can bind to myoglobin that stores a small amount of oxygen or can be used in cellular respiration by mitochondria. This paper presents a Simulink model of oxygen distribution in skeletal muscle, based on previously published mathematical models. Different parameters for supply concentration of oxygen at capillary wall and consumption rate of the muscle tissue simulate oxygen distribution when the body is under certain conditions: in rest state, dysoxia and exercise state.**

Keywords—skeletal muscle, ATP, mass transport, oxygen distribution, Simulink

I. INTRODUCTION

Cellular respiration is a biochemical pathway in which cells release energy from chemical bonds of food molecules, and provide energy for processes of life [1]. In humans it is done in mitochondria, in three steps: glycolysis (breaking of sugars), Kreb's cycle (production of CO2), electron transport chain (production of adenosine triphospahte-ATP) [2]. Inhaled oxygen it is taken up by haemoglobin (Hb) that carries oxygen and transports it through the body. The shape of the hemoglobin molecules is based on how many O2 molecules are bound to it, and the more oxygen is bound, the higher the molecule's affinity for oxygen is. Percentage of oxygen that is bound to Hb molecules, is related to the partial pressure of oxygen (PO_2). When hemoglobin with attached oxygen moleculesoxyhaemoglobin reaches a tissue that uses oxygen like skeletal muscle, the oxygen dissociates because of the lower local PO_2. This allows for oxygen to be delivered to the tissues and cells that need it the most [3]. Oxygen in skeletal muscle can bind to myoglobin that stores a small amount of oxygen or can be used in cellular respiration by mitochondria [4]. Mass exchange occurs from blood vessels to cells and organelles within cells. In this exchange molecules such as oxygen and other metabolic products cross different macro and micro environments, blood extracellular/intracellular space and interior of organelles and biological barriers such as walls of blood vessels and membranes of cells and organelles [5].

Skeletal muscle is a heterogeneous tissue with different fiber types and metabolic responses of muscle fibers during exercise can reveal regulatory mechanisms that govern metabolism in these cells both in healthy organism and in a disease state. During the transition from rest to exercise state, muscle blood flow increases and muscle fibres are recruited to match the required energy demand. In the exercise state the rate of ATP production increases, so intracellular metabolic processes are activated to fulfill the needs of an energy demand, while the rate of oxygen drops in the cell. Quantitative studies of the geometry of the microvascular network are needed in order to reveal a mechanism by which microvasculature regulates the oxygen supply in relation to metabolic demands of a tissue [6]. Stretched muscles have capillaries that are more or less straight but in contracted muscle they form long meshes about the fibres. Isometric contraction of skeletal muscles causes capillaries to bend and coil [7].

Development of the abnormal tissue oxygen metabolism with changes in oxygen supply and oxygen utilization is the cause of many diseases. Cause of dysoxia, abnormal tissue oxygen metabolism, is high oxygen concentration [1].

Goldman and Popel [7] provided the model based on the experimental data on the hamster cheek-pouch retractor muscle. The model includes geometric components since there is no data available on the structure of capillary networks in this muscle. This model also contains a simulation of two-phase blood flow in the network. They state that their computational model is sufficient to allow application to the study of transport in other skeletal muscles. Their computational model was obtained by using relevant transport equations and physiological data. The most general part is accurately simulation of distribution of oxygen in block of tissue, given geometry and blood flow distribution. Second part of a model was designed to describe both total blood flow and hematocrit distribution through a capillary network. Third part captures certain known anatomical features like diameter and capillary number density. In another study [8], a mathematical model of oxygen transport to tissue determine the way each mass transport parameter affects delivery of oxygen. Most of the models for oxygen delivery to skeletal muscles includes same geometry of long straight capillaries which are parallel to the muscle fibers. The Krogh cylinder model describes the mechanism of oxygen exchange in working skeletal muscle and it was one of the most important models used

which shows the capillary bed of skeletal muscle as being composed of individual capillaries. Each of these capillaries supply a definite cylinder of tissue and is not influenced by other capillaries. Elis et al [8] reported an evidence from scanning electron microscopy that around shortened muscles capillaries form a mesh and under these conditions the Krogh model is no longer appropriate. Also, they display the evidence for striking change in capillary network during contraction of the muscle.

In our model we are representing the difference in oxygen concentration across the skeletal muscle in different stages (rest and exercise state) and in muscles affected by a certain medical condition (dysoxia). The equations used in this model are taken from King and Mody [9]. Using this kind of a model we can determine regulatory mechanism that is responsible for the metabolism in skeletal muscle cells that directly influences an ability to move and participate in normal life activities

II. MATERIALS AND METHODS

Table 1: Parameters for determining the oxygen transport from capillaries to the skeletal muscle [11][12][13].

Abbreviation	Description	Value
D	O2 diffusivity in tissue	2.41 x 10^{-3} mm^2/s
Γ	Rate of O2 consumption (average)	180-200 ml/min 3-4.6 ml/s
No	Supply concentration of O2 at capillary wall	60-90%
n	Number of capillaries (average)	319/ mm^2
L	Length of muscle (average)	30 mm
x	Range of Muscle length (average)	20 mm
t	Time	10 s
e	The base of the natural logarithm	2.71828
π	Mathematical constant	3.14159
Ns	Initial O2 distribution	Obtained by model
N	O2 concentration in tissue	Obtained by model



Figure 1: Block diagram representing the system with input parameters (from top to bottom order: supply concentration of O_2 in capillary wall, rate of O_2 consumption (Γ), length of muscle, O_2 diffusivity in tissue, range of muscle length, rate of O_2 consumption (Γ1), rate of O_2 consumption (Γ2), number of capillaries, mathematical constant π, mathematical constant e and time) and O_2 concentration in tissue as an output.

A. Equations

$$N_s = N_0 - \left(\frac{\Gamma x}{2D}\right)(L - x) \quad (1)$$

$$N = N_s + \left[\frac{4(\Gamma 2 - \Gamma 1)L^2}{D}\right]\left[\left(\frac{1}{(n\pi)^3}\right) e^{-(n\pi)^2 Dt/L^2} \sin\left(\frac{n\pi x}{L}\right)\right] \quad (2)$$

The equation (1) is part of (2) that is modelled in Matlab and presented in the results.

The equation (1) represents the initial O_2 distribution in skeletal muscle, while equation (2) represents the final O_2 consumption in the muscle. Since the equation has Γ1 and Γ2 and their range is the same as Γ, 3 ml/s or 180 ml/min is used for Γ1, and 4.6 ml/s or 280 ml/min is used for Γ2 in normal condition.

III. RESULTS

SIMULINK model was developed for oxygen concentration in the tissue and different concentrations are presented in different conditions in two-dimensional coordinate system. X-axis represents the time while the y-axis is oxygen concentration in the tissue. Parameters used for modelling are used from Table 1 and the results are presented in the following figures. Time was set to be 10 seconds. In order to represent different conditions different values of supply oxygen concentration at capillary wall and rate of oxygen consumption were used, while the rest parameters were kept constant.

Figure 2 represent the initial oxygen distribution in skeletal muscle cells in normal condition (N_0=75% and Γ=3.83 ml/s), exercise state (N_0=100 % and Γ=5 ml/s), low saturation of oxygen state (N_0=45% and Γ=2 ml/s) and disease state (N_0=105% and Γ=2 ml/s).

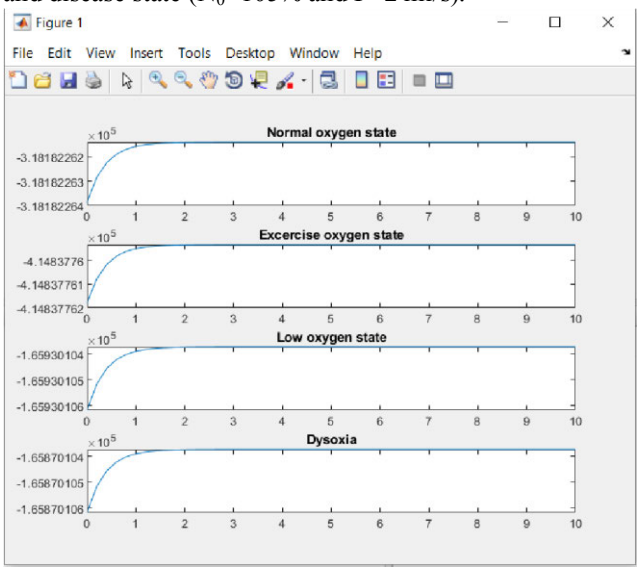


Figure 2: Graph showing initial oxygen distribution in tissue in four different states: normal, exercise, low O_2 and dysoxia.

Figure 3 represents oxygen concentration in skeletal muscle tissue during normal state (N_0=75% and Γ=3.83 ml/s), exercise state (N_0=100 % and Γ=5 ml/s), low state (N_0=45% and Γ=2 ml/s) and disease state (N_0=105% and Γ=2 ml/s).

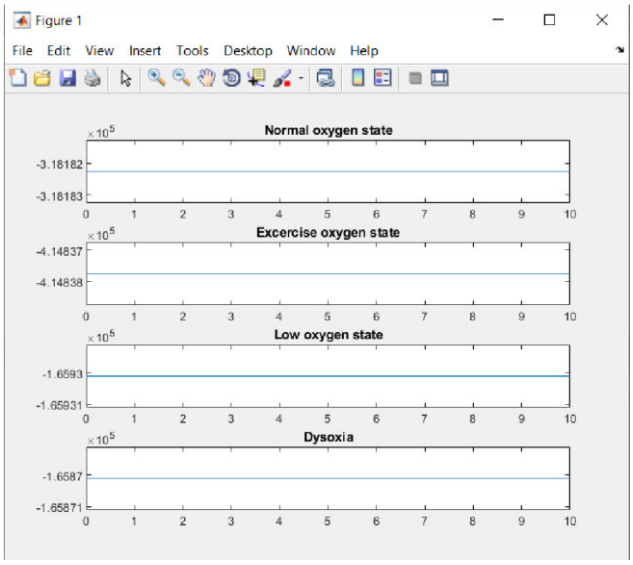


Figure 3: Graph showing oxygen concentration in tissue in four different states: normal, exercise, low and dysoxia state

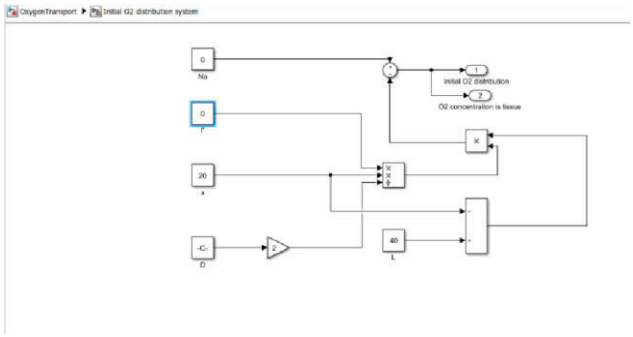


Figure 4: Developed Simulink subsystem of a model based on previously defined mathematical models.

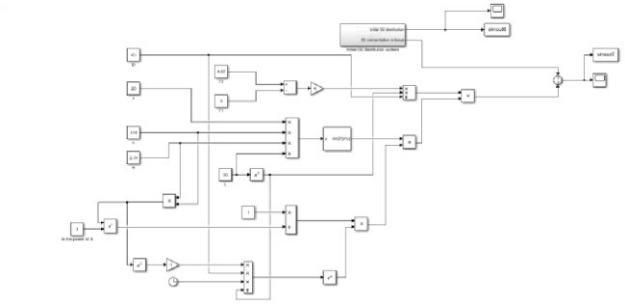


Figure 5: Developed Simulink model based on previously defined mathematical models for calculating oxygen concentration in skeletal muscles.

Figures 4 and 5 present developed Simulink model for oxygen distribution in tissue. Based on previous mathematical models, different Simulink blocks are used to present various components that are part of the oxygen transport and distribution in skeletal muscle tissue. By usage of equations (1) and (2) model present calculation of oxygen concentration in skeletal muscles. As it can be seen from figures 2 and 3, different states of oxygen distribution in skeletal muscles in human body are covered through model.

IV. CONCLUSION

Simplified Simulink model for presentation of oxygen distribution in skeletal muscle cells has been developed and presented in this paper. Using equations from [8] different conditions that affect oxygen distribution were simulated using range variables and changing the rate of O_2 consumption and supply concentration of O_2 at the capillary wall. The dysoxia state, rest state and exercise state of the muscle cells oxygen metabolism were modeled and this information can be used for better understanding of oxygen metabolism as a major source of energy production pathway in mitochondria.

Previously reported studies were using mathematical and computational models to determine mass transport parameters affecting delivery of oxygen [7], while in this study the model is based on the oxygen concentration in the tissue after the mass transport from blood vessels to the tissue. Another paper has taken into consideration the geometric components and two-phase blood flow as well as capillary density and network to provide computational model by transport equation and physiological data to provide details on oxygen distribution on hamster-cheek muscle [6].

This model uses the number of capillaries to simulate oxygen transport in skeletal muscle. In order to improve the model, the network and different numbers of red blood cells may be used. Future research should be focused on including information about blood components, and different density of capillaries in different movements in order to develop a more detailed and accurate model that would represent oxygen distribution in certain activity.

V. REFERENCES

[1] Robin, E. D. (1977). Dysoxia: abnormal tissue oxygen utilization. Archives of internal medicine, 137(7), 905-910.

[2] Wolfe, E. (2016). 3 Simple Stages in Cellular Respiration and How They Work. Available: https://medium.com/countdowneducation/3-simple-stages-in-cellular-respiration-and-how-theywork-4de3abc5f819. Last accessed 12th December 2019.

[3] Clay, W.. (2018). Oxygen Transport in The Blood. Available: https://teachmephysiology.com/respiratory-system/transport-inthe-blood/oxygen-transport/. Last accessed 12th December 2019.

[4] Rowland, T. (2012). Iron deficiency in athletes: an update. American Journal of Lifestyle Medicine, 6(4), 319-327.

[5] Kojic, M., Milosevic, M., Simic, V., Koay, E. J., Kojic, N., Ziemys, A., & Ferrari, M. (2018). Multiscale smeared finite element model for mass transport in biological tissue: From blood vessels to cells and cellular organelles. Computers in biology and medicine, 99, 7-23

[6] Goldman, D. (2008). Theoretical models of microvascular oxygen transport to tissue. Microcirculation, 15(8), 795-811.

[7] Goldman, D., & Popel, A. S. (2000). A computational study of the effect of capillary network anastomoses and tortuosity on oxygen transport. Journal of Theoretical Biology, 206(2), 181-194.

[8] Ellis, C. G., Potter, R. F., & Groom, A. C. (1983). The Krogh cylinder geometry is not appropriate for modelling O 2 transport in contracted skeletal muscle. In Oxygen Transport to Tissue—IV (pp. 253-268). Springer, Boston, MA.

[9] King, M. R., & Mody, N. A. (2010). Numerical and statistical methods for bioengineering: applications in MATLAB. Cambridge University Press.

[10] Nave, R. (N.D.). Cellular Respiration. Available: http://hyperphysics.phy-astr.gsu.edu/hbase/Biology/celres.html. Last accessed 12th December 2019.

[11] Li, Y., Lai, N., Kirwan, J. P., & Saidel, G. M. (2012). Computational model of cellular metabolic dynamics in skeletal muscle fibers during moderate intensity exercise. Cellular and molecular bioengineering, 5(1), 92-112. Cindy S. Cheng, Timothy R. Koves, William E. Kraus, and George A. Truskey. (2017).

[12] Davis, B. N., Santoso, J. W., Walker, M. J., Cheng, C. S., Koves, T. R., Kraus, W. E., & Truskey, G. A. (2017). Human, tissueengineered, skeletal muscle myobundles to measure oxygen uptake and assess mitochondrial toxicity. Tissue Engineering Part C: Methods, 23(4), 189-199.

[13] Basic Principles of Oxygen Transport and Calculations. Available: https://nataonline.com/np/437/basic-principles-oxygen-transportand calculations?fbclid=IwAR0j7CRnZnwYim7V_ogb3kakeCFA1Ohf sFhVQR9BhtlpNWJ3yUjSb80GXuM. Last accessed 12th Dec 2019.

Multi-Model Databases - Introducing Polyglot Persistence in the Big Data World

I. Košmerl*, K. Rabuzin* and M. Šestak**
* University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
** University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia
ikosmerl@foi.hr

Abstract -**The popularization of the NoSQL movement has led to the development of non-relational data models with different approaches to managing Big Data. As each approach has its advantages and disadvantages, there is a need to incorporate certain concepts of the relational data model. With the combination of relational and non-relational data models, monolithic systems based on a single data model began to transform into the first multi-model systems. Most currently used database management systems can be seen as multi-model systems (e.g. Oracle, MySQL, Microsoft SQL Server, etc.). The aim of this paper is to describe the basic concepts of non-relational data models and their advantages and disadvantages. The focus will be put on describing multi-model databases and discussing how multi-model databases contribute to achieving polyglot persistence in modern Big Data environments. The properties of multi-model systems will be compared to those of relational and NoSQL systems, and demonstrated by building a sample college scheduling application.**

Keywords - Multi-model databases; Polyglot persistence; Big Data; NoSQL; Non-relational data models; OrientDB.

I. INTRODUCTION

Nowadays, the availability of a growing number of database vendors on the market further complicates the selection of a given database solution for a given use-case. Various elements need to be taken into account when selecting the most appropriate database solution for a given application domain, which include the properties and requirements of the domain itself, as well as the properties of these database solutions. The properties of Big Data (volume, velocity, variety, veracity, etc.) bring new challenges to existing data storage solutions. Relational databases show some difficulties in terms of flexibility and scalability, while NoSQL solutions still did not achieve the maturity level and robustness of relational databases.

For years, system architects built their solutions around a single Database Management System (DBMS) with the most appropriate underlying data model. However, in Big Data environment, which require extensive flexibility and scalability, neither SQL nor NoSQL databases can fully meet these challenges without certain repercussions related to horizontal scalability, handling data with different formats, consistency, etc. As a solution to this problem, the concept of polyglot persistence systems has been introduced. The entire paradigm is based on using the most appropriate database solution for a given part of the system, which results in multiple data storage technologies being used for the same use-case. For instance, both relational and graph databases might be used to implement a web shop, where relational databases would be in charge of handling transactions, whereas graph databases would be used to implement product recommendations. By following the polyglot persistence paradigm, we ensure that the strengths of different database solutions are combined to satisfy the needs of a Big Data environment. However, polyglot persistence systems also bring some challenges to the mentioned environments (e.g. security, logical redundancy, etc.).

In recent years, multi-model databases have emerged as an alternative, which can resolve the polyglot persistence issues, while still providing enough flexibility. Namely, instead of using multiple systems to manage data, multi-model databases introduce the usage of a single database solution, which supports multiple models. This idea has been accepted by relational DBMS vendors, such as Microsoft and Oracle, which upgraded their "pure" relational DBMSs (SQL Server and Oracle) to support multiple data models, so now it is more precise to consider these DBMSs as multi-model as well [1]. However, note that every multi-model database solution has a "native" data model, while others are used to complement that model (e.g. the relational data model is the native model of SQL Server, while graph data model is implemented on top of this model).

In this paper, we discuss the requirements of modern Big Data applications in order to reason about the importance and usability of the polyglot persistence paradigm. Specifically, the following contributions are presented:

- We provide an overview of strengths and weaknesses of different data storage solutions in the context of Big Data requirements,
- We discuss the properties of polyglot persistence systems, and how they meet the requirements of Big Data properties,
- We discuss the properties of multi-model databases to justify their increasing usage in modern Big Data environments, and
- We demonstrate the applicability of multi-model databases (specifically, OrientDB) by implementing a college scheduling application.

The rest of this paper is organized as follows: in Section 2, we present an analysis of the applicability of relational and NoSQL database solution in managing Big Data, and we study characteristics of polyglot persistence systems in

general. In Section 3, we discuss the characteristics of multi-model databases, and how they fit into the polyglot persistence paradigm. Section 4 describes a prototype college scheduling application built to demonstrate the usage of multi-model databases in different domains. We conclude the paper by summarizing our discussions and the possibilities of multi-model databases as an alternative to build polyglot persistence systems in the Big Data world.

II. POLYGLOT PERSISTENCE AND BIG DATA

Modern Big Data applications are built on the premise of being distributed systems able to store structured, semi-structured and unstructured data at once. Even though relational databases have provided a standardized approach to manage consistent data for years, their lack of flexibility due to fixed schema and difficulties in horizontal scaling required for distributed systems has become a significant hurdle over the years, especially in modern dynamic environments. Additionally, the rise of machine learning has brought new requirements regarding the analytical capabilities of database solutions, i.e., their mechanisms for gaining new insights from stored data.

To deal with these challenges, non-relational solutions have emerged, known as NoSQL data stores. The NoSQL paradigm is based on not using SQL language to manage data, as well as efficiently managing large volumes of schema-less data (not necessarily structured)[2], which makes them more adapt to modern data management requirements in the Big Data era. NoSQL data stores include four different types of data storage solutions, which differ in the underlying data model:

1. Key-value databases,
2. Document-oriented databases,
3. Column-oriented databases, and
4. Graph databases.

Each of these NoSQL categories focuses on specific data characteristics (e.g. data structure, connectedness, completeness of data entries, etc.), which allows database architects to select the NoSQL solution, which best suits their business needs (e.g., graph databases would be a good solution to store social network data). In general, NoSQL data stores as alternatives to their relational counterpart bring many benefits, more important of which are the following:

- The absence of a fixed schema brings greater flexibility for future changes,
- Query processing does not include resource-intensive JOIN operations, which results in better query performance, and
- The ability to scale horizontally due to data replicated across more nodes in the cluster and the underlying BASE consistency model (instead of ACID), which guarantees eventual (and not instant) consistency of data in favor of its availability.

Nevertheless, implementing BASE instead of the ACID consistency model appeared to bring issues regarding consistency and validity of stored data, which should not be ignored in certain domains (e.g. banking). Hence, NewSQL solutions were introduced, which combine the strengths of both SQL (e.g. standardized query language and ACID model) and NoSQL databases (e.g. clustering and scalability), but are still in their early years of development.

As discussed in [2], both SQL and NoSQL solutions have their disadvantages related to their scalability, flexibility, ability to ensure data consistency, and others. Nonetheless, Big Data environments require a data storage solution able to handle both structured and unstructured data, scale depending on the workloads, adapt to changing requirements, quickly recover in the event of failures, deliver query results fast and accurately, and provide mechanisms for creating value from stored data. This is where it became apparent that Big Data systems with a "one-size-fits-all" single data storage solution for the so-called monoglot persistence may not be the best choice anymore. Namely, both SQL and NoSQL solutions can only show flexibility until a certain point given their capabilities, beyond which there is a risk that some domain requirements may not be fulfilled.

The polyglot persistence paradigm is built on mitigating this risk by using different data storage solutions in a given part of the Big Data system based on their suitability. By following this paradigm, we ensure (or at least increase the probability) that the system will be able to adapt to any future changes or requirements. For instance, the paradigm might be applied in build an e-commerce system, in which a key-value store is used to manage customers' shopping carts, whereas a database based on the ACID model is used to handle payments [2]. In [3], the authors even developed an approach to model polyglot persistence systems by using the well-known Entity-Relationship (ER) model. Their approach includes dividing the entire system into segmentation units (e.g. shopping cart, customer registration) and consistency units (e.g. order status, completed orders), which represent subsystems with different data storage needs regarding functional independence for replication purposes and data consistency. Then, the most appropriate database solution is selected for each identified unit, based on which the target data model is determined.

Nevertheless, using multiple data storage solutions within a single system also brings some challenges, since these solutions often differ in their properties, query processing techniques, data models, etc. One of the biggest challenges is how to query data across different databases, which has been studied in the literature [2], [4], [5]. Most research efforts are directed towards developing a global query language, which would be compatible with both SQL and NoSQL systems. Ensuring cross-database consistency and interoperability are also important challenges, which might bring additional costs if not enough attention is given to them [6]. Certainly, the communication between diverse storage architectures within the same environment pose some security challenges as well.

III. MULTI-MODEL DATABASES

The ascend of multi-model database systems brought to attention many of their characteristics, which present their sole purpose of existence. The first problem eliminated with the emergence of multi-model databases is determining the best (most efficient) choice of the data model, depending on the domain. The mentioned problem is solvable by two different data manipulation approaches, which define their corresponding architectural viewpoints: (1) approach based on *polyglot persistence* discussed earlier, and (2) *multi-model databases* [7]. The mentioned approaches differ in the availability of different data models respectively through multiple or a single DBMS. The *multi-model databases* approach initially served as an answer to the „expensiveness" of multiple DBMS usage [8].

Furthermore, taking into account different data models supported by a multi-model DBMS, one does not have to consist strictly of multiple non-relational models. In contrast, most of the currently popular DBMS are multi-model databases, converted from a strictly relational to a combination of relational and non-relational data models. Most known examples of such cases are Oracle, MySQL, Microsoft SQL Server, PostgreSQL and IBM DB2, all of which support document-oriented (JSON) data model, alongside other models (XML, geographic data, graph model) [9]. According to [7], it is the universality and simplicity of the relational databases, which allows the extension to other data models, thereby becoming multi-model.

A substantial characteristic of multi-model databases is the yield of the characteristics of used data models. In other words, the co-existence of relational and non-relational data models within the same DBMS requires the removal of strict relational data model properties, and the introduction of analogous properties used in simple non-relational single-model databases [10]. The most impactful property restricted by the co-existence of different data models are the *ACID* properties (*Atomicity, Consistency, Isolation, Durability*), which guaranteed safe and reliable transaction execution. However, in modern application development, it is often a case that consistency is less important than data availability and database efficiency [11].

The next concept mutual to both relational databases and multi-model databases are indexes. The reason behind their continuous usage is their direct impact on query execution time [12]. On the other hand, most multi-model DBMSs do not implement the JOIN query operation, which is crucial to the relational data model, because of its inefficiency [10].

The standardization and popularity of SQL query language trough the evolution of databases made it nearly impossible to introduce and popularize new query languages. With that being the crucial part of multi-model databases, new ways of querying data had to be introduced because of the variety of co-existing data models. Although SQL is still the most used database query language, new languages like AQL (ArangoDB query language) [13], Apache TinkerPop Gremlin (graph traversal language) [14], [15], Cypher (also a graph traversal language used by Neo4j) [16], and many more emerged and became more popular within the community. The reason behind their market usage growth is them being SQL-like. Regarding the multi-model DBMSs, popular SQL alternatives are such SQL-like languages and SQL extensions [7].

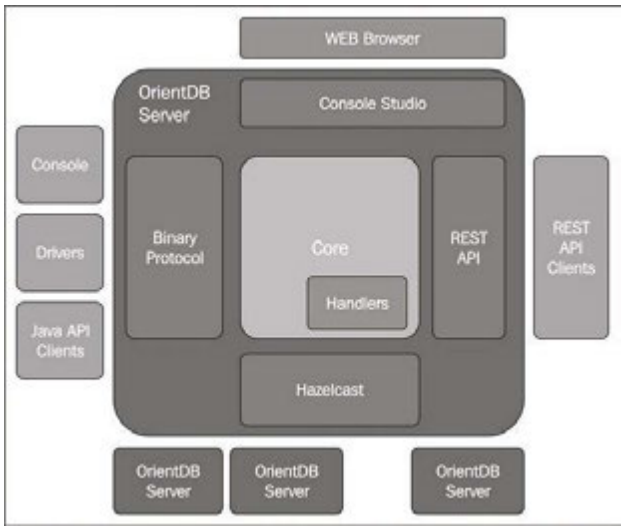


Figure 1. OrientDB system architecture [19].

A. OrientDB

The combination of graph and document-oriented data model's advantages within the same system brings OrientDB on the list of open source multi-model NoSQL databases. Many NoSQL DBMSs find frequent usage as secondary data storage, but OrientDB is claimed to be flexible enough to be used for primary (operational) data storage. Alongside two versions (*Community Edition* and *Enterprise Edition*), OrientDB offers many other user and administration tools, making it convenient for its new users to either start from scratch, or convert an existing database into a multi-model one [17].

Besides the usual multi-model database properties, OrientDB DBMS is enriched with some specific additional ones. Firstly, it is important to mention a specific kind of a *master-slave* server organization used by OrientDB, also known as the *master-less* organization, with the main difference being that each server is able to read and write data. This property is achieved by using so called *replica sets*, which remove the need of former *slave* organization [18].

To avoid JOINs, OrientDB supports relationships. Considering that all of the data stored in a OrientDB database is formed into *records* with unique RIDs (Record IDs), which correspond to documents, BLOBs, vertexes or edges [7], the reason for record usage is the ability to define all coexisting data model concepts [19].

Another property which can be put into comparison with storage organization in other single-model graph or document-oriented databases is the inter-connectedness of data within the system [18]. OrientDB supports two types of data relationships: *Referenced* (physical links similar to pointers) and *Embedded* (strong relationships embedded into the record). Furthermore, each of the two relationship

types split into three separate kinds of relationships. The referenced relationships are:

- *LINKSET* as an unordered set of links,
- *LINKLIST* as an ordered list of links and
- *LINKMAP* as a map of links keyed to the source record,

and embedded relationships are:

- *EMBEDDEDSET* as an unordered set of records,
- *EMBEDDEDLIST* as an ordered list of records and
- *EMBEDDEDMAP* as a map of records keyed to the source record.

All of the mentioned concepts are important, because they improve the performance of queries with high depth level [20].

Furthermore, the architecture of OrientDB is also one of the reasons of uniqueness of the mentioned multi-model DBMS (depicted in Figure 1). OrientDB is a distributed system, meaning it is divided into inter-connected servers which communicate with outer clients, drivers, the console and the web browser. The center processing part of each server is the core, and its handlers (plugins) with the main purpose of functionality enhancement of the core. Another component of the server specific for the OrientDB technology is the *Hazelcast*, responsible for the communication between nodes in the cluster (in the transport layer) [19].

Alongside the already mentioned graph and document-oriented data models, OrientDB engine supports key-value and object-oriented data models, as well as their features. The document-oriented model's main concepts are classes (analogous to the concept of classes in object-oriented programming) and clusters (similar concept as MongoDB collections or relations of the relational data model). In contrast to that, the graph data model embedded in OrientDB consists of data types classified into classes which extend two main classes *Vertex* and *Edge*. Every node (vertex) or edge can be extended with its set of unique properties, making them similar to the concept of document (in the document-oriented data model). Furthermore, the OrientDB graph model differs two edge types: *regular edge* and *lightweight edge*.



Figure 2. Graph representation of the data inside the application database.

Although in the graph model all edges are bidirectional, the difference between those edge types lies in the ability of the regular edge to store documents (node with its properties). The goal of such concept implementation is to allow the users of the database to determine whether they need additional memory to store attributes of a connection. In case they do not, then the usage of lightweight edges is recommended to maintain the efficiency of the database [19].

OrientDB also includes a solid security system based on users and roles. In terms of its flexible multi-model nature, OrientDB allows schema-less, schema-full and schema-mixed implementation. The popularity of OrientDB lies in its SQL layer with extensions for graph traversal, which allows for skilled relational database professionals to quickly switch to OrientDB database [7]. Furthermore, OrientDB DBMS is also highly customizable, making it possible to build custom plugins and add custom commands [21].

Another feature crucial for modern cloud database deployment is that OrientDB is deployable and manageable in most cloud environments [7].

B. *Comparison of OrientDB with other database solutions*

Having mentioned all the key characteristics of OrientDB, it is important to mention its current position on the market as the most popular primarily graph multi-model database [1].

There are a few related works which contain performance benchmarks containing OrientDB, one of them being between two multi-model and two single-model databases, which shows that OrientDB performs better than certain other databases in terms of execution of deeper query levels amongst a relatively small dataset [20]. Another domain specific work concludes that OrientDBs graph model and multi-model characteristics perform better than the MySQL relational model in terms of storing and querying complex multimedia data and are therefore better suited for the tested domain [22].

The lack of adequate multi-model database benchmarks is also already stated and therefore also implemented in form of a unique novel benchmark called *UniBench*. The benchmark has also been tested on implementations of ArangoDB and OrientDB workloads, which aim to show the performance advantages between current multi-model databases [23].

IV. IMPLEMENTATION

To demonstrate the applicability of multi-model databases, a prototype application using a multi-model database was built, which properties will be presented in this section. For the multi-model DBMS, OrientDB was chosen because of the supported data models, which correspond to the chosen polyglot persistence application domain.

To show the full potential of OrientDB, the domain should not be arbitrary. Moreover, the combination of the multiple data models is ideal for implementation of content management systems (CMS), which have the potential to grow and merge with the domain of social media and

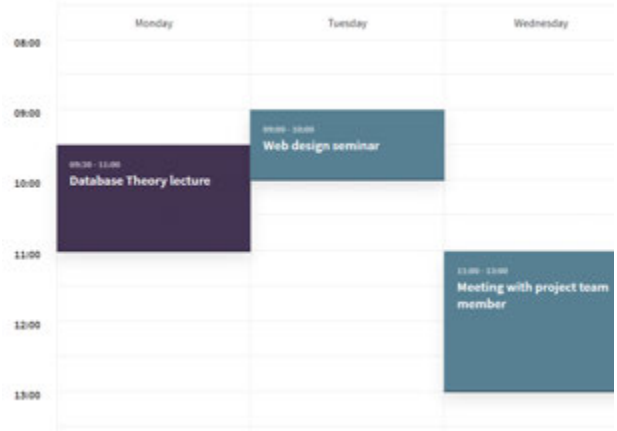


Figure 3. The main sheduling application interface.

recommendation systems. The combination of mentioned systems excellently describes a college scheduling application, which will be implemented. The selected domain's data could certainly be organized within a graph or document single-model system, in which case the performance might be sacrificed [20].

A. College scheduling application

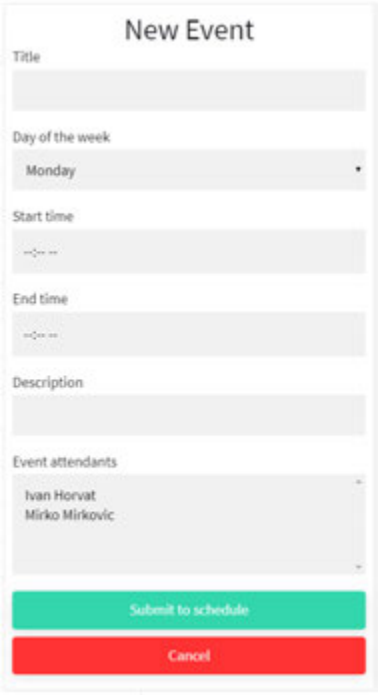
Programming languages which were used for the implementation of the online scheduling application are PHP, HTML and CSS, even though OrientDB provides drivers and APIs for other programming languages, like Java, NodeJS, .NET, Python and many more. For the visual design, a template was used, because the main focus was put on the database implementation [24].

The two domains intended to inter-connect are scheduling and social networking. The advantage of a multi-model database while working with the two mentioned domains is that scheduling is simply manageable with document-oriented data model, while the social networking is manageable with the graph data model. A multi-model database may bring additional benefits compared to the relational database, as it offers a wider set of functionalities for such an application. Namely, it is possible to store information about given time slots in the schedule extended by attached notes (documents) to each slot, but also share the schedule or documents through its graph capabilities.

The simple database schema consists of two classes which extend the vertex class (*Event* and *User*) and three classes which extend the edge class (*EventOwner*, *Friend* and *EventAttendant*). The *User* class has its properties *NameSurname* and *Username*, while the *Event* class consists of properties *Title*, *Day*, *TimeFrom*, *TimeTo* and *Description*.

The final version of the implemented application allows functionalities like:

- logging into accounts,
- viewing of all weekly events (Figure 4),
- viewing of details about a selected event,



(a)

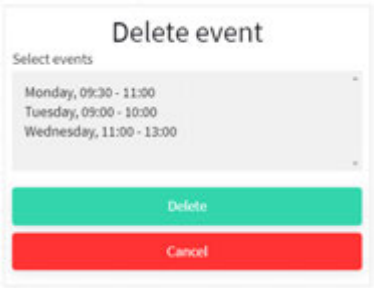


Figure 4.(a) Form for adding new events, (b) form for removing events.

- adding or removing events from the weekly schedule (Figures 3-a and 3-b), and
- adding friends to the event while creating it.

The before mentioned inter-connection of the document-oriented and the graph data model mirrors the implementation in a way that the event data is stored in the document-oriented model and the user-related information in the graph model. The usage of both models can be observed trough SQL projection examples of:

- all currently active events of the specified owner (as edges), ordered by the day and event start time (graph storage projection):

SELECT FROM Event WHERE OUT('EventOwner').Username = 'mmarkovic' ORDER BY Day, TimeFrom

- all event titles and timespans where any event property field contains the specified string (document storage projection; returns JSON format):

SELECT Title, TimeFrom + ' - ' + TimeTo as Timespan FROM Event WHERE ANY() LIKE '%Mirko%'

V. CONCLUSION

The importance of multi-model databases in the context of Big Data storage has grown immensely trough the past decade, the main initiators being the complexity and diversity of data formats, and the trend of combining properties of multiple domains in building modern applications. The "one-size-fits-all" database solutions are no longer suitable for Big Data applications, as existing relational and non-relational database solutions are limited in how much they can be extended and flexible. Even though polyglot persistence systems appear as a viable option, which can combine benefits of both worlds (relational and NoSQL), such systems bring certain challenges in terms of security, querying and consistency. As an alternative, multi-model databases mitigate some of these problems by using different data models within the same database solution. To demonstrate their usability, we presented a prototype college scheduling application, which combines the benefits of the document-oriented and graph data models to enrich the features of the application. As part of our future work, we plan to run a benchmark test on a large social network to compare the performance and further analyze the benefits and drawbacks of OrientDB, Neo4j and a selected relational database.

REFERENCES

[1] solid IT, "DB-Engines Ranking," 2020. [Online]. Available: https://db-engines.com/en/ranking. [Accessed: 14-Feb-2020].

[2] P. P. Khine and Z. Wang, "A review of polyglot persistence in the big data world," *Inf.*, vol. 10, no. 4, 2019.

[3] C. Zdepski, T. A. Bini, and S. N. Matos, "An approach for modeling polyglot persistence," *ICEIS 2018 - Proc. 20th Int. Conf. Enterp. Inf. Syst.*, vol. 1, no. Iceis 2018, pp. 120–126, 2018.

[4] B. Kolev, P. Valduriez, C. Bondiombouy, R. Jimenez-Peris, R. Pau, and J. Pereira, "CloudMdsQL: querying heterogeneous cloud data stores with a common language," *Distrib. PARALLEL DATABASES*, vol. 34, no. 4, pp. 463–503, Dec. 2016.

[5] L. H. Zambom Santana and R. Dos Santos Mello, "A middleware for polyglot persistence of RDF data into NoSQL databases," *Proc. - 2019 IEEE 20th Int. Conf. Inf. Reuse Integr. Data Sci. IRI 2019*, no. July, pp. 237–244, 2019.

[6] L. Wiese, "Polyglot Database Architectures= Polyglot Challenges.," in *LWA*, 2015, pp. 422–426.

[7] J. Lu and I. Holubová, "Multi-model Databases: A new journey to handle the variety of data," *ACM Comput. Surv.*, vol. 52, no. 3, 2019.

[8] Z. H. Liu, J. Lu, D. Gawlick, H. Helskyaho, G. Pogossiants, and Z. Wu, "Multi-model Database Management Systems - A Look Forward," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11470 LNCS, pp. 16–29, 2019.

[9] "DB-Engines Ranking," 2020. [Online]. Available: https://db-engines.com/en/ranking.

[10] N. Jatana, S. Puri, M. Ahuja, I. Kathuria, and D. Gosain, "A Survey and Comparison of Relational and Non-Relational Database."

[11] A. Meier and M. Kaufmann, *SQL & NOSQL Databases: Models, Languages, Consistency Options and Architectures for Big Data Management*. Morgan Kaufmann, 2019.

[12] M. Šestak, K. Rabuzin, and M. Konecki, "Indexing in relational and graph databases as a mean of dealing with increasing amount of business data," *Proc. IAC-MEBM Conf. Vienna 2016*, pp. 181–185, 2016.

[13] "ArangoDB Query Language," 2020. [Online]. Available: https://www.arangodb.com/docs/stable/aql/. [Accessed: 11-Feb-2020].

[14] M. A. Rodriguez, "The gremlin graph traversal machine and language (Invited Talk)," *DBPL 2015 - Proc. 15th Symp. Database Program. Lang.*, pp. 1–10, 2015.

[15] K. Rabuzin, M. Maleković, and M. Šestak, "Gremlin By Example," *Proc. 2016 Int. Conf. Adv. Big Data Anal.*, pp. 144–149, 2016.

[16] N. Francis *et al.*, "Cypher: An evolving query language for property graphs," *Proc. ACM SIGMOD Int. Conf. Manag. Data*, pp. 1433–1445, 2018.

[17] "OrientDB," 2020. [Online]. Available: https://orientdb.com/. [Accessed: 11-Feb-2020].

[18] "OrientDB Official Documentation," 2020. [Online]. Available: http://orientdb.com/docs/3.0.x/. [Accessed: 11-Feb-2020].

[19] C. Tesoriero, *Getting started with OrientDB : a practical guide to learn, deploy, and customize OrientDB*. Packt Publishing, 2013.

[20] F. R. Oliveira and L. Del Val Cura, "Performance evaluation of NoSQL multi-model data stores in polyglot persistence applications," *ACM Int. Conf. Proceeding Ser.*, vol. 11-13-July, pp. 230–235, 2016.

[21] A. Messina, P. Storniolo, and A. Urso, "Keep it simple, fast and scalable: A multi-model NoSQL DBMS as an (eb) XML-over-SOAP service," *Proc. - IEEE 30th Int. Conf. Adv. Inf. Netw. Appl. Work. WAINA 2016*, pp. 220–225, 2016.

[22] C. Küçükkeçeci and A. Yazıcı, "Big Data Model Simulation on a Graph Database for Surveillance in Wireless Multimedia Sensor Networks," *Big Data Res.*, vol. 11, no. September, pp. 33–43, 2018.

[23] C. Zhang, J. Lu, P. Xu, and Y. Chen, "UniBench: A benchmark for multi-model database management systems," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11135 LNCS, pp. 7–23, 2019.

[24] CodyHouse, "Scheduling Template," 2019. [Online]. Available: https://codyhouse.co/gem/schedule-template. [Accessed: 12-Feb-2020].

Hot Topic Detection Using Twitter Streaming Data

Teodor Jagić*, Ljiljana Brkić *
* *University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
teodor.jagic@fer.hr
ljiljana.brkic@fer.hr

***Abstract* - With the increasing popularity and widespread use of social networks, it is becoming increasingly beneficial to analyse the data being shared to identify topics of public interest and specific social phenomena. This paper analyses the possibilities, challenges and difficulties of automatic periodic collection and analysis of data from popular social networks and explains in detail the acquisition and analysis of data from Twitter. The paper explains an implementation of a simple hot topic detection algorithm based on texts acquired from the Twitter's official API. The texts collected are being pre-processed by removing stop-words and stemming the remaining words using Porter's stemming algorithm. Words from pre-processed text are assigned ranks depending on a large-scale analysis using TF-IDF weight and grouped into a hot topic. The algorithm accuracy was evaluated by comparison with Twitter's official hot topic detection algorithm. Appropriate user interface enabling configuring the process of data acquisition, analysis and viewing results in a geographic fashion was implemented.**

Key words - hot topic detection, social networks, text analysis, geographic clustering

I. INTRODUCTION

Since its inception in the early years of the 20th century, social networks have evolved to such an extent that of the 7.7 billion people in the world today, 3.5 billion are actively using social networks. People have 7.6 social media accounts on average, with social network users growing by 328 million from October 2018 to October 2019. This information was taken from the official website of Brandwatch, which describes itself as the world's leading social intelligence company [1]. This data makes clear the fact that the influence of social networks is still growing rapidly, as is the number of users on social networks. An increase in the number of users on social networks also means increase in the amount of data shared on these networks. The information that people post on social networks comes from different aspects of life. They range from banal everyday events to essential crisis-preventing information beneficial for the general public such as information regarding a fire, a flood or an epidemic.

Due to the amount of information constantly being shared on social networks and the potential relevance of this information to society, analysing and archiving the most relevant and popular data from social networks can be very useful for discovering natural disasters as early as possible, understanding social phenomena and detecting and preventing potentially dangerous events.

This paper describes our work on detecting hot topics in tweets we have been collected periodically using Twitter API. An overall architecture of the system is outlined, as well as the data collecting and data analysing phases. The way we evaluated accuracy of our algorithm and results of evaluation phase are presented also.

II. RELATED WORK

Many techniques have been proposed for topic detection, some of them beeing clustering, probabilistic, and frequent pattern mining. One of the common representations of the topics is by a set of keywords. This set is usually a weighted set of words, where weights represent the importance of the word in the topic [2].
The clustering techniques used for topic detection assume that texts (eg. tweets from Twitter) in the same cluster discuss the same topic. Centroid of the cluster is used as a representative where the top n words (in terms of TF-IDF [3] weights) are used as the keywords of this topic. For the purpose of detecting topics, each text is represented using TF-IDF scheme and the number of topics to be discovered is used as the number of clusters (n). This is mainly the idea we followed when developing our hot topic detection algorithm.

Similar to our work is the work of Ishikawa et al [4] in which the authors present a system that detects a hot topic in a local area and over a period. The architecture of their system consists of pre-processing text that eliminates unwanted tweets and text, semantic classifications, clustering that groups the tweets by a similarity function, and finally burst detection, which is a method for detecting a period of time in which events are unusually popular.

Approach using a probabilistic method is presented in [5]. After pre-processing the text, authors perform a probabilistic latent semantic analysis to determine a parameter matrix that provides information about the distribution of topics in a single document, and then performs K-means clustering based on those parameters.

Both algorithms presented in these two articles rely on a time period to receive some information about the importance of tweets for a topic and have proven too complex to fully implement, but the part of the algorithms related to word processing, which is a standard procedure in text analysis, was helpful when implementing our own hot topic detection algorithm.

III. ACQUIRING DATA

There are several techniques of acquiring data from the Internet, the most popular being: crawling, scraping and using third-party APIs. Crawlers follow links from HTML pages and download their content. They are extensively used to gather data for search engines. Scrapers act in a similar fashion but they are more convenient for extracting structured data from a web page. This data can then be processed and stored into a data repository like eg. database. However, web scraping is not allowed by some social networks and would have to be adjusted to each website that we would want to extract data from [6]. Using the official social network API is more favourable in many ways, one of them being that it is directly supported by the social network thus eliminating any worry of accessing prohibited data. The other being that we are using the intended way of acquiring data from the social network as imagined by the social network developers. Social network APIs are usually built using the commonly used REST (Representational State Transfer) architecture making them reliable and straightforward to use. In the rest of this section we explain why we base our research on Twitter data and what difficulties we encountered when trying to retrieve data from Facebook and some other social networks.

A. Facebook

One of the most widely used social networks is Facebook, boasting an impressive number of 1.62 billion active daily users as of 3rd quarter 2019 [7]. Facebook also has an official API that would allow for easy acquisition of data, but that comes with difficulties. Due to the General Data Protection Regulation (GDPR), implemented in the EU 2018, after the Facebook and Cambridge Analytica scandal [8], access to data on Facebook has become significantly impeded. The Facebook Cambridge Analytica data scandal was a major political scandal in early 2018 when it was revealed that Cambridge Analytica had collected the personal information of millions of people's profiles without their consent and used it for political advertising purposes. It has been described as a turning point in public understanding of personal information and caused a massive fall in Facebook stock prices [9]. Today, Facebook has many restrictions regarding data access, and explicitly asks users to give permission to access private data to any application that uses the Facebook Graph API. This change has made it difficult to collect and later analyse data for applications that need a large amount of anonymous user data.

The GDPR made it so that Social networks must pay close attention to the privacy of their users when choosing how much of their data they make accessible to the public. Even though the GDPR is mainly focused on the protection of user data in the EU, many social media companies revamped their policies concerning public user data availability in general. Facebook is no exception to that rule. Following the GDPR, Facebook implemented stricter rules and regulations to the way users can access its official graph API making it more difficult to acquire vast amounts of user data necessary for getting useful information about social developments and activities.

B. Twitter

Twitter is another social network that can serve as a source of large amount of data. What makes Twitter a good data source candidate is it's official API [10], huge amount of tweets (around 9000 tweets per second [11]), and most of all its policy expressed in the following sentence "Twitter's purpose is to serve the public conversation" [12]. Twitter believes that the public information posted on their network is truly public. Everyone can access, comment on and share this information if they adhere to Twitter's terms of use. This attitude is favourable for an application that seeks to collect and analyse public data from social networks, and therefore, Twitter was the source of data our hot topic detection was focused on. Although the length of tweet may not exceed 280 characters, Twitter has several advantages. For example, people use Twitter not only to express their opinions on any topic but also as a social media to disseminate useful information and breaking news. In addition, the extensive user base of over 300 million [13] is an advantage of using Twitter as a research tool. For these reasons, there are many papers where data, gathered from Twitter, are used for social topics detection [14][2][15] or studying human interactions and communications [16][17].

The Twitter API enables collecting tweets, Twitter trends and Twitter user information such as the location the tweet was posted from, the language of the tweet, etc. The API implements a REST architecture enabling interaction via a simple HTTP request. The following is the example of the API's response (with minimum information for the sake of brevity) in JSON format when requesting latest tweets in an area.

```
{
   "created_at":       "Thu may 10 15:24:15 …",
   "id_str":           "85000624512695744",
   "text":             "Here goes the Twitt ...",
   "user":             { ... },
   "place":            { ... },
   "entities":         { ... },
   "extended_entities":       { ... },
}
```

C. Other social networks

Although there are many more social networks with a large number of users, such as WeChat which is a Chinese social network with 1.15 billion active users on a daily basis as of 3rd quarter 2019 [18], they are not used as data sources in our research due to various factors that make it difficult to analyze or collect data such as data download bans, restrictions on the amount of data that can be collected, foreign language other than English used in texts, use of non-Latin script, high user concentration, but only in a small area, lack of an official API to access the data, and the like.

IV. SYSTEM ARCHITECTURE

Figure 1 represents the overall architecture of the software system developed. Firstly, tweets are collected from the Twitter API using the collection algorithm and stored in the database.

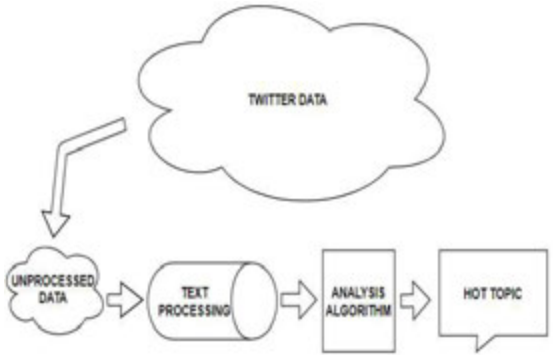


Figure 1 System architecture

The data collected is then parsed, processed and analysed to obtain hot topics. Details regarding collecting and analysing tweets are described in the next two sections.

To evaluate accuracy of the developed hot topics detection algorithm we compare Twitter's trends with our generated hot topics. Trends are hot topics determined by Twitter and can be obtained from Twitter API which accepts the geographic location of an area as a parameter and returns relevant, newest or most popular Twitter trends from that geographic location. The geographic location required to acquire trends must be supplied in the format of a "where on earth id" (WOEID). Where on earth id is a number representing any feature on Earth (a city or a state for example) and it was originally defined by GeoPlanet, a computer platform for coordinating world-wide geographic information.

V. COLLECTING THE DATA

Tweets are collected according to the collection algorithm (Algorithm 1) and stored into a relational database with schema displayed on Figure 2. The process begins by collecting Twitter trends (lines 4 to 6) using the API provided with parameters of WOEID type previously determined and stored in our database (*collection_region* table).

```
Algorithm 1: Collection algorithm
1   Input:  collectionRegions;
2   Output: rawTweetsAndTrends;
    Begin
3   for (each active collection){
4      for(each associated geografic region){
5         get trends from the Twitter API;
6         save trends to database;
7      }
8      for(each trend received){
9         request N tweets from the Twitter API;
10        for(each tweet received){
11          if( tweet showedUp in prev API calls)
12             continue;
13          elif(tweet lang is not english)
14             continue;
15          else{
16             if(the geolocation params exist)
17               extract geoloc using jackson;
18             take all other tweet data;
19             save the tweet into the database;
20           }
21        }
22     }
23  }
24  End
```

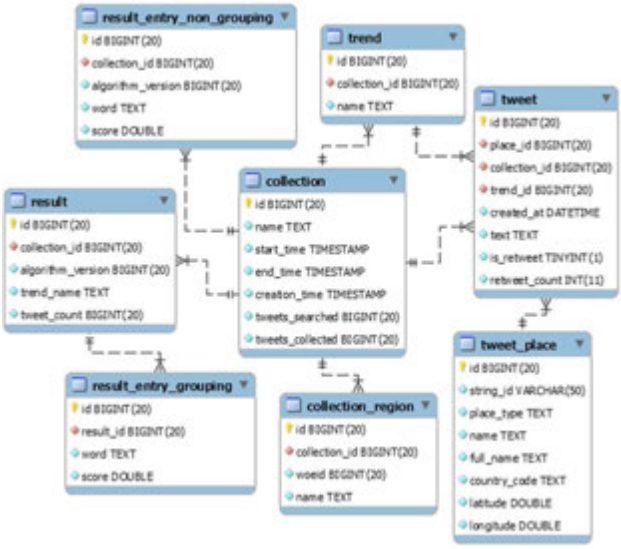


Figure 2 Relational database schema

After acquiring trends (*trend* table), they are used to get tweets associated with those trends (lines 8 to 17). Periodically (set up with system parameter) we request a certain number (N in line 9 of the Algorithm 1) of tweets from the Twitter's API. Twitter's API always returns the requested number of tweets, regardless of whether sufficient number of new tweets connected to the requested words appeared in the time between 2 calls or not. This can lead to many duplicate tweets being acquired on topics that don't have many people tweeting. We ignored tweets that had already appeared in one of the previous API calls and were therefore stored in the database (lines12 and 12). The more new tweets are being stored on the topic we are acquiring data on, the more "efficient" our collection will be, meaning we are going to get fewer duplicate tweets and more information overall in less amount of time.

For the sake of simplicity, we have limited our research to English tweets written in Latin alphabet. Non-English tweets are ignored (lines 13 and 142) since text analysis would need to be adjusted for each language the system would support. For Tweets with known geolocation attribute (lines 16 and 17) we determine geographic details, if possible, and store tweet data into the database (*tweet* and *tweet_place* tables). This way we create a large database of tweets associated with Twitter trends that will subsequently be used in evaluation phase. To be able to compose different datasets tied with different locations or/and time periods we have brought together tweets into so called collections (*collection* table).

The database also serves as a repository for results of analysis that will be produced subsequently (*result*, *result_entry_grouping* and *result_entry_non_grouping* tables).

The part of the user interface, displaying the list of collections with all accompanying data, is shown on Figure 3.

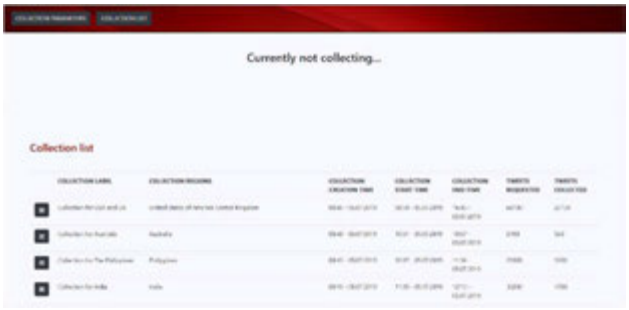


Figure 3. Status of collections

VI. ANALYSING THE DATA

After acquiring a sufficiently large amount of data from the Twitter API, the system can analyse it to detect hot topics (Algorithm 2). A various techniques of detecting hot topics from a large amount of text are reported in the literature, many of them including some sort of clustering algorithm or artificial intelligence. The system we developed chooses a simpler, more straightforward way – use of term frequency-inverse document frequency (*TF-IDF*) numerical weight [3] to evaluate importance of a word to a document in a collection. The TF-IDF weight increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. The formula (1)-(3) given below explain the basic principle of calculating TF-IDF weight.

$$TFIDF\ for\ term\ i\ in\ document\ j = TF(i,j) * IDF(i) \quad (1)$$

where

$TF = Term\ Frequency$

$IDF = Inverse\ Document\ Frequency$

$$TF(i,j) = \frac{Term\ i\ frequency\ in\ document\ j}{Number\ of\ words\ in\ documents} \quad (2)$$

$$IDF(i) = \log_2 \frac{Number\ of\ documents}{Number\ of\ documents\ containing\ term\ i} \quad (3)$$

Tweets are analysed in a context of the corpus to which they belong (lines 4 and 5). As will be explained in the next section, corpus is sometimes made up of tweets having the same Twitter trend, while sometimes of all tweets in the same collection (*collection* table). Before determining the TF-IDF weight of each word in a tweet, the tweet text is parsed and non-alphanumeric characters are removed as well as stop words (lines 6, 8 and 9). Stop words removal is a standard procedure in text processing as they regularly don't contain any useful information. Every word is also transformed into its normalised (root) form using Porter's stemming algorithm [19] (lines 10 and 11). Stemming is the process of reducing inflectional (or sometimes derived) forms to their word stem, base or root form. Porter's stemming algorithm transforms a word into its root form allowing the system to handle a word like "cat" and a word like "cats" as the same word. This is desirable because if for example tweets contain a lot of information on cats, the algorithm shouldn't differentiate between the different forms of the word "cat" considering they all refer to the same animal.

```
    Algorithm 2: Analysis algorithm
1   Input: rawTweetsAndTrends;
2   Output: analysisResults;
3   Begin
4   For (each unanalyzed corpus) {
5     For (each tweet from the corpus) {
6       remove non alphanumeric characters
        excluding: hashtags and underscores;
7       split the tweet text into words;
8       remove words of length 1;
9       remove stop words;
10      For (each tweet word) {
11        calc normalized form of the word;
12        calc TF-IDF for normalized form;
13        save the word with calc measures;
14      }
15    }
16    order analyzed words by TF-IDF;
17    make groups of 10 words;
18    eliminate duplicates
19    save the hot topics as the result;
20  }
21  End
```

The lower the TF-IDF of a word, the more value it has in determining a hot topic. Words are then grouped into groups of 10 words (line 17), making a hot topic. It is worth noting that the same words in different tweets are regarded as different when detecting a hot topic, so it could happen that out of the 10 words making a hot topic multiple words are the same. For example, out of the 10 words, 6 could be the word "cat" because "cat" appeared in many different tweets, many times. Duplicate words are removed (line 18) resulting in less than 10 words for some hot topics and results are stored into the database (line 19).

VII. RESULTS AND DISCUSSION

In the system developed, a corpus can be formed in one of the following two ways: (i) of all tweets that have the same Twitter trend (ii) of all tweets belonging to the same collection. The first type of corpus is possible to form thanks to the link of each tweet with its trend stored in database during the data collection phase. Determining hot topics for tweets that belong to the same, known beforehand, Twitter trend serves to evaluate the usability of the procedure and algorithm in the early stages of its development. If the comparison of our hot topics with Twitter's trends showed that there were absolutely no similarities between them then, the whole approach had to be changed. But fortunately, that did not happen. The second type of corpus allows the analysis of tweets belonging to different Twitter trend. Because of the way TF-IDF works, giving more importance to words that appear more frequently and, in more tweets, this type of analysis results with the most prevalent hot topic in the collection.

To evaluate the accuracy of our algorithm, we designed and calculated a measure to detect accuracy for the twitter trend from the corpus. The measure takes values from the interval [0-100] and is calculated as follows. If a word from the list of words that makes a hot topic of a group of tweets completely matches the word from the trend determined by the Twitter API, it contributes with 100 points to the rating. If the word is only partially

matching a word from a trend (eg. the word being singular and the trend being plural), it contributes with 90 points, and if the word is not in the trend at all it contributes with 0 points. Points are then divided by the number of words in the word list that make the hot topic and scaled to an interval of 0-100 giving us a percent value of how successful the detection was. We are aware that this measure is far from perfect but non the less gives us a general idea of how well a topic is detected.

We were interested whether the word normalization (stemming with Porter's algorithm) contributes to a more accurate determination of the hot topics or not, so we performed the procedure with and without word stemming. We were surprised with the results because the answer to the previous question is - sometimes.

Figure 4 shows the results of evaluation using measure described above. The red graph presents results of the algorithm that uses stemming, while the results of the algorithm that does not use stemming are shown with blue graph. On the abscissa of the graph is the number of topics having the same value of used measure which is on the ordinate. The larger the area under the graph to the left of the graph, the more topics were detected with a high accuracy and the more successful the detection. This forthcoming example shows that out of 50 analysed groups of tweets, the algorithm can detect 23 hot topics when using Porter's stemming and 47 when not using it. This could lead us to believe that Porter's stemming algorithm actually obstructs our hot topic detection but it is also visible on the graph that the hot topics detected by using Porter's stemming are more similar to Twitter's trends (displayed on the x axis) which we are regarding as the correct. Stemming helps in detecting hot topics similar to Twitter's trends and makes them more accurate, but also makes detection of topics that are detected with less accuracy even worse.

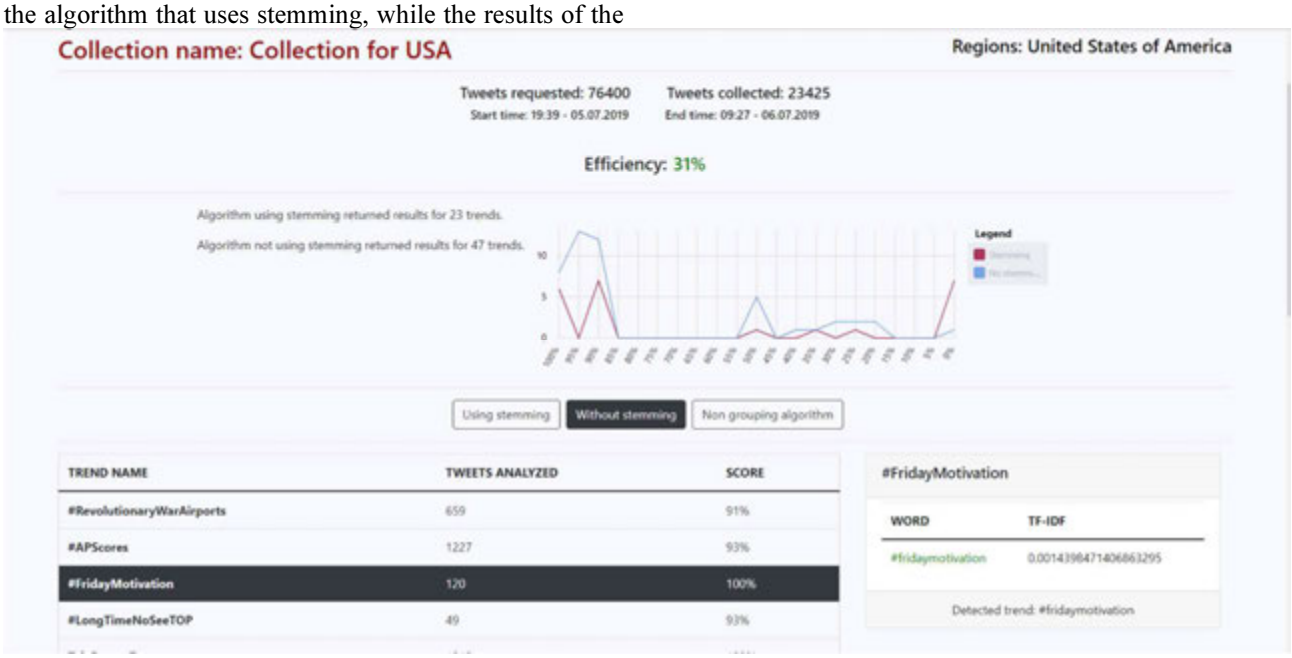


Figure 4 Results and evaluation

Figure 5 shows the heat map displaying the locations from which the tweets in the selected group were tweeted. This allows us to see easily from which areas of a particular region, most of the tweets on a topic come from. The heatmap is an asset when wanting to see an area at a glance in which a potentially important topic, for example information on a influeza, fire or hurricane, is being shared the most.

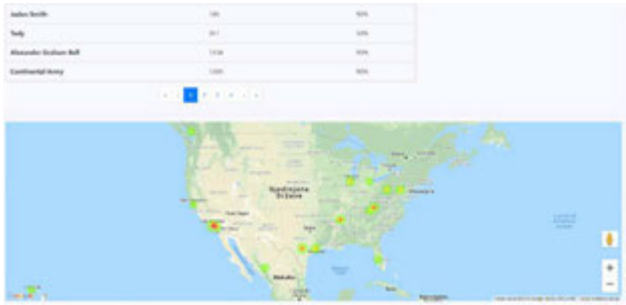


Figure 5. Tweet locations heat map

Figure 6 shows an example of analysis of the results for a non-grouping type of corpus (which pays no attention to Twitter's trends) and gives us the hot topic about which we collected the most tweets or the most prevalent hot topic.



Figure 6 Results of the non-grouping algorithm

CONCLUSION

This paper explores the possibilities of colecting and analysing data from social networks and describes an implementation of hot topic detection for tweets from Twitter. Analysing data from social networks has great potential for improving the detection of patterns and social phenomena varing from banal everyday occurrences to important information about accidents, epidemics and the like. Retrieving data from social networks is more difficult today than a few years ago because of the many rules introduced to improve the protection of social network users' data. Some social networks disable the download of information about their users, and some place major restrictions on the amount of data that can be downloaded. The implemented algorithm for determining topics, detects topics satisfactory good in the selected area by using TF-IDF weights of words in tweets. It can be improved with more sophisticated mehods such as clustering and clasification.

REFERENCES

[1] "Customer Success | Brandwatch." [Online]. Available: https://www.brandwatch.com/customer-success/. [Accessed: 24-Jan-2020].

[2] R. Ibrahim, A. Elbagoury, M. S. Kamel, and F. Karray, "Tools and approaches for topic detection from Twitter streams: survey," *Knowl. Inf. Syst.*, vol. 54, no. 3, pp. 511–539, 2018.

[3] "Tf-idf :: A Single-Page Tutorial - Information Retrieval and Text Mining." [Online]. Available: http://www.tfidf.com/. [Accessed: 31-Jan-2020].

[4] S. Ishikawa, Y. Arakawa, S. Tagashira, and A. Fukuda, "Hot topic detection in local areas using Twitter and Wikipedia," *ARCS Work. ARCS 2012*, pp. 1–5, 2012.

[5] Y. Sun, H. Ma, M. Jia, and W. Peiqing, "An efficient microblog hot topic detection algorithm based on two stage clustering," *IFIP Adv. Inf. Commun. Technol.*, vol. 432, pp. 90–95, 2014.

[6] "Is Web Scraping Illegal? Depends on What the Meaning of the Word Is | Imperva." [Online]. Available: https://www.imperva.com/blog/is-web-scraping-illegal/. [Accessed: 29-Jan-2020].

[7] "• Facebook: global daily active users 2019 | Statista." [Online]. Available: https://www.statista.com/statistics/346167/facebook-global-dau/. [Accessed: 29-Jan-2020].

[8] "Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens - The New York Times." [Online]. Available: https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html. [Accessed: 29-Jan-2020].

[9] "Facebook's scandals in 2018: effect on stock." [Online]. Available: https://www.cnbc.com/2018/11/20/facebooks-scandals-in-2018-effect-on-stock.html. [Accessed: 29-Jan-2020].

[10] "API reference index — Twitter Developers." [Online]. Available: https://developer.twitter.com/en/docs/api-reference-index. [Accessed: 30-Jan-2020].

[11] "1 Second - Internet Live Stats." [Online]. Available: https://www.internetlivestats.com/one-second/#tweets-band. [Accessed: 30-Jan-2020].

[12] "The Twitter Rules." [Online]. Available: https://help.twitter.com/en/rules-and-policies/twitter-rules. [Accessed: 30-Jan-2020].

[13] "• Twitter by the Numbers (2020): Stats, Demographics & Fun Facts." [Online]. Available: https://www.omnicoreagency.com/twitter-statistics/. [Accessed: 24-Jan-2020].

[14] H. G. Kim, S. Lee, and S. Kyeong, "Discovering hot topics using Twitter streaming data social topic detection and geographic clustering," *Proc. 2013 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2013*, no. August 2013, pp. 1215–1220, 2013.

[15] W. Xie, F. Zhu, J. Jiang, E. P. Lim, and K. Wang, "TopicSketch: Real-time bursty topic detection from twitter," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 2216–2229, 2016.

[16] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi, "Sentiment analysis of Twitter data for predicting stock market movements," *Int. Conf. Signal Process. Commun. Power Embed. Syst. SCOPES 2016 - Proc.*, pp. 1345–1350, 2017.

[17] V. A. and S. S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," *Int. J. Comput. Appl.*, vol. 139, no. 11, pp. 5–15, 2016.

[18] "• WeChat: number of users 2019 | Statista." [Online]. Available: https://www.statista.com/statistics/255778/number-of-active-wechat-messenger-accounts/. [Accessed: 30-Jan-2020].

[19] "The Porter stemming algorithm." [Online]. Available: http://snowball.tartarus.org/algorithms/porter/stemmer.html. [Accessed: 31-Jan-2020].

Processing and Visualization of Collected Data Based on Open-Source Tools and Principles

Sanja Grbac Babić*, Kristijan Cetina *
* Istarsko veleučilište - Università Istriana di scienze applicate, Pula, Croatia
sgrbac@iv.hr, kristijan.cetina@gmail.com

***Summary* - This paper demonstrates the use of open-source tools and libraries to process collected data and visualize analysis. These data were collected using a built-in assembly with a GPS receiver and an Arduino microcomputer, and in a second example with the help of a mobile device and a free Phyphox app. Using the Python programming language, the Matplotlib, NumPy and Pandas libraries analysis and visualization the collected data are made within the Jupyter notebook.**

In this paper, we demonstrate how it is possible to do a simple data analysis and apply it for learning purposes, but also in further research. Different graphical data representations are used in the paper to demonstrate the importance of the correct representations selected for different types of data. In an age when the amount of data collected is growing exponentially, it is important to be able to extract useful information from them, to make the data analysis results as accurate as possible.

The complete software part of the work was created using an open-source tool and all the details and results of this work are freely available on GitHub[1].

Keywords - Python, Matplotlib, open-source, data analysis

I. INTRODUCTION

Today, data collection is increasingly important. As technological development made it possible to collect and store data, it became necessary and very useful to collect, but also to process and analyse data. In the ever-increasing number of available data, the challenge is to analyse that data and visualize it as properly as possible, so that the data can be used further. In this way, future trends can be predicted by data processing, but also lots information can be deduced from proper analysis. To obtain as accurate information as possible appropriate tools should be used.

This paper describes a possible way to build a data analysis using Jupyter notebook with Python and help of the NumPy and Matplotlib libraries.

The paper is organized in six sections. In section II. we discuss technology stack. In section III we describe data acquisition methods, giving the results of data processing and visualization in section IV. In section V. is presented a case study and intentions for future work. Finally, the conclusion is found in section VI.

II. TECHNOLOGY STACK

A. Jupyter notebook

The Jupyter Notebook is an open-source web application that can be used to create and share documents that contain live code, equations, visualizations, and text. Project Jupyter is a non-profit, open-source project, born out of the IPython Project in 2014 as it evolved to support interactive data science and scientific computing [1]. Beside Python, Jupyter Notebook supports Julia, R and many other languages [2].

One of the benefits of using interactive notebook over some other IDE (Integrated Development Environment) code editor for Python is ease of use and ability to quickly iterate over certain cells before continuing with analysis without the need to run the whole script every time. This greatly facilitates the development of the script for beginners and less experienced coders.

Other benefit is easy sharing and reproducibility of the work. The Jupyter Notebook includes export to PDF as final document or LaTeX for future work and publishing of the results as part of a research project. It is even possible to export as HTML.

Several papers have been published with supporting notebooks to reproduce the analysis, or the creation of key plots. The detection of gravitational waves by the LIGO experiment (LIGO Scientific Collaboration and Virgo Collaboration et al., 2016), is one such: the researchers posted a notebook on their website illustrating in detail how to filter and process the data to reveal the signature of a distant black hole merger (LIGO collaboration) [3].

Jupyter Notebook contains series of cells which supports adding rich content to it. There are four types of cells: Code, Markdown, Heading and Raw NBConvert. As the Heading cell is no longer supported, using Markdown cell is preferred way to do it.

The Raw NBConvert cell type is only intended for special use cases when using the *nbconvert* command line tool. Basically, it allows control over the formatting in a very specific way when converting from a Notebook to another format.

[1]https://github.com/KristijanCetina/BachelorThesis/tree/master/dataAnalysis

The most used cell types are Code in which is used to run a source code and Markdown for headings, user comments and explanations.

B. NumPy library

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities [4]

In addition to significantly facilitating the numerical and statistical analysis of the data set, NumPy, due to its structure and representation of the data field, enables a significant improvement in data processing performance. To demonstrate the difference in performance between a pure Python and a NumPy library, it was done a simple experiment consisting of summing elements in a 10^6 element size array and measuring the time it takes to complete a task. Our test shows result of 6.07 ms ± 117 µs per loop for pure Python and 749 µs ± 6.74 µs per loop for NumPy[2]. There is a significant difference in the time it takes to complete a task, and industry experience shows an even greater difference in more complex tasks [5].

C. Plotting and version control

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms [6]. All plots in this paper are made utilising Matplotlib, whose documentation with hundreds of examples it's possible to use for this type of application.

For version control is used Git. Git is a free and open-source distributed version control system designed to handle everything from small to very large projects with speed and efficiency [7].

III. DATA ACQUISITION

The data used in this paper were obtained using a custom-made Arduino based hardware, made by authors in their previous work[3] and the free phyphox app [7]. This paper does not provide details on the hardware used since the data analysis and visualization is independent of data source and acquisition method. Instead, is given just a brief overview of used hardware to show a various possibility.

Used data logger was created as part of author's final thesis. The device is based on Arduino UNO microcontroller with added TMP36[4] temperature sensor on top of Adafruit Ultimate GPS Logger Shield[5]. The shield has an integrated SD card slot which is used to write a data to a card.

An example of the record from Arduino microcontroller is:

```
$GPRMC,052730.000,A,4458.8300,N,01356.1
724,E,12.41,283.84,310719,,,A*54,26.38
```

Where is:

- $GPRMC - Sentence Identifier,
- 052730.000 – Time stamp (UTC),
- A - validity (A-ok, V-invalid),
- 4458.8300 – Current Latitude,
- N – North / South,
- 01356.1724 – Current Longitude,
- E – East / West,
- 12.41 – Speed in knots,
- 283.84 – True course,
- 310719 – Date stamp,
- A*54 – checksum,
- 26.38 – temperature in °C.

Other data source was smartphone with installed phyphox application which allows to use the sensors of the smartphone for various experiments. In this case, accelerometer data were recorded during a travel of a lift in apartment building.

All the data were recorded in comma separated values (.csv) file and transferred to a computer for further analysis.

Phyphox accelerometer data are given as:

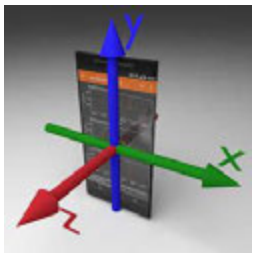


Figure 1. Coordinate system used in phyphox application

- Time (s);
- Linear Acceleration x (m/s^2);
- Linear Acceleration y (m/s^2);
- Linear Acceleration z (m/s^2);
- Absolute acceleration (m/s^2).

Figure 1. Coordinate system used in phyphox application

Coordinate system used in application is shown in figure 1.

[2] Results may vary depending on used system

[3] https://github.com/KristijanCetina/BachelorThesis

[4] https://github.com/KristijanCetina/BachelorThesis/blob/master/resources/TMP35_36_37.pdf

[5] https://learn.adafruit.com/adafruit-ultimate-gps-logger-shield?view=all

IV. Data Processing and Visualisation

As a first example of plotting data, it has been plotted speed of a car caring datalogger as function of time. To be able to use libraries in code first it was needed to import them as follows:

```
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib as mpl
import numpy as np
```

It is common naming convention to name imported libraries as shown and use aliases in code. Next step is import data used for analysis. As there is no header in the used.csv file, we must name the attributes (columns):

```
filename='GPSLOG10.CSV'
data=pd.read_csv(filename,
header=None, delimiter=',' ,names=
['Sentence', 'Time', 'Validity',
'Latitude', 'NS','Longitude', 'EW',
'Speed', 'Direction', 'Date', 'NA1',
'NA2', 'Checksum', 'Temperature'])
```

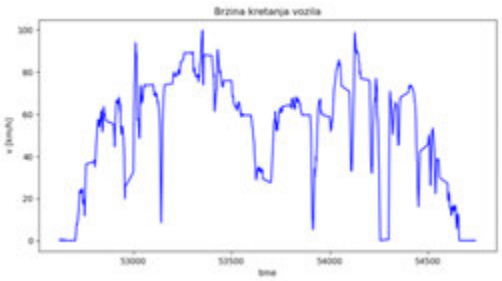


Figure 2: Plot of vehicle speed

and finally, we can plot our data as shown in figure 2 with:

```
plt.plot(data['Time'],data['Speed']*1.
852, 'b-',label='Speed')
plt.legend(loc='upper left')
plt.xlabel ('time')
plt.ylabel ('v [km/h]')
plt.title('Vehicle speed')
```

Should be noted that data are in knots (nautical miles per hour), but they are plotted in km/h, so it was needed to multiply raw data with 1.852 to convert units. Statistical analysis can be done, though for more advanced ones additional libraries may be needed. One of the simpler data is the minimum and maximum temperature which we can get with:

```
print ('Min temp: ', np.min (data
['Temperature']), '°C')
print ('Max temp: ', np.max (data
['Temperature']), '°C').
```

As a response we get „Min temp: 25.73 °C; Max temp: 27.99 °C “for a given dataset.

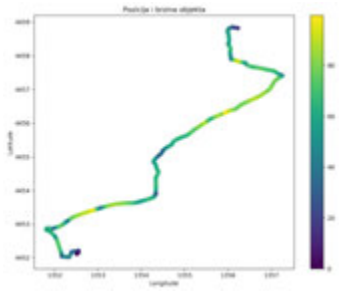


Figure 3: Position and speed of vehicle

Using colormaps it’s possible to encode additional information on a plot. As an example, on figure 3 is shown position of vehicle on x-y axes and its speed in colour with scale aside.

One should be mindful when choosing colormap what data will be shown as explained in [9] and [10].

For the second example we plotted accelerometer data acquired using before mentioned phyphox app on

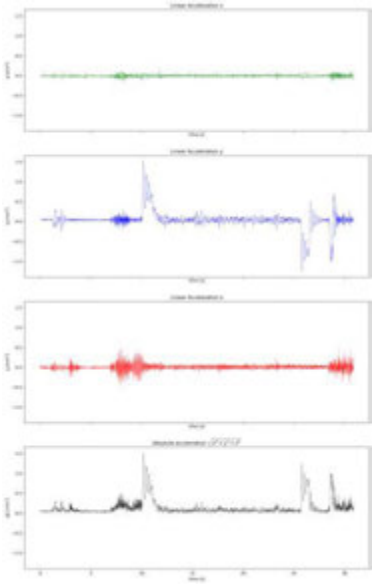


Figure 4: Accelerometer data

smartphone. The smartphone was firmly attached to the lift wall taking care to align the axis of motion with the axis of the device. Result is shown in figure 4.

The key point of this plot is to show the influence of the individual components of the axis movement on the total force exerted on the body. The first three plots show linear movement on x, y and z axis respectively, while fourth plot show linear combination of them. It's easy to notice a very small amplitude in movements in x-axis and majority of total force is from y-axis. That is to be expected as this axis is aligned with lifts travel path.

Subplots of linear movements can be plotted independent of each other, but in this case is important to have them plotted on same scale to avoid misunderstanding due to different scales. In setup we used the following code:

```
fig,axs=plt.subplots(4,sharex=True,
sharey=True)
```

By default, parameter `sharey` is set to `False` which scales each subplot to full range of its motion component which gives a wrong misconception about amplitude of motion. By setting it to `True` all subplots are using same range which gives a realistic representation of actual forces.

V. CASE STUDY AND FUTURE WORK

In everyday life, one often requires monitoring of events, that is, collecting data that can help us find useful information and predict future situations and needs, and is especially needed in technical systems. As today's technology has enabled real-time remote monitoring, the authors of the paper would put these principles into practice in their future work.

In his work [11], Grgin implemented remote monitoring of the battery on board using a microcontroller. The lead-acid battery on board the vessel is used to start the main propulsion and to supply various electrical and electronic devices, which may be damaged during the non-use period due to insufficient maintenance. Lead batteries have a relatively high degree of self-discharge, especially in colder winter conditions. Replacing a rechargeable battery results in a financial and environmental aspect.

Since they are already polluting the environment during their production, it would be logical to think that such devices provide adequate maintenance to prevent unnecessary waste of funds and to prevent additional environmental pollution. The expected life of lead-acid batteries commonly used on board vessels varies greatly depending on the operating conditions and the level of maintenance. It is certainly not advisable for lead batteries to remain relatively low for long periods of time as their battery life is rapidly reduced, so it is advisable to check the battery status regularly.

That is why it's advisable to monitor in real time remotely. In addition to monitoring, it is also desirable to collect and subsequently process the collected data.

The purpose of the work [11], was to create, with the help of a microcontroller, a device for remote monitoring of the battery voltage and any charging current so that the battery could be adequately maintained.

Continuing on the work [11], it's planned to monitor and analyse the data over time, and based on the obtained data, it would be possible to predict the maintenance cycle and ultimately extend the life of the battery itself.

In the future, we will explore possibilities of used tools and techniques for visualizing more complex data and structures as well as their statistical properties.

VI. CONCLUSION

In this paper, we explored the possibilities of data analysis and visualization using open-source tools. With the main focus on ease of use and sharing and reproducibility of analysis. With the availability of open-source tools and resource this is possible to achieve and use by novice for similar work as well as an expert for more complex analysis.

REFERENCE

[1] Project Jupyter. [Online]. Available at: https://jupyter.org/, [Accessed: 8.1.2020.].

[2] List of supported kernels. [Online]. Available at: https://github.com/jupyter/jupyter/wiki/Jupyter-kernels, [Accessed: 9.1.2020.]

[3] Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian E. Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley et al. "Jupyter Notebooks-a publishing format for reproducible computational workflows." In ELPUB, pp. 87-90. 2016.

[4] NumPy library homepage. [Online]. Available at: https://numpy.org, [Accessed: 8.1.2020.]

[5] Van Der Walt, Stefan, S. Chris Colbert, and Gael Varoquaux. "The NumPy array: a structure for efficient numerical computation." Computing in Science & Engineering 13, no. 2 (2011): 22.

[6] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

[7] Git homepage. [Online]. Available at: htto://git-scm.com, [Accessed: 8.1.2020.]

[8] Phyphox app project homepage. [Online]. Available at: https://phyphox.org/, [Accessed:8.1.2020]

[9] Moreland, Kenneth. "Why we use bad color maps and what you can do about it." *Electronic Imaging* 2016, no. 16 (2016): 1-6.

[10] A Better Default Colormap for Matplotlib | SciPy 2015, [Online], Available at: https://www.youtube.com/watch?v=xAoljeRJ3lU, [Accessed: 10.1.2020.]

[11] Grgin, S. „Daljinsko praćenje stanja baterije na plovilu pomoću ESP8266 mikrokontrolera", Završni rad, Istarsko veleučilište, Pula, 2019.

Web application dashboards as a tool for data visualization and enrichment

Maria Holjevac*, Tomislav Jakopec**
* Faculty of Humanities and Social Sciences/Department of Information Sciences, Osijek, Croatia
mholjevac@ffos.hr
** Faculty of Humanities and Social Sciences/Department of Information Sciences, Osijek, Croatia
tjakopec@ffos.hr

Abstract - **Working place of an individual that manages data analysis and data processing can involve one person or a whole team. The tools that can make their job easier, reduce the work hours and ultimately reduce the number of people needed in the team are dashboards of web applications. With the help of dashboards one can manage, analyze and visually monitor their key metrics related to their business or to an individual. There is no universal dashboard to suit every form of business. Equally, there is no universal interface that can be declared as the only suitable solution for all businesses, therefore different dashboards are being developed in different interfaces. Individuals or teams are the users of dashboards. There is a number of dashboard templates for website interfaces that can help developers create their own dashboards for accessible data more easily and quickly. The aim of the paper is to analyze and evaluate various dashboard templates according to the defined criteria in order to demonstrate which template is the most appropriate for a certain domain and field of work. It is expected that the results will help developers with the choice of a dashboard template for their new projects. This paper gives basic guidelines regarding responsive web design and some of the frameworks that can be used to build a responsive web design page. It also briefly describes the CSS Grid. Furthermore, this paper analyzes five different dashboard templates created within four different frameworks and CSS Grid. They were analyzed with five given criteria: the 5 second rule, the inverted pyramid, data visualization, less is more and interactive elements. The results would emphasize the templates created within Bootstrap and Semantic UI because they were not evaluated by any criterion with a score less than 4 out of 5.**

Keywords – dashboard, data, visualization, web application

I. INTRODUCTION

The large amount of data has led to the need for an efficient way of monitoring. For the most part, business-related information is written within tables that have often been opaque and monotonous. In addition, updating spreadsheets and tracking individual data took time that could be used to solve business problems. This is where dashboards come into play. With dashboards, data tracking is neither monotonous nor opaque.

Veronica S. Smith, in her work 'Data Dashboard as an Evaluation and Research Communication Tool', states that dashboards are visual representations that track, analyze, and present the most important information needed to achieve specific goals, and are as such recorded on a single screen. The convenience of dashboards is that the data is constantly updated and provides up-to-date insight into the state of business [1].

There are a number of dashboard templates available on the Internet that are created within Frameworks or with the help of CSS Grid that create Responsive Web Design. Such templates are easy to install and customize, depending on the wishes and needs of end users.

II. RESPONSIVE WEB DESIGN

Responsive web design means that web pages are being able to adapt to different screen widths, so a responsive web design is a way for a programmer to design a web page to match the different widths of the devices from which it was viewed. This way of creating a web page ensures that the content of the page is easily viewed and that the data on the web page is readable, regardless of the dimension of the device from which it was viewed [2].

With the increasing production of devices of different widths, there was a need to develop responsive web pages. In the last few years, mobile sales have outpaced desktop sales, and therefore Google has prioritized mobile-friendly pages in search results.

The term 'Responsive Design' was first used by a web designer Ethan Marcotte. An article from 2010 called 'On A List Apart' talks about how web design should not be stagnant and that new devices must be accompanied by appropriate design. He decided that creating new pages for each type of device would not be a good solution, but that there should be a design that would adapt with its flexibility and fluidity to different screens [3].

In his work 'Responsive Web Design: Present and Future', Muhammad Rizwan Pasha mentions four basic principles of adapting web pages to different screen widths.

The first fundamental principle applies to a Fluid / Flexible Grid. In the context of web pages, a grid is a set of rows and columns and is used to organize the appearance of individual elements within a web page. The flexible grid expresses values in percentages by the following formula: (target / context) x 100 = desired value % (result rounded to 0 decimal place)

Another fundamental principle applies to flexible media such as videos and images. The objective of this principle is to avoid assigning a fixed width to the specified media, but to place them in a container to which the relative value that is expressed as a percentage is added.

The third fundamental principle applies to media queries containing individual CSS rules that apply to width, height, color, resolution, etc. There are two types of pixels in this context: device pixels and CSS pixels. CSS pixels change by zooming in or out. For example, if a user enlarges the page to 150%, then CSS pixels will increase by 50% in width and height, but the device pixels do not change by zooming in or out of individual content. A media query consists of one or more expressions for a given screen width that may be true or false. If the media query is true, the terms given for the appropriate screen width are applied along with the rest of the CSS elements listed.

The fourth fundamental principle relates to dynamic content. Dynamic content refers to content that changes based on the behavior, interests, or preferences of an individual user [4].

A. Responsive Web Design Frameworks

Regardless of the developer's knowledge, creating a customizable website from the start is a very challenging and complicated task. That is why developers tend to choose a predefined work environment or the so-called Responsive Web Design Frameworks.

A Framework is defined as a package consisting of the structure of files and folders of standardized code (HTML, CSS, JS, etc.). Creating a design through a chosen Framework is simple. The developer should be familiar with the documentation of the particular Framework within which the site is built. The goal of the Framework is to save time and enable some kind of universality. It provides developers with a well-known code structure so that they do not waste time on writing the structure of the page, and it reduces much of the working hours [5].

Certainly, there are disadvantages that come with this way of creating a responsive website. Within most frameworks, the vast majority of possible scenarios that can occur when designing a site are predicted. So, in addition to what it takes for a designer to create a customizable website, there is often a surplus of code. Another drawback is that the flexibility to create is reduced because the work environment comes with predefined styles and standards. So if an individual wants to change something within the predefined files, it will take a lot of time, and there is also the possibility of inadvertently changing the structure of the predefined files [6]. There are various types of responsive web design frameworks, and those that will be covered in more detail in this paper are: Foundation, Bootstrap, Materialize and Semantic UI.

Foundation is a responsive frontend framework that consists of HTML, CSS, and JavaScript files to help developers build a responsive web page. Foundation was created by ZURB in 2008 and was exclusively used to build websites for their customers. Later, they combined what they thought was a good foundation and a great relief to any developer, and presented the Foundation to the public for the first time in 2011. From the very beginning, even today, the Foundation is free of charge. The fact that the Foundation has become one of the most famous frameworks is evidenced by the fact that it is used by millions of designers and developers worldwide. The current version of this framework that is currently up-to-date is Foundation 6. This updated version allows you to create quality optimized pages for all types of screens. Within the Foundation there is there is also the Business Foundation that offers professional help, counseling, tools and training for individual company workers who choose to use this framework when designing and developing pages for their clients. The documentation and support they provide shows how much they care about their customers. With the latest update, they have improved the documentation and added a "Getting started" document that offers quick insight into how the Foundation works. They have created a Foundation Forum for all users to post potential questions and share their experience [7].

A responsive web design framework called Bootstrap was designed and implemented by Twitter designers and developers in 2011. Before being known as responsive web design framework, Bootstrap was initially known as "Twitter Blueprint" (2010). In the first months of its existence, Twitter organized a project called "Hack Week" which was supported by various developers who contributed greatly to the further development of these platforms. In this way, developers of these platforms were able to decide in which direction they wanted to develop the style of both Twitter and Bootstrap. Bootstrap is currently maintained by a team of developers on GitHub. Since its release to the public, Bootstrap has had over twenty smaller editions and three large ones. For the first time, a larger release called Bootstrap 2 has been able to create responsive web pages. The current release is Bootstrap 4, which includes the move to Sass and CSS flexboxes. Bootstrap developers encourage the development community to use the newer CSS features and make new technology available through modern browsers [8].

Materialize is a responsive web design framework created and designed by Google. This framework was created to integrate classic design principles with innovation and technology. When building this design language, Google's developers were guided by research on paper and ink, and with the help of tactile reality, created so-called material design. Materialize aims to provide a unique user experience across all Google products. Through various elements such as images, typography and colors, developers place emphasis on visual comfort and put focus on specific content. Materialize comes in two different forms. Depending on knowledge and needs, the developer chooses between Materialize and Sass. Materialize contains a standardized version of the code that includes both the full and minimized versions of CSS and JavaScript files. The Sass option includes SCSS files that require a Sass compiler, but with this option, the developer can choose which components to include when creating the site [9].

Semantic UI is also one of responsive frontend frameworks that is supported by Leaner Style Sheets and jQuery. Semantic UI is different because instead of abbreviations, which are commonly used by other frameworks, it uses real words that resemble ordinary English. Its uniqueness is also evident through the use of five categories to define behavior within the workspace,

namely: UI Elements, UI Collection, UI View, UI Module and UI Behavior [10].

B. Cascading Style Sheet Grid

Cascading Style Sheet Grid, commonly known as CSS Grid, is a way of organizing content on web pages. CSS Grid is a two-dimensional system that simultaneously manages columns and rows. CSS was always used to set the layout of elements on a web page, but it wasn't always practical or user-friendly. It is for this reason that frameworks have been created. However, there are also some disadvantages within frameworks regarding the performance and size of the libraries that need to be included while working on a project within a certain framework. After learning about the CSS Grid, there was a concern that responsive web design frameworks would go down in history, but that's certainly not the case. Flexboxes are also very helpful when building a customizable web site, but they are more intended for one-dimensional content management [11]. CSS Grid is specifically designed to address the appearance and positioning of content in a two-dimensional system across different device widths.

As mentioned previously, before the CSS Grid, various methods were used to define elements using CSS. As technology has evolved, so has the need for a new way of defining elements. Since this type of defining elements (inline-block, float, etc.) is not fixed, if the page were viewed at different sizes, these elements would take up the space depending on the size of the device from which they were viewed, contrary to what the designer intended. This is why most designers and developers opt for a fixed display so that the elements cannot be expanded or overlapped in a given space. CSS Grid provides designers and developers with a mechanism that divides available space into columns and rows and intelligently resizes elements across web pages and applications. For mobile devices and tablets, it is extremely important to tailor web pages and applications to the vertical and landscape orientation of the device. CSS Grid combines the appearance of a grid with media queries so that the author does not have to reorder the elements within the .html file, but defines the location of the element within the CSS file to achieve the desired layout in both orientations [12].

Creating a responsive web page with CSS Grid is easy. The developer builds the grid and organizes the content within it, applying CSS rules to the parent element, which becomes a grid container, and to the children of that element. At the very beginning of document creation, we define the parent element, which has become a grid container, with display: grid. Columns and rows are defined using grid-template-columns and grid-template-rows, and are included in grid-column and grid-row elements. The order in which the elements are written in an HTML document is irrelevant, because the same order can be completely altered by a few lines of code in CSS, which is why CSS Grid is a very powerful tool [13].

III. DATA DASHBOARDS

Dashboards are customizable for the diverse needs of individual companies or institutions. The dashboard works by connecting to a variety of databases, attachments, services and APIs, and ultimately the on dashboard user sees it all in the form of tables, charts, and other ways of displaying information. The great advantage of a dashboard is that it gathers data from different sources into one place and gives you an effective overview of them. In the former way of doing business, much of the employee's time was spent analyzing data. Now they don't spend time on data analyzing because dashboards offer real-time data monitoring so employees can deal with other business issues.

There is no universal dashboard to suit every type of business. That's why the important part of customizing your dashboard is to choose the key performance indicators [14]. A quality dashboard answers important business questions and provides an easy-to-use analysis of those answers. Designing an efficient dashboard is greatly aided with the principle of an inverted pyramid. The principle of the inverted pyramid indicates that the most important business indicators or individual activities that the user wants to monitor are placed at the very beginning. In this way, the most important indicators are in the first place and thus attract the necessary attention form the user. The rest of the pyramid is filled with details consisting of more general information for a deeper understanding of the data [15]. However, it is very important to consider how people of different functions use the dashboard. Therefore, it would be best to have one dashboard for specific users who are in certain functions in the business, because not all users are interested in the same business-related information. Therefore, it is recommended that multiple dashboards are created for the targeted group so that everyone can quickly and efficiently access the data they need to continue their quality work. In order for users of dashboards to know at a glance what direction their business is headed, choosing the right colors is important. It has become generally accepted that green indicates something positive, safe, and encourages people, while red indicates danger and indicates caution. Proper use of these two colors can give a quick insight in the state of the business. Also, other colors should contribute to the dashboard's harmony, rather than distract users. An important factor in building an effective dashboard is the amount of data that is visible on the dashboard. The average person can hold between five and nine data in their short-term memory. Anything that exceeds a given number requires more analysis and reflection on the part of the individual [16].

An important factor in creating a dashboard is how you visually present your data. That means that the designer should know what is the best way to visualize each type of data. The purpose of data visualization is to clearly convey the values of the data in graphs. In order for a single visualization to truly convey and display individual data, it is important to find the best possible way in which that data can be displayed, rather than using individual forms of data presentation simply because they are "cool". When using forms that present data which are not thoughtfully picked, it confuses the users and based on the information presented they come to conclusions more difficult. Line charts are used to give users insight into the trends of different categories that take place over the same period.

Pie charts are useful to show what the whole is made up of and how much space does each item occupy. It is important to emphasize in pie charts that it is best to use up to six or fewer different values of the whole to give a good quality data view, so the differences between the values are more visible [17].

These were just some of the ways to visualize data. But not everything is in appearance. The most important part is that a variety of data can be inserted into the dashboard, which, apart from beauty in appearance, ads on functionality in every sense of the word. The goal of the dashboard is to give users an instant insight into the state of the business, not to think extra. It is also important to show the context of the information provided. For the most part, the current data includes relevant data from the previous period, depending on the type of data being displayed. In order for the dashboard to be as efficient as possible, it is important to provide the end user with an effective view of the data on which his business depends [18].

IV. THE RESEARCH

The research includes finding and installing dashboard templates that are created within mentioned responsive web design frameworks and with the help of CSS Grid. Subsequently, individual dashboard templates were evaluated by defined criteria.

A. Criteria for evaluating dashboards

Noor Nashriq Ramly, Fazli Mat Nor, Nurul Haszeli Ahmad and Mohd Haris Aziz in their paper "Comparative Analysis on Data Visualization for Operations Dashboard" [19] set out certain criteria by which dashboards are being valued. The first criterion involves understanding and refers to the correct definition of the most relevant data. The second and third criteria are related to the visual aspect of the dashboard. They mostly deal with interactive and dynamic elements. The fourth criterion discusses the importance of using the correct visual representations for each type of data. The fifth criterion is about improving the ability to remember certain data that is of utmost importance to the business.

The criteria by which dashboard templates that were chosen are evaluated as follows:

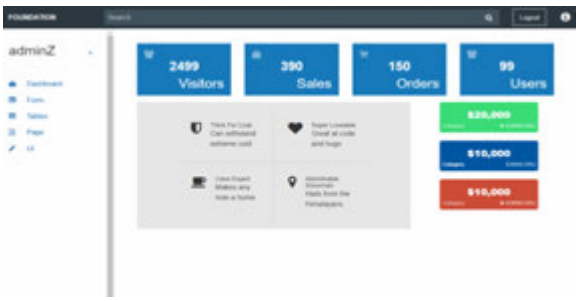
1. The 5 second rule - The template simplifies complex and gives an insight into the most relevant business information at a glance,
2. The inverted pyramid – the most important data should be placed at the very top so that there is a clear hierarchy among the data that is presented
3. A profound way of visualizing data - not every graph is good for every type of data, so you need to tailor the visualization to specific types of data,
4. Less is more – a single dashboard should have no more than 5-9 different ways of visualizing data.
5. The use of interactive elements

B. Choosing the dashboard template

For each of these responsive web design frameworks and for the CSS Grid, one representative is selected, based on Google search popularity [20].

	Dashboard template				
	AdminZ	***Material Dashboard***	***Material Design Admin Template***	***CSS Grid - Matt Holland***	***Stagb Admin***
The search query	Foundation dashboard template free	Bootstrap dashboard template free	Materialize dashboard template free	CSS Grid dashboard template free	Semantic UI dashboard template free
Date of the search	07.01.2020.	07.01.2020.	07.01.2020.	08.01.2020.	12.01.2020.

The first dashboard, called "AdminZ", was created within the Foundation responsive web design framework.



Picture 1. Dashboard template layout created within Foundation framework

The second dashboard called "Material Dashboard" was created within Bootstrap responsive web design framework.



Picture 2. Dashboard template layout created within Bootstrap framework

The third dashboard called "Material Design Admin Template" was created within the Materialize responsive web design framework.



Picture 3. Dashboard template layout created within Materialize framework

The fourth dashboard was made by using CSS Grid and was created by Matt Holland.



Picture 4. Dashboard template layout created using CSS Grid

The fifth dashboard called "Stagb Admin" was created using Semantic UI responsive web design framework.



Picture 5. Dashboard template layout created within Semantic UI framework

C. Results and evaluation

Evaluation criteria	Dashboard template				
	AdminZ	***Material Dashboard***	***Material Design Admin Template***	***CSS Grid - Matt Holland***	***Stagb Admin***
The 5 seconds rule	5	5	5	3	4
Inverted pyramid	5	5	5	5	5
Data visualization	2	5	4	3	5
Less is more	3	5	4	4	4
Interactivity of elements	1	3	4	2	5
	16	23	22	17	23

It should be emphasized that within this paper, the example / home pages that the authors of these dashboards have compiled to show the capabilities that exist within their template have been evaluated. It should also be borne in mind that these are templates that need to be tailored to particular type of business. As mentioned above, each dashboard comes with various elements that are not necessarily implemented on the homepage, but the user can easily implement them and assign them the desired values. The user can certainly arrange the data himself and depending on the needs, give them the appropriate type of data visualization. The code itself that comes with each dashboard template created within the responsive web design framework is easy to read and the individual elements can be easily implemented within it. The grades were given on a scale from 1 to 5, with 5 being the best scoring result. Some of the criteria demand grades based on the subjective opinion and those were given by the authors themselves as well as other grades in this paper.

A dashboard template called "AdminZ" created within Foundation framework does not exude a modern design. The template brings together business-critical data in one place, but presents them with only two non-interactive visualization modes. The most important business-related data is prioritized at the top of the dashboard, but the template does not offer the best solution for visualizing that data.

A dashboard template called "Material Dashboard," created within Bootstrap framework exudes a very modern design. At a glance, it gives you an insight into the most important business-related data, and with the help of interactive graphs gives you an overview of the past and present state of the business. Data visualization modes are meaningfully selected and placed within the template. This template also has an interactive part where an individual can mark or cross the tasks he or she has completed. This template offers a handful of visualization modes, but on the example of their template, they have chosen 5 different modes that perfectly fit the data presentation needs. This template also offers ready-made building blocks that are easy to implement with minimal HTML knowledge.

A dashboard template created within Materialize framework called "Material Design Admin Template," is also a very modern example. The most important business data can be easily read by the first glance at the template, but the visualization methods are not the most cleverly selected because they only give an insight into the current state of business, and a comparison with the previous data would be useful in determining the actual state of business. Within this template we come across two interactive elements, one of which is the same as the previous example of the dashboard template, while the other provides the ability to edit a personal profile and enter personal information. This template also offers a large selection of visualization modes, but six examples have been chosen based on their template. This dashboard offers the ability to change the typography, color, incorporate ready-made tables and forms, offers various tabs within which the user can place individual interactive elements and various buttons that are used as a navigation through the dashboard. With the ready-to-go examples of the elements comes a code that needs to be placed in the desired location, which greatly facilitates the job of customizing this dashboard template.

The fourth dashboard template created by Matt Holland was created with the help of CSS Grid. This template is very modern, but at first glance it does not give

the clearest picture related to business. Data is arranged in order of importance, but the visualization of the data is not best chosen. Some data visualizations do not offer insight into the past, and based on the present data alone, one cannot get a clear insight into the specific state of things related to the business. The template offers minimal inactivity where the user gets more information related to a particular data view by moving the mouse over individual items. The template contains a very small number of different data visualizations.

The fifth template called "Stagb Admin" was created within the Semantic UI framework. This template has a very simple yet stylish design that gives an airy impression. At first glance, this template provides essential data, but much of it remains in the shadow because it does not fit on top of the dashboard. The information that is most relevant to the business is at the very top and in a clear and interactive way gives the user an insight into the details of particular data. Visualization modes are carefully chosen and faithfully represent the current business situation. The dashboard offers a handful of embedded blocks such as various graphs, tables, forms, and the like.

V. CONCLUSION

Dashboards have greatly facilitated business around the world. With the help of dashboards, all business-relevant information is in a single place. The fact that dashboards save time goes in favor of employees because they no longer having to manually summarize and track data, but can monitor it in real time. Different types of business require a different view of the hierarchy among the data, which is where dashboard templates come into play. The dashboard template can be easily installed and customized by the developer who organizes and presents the data by the user's requirements.

When choosing an effective dashboard template, it is important to determine if there are different ways of visualizing the data that is needed for the business, then it is important to set the desired hierarchy among the data and thus identify the most important ones. It is equally important to briefly summarize the data so that the dashboard users will at first glance get specific information about the current state of business, not to further confuse them with many different visualization modes with too much detailed data. Also, interactive elements that allow users to get a closer look at particular data or the ability to capture specific notes are of great importance.

Many dashboard templates are built within responsive web design frameworks that make it easy to customize your dashboard design. A good number of dashboard templates have also been created with the help of CSS Grid, but there are certainly more that are created within responsive web design frameworks. Free dashboard templates that are easily accessible and high quality are mostly created within the Bootstrap framework, while free templates within other frameworks and those created with the help of CSS Grid languages are somewhat more difficult to come by.

On the chosen samples of dashboards, according to the given criteria the best results were achieved by "Material Dashboard" and "Stagb Admin", however the others are not far behind. The results within the given criteria would nevertheless emphasize "Material Design Admin Template" and "Stagb Admin" because they were not evaluated by any criterion with a score less than 4. It can be concluded that in a given sample, according to the given evaluation criteria, "Stagb Admin" represents a great choice when selecting a dashboard template.

The gathered results are general, without taking in consideration conditions of a particular sector of industry, business world, institutions of social importance or specificity of any domain. Based on the gathered results, it is possible to test observed dashboard pattern and to confirm or refute the values presented in this paper.

REFERENCES

[1] S. Smith, Veronica. Data Dashboard as Evaluation and Research Communication Tool URL: https://doi.org/10.1002/ev.20072 (2020-01-18)

[2] Pilon, Annie. What is responsive web design? URL: https://smallbiztrends.com/2013/05/what-is-responsive-web-design.html (2020-01-12)

[3] Wright, Nicholas. What is responzive web design & why do you need it? URL: https://www.upwork.com/hiring/mobile/what-is-responsive-web-design/ (2020-01-12)

[4] M. Rizwan Pasha, Responsive web design: present and future, MS (CS) 1-A, Enrolment Number: 01-243151-006, 2015.

[5] Marcotte, Ethan. Responsive Web Design. URL: http://alistapart.com/article/responsive-web-design/ (2020-01-12)

[6] Introduction to RWD frameworks. URL: https://hub.packtpub.com/introduction-rwd-frameworks/ (2020-01-12)

[7] ZURB Foundation: About. URL: https://foundation.zurb.com/showcase/about (2020-01-12)

[8] Bootstrap: About. URL: https://getbootstrap.com/docs/4.3/about/overview/ (2020-01-12)

[9] Materialize: About. URL: https://materializecss.com/about.html (2020-01-12)

[10] Gerchev, Ivaylo. Introducing: Semantic UI Component Library. URL: https://www.sitepoint.com/introducing-semantic-ui-component-library/

[11] A complete guide to Grid: Introduction. URL: https://css-tricks.com/snippets/css/complete-guide-grid/#grid-introduction (2020-01-12)

[12] CSS Grid Layout Module Level 1. URL: https://www.w3.org/TR/css-grid-1/ (2020-01-12)

[13] A complete guide to Grid: Basics and Browser Support. URL: https://css-tricks.com/snippets/css/complete-guide-grid/#grid-browser-support (2020-01-12)

[14] What is a data dashboard? URL: https://www.klipfolio.com/resources/articles/what-is-data-dashboard (2020-01-12)

[15] Ilan Hertz, Dashboard design best practices – 4 key principles. URL: https://www.sisense.com/blog/4-design-principles-creating-better-dashboards/ (2020-01-12)

[16] 5 zlatnih pravila za stvaranje efikasnog dashboard-a. URL: https://oxidian.hr/5-zlatnih-pravila-za-stvaranje-efikasnog-dashboard-a/ (2020-01-12)

[17] Blitz, Shelby. 14 Useful Ways to Visualize Your Data (with Examples). URL: https://www.sisense.com/blog/10-useful-ways-visualize-data-examples/ (2020-01-12)

[18] 5 zlatnih pravila za stvaranje efikasnog dashboard-a. URL: https://oxidian.hr/5-zlatnih-pravila-za-stvaranje-efikasnog-dashboard-a/ (2020-01-12)

[19] Noor Nashriq Ramly, Fazli Mat Nor, Nurul Haszeli Ahmad, Mohd Haris Aziz. International Journal of Information and Education Technology: Comparative Analysis on Data Visualization for Operations Dashboard, Vol. 2, No.

[20] How Search algorithms work. URL: https://www.google.com/search/howsearchworks/algorithms/ (2020-01-22)

Razvoj sustava za upravljanje rasporedom sati

Adriana Šturlan, Klara Vučković i Tihomir Orehovački
Fakultet informatike u Puli
Sveučilište Jurja Dobrile u Puli, Pula, Hrvatska
{asturlan, kvuckov, torehov}@unipu.hr

***Sažetak* – Raspored sati jedan je od klasičnih modela upravljanja vremenom kod kojeg se javlja optimizacijski problem alociranja resursa u vremenske jedinice pod uvjetom zadovoljavanja postavljenih ograničenja. Sustav za upravljanje rasporedom nudi svojevrsno rješenje i djelomičnu automatizaciju ovog procesa te je namijenjen korištenju na sveučilišnoj razini sa ciljem olakšavanja izrade semestralnog modela rasporeda sati i optimizacije organizacije resursa potrebnih za izvođenje nastave. U ovom je radu opisan logički dizajn sustava, njegove funkcionalnosti, programsko rješenje te web tehnologije koje su korištene u implementaciji.**

Ključne riječi – web aplikacija; raspored sati; upravljanje vremenom; problem optimizacije; NP-težak problem; NP-potpun problem; JSON; REST; Vue.js; Node.js; Express.js; PostgreSQL

I. UVOD

Trenutno na tržištu postoji nekoliko sustava čija je svrha optimizacija i olakšavanje izrade rasporeda sati. Neka od ponuđenih rješenja su kako slijedi:

- „ascTimetables“ [1] je komercijalna aplikacija koja sadrži funkcionalnosti kreiranja vlastitih bilješki, automatiziran odabir plana, filtrirani prikaz rasporeda te mogućnost ručnog i automatskog generiranja rasporeda.
- „Wise Timetable“ [2] je komercijalna aplikacija koja nudi mogućnost generiranja rasporeda temeljem ručno unesenih podataka, korištenje postojećeg generiranog rasporeda za dodavanje izmjena te automatsko popunjavanje preostalog dijela rasporeda. Aplikacija pruža mogućnost rezervacije dvorana i ispita te generiranja tjednih rasporeda.
- „Raspored sati“ [3] je sustav razvijen u Sveučilišnom računskom centru Sveučilišta u Zagrebu koji pruža mogućnost odabira akademske godine, semestra, tjedna i grupe pri izradi rasporeda. Ovaj sustav je bio aktivno korišten do 2020. godine kada je uklonjen sa službenih stranica SRCE-a.

Spomenute aplikacije dijele iste funkcionalnosti ali i mane, od kojih se mogu istaknuti odužen period prilagodbe na aplikaciju, kompliciranog korisničko sučelje te optimizacijski problem izrade i usklađivanja rasporeda gdje aplikacija sa porastom broja ograničenja i ulaznih podataka postaje sve manje upotrebljiva [4]. Osim toga, većina se spomenutih rješenja sustavno naplaćuje ili su prilagođene za specifične potrebe pojedine ustanove.

Problem optimizacije proizlazi iz velike količine nezavisnih varijabli te specifičnosti područja uporabe. U slučaju izrade rasporeda sati unutar obrazovne ustanove problemi se javljaju kada se kao ulazni podaci uključe varijable dostupnosti dvorana ili nastavnika u određenoj vremenskoj jedinici te kapacitet dvorana i broj upisanih studenata u tekućoj godini. Upravo zbog međuovisnosti varijabli te velikog broja kombinacija i mogućih rješenja, donošenje odluka tijekom kreiranja rasporeda slovi kao NP-potpun problem dok kreiranje optimalnog rasporeda kao NP-težak problem jer se optimalno rješenje ne može pronaći u polinomijalnom vremenu [5].

Predloženi sustav za upravljanje rasporedom sati nudi potporu u izradi semestralnog rasporeda sati na primjeru fakulteta, a temeljna su mu obilježja lakoća pristupa, jednostavnost korištenja i ubrzavanje procesa izrade rasporeda. Sustav omogućava reprezentaciju korisnički kreiranih rasporeda popraćenih vizualnim asistencijama te nastoji svojim jednostavnim korisničkim sučeljem pružiti ugodno iskustvo rada.

Kreiran sa ciljem djelomične automatizacije cjelokupnog procesa izrade rasporeda projekt nastoji otkloniti navedenu problematiku te olakšati pronalazak optimalnog rješenja problema rasporeda.

II. IDEJA I MOTIVACIJA

Izrada rasporeda sati dugotrajan je i zahtjevan proces, a upravo je navedena problematika glavni motiv izrade ovog sustava. Mnoštvo je elemenata i ograničenja potrebno uzeti u obzir tijekom izrade, poput broja upisanih studenata, dostupnosti nastavnika, zauzeća, kapaciteta i opremljenosti dvorana, te trajanja izvođenja različitih oblika nastave. Ovo su neki od problema koji se ovim sustavnom nastoje eliminirati:

1. problem vremenskog ograničenja nastavnika te njegova dostupnost u određenoj jedinici vremena,
2. problem optimalnog razmještaja različitih oblika nastave uzimajući u obzir lokaciju dvorane održavanja,
3. problem integriteta rasporeda sati,
4. problem optimizacije zauzeća dvorana u vidu iskoristivosti kapaciteta i opremljenosti.

Svaka obrazovna ustanova ima svoja unutarnja ograničenja što može donekle otežati posao izrade rasporeda kao što je primjerice održavanje nastave na dvije različite lokacije. Stoga je jedan od ciljeva ovog sustava i obavještavanje korisnika u slučaju narušavanja postavljenih pravila i ograničenja.

III. ZAHTJEVI NA SUSTAV

Zahtjevi prema kojima je sustav dizajniran i pripremljen za uporabu uključuju sljedeće funkcionalnosti:

- mehanizam prijave i odjave iz sustava,
- pregled i uređivanje godišnjih rasporeda sati
- pregled i uređivanje pojedinih semestralnih rasporeda sati.

Uređivanje godišnjih rasporeda kao glavni aspekt sustava pruža mogućnost odabira i uređivanja pojedinih rasporeda filtriranih po razini studija, načinu izvođenja studija te broju semestra, čije dostupne vrijednosti ovise o tome pripada li godišnji raspored zimskom ili ljetnom semestru. Odabirom željene kombinacije, korisnik pristupa listi dostupnih predmeta, odvojenih po kategorijama (predavanja, vježbe, seminari), koje je potrebno, uz odabir željene dvorane izvođenja, smještati u raspored sati. Kako bi se zadovoljio kriterij integriteta te kako bi raspored bio upotpunjen, potrebno je iskoristivu semestralnu satnicu svesti na nulu.

Navedeni zahtjevi na sustav primjenjuju se na vremenske jedinice (polja) unutar pojedinog rasporeda objedinjujući sljedeće funkcionalnosti:

- vizualni prikaz popunjenih polja i pripadajućih detalja,
- označavanje polja nedostupnih za popunjavanje odabranom kombinacijom predavanja i dvorane,
- označavanje polja dostupnih za popunjavanje odabranom kombinacijom oblika nastave i dvorane,
- oslobađanje popunjenih polja od zauzeća,
- premještanje popunjenih polja na drugu dostupnu vremensku jedinicu izvođenja.

Uz navedene zahtjeve na sustav, u obzir treba uzeti i potrebu prikazivanja presjeka dostupnosti polja kroz sve rasporede unutar godišnjeg rasporeda kako bi se spriječilo zauzimanje dvorane već zauzete na jednom od ostalih rasporeda u istoj vremenskoj jedinici ili zauzimanje polja u čijoj je vremenskoj jedinici odabrani nastavnik već zauzet [6]. Sustav je dizajniran temeljem dijagrama slučaja korištenja prikazanog na slici 1. čije su slučajevi izvedeni iz spomenutih zahtjeva na sustav.



Slika 1. Dijagram slučajeva korištenja [5]

IV. LOGIČKI DIZAJN SUSTAVA

Sustav je izgrađen kao moderna web aplikacija, stoga joj je moguće pristupiti sa bilo kojeg uređaja putem web preglednika. Podaci se u bazu podataka spremaju u stvarnom vremenu i nakon svake izmjene što rezultira rasterećenjem korisnika od samostalnog spremanja promjena te sprječava gubitak podataka.

Tijekom rada, sustav pored korisnički kreiranih podataka koristi vanjski definirane podatke koji, ovisno o korisničkom izboru, selektivno uključuju: dostupne predmete i nastavnike, dvorane, broj upisanih studenata, broj semestara, razinu studija te pristup studiranju. Podaci koje korisnik generira svojim radom u aplikaciji su godišnji rasporedi te predavanja unesena u polja rasporeda tijekom kreiranja i uređivanja rasporeda. Takav pristup aplikaciji daje slobodu integracije s drugim sustavima u budućem razvoju te rasterećuje korisnika od nepotrebnih koraka samostalnog unosa podataka potrebnih za rad.

A. Model klijent-poslužitelj

Logički dizajn distribuiranog sustava temelji se na nezavisnoj komunikaciji klijenta i poslužitelja gdje klijent od poslužitelja očekuje odgovor, dok poslužitelj odgovara na zahtjeve klijenta. Niti jedna strana nije svjesna postojanja druge, ali zato obje strane moraju imati uniformni jezik kako bi postigle međuprocesnu komunikaciju. Funkcije klijenta u ovom sustavu su sljedeće:

- isporuka korisničkog sučelja,
- validacija korisnički unesenih podataka,
- slanje formatiranog upita poslužitelju,
- zaprimanje i obrada odgovora,
- prikaz obrađenih podataka korisniku.

Klijent u ovom sustavu snosi funkcionalnu težinu obrade podataka (debeli klijent), dok se poslužitelju pristupa isključivo za dohvat podataka, autentikaciju te verifikaciju.

Poslužiteljski dio aplikacije odgovara na zahtjeve klijenta vračajući podatke u formatu JSON. Zahtjevi klijenta šalju se na definirane REST servise, a pristigli podaci se zatim obrađuju ili se dohvaćaju iz baze podataka. Funkcionalnosti poslužitelja ovog sustava su sljedeće:

- prihvaćanje zahtjeva klijenta,
- upravljanje pristupom,
- obrada nadolazećih zahtjeva,
- interakcija s bazom podataka,
- slanje formatiranih podataka klijentu.

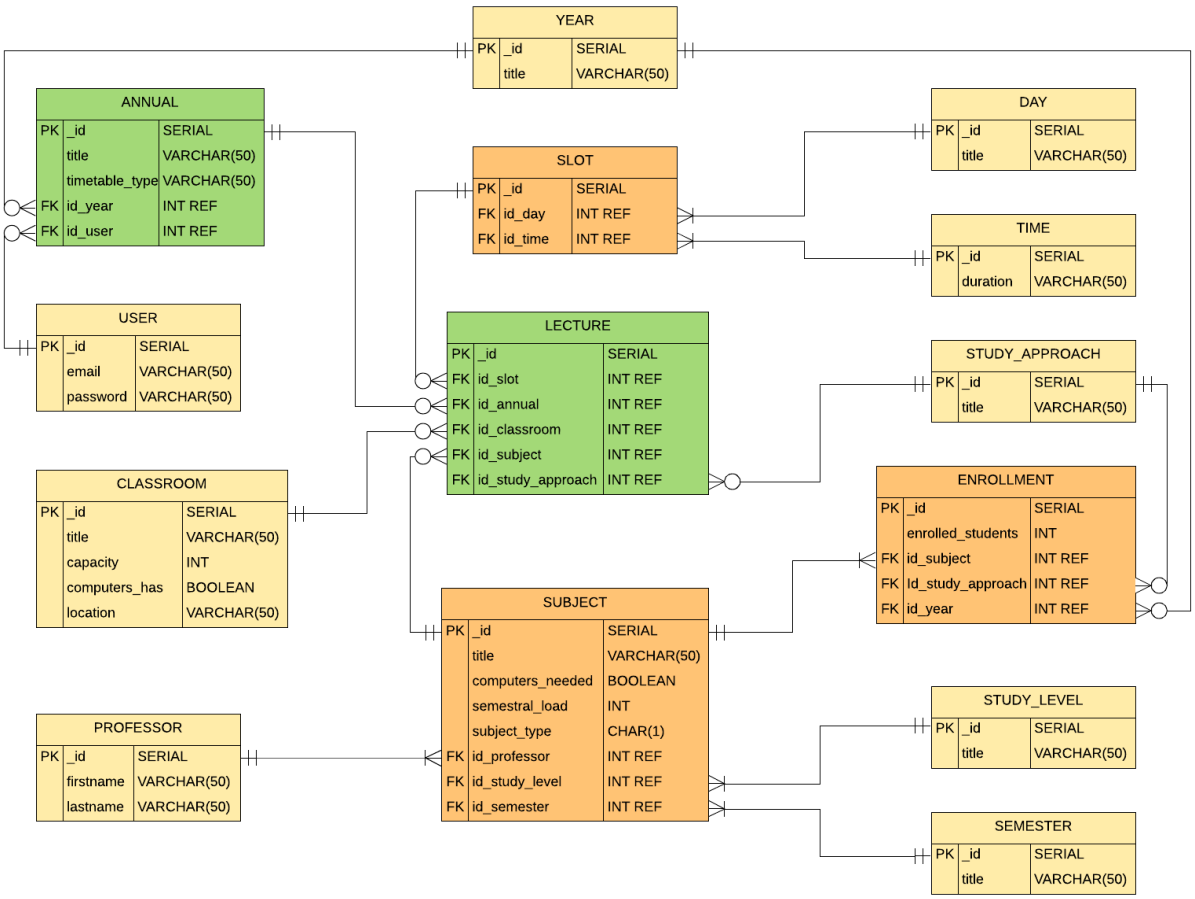
Svaki zahtjev upućen poslužitelju pogađa specifičnu putanju čije se nazivlje nastoji održati opisnim radi lakšeg razumijevanja njezine funkcije. Zahtjevi zauzvrat uz odgovor očekuju i određene standardne statusne kodove, na koje adekvatno reagiraju.

B. Model baze podataka

Pri svakom zahtjevu, poslužitelj autentificira korisnika, a u slučaju potrebe za dostavljanjem podataka, poslužitelj se povezuje sa bazom podataka putem ORM alata za transakcijsku podršku. Logički model podataka prikazan na slici 2 ilustrira veze unutar baze podataka te pripadajuće kardinalnosti i odgovarajuće tipove podataka. Isti se sastoji od 14 povezanih relacija od kojih svaka relacija sadrži primarni ključ (jedinstveni id) te ostale atribute koji su preslikani u fizičku implementaciju baze podataka. Primarni se ključevi u bazi generiraju sekvencijalno pa stoga na sebi nose tip podatka SERIAL, a tipovi ostalih podataka usklađeni su sa specifikacijom odabrane baze podataka. Strani ključevi i veze prikazani su primjenom Martinove notacije.

Iz logičkog modela podataka se mogu izdvojiti relacije obojane zeleno jer se CRUD operacije izvršavaju isključivo nad njima. Spomenute se relacije, građene po modelu pahuljice, vežu za okolne relacije (dimenzije) putem stranih ključeva. Tako se primjerice relacija „Lecture“ sastoji od stranih ključeva svih relacija koje sadrže detalje o predavanjima. Spomenuta je relacija kreirana u normaliziranom obliku i s minimalnom redundancijom podataka jer korisnički kreiranih predavanja može biti relativno mnogo.

Žuto obojane relacije sadrže podatke koji se koriste u drugim relacijama (obojenima narančasto) te ne sadržavaju strane ključeve. Svaka je relacija u fizičkoj implementaciji prikazana pripadajućim Sequelize modelom [6].



Slika 2. Logički model podataka [6]

V. FUNKCIONALNOSTI SUSTAVA

Korisnik se pri ulasku u sustav najprije susreće sa obrascem za prijavu kojeg popunjava podacima predefiniranim od strane administratora, čime se sustav štiti od neovlaštenog pristupa. Nakon uspješne prijave, moguće je pristupiti sljedećim dvama modulima:

1. upravljanje godišnjim rasporedima,
2. uređivanje godišnjih rasporeda.

Modul za upravljanje godišnjim rasporedima (prikazan na slici 3) obuhvaća dodavanje i brisanje rasporeda te je usko vezan za uređivanje pojedinih rasporeda. Odabirom aktivnog godišnjeg rasporeda, korisnik dobiva pristup modulu uređivanja semestralnih rasporeda sati.



Slika 3. Modul za upravljanje godišnjim rasporedima [6]

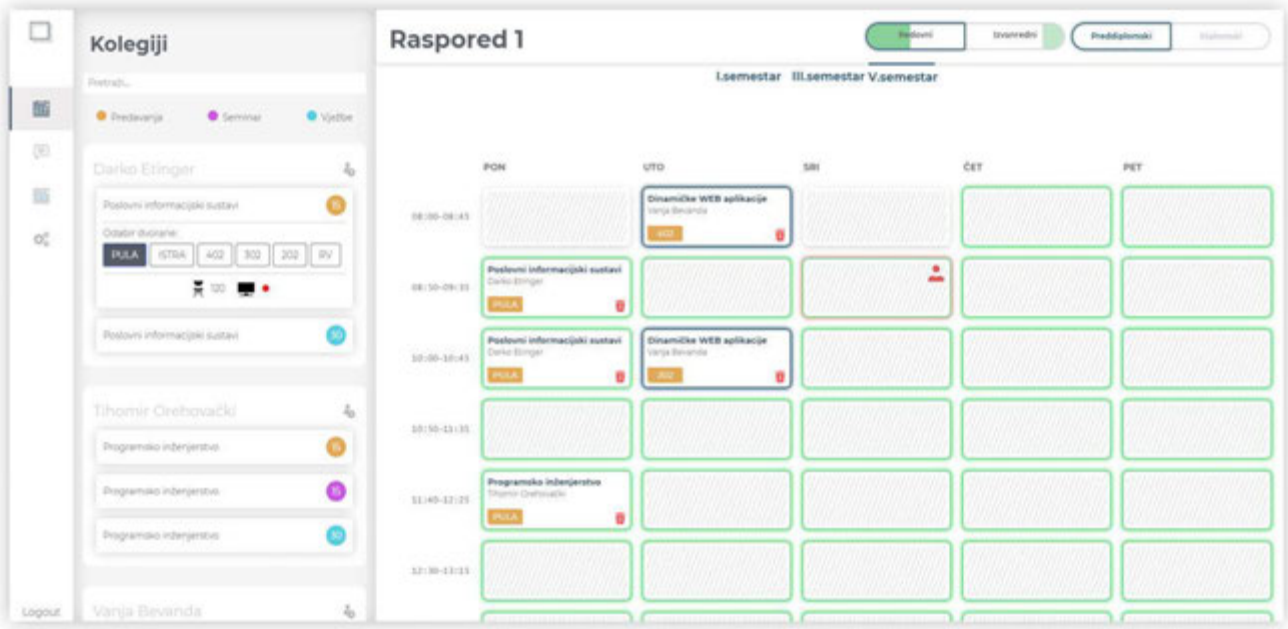
Odabirom ikone rasporeda u glavnom izborniku, korisnik pristupa modulu za uređivanje pojedinih rasporeda sati (prikazan na slici 4). Ovaj modul obilježava lagan pristup svim rasporedima unutar godišnjeg rasporeda pomoću navigacije u gornjem desnom uglu aplikacije, čijom se kombinacijom prikazuje željeni raspored u načinu rada za uređivanje.

Ovisno o odabranoj kombinaciji, na lijevoj traci aplikacije prikazani su svi predmeti koje je potrebno unijeti u raspored. Svaki se predmet može podijeliti na maksimalno tri kategorije izvođenja nastave – predavanja, seminari, vježbe - od kojih svaka kategorija ima svoje semestralno opterećenje te specifične zahtjeve na opremljenost dvorane izvođenja. Svaki predmet i kategorija predmeta, kao preduvjet izvođenja, može zahtijevati da dvorana izvođenja bude opremljena računalima, što se uzima u obzir kod ponude dvorana. Ukoliko dvorana nije opremljena računalima, a predmet zahtjeva da bude, dvorana neće biti prikazana u listi dostupnih. Također, svaki predmet ima određeni broj upisanih studenata te ukoliko je kapacitet dvorane manji od broja upisanih studenata, dvorana neće biti prikazana na listi dostupnih dvorana.

Podno odabrane dvorane, vizualno se prikazuje njezin kapacitet i opremljenost. Bitno je napomenuti da ukoliko odabrani predmet ne zahtijeva dodatnu opremljenost, korisnik i dalje kao izbor ima odabir dvorane opremljene računalima.

Nakon odabira kombinacije, klikom na odabrano zeleno polje, ono postaje popunjeno detaljima predavanja te se od ukupne satnice oduzima broj jednak zbroju tjedana u odabranom semestru. Slika 5 prikazuje stanja u kojima se polja mogu nalaziti. Zelena boja označava dostupnost dvorane i nastavnika, dok crvena, sa dodatkom ikone čovječuljka, nedostupnost odabranog nastavnika u određeno vrijeme.

Ukoliko je polje uokvireno i crvenom i zelenom bojom, to znači da je odabrana dvorana u to vrijeme dostupna, no odabrani nastavnik nije. Kada je polje popunjeno predavanjem, isto se uokviruje plavom bojom. Prazno neuokvireno polje označava da je u to vrijeme odabrana dvorana nedostupna te da se ne održava niti jedno predavanje u trenutnom rasporedu. Također, bitno je naglasiti da se u bilo kojem trenutku popunjeno polje može osloboditi od predavanja.



Slika 4. Modul za uređivanje pojedinih rasporeda [6]



Slika 5. Moguća stanja pojedinih polja

Uređivanje jednog rasporeda unutar godišnjeg rasporeda reflektira izradu ostalih pojedinačnih rasporeda. To rezultira različitim kombinacijama okvira pojedinih polja te predstavlja glavnu funkcionalnost ovog sustava.

Dodatna funkcionalnost sustava je i prikaz predavanja po određenom nastavniku. Taj prikaz služi kao pomoć pri izradi uzajamno optimalnog rasporeda te ga nije moguće uređivati [6].

VI. PROGRAMSKO RJEŠENJE

A . Prijava u sustav

Prilikom ulaska u sustav, od korisnika se zahtjeva unos korisničkog imena i lozinke kako bi se eliminirala mogućnost neovlaštenog pristupa istoj.

Neovlašteni pristup na strani klijenta rezultira zabranom pristupa bilo kojoj ruti sve dok poslužitelj ne vrati odgovor uspješne autentikacije na klijentov zahtjev tipa POST. Poslužitelj, nakon zaprimanja zahtjeva, koristeći biblioteku *express-validator* usmjerava zahtjev autentikacijskom posredniku pod imenom *passport* čija je zadaća utvrđivanje prava pristupa korisnika sustavu. Bitno je naglasiti da se ispravnost lozinke provjerava odvojeno od provjere ispravnosti e-mail adrese (vidljivo na slici 6).



Slika 6. Modul za prijavu u sustav [5]

Kada je korisnik uspješno autoriziran, poslužitelj generira jedinstveni token čija je uloga identifikacija skupa zaprimljenih podataka. Objekt modela „Korisnik“ se zatim potpisuje BASE64 tekstualnom reprezentacijom rezultata SHA256 algoritma čime dobivamo token.

Generirani se token zatim vraća klijentu u obliku odgovora. Nakon što klijent zaprimi odgovor od poslužitelja, on provjerava uspješnost prijave. Kreirani se token otpušta onog trena kada istekne vrijeme njegovog trajanja definirano kroz neobavezni parametar funkcije za kreiranje tokena.

Sukladno odgovoru, korisnički se token sprema u obliku kolačića čime korisnik dobiva pristup rutama poslužitelja. Nadalje, svaki zahtjev klijenta prolazi kroz posebni posrednik čija je svrha utvrđivanje postojanja i validnosti tokena.

B. Popunjavanje rasporeda postojećim podacima

Svakim se klikom na dvoranu, pristup, semestar ili razinu studija, raspored iznova generira, čime prikazuje aktivno stanje odabranih polja. Posebna metoda zatim kreira objekt koji predstavlja jedan redak u rasporedu. Objekt se sastoji od te liste polja, dok se pojedino polje sastoji od sljedećih podataka: dostupnost dvorane, dostupnost nastavnika te, ukoliko je polje popunjeno, objekta koji sadrži informacije o nastavniku i predmetu.

Nakon kreiranja spomenutog objekta, druga metoda dohvaća sva vremena te prolaskom kroz svaki redak rasporeda, označava polja čije vrijeme održavanja odgovara ograničenjima. Metoda zatim pronalazi sva predavanja koja su već dodijeljena svakom pojedinom polju te se vrši provjera odgovara li odabrana dvorana dvorani održavanja predavanja u polju. Ukoliko nije odabrana nijedna dvorana, polje je uokvireno plavom bojom, a ukoliko je dvorana odabrana, vrši se pretraga predmeta koji odgovaraju odabranim vrijednostima. Zatim se vrši provjera dostupnosti nastavnika u određenom terminu polja. Ako su sva ograničenja zadovoljena, polje se uokviruje zelenom bojom. Nakon toga slijedi drugi prolazak kroz popunjena polja pri čemu se provjerava status zauzeća dvorane ovisno o neaktivnim predmetima koji pripadaju tom istom polju. Ukoliko je dvorana izvođenja bilo kojeg od tih predavanja ista kao i odabrana, polje postaje zauzeto i sa njega se uklanja okvir. Međutim, ukoliko to predavanje pripada odabranom nastavniku, tada je moguće klikom na crvenu kanticu osloboditi polje od predavanja, a time i osloboditi dvoranu.

Osim navedenoga, za svako je polje potrebno pregledavati posjeduje li barem jedno predavanje. Ako polje ne sadrži niti jedno predavanje, ono se automatski označava kao dostupno [5].

C. Popunjavanje rasporeda novim podacima

Prilikom prikaza trenutnog stanja raspoloživih dvorana i nastavnika, odabirom dvorane izvođenja te klikom na zeleno polje u rasporedu, predavanje se unosi u raspored. Semestralna satnica tog predmeta tada se umanjuje za iznos jednak broju tjedana u semestru. Akcija koja umanjuje broj preostalih sati okida se nakon dodavanja predmeta u raspored, dok se akcija uvećavanja broja preostalih sati izvršava prilikom brisanja predmeta iz rasporeda.

Ažuriranje predavanja u bazi vrši se na način da se poslužitelju prosljeđuje aktivni predmet, dvorana, aktivni raspored, pristup studiranju te id polja, čiju zadaću snosi jedna od primarnih funkcija klijenta [5].

D. Oslobađanje popunjenih polja

Prilikom brisanja predmeta iz rasporeda, poslužitelju prosljeđujemo id predavanja kojeg je potrebno osloboditi, te se iz baze podataka briše predavanje pod primarnim ključem te vrijednosti. Broj predavanja tog istog predmeta povećava se za „uklonjenu“ vrijednost te se ono ažurira u listi dostupnih predavanja koje je potrebno smjestiti u raspored. Prema tome, ukoliko je predavanje bilo smješteno u raspored, ono ponovno postaje aktivno u navedenoj listi [5].

VII. KORIŠTENE TEHNOLOGIJE

Moderan i fluidan dizajn aplikacije ostvaren je korištenjem CSS3 i Bootstrap tehnologija, dok su funkcionalnosti na klijentu implementirane kroz Vue.js Javascript okvir uz potporu Vuex i Vue-Router tehnologija te ostalih dopunskih biblioteka. Poslužiteljska strana sustava građena je u Node.js programskom okruženju, a komunikacija između klijenta i poslužitelja ostvaruje se RESTful pristupom.

Suštinski Node.js upotpunjen je Express.js aplikacijskim okvirom čija je svrha apstrahiranje procesa izgradnje sustava te ubrzavanje razvoja servisa. Među ostalim tehnologijama nalaze se i Ekspress-validator kao posrednik u validaciji podataka na strani poslužitelja, zatim Passport i JWT za ovjeru autentičnosti pristupa te Babel i Webpack za ES6 podršku i objedinjenje modula.

Za interakciju s bazom podataka korišteni su PostgreSQL moduli sa svojom kolekcijom funkcija i predefiniranih SQL upita, te Sequelize ORM (eng. *object relational mapping*) moduli koji, bazirani na obećanjima, pružaju podršku izvođenju CRUD operacija kroz odabrani dijalekt [6].

VIII. ZAKLJUČAK

Predloženi sustav za upravljanje rasporedima sati nastoji svojim jednostavnim i intuitivnim sučeljem, fluidnim dizajnom te poboljšanim sustavom rješavanja problematike optimizacije rasporeda vremena i resursa, staviti točku na prvu fazu razvoja te sveučilištima olakšati izradu rasporeda sati kako za nastavnike tako i za studente, te ublažiti mane navedenih postojećih sustava za kreiranje rasporeda sati. U narednim iteracijama fokus će biti stavljen na proširivanje funkcionalnosti u vidu generiranja tjednih rasporeda (kao komplement semestralnoj izradi rasporeda) te razvoj korisničkog sučelja za nastavnike koji će omogućiti dostavljanje ograničenja administratoru putem modula „zahtjevi“. Ova bi funkcionalnost sa sobom donijela i potrebu izrade modula za uređivanje rasporeda u realnom vremenu koji bi služio za vršenje izmjena na postojećem, aktivnom rasporedu sati. Funkcionalnost bi bila popraćena korisničkim sučeljem i pogledom za studente, koji bi imali mogućnost pratiti izmjene u rasporedu u realnom vremenu. Trenutni sustav nudi čvrsti temelj za buduću nadogradnju i razvoj te ostavlja širok prostor za integraciju s drugim komplementarnim sustavima.

LITERATURA

[1] A. Tus, „Analiza zahtjeva i postojećih rješenja u izradi rasporeda sati za škole“, diplomski seminar, Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, 2011, https://ferko.fer.hr/ferko/EPortfolio!dlFile.action?id=54

[2] Wise Technologies URL:http://www.wise-t.com/product-wise-timetable.php , pristupano 05.02.2020.

[3] Javna dokumentacija usluga Srca, https://wiki.srce.hr/, pristupano 06.02.2020.

[4] A. L. Lovelace, „On the complexity of scheduling university courses“, magistarski rad, Faculty of California Polytechnic State University, San Luis Obispo, 2010, <https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1255&context=theses>

[5] K. Vučković, „Razvoj klijentskih komponenti aplikacije za upravljanje rasporedom“, završni rad, Fakultet informatike u Puli, Sveučilište Jurja Dobrile u Puli, 2019.

[6] A. Šturlan, „Razvoj poslužiteljskih komponenti aplikacije za upravljanje rasporedom“, završni rad, Fakultet informatike u Puli, Sveučilište Jurja Dobrile u Puli, 2019.

Enhancing Performance of Cloud-based Software Applications with GraalVM and Quarkus

M. Šipek, D. Muharemagić, B. Mihaljević, and A. Radovan
Rochester Institute of Technology Croatia, Zagreb, Croatia
matija.sipek@mail.rit.edu, dino.muharemagic@mail.rit.edu,
branko.mihaljevic@croatia.rit.edu, aleksander.radovan@croatia.rit.edu

***Abstract* - Increased complexity of network-based software solutions and the ever-rising number of concurrent users forced a shift of the IT industry to cloud computing. Conventional network software systems commonly based on monolithic application stack running on costly physical single-purpose servers are affected by significant problems of resource management, computing power distribution, and scalability. Such implementation is restricting applications to be reduced to smaller, independent services that can be more easily deployed, managed, and scaled dynamically; therefore, embellishing environmental uniformity across development, testing, and production. Current cloud-based infrastructure frequently runs on containers placed in Kubernetes or Docker-based cluster, and the system configuration is considerably different compared to the environment prevailed with common virtualizations. This paper discusses the usage of GraalVM, a polyglot high-performance virtual machine for JVM-based and other languages, combined with new Kubernetes native Java tailored stacked framework named Quarkus, formed from enhanced Java libraries. Moreover, our research explores GraalVM's creation of native images using Ahead-Of-Time (AOT) compilation and Quarkus' deployment to Kubernetes. Furthermore, we examined the architectures of given systems, various performance variables, and differing memory usage cases within our academic testing environment and presented the comparison results of selected performance measures with other traditional and contemporary solutions.**

Keywords – GraalVM, Kubernetes, Quarkus, Cloud computing, native image

I. INTRODUCTION

An ever-increasing number of requirements and users demands ubiquitous availability of network-based software solutions, both server-based and serverless. Server applications need to be stable, responsive, and capable of elastically adjusting performance in order to provide well-timed and adequate complex services. Traditional service-oriented architecture (SOA) focuses excessively on using heavyweight middleware technologies, protocols, and formats such as Simple Object Access Protocol (SOAP) and Web Services Description Language (WSDL), thus becoming too interdependent and not loosely-coupled. These integrated solutions are not as flexible and problems within one part of the service can make the whole inaccessible and consequently inappropriate [1]. The current architectural shift to microservices emphasized scalability, autonomy, faster delivery, and lightweightness of individual independent services [2]. In order to fully use the advantages of microservices, they should be deployed within a cloud-based environment, using lightweight container technologies such as Docker[1] or Kubernetes[2].

One of the clouds' purposes is to set aside complex implementations presented by ununiformed IT-related impediments and help software developers focus on the implementation of core business functionalities while reducing the cost of deployment in comparison to traditional server-based applications. Cloud computing is a software deployment model that embodies several qualifications required by modern software, which allow omnipresence, rapid scalability, and on-demand access to network-based services [3]. As customer demand varies, consequently, various virtual and physical resources are reallocated in real time. In addition, one of the essential features leveraged in cloud systems is the ability of the service to track the number of consumers, bandwidth, server time, and memory processing, hence using a charge-per-use basis model of payment.

The most common cloud computing models are used in three forms: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) [4]; and the main deployment models: Public cloud, Private cloud, and Hybrid cloud [5]. Different software models are allowed to be deployed by various cloud providers such as leading IaaS vendors[3] Amazon Web Services (AWS), Microsoft Azure, Alibaba Cloud, Google Cloud Platform (GCP), Oracle Cloud, and IBM Cloud. Process Virtual Machines are virtualized and managed by a hypervisor on the hardware level [6]. The host processor is communicating with a thin layer of software systems also called Virtual Machine Management (VMM). As presented in Fig. 1., VMM manages sharing the physical resources between the several guest OSs running. While hardware virtualization emphasizes efficient resource utilization and lower cost, it still has limits in scalability, thus when a CPU overhead occurs on a particular VM, additional resources cannot be easily allocated to maintain the efficiency of a given server.

[1] Docker, https://www.docker.com/
[2] Kubernetes, https://kubernetes.io/
[3] Gartner 2019 Magic Quadrant for Cloud Infrastructure as a Service, Worldwide, Gartner, 2019

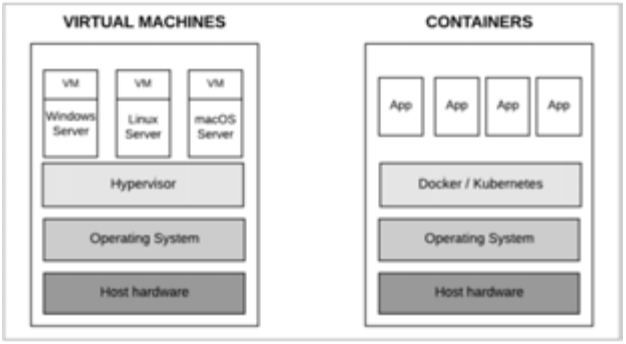


Figure 1. Common structure of containers and virtual machines

The separation of services and containerization is realized by operating system-level virtualization. The kernel allows multiple isolated instances to run simultaneously, while synchronously scaling computing resources. From the application's point of view, they are running on individual computers allowing common expensive resources to be shared.

As client requests and thus, overhead increases, the underlying cloud service providers spawn new work processes and increase computing resources in order to handle to incoming workload [7]. Initial warmup phase of VMs and the spawn of new workers in order to handle the request is named cold start, and it requires initialization of system service configurations and language runtime subsystems before handling requests. Cold start is usually a slow and unpredictable process [8], and this is even more enhanced in cloud environments. High-level object-oriented programming (OOP) languages running on Common Language Runtime (CLR) or Java VM are sometimes slower as they perform several expensive optimization and initialization processes such as class loading, profiling, and dynamic class loading [8].

In our previous research [9] we explored the essential aspects of the GraalVM[4], a high-performance universal polyglot virtual machine, which presented features with high industry potential such as embeddability and the usage of native images in serverless services. As defined in [10] and [11], usage of open source cloud in enterprise systems can excel in the utilization of several crucial business resources, and with the usage of GraalVM this facilitates further. Furthermore, in this research paper, we are exploring GraalVM's ahead-of-time (AOT) compilation mechanism of the Java code to a standalone native executable. In combination with Kubernetes and Java tailored stacked framework Quarkus [5] we tested and analyzed several performance variables, as well as elaborated common system components.

II. GraalVM-based System Overview

A. GraalVM system

GraalVM was developed as a part of OpenJDK's Graal Project [6] initiated to build a next-generation high-performance polyglot virtual machine that offers extensive ecosystem incorporating several important subsystems, including Graal Compiler, GraalVM Native Image mechanism, Truffle Language Implementation Framework and Sulong (LLVM), as presented in Figure 2. and in detail in [9].

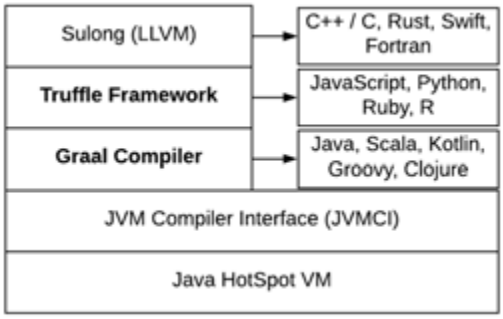


Figure 2. Simplified representation of GraalVM Architecture

GraalVM architecture starts with Java HotSpot VM at the bottom of the system stack. HotSpot VM's components that are not directly connected to the compilation process are being reused, comprising of Java Native Interface (JNI), interpreter, class libraries, class loading, as well as Garbage Collectors (GC) which efficiency is suitable and sufficient for the next generation of high-performance compilers [12], as tested in [13] and [14].

JVM Compiler Interface [7] (JVMCI) is connecting the Graal compiler with the underlying HotSpot VM in order to allow this new compiler written in Java to be used as the just-in-time (JIT) compiler, allowing JVM-based languages such as Java, Scala, Kotlin, Groovy and Clojure to be run, as well as a static, AOT compiler, used for creation of native images. Graal's modular architecture allows the reusability of common components; furthermore, taking the advantage of straightforwardness and affability of developing a compiler entirely in Java.

Truffle [8] is a framework used for implementing managed languages that are to be partially evaluated by the Graal compiler, thus allowing natural polyglotism within the system. It uses Abstract Syntax Tree (AST) as a simple way to interpret the program, but it is sometimes slower than expected and expensive. GraalVM recognizes interpreters written using the Truffle and is able to translate ASTs in highly optimized native executables. In order to eliminate overhead created by using cross-language scripts, any number of required ASTs can be joined together and interpreted as one [15].

Finally, on the top of the stack is Sulong, a compiler infrastructure engine for bitcode execution used for low-level programming languages interpretation on GraalVM, supporting common languages like C, C++, Fortran, Rust, and Swift, hence further extending GraalVM's versatility and interoperability.

[4] GraalVM, https://www.graalvm.org/
[5] Quarkus, https://quarkus.io/
[6] Graal Project, http://openjdk.java.net/projects/graal/
[7] JEP 243: Java-Level JVM Compiler Interface, http://openjdk.java.net/jeps/243
[8] The Truffle Language Implementation Framework , http://www.ssw.uni-linz.ac.at/Research/Projects/JVM/Truffle.html

B. Native Image mechanism

In order to handle microservice requests within the cloud, the deployment of work processes requires an initialization subsystem, which would reduce the traditional system's startup time and memory footprint. One approach is Checkpoint/Restore In Userspace (CRIU) [9] used for freezing an application, saving it to persistant storage and when required, restoring and running application from that point. One of the issues with the checkpoint/restore concept is that it does not reduce the memory footprint that is systemic due to dynamic class loading and optimization, necessary for HotSpot VM memory needs for bytecode, dynamically compiled code, and metadata. GraalVM offers its Native Image mechanism completely written in Java, nevertheless offering applicability to other managed languages. A sequence of complementing mechanisms, such as point-to-analysis, build time initialization, heap snapshotting, and AOT compilation, achieves the preservation of integral Java benefits [7]. The complete system used for native image creation is named *image build time* and presented in Fig. 3.

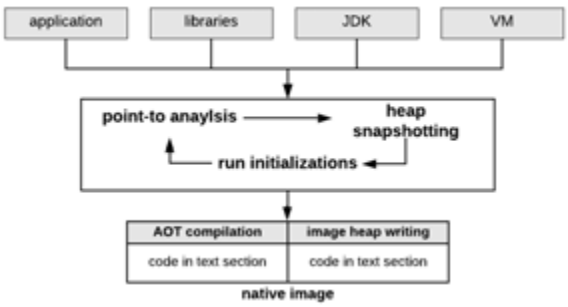


Figure 3. Image build time process

The aggressive static *point-to analysis* iteratively finds all the parts of the application which are reachable during runtime before a fixed point is reached. As analysis performs detailed processing, every field, method, and class is tracked individually, hence all parts of libraries on which the application depends are not included, but rather the necessary parts. The point-to analysis utilizes the frontend of the Graal compiler, using Graphbuilder [12] to turn Java bytecode into Graal Intermediate Representation (Graal IR) [7].

Execution of the code initialization starts when there are no more types to be added to type states, that is when the local fixed point is reached. The defining of the time at which classes are initialized in regards to image build time is up to the developer. Accordingly, information from methods, types or fields marked as reachable by point-to analysis facilitates the build of data structures that are optimized for reachable pieces of the application.

Afterward, the *heap snapshotting* mechanism searches for the objects created during the initialization and builds an object graph. Iterative analysis re-takes point-to analysis including heap snapshotting, until a point where the system does not find changes compared to the previous iteration. The latest graph instance is used to assume the role of image heap, and these objects are consequently serialized and placed into the data section of native image [7].

Lastly, the compiler used in the AOT Compilation mechanism is not the equivalent to the Graal compiler described in [9], yet it is its adjusted counter-part. The Graal compiler has modular architecture [16] and speculative compiler optimizations, which require additional consecutive deoptimizations to be run. However, deoptimizations cannot be used for the AOT compilation; henceforth, the compiler adapts point-to analysis results to optimize code performance, as presented in Fig. 3. Consequently, the part of the image heap was stacked during this process resulting in an executable image with a pre-populated heap.

C. Quarkus framework

Native and serverless environments require additional adaption of containers and native images as larger applications consume more memory thus inducing significantly higher cost. Further enhancement of the Java system is supported by Quarkus, a Kubernetes Native Java stack framework, designed to work both with HotSpot VM and GraalVM. As Quarkus presents cloud-optimized Java libraries, constraints presented with the translation of Java into native images are reduced, thereby benefiting the image creation mechanism [17].

The configuration and automatization properties of Quarkus improves developers' workflow and productivity as Quarkus and its extensions can be easily configured. While creating an application, Quarkus' development mode offers automatic detection of changes and updates to the application. Thus, Quarkus opens HTTP request, reboots the application, presenting immediate changes accordingly allowing the developer to save time compared to common server solutions [17]. As one of the extensions, Quarkus uses Undertow[10], a flexible performant web server written in Java that further enhances the development process. It contributes to the usage of microservices as well as asynchronous programming in order to endure the modern standards of business applications which require high responsiveness.

The security of Quarkus applications is presented with extendability and the ability to reuse certain good practices. With the Quarkus core system, a security base extension is provided, which gives concrete implementation choices. A developer can choose to manually handle data by using properties file Security Extension. The next level of protection is the Java Database Connectivity (JDBC) Security Extension, setting data into developer maintained database. Nevertheless, the suggested practice is to delegate the security to a third-party provider that would handle key business components such as authentication, authorization, and user management.

Ultimately, Quarkus is a stable platform as it uses features provided by very stable ecosystems comprising of Hibernate ORM, Eclipse Vert.x, Netty, and RestEasy. It provides a unified configuration with a fast live reload, currently making it one of the leading platforms for effective Java solutions running in serverless environments.

[9] Checkpoint/Restart in Userspace (CRIU), https://www.criu.org/

[10] Undertow, http://undertow.io/

D. Kubernetes platform

Kubernetes is one of the popular containerization platforms used for microservice application bundling, enabling continuous development, deployment, and integration, as well as DevOps-like separation of concerns. This stable continuously building container image platform is able to create application container images at build time; however, as a result of running on a container level rather than on a traditional, hardware level with PaaS systems. Nevertheless, it provides key features for managing the deployment of a given system such as service discovery, storage orchestration, and load balancing with additional security and configuration mechanisms.

E. Other competitors: Micronaut and Oracle Helidon

Helidon[11] is an open-source set of Java libraries for writing microservices that offers two different programming models: Helidon SE and Helidon MP. Namely, Helidon SE is a microframework that supports the reactive programming model, while Helidon MP is an Eclipse MicroProfile based, allowing the Jakarta EE (former Java EE) style of programming and running microservices with CDI, JAX-RS, and JSON-P in a portable way.

Micronaut[12] is, similarly, a complete polyglot full-stack application framework for any type of application, while being focused on microservices and serverless applications. It uses AOT for all dependency and configuration injections for precomputing annotation metadata during compilation time, consequently removing all the metadata work at runtime. Hence, anything that is framework infrastructure is performed at compilation time to avoid higher memory and startup time at runtime.

III. The Benchmark and Preliminary Results

A. The benchmark suite

For this research we selected the Renaissance benchmark suite[13], composed of modern, real-world, concurrent, and object-oriented workloads that exercise various concurrency primitives of the JVM. This new open-source suite proposes a representative set of 21 benchmarks that bring into focus concurrency and parallelism paradigms, as well as newer language and JVM features, such as Java Lambdas, invokedynamic, and method handles [18], giving it the advantage over other benchmark suites like DaCapo[14] or SPECjvm2008[15]. Its main intention is to optimize JIT compilers, interpreters, and garbage collectors for tools such as profilers, debuggers, and static analyzers, by aggregating prevalent modern JVM workloads, including Big Data, Machine Learning, and functional programming. We selected a set of several benchmark tests from Renaissance Suite version 0.10.0 to gain a better point of view on the benefits of GraalVM and standalone native executables.

The benchmark tests are described in [18] and we selected the following in our testing process:

- *future-genetic* – genetic algorithm function optimization using the Jenetics library and futures
- *philosophers* – a variant of the dining philosophers problem using ScalaSTM
- *movie-lens* – movie recommender using the ALS algorithm
- *par-mnemonics* – solves the phone mnemonics problem using parallel JDK streams

This set of benchmark tests we found to be the most suitable for our academic testing environment while at the same time offering a wide spectrum of paradigms, enabling us to obtain relevant information in numerous scenarios.

Most commonly, benchmark tests have a warm-up stage, which takes either a certain number of iterations (Renaissance, DaCapo) or execution time (SPECjvm2008) in order for the JVM to stabilize. Only after that criterion is met we can obtain the credible benchmark results from the execution, called the steady-state.

We have also implemented a RESTful API system[16] featuring CRUD operations in several modern frameworks, including Spring Boot (release 2.2.4), Quarkus (version 1.1.1), and Micronaut (version 1.2.10). Each of the implementations operate with a MySQL database through the Hibernate framework on a table. The table data was randomly generated and consists of approximately 250 rows. The applications have a set of 3 endpoints exposed, namely a GET mapping to /products, which retrieves a collection of products, a GET mapping to /products/{id}, which retrieves a specific product determined by its unique identifier, and a POST mapping to /products which creates a new product. The REST applications have been hit with 100 requests before collecting the results and calculating the average result of both the Resident Set Size (RSS) and response times.

B. Testing environment configuration

The benchmark tests were obtained from our testing environment consisting of several different JDKs:

- OpenJDK 64-Bit Server VM (AdoptOpenJDK) (build 25.242-b08) in mixed mode OpenJDK version 1.8.0_242
- OpenJDK 64-Bit Server VM (AdoptOpenJDK) (build 13.0.1+9) in mixed mode with the same javac version
- OpenJDK 64-Bit Server VM GraalVM CE 19.3.1 (build 11.0.6+9-jvmci-19.3-b07) in mixed mode with OpenJDK version 11.0.6

The operating system utilized for testing was Mac OS X 10.15.2 x86_64, with 2.3 GHz Quad-Core Intel Core i5 and 8 GB 2133 MHz LPDDR3 RAM.

---

[11] Helidon, https://helidon.io/
[12] Micronaut, https://micronaut.io/
[13] Renaissance Suite, https://renaissance.dev/
[14] DaCapo benchmark suite, http://dacapobench.org/
[15] SPECjvm2008 benchmark suite, https://www.spec.org/jvm2008/
[16] RESTful API source code, https://github.com/dmuharemagic/rest-api-benchmark-mipro

C. Selected test cases and discussion of results

Our preliminary test results, as presented in Fig. 4., were gathered in the steady-state after the warm-up has occurred and the Renaissance benchmark tests were stabilized, thus recording and observing the best execution times. The figure presents the average test time expressed in milliseconds (ms), and each benchmark test was performed in 250 iterations. The tests were run on OpenJDK 8, OpenJDK 13, and Graal CE 19.3.1 environments, with their individual respective versions being referenced in the previous segment.

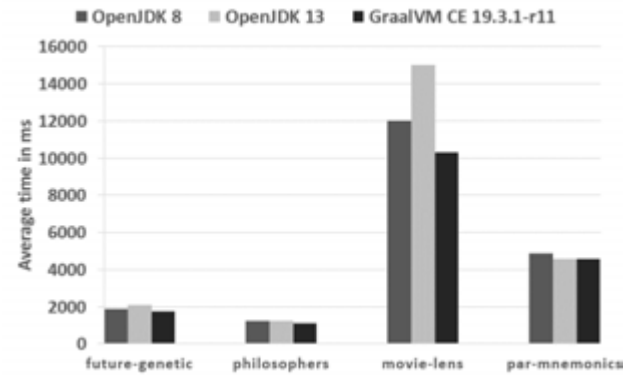


Figure 4. Preliminary benchmark test results for selected Renaissance benchmarks on Graal CE, OpenJDK 8, and OpenJDK 13

In the scenario of running the *movie-lens* benchmark test, Graal CE outperforms OpenJDK 13 by as much as 31.22% in terms of the decreased amount of running time. When speaking of f*uture-genetic* and *par-mnemonics* benchmarks, Graal CE performs better by a small difference in comparison to both OpenJDKs 8 and 13. Regarding the *future-genetic* benchmark test, Graal CE seems to surpass OpenJDK 8 by 6.95% and OpenJDK 13 by 15.18% in terms of speed.

Preliminary test results illustrated in Table 1. demonstrate the overall advantage of running native images in comparison with a traditional cloud-native stack. When measuring the memory footprint of our RESTful applications, in order to obtain complete results, we measure RSS, rather than the JVM heap size itself. Besides allocating native machine memory for the Heap, the total allocated memory for an application may also include class metadata, thread stacks, compiled code and garbage collection (varying on the JVM implementation). A standalone executable produced by GraalVM shows an 82.4% decrease in the usage of RSS memory in comparison to the same application implemented in Spring Boot running OpenJDK 13. Likewise, it is interesting to see that Micronaut outperforms Quarkus when running OpenJDK 13 by 22.92% in terms of memory allocation.

TABLE I. RESIDENT SET SIZE (RSS) IN SIDE-BY-SIDE COMPARISON LISTING COMMON MODERN FRAMEWORKS

Benchmark	Resident Set Size (RSS) in kilobytes
Quarkus + Native (via GraalVM)	31404
Quarkus + OpenJDK 13	171940
Micronaut + OpenJDK 13	132536
Spring Boot + OpenJDK 13	178484

Similarly, the results presented in Figure 5. manifest the massive gain in response time when querying our RESTful API for a collection of objects. Quarkus running as a standalone native image produced by GraalVM runs 81.74% faster than the Spring Boot application with the same implementation. Quarkus implementation running on OpenJDK 13 runs 17.65% faster than the same Spring Boot application. Once again, Micronaut seems to outperform by 16.82% in terms of faster response times when running OpenJDK 13.

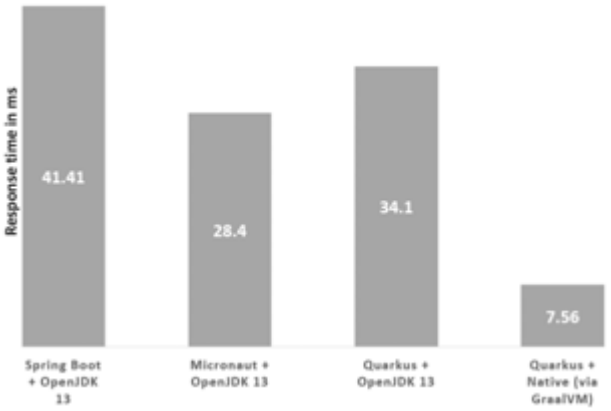


Figure 5. Response time in milliseconds obtained from a custom benchmark test (featuring a small RESTful API with basic CRUD operations) in selected frameworks

IV. FUTURE WORK

This paper presents our continuous research [9] related to GraalVM, and in this paper we focused more on an AOT compiler. We believe that more features will further improve its extensibility and thus, industry usage. Serverless and microservice-based software systems are becoming the standard and will possibly take over the majority of network-based applications. Eventually, more focus will be needed in security aspects, as well as hardening the newly created system for improved resilience. As noted above, there are several competitors which we intend to compare in more detail together with OpenJDK 14 (in Early Access) and others. Additionally, other AOT compilers and cloud providers with adding new features, such as JVM-level threads named fibers, try to further speed up our system, and we plan to cover them in our future research.

V. CONCLUSION

GraalVM's progress though years is visible and noticeable. In our first research [9] we mainly focused on testing and comparing GraalVM to the standard JDK version without deployment to server environments results confirmed our assumption and GraalVM's performance was surpassing competitor's. In this paper we presented preliminary results of our tests conducted considering GraalVM's native image compilation performance in combination with Quarkus gave good results, showing significantly reduced RSS.

GraalVM preliminary results on newly created benchmarking tool Renaissance, which focuses on

increased service workload and asynchronous programming practices, slightly exceeds the current competitors. Since version 19 was released, we consider GraalVM ready for usage in the production environments, additionally as Oracle labs claim its stability and maturity, particularly in GraalVM EE.

REFERENCES

[1] N. M. Josuttis, "SOA in Practice: The Art of Distributed System Design", O'Reilly Media, Inc., 2007.

[2] P. Jamshidi, C. Pahl, N. C. Mendonca, et al., "Microservices: The journey so far and challenges ahead," IEEE Software, vol. 35, issue: 3, pp. 24-35, 2018.

[3] P. Mell, T. Grance, "The NIST Definition of Cloud Computing," NIST Special Publication 800-145, 2011.

[4] R. Ch Chakradhara, M. Leelarani, and Y. R. Kumar, "Cloud: Computing Services And Deployment Models," International Journal of Engineering and computer science, pp. 3389-3392, 2013.

[5] D. Rani, and R.K Ranjan, "A Comparative Study of SaaS, PaaS and IaaS in Cloud Computing," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, Issue 6, June 2014.

[6] M. Sudha, G. M. Harish, A. Nandan, and J. Usha, "Performance Analysis of Kernel-based Virtual Machine," Int. Journal of Computer Science & Information Technology (IJCSIT), vol. 5, no. 1, 2013.

[7] C. Wimmer, C. Stancu, P. Hofer, et al., "Initialize once, start fast: application initialization at build time," Proc. of the ACM on Programming Languages 3, (OOPSLA), Article 184, pp. 1-29, 2019.

[8] E. Barrett, C.F. Bolz-Tereick, R. Killick, et al., "Virtual machine warmup blows hot and cold," Proceedings of the ACM on Programming Languages 1, (OOPSLA), Article 52, pp. 1-27, 2017

[9] M. Šipek, B. Mihaljević, and A. Radovan, "Exploring Aspects of Polyglot High-Performance Virtual Machine GraalVM," Proc of 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1671-1676, May 2019.

[10] I. Voras, B. Mihaljević, M. Orlić, et al. "Evaluating open-source cloud computing solutions, " Proceedings of the 34th International Convention MIPRO, pp. 209-214, May 2011.

[11] I. Voras, M. Orlić, B. Mihaljević, "An early comparison of commercial and open-source cloud platforms for scientific environments," Lecture Notes in Computer Science (LNCS). Subseries Lecture Notes in Artificial Intelligence (LNAI), 7327, pp. 164-173, June 2012.

[12] L. Stadler, "Partial Escape Analysis and Scalar Replacement for Java," Proc. of Annual IEEE/ACM Int. Symposium on Code Generation and Optimization CGO '14, pp. 165-173. February 2014.

[13] H. Grgić, B. Mihaljević, and A. Radovan, "Comparison of Garbage Collectors in Java Programming Language," 41st Int. Conv. on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2018), pp. 1539-1544, May 2018.

[14] P. Pufek, H. Grgić, B. Mihaljević, "Analysis of Garbage Collection Algorithms and Memory Management in Java," 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1677-1682, May 2019.

[15] C. Wimmer and T. Wurthinger, "Truffle: A Self-Optimizing Runtime System," Proc. of the 3rd Annual Conf. on Systems, Programming, and Applications: Software for Humanity SPLASH'12, pp. 13-14, October 2012.

[16] G. Duboscq, L. Stadler, T. Würthinger, et al. "Graal IR: An Extensible Declarative Intermediate Representation," Proc. of the Asia-Pacific Programming Languages and Compilers Workshop, pp. 1-9, February 2013.

[17] M. Bizjak, "Mikrostoritve v ogrodju Quarkus," Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, 2019. Available:https://repozitorij.uni-j.si/IzpisGradiva.php?id=110312 [Accessed: 2020-01-20]

[18] A. Prokopec, A. Rosà, D. Leopoldseder, et al. "Renaissance: benchmarking suite for parallel applications on the JVM," Proc. of 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, pp. 31-47, June 2019.

Achieving Efficient Structured Concurrency through Lightweight Fibers in Java Virtual Machine

P. Pufek, D. Beronić, B. Mihaljević and A. Radovan
* Rochester Institute of Technology Croatia, Zagreb, Croatia
paula.pufek@mail.rit.edu, dora.beronic@mail.rit.edu,
branko.mihaljevic@croatia.rit.edu, aleksander.radovan@croatia.rit.edu

Abstract - **Contemporary concurrent server applications, commonly built of smaller and independent services, are using concurrent threads to serve many incoming requests and often have to perform under excessive load. Those applications are relatively easy to develop in general-purpose, imperative programming languages such as Java, and have great tooling support. However, such applications are not easily scalable, mostly due to relying on oversized OS kernel threads, which can be created only in a limited number on finite hardware resources. Furthermore, heavyweight OS threads are implemented so robustly and generically to support various usage scenarios, and are, therefore, wasteful in resources and often inefficient in addressing specific application demands.**

Those challenges in software development resulted in various asynchronous programming techniques. This paper presents an exploration of a novel structured concurrency model in the Java Virtual Machine (JVM), introduced within OpenJDK's Project Loom. It is focused on the exploitation of fibers, new lightweight implementation of virtual threads within the JVM depending on delimited continuations. Furthermore, on several applications' benchmark cases, we analyzed performance with traditional threads and new fibers in different configurations. Finally, we discussed the current challenges of implementing fibers as a feasible approach for the more efficient future of Java.

Keywords - Structured Concurrency, Fibers, Java, Java Virtual Machine, Threads, Delimited Continuations

I. INTRODUCTION

One of the most important aspects of writing applications that concurrently respond to a large number of incoming requests is the achievement of optimal efficiency. Applications that provide such services could be developed with simple, blocking, synchronous code or might use asynchronous libraries that are more scalable but harder to implement, profile, and debug. In both cases, kernel threads are assigned for the services and are frequently wasteful in resources. To increase the extent of optimization in terms of scalability and preserving the utilization of external memory, software developers should not depend on the underlying system's memory allocation. Moreover, to demonstrate more convincing correctness and manageability, the internal structure of a program should not be treated as a black box [1].

Several programming languages and runtimes started exploring *Structured Concurrency* through independent mechanisms for asynchronous programming. Structured Concurrency might be defined as an organization of computations in a way that the limited hardware powers are sufficient to guarantee that computations will establish the desired effect [2]. For example, in Python, asynchronous I/O library named Trio[1] implements scopes of nurseries that launch coroutines, and in C programming language scopes that use bundles of coroutines are implemented within *libdill* tool[2].

The concept of Structured Concurrency implements continuations that represent a form of control flow that can be captured and manipulated [3] within a given scope with the aim of reducing the limited number of OS threads allocated for each control flow in a program. Moreover, modern programming languages such as Java, Scala, and Python currently support different systems for achieving concurrency.

In contrast to the common threads model, the *Actor* model in Scala implements asynchronous communication and enables structuring controls with Actors [4]. In this case, Actors are treated as primitives of concurrent digital computation, can send messages to other Actors, create new Actors, and decide how to handle the next method call. All of these actions might be executed concurrently, as the order of the system's tasks is not determined in advance. Another example is the *coroutine* concept thoroughly explored and used in Python, which supports the persistence of the coroutine's data between calls while temporarily suspending the execution flow and continuing where it was stopped when it is required for later use. However, to be able to implement a similar approach for modeling concurrency in Java, each coroutine might be assigned to a heavyweight thread [5]. That could consequently make a system less scalable as more operating system-native threads are required to efficiently perform the operation that might be expensive with regard to memory usage. With the intention of returning back to a simple, synchronous way of writing the code, without requiring such a great amount of power, a novel concept

[1] Trio: a friendly Python library for async concurrency and I/O, https://trio.readthedocs.io/en/stable/
[2] libdill: Structured Concurrency for C, http://libdill.org/

of *Continuations* might promise efficient flow of controls with trivial synchronization [6].

II. Novel Concurrency Concept in Java

To solve the problem of heavy-load thread utilization in Java, an early prototype is proposed within OpenJDK's Project Loom[3], which investigates Java Virtual Machine's (JVM's) features and APIs with the intention to support high-throughput lightweight concurrency within the Java platform. It suggests the increase of the number of lightweight threads, so-called *fibers*, scheduled in the JVM runtime so that the program does not rely directly on the availability and sufficiency of heavyweight OS kernel threads. In contrast to threads, fibers allow organizing concurrency in a coherent manner as efficient higher-level constructs built on top of continuations, which are implemented within the JVM.

In Java, the Continuation object commonly is associated with a Runnable object that represents the body of continuation, and a *ContinuationScope,* which allows nesting different continuations inside each other. When Continuation, which has its own stack, is executed, it either runs until it is not completed or yields. If it yields, it is returned to the caller, but since it remembers its state the next time it is run it continues from where it left off. Continuations are not supposed to be used directly, thus FiberScope API has been implemented to enable utilization of continuations.

III. Structured Concurrency

The main idea behind structured concurrency [2] is the ability to ensure that every time an application concurrently runs any number of tasks required by a developer, they all execute successfully and join up together after completion of each. With structuring the layout of the code to reflect the structure of execution flow, software developers might be able to make assertions about the next computations and control all concurrently running tasks that are working on a process.

Similarly to structured programming, structured concurrency structures the code as a set of nested code blocks that do not overlap [7]. To achieve more efficient control flow, the code is structured so that the tasks are bound to a scope with predefined start and end of operation's execution. If any of the tasks within the container (scope) are unsuccessful, the whole scope is canceled, and the stack frame where the scope was opened reports an exception. If several threads are performing a single operation, a developer should have an ability to know where each part of the task is executed within the code – which thread performs which operation, with the aim to enhance overall performance and decrease the possibility of exceptions.

A. Coroutines

Thus far, several variants of coroutines have been implemented in programming languages, under different names such as generators, fibers, and green threads. As they do not have to be synchronized and they explicitly pass control to each other, typically by invoking *yield* method [8], several high-level abstractions have been developed to make use of these low-level features.

The Abstract Behavioral Specification (ABS) modeling language, presented in [9], integrates actor-based programming with futures and coroutines. In the ABS model, the future is generated when a method is called asynchronously. It can be passed via actors that hold a reference to it and can check whether the corresponding method has completed and get a return value. In the case of the method's failure, the coroutines allow explicit suspension statements that enable multiple control flows in only one actor. Java library JAAC[4], which generates programs from ABS sources and provides a scalable implementation of asynchronous programming, currently supports coroutines that operate on the source and bytecode level in the JVM. At the source-code level, every thread could be allocated to each coroutine, which might not be sufficiently scalable. However, the architecture of this system allows an actor to run multiple tasks within its stack by storing its tasks in a queue. Each thread – actor, is able to select a task from its queue and run it as a separate process, concurrently with already running processes.

To adapt to JVM's specification of automatic memory management [10], the Garbage Collection algorithms had to be modified to support the implementation of lightweight threads. This process requires searching through the whole heap and finding all root references of live objects when freeing up the unused memory [11]. Thus, when a coroutine is located on the thread's stack, coroutine's stack is searched in exactly the same manner as threads' stacks are, and is investigated to check if there are parts of the heap ready for reclamation [8].

B. Continuations

The Continuation can be interpreted as the state of a running thread that can be captured and altered. It is commonly referred to as "*the rest of the computation*" [12] as it runs from the point in time when it was instantiated. According to [3], there are three categorizations of continuations:

- *First-class continuations* – fully store computation's state and can be reinstated many times
- *Delimited continuations* – part of continuation used to perform a function that can be invoked multiple times
- *One-shot continuations* – act as exit statements and can exist at multiple method scopes at once

Delimited continuations might be considered as functions that can be called multiple times. Furthermore, they are only paused for a particular amount of time and do not span throughout the entire life span of a program [14].

The two main functions of continuations are capturing (creating a continuation and storing the dynamic

[3] Project Loom, https://openjdk.java.net/projects/loom/

[4] JAAC, https://github.com/JaacRepo/JAAC

computation context of the currently running thread) and resuming (restoring the stack to exactly the same state as when the continuation was captured). The capture method associates the continuation with the currently running thread and multiple method invocations might be made on the same continuation object. The method resume is allowing a continuation to be reinstated in case if the capture method has already been invoked on that particular continuation object. If not, a runtime exception is thrown.

C. Fibers

Currently, Fibers are experimentally implemented as the JDK library on top of delimited Continuations combined with Java Schedulers in the Project Loom. Fibers are structured as they are constrained to FiberScope API; thus, every fiber is alive as long as the scope that created it, and multiple Fibers can be started that will be managed by the JVM. As soon as scope is exited, the close method is automatically called on FiberScope and it blocks every other method call until all fibers that are inside the block are terminated.

This core concept of Fibers allows the usage of structured concurrency in a way that all next operations would have to wait until Fibers scheduled in a particular scope have all finished their tasks. As expected, after each fiber has executed its assigned task, the following operations can proceed.

However, there are plenty of differences between Threads and Fibers. In the implementation of lightweight threads in the Quasar library[5], each fiber occupies approx. 400 bytes of RAM and decreases the use of CPU when switching the tasks. A fiber might have a return value, but implementation in the Project Loom also supports *Fiber<Void>* for type *V* which returns *null*. Furthermore, the *run* method is allowed to throw an exception, either InterruptedException or SuspendException; however, thread's run method does not throw any exceptions. Just as a thread might have a higher or lower priority, we are able to decide the order and importance of fiber's tasks by structuring the code with making efficient use of several Fibers in one or multiple FiberScopes.

As Fibers are scheduled by Java schedulers, they do not have to be Garbage Collection (GC) roots; thus, each Fiber does not have to be bound to a new thread. Fibers are runnable or are in a state of waiting – blocked. In both cases, their reference is held by a scheduler, and in case of blocking, the reference points to the object which caused the blocking. Existing Java schedulers, including ForkJoinPool and ThreadPoolExecutor [6] serve as Thread schedulers. Additionally, third-party optimized schedulers for implementing fibers were developed to enable manipulation of stacks' calls within JVM. The default scheduler is an object of FiberForkJoinScheduler that schedules Fiber in ForkJoinPool. When scheduling fibers in a thread pool of a particular architecture or performing operations on special, event dispatching thread and not daemon thread, such as Swing, FiberExecutorScheduler might be used.

[5] Quasar, http://docs.paralleluniverse.co/quasar/

IV. Testing Environment

A. Testing Environment Setup

The testing environment which hosted the virtual machine on which we have performed the testing was Windows 10 64-bit OS with an x64-based processor and 16 GB of RAM. The tests were run on Ubuntu 18.04.3 64-bit Virtual Machine with a base memory of 9 GB. The testing environment relies on OpenJDK Runtime Environment, OpenJDK 64-Bit Server VM (build 11.0.6+10-post-Ubuntu-lubuntu118.04.1, mixed-mode, sharing). However, for the JDK version that supports the utilization of Fiber and FiberScope API, integrated within Project Loom build, we used the early access OpenJDK 64-Bit Server VM (15-internal build mixed mode, sharing).

B. Test Cases

To record and compare the differences between each runtime, we used Vegeta[6], an open-source command-line utility application written in Go programming language, which provided the detailed reports on requests and responses. Vegeta is a load testing HTTP-request based tool and library, built to consume HTTP services by constantly sending requests to the given addresses. Also, we used the implementation of the HTTP server provided within the JDK[7], bounded to a given IP address and port number and listens for the incoming TCP client connections. Dealing with requests within the implementation of this virtual server is supported by HttpHandler, which is associated with the HttpServer object. Each server could be assigned one or multiple handlers which are mapped to the server via HttpContext object.

V. Preliminary Results

After warm-up pre-runs, which were repeated 3 times and had a duration of 10 seconds, we were able to record the results and maintain the stable behavior of the program execution of tests with both fibers and threads. For the purposes of our research and to find the optimal cases for thread and fiber performances, we analyzed several configurations of test parameters.

We maintained the optimal load to test their performances and experienced the most conclusive results by declaring a request rate per time unit value of 75, which represents a number of clients trying to connect to the server in 1 second. The actual request rates slightly varied due to internal JVM's processes such as the Garbage Collection process. The internal number of CPUs of the virtual machine was 3. Additionally, we specified the constant number of 64 workers which are used in the attack to control the level of concurrency as otherwise they would be increased as needed by the JVM.

[6] Vegeta, https://github.com/tsenart/vegeta

[7] HttpServer, https://docs.oracle.com/en/java/javase/13/docs/api/jdk.httpserver/com/sun/net/httpserver/HttpServer.html

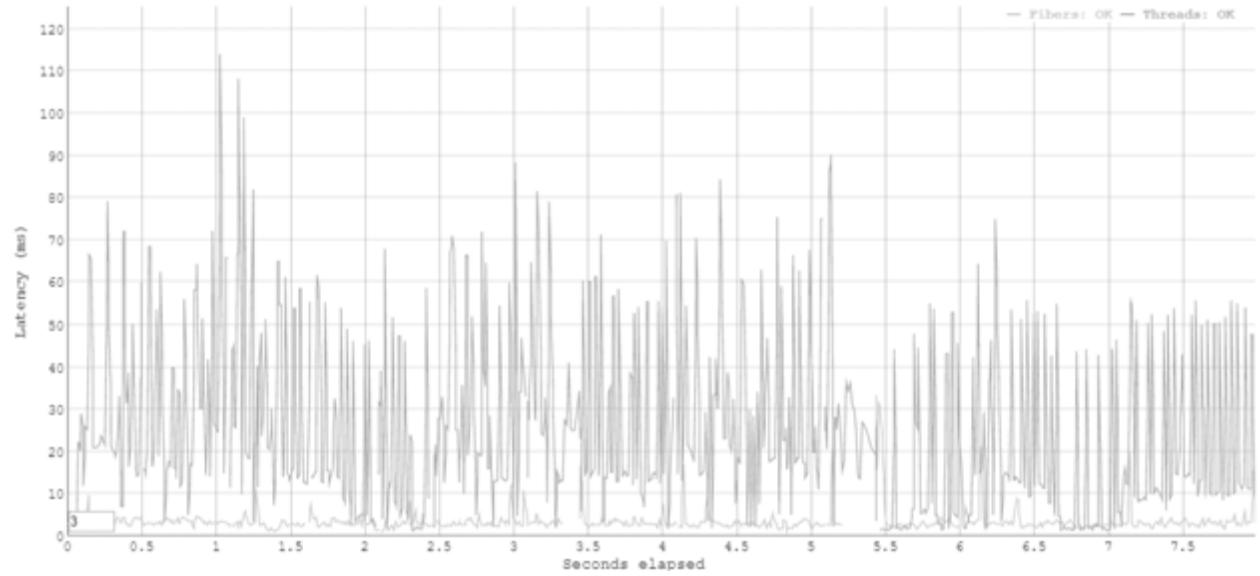


Figure 1. Variations of latencies in attacks of 8 seconds duration with scheduled fibers and threads (1 millisecond sleep)

Both threads and fibers were assigned one Runnable object, which represented a task that they would perform. In our first test case, the body of Runnable set a Thread to sleep for 1 millisecond. Similarly, the wait method was invoked on Fiber which correlates with the sleep method invoked on Thread object. In the second case, Thread and Fiber waited for 10 milliseconds. The sleep period could represent a task with constant duration. With this kind of test that performs a simple task, we were able to investigate the speed of scheduling and execution by the JVM and underlying OS.

In the figures, the vertical axis represents latency in milliseconds, and the horizontal axis represents the duration of each request, which was 8 seconds for every run. Based on the conducted results, the total execution of the task with fibers had on average 6.15 times lower latency than threads in heavier-load tests (with passing the greater number as a parameter to either sleep or the wait method) and around 1.02 times faster on tests where the duration of waiting was 10 times shorter. On average, fiber latency is approximately 16.26% of thread latency, which makes them more efficient in heavier-load cases. In both cases, fibers' amounts of latencies are more constant than threads and do not have significant variations in latency.

In the tests with shorter sleep periods, fiber's startup period is significantly longer, and more time is needed for them to reach stable execution. However, in a heavier-load test case, the startup difference between threads and fibers at the beginning of the runtimes is significantly lower. According to these results, we might conclude that fibers are able to handle an extensive workload more efficiently in comparison to lightweight scheduled tasks.

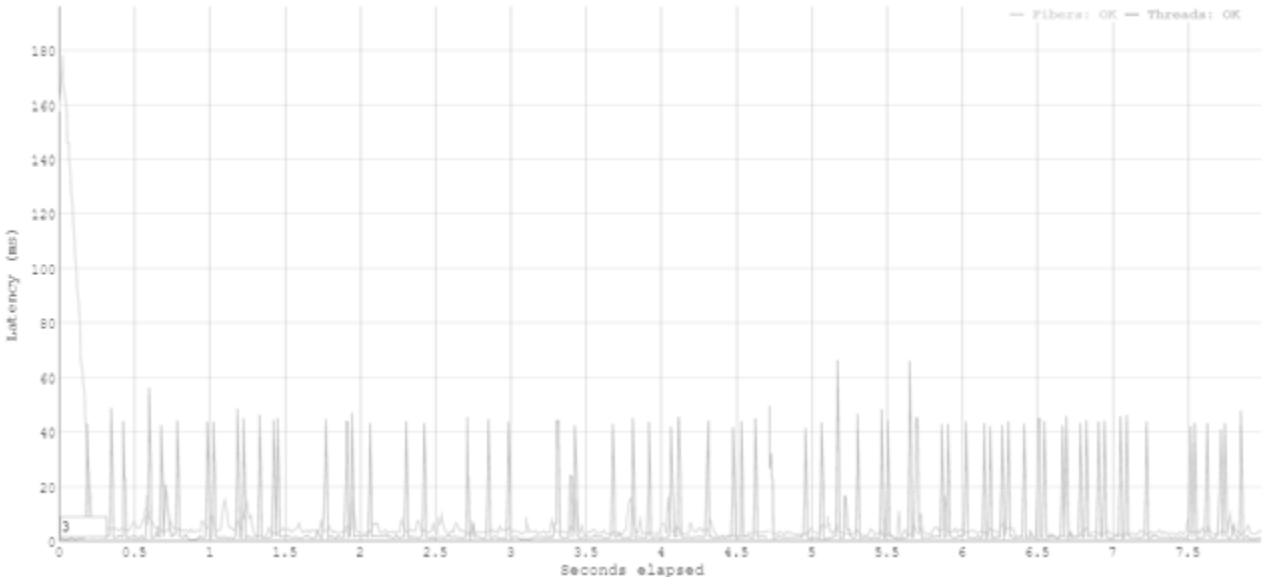
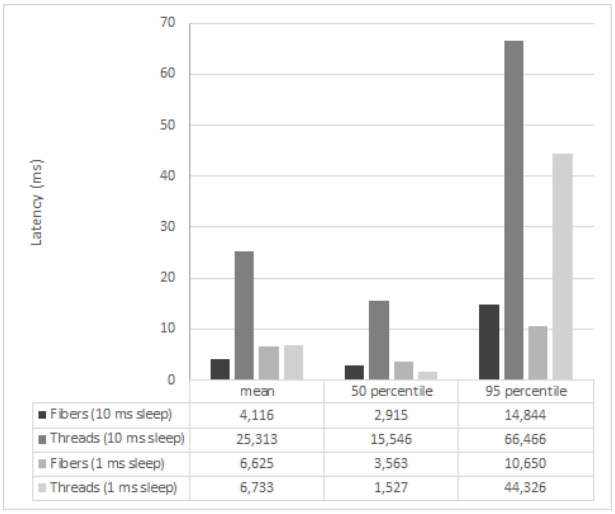


Figure 2. Variations of latencies in attacks of 8 seconds duration with scheduled fibers and threads (10 milliseconds sleep)



	mean	50 percentile	95 percentile
Fibers (10 ms sleep)	4,116	2,915	14,844
Threads (10 ms sleep)	25,313	15,546	66,466
Fibers (1 ms sleep)	6,625	3,563	10,650
Threads (1 ms sleep)	6,733	1,527	44,326

Figure 3. Amount of latency for fibers and threads (with 1 and 10 milliseconds sleep) - mean value, 50th percentile, and 95th percentile

When compared to different amounts of time taken for the waiting for responses in all test cases, the arithmetic mean of the latencies of all requests in an attack was the longest in case of threads execution with a sleep of 10 milliseconds. The average latency rises up to 20 ms, which takes the longest period of time to receive responses. In comparison to fibers, scheduled to wait for an equal time period of 10ms, fibers performed remarkably better with an average latency of 4.116 ms.

The only test case in which threads outperformed fibers is visible in the case with regard to the maximum duration of waiting for the responses in one attack, where latency rises up to 170ms in cases of fibers with a wait time of 1 millisecond. However, on the 50th percentile of all requests' latencies, the amount was the highest in the thread-case scenario in which each thread sleeps for 10 milliseconds. A similar report is visible on the 95th percentile, where threads scheduled to wait for 1 and 10 milliseconds had a notably higher amount of latencies.

The figures represent the average results of the results gathered after 3 iterations of each test case. The fibers had constant request rates with a quantity of 75.12, but with the increase of sleep duration by 10 times, in the case of threads, the rate increased by 1%. When performed lighter-load tests, with a shorter wait period, fibers on average performed slightly better than threads with average latency that is 0.11 milliseconds lower. However, heavier-load tests present results with a difference in the average latency of 21.2 milliseconds.

VI. DISCUSSION

In [14], Cuadrado and Aracil presented scheduling techniques for model-to-model transformations (M2M) with which they had performed reverse engineering based on continuations. Whenever a particular rule requires a target element that is not yet created, it saves its state in a continuation which might be resumed later, when available for use. During the explained event, other nondependent parts of transformation are being executed concurrently. Based on conducted performance testing, their model transformation based on continuations has shown to be efficiently scalable and brought useful methodologies to deal with language interoperability.

Stadler's, Würthinger's, and Wimmer's implementation of fibers in JRuby [8], the second most used implementation of Ruby programming language, uses the underlying system of coroutines that handles a great amount of workload during the applications' runtimes. The API for coroutines was introduced by modifying the Java HotSpot™ VM. This system uses two strategies that enable the implementation of coroutines within JVM: each coroutine is allocated a separate stack and the stacks are copied to support larger amounts of coroutines to be allocated for the execution of the necessary tasks. Minimal changes in JVM were required to achieve an efficient performance of this novel system, including the addition of stack management code and new code in the Garbage Collection system. Different benchmarks in Ruby environments were used to evaluate the outcome of the coroutine support. The results prove that JRuby's implementation with coroutines is 7 times faster than JRuby without them. Therefore, the Java API for coroutines has shown to be beneficial to this programming language implemented on top of the JVM.

A benchmark that performs recursive method calls in which each function captures and resumes a continuation was performed to measure the performance of the algorithm that lazily stores the contents of continuations in [13]. This approach is implemented for the interpreter and the compiler for Java HotSpot™ VM, but could be applicable in other Java VMs and might operate in runtimes of other programming languages. The results show that runtimes increase linearly with the addition of local variables handled within each continuation. Also, the duration of resuming a continuation is lower than capturing because there is no need for memory allocation. The algorithm for the execution of storing contents in continuation and resuming a continuation uses an optimized approach by restoring only stack frames that are not bound to the current thread's state. In comparison to continuation libraries that do not modify JVM, the performances of VM-based implementation of continuations are shown to be more efficient.

The coroutine support in the Abstract Behavioral Specification (ABS) system performs resuming of execution by the underlying scheduler, which provides the avoidance of errors [9]. With implementing such a feature, the explicit command to invoke the resuming of an operation is not allowed and the program quality is improved. Instead of creating a thread for every asynchronous function call, such calls are stored within Callable objects that represent tasks. The architecture consists of a thread-pool in which at most one thread is assigned to an actor (a lightweight representation of a thread), a queue assigned to each actor and each thread that holds an actor executes the main task that iteratively runs tasks from the assigned queue. The instances of ABSTask objects are created by calling send or spawn method and are stored in a queue of an actor which was generated to hold a group of tasks. When a particular task terminates, its return value is saved in the field of ABSFuture object which keeps track of completed tasks. ABSFuture implements a mechanism that notifies the

actors about the tasks' completions. A test-case example that creates a stack of synchronous, recursive calls was used to measure the impact of coroutine implementation on the overall performance. Storing the tasks in heap memory instead of using native threads for each function call resulted in a decrease of memory overhead.

As explained before, Fiber and FiberScope API is still experimental and under construction. However, after research about available solutions with similar approaches to structured concurrency in the form of concrete lightweight thread implementations, we believe the further development of such features might show promising results in future updated versions of JDK. As most of our tests' results showed a better performance of fibers in terms of low latency and higher throughput, we support the claim that suggests handling requests concurrently without reliance on the memory provided by the operating system. We believe that the cost of external resources might decrease by the replacement of the first version of threads with the integration of some ideas of these novel approaches into future Java releases.

VII. CONCLUSION

Due to the insufficient amount of support that Java receives from kernel native threads, this novel approach to achieving concurrency with the concrete implementation of JVM's internal threads in the form of fibers brings an appropriate solution for improving the performance. Based on conducted reports and our preliminary results, the system which incorporates these lightweight threads promises future progress in handling heavy workloads that might appear during the execution of multi-threaded server applications. Furthermore, Project Loom's features enable the code scalability and Fibers might allow Java programs to manage concurrency under the heavy-load in a more efficient way. Although this fiber-based mechanism is still declared as experimental, we believe that further development could be promising and definitely worth exploring. With this novel approach, we might be able to increase the level of optimization and reduce costs when providing services to a large number of users.

REFERENCES

[1] N. J. Smith, "Notes on structured concurrency, or: Go statement considered harmful," April 25, 2018. https://vorpus.org/blog/notes-on-structured-concurrency-or-go-statement-considered-harmful/

[2] E. W. Dijkstra, "Notes on Structured Programming," 2nd ed., T. H.-Report 70-WSK-03, April 1970. https://www.cs.utexas.edu/~EWD/ewd02xx/EWD249.PDF

[3] C. Tismer, "Continuations and Stackless Python," in Proc. of the 8th International Python Conference, vol. 1, 2000.

[4] Z. Li and E. Kraemer, "Programming with Concurrency: Threads, Actors, and Coroutines," in Proc. of the 2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and PhD Forum (IPDPSW), Boston, USA, May 2013, pp. 1304-1311.

[5] R. Malhotra, "Java Language and Ecosystem Recap," in Rapid Java Persistence and Microservices: Persistence Made Easy Using Java, EE8, JPA and Spring, Apress, Berkeley, 2019.

[6] R. Pressler, "Project Loom: Fibers and Continuations for the Java Virtual Machine," August 23, 2018. https://cr.openjdk.java.net/~rpressler/loom/Loom-Proposal.html

[7] M. Sústrik, "Structured Concurrency," February 7, 2016. http://250bpm.com/blog:71

[8] L. Stadler, T. Würthinger, and C. Wimmer, "Efficient Coroutines for the Java Platform," in Proc. of the 8th International Conference on the Principles and Practice of Programming in Java, pp. 20-28, 2010.

[9] V. Serbanescu, F. de Boer, and M. M. Jaghoori "Actors with Coroutine Support in Java," in Formal Aspects of Component Software (FACS 2018), K. Bae and P. Ölveczky, Eds. Lecture Notes in Computer Science, vol. 11222, pp. 237-255, Springer, Cham, October 05, 2018.

[10] P. Pufek, H. Grgić, B Mihaljević, "Analysis of Garbage Collection Algorithms and Memory Management in Java," in Proc. of the 42nd Internationa. Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2019, pp. 1677-1682.

[11] H. Grgić, B. Mihaljević, and A. Radovan, "Comparison of Garbage Collectors in Java Programming Language," Proc. of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2018, pp. 1539-1544.

[12] P. Sestoft, "Programming Language Concepts," 2nd ed., Undergraduate topics in computer science, Springer, Berlin, 2017.

[13] L. Stadler, C. Wimmer, T. Würthinger, H. Mössenböck, J. Rose, "Lazy Continuations for Java Virtual Machines," in Proc. of the 7th International Conference on Principles and Practice of Programming in Java (PPPJ '09). Association for Computing Machinery, New York, USA, 2009, pp. 143-152.

[14] J. S. Cuadrado and J. P. Aracil, "Scheduling model-to-model transformations with continuations," in Software: Practice and Experience, vol. 44, issue 11, November 2014, pp. 1351-1378.

Use of keystroke dynamics and a keystroke-face fusion system in the real world

J. Stavanja *, P. Peer * and Ž. Emeršič *

* Faculty of Computer and Information Science, University of Ljubljana, 1000 Ljubljana, Slovenia
E-mail: js8927@student.uni-lj.si, {peter.peer, ziga.emersic}@fri.uni-lj.si

***Abstract* - In the field of keystroke biometrics, many systems have been developed using more traditional approaches and evaluated based on their accuracies. However, with the rise in popularity of neural networks, the research in the recent years focuses mainly on deep learning approaches. We test the accuracy and maintainability of both types of classifiers and try to test a new vector distance approach using a custom distance metric to see if there could still be room for improvement in the field of more traditional systems for keystroke-based authentication. We also test a keystroke dynamics and face recognition fusion classifier to see if a multi-modal system is perhaps the best compromise in terms of maintainability and performance. We find that there is still room for research on more traditional keystroke rhythm comparison techniques and conclude that by fusing those together with a face recognition pipeline we can achieve very good results in web authentication systems.**

Keywords - keystroke dynamics, biometrics, biometry, face recognition, authentication

I. INTRODUCTION

Various techniques for identity recognition have entered wide stream popularity, e.g. face recognition, fingerprint recognition and others. However, the behavior-based recognition is arguably not that widely used, despite many advantages. This can be contributed to the fact that behavior-based approaches proved not to be that accurate and especially hard to introduce to users as shown in an article by Pisani and Lorena [1]. Their work also shows that some techniques do not perform well enough to comply with the European law. Their article also emphasizes the problem of identity theft and easy spoofing of typing rhythms between users. Recognizing people by analyzing keystroke dynamics is very accessible on desktop computers - it requires no additional hardware, whereas with the majority of other techniques at least some sort of camera or microphone is needed. To verify the quality of a more traditional approach, we need to test different methods and measurements for authentication via subjects' typing patterns. Then we can observe how maintainable and easy to setup and to use these systems are and conclude whether they perform well enough for real world use.

Usually, recognition systems based on keyboard dynamics work as verification systems: a binary classificator that accepts an input timing vector and the identifier of the person that the person who is logging in is acting as and return 0 if they detect a potential imposter and 1 otherwise. We can then implement different keystroke dynamics classifiers from other papers, put them side by side and see how well they perform and how maintainable they are. We start by researching and implementing some of the classifiers that aren't based on deep learning approaches and a good one that is and create a table of accuracies using a certain evaluation technique (in our case, the Area Under Curve). We can also choose different features from the dataset to see what kind of impact they have on the final scores. Thus, we can establish if any of the keystroke dynamics approaches are secure and are performing enough for professional use. Another possibility is to add other approaches for biometric authentication and make a fusion of systems to try and improve the accuracy of some biometric authentication software.

In this work we also present a case with the use of simple neural networks that offer superior performance to some of the more traditional approaches, while still retaining a small footprint.

A. Related work

Research in the field of keystroke biometrics started emerging in the early 1990s. The first significant paper that covered this field was published by Joyce and Gupta [2], where they constructed vectors of keystroke latencies and used simple distance metrics to construct a classifier. Another major milestone was a paper from Killourhy and Maxion in 2009 which featured a comparison of different distance metrics including a simple neural network approach. We can try to implement some of those algorithms to test the importance of feature selection. Due to the rise in popularity of neural networks, the research in the recent years focuses mainly on deep learning approaches. Research has also been conducted on multi-modal systems that include keystroke biometrics, the most cited being the paper from Montalvao et al. [3] which uses speech recognition as the second trait. However, only the paper by Giot et al. [4] features a keystroke and face recognition approach. We can experiment with a different distance metric for keystroke comparison in our multi-modal system and test its quality.

II. METHODOLOGY

As stated in the introduction, we research and implement some classifiers and test their performance. For the keystroke-only part of our experiments we use a publicly available keystroke timing vector dataset collected for Killourhy and Maxion's experiments. The dataset used to test our multi-modal system was collected during our previous work [5] however the original timing vectors were lost and what remains is only the ratio results table. All the backend code used for the calculations in our experiments is publicly available on https://github.com/jstavanja/quiz-biometrics-api while the plugin for Moodle that includes the frontend forms is available on https://github.com/jstavanja/quiz-biometrics-plugin.

A. Tools

We implement our classifiers and test others in Python along with some libraries (i.e. Numpy, Pandas, SciPy). We also implemented a web form where users can register his pattern and test out the system's quality.

B. The foundation and the framework

The dataset we use during our experiments was collected during Killourhy and Maxion's experiments. This dataset from the Carnegie Mellon university features 51 subjects, which have repeated the typing of a secure password 400 times each. Gathering a sufficient amount of data ourselves (at least the same amount as we can get from this dataset) is very hard and time consuming, so at this stage, we avoid doing that.

The idea for our experiment is to construct a simple framework in Python that can accept an arbitrary classifier and then compute the AUC score for it and print it out. Our classification models accept the input data as a table which contains the subjects' identifiers and their timing vectors.

As a simple example to show how the timing vectors are constructed, we choose the word "lol" and measure different times, which we denote by H, DD and UD. H (hold) denotes the time between pressing and releasing a key, DD (down-down) denotes the time between two consecutive presses of keys and UD (up-down) denotes the time between the release of the previous key and the down-press of the next one. We put all of these timings in a vector that looks like this:

$$[H_l, H_o, DD_{lo}, UD_{lo}, H_l, DD_{ol}, UD_{ol}].$$

We can now make a classifier that takes in a table of these vectors accompanied by subjects' identifiers and trains a model based on recorded timings. The testing and evaluation algorithm comprises of a few steps and is best described with some simple pseudo-code, shown in Algorithm 1.

Data: classifier, dataset
Result: AUC score
all scores = [];
for *subject in dataset* **do**
  train data = first 200 rows of current subject's data;
  train model on the train data;
  legitimate test data = last 200 rows of subject data;
  imposter test data = random sample of vectors from other subjects;
  test model on the legitimate and imposter data;
  roc = compute ROC curve;
  calculate AUC score on the computed ROC curve;
  append score to all scores;
end
return mean of all scores

Algorithm 1: Testing a classifier.

We essentially test how well the classifiers distinguish authentication attempts that are made by imposters from the attempts made by the people who have recorded the original keystroke patterns. We then cross-validate scores from all of the subjects by computing the distances between their trained vector and their legitimate and random imposters input vectors. We use those distance scores to compute the Receiver Operating Characteristic curve and the area under it (AUC), which is our final metric to describe the performance of the binary classifier.

We also add the option to choose which features we wish to include when training a model (e.g. only UD and DD) and compute accuracies based on those constraints. Finally, we display all the mean AUC scores for our classifiers next to their names and construct a table of scores, show in Section 3, Table 1.

C. Implementation of non-deep classifiers

We implement one of the best performing techniques from the paper by Killourhy and Maxion [6] and check the importance of feature selection. We then also use a deep learning approach from an open source implementation by Janakiev [7].

1) Euclidean-Distance-Based Classification

To train our model, we obtain our training data from the current subject and construct a vector of the means of times, captured in the cited experiment. That means that for each repetition of the password input, we measure the H, DD and UD times and create a vector of their averages. To compute the distance score from the model of the current subject and the input vector, we use the Euclidean distance formula:

$$d_{euclidean} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2},$$

where q_i is the i-th measured time in the input vector and p_i is the i-th measured time in the mean model vector. We will denote those times with the same letters in the following formulas as well.

2) Scaled-Manhattan-Distance-Based Classification

The best equal error rate in the cited experiment is achieved by the Scaled Manhattan classifier, which computes the training model in the exact same way as the Euclidean classifier (i.e. constructs a mean timing vector for a subject). The difference is the distance metric we use and that is the Manhattan distance, which we then also divide by the mean absolute deviation.

Our distance is computed using the equation:

$$d_{scaledmanhattan} = \sum_{i=1}^{n} |q_i - p_i| / msd_i,$$

where msd_i is the mean standard deviation for the timing at the i-th position.

3) Match-Ratio-Distance Classifier

As another distance metric we also create our own Match-Ratio-Distance. It compares pairwise timings in vectors and computes the ratio between them to essential-

ly describe how much different the times measured are. We will be using this metric in our multi-modal system because of it's interpretability. To transform a simple distance into a more understandable metric for the person who is supervising authentication attempts for e.g. a quiz or an exam (which is the scenario for the usage of our multi-modal system), we will use the match ratio metric instead of the Euclidean distance.

The match (which is opposite to the "difference") between two vectors is calculated using the following equation:

$$match_{\%} = \sum_{i=1}^{n} \frac{min(q_i, p_i)}{max(q_i, p_i)}.$$

This equation provides ratios between timings for two measurements (e.g. 20ms versus 40ms gives us a ratio of 0.5 and likewise 40ms versus 20ms). The final output of the classifier is a mean ratio for all the pairs in two vectors.

4) Testing deep classifiers

We base our deep learning approach on the implementation of the neural network by Nikolai Janakiev [7], which uses the same dataset as we did before and reports the accuracy (accuracy metric from the Keras library) of 94%.

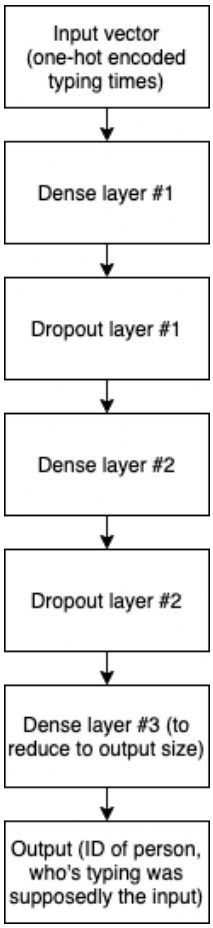


Figure 1: Janakiev's proposed neural network model structure.

We create the same sequential model using the Keras library, which features two hidden layers with 300 nodes each and two dropout layers to prevent overfitting (the architecture is shown in Figure 1. But this proposed architecture here will be trained on the predefined dataset and will output the subject's identifier, which makes it a non-binary classifier. We selected the number of nodes and layers based on the best performing classifier according to Janakiev. We made all the codes available online at https://github.com/jstavanja/kd_and_fusion.

Using the more traditional approaches, we measure distances between vectors to make a judgement whether someone's typing rhythm is similar enough for a successful authentication attempt. As we implement Janakiev's architecture [7], we immediately see a potential flaw in the maintainability of the system because of the issues at enrolment due to the shape of the last fully-connected layer. Because the proposed model uses one-hot encoding to store each person's typing rhythm vectors, we say that an authentication attempt is successful if the last layer yields the one-hot encoded identifier of the person currently trying to log in. As we enrol more people into the system, we would need to reshape the last layer and retrain the network every time, which is time consuming.

5) Fusion of keystroke dynamics and face recognition

We know that we can get very accurate classifiers by using deep learning approaches, but those can be complicated to manage in production. As we will show in the next section with results, some of the accuracies achieved by the basic classifiers are fairly good. We use a fusion of different modalities to create a more accurate system. One of the most widely used approaches nowadays is the use of facial recognition. For example, the Face ID system from Apple is in use over the whole word for authenticating subjects on a mobile phone.

In our previous work, we created a system for authenticating students [5] into the quizzes on the Moodle platform which compares keystroke patterns and uses OpenFace [8], a face recognition deep learning library, for measuring differences between faces.

In this system we inject a registration from into a Moodle class where all the participants record their typing patterns and submit a picture of their face taken with a webcam. Before every quiz we then inject a form into which a person that is solving the quiz has to type in the same word and submit the current picture of their face. The data gets sent to the API, which computes a distance between the user's trained pattern (which is just a mean vector, constructed with the timing vectors recorded in the registration process) and the current keystroke record and the OpenFace scripts do the face comparison. To set the thresholds for allowing people to solve a quiz, we perform some testing of the system on ten volunteers. We use the ratio metric here as it is more interpretable for humans. For example, the supervisor of some quiz would (in the administration dashboard) rather see a percent match (e.g. 80\%) between two patterns than a distance (e.g. 69.420).

We recorded their typing patterns and took pictures of their faces on a MacBook Pro 13" using our Moodle plugin's forms. Then we make them try to log in to other peoples' accounts and collect the distances between keystroke patterns and faces for as many authentication attempts as possible. The results we obtained for both

Name of classifier / Measure (Features)	AUC (H)	AUC (DD)	AUC (UD)	AUC (H, DD)	AUC (H, UD)	AUC (DD, UD)	AUC (all)
Euclidean	**0,884**	0,777	0,800	0,798	0,798	0,780	0,794
Scaled Manhattan	0,905	0,819	0,815	**0,922**	0,916	0,826	0,910
Ratio	**0,901**	0,760	0,615	0,864	0,661	0,644	0,679
NN (Janakiev)							**0,998**

Table 1: AUC scores for classifiers, tested on the Killourhy and Maxion dataset.

	A	B	Γ	Δ
A	92,44%	61,00%	55,03%	63,81%
B	51,55%	88,51%	67,58%	71,80%
Γ	55,27%	78,65%	80,05%	78,55%
Δ	52,71%	75,47%	64,61%	81,68%

Table 2: Ratio classifier match outputs for the first part of the testing process (comparing typing rhythms of 4 volunteers).

modalities and the AUC score can be found in the Section 3. All of the code is available in the Moodle plugin [9] and backend API [10] repositories online (https://github.com/jstavanja/quiz-biometrics-plugin and https://github.com/jstavanja/quiz-biometrics-api).

III. Results

For our classifiers, we test the impact of feature selection on the final AUC score and write all the test results in Table 1. As shown in the research done by Janakiev, the best outcome for the neural network based classifier is when we use all of the available features (H (hold time), DD (down-press to down-press time) and UD (release to down-press time)). Thus, we train the network on those features only to save time, as these types of networks have to be trained for long periods of times. We use the dataset from Killourhy and Maxion.

By simply looking at the scores it is evident that the best approach for the Euclidean and Ratio classifiers is to only use the H features and to use H and DD for the Scaled Manhattan classifier. Those features will in real life be recorded on the frontend on some website's login form and sent to a backend system that would compute a distance and try to see if it is acceptably small. We would set the smallest acceptable distance by thresholding our ROC curve, obtaining the point that is closest to the upper left corner of the ROC plot and taking that threshold as the limit.

We can see how well the neural network performs with all the features, but as explained it is a lot harder to maintain and not so successful. For example, the training on the dataset using the Scaled Manhattan distance approach takes up to 10 seconds where as the training of the neural network take about half an hour.

We now record the rhythms for four volunteers on our own. The ratio classifier results are shown in Table 2. We now use the ratio metric because we wish to preserve the idea of the score being understandable for the person verifying identities and also judging by the fact that it doesn't yield a much smaller accuracy than the other classical classifiers. Using the ROC curve, we determine the thresholds that separate still acceptable matches (between the input and registration vectors) from the non-acceptable. We set the limit for the score to be a match if it is bigger than 78,65%., which would classify every attempt of the keystroke login correctly on our test run. We use this threshold in the multi-modal system later. Then, for the face comparison part of the multi-method system, we set the threshold to 75,25% as recommended by the authors of the OpenFace library.

We then take a different group of volunteers and test our whole system with the predetermined thresholds and denote a score with P if the check is a pass and S if the score is suspicious (e.g. the match is lower than the threshold). The scores we obtain using our keystroke ratio metric can be seen in Table 3 and the scores from the face comparison are shown in Table 4. We label the participants with Greek letters to keep them anonymous.

Even though our dataset is small, we can see how powerful a system can be when we use different modalities and fuse them together. If we take into consideration both of our scores (keystroke and face match) and the set thresholds, we can see that our system make does not make any errors.

IV. Conclusion

We see that our simple and quick classifiers achieve a fairly good result in comparison to neural networks and are easier to maintain. But in case we are creating a system where one could potentially want an imposter to act as the current user instead of him (e.g. a school quiz, where the imposter is more skilled in a certain subject than the owner of the account), we need to take into consideration that a typing rhythm is a potential risk as it is easy to teach someone and spoof. If in a system one wishes to protect his data using keystroke typing patterns, the described methods can be a great addition to an existing authentication service.

Multi-modal systems like the one we tested and implemented for Moodle can be a great addition to an already implemented security system, e.g. a username and password authentication service. What we tested in our plugin for Moodle implementation essentially does exactly that. In order to login in to a student's account, we

	A	B	C	D	E	F
A	85.13 % P	68.83% S	76.57% S	67.13% S	68.37% S	68.80% S
B	65.23% S	83.66% P	66.19% S	67.22% S	68.16% S	72.31% S
C	70.20% S	71.20% S	83.03% P	68.49% S		64.11% S
D	69.83% S	76.72% S	72.45% S	87.76% P		75.17% S
E	57.91% S	72.49% S	68.28% S		87.11% P	70.27% S
F	59.79% S	80.08% P				82.90% P

Table 3: Matches for keystroke pattern comparison for 7 attempts of writing ‘':ok_hand:" as a password on 6 volunteers that don’t overlap with the first group of volunteers.

	A	B	C	D	E	F
A	97.83% P	56.06% S	71.97% S	64.33% S	67.81% S	74.49% S
B	49.47% S	97.59% P	54.03% S	29.99% S	38.35% S	57.80% S
C	78.39% P	65.98% S	98.99% P	76.38% P		56.98% S
D	70.65% S	37.55% S	68.41% S	91.30% P		39.91% S
E	62.50% S	40.26% S	56.73% S		97.06% P	42.38% S
F	64.68% S	70.60% S				92.34% P

Table 4: Face comparison matches on 6 volunteers that don’t overlap with the first group of volunteers.

need a username and password combination. Then, to recheck the identity of a student, we add the keystroke pattern and face comparison at the start of every quiz. All of that together can be very powerful and provide a safe system for checking student's knowledge on the computer instead of them needing to write an exam on paper.

In conclusion, we can see that the keystroke pattern modality is not as bad as we might think, but in order to achieve good performance, we almost certainly need to fuse it with different modalities or use a lot of processing power and memory and carefully maintain the authentication system. As we saw by creating the Match-Ratio-Distance classifier, it is easy to come up with a new way of differentiating two patterns. That means that there may still be room for research on different metrics that could provide us with a very simple and performant keystroke dynamics classifier without the need for neural networks.

ACKNOWLEDGMENT

Sincere thanks for making the data collection part of our experiments easier go to family members (Nuša, Špela and Iztok) as well as to David J. and other colleagues (Didka, Matija, Sandi, Sašo and Matevž). Without your valuable input and time, our experiments would not be possible.

REFERENCES

1. P.. H. Pisani and A. C. Lorena, “A systematic review on keystroke dynamics”, Journal of the Brazilian Computer Society, vol. 19, no. 4, pp. 573–587, Nov 2013. [Online]. Available: https://doi.org/10.1007/s13173-013-0117-7
2. R. Joyce and G. Gupta, “Identity authentication based on keystroke latencies, ”Communications of the ACM, vol. 33, no. 2, pp. 168–176, 1990.
3. J. R. Montalvao Filho and E. O. Freire, “Multimodal biometric fusion — joint typist (keystroke) and speaker verification,” in Telecommunications symposium, 2006 international. IEEE, 2006, pp. 609–614.
4. R. Giot, B. Hemery, and C. Rosenberger, “Lowcost and usable multimodal biometric system based on keystroke dynamics and 2d face recognition,” in Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, 2010, pp. 1128–1131.
5. J. Stavanja, “Use of biometric techniques in online exams. engd thesis.” University of Ljubljana, Faculty of Computer and Information Science, 2018.
6. K. S. Killourhy and R. A. Maxion, “Comparing anomaly-detection algorithms for keystroke dynamics,” in Dependable Systems & Networks, 2009. DSN’09. IEEE/IFIP international conference on. IEEE, 2009, pp. 125–134.
7. “njanakiev/keystroke-biometric: Biometrics on keystroke dynamics.” Github URL:https://github.com/njanakiev/keystroke-biometrics, [Accessed on: 12.1.2019].

8. B. Amos, B. Ludwiczuk, and M. Satyanarayanan, “Openface: A general-purpose face recognition library with mobile applications,” CMU-CS-16-118, CMU School of Comput
9. er Science, Tech. Rep., 2016.“jstavanja/quiz-biometrics-plugin: Moodle plugin for question type, that communicates with the quiz-biometrics-api.” Github URL: https://github.com/jstavanja/quiz-biometrics-plugin, [Accessed on: 12.1.2019].
10. “jstavanja/quiz-biometrics-api: Bachelor’s thesis backend api code.” Github URL: https://github.com/jstavanja/quiz-biometrics-api, [Accessed on: 12.1.2019].

Digital forensics appliance in corporate ecosystem considering limitations in the EU legal framework

Vinko Rajič univ. bacc. ing. traff., Melita Milenković, LL. M. and Goran Vojković, Ph.D.
Faculty of Transport and Traffic Science, Department of Transport Planning, Chair of Transport Law and Economics
Zagreb, Croatia
gvojkovic@fpz.hr,
mmilenkovic@fpz.hr,
rajicvinko@gmail.com

***Abstract* - This paper presents appliance of digital forensics in a corporate ecosystem i.e. appliance of digital forensics investigation in a corporate or internal investigation as a result of an incident caused by internal or external threat while considering necessity of coexisting corporate policies and legal regulations. Conducting digital forensic investigation refers to the process of acquiring and processing data for the purpose of investigation which are commonly performed by the corporate investigation team, i.e. by the third-party investigators. Today's organizations need to be forensically ready to minimize a response time necessary for timely reaction, both in the cases of internal or external digital forensics incidents. The goal of this paper is to provide oversights of various companies as well as deficiencies of EU legal system.**

Key words - digital forensics, forensic investigation, data protection, data breaches, corporate ecosystem, GDPR

I. INTRODUCTION

Main goal of this paper is to allude on the insufficiency of EU legal framework when investigation is conducted in corporate ecosystem. After defining digital forensics and reviewing its existing impact on corporate ecosystem, recommendation for legal framework will be proposed. In the procedure of collecting, processing, interpreting and presenting digital evidence, and information about how the evidence was handled it is more important than the content of the evidence itself. This is due to a vaguely defined and/or non-transparent legal framework, in the context of the processing of an individual's data. This incompleteness of legal regulation limits the ability of forensic scientist to collect and process possible evidence, which can directly affect the outcome of the investigation.

Before laying the foundations of modern forensics, it was difficult to pinpoint the offender of any act. The reason for this was that the trials often relied on forced confessions and on the witnesses' testimony, and it wasn't uncommon that the actual offender was left unpunished. [1] For this reason, the need for examination of the very circumstances of a crime has occurred, instead of drawing conclusions based on incomplete statements of a witnesses. Forensic science has begun to be applied in order to enhance the validity of the conclusion reached on the basis of the investigation carried out by the competent authorities.

First instances of forensics development were seen on the local level when authorities could identify the prime suspect just on the M.O.[1] However, the first significant step in identification of an offender was introduction of fingerprinting method. Afterwards, forensics as a science has started to couple a variety of methods for identification of evidence at the scene of the crime such as hair, soil, stains, etc. [2]

Forensics science involves the application of scientific methods and processes, such as biology, chemistry, physics, computer science, etc.; in accordance with the law for the purpose of conducting a criminal investigation [3]. It is a set of carefully coordinated activities and processes by which evidence is acquired, analysed and interpreted. The moment when object started to be considered an evidence, it becomes a mute witness to the crime [4]. Forensics is a process of questioning a mute witness performed by the authorized personnel. Forensics science has had a crucial role in the legal framework by providing scientifically based information through analysis of physical evidence since the legal and professional aspect of forensics are interrelated, i.e. best evidence is useless if it is acquired illegally [5].

Forensic science has a various application on a vast field such as geology, accounting, botany, serology, etc. However, those fields started from the more common area of forensics: pathology, anthropology, odontology, entomology, behavioural forensic science and computer or digital forensics. [6] Main focus of this paper is on digital forensics which will be explained more in depth in the next chapter.

II. THE FIELDS OF APPLIANCE OF DIGITAL FORENSICS

Although widespread use of computers began in the second half of the 20th century, digital forensics did not have any significant development until the end of the 20th century. The reason is that computers had been operating

[1] A modus operandi – someones habit of working in the context of criminal investigations. Term is used when addressing the methods applied by the criminals.

in the industrial production and were owned by corporations, institutes, government organizations, etc. Although computers had limited functionality at the time, they performed data processing that allowed organizations to perform their core business functions. In order to avoid data manipulation, information security and methods that could trace the perpetrator began to be considered [7]. While taking into consideration that digital forensics is relatively a new branch of forensics science, it can be said that the digital forensics went through three stages of development which can be seen on the Figure 1. Those are:

- The Ad Hoc Phase
- The Structured Phase
- The Enterprise Phase [8]; as seen on the Figure 1.

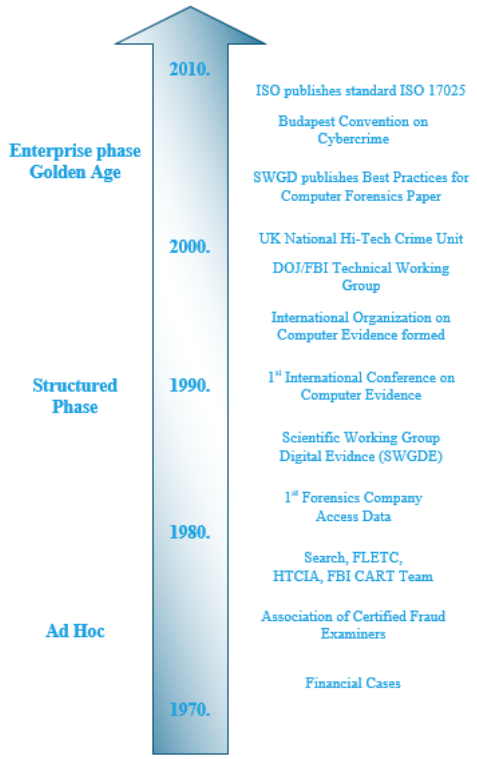


Figure 1 Digital Forensics Timeline [11]

Digital forensics, as a branch of forensics science involves identifying, acquiring, processing and interpretation of digital evidence that can be stored on various types of memories in different formats. Main purpose of digital forensics is extraction of data from device, its interpretation, and presenting the evidence to the authorized institutions in an intelligible form [9]. Digital forensics can be classified by different criteria, although the most common classification is by its usage on different aspects of information-communication technology:

- Computer forensics
- Network forensics
- Mobile device forensics
- Database forensics
- Other (cloud, e-mail, IoT forensics, etc.) [10]

Every mentioned sub-branch of forensics encounters the same challenges regarding the integrity of digital evidence. There are instances where data extraction cannot be done without alteration or contamination which renders evidence useless. All of the aforementioned points to standardization of methods and tools as mandatory in a manner acceptable by the law. Meaning that data is collected, handled, and stored using tools and methods that are verified to be in compliance with the law. This is also known as a Forensic Soundness [12]. From the moment when data begins to be considered as an evidence in forensic investigation, standardized measures should be considered.

Furthermore, careful handling regarding the evidence must be taken into account due to volatile nature of digital evidence. This means that the evidence must be documented from the moment when identified, acquired, processed, interpreted and presented in the court, i.e. Chain of custody[2]. Moreover, clear and structured process step-by-step can help investigator to determine if there was any unallowed action that can compromise the integrity of evidence. [13]

While digital forensics can have wide application regarding its variety of tools and methods, forensics should be performed using the forensic process shown on the Figure 2. It consists of four steps where each must handle the evidence without altering or contaminating them. These are: Collection, Examination, Analysis and Reporting, [14]. In order not to deviate from the procedures, there is a need to define the methodologies that can be followed when managing the evidence. Carefully considered and defined methodology can help with proper handling regarding the evidence.



Figure 2 Forensic Process [14]

The data is transformed into information when analysis is being conducted and finally information becomes evidence when it is reported since a decision is made based on the evidence. [14]

III. DIGITAL FORENSICS IN CORPORATE ECOSYSTEM

A. Digital Forensics Process

Purpose of forensic process is to acquire unbiased evidence of criminal activity and as such, it can give answers to the following:

[2] Chain of custody refers to as chain of possession of evidence from the moment the evidence is collected, examined, analyzed and reported.

- Association of criminal activities with the individual
- Assessing alibis and statements
- Determining intent
- Credibility of sources
- Document Authentication
- Data gathering and observation [12]

Digital forensics has a wide application in different areas of legal framework:

- Criminal investigation - Forensics methods and tools that are used in data extraction process depends on which type of device or media the data is being extracted from. Main purpose of digital evidence is to support or rebut a thesis or argument on which court decision is based on.
- Civil Litigation - In civil litigation, both parties have the right to review the evidence that will be used against them before the trial itself. The most common cases of civil litigation are divorce or child custody.
- Public sector security and operation - Security and public sector operations include stopping hacking attacks, investigating them, intelligence, and counterintelligence.
- Corporate investigation - violations can be committed in various ways, such as: disclosing sensitive information to competitors, intentional or unintentional sabotage, etc. In such cases, digital forensics is used to determine where the omission occurred., which may not necessarily be illegal, but if they contravene with the company policy, sanctions are imposed by the employer. [10]

When discussing digital forensics in a corporate ecosystem, two aspects are considered when it comes to a potential incident:

- Preventive measures as a mean of preventing an incident
- Corrective measures as a mean of repairing exploited weakness

B. Preventive measures as a mean of preventing an incident

Preventive measures are a set of measures and rules designed to protect and secure the information-communication system. This means defining and applying security policies and principles to prevent the intentional or unintentional damage to an organization's assets, financial power, or reputation. Information-communication systems are at risk of every-day exposure to threats which could easily compromise information on which organizations rely their core business function.

For this reason, three basic principles of information security have been defined: confidentiality, integrity, and accessibility [15]. Confidentiality ensures that classified information is available only to authorized personnel. Integrity implies that confidential information is protected against unauthorized changes. Availability ensures that the information is accessible to authorized users [16].

The stated principles of information security are implemented and enforced with security policy. Security policy is a set of rules of conduct when managing information-communication system and it applies to all employees of the organization. It defines acceptable and unacceptable behaviour as well as sanctions, i.e. consequences in case of non-compliance [17] Security policy is based on organizations requirements of the risk management as well as the balanced cost/benefit ratio which implies the question: Is it worthwhile to protect information that is not of great value to the organization? This means assessing whether it pays to invest in protecting information that has a little value to the organization [18].

Concerning organizational security in the EU companies it would be very hard to imagine such a system being operational without adhering to rules of a certain professional certification. Therefore, it has been noted that most appropriate system for standardization of information security practices would be ISO/IEC standard 27001, [19] although ENISA [20] currently works on forming of new cyber standards which will improve existing ISO/IEC 27001 standards. Standardization with detailed risk analysis has proven to be useful in cases of recovery from security incidents and it allows the organization to continue to operate undisturbedly.

Taking politics for granted is very often the reason why organizations find themselves in incident situations. Most common mistakes that are made are:

- Not having a policy
- Not updating the security policy
- Not tracking compliance with the security policy
- Having a “tech only“ policy
- Having a policy that is large and unwieldy [21]

Security policy is a process, not a one-time solution which results in corrective measures to be taken.

C. Corrective measures as a mean of "repairing" exploited weakness

Corrective measures are a set of activities undertaken after an incident has occurred. The purpose of remedial action is to eliminate the weakness that has been exploited in the system and to identify the perpetrator who has exploited the same weakness. The perpetrators may take the form of an external and internal threat. The most common use cases of digital forensics in a corporate ecosystem can be divided into:

- Intellectual property (IP) theft and Internal investigations
- Data recovery

- Damage control [22]

Internal investigations within the organization aim to identify the perpetrator who has violated the organization's policies. Violation of an organization's policy implies intentional or unintentional disregard for the interests of the employer or the employment contract. The most common case is a breach of the privacy statement. (proprietary files associated with a revenue model, customer/customer database or trade secret). Furthermore, digital forensics can also respond in cases of fraud, injury at work or sexual harassment. All digital media and devices used within the organization can be used as potential evidence. [23]

The most common cases of ‘’misconduct‘’ in organizations are: accounting fraud, asset misappropriation, unauthorized use of organization assets, threats, sexual harassment and other forms of inappropriate behaviour and other violations of policies, misconduct or criminal activity [23]. For an internal investigation to be conducted, it is mandatory that the organization prepares in advance to conduct an effective investigation. To make the internal investigation more effective, it is necessary to: plan and organize, fast acting, thoroughness, honesty, accuracy and precision, confidentiality, and proper documentation. [23]

Data recovery is a process of recovering data from different types of memory that is damaged, inaccessible, lost or erased. The process itself often involves disassembling memory as well as repairing damaged parts. Data recovery does not fall directly under digital forensics but becomes part of it when recovery is done for the purpose of gathering evidence for criminal or internal investigations. This includes instances where there has been a violation of an organization's policy and the return of data is for the purpose of identifying the perpetrator and/or recovering sensitive data. In addition, the difference between a normal data recovery process (such as a memory failure) and data recovery due to an investigation is the documenting process. A recovery forensic scientist is required to follow a rigorous procedure, record, the extraction process and to submit a report. [24]

Damage control is a set of measures or activities undertaken after an incident has occurred. There is a possibility that sooner or later, every organization will be compromised by an internal or external threat. When this happens, organization must remain in "control" of the incident. When incident occurs, the organization must have in place procedures which should be prepared to apply. Damage control means controlling the loss of reputation, profits, and business partners. The methods of damage control depend on the current situation and there is no unique method of damage control. In order to minimize damage, the damage control team must find the right approach to gain "control" of the situation. Damage control is divided into three basic items:

- Communication
- Responsibility, and
- Solution [25]

For damage control to be successful, organization must be direct and up to date in dealing with the incident. This implies taking corrective action, accepting responsibility, and implementing the appropriate solution. When planning and anticipating potential threats, crisis management checklist should be made. The checklist should include: pre-incident planning, security measures, emergency contacts, establishment and implementation of a security policy and creation of a crisis management team. [24]

IV. EXISTING EU LEGAL FRAMEWORK

A. Member States' insufficient regulation in the field of digital forensics

The conduct of digital forensics investigation resulting from the data breach in the company either by the internal or external perpetrator is not regulated by any particular EU Regulation or Directive. Rather, each company should have regulated its own Ordinance on data breach policy and have a dedicated department that handles such cases or it could outsource a company to investigate and resolve each case in the event of an attack. Each company independently decides how to conduct a digital forensics procedure in the event of an attack on the system, whether by internal factors or external ones. There is no legal binding act that fully defines data breaches and/or answers to question, such as: who can request data analysis when a security attack occurs and in which cases digital forensic data processing takes place. As seen in the Table 1, the following legal regulations regulate these issues, although only in the field of free movement of personal data and processing of personal data protection. It should be noted that the ePrivacy Regulation is intended to supplement and refine the GDPR, especially in the field of user metadata and cookie consent obligation, which is neither regulated by the GDPR or any other Regulation, so that the rules of the GDPR are always relevant and are an overall part of the legislative aspects of ePrivacy.

The EU's ePrivacy Regulation was proposed to extend the scope of the current ePrivacy Directive and to harmonize the various online privacy rules that exist in EU Member States.

Table 1 Overview of key regulations/directives and their implication

The EU legal framework	**Implications**
General Data Protection Regulation (EU) 2016/679	*Processing of personal data and ensuring the free movement of personal data*
ePrivacy Directive 2002/58/EC	
Directive 2000/31/EC	
Directive 98/34/EC	
Regulation 2018/1725	
Directive (EU) 2016/680	

B. Proposal for the new ePrivacy Regulation

On 10 January 2017, a proposal for a new ePrivacy Regulation (hereinafter: ePR) was adopted. The proposed ePR was intended to replace the existing ePrivacy Directive 2002/58 [26]. As well as updating the current ePrivacy framework in the EU. The Commission has qualified the proposal as a *lex specialis*[3] in regard to General Data Protection Regulation 2016/679 (hereinafter: GDPR), which is *lex generalis*[4]. The original aim has been for the proposed ePR to become enforceable on 25 May 2018 (at the same time as GDPR). [27]

Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications) should particularize and complement GDPR regarding electronic communications data that qualify as personal data. This Regulation on the one hand does not apply to electronic communications services which are not publicly available; which on the other hand represents an issue for the companies in the real sector; the companies that don't have an obligation to the processing of electronic communications data in accordance with the Regulation. Such companies can only be guided by the GDPR and their internal rules of procedure, that is, national laws, assuming that they have been enacted, given that this area is fairly new and still deregulated.

Therefore, Member States should lay down in their national laws specific guidelines for dealing with attacks against electronic communications metadata, as such security attacks result in the loss of a significant amount of classified information. Real-sector companies whose electronic communications services are not publicly available are left to resolve such security attacks themselves without any legal guidance from the state.

Also, it is important to notice that Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000, on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (hereinafter: Directive on electronic commerce) is giving definition of information society, which, according to its recital 17, states: "The definition of information society services already exists in Community law in Directive 98/34/EC of the European Parliament and of the Council of 22 June 1998 laying down a procedure for the provision of information in the field of technical standards and regulations and of rules on information society services and in Directive 98/84/EC of the European Parliament and of the Council of 20 November 1998 on the legal protection of services based on, or consisting of, conditional access; this definition covers any service normally provided for remuneration, at a distance, by means of electronic equipment for the processing (including digital compression) and storage of data, and at the individual request of a recipient of a service, except services in the indicative list in Annex V to Directive 98/34/EC which do not imply data processing and storage are not covered by this definition." The aim of this Directive is to create a legal framework to ensure the free movement of information society services between Member States, and not to harmonize criminal law as such.

C. Analysis of data protection according to the EU legal regulation

Moreover, all matters concerning the processing of personal data not specifically addressed by the proposal are covered by the GDPR. The alignment with the GDPR resulted in the repeal of some provisions, such as the security obligations of Article 4 of the ePrivacy Directive. Unfortunately, the proposal does not include any specific provisions in the field of data retention. It maintains the substance of Article 15 of the ePrivacy Directive and aligns it with specific wording of Article 23 of the GDPR, which provides grounds for Member States to restrict the scope of the rights and obligations in specific articles of the ePrivacy Directive. Therefore, Member States are free to keep or create national data retention frameworks that provide, *inter alia*[5], for targeted retention measures, in so far as such frameworks comply with Union law, taking into account the case-law of the Court of Justice on the interpretation of the ePrivacy Directive and the Charter of Fundamental Rights. [28]

According to the Regulation (EU) 2018/1725 [27] of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data in the preamble (1) states: ''The protection of natural persons in relation to the processing of personal data is a fundamental right.'' Article 8(1) of the Charter of Fundamental Rights of the European Union (the "Charter") and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU) provides that everyone has the right to the protection of personal data concerning him or her.

This right is also guaranteed under Article 8 of the European Convention for the Protection of Human Rights and Fundamental Freedoms. GDPR and Directive (EU) 2016/680 of the European Parliament and of the Council were adopted on 27 April 2016. While the Regulation lays down general rules to protect natural persons with regard to the processing of personal data and to ensure the free movement of personal data within the Union, the Directive lays down the specific rules to protect natural persons with regard to the processing of personal data and to ensure the free movement of personal data within the Union in the fields of judicial cooperation in criminal matters and police cooperation.

Also, there is OLAF [29], European anti-fraud office which has a legal basis for conducting a digital forensics operation in internal investigations under the Article 4(2)

[3] *Lex specialis* is a Latin phrase which means ''law governing a specific subject matter'', it comes from the legal maxim ''lex specialis derogat legi generali''. This doctrine relates to the interpretation of laws. It can apply in both domestic and international law contexts

[4] *Lex generalis* refers, literally means, to the ''general law''.

[5] Inter alia is a Latin phrase that means ''among other things''.

of Regulation (EC) 883/2013 empowering OLAF to take a copy of, and obtain extracts from any document or the contents of any data medium held by the institutions, bodies, offices and agencies, and if necessary, assume custody of such documents or data to ensure that there is no danger of their disappearance, [29] and with responsibility for conducting administrative fraud investigations. The Commission has given the Office full independence to exercise its investigative function. Under Decision 1999/352/EC, ECSC, Euratom provides that for the purposes of investigations the Office is to exercise the powers conferred by Union law. [29]

Moreover, European Union Agency for cybersecurity (ENISA) has been working on making Europe cyber space secure. What ENISA advocates is development of the EU Cybersecurity Certification Framework which will make it easier for ICT manufacturers and developers to serve the EU market. A unified Cybersecurity Certification Framework across the EU will reduce the effects that a fragmented market has on the economy. To support the creation of certification schemes under this framework the role of standardisation bodies is key. [31]

V. CONCLUSION

As mentioned in the paper, the most typical case of corporate use of digital forensics is investigating internal policy violations such as intellectual property (IP) theft, e.g. if an employee leaves the organization and starts working for a competitor - and takes ownership files related to revenue model, customer databases or trade secrets. Digital forensics can also support cases against employees in cases of fraud, wrongful death or personal injury or even sexual harassment. Laptops, social networking applications and mobile phones can provide evidence in the form of conversation history as well as geolocation data. Emerging technologies will continue to shape digital forensics, and experts say it is crucial to keep up with the global trends in technology - such as newer and more secure mobile devices, Internet of Things (IoT) platforms and artificially intelligent systems.

Therefore, all of the mentioned Regulations and Directives make it easier to handle cyber-attack cases, however it can be deduced that none of the above legal acts clearly delineates how to proceed in the event of an attack and/after the attack has occurred, except for OLAF, which Commission has set up for this purpose.

It should also be noted that EU regulation is also desirable from the point of view of criminal prosecution, as cybercrime is mainly linked to more than one country, so that perpetrators are often prosecuted in a completely different environment from the place of the perpetration. Standardization is what makes it easier to prove the nature of the crime.

Therefore, in this paper we propose that services of the mentioned companies (real-sector companies engaged in virtual business and providing services online and for the ones who could easily become victims of cyber-attacks); would be appropriate to fall within the scope of the Directive on electronic commerce, due to the fact that all the services companies perform fall within the definition of the information society and fall within the activities of the information society, according to the Recital 17 of the Directive.

The next necessary step in regulating security risks is certainly adopting an ePrivacy Regulation which should be overregulated in relation to GDPR, and to harmonize standardization, which will regulate current deficiencies and legal omissions in the Member States' legal framework. It still remains unknown if the ePrivacy Regulation will be adopted and accepted, instead there is a possibility of amending ePrivacy Directive within the matters not specified with the scope of GDPR.

VI. REFERENCES

[1] A. Embar-Seddon, A. D. Pass, "Forensic science", Salem press, 2008

[2] W. G. Eckert (edited), "Introduction to Forensic Sciences, Second edition", CRC Press, 1997

[3] National Institute od Justice, Forensic Sciences https://nij.ojp.gov/topics/forensics (Retrieved 1 of February 2020)

[4] J. E. Girard, Criminalistics: Forensics science, Crime, and Terrorism, 3rd Edition, American University, Department of Chemistry, Washington, DC,

[5] Z. Durdevic, "Admissibility of evidence, judicial review of the actions of the European Public Prosecutor's Office and the protection of fundamental rights", Head of the Department of Criminal Procedural Law, University of Zagreb, http://www.europeanrights.eu/olaf/pdf_eng/4-Admissibility%20of%20evidence-zd.pdf (Retrieved 1 of February 2020)

[6] K. Mirakovits, J. A. Siegel, Forensic science, the basics, Second edition, 2010, CRC Press

[7] M. Pollit, A History of Digital Forensics, 6th IFIP WG 11.9 International Conference on Digital Forensics (DF), Jan 2010, Hong Kong, China. pp.3-15

[8] I. Chartes, M. Smith, G. McKee, The Evolution of Digital Forensics: Civilizing the Cyber Frontier, 2009, http://www.guerilla-ciso.com/wp-content/uploads/2009/01/the-evolution-of-digital-forensics-ian-charters.pdf?fbclid=IwAR0v5iqcnlEhs9Y0UhgzZenhD5hcgpRHMuFt3-vxgh3UpdMVQUgwRsbl9Tk (Retrieved 11 of February 2020)

[9] Interpol, Global guidelines for digital forensics laboratories, 2019.

[10] D. Peraković, Authorized course lectures: Forensic analysis of information and communication system, University of Zagreb, Faculty of Transport and Traffic Sciences, Department of Information and Communitcation Traffic, "Basic of digital evidence" (in Croatian), 2019

[11] Overview of Digital forensics, Cybersecurity Nexus, ISACA, 2015

[12] E. Casey (edited), Handbook of digital forensics and investigation, Published by Elsevier Inc., 2010

[13] D. Peraković, Authorized course lectures: Forensic analysis of information and communication system, University of Zagreb, Faculty of Transport and Traffic Sciences, Department of Information and Communitcation Traffic, "Forensic analysis methodologies for information communitaction system" (in Croatian), 2019

[14] K. Kent, S. Chevalier, T. Grance, H. Dang, Guide to Integrating Forensic Techniques into Incident Response, Recommendations of the National Institute of Standards and Technology, Special Publication 800-86, Gaithersburg, National Institute of Standards and Technology, 2006, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-86.pdf (Retrieved 11 of February 2020)

[15] D. Peraković, Authorized course lectures: Security and protection of the information and communication system, University of Zagreb, Faculty of Transport and Traffic Sciences, Department of Information and Communitcation Traffic, "Information security terminology" (in Croatian), 2017

[16] I. Kokot, "Criminal-law Protection of Computer Systems, Programs and Data" (in Croatian), Zagreb Law Review, e-Journal of Postgraduate Studies of Faculy of Law, University of Zagreb, Faculty of Law of the University in Zagreb, vol. 3, p.p. 301-327, 2014.

[17] Croatian academic and research network, Security policy (in Croatian), CCERT-PUBDOC-2009-05-265 https://www.cert.hr/wp-content/uploads/2009/05/CCERT-PUBDOC-2009-05-265.pdf (Retrieved 1 of February 2020)

[18] M. S. Corpuz, The Enterprise Information Security Policy as a Strategic Business Policy within the Corporate Strategic Plan, Information Security Institute, Queensland University of Technology, Brisbane, Queensland/4000, Australia, 2011. Available: SemanticScholar, https://eprints.qut.edu.au/51493/1/RA717ZB.pdf (Retrieved 1 of February 2020)

[19] S. Aksantijević, E. Tijan, B. Hlača, "Importance of organizational information security in port community systems", MIPRO 2009, 32nd International Convention on information and communication tecnology, electronic and microelectronics, Proceeding, Opatija, pp. 105-110

[20] European Union Agency For Cybersecurity, https://www.enisa.europa.eu/news/enisa-news/standardisation-and-the-eu-cybersecurity-act-1, (Retrieved 7 of July 2020)

[21] A. Chuvakin,Five basic mistakes of security policy, https://www.computerworld.com/article/2537565/five-basic-mistakes-of-security-policy.html, (Retrieved 5 of February 2020)

[22] S. Walden, Three Business Use Cases for Digital Forensic, https://www.delltechnologies.com/en-us/perspectives/three-business-use-cases-for-digital-forensics/, (Retrieved 5 of February 2020)

[23] D. Adu, A. Oyeola, Nigeria: Conducting Internal Investigations In Organisation, http://www.mondaq.com/Nigeria/x/860156/Health+Safety/Conducting+Internal+Investigations+In+Organisation, (Retrieved 7 of February 2020)

[24] P. Lohberg, The Data Recovery Expert Vs The Computer Forensics Specialist, https://www.ontrack.com/uk/blog/pieces-of-interest/what-sets-the-data-recovery-expert-apart-from-the-computer-forensics-expert/, (Retrieved 8 of February 2020)

[25] Damage Control in Crisis Management, https://www.universalclass.com/articles/business/crisis-management/damage-control-in-crisis-management.htm, (Retrieved 9 of February 2020)

[26] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), Official Journal L 201 , 31/07/2002

[27] Study of proposal for an ePrivacy Regulation, https://www.digitaleurope.org/resources/study-of-proposal-for-an-eprivacy-regulation/, (Retrieved 11 of February 2020)

[28] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications), Brussels, 10.1.2017 COM(2017) 10 final, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52017PC0010&from=EN, (Retrieved 11 of February 2020)

[29] OLAF – European Anti-fraud Office Digital Forensics Operation Leaflet https://ec.europa.eu/anti-fraud/sites/antifraud/files/digital_forensic_leaflet_en.pdf, (Retrieved 7 of July 2020)

[30] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC (Text with EEA relevance.), Official Journal of the European Union L 295/39

[31] Regulation (EC) No 1073/1999 of the European Parliament and of the Council of 25 May 1999 concerning investigations conducted by the European Anti-Fraud Office (OLAF), Official Journal L 136 , 31/05/1999

Using Convolutional Neural Network for Chest X-ray Image classification

Matija Sorić [1,2], Danijela Pongrac [1], Iñaki Inza [2]
[1] Zagreb University of Applied Sciences, Zagreb, Croatia
[2] University of the Basque Country, Donostia-San Sebastián, Spain
msoric@tvz.hr

***Abstract* - Chest X-ray is an imaging technique that plays an important role in pneumonia diagnosis. Owing to the high availability of medically-oriented image datasets, great success can be achieved using convolutional neural networks (CNNs) in the recognition and classification of these images. Since previous research has shown CNNs to perform as well as the best clinicians in diagnostic tasks, they caused great excitement among researchers. In this paper, convolutional neural network (CNN) machine learning (ML) model was built using a supervised dataset. The dataset used contained both pneumonia and non-pneumonia images, which the model had to classify correctly. In the end, the model is demonstrated to have achieved satisfactory results, with the high accuracy of 90.38%, 98.21% recall and 87.84% precision.**

Keywords – convolutional neural network, classification, deep learning, X-ray imaging

I. INTRODUCTION

Due to the rapid development of the X-ray imaging technique, the physiological state of tissues in the human body can now be examined noninvasively. Since every imaging examination should always begin with conventional radiography, X-ray imaging has been the main diagnostic method for diagnosing pneumonia since its inception.

The growing availability of electronic medical data is a great opportunity to discover and develop new healthcare-related technologies. However, as standard systems and programs are incapable of handling such large datasets, new computational techniques are required for data scientists to use this data. This is where artificial intelligence (AI) and machine learning can be of assistance.

Image classification is one of the most challenging problems in healthcare. It essentially aims to classify medical images into different categories which would help physicians with diagnosis and further examinations. Classification is usually performed by extracting certain features from the image and classifying images based on those features.

Until now, physicians have relied on their professional experience to extract important features - usually a laborious and time-consuming task. In that respect, *deep learning* (DL) has become one of the hottest research areas in computer science and computer applications. [1]

Hence, the goal of this study was to evaluate the performance of a *neural network* (NN) using X-ray images, and most importantly, to determine whether a patient has pneumonia or not based on the images fed to the model. For this concept to be tested, 5856 X-ray images were used, previously graded by expert physicians. The results presented in the following sections, demonstrate the proof-of-concept principle using DL method for radiological image feature extraction, which can later be used for x-ray image classification.

II. DATASET ANALYSIS

Images used in this project were taken from the *Kaggle* website. Kaggle is an online community of data scientists, where users can download or publish datasets, as well as build ML models in a web-based environment and collaborate with other data scientists. [2]

Dataset is divided into 3 sets: *'train'*, *'test'* and *'validation'*. Each of these sets contains subfolders Pneumonia/Normal. Dataset structure is illustrated in Figure 1.

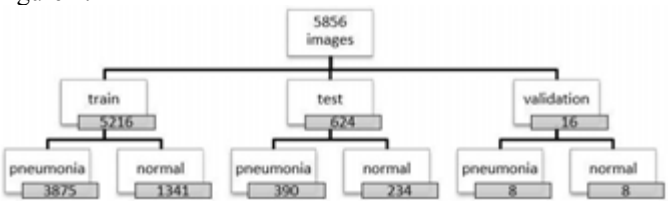


Figure 1. Dataset structure

It needs to be emphasized that patients whose images were used in the dataset are children aged 1-5. Images were taken in the Guangzhou Women and Children's Medical Center in China. Prior to being included in the Kaggle dataset, images were preprocessed, quality screened and all low quality scans removed. Furthermore, they were graded by expert physicians before taking them into account for the Kaggle upload.

Folder 'train' has the size of 1.07 GB, and, as visible in the Figure 1., 1341 and 3875 images in the subfolders 'PNEUMONIA' and 'NORMAL', respectively. Folder 'test' has the size of 75.3 MB, with 234 images in folder 'NORMAL' and 390 images in folder 'PNEUMONIA'. Folder 'val' is the smallest, having the size of 2.88 MB,

containing 8 images in the folder 'NORMAL' and 8 images in the folder 'PNEUMONIA'. As these statistics indicate that the dataset structure did not correspond to the recommended 60%-20%-20% ratio of training-test-validation data, further dataset balancing could presumably improve evaluation metrics. Likewise, the dataset was skewed since there are considerably more pneumonia than normal lung images. The ratio is approximately 72% to 28% in favor of pneumonia images. A 50% to 50% ratio would, in all likelihood, provide a better model.

III. TECHNOLOGY BEHIND MACHINE LEARNING

ML algorithms can be divided into two categories: unsupervised and supervised. Unsupervised learning is known for feature extraction, and supervised learning builds the relationship between input (e.g. patient traits) and output (outcome of interest), and is used in predictive modelling. There are also other types like semi-supervised and weakly-supervised learning. [3]

The model used in this project was built using supervised learning that requires „training" data to be labeled (tagged with the correct answer). E.g. if there is known set of columns with a new column to be predicted, a reference column is needed, for comparing predicted values with the new column values.

Moreover, the model is built using the CNN, since it is the most commonly used algorithm today for computer vision, combining high performance and execution speed. Figure 2 shows the comparison between various ML algorithms used in computer vision.

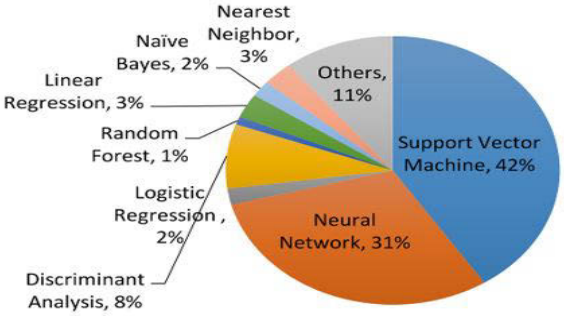


Figure 2. The most popular ML algorithms used for medical purposes [3]

A. Neural Network architecture

NN is actually a set of algorithms, operating similarly to the human brain (emulating brain's cognitive, i.e. pattern-recognition skills). Just like the human brain interprets sensory data, NN functions like a black box, taking inputs like x-ray images and processing them into one of the multiple outputs, in this case, classes. The inputs taken are numerical, contained in vectors, into which all data (e.g. images, sound, text, etc.) entering an NN must be translated. NN has numerous small units called neurons, grouped into layers mutually connected by neurons. NN architecture is depicted in Figure 3, where the first layer is the input layer which takes inputs such as images, hidden layers (all the layers between input and output layer) where computation is done, and the output layer that produces the result for an input:

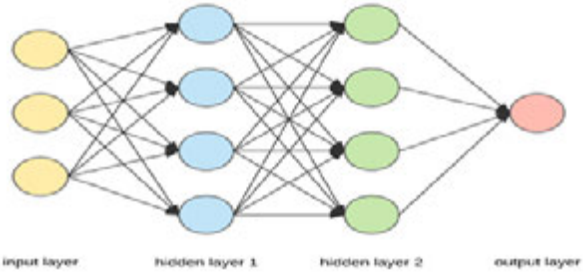


Figure 3. Neural network architecture [4]

Neurons are connected with neurons from other layers through weighted connections. As every weighted connection has a real-valued number attached, the neuron actually takes the value of the connected neuron and multiplies it with the connections weight. That sum of all connected neurons is called neurons' bias value. Bias value is mathematically transformed by putting it through the activation function, which is then assigned to the connected neuron in the connected layer. Activation function 'decides' whether a neuron should be activated or not (based on its weighted sum), thus introducing non-linearity. [5]

Operations performed by neurons are presented in Figure 4:

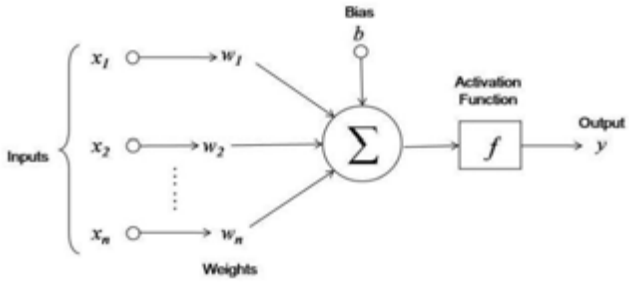


Figure 4. Operation performed by neurons [6]

In this respect, NN can be thought of as a filter which calculates possibilities, and the challenge is to obtain the correct neuron values (weights) that would allow the correct results to be computed. These weights are found by *optimization functions*.

B. Convolutional Neural Network architecture

Deep Learning (DL) is a subfield of ML based on learning multiple levels of representations by making a hierarchy of features where the higher levels are defined from the lower levels and the same lower level features can help in defining many higher level features. [7]

In the same manner, *Deep Neural Network* (DNN) is a variation of an NN with more than one hidden layer. Each of these hidden layers trains on a distinct set of features based on the previous layer's output. The more hidden layers there are in the DNN, the more complex are the features the nodes (neurons) can recognize, thus creating a better network.

One of the best DL algorithms that continuously proves its ability of diagnosing medical images with substantial performance is CNN. Its main characteristic is

its ability to learn features automatically from domain-specific images (unlike the classical ML methods). [8]

For example, a CNN was used to detect patients suffering from COVID-19, using computorized tomography (CT) images. [9]

CNNs are a class of DNNs, and are designed specifically for computer vision. Their name is derived from convolutional layers. The goal of the CNN is to shrink images into a more easily processed form, without losing the essential features critical for making predictions. The aforementioned convolutional layer is the main building block of the CNN. Although its parameters include learnable filters with a small receptive field (similar to the human visual cortex neurons), they span full input volume depth. [10]

This process is best illustrated if one imagines a flashlight illuminating only a part of an image, like in Figure 5. It is important to note that for a computer, an image is nothing more than a pixel of values:

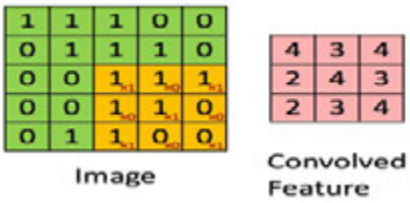


Figure 5. Convolution operation [11]

Therefore, input image is represented by a 5x5 matrix on the left. Our kernel/filter is a 3x3 matrix (flashlight) that shifts along the input matrix (stride value) and produces a new matrix on the right. Convolutional filters can be abstractly described as feature identifiers, with features represented by edges, colors, curves, etc. When an image is too large, a pooling layer is used which essentially reduces parameter number, and is represented in Figure 6:

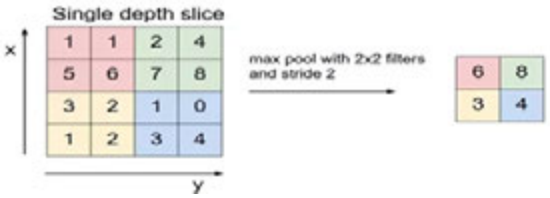


Figure 6. Max pooling [11]

Another important layer is ReLU layer, using an activation map to set negative values to zero (or another number, depending on ReLU type), thus increasing the nonlinear properties of the decision function. [12]

Hence, at first, convolutional layers detect edges and curves, but for the network to be able to detect higher level features (such as eyes or ears - if a human face has to be detected), a *fully-connected* (FC) layer is needed. It takes the inputs and outputs of an N-dimensional vector, where N is the number of classes to be predicted (in this case two). Consequently, the FC layer determines which features correlate the most with a particular class. A classic CNN architecture would look like this:

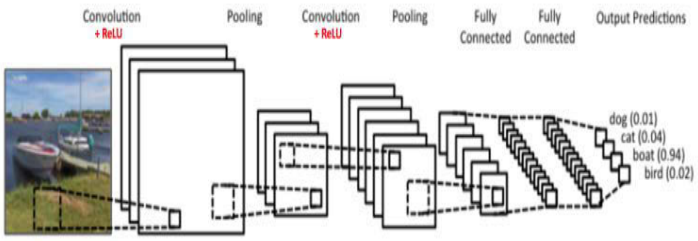


Figure 7. CNN architecture [13]

Finally, there is an FC output layer which makes the final prediction, as depicted in the Figure 7. [13]

IV. BUILDING THE MODEL

Our program code for this project, written in Python using Jupyter Notebook environment, which will be described here, can be found on *Github*. [14]

Model building commences with the libraries importation. Table I describes the standard libraries used.

TABLE I. IMPORTED STANDARD LIBRARIES

Library	Description
os	OS operations: reading files, changing or making a directory, etc.
random	generates random numbers
matplotlib	library for creating visualizations
cv2	used for reading, displaying images, saving
pandas	used for data analysis and modeling
numpy	support for arrays and matrices

Then ML libraries like Keras and Tensorflow were imported.

- TensorFlow - library for DL model production.
- Keras - an API built on TensorFlow, used to build and test NNs.

As for image preprocessing, images are resized and casted to float datatype for normalization, labels (0 and 1) are added subsequently and added to an array. Figure 8. illustrates this process (pseudo-code):

```
test_data   = []
test_labels = []

for cond in ['NORMAL', 'PNEUMONIA']:
  for img in path:
    img = read(path + 'test' + cond + img)
    img = resize(img, dimensions)
    img = cast(float)

    if cond == '/NORMAL/':
      label = 0
    elif cond == '/PNEUMONIA/':
      label = 1
    test_labels.append(label)
    test_data.append(img)

test_data = np.array(test_data)
test_labels = np.array(test_labels)
```

Figure 8. Image preprocessing (pseudocode)

The data is augmented with ImageDataGenerator class, usually used for artificial dataset expansion and the improvement of a model's ability to generalize what it has learned to new images. This is achieved through the use of image rescaling, flipping and viewing at a different magnification scale (zoom).

Following preprocessing and data augmentation, an input function is required to return the tensor, i.e. to instantiate the Keras tensor. In Table II, convolutional layers and pooling attributes are explained.

TABLE II. IMPORTED CONVOLUTIONAL LAYERS

Function	Description
conv2D	creates 2D kernel convolved with input to produce an output. It is a traditional convolution - filter sliding through the image
SeparableConv2D	variation for faster computing
MaxPool2D	used to down-sample an input (image) and reduce dimensionality.

Figure 9. shows the first convolutional layer:

```
model = Sequential()
model.add(Conv2D(16, kernel_size=(3, 3), padding = 'same', input_shape=(150,150,3)))
model.add(Conv2D(16, kernel_size=(3, 3), padding = 'same'))
model.add(MaxPool2D(pool_size=(2, 2)))
```

Figure 9. Instantiating the first convolutional layer

Another four convolutional layers were used, as well as functions that enable the use of higher learning rates; which in return accelerates the learning process. This architecture is in many respects similar to the architecture advocated by Yann LeCun in the 1990s for his image classification model (with the exception of LeakyReLU).

Activation function used was *PReLU* (*Parametric ReLU*). As previously mentioned, activation functions are an important NN feature. They decide whether a neuron should be activated. 'Activation' is used to determine an output (like yes/no). ReLU is the most commonly used activation function, especially in CNNs. Since ReLU is linear for all positive values and zero for all negative values:

- model training is computationally cheaper
- convergence is accelerated

By contrast, Leaky ReLU has a small slope for negative values, instead of zero, which accelerates the training process. Although Parametric ReLU (PReLU) is a variant of LeakyReLu, instead of having a predetermined alpha value of e.g. 0.3, it is learned automatically by NN. The difference between ReLU, LeakyReLU and PReLU is illustrated in Figure 10, where ReLU (presented as a blue line) follows the y=0 axis, LeakyReLU is depicted as a red line, and PreLU as a yellow line.

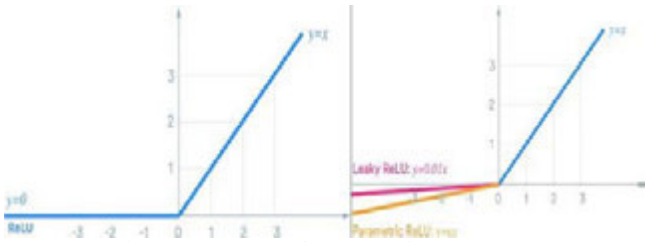


Figure 10. ReLU, LeakyReLU and PreLU [15]

Adam, an adaptive learning rate optimization algorithm designed specifically for training DNNs, was used as an optimization function for identifying the best possible weights. [16]

The Keras library implements numerous functions for creating and performing various operations with the ML model. Core layers like Dense, Flatten and Dropout are imported, and explained in Table III.

TABLE III. IMPORTED CORE LAYERS

Layers	Description
Dense	Used when association exists between features. Feeds all outputs from the previous layer to all of its neurons - each neuron provides one output to the next layer
Flatten	Converts pooled feature map to a single column that is passed to the FC layer. Flattening removes all but one dimension. Later it is passed to the Dense layer.
Dropout	A regularization technique, random neurons are ignored (dropped-out) during training. When a neuron is dropped, another takes its place, reducing the network's sensitivity to neuron weight and improving generalization

BatchNormalization, which accelerates the learning process by maintaining the activation standard deviation close to 1 and the mean activation close to 0, is also used. Pixel intensity in grayscale images (used in the project) varies from 0 to 255. Prior to inputting an image to NN, every image is transformed to an array (1D), hence every pixel enters one neuron from the input layer. If each neuron output is passed to a function, every value other than 0 (i.e. 1 to 255) will be reduced to a number between 0 and 1. Normalizing pixel values before training is common practice. [17]

The FC layer is the last layer in the model. Overfitting was reduced by regularization, i.e. *L2 regularization*, which forces the weights to take smaller values. In the output layer, activation is set to *'sigmoid'*, because the problem is a binary (two-class) classification. Sigmoid activation function exists only in the area between 0 and 1 and is therefore suitable for binary classifications.

Upon its completion, the model is compiled and run using the *model.compile( )* function. Before that, certain functions are applied during the training, called callbacks. *Callback* is a class called at various training stages. It interacts with the model during the training process. Two attributes of this class are *ModelCheckpoint*, which saves the best model obtained during training, and *ReduceLROnPlateau* which monitors a metric and reduces the learning rate if the metric stops improving. Callbacks are used because learning rate reduction is frequently beneficial for ML models.

V. RESULT EVALUATION

Once the model is built, prediction model training takes place, shown in Figure 11, with some additional arguments set, e.g. the number of training epochs, etc.

```
hist = model.fit_generator(train_gen,
                           steps_per_epoch = train_gen.samples // batch_size,
                           epochs = epochs,
                           validation_data = test_gen,
                           validation_steps = test_gen.samples // batch_size,
                           callbacks = [checkpoint, lr_reduce, early_stop, tensor_board])
```

```
Epoch 1/10
163/163 [==============================] - 209s 1s/step - loss: 0.3954 - accuracy: 0.8115 - v
al_loss: 0.7950 - val_accuracy: 0.6234
Epoch 2/10
163/163 [==============================] - 204s 1s/step - loss: 0.2953 - accuracy: 0.8763 - v
al_loss: 1.1236 - val_accuracy: 0.6334
```

Figure 11. Model training

The results were evaluated using two graphs presented in Figure 12. The graphs show the accuracy and loss of the model over a fixed number of epochs.

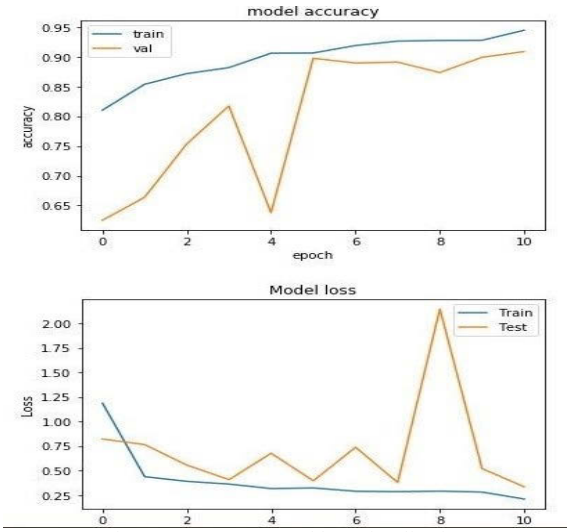


Figure 12. Accuracy and loss visualization

Figure 13 shows model performance evaluation using a series of metrics, like accuracy, precision, recall, etc.

```
Test metrics
----------------
Confusion matrix
 [[181  53]
 [  7 383]]
Accuracy: 90.38%
Precision: 87.84%
Recall: 98.21%
F1-score: 92.74%
ROC AUC: 0.952104
```

Figure 13. Model metrics

In this respect, the model given was concluded to be converging and classify X-ray images with 90.38% accuracy. The model's precision was 87.84%, and a model producing no false positives would have the precision of 1, or 100 %. Therefore, a higher number of false positives is clearly present. There were 53 false positive results, i.e. 53 people diagnosed with a disease were actually healthy. Recall lies at 98.21%, indicating a very low number of false negative diagnoses – only seven. F1 score is 91.68%, which is usually considered an excellent result.

The model could definitely be enhanced by further adjustment of parameters, layers etc. Since train accuracy was 94.45% and test accuracy 90.38%, there was probably some model overfitting. As techniques like ReduceLROnPlateau, dropouts and L2 regularization (which prevent overfitting) were used, other techniques, like training the model with more data or dataset balancing (with respect to the number of images) could be applied.

VI. COMPARING ACTIVATION FUNCTIONS PERFORMANCE

In this section, the difference between applying different rectified activation functions in CNN, namely, between ReLU, Leaky ReLU and PReLU, will be presented. Since ReLU is one of the most frequently used activation functions for ML applications, LeakyReLU can usually be found in the best Kaggle competition projects and PReLU is a variant of LeakyReLU. Performance differences were established using different activation functions. Comparison is given in Table IV.

Table IV. Activation function comparation

Activation function	**Training accuracy**	**Test accuracy**	**Training loss**	**Test loss**
ReLU	0.9460	0.8750	0.1952	0.5904
LeakyReLU ($\alpha = 0.3$)	0.9378	0.8894	0.2106	0.2640
LeakyReLU ($\alpha = 30.0$)	0.9372	0.8702	0.2296	0.2754
PReLU	0.9445	0.9038	0.2088	0.3339

A total of four models implementing different activation functions were built. The same preprocessing, regularization techniques, layers and the same number of epochs were used in all models. Therefore, the only difference between them was the usage of different rectified activations. Their performance was compared and the best model chosen for the final model.

As anticipated, the results were extremely similar, with PReLU achieving the best result. Other types of rectified activations were found to outperform ReLU in the image classification task, achieving similar performance as models created by Xu, Wang et al., described in 'Empirical Evaluation of Rectified Activations in Convolution Network' on CIFAR-100 and CIFAR-10 networks. [18]

Since PReLU outperformed the other 2 functions, achieving the highest accuracy and satisfactory loss, the activation function used in the project was PReLU. Hence, although ReLU is the most commonly used function, better alternatives exist, depending on the dataset.

Since there exist considerably larger datasets than the one used here, the performances of these functions on them are still to be investigated. However, the results suggest that this issue is worth pursuing in the future.

VII. COMPARING RESULTS WITH THE REFERENCE MODEL

To check the performance of the created model, it had to be compared with a reference model. As there are neither ongoing nor completed Kaggle competitions for this dataset, this type of comparison was impossible. The reference model was found online. A Kaggle user used a CNN based on residual NN, or ResNet. A residual NN is an NN, the architecture of which is based on constructs known from pyramidal cells in the cerebral cortex. ResNet is typically implemented with double or triple layer architecture using a ReLu activation function and batch normalizations. In this reference model, the data were augmented in a different manner and trained using a different DL library - *fastai*. The exact same dataset was used.

After preprocessing images in the similar manner as in our project, and conducting a 12-epoch training, the aforementioned reference model achieved the accuracy of 90.86 %, as seen in Figure 14: [19]

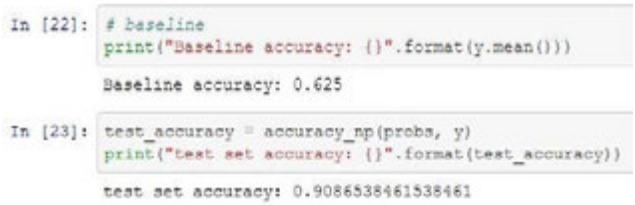


Figure 14. Reference model performance

Since the original model created in the project achieved the accuracy of 90.38%, it was found to be satisfactory and achieve acceptable metrics.

CONCLUSION

In this paper, a medical image classification algorithm based on neural networks was presented. This project shows the true power of machine learning in real-life use. Since the model clearly has satisfying metrics such as accuracy, precision, F1 score etc., it could, with further parameter adjustment (which would resolve the issue of overfitting), be used in a real-life environment. The findings of published papers that neural networks can be of exceptional help in healthcare, have been repeated and confirmed.

REFERENCES

[1] ZhiFei Lai, HuiFang Deng, "Medical Image Classification Based on Deep Features Extracted by Deep Model and Statistic Feature Fusion with Multilayer Perceptron", 12 September 2018.;.

[2] Paul Mooney, "Chest X-Ray Images (Pneumonia)", online, accessed 5/25/2020, March 2018, https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia,

[3] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, et al., "Artificial intelligence in healthcare: past, present and future, SVN – Stroke and Vacular Neurology", 0:e000101., DOI: 10.1136/svn-2017-000101, June 2017.

[4] Sai Kambampati, "What's New in Core ML 2", online, accessed 5/25/2020, June 2018, https://www.appcoda.com/coreml2/,

[5] Armaan Merchant, "Neural Networks Explained", online, accessed 5/23/2020, December 2018, https://medium.com/datadriveninvestor/neural-networks-explained-6e21c70d7818

[6] Arthur Arnx, "First neural network for beginners explained (with code)", online, accessed 5/25/2020, January 2019, https://towardsdatascience.com/first-neural-network-for-beginners-explained-with-code-4cfd37e06eaf

[7] TM Capital, "The Next Generation of Medicine: Artificial Intelligence and Machine Learning", November 2017.

[8] Asmaa Abbas, Mohammed M. Abdelsamea, and Mohamed Medhat Gaber, "Classification of COVID-19 in chest X-ray images usingDeTraC deep convolutional neural network", Mathematics Department, Faculty of Science, Assiut University, Assiut, Egypt , April 1, 2020.

[9] Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, et al. "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)", medRxiv.2020.

[10] Adit Deshpande, "A Beginner's Guide To Understanding Convolutional Neural Networks", online, accessed 5/23/2020, July 2016, https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/,

[11] Moazzem Hossain , "Easy way to understand Convolutional Neural Network", online, accessed 5/23/2020, April 2019, https://medium.com/@_moazzemhossain/easy-way-to-understand-convolutional-neural-network-its-easy-d9b7c0c5adb3

[12] V.G. Ivancevic. Darryn J Reid, Michael J Pilling, "Mathematics of Autonomy, Mathematical Methods for Cyber-Physical-Cognitive Systems", Defence Science & Technology Group, Australia, 2017

[13] Camron Godbout, "TensorFlow in a Nutshell — Part Three: All the Models", online, accessed 5/25/2020, October 2016, https://medium.com/hackernoon/tensorflow-in-a-nutshell-part-three-all-the-models-be1465993930

[14] Ayperos23, "Github (ayperos23)- pneumonia classification", online, accessed 5/25/2020, May 2020, https://github.com/ayperos23/ML-pneumonia-classification,

[15] Danqing Liu , "A practical guide to ReLU", online, accessed 5/23/2020, November 2017, https://medium.com/@danqing/a-practical-guide-to-relu-b83ca804f1f7

[16] Algorithmia, "Introduction to Optimizers", online, accessed 5/23/2020, May 2018, https://algorithmia.com/blog/introduction-to-optimizers,

[17] Cory Maklin, "Batch Normalization Tensorflow Keras Example", online, accessed 5/23/2020, June 2019, https://towardsdatascience.com/backpropagation-and-batch-normalization-in-feedforward-neural-networks-explained-901fd6e5393e

[18] Bing Xu, Naiyan Wang, Tianqi Chen, Mu Li," Empirical Evaluation of Rectified Activations in Convolution Network", arXiv:1505.00853v2, November 2015

[19] Yuan Tian, "Detecting Pneumonia with Deep Learning ", online, accessed 5/25/2020, June 2018, https://becominghuman.ai/detecting-pneumonia-with-deep-learning-3cf49b640c14

SSRCI[1]

International Conference on

SMART, SUSTAINABLE AND RESILIENT CITIES AND INFRASTRUCTURE

Steering Committee

Chairs: Adrian Boukalov, EAI Belgium, Gent, Belgium
Dragan Čišić, University of Rijeka, Croatia

Members: Miimu Airaksinen, Finnish Association of Civil Engineers, Helsinki, Finland
Darko Jardas, Institution Regional Energy Agency Kvarner, Rijeka, Croatia
Pekka Kess, University of Oulu, Finland
Srećko Krile, University of Dubrovnik, Croatia
Paul Timmers, University of Oxford, UK

Program Committee

Members: Radoslav Delina, Technical University of Košice, Slovakia
Bojan Hlača, University of Rijeka, Croatia
Goran Krajačić, University of Zagreb, Croatia
Thor Moen, Smart Innovation Norway, Halden, Norway
Dario Ogrizović, University of Rijeka, Croatia
Mile Pavlić, RIS, Kastav, Croatia
Andreja Pucihar, University of Maribor, Slovenia
Vittorio Rosato, ENEA Centro della Casaccia, Roma, Italy

[1] Conference held at the Mipro 2019

PAPERS

Creating Sustainable Solutions for Photovoltaics

W.Brenner*, N. Adamovic*
* Technische Universitaet Wien/Institute of Sensor and Actuator Systems, Vienna, Austria
werner.brenner@tuwien.ac.at

Abstract – **Aiming to achieve a low-emission to zero-emission building stock across the European Union by the year 2050, the legislative bodies of the EU have established new requirements regarding the energy efficiency performance of buildings. Recently conducted research on solar energy systems and materials has yielded promising opportunities for sustainable energy solutions, which do not only consider environmental issues but also allow aesthetic design possibilities.**

These new research findings are expected to play a crucial role in coping with the issue of Waste Electrical and Electronic Equipment (WEEE) including Photovoltaic (PV) waste. While ever-increasing amounts of PV panel waste constitute a serious environmental threat, they can also be seen as a potential source for new business fields and future ways to create value. This paper covers the visions of two European projects: FP7 SolarDesign (regarding Building Integrated Photovoltaics solutions for both new and historic buildings as well as freely shaped building surfaces) and H2020 CABRISS (which aims to establish a more circular economy and reuse for photovoltaic products). Proactive involvement of European standardization bodies helped the market obtain substantiated results.

Keywords – photovoltaics PV; building-integrated photovoltaics BIPV; standardization; sustainability; innovation; circular economy

I. INTRODUCTION

The ongoing growth of human population results in an increasing demand for industrial and agricultural products and causes the exhaustion of fossil energy sources. It induces rising levels of CO^2 in the atmosphere and, consequently, the greenhouse heating effect and environmental degradation. Up to now, Renewable Energy Sources (RES) account for only 2% of the 150.000 TWh of energy used globally [3]. 30% of the energy consumption results from industry and agriculture, 30% from transportation, and 40% from buildings. Solar energy utilization occurs in the form of PV parks and PVs installed on building roofs and facades to create electrical power as well as in the form of solar thermal collectors to heat water and space [9]. In the last decades, solar energy has been advertised as a sustainable solution and its prevalence has increased significantly. In the light of the growing importance of solar energy, extensive research and development has been conducted in this field, aiming to stimulate an efficient and wide implementation.

Solar energy and yet-to-develop technologies will play an important role in coping with growing energy demand. The energy targets set for 2020 are important factors in achieving the necessary CO^2 reduction. A steady shift towards an increased use of RES and reduction of hydrocarbon use is the target of the next decade. Despite strong promotion through support policies, this transition to renewables will be difficult and bear different challenges. Among these challenges are issues regarding optimization of energy saving and reduction of energy demand as well as the implementation of aesthetic energy solutions. Modern solar energy systems and materials seem to be a key factor in addressing these challenges. Especially photovoltaics integrated into buildings will presumably play an important role in future smart cities.

TABLE I. SMART CITY COMPONENTS

	economic attractiveness
	citizen-friendly government
	open data
	smart building technology
	digital citicens
	mobility and infrastructure
	resource friendly (contribution by H2020 CABRISS)
	efficient use of energy (contribution by FP7 SolarDesign)

Not only PV solar farms in rural areas are increasing in number, but also PV systems which are integrated into the envelope of buildings within urban environments [9]. Development of PV systems consisting of components that enable extensive design possibilities are of great importance in the field of architecture.

Wide freedom in formal design is a crucial requirement in architecture. In the context of building, photovoltaic installations which are fully integrated into the envelope of the building and function as building material are classified as BIPV (building-integrated photovoltaics). On the other hand, systems that are simply linked to the building through a support structure, e.g. on the roof, are classified as BAPV (building-attached photovoltaics).

Through promotion of collective research and development, Europe's building sector can potentially evolve into energy innovative, energy efficient industry with great global competitiveness by the year 2030.

TABLE II. CHALLENGES FOR THE BUILDING SECTOR WHEN INTEGRATING PV SYSTEMS IN ENVELOPES [23]

	Challenges
Technological	Meeting the design performances of integrated envelopes; Long term assessment of the contribution of each envelope component to the system performance Availability of data on material characteristics and energy needs Industrialization of reasonable building envelope solutions; Characterization methods and standards covering high volume / high performance eco construction materials Guidelines for dismantling the building already during its design (with regards to lifecycle assessment approaches [16])
Economic	Integration of reasonable envelope components meeting the final performance targets; Return on investment for refurbishment projects is not attractive;
Societal	Building owners are not open enough to make sustainable choices; Design can show varying traits due to different regional styles. Involvement of product designers and specialists in the field of PVs from multiple EU member states will result in a wider inclusion of different design approaches; High intrusiveness in refurbishment intervention; New multifunctional components require continuous training of craftsmen. Lack of guidelines for the evaluation of the historic value and quality of existent building Dispersed ownership in large housing/buildings prevents from refurbishment; Valuing not only energy performances but also environmental, aesthetics, historic value, acoustics, safety, accessibility or comfort as purchase criteria for end users;
Regulatory	Nationally distinct legal requirements across Europe like building regulations impede prefabrication of standardized components; Regulations for public decision often promote lowest prices, not best performances Regulations for historic buildings preservation may be complex and vary considerably from country to country Existence of quality standards that encompass the whole life cycle of any building, thus guaranteeing durable building performances

II. BUILDING INTEGRATED PHOTOVOLTAICS (BIPV) – DESIGN, INNOVATION AND ACCEPTANCE

SolarDesign (On-the-fly alterable thin-film solar modules for design driven applications [2, 5, 19]), an FP7 project carried out with EU-funds, puts a focus on the development of flexible and easily used PV elements which utilize modern solar cell materials and newly evolved production processes. The project's outcome are novel printing and scribing methods which allow the manufacturing of flexible monolithically interconnected photovoltaic modules. These new monolithic interconnection technologies facilitate the manufacturing of thin-film PV modules. SolarDesign's new developments make on-the-fly alterations of the solar element's electrical properties possible. Furthermore, they allow extensive customization of the PV modules' shape, size and voltage. Through application of this technology, SolarDesign achieved to manufacture a solar foil with a width of 300mm, which allows curved solar cells with a minimal radius of 10mm.

The project addressed creative industries and adjacent industries alike. The creative industry receives novel materials and a new freedom-of-design for the development of solar powered products. Manufacturing industries can upgrade their products and productions by full integration of PV materials. The innovation of the SolarDesign BIPV prototype was in a sophisticated solution for the customized modules (in shape and area) for a geodetic or parametric architecture. This is possible thanks to the process, that studied the needs of the clients (geodetic or parametric structure developers companies, architecture practices or building companies specialized in cladding) and developed the necessary steps to design the required on-the-fly modules. The final prototype developed within the SolarDesign consortium was the demonstrator of the feasibility for the specific custom designed module technic.

The principal advantages of SolarDesign approach are: overcoming the square/rectangular solar module by using PV as a versatile building integrated component, improving the efficiency of the PV module by integrating in new morphologies, increasing the receiving surface by escaping from the concept of "optimal orientation" and increasing and improving the participation of the PV manufacturer in all the process. The SolarDesign approach provides the characteristics that are not available in the same product or service: PV modules on-the-fly that could be integrated in complex architecture. In addition to producing electricity, this kind of BIPV installations adds structural value and can enhance a building's aesthetic quality and design (colors, customization and at the same time remain cost competitive). Aesthetics and enhanced freedom of design are key stimulants for an increasing prevalence of Building-Integrated Photovoltaics (BIPVs) [19]. The possibility to generate electricity by replacing conventional building elements with photovoltaic elements makes BIPVs a promising green energy concept. This technology has the potential to meet the demand for different shapes, sizes, colors and transparency levels which cannot be met by conventional, rigid PV elements. Previous PV modules are limited to fixed shapes and certain electric outputs. While these characteristics might not be detrimental in large scale solar power plants, they bring limitations and can cause increased costs and lower efficiency when applied to buildings or other solar-powered devices.

Several studies show the environmental advantages of photovoltaic energy generation [24, 25, 9]. In accordance with these studies, solar energy generally is perceived as a green and sustainable energy source, although the level of the user's environmental concern must be taken into account [26, 16]. On the contrary, other studies point out certain adverse environmental effects to be considered [28]. In the field of architecture, aesthetic aspects are of

utmost importance in designing. Pure functionality is not always sufficient, as the form must follow both functional and aesthetic aspects. Acceptance by society is a definite prerequisite for a successful furtherance of solar energy as a key future energy source. [30]. Especially in urban landscapes, a key decisive influence on the end consumer's preference is the visual appearance of the various options [26, 31]. Thus, it is necessary to bear in mind not only functional, but also aesthetic viewpoints when aiming to strengthen the use of PV. Nevertheless, studies on the aesthetic perception of integrated photovoltaic elements indicated wide acceptance results [26, 32].

Architectural design objectives sometimes can decrease energy efficiency, which potentially makes a certain degree of customization advisable. PV material use is not limited to elements that form the building's roof or façade. By application of semitransparent PV materials, even windows with PV characteristics and luminescent solar concentrators can be achieved. Semitransparent PV windows can create solar electricity while simultaneously reducing the heating-up of the building. This field incorporates a wide range of different PV technologies, such as specially arranged Si-based (which allow light to permeate through the spaces between the opaque cells) or thin films with see-through quality, such as a-Si, CIGS and organic PV [19].

Since buildings in urban areas collectively shape a city's appearance and thus the daily environment of many people, their aesthetic design is not only of relevance for the buildings' inhabitants or the people working there, but also for the entirety of citizens of the city. Installations of new energy generation structures within an urban landscape could cause disapproval if they seem to impair the city's overall appearance. Finding ways to conceal PV installations or to integrate them into buildings discreetly, seamlessly or in an aesthetically pleasing manner is an important task.

It should not be neglected that the professional background of the observer influences how positively PV structures are perceived [29, 27, 18]. How well PVs can be implemented into a building is not only dependent on the technical feasibility of incorporating the PV into other construction materials, but also on factors for harmonious inclusion, such as color and texture.



Figure 1. Color variations providing more freedom of PV design, using different novel encapsulants allows custom designed optical appearance (source: CEA) [22]

Especially color is a key creative variable and an important tool for architects to achieve harmonious aesthetic design. Color characteristics of BIPV should therefore be highlighted in upcoming research. Similarly, (surface) texture is an important architectural design medium which could be examined further in BIPV research.

III. THE CREATION OF SUSTAINABLE SOLUTIONS FOR PHOTOVOLTAICS

In this context, Europe is paying increasing attention to innovation issues as well as commercialization of research findings. Aiming to enhance Europe's competitiveness through promotion of innovation, bridges between research and standardization were built within the last years [12]. While the outcomes of research projects constantly need to be regarded by standardizers, researchers are, vice versa, dependent on information on standards that are already applicable or currently being developed. The implementation of new standards itself may even necessitate additional research activity, for example when appropriate test methods are required. A target of European legislation is the development of environmental-friendly solutions which aim at the issue of Waste Electrical and Electronic Equipment WEEE [5, 6] which also entails PV waste [4]. WEEE covers a broad variety of materials which act as a major potential threat to the environment if not handled appropriately.

CABRISS [7] aims at stimulating a circular economy that manages the critical situation of recycling the significant volume of photovoltaic waste. The project aims to benefit also to electronics, metallurgy and glass industries by these means. The H2020 project CABRISS bundles the efforts of 16 European companies and research institutions.

The project comprises the development of: (i) recycling technologies to recover In, Ag and Si for the sustainable PV technology and other applications; (ii) a solar cell processing roadmap, which will use Si waste for the high throughput, cost-effective manufacturing of hybrid Si based solar cells and aims at demonstrating the possibility for the re-usability and recyclability at the end of life of key PV materials. From the begin the CABRISS consortium decided to have all results in accordance with valid European standards to ease access to market.

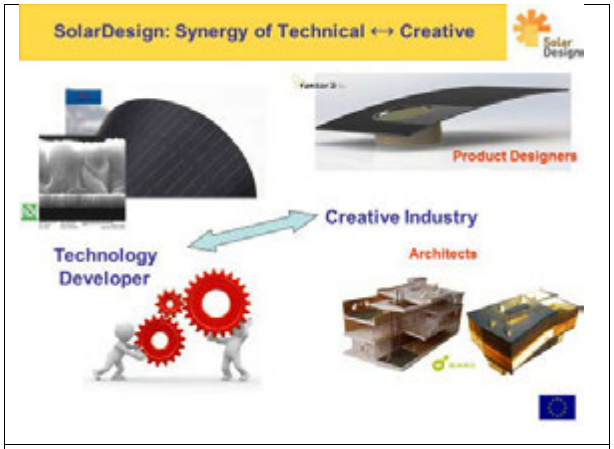
TABLE III. CABRISS' MAIN OBJECTIVES

Task	Activity
Collection	Achieving a collection of up to 90% of the PV waste throughout Europe
Retrieving	Retrieving up to 90% of the high value raw materials from the PV cells and panels: Si, In, Ag
Symbiosis	Establishing industrial symbioses by providing raw materials such as glass or silver as feedstock for other industries
Costs	Lowering the cost of manufacturing PV cells and panels from the recycled raw materials
Involvement	Fostering involvement of the EU citizens and industry

To achieve a long-lasting shift towards a more sustainable future, changes in consumer awareness, consumer behavior and consumer choices must be accompanied by supportive frameworks.

Different aspects are to be considered when assessing the relationship between product aesthetics and sustainability. How the design of a product is aesthetically perceived by users can indirectly have significance from a sustainability-focused viewpoint.

TABLE IV. FOUR STATES OF SUSTAINABILITY-RELATED EFFECTS ON THE USER [19, 37]


Product aesthetics directly influence (potential) buyers before and during the phase of purchasing
In the phase of usage, the product's design should keep the user satisfied. How well the product is perceived in this phase influences the user's appreciation for and relation with the product, which psychologically results in prolonged product durability.
The aesthetic characteristics of a product can emotionally influence in what way the user interacts with the product.
User's attitude regarding both the product itself and its brand as well as the environment can be directly affected by the product's appearance and design.

Main target stakeholders to be addressed: standardization bodies, certification entities, photovoltaic systems suppliers, installers and service providers for energy efficient buildings and solar-powered consumer products, training providers, local authorities & national/regional public bodies, architects' associations, solar-powered consumer users, construction companies associations, public and private real estate promoters, clients and users (citizens), network operators, energy management agencies.

A concept for a PV driven street lamps was developed which illustrated the interdependence between design, standards and the resulting need for minimum efficiency of involved PV materials to guarantee the given luminescence.

Shape and size of the street lamp's reflector define the aesthetic appearance. Thus, the diameter is limited to avoid a clumsy or bulky appearance. This restricted surface defines the available PV bearing area and consequently the affordable electrical energy which can be converted into light.

SolarDesign had to solve the research problem to increase the efficiency of the PV material and especially its unfavourable temperature coefficient to guarantee the standardized luminescence in the street.



Figure 2. PV driven street lamp, illustrating the need for research concerning a reduced temperature coefficient of the involved photovoltaic material to guarantee minimum values of luminescence (standardized for street lighting in Standard "Road lighting CEN/TR13201-1,2,3,4), source: SIARQ)

The efficiency decrease of ZnS-buffered CIGS (Copper, Indium, Gallium, Diselenide) is smaller than for CdS-buffered CIGS. With further improvement of alternative buffers (like ZnS, ZnMgO, Zn(O,S)), the efficiency values of CdS-buffered CIGS at 25°C can be reached. Due to lower temperature coefficient even higher efficiencies at operating temperatures (typically ~60°C for midsummer day) can be expected for alternative buffers – without using toxic Cd.

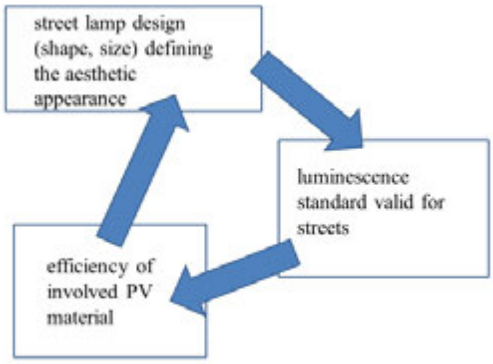


Figure 3. Example of the interdependence of design, standards and PV efficiency

The developed model is used to meet the needs of the designers/architects. Specific user defined shapes of modules can be easily achieved, while respecting the electrical requirements. The simulations of the solar modules under different illumination levels have been taken into account, thus enabling the calculation of the daily or yearly energy performance of the considered modules [4].

Since FP6 researchers are encouraged by EC to involve standardization in their projects and in the development of proposals. Furthermore, taking part in the work of Technical Committees (TCs) is expected. Thus standardization bodies and electro-technical committees can act as a bonding link between researchers and industry. This can be achieved by spreading the outcomes of research and by assuring communication and close cooperation between the key actors when standards are being developed. The following benefits will be achieved [10] : better networking and interaction with other communities;

new areas for standardization and new markets for commercialization of research findings can be identified; increase of the number of competent researchers in Technical Committees will increase; involvement in research projects will be a benefit for the competence of standardization experts; awareness amongst policymakers and stakeholders about standardization organizations and their impact on innovation processes will increase.

The model of the considered PV recycling system is presented in Fig. 4. It starts with the collection, the pre-sorting and transportation of end-of-life PV modules. The next step is a recycling plant where disassembling and separation of the main components (frame, wires, eventually glass, cell material etc.) are carried out. Material products of these manufacturing steps can be sold to the market as basic materials and pre products, but could also stay in the recycling loop (as it will be assumed here). Since we are focusing on specific valuable materials like silver and indium and also silicon semiconductor technologies, it can be assumed that the frame metals (mainly Al) and glass will be delivered to the respective markets and not stay in the loop. Plastic films like back side films (back sheet) or intermediate plastic layers (encapsulant) will be used as an energy carrier for thermal processes. At this stage of the recycling process, more than 96 % of total EoL PV mass leaves the PV recycling process or was already used in the process (e.g. thermal energy).

Materials for further use in the recycling process are mainly Si cell materials or valuable materials from thin film modules. After purification and pulverization, Si is used in different processes to provide crystalline Si ingots, hot pressed ingots, wafers or low cost substrates with the following cell and module production. In a parallel pathway to the innovative c-Si PV production, the production of new thin film PV modules based on recycled materials and a roll-to-roll process is analyzed. Once again, the valuable materials like In or Ag which are an outcome of the first recycling steps can be sold to the market or can be re-entered into the recycling process.

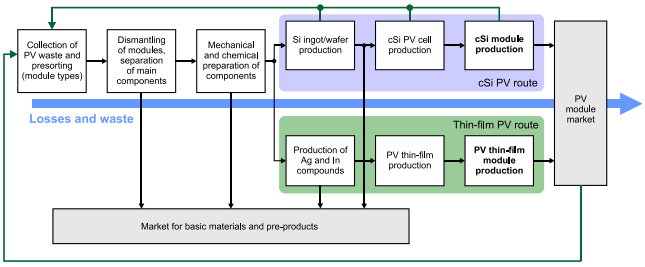


Figure 4. The PV recycling system with its main processes and pathways [28]

An additional generator for innovations is the Raw Materials Initiative [13] which addresses the challenges Europe is facing with regard to access to raw materials. Such critical raw materials are a main focus of the EU Circular Economy Action Plan [15]. CABRISS and SolarDesign in an early phase established contacts to SEMI [10] and CEN/CENELEC [11] for a closer cooperation and knowledge exchange. SEMI is the global industry association which serves the manufacturing supply chain for the micro- and nano-electronics industries. CABRISS proactively cooperates with SEMI's "TC Photovoltaics", CENELEC's CLC/TC 111X "Environment" and CLC/TC 10 "Energy-related Products". The CEN/CENELEC 45550 to 45559 series cover the aspects of recyclability, product durability, reparability, reusability, recycled content, ability to remanufacture, and product lifespan.

IV. CONLUSION

Smart systems and control can allow energy usage optimization whilst guaranteeing optimal comfort, a healthy environment and numerous other services. Existing buildings could have high insulating envelopes to reduce energy use much below 50 kWh/m²/year while achieving thermal comfort Two projects are presented which contributed to these requirements. The focus of CABRISS lies on the collection, de-pollution and treatment technologies for end-of-life PV as prerequisites for recovery and recycling. By proactive involving European standardization, validity of results is guaranteed for the market. While growing photovoltaic (PV) panel waste embodies a new environmental challenge, it also bears unprecedented opportunities to create value and to open new economic paths. On the one hand, SolarDesign addressed the need for customized PV cells that are flexible and easy to integrate through the development of novel solar cell materials and manufacturing processes. SolarDesign developed new scribing and printing technologies that enable monolithically interconnected flexible solar modules.

ACKNOWLEDGMENT

The FP7 NMP SolarDesign Project has received funding from the European Union's Seventh Framework Programme under grant agreement n° 310220.

CABRISS has received funding from the European Union's Horizon 2020 research and innovation program, under grant agreement No 641972.

REFERENCES

[1] CEN/CENELEC, "How to link standardization with EU research projects: advice for CEN and CENELEC members," February 2015, https://www.cencenelec.eu/research/news/publications/Publications/BRIDGIT-members-guide.pdf

[2] CEN/CENELEC, "Research Study on the benefits of linking innovation and standardization," Dec 2014, Ref: J2572/CEN, CEN/CENELEC, http://www.cencenelec.eu/research/news/publications/Publications/BRIDGIT-standinno-study.pdf

[3] Y. Tripanagnostopoulos, et al., "Solar energy materials and systems for and aesthetic and sustainable future," Contemporary Materials, V−2 (2014), pp. 172-185, Review UDK 620.91:66.017−022.53 , doi: 10.7251/COMEN1402172T

[4] W. Brenner, N. Adamovic, "A circular economy for photovoltaic waste – the vision of the European project CABRISS," in P Biljanovic (ed.): mipro2017 proceedings, ISSN 1847-3946, ISBN 978-953-233-093-9, pp. 154-159, doi: 10.23919/MIPRO.2017.7973407

[5] IRENA and IEA-PVPS, "End-of-life management: solar photovoltaic panels," International Renewable Energy Agency and International Energy Agency Photovoltaic Power Systems. IRENA and IEA-PVPS (2016), http://iea-

pvps.org/index.php?id=95&eID=dam_frontend_push &docID=3222

[6] Directive 2012/19/EU of the European Parliament and of the Council of 4 July 2012 on waste electrical and electronic equipment (WEEE) Text with EEA relevance, Official Journal of the European Union L 197/38 , Document 32012L0019 .

[7] H2020 project CABRISS **"**Implementation of a circular economy based on recycling, reused and recovered indium, silicon and silver materials for photovoltaic and other applications", Website https://www.spire2030.eu/CABRISS

[8] European Commission, "The role of waste-to-energy in the circular economy," communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee oft the Regions, Brussels, 26.1.2017, COM(2017) 34 final

[9] N. Sánchez-Pantoja, R. Vidal, M. Carmen Pastor, "Aesthetic impact of solar energy systems," Renewable and Sustainable Energy Reviews 98 (2018), pp. 227–238, https://doi.org/10.1016/j.rser.2018.09.021

[10] SEMI, http://www.semi.org

[11] CEN/CENELEC, "Why standards," http://www.cencenelec.eu/research/WhyStandards/Pages/default.aspx

[12] STAIR "An integrated approach for standardization, innovation and research," CEN/CENELEC, https://www.cencenelec.eu/research/news/publications/Publications/STAIR.pdf

[13] Commission's Communication "List of Critical Raw Materials for the EU", Brussels, COM(2017) 490

[14] Commission's Communication "Investing in a smart, innovative and sustainable industry: A renewed EU industrial policy strategy," COM(2017) 479

[15] Commission's Communication "Closing the loop – An EU action action plan for the circular economy," COM(2015) 614

[16] P. Zhai, E. Williams, "Analyzing consumer acceptance of photovoltaics (PV) using fuzzy logic model," Renewable Energy 2012;41:350–7. https://doi.org/10.1016/J.RENENE.2011.11.041

[17] CEN/CENELEC, "IPR and Standards in Research," https://www.cencenelec.eu/research/innovation/IPR/Pages/default.aspx

[18] S. J. Zafarmand, K. Sugiyama, M. Watanabe, "Aesthetic and sustainability: the aesthetic attributes promoting product sustainability," The Journal of Sustainable Product Design *3*:173–186, 2003 , DOI 10.1007/s10970-005-6157-0 ,

[19] Final Report Summary – FP7 SolarDesign (On-the-fly alterable thin-film solar modules for design driven applications) https://cordis.europa.eu/project/rcn/105949/reporting/en

[20] M.-Ch. Hoffmann, H. Suitner, R. Thomas, "H2020 CABRISS public businessplan," https://doi.org/10.5281/zenodo.998558

[21] https://internetofthingsagenda.techtarget.com/definition/smart-city

[22] W. Brenner, N. Adamovic, "Thin-film CIGS solar modules for design driven applications in the frame of the FP7 NMP project SolarDesign" in D. Sabolic (ed.): mipro2015 proceedings, ISSN 1847-3946, ISBN 978-953-233-083-0, pp. 137-142, DOI: 10.1109/MIPRO.2015.7160252

[23] Publications Office of the European Union, 2013, , European Union, 2013, Energy efficient buildings, Multi-annual roadmap for the contractual PPP under Horizon 2020, ISBN 978-92-79-31239-7 , doi: 10.2777/29993

[24] R. Wiser, D. Millstein, T. Mai, J. Macknick, A. Carpenter, S. Cohen et al. „The environmental and public health benefits of achieving high penetrations of solar energy in the United States," Energy 2016; 113, pp. 472–86. http://www.sciencedirect.com/science/article/pii/S0360544216309938

[25] D. Turney, V. Fthenakis, "Environmental impacts from the installation and operation of large-scale solar power plants," Renew Sustain Energy Rev 2011; 15, pp. 3261–3270. https://doi.org/10.1016/J.RSER.2011.04.023.

[26] M. Lu, A. Lin. "The impact of photovoltaic applications on urban landscapes based on visual Q methodology," Sustainability 2018; 10, pp.1–15. https://doi.org/10.3390/su10041051.

[27] S. Naspetti, S. Mandolesi, R. Zanoli, "Using visual Q sorting to determine the impact of photovoltaic applications on the landscape," Land Use Policy 2016; 57, pp. 564–573, https://doi.org/10.1016/j.landusepol.2016.06.021.

[28] W. Brenner, N. Bednar, P. Biermayr, N. Adamovic, "Standardization and Life Cycle Cost Assessment Approach in Circular Economy for Photovoltaic Waste," in: Proceedings of 3rd International Conference on Smart and Sustainable Technologies (SpliTech) 2018, IEEE Conferences

[29] R. Gifford "Decoding modern architecture. A lens model approach for understanding the aesthetic differences of architects and laypersons," Environ Behav 2000; 32, pp. 163–187

[30] J. Zoellner, P. Schweizer-Ries, C. Wemheuer, "Public acceptance of renewable energies: results from case studies in Germany," Energy Policy 2008; 36, pp. 4136–4141, https://doi.org/10.1016/J.ENPOL.2008.06.026.

[31] Q. Bao, T. Honda, S. El Ferik, M.M. Shaukat, M.C.Yang, "Understanding the role of visual appeal in consumer preference for residential solar panels," Renewable Energy 2017; 113, pp. 1569–1579,. https://doi.org/10.1016/j.renene.2017.07.021

[32] E. Strazzera, V. Statzu, "Fostering photovoltaic technologies in Mediterranean cities: consumers' demand and social acceptance," Renewable Energy 2017; 102, pp. 361–371. https://doi.org/10.1016/j.renene.2016.10.056.

Use of Drone to Improve Healthcare Efficiency and Sustainability

L. Faramondi[a*], G. Oliva[a], L. Ardito[a], A. Crescenzi[a], M. Caricato[a], M. Tesei[b], A. Onetti Muda[c] and R. Setola[a]

Abstract— **Following the Industry/Hospital 4.0 paradigm, the cost and effectiveness of the services offered by a modern Hospital can be dramatically improved by resorting to technology and by putting the patients (and the medical practitioners) at the center, thus improving their care experience (and work time), improving efficiency and paying more attention to the environment. In this paper, we focus on how to improve efficiency and cost-effectiveness of intra-operative consultations; instead of letting the pathologist roam around several hospitals, each provided with complex and costly equipment (e.g., cryostat) and professional resources (e.g., specialized technicians), we propose an approach where drones are used to move biological material (non-fixed biological samples) from small hospitals and clinics to a larger hospital, provided with the necessary infrastructure for conducting intra-operative consultations. In particular, we aim at demonstrating that technology can help creating value by minimizing inefficiencies due to under-usage of professional resources (e.g., the pathologist should maximize the amount of time spent doing his/her job, and such a job should be done in an environment where the pathologist is not alone) and assets (e.g., the complex infrastructure required to manage intraoperative consultations should not be replicated in every facility), and transportations of goods (e.g., increasing efficiency by using lightweight devices to transport samples instead of medical cars to transport specialized practitioners).**

I. Introduction

Intra-operative consultations are a critical, life-saving tool during surgeries. However, for economical and organizational reasons, hospitals may not have technological (e.g., cryostat, microscope) or professional (pathologists or technicians) means to conduct them. Classical ways to face such issues include medical cars bringing personnel from one hospital to another. Such solutions have high costs for equipment, under-usage of resources, and high environmental impact. Telemedicine could represent a valid solution, but it does not avoid the costs for equipment and technicians in each hospital. Furthermore, it is worth noting that the healthcare system has been largely changing in the last years moving towards a structure where high specialized centers are surrounded by several no-specialized small structure dispersed in the territory, which often ask high specialized centers for i intra-operative consultations.

According to the foregoing discussion, following the Healthcare 4.0 paradigm, this paper aims to present a new system of hospital-to-hospital drone transportation applied to the transportation of ***non-fixed***[1] samples for intra-operative consultation. This goal will be handled from both a technological and managerial perspective. Indeed, on the one hand, the technological feasibility of the system will be presented, and preliminary application case will be proposed. On the other hand, the value for individual beneficiaries (e.g. the hospital, the patient, and others stakeholders) - including economic, environmental, social aspects - will be unveiled through a value Proposition Canvas (VPC), eventually proposing a strength, weaknesses, opportunities, threats (SWOT) analysis for the proposed system.

We would like to point out that the application of drones to the transportation of non-fixed biological samples is ***completely new*** and, to the best of our knowledge, there is ***no literature on this topic***.

The reminder of the paper is structured as follows. After a last section containing an illustrative example of the proposed approach, which concludes this section, Section 2 reviews from one side existing approaches and attempt to transport biological material via drones and from another side reviews emerging literature on digital transformation, especially in the healthcare sector. Section 3 describes the drone-based system from a technological perspective. The fourth section highlights beneficiaries of the system and related pains and benefits, ultimately proposing a SWOT analysis. Finally, discussion and future development are presented.

II. State of the Art

A. Drone-based transportation of biological material

In recent years, several initiatives have been carried out regarding drone-based transportation of biological material. In particular, the Johns Hopkins University School of Medicine has recently realized a preliminary proof-of-concept study related to drone transportation of blood samples by analyzing the relations among the flight parameters, the destruction of cells, and the coagulation aspects [1], [2].

Another research project involves the "Ospedale Civico di Lugano", the "Ospedale Italiano di Lugano", and the Swiss Posts, this project takes into account also the transportation of medicines [3]. Recently, also Sant'Anna University of Pisa started a similar experimentation [4].

In May 2018, the Italian air navigation service provider (ENAC) announced the completion of the tender process for the selection of its industrial partner in the development

[a] Università Campus Bio-Medico di Roma, via Álvaro del Portillo 21, 00128, Rome, Italy.
[b] Consorzio Nazionale Interuniversitario per i Trasporti e la Logistica (NITEL), Rome 00198, Italy.
[c] Department of Laboratories, Bambino Gesú Pediatric Hospital – IRCCS, Rome, Italy.
* Corresponding author, email: l.faramondi@unicampus.it

[1] Note that non-fixed samples have to be considered because, due to the strict time limits of intraoperative consultation, there is not enough time to fix the sample in formalin or equivalent medium (a procedure that would require hours).

of a platform to provide Unmanned Aerial Vehicles Traffic Management (UTM) services, i.e. to manage the so-called "drones" traffic.

Moreover, the Local Sanitary Agency of southern Tuscany has recently started a project aiming at transporting sanitary material [5].

To conclude, it is worth mentioning that, in collaboration with WindTre (Italian telecommunication company) and Archon (startup for the development of innovative drones), we have participated to the *Drones for Healthcare* project, aiming at designing a drone able to bring defibrillators in emergency scenarios (e.g., in geographically dispersed areas). In this context, WindTre had the duty to enable the drones produced by Archon to use their 4G network for navigation and live streaming of an operator providing indications on how to use the defibrillator [6], [7].

The project, ended on September 2018, emphasized the large attention about this topic, both for authorities and stakeholders, and a concrete interest of telecommunication companies in the usage of drones as a useful scenario in the 5G network.

B. Digital transformation in the healthcare sector

In the recent years, the term Industry 4.0 has become more and more important. It calls for the confluence of many new digital solutions (e.g., additive manufacturing, augmented reality, big data analytics, Internet of Things) [1], [8] in the industrial context. This digital transformation of companies is deemed innovative and useful to create novel business opportunities and improves efficiency and effectiveness of firms by improving their financial and innovative performance. For instance, predictions reveal that benefits in growth performance of 90-150 billion over the next 5-10 years and productivity improvements of about 15-25% [2]. As a result, many governments (e.g., German, Italian, UK) are pushing firm towards a digital transformation (e.g., [3], [4]).

More recently, the concept of Industry 4.0 has been translated to the healthcare sector. That is, it has also been called for a digital transformation within hospitals; thus, by analogy to Industry 4.0, "Healthcare 4.0" and "Hospital 4.0" have emerged as new concepts to denote that the novel approaches in the industrial context driven by Industry 4.0 can be adopted in hospitals and healthcare services (e.g., [5]). Especially hospitals are under tremendous pressure for digitalization, in that it has the highest potential to be (positively) revolutionized. In line with this reasoning, digital solutions are (slightly) entering hospitals. For example, there exist attempts to implement automated logistics services [6], systems for gathering medical data, anonymizing it, cleaning it, making it interoperable [7], and automated surgery robots. Still, many other potential applications can be adopted, and several types of technologies can be used, so making the process of digital transformation in hospitals as one of the most complex, albeit with the highest growth potential. With this in mind, we propose drone-based system for non-fixed intra-operative consultations since, so far, the use of drones has been widely theorized but scantly applied to this context.



Fig. 1. Low-cost initial experimental setup.

III. Drone-based system: first experimental setup

In this section, we describe a simple, low-cost setup to demonstrate the potential of the proposed approach, as well as a set of initial tests.

Figure 1 shows the setup, which consists of a DJI advanced phantom 3 quadcopter. Before starting the customization phase, it has been necessary to analyze the maximum payload considering the addition of the transportation compartment and the material for its cooling. Due to reasons of space, weight, and stability of the drone during the flight operations, the additional compartment has been placed on the bottom of the drone by removing the HD camera and the gimble, paying attention to not interfere with the ultrasonic sensors used to estimate the quote of the drone during the takeoff and landing phases. The additional compartment consists of a 70 x 90 x 140 mm ABS (Acrylonitrile Butadiene Styrene) box, 130 g total weight. The thickness of the walls is about 6 mm. With the aim to maintain the internal temperature of the compartment as low as possible, the box is internally lined with Aerogel panels. The case contained the biological sample (weighting around 5g) and three cells of synthetic ice (60 x 50 x 10 mm) in order to keep the temperature as low as possible, and a temperature data logger was used to record the internal temperature of the box during the flight.

Figure 2, shows the trajectory of the drone during one out of several test flights, which were performed in a private area for about 20 minutes (including the takeoff and the landing operations), in different operative and flight conditions. The monitored parameters were: accelerations, internal and external temperatures, ground and vertical speeds.

Actually, we performed six tests on sample tissues from gallbladder, thyroid, breast and cutaneous tissue. In five cases no degradation in the sample have been noted, both from a morphological and histoenzymatic points of view. On the other side, as illustrated in Figure 3 and Figure 4, in one case the sample undergone flight showed signs of degradation. In the figures, we report the results of different tests conducted on a sample of gallbladder, which has been divided in two samples, the first one analyzed immediately

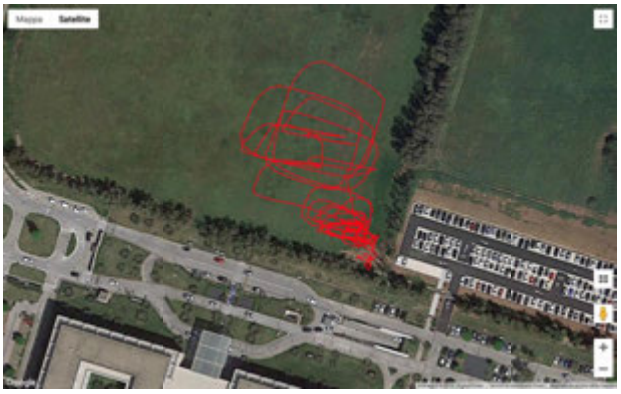
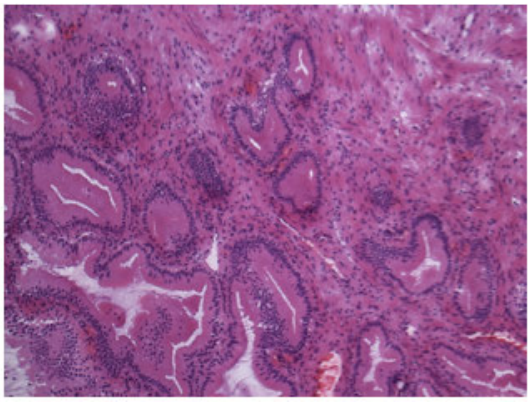
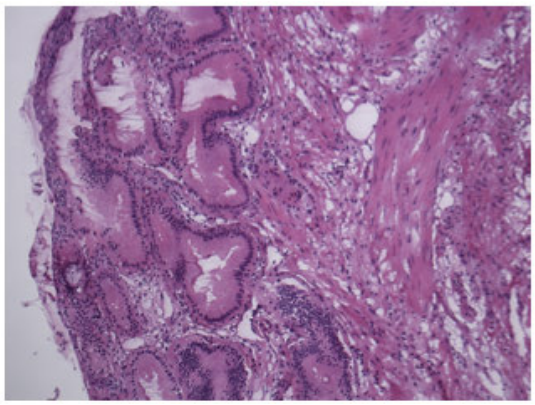


Fig. 2. Trajectory of a drone flight experiment carried out at a private area belonging to the University.

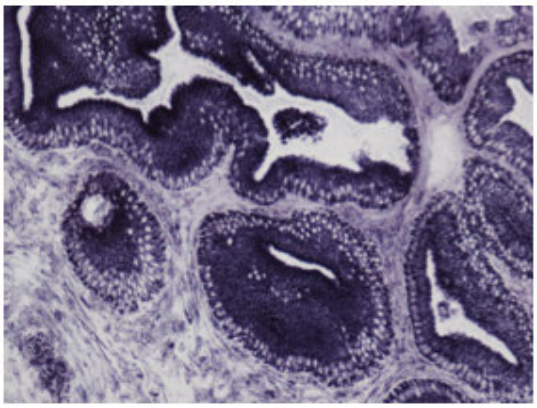


(a)

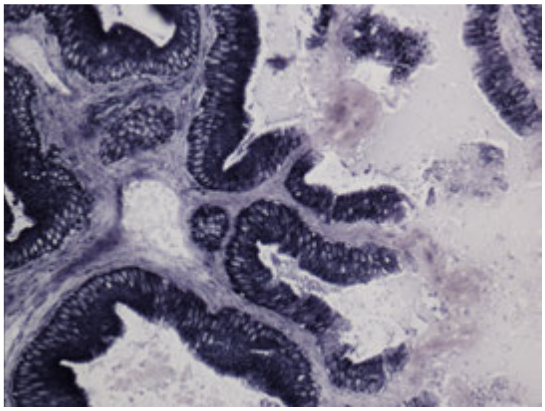


(b)

Fig. 3. Morphological analysis of a gallbladder sample that has undergone flight (upper panel) vs a sample from the same tissue that has been analyzed immediately after sampling (lower panel).



(a)



(b)

Fig. 4. Histoenzymatic analysis of a gallbladder sample that has undergone flight (lower panel) vs a sample from the same tissue that has been analyzed immediately after sampling (upper panel).

after being surgically removed (left panel in the figures) and the second one analyzed after flight. Specifically, the figures show on the left the sample analyzed immediately after being surgically removed, while on the right it is shown the sample after flight. According to the figures, the morphological properties of the samples are comparable, while there is a slight degradation in enzymatic activity in the sample undergone flight.

Overall, in spite of the cheap operational setting, the experimental results support the feasibility of the adoption of drones to transport non-fixed biological material between close hospitals (i.e., 20 minute flight) for intra-operative consultations.

A. Example of application

In this subsection we provide an illustrative example to show the potential of the proposed approach.

Traditional Timeline

Rome, October 12th, 2018, 10:00 AM. A public transportation strike is causing large turmoil. A car leaves a major hospital in the city center, bringing a pathologist. An hour from now a patient in a small-size clinic in the outskirts of Rome is about to undergo lung surgery, and the pathologist is required for intraoperative frozen section consultation, to

quickly assess pathology of a lung lesion as soon as it has been removed. The surgeon will decide, on the basis of this information, if the procedure must be completed with lobectomy and lymphadenectomy. If the lesion is a lung carcinoma, the lymphadenectomy is mandatory to give the best chances of cure to the patient. If the lesion is an amartoma, a benign lesion, the lymphadenectomy will increase the risk of the surgery without any benefit. Due to the strike, which results in traffic jams, the path (ĂŤjust 20 km) requires 73 minutes instead of the 20 minutes that would be necessary in normal conditions. As a result of the delay, the lung lobectomy, planned at 11:00 AM, is rescheduled at 12:30 AM. The pathologist arrives at 11:13 AM at the clinic and has to wait until 2:00 PM to be able to evaluate a frozen section of the lung. The intraoperative evaluation is difficult due to unusual features of the lesion and the pathologist is alone in taking the decision. The procedure is completed at 2:32 PM and the pathologist jumps in the car in order to reach a third hospital, where a surgery planned for 1:00 PM has been rescheduled at 4:00 PM. Due to the cascade of delays, two other planned surgeries in two different hospitals have been rescheduled for the subsequent week. The pathologist has spent 128 minutes in the car, 166 minutes idle and just 83 minutes actually performing his job. The overall costs/consequences are severe: each hospital must have and maintain a cryostat and a specialized technician able to use it, the pathologist is underused, potentially life-saving surgeries have to be delayed or re-scheduled. Moreover, carrying the pathologist contributes to increase the traffic jam and has an impact of about 20kg of CO_2 on the environment[2]. In order to transport the pathologist (e.g., 70 kg) a 1300 kg car is used, with an efficiency of about 0.05. Notice that such an issue would remain even if we adopted medical cars to deliver the biological samples (e.g., 0.1 kg) to the major hospital, resulting in an efficiency of 0.00007. However, the latter solution is not feasible in the presence of traffic jams, due to the strict time limit requirements of intraoperative consultation.

Timeline with the proposed system

Rome, October 12th, 2018, 10:00 AM. A public transportation strike is causing large turmoil. At 10:30 AM, a lung lobectomy starts in a small-size clinic in the outskirts of Rome. At 11:35 a drone departs from the small-size clinic bringing a non-fixed sample of the patientĂŹs lung to be evaluated by a pathologist in a major hospital in the city center, 20 km away from the clinic. The drone safely delivers the sample at 12:00 AM. At 12:22 the surgeon receives a call from the pathologist: everything is fine, the tumor is benignant. During the diagnostic procedure, the pathologist had the possibility to discuss with the colleagues some challenging features of the cases. The surgery ends successfully at 1:40 PM. At 1:40 PM another drone reaches the major hospital. At the end of the day, all the four planned surgical operations have been successfully carried out. In particular, the last one, seemingly less critical than the others, has instead a life-saving effect. While inspecting the patientĂŹs abdomen, the surgeon finds an internal bleeding and stops it, avoiding the patientĂŹs death. The cost of the STEADY system is extremely reduced: only the major hospital has to buy and maintain a cryostat and to employ a specialized technician. The pathologist is able to spend most of his work shift actually doing his job and can interact with his/her peers in order to handle complex cases. The car does not contribute to the traffic jam and to the air pollution, while the drone flights cause the emission[3], overall, of 0.46 kg of CO_2, i.e., a reduction of about 97.7% with respect to the scenario involving the car. Moreover, biological samples weighting less than 100 g are transported by means of a 5kg drone, resulting in an efficiency of at least 0.02, i.e., about 286 times more efficient than transporting the samples via car and is not dependent on traffic conditions.

But, most important, delays are avoided, and, at the end of the day, a life is saved.

IV. VPC AND SWOT ANALYSIS

The VPC was initially developed by Osterwalder et al. [9] as a part of the business model canvas. It was conceptualized as a framework to ensure that there is a fit between a product and customers. Afterward, the VPC has been adopted to highlight the beneficiaries (there are usually more than one beneficiary) of a new product/service as well as the gains and pains provided by the new product/service to each beneficiary, in general and respect to alternative products (if any). In light of these characteristics, we adopt this tool to assess the value of the proposed system to multiple beneficiaries, distinguishing between economic, environmental, and social aspects of the system and respective implications.

In detail, the VPC is formed around two building blocks (see Figure 5): (i) customer/beneficiary profile and (ii) product/service value proposition. The former refers to the objectives/requirements of the beneficiary, what it thinks may meet its requests, and potential drawbacks in addressing its objectives/requirements. Instead, the latter highlights the proposed product/service, how and to what extent it may meet the objectives/requirements of the beneficiary, and how and to what extent it may alleviate the beneficiaries' pains.

[2]The estimate is obtained by considering a medium-size car emitting 200 g/km and an overall travel of 100 km.

[3]This estimate represents a worst case and is obtained by considering 8 flights (4 round trip flights, one for each surgery) of a drone with a 150Wh battery that is completely discharged at each flight, and assuming that the CO2 emissions per kWh (i.e., the amount of CO2 that would be emitted at power generation facilities due to drone electricity demand) are 0.3773 kg, as estimated in [8].

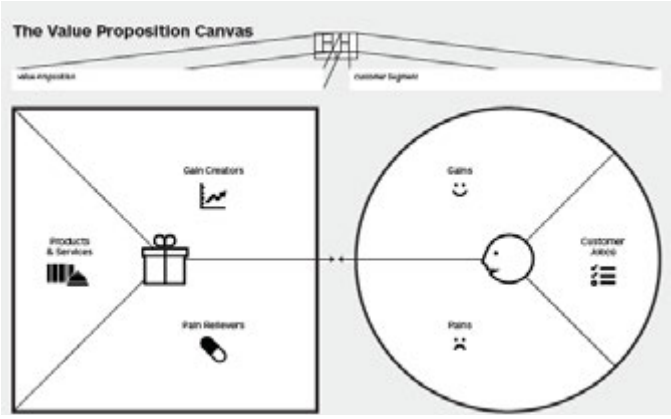


Fig. 5. The VPC. Source: Osterwalder et al. [11]

Following the principles of the VPC model and based on the what proposed in Section 2, firstly, we can identify three main beneficiaries of our drone-based system, namely the hospital with no department of pathological anatomy or effective equipment (so requiring intra-operative consultations), the department of pathological anatomy and pathologists providing the intra-operative consultations, and the society. Second, we reveal main advantages of the system.

The hospital requiring intra-operative consultations asks for a way to quickly let surgeons receive information on the tissue to analyze in order to decide how to proceed with the surgery. To do so, it may set an internal department of pathological anatomy and, hence, hire more staff and buy complex and costly equipment (e.g., cryostat), which represent new fixed costs. The envisaged drone-based system, instead, can fulfill the need of providing the hospitals intra-operative consultations in a timely manner upon request, making almost all costs for this service as variable instead of fixed.

The department of pathological anatomy of the hospital providing intra-operative consultations is often asked to send an internal pathologist to another hospital, temporarily diminishing the human resources available internally. This situation can be avoided through our system since the tissue directly comes to the department. In turn, it may be considered as an enabler of a new revenue model for the hospital. Regarding the pathologist who will actually provide the consultation service, he/she often needs to drive to another location in condition of time stress to accomplish its job. In this case, the drone-based-system saves pathologists time in that he/she does not need to move, with positive repercussion also to the safety of the pathologist who will not drive in condition of stress. Moreover, being in its department, should the pathologist require any type of support, he/she is allowed discussing with his/her colleagues.

Finally, the society, including whoever could be on the operating table "awaiting" for intra-operative consultations, will have a faster diagnosis and treatment. Furthermore, indirectly, the society will benefit from the drone-based system since its greenhouse-gas emissions are lower than the car solution.

Born as a framework to evaluate a company's competitive position, the SWOT analysis is now a more comprehensive foundational assessment model that helps a person or organization identify strengths (S), weaknesses (W), opportunities (O), and threats (T) related to defined product, service, project, etc [10]. Strengths and weaknesses refer to the intrinsic property of the product/solution (the drone-based system in our case), while opportunities and threats refer to how the external environment may influence the functioning of the proposed system.

By integrating application cases and the VPC, we developed a SWOT analysis of the proposed system. Figure 6 schematizes results.

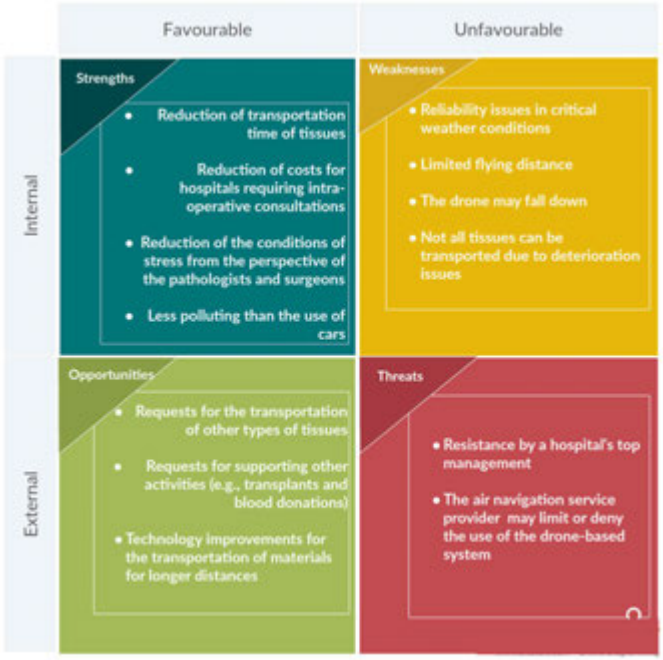


Fig. 6. SWOT analysis

V. Conclusions and Future Development

The proposed approach carries a seed of groundbreaking innovation, as it represents the first step in the implementation of a suite of innovative services with high added value, opening new markets and commercial opportunities. Implementing the proposed solution will also have consequences on the way society perceives drones, helping the diffusion of new, non-necessarily medical, services (e.g., package delivery). Moreover, the proposed solution may represent a collector for further innovation in the future. In particular, we expect that the following innovative aspects will naturally blend with it in the future:

- Advanced automatic coordination of conventional (e.g., medical cars) and non-conventional transportation means (e.g., drones), opening new scenarios for innovative management and routing policies (e.g., vehicle routing and optimization problems aiming at minimizing shipment time while saving resources).

- Management of the services as transitions in a blockchain, in order to guarantee non-repudiation and security of the collected data, accountability of the operators and redundancy.

REFERENCES

[1] https://hub.jhu.edu/2017/09/19/drone-transport-blood-samples-record/

[2] Amukele, Timothy K., et al. Drone Transport of Chemistry and Hematology Samples Over Long Distances. American journal of clinical pathology 148.5 (2017): 427-435;

[3] https://www.post.ch/en/about-us/company/media/press-releases/2017/further-drone-flights-in-lugano

[4] https://www.santannapisa.it/it/multimedia/sky-tg24-da-una-spinoff-del-santanna-il-primo-drone-al-mondo-grado-di-trasportare-il

[5] https://www.enac.gov.it/sicurezza-aerea/droni/progetti-di-ricerca-e-sviluppo-settore-sapr

[6] http://www.windtre.it/press-and-events/comunicati-stampa/sperimentazione_healthcare

[7] https://www.unicampus.it/news-blog/71191-la-salute-viaggia-con-i-droni

[8] Goodchild, Anne, and Jordan Toy. "Delivery by drone: An evaluation of unmanned aerial vehicle technology in reducing CO2 emissions in the delivery service industry." Transportation Research Part D: Transport and Environment 61 (2018): 58-67.

[9] A. Osterwalder, Y. Pigneur, G. Bernarda, A. Smith, and V. B. Sala, Value Proposition Design. Come creare prodotti e servizi che i clienti desiderano. Edizioni LSWR, 2015.

[10] N. Pahl and A. Richter, Swot Analysis - Idea, Methodology and a Practical Approach. GRIN Verlag, 2009.

A Privacy-Oriented Solution for the Improvement of Workers Safety

L. Faramondi**, P. Bragatto*, C. Fioravanti**, M. G. Gnoni***, S. Guarino** and R. Setola**
* Dept of Technological Innovations, INAIL Italian Workers Compensation Authority, Monteporzio Catone, Italy
** Unit of Automatic Control, University Campus Bio-Medico di Roma, Via Alvaro del Portillo 21, 00128, Rome, Italy
*** Department of Innovation Engineering, University of Salento, Via Monterondo, 73100, Lecce, Italy
{l.faramondi, r.setola}@unicampus.it, p.bragatto@inail.it, {camilla.fioravanti, simone.guarino}@alcampus.it, mariagrazia.gnoni@unisalento.it

***Abstract* - The improvement of workers' safety, especially in complex and potentially dangerous scenario is a mandatory task. To this end, in recent years, there has been a large interest to complement actual passive solutions with IoT based system in order to provide a context related awareness information. Indeed, well informed workers can better manage anomalous situations avoiding dangerous situations. Unfortunately, the large part of the solutions proposed in the literature have difficulties to be accepted because they have implicitly mechanism that can, even if only in potential, be used to monitoring the workers' activities. With the aim to overcome such drawback, in the context of Smart-Bench project we propose a fully distributed solution where all the detailed information is managed in order to avoid any possible incorrect abuse of the data. In more detail, using an architecture based on deployed and wearable sensors, the system is able to provide information about the environmental status directly and exclusively to the worker. Only in case of anomalous situation the information is shared in order to better manage the crisis. With the aim to design a solution easy to be implemented with limited effort and fully integrated with the IoT standards, we propose a distributed infrastructure based on the BLE standard.**

Keywords - IoT; Situation Awareness; Safety; Sensors

I. INTRODUCTION

The wide distribution of innovative data transmission technologies has allowed the integration of new devices and items in the cloud of companies and systems. Nowadays, the affirmation of Internet of Things (IoT) concept in industries leads to an improvement of services efficiency, a smart monitoring of the production, and an accurate tracking of the products from the initial phases to the delivery [1]. In this context, the development of the Industry 4.0, which integrates new production technologies to improve working conditions, offers to academic and industrial research communities the opportunity to use these novel items for enhancing the security of human-related environments. The issue concerning the safety of companies appears to be particularly relevant since, in many cases, profitability increases by improving availability and accident prevention, and so there is theoretically no conflict between safety and profitability.

The "Seveso II" Directive (96/82/EC) aimed at preventing and controlling the risk of major-accidents, established that operators of certain establishments holding substantial quantities of hazardous substances are required to put into effect a Safety Management System. Methodical experience in drawing up emergency plans led to a procedure divided into three main steps: analysis of information on hazardous substances, hazard analysis including all plants, more specific hazard analysis for representative plants [2]. In particular, the data on temperature, pressure, concentration and state of aggregation of the dangerous substances provide essential elements for the classification of plants into groups with different hazard potentials. Furthermore, as stated in [3], HSE management would benefit greatly from the improvement of environmental performance indicators, that identify the conditions prior to accidents through the measurement of physical parameters related to relevant conditions for occupational health.

The possibility of ensuring necessary interventions which are needed if wireless sensor data goes across pre-defined thresholds, is allowed by integrated solutions based on Building Information Modeling (BIM) and wireless sensor technology, which eliminate manually monitoring [4]. The authors focused on monitoring oxygen deficiency in confined spaces, explaining how it is possible extracting the location of each sensor embedded in the building.

Taking into account the IoT technologies and products applied in the context of industries, in [5] is presented an overview about the most applied communication technologies in Industry 4.0, underlining the efficiency of radio-frequency identification (RFID) tags, Bluetooth standard, and near-field communication (NFC).

In the last decade the development of all these devices led to the implementation of the concept of Smart Factory, enabled by means of different key technologies [6] like WSN, which mainly use interconnected intelligent sensors to perform real time environmental monitoring, industrial asset tracking, and localization of both people and equipment.

The widespread diffusion of this kind of technologies in the industrial context is unrestrainable: as reported in [7], in recent years the contribution of IoT is increasing especially in the field of Industry 4.0, despite the

presence of issues related to cyber threats, safety, and security aspects, nowadays considered a hot topic. In [8] the authors face the problem of cyber threats, data integrity, and intrusion detection from an IoT point of view.
Recent research developments, as described in [9], extend the concept of smart object to the workers by including them in a large environmental network composed by cyber, physical, and human elements; this new kind of system is identified by the authors as Cyber-Physical Human System (CPHS).

Starting from this kind of system, in this paper we propose a framework based on classical IoT technologies with the aim of proposing a CPHS able to assist workers and inspectors in industries or hazardous environments by analyzing and merging heterogeneous information.

A. *Paper Outline*

The outline of the paper is as follows: In Section II we introduce the aim of the Smart-Bench project. In Section III we illustrate the main aspects of the project with a focus on the proposed infrastructure and the main services. Some relevant details about the hardware used in the project and the privacy aspects are resumed in Section IV with the aim to highlight the privacy-oriented solution. Finally. some conclusive discussions about possible future works are reported in Section V.

II. PROJECT OVERVIEW

The Smart-Bench project starts in 2017, it is an Italian research project co-funded by INAIL, the Italian Workers Compensation Authority. It involves five engineering departments from five Italian universities with the aim to realize a set of services able to improve the safety of the workers in process industries. As introduced in Section I, the safety of workers in this kind of industries is a hot topic due to large number of activities characterized by an high risk level.

The aim of the project is the realization of a risk assessment system in order to prevent accidents. The key idea is the realization of a distributed infrastructure on the field where, machines, environment, and workers are connected each other in order to retrieve relevant data from the field and provide an early warning notification to the workers or remote control rooms in case of emergency.

The project consists in three main services: a *localization service* necessary to estimate the position of the worker in case of emergency, an *indirect shared warning system* devoted to the exchange of relevant notions about human and environmental conditions, and a *state estimation tool* responsible for the estimation of the anomalous situations. Notice that, in line with the regulations about the privacy of the workers, each service has been implemented with the aim to respect the privacy constraints by ensuring the access to the information only in case of emergency and by adopting secure encryption algorithms for the data storage.

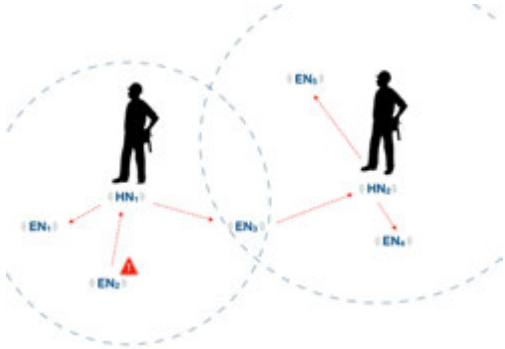


Figure 1. Smart-Bench infrastructure and data sharing strategy

In more details the framework proposed in this project is specially designed with the aim to provide to the workers a real-time context awareness data about the environment. The proposed project merges the innovations which characterize the Industry 4.0 applications such as communication protocols and technology. One of the main guide lines of the proposed project is related to its modularity and low energy requirements. Indeed, the proposed solution is easy to install without require the presence of a pre-existent wired network on the field, it is based on the development of a Wireless Sensor Network (WSN) according to the IoT paradigm using Bluetooth Low Energy (BLE) protocol. A few preliminary results about the application of the Smart-Bench framework in the context of chemical and petrol industries have been recently presented in [10].

At the same time, concerning the aspect about the integration of the framework and the worker activities, the solution is easily integrated with the worker equipment, and, at the same time, without interfering with the daily activities of the workers.

III. SMART-BENCH INFRASTRUCTURE

As explained in the previous section, the proposed infrastructure is compliant to the most common IoT standards in order to ensure the largest degree of interoperability with the pre-deployed infrastructure, the industrial machines, and the workers equipment. To this end, the solution provides for a mobile application, which runs on a smartphone/tablet, able to exchange data with the environment and remote control room. In more details, the mobile application is implemented as a background service therefore it is able to communicate with the devices deployed in the environment without interfering with the worker activities and the other mobile applications (e.g. for inventory, report, or inspection activities). As depicted in Figure 1, the infrastructure is composed by a set of environmental nodes (EN), each of them consists of a smart device based on a microcontroller board equipped with a communication module and a set of environmental sensors (e.g. CO, CO2, humidity, temperature, pressure, etc). The ENs, deployed in the environment in strategic positions, are able to send data via BLE protocol when a worker detects their presence. Concerning the user equipment, in addition to the tablet, the proposed solution provides also for the presence of a Human Node (HN). Each worker is

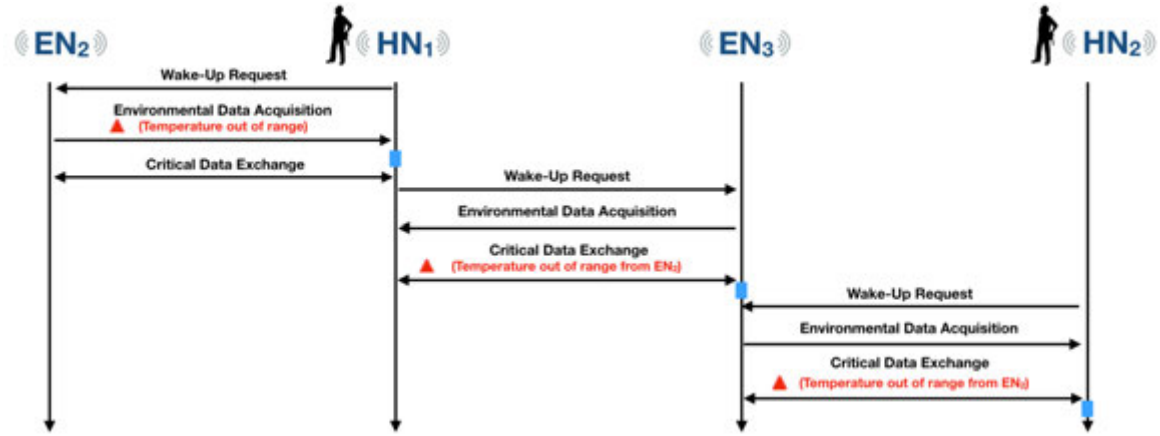


Figure 2. Indirect information sharing strategy. (Blue square represents the update of the internal report of the devices)

equipped with a smart device with the same characteristics of the EN. Notice that HN further collects data about the environment is able to monitoring the state of health of the worker.

The basic idea is that the worker, during his/her daily activities, is able to automatically retrieves data about the environment when he/she is close enough to an EN (see Figure 2). At the same time, information provided by EN are used to estimate the position of the worker. All these data, merged with data retrieved from the HN, form the basis for the analysis of correlation among the health of the worker and the presence of abnormal environmental parameters. With the aim to preserve the privacy of the worker, both the information retrieved from HN and ENs are exclusively locally managed and processed by the worker's tablet. The proposed framework provides a system of data and alarm sharing only in case of emergency. To this end, the system adopts several communication channels such as Wi-Fi and SMS, however, because of possible lack of connectivity there is no guarantee to have adequate coverage. With the aim to overcome this issue, the system uses also the ENs in order to share the emergency alarm or critical data.

More precisely, ENs devices can acts also as a distributed shared memory where the system stores information about anomalous conditions retrieved by the workers in the neighborhood area.

A. State Estimation Service

The state estimation service consists of a real-time algorithm implemented as an Android background service. Starting from the data retrieved from the inertial sensors placed in the HN, the algorithm performs a preliminary filtering operation on the signals and compares the acquired measures with fixed thresholds in order to recognize potentially dangerous activities of the worker. The algorithm is able to detect if the user is running, is walking, is down on the floor, or in a stay in a stop state or he fell down.

In more details, the discrimination of the actions performed by the workers are based on the analysis of accelerometer and gyroscope measures. Concerning the identification of the patterns related to the walking action, it is based on the study of the Euclidean norm of the accelerometer signal considering all the three components. In a fixed time window, the algorithm compares the values of the peaks in the signal with a fixed threshold in order to estimate the walking activity. A similar approach has been adopted in order to identify the running activity. In this case, addition to the analysis of the Euclidean norm of the accelerometer signal, also the gyroscope measures are compared with thresholds. A more accurate analysis has been performed in order to identify the pattern that describe the fall of the worker. In this recognition phase, the algorithm analyzes also the first derivative of the signal which describe the orientation of the worker in addition to the accelerometer and gyroscope signals. Finally, concerning the estimation of lying position, the identification is based on the observation of the gravity acceleration and its projection on the three axes.

With the aim to perform such estimation, one of the eight available BLE connections of the tablet is permanently devoted to the exchange of data with the HN.

B. Localization Service

The indoor localization system is the second relevant feature of the proposed project. It is essential in order to estimate the positions of the workers in case of emergency.

The position is calculated on the basis of the RSSI (Received Signal Strength Indicator) of the connection among the worker's tablet and the ENs deployed in the field in known positions.

The coordinates of the worker are calculated using weighted centroid localization as it follows:

$$X_0 = \sum_{i=1}^{N} w_i X_i$$

$$Y_0 = \sum_{i=1}^{N} w_i Y_i$$

where X_0 and Y_0 are Cartesian coordinates of the worker, N is the number of ENs deployed in fixed known positions (X_i, Y_i) on the field, and w_i is the RSSI weight assigned to the particular EN.

The estimation of the worker position can be evaluated by implementing a Fuzzy Logic ([11]) based on the fuzzification of RSSI values and by applying triangular membership functions, as represented in Figure 3. Here we consider four different signal levels: strong, intermediate, weak, and unusable.
The implemented fuzzy inference engine is the Mamdani min-max ([12]) method based on the following rules:

1. **If** RSSI is STRONG **then** w_i is HIGH
2. **If** RSSI is INTERMEDIATE **then** w_i is MEDIUM
3. **If** RSSI is WEAK **then** w_i is WEAK
4. **IF** RSSI is UNUSABLE **then** w_i is SMALL

Once obtained the fuzzy representation of the weight for a given RSSI value, it is necessary to convert the result to a crisp value.
In more details, in this case our approach is based on the "center of gravity" method.

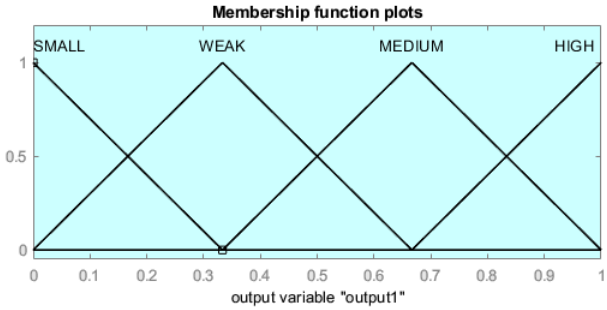


Figure 4. Triangular membership functions of weight

So, for example, the corresponding weight for an RSSI value equals to -50 is 0.56, as reported in Figure 5.

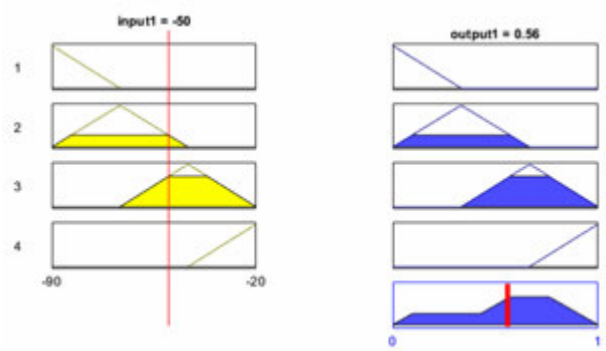


Figure 5. Evaluation of weight for an RSSI equals to -50

The fuzzy interpretation of the RSSI Measure is useful in order to better manage the intrinsic uncertainty of this kind of measure. Once the crisp value is obtained, the localization service considers the possible presence of multiple ENs and applies a triangulation pocess with the aim to compute a precise estimation of the position.

C. *Indirect Shared Warning System*

The last main service of the framework is the indirect shared warning system. Also this feature is implemented as an Android background service to disseminate relevant data about workers and environmental anomalous conditions. Notice that, as introduced in Section II, the exchange of information between the tablet and the ENs is based on BLE connections and inherits the limits of this technology, in fact, as defined in the BLE protocol, the tablet is able to manage at the same time a maximum of eight connections. With the aim to overcome this limit, the indirect sharing warning systems implements a sort of Round Robin technique in order to manage connections with more than eight ENs simultaneously. When the tablet perceives the presence of more than eight ENs, it initializes seven connections (one connection is always active and dedicated to the exchange of data between tablet and personal HN) with seven random ENs, it retrieves the data from the first seven ENs, closes the connection to the first connected devices and establish the connections to the rest of disconnected devices in order to complete the data collection from all the discovered ENs.
Due to the lack of a priori knowledge about the placement of the ENs in the field, and with the aim to preserve the energy required by the devices to exchange information, the proposed solution does not provide for a direct communication among ENs or tablets. The allowed connections involve always one, or more, ENs and a tablet, while connections between two ENs or two tablets are not allowed. In this way, the worker, moving around the ENs deployment area, operates as a gateway for the exchange of information by disseminating data retrieved from different sources. Note that each tablet elaborates a smart report which collects the activities of the user, as a result of the State Estimation Service and the most relevant data about gas presence, temperature, and humidity retrieved from the ENs in the field. At the same time, also the ENs compiles a report by keeping track of the connections to the tablet. It consists in a summary of the most relevant measures collected by the workers during their activities. In this way, each device acts as a shared memory where each worker is able to store relevant data about his/her health conditions, the detection of environmental values out of a safe range, and

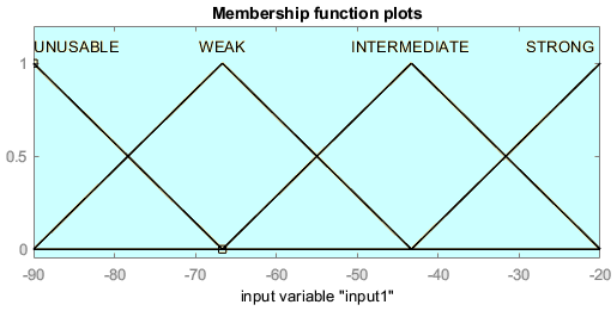


Figure 3. Triangular membership functions of RSSI.

the list of the visited areas of the environment.

As depicted in Figure 2, the exchange of information between tablet and EN consists of three phases:

TABLE I: HN AND EN HARDWARE DETAILS

Included in	*Sensor*	*Measure Range*
HN	3D Accelerometer	±2/±24/±28 g
HN	3D Gyroscope	±2245/±2500/±22000 dps
HN	Magnetometer	±24/ ±28/ ±212/ ±216 gauss
HN/EN	Barometer	260 - 1260 hPa
HN/EN	Temperature	- 40 – 120 °C
HN/EN	Relative Humidity	0 – 100 %
HN/EN	Carbon Monoxide	0 – 5000 ppm

1. Connection Establishment: usually, due to power saving reasons, the EN is in sleep mode. When the tablet perceives the presence of the node, it sends a wake-up request to the node and establish a new connection.
2. Environmental Data Acquisition: Once a connection has been established between tablet and EN, the node increases the sampling frequency of the environmental measures and send the real time data to the tablet for a fixed period.
3. Critical Data Exchange: During this phase the two devices share relevant information collected in their internal reports in order to disseminate relevant data among the ENs without a direct connection.

For the sake of clarity, please consider Figure 1 and Figure 2. As depicted in the first figure, the first worker detects the environmental node EN_2, notifies his/her presence to the node and acquires the environmental data. In this case, as detailed in Figure 2, the temperature measured by EN_2 is over the safe threshold, due to this reason the first worker updates automatically the field about the highest measured temperature into his/her report and, when it perceives the presence of EN_3, during the phase of the critical data exchange, the environmental node EN_3 is informed about the excessive temperature measured from EN_2.

As represented in Figure 2, thanks to the indirect data sharing process, also the second worker is informed about the high temperature measured by EN_2, even if he/she has never directly communicated with EN_2 or with first worker.

IV. IMPLEMENTATION DETAILS

A. Hardware Details

The proposed solution has been implemented according to the guidelines described in Section II and Section III. In more details, concerning the hardware aspects, each EN and HN has been implemented on a Nucleo STM32 microcontroller board, equipped with an ARM 32 bit Cortex M4 CPU, the BLE expansion board, the inertial platform, and the electrochemical toxic gas sensor expansion board with CO sensor, in addition, the HN involves also the wearing module for sweat monitoring.



Figure 6. Overall view of the hardware devices

The details about the hardware are resumed in Table I. At the current state, the proposed solution provides for 10 EN and 2 IN. A prototypal GUI has been deployed in order to debug the implemented services. The overall view of the framework is shown in Figure 5.

B. Software and Privacy Aspects

Concerning to the implementation details about the Android services, the user is able to setup the frequency with which the system produces a report. Usually the report consists in a Json file which contains the mean and the highest/lowest values about environmental measures (temperature, gas presence, etc), the physical state of the worker, computed by the state estimation service, the historian about the connected ENs, and the position estimation based on the algorithm described in Section III.B.

The management of privacy and data is the crucial point of this project. To guarantee an adequate level of privacy we propose an approach based on the following principles.

The Json files produced during the activities of the worker which contain the report of the user activities are stored only in the tablet local memory and finally deleted if the activities are completed without any accidents. Only in case of anomalous conditions, the report is shared with the control room as soon as a data connection is available

Note that, when a Json file is generated, it is encrypted using a private random key which is saved as an internal variable of the service instance. In this way, the control room must require the key in order to obtain the access to the report contents. When the control room require the key, a notification message is shown on the screen and the worker can allow to send the key. An automatic send is performed in case of particular health conditions (e.g. no movements detected, worker lying on the floor for too long, etc).

Another fundamental principle is about the establishment of BLE connections among ENs, HNs, and tablets. In our approach the communication about relevant data stored in the tablets is always managed as an indirect communication: the tablets use the ENs deployed on the field as shared memories where they are able to store data in anonymous way, without exchange data directly among them. In this way the traffic analysis of the incoming/outcoming BLE connections does not allow to identify the source of the acquired data.

V. Conclusion and Future Works

The approach presented in this work makes it possible to merge heterogeneous data among workers, environmental smart sensors, and industrial machines with the aim to provide a useful tool able to provide a risk assessment analysis and an early warning notification to remote control rooms. Moreover, one of the advantages of this work is the complete absence of dependencies by power or communication networks. Finally, the indirect exchange of data between workers ensures the respect of the privacy policy. Future works will be devoted to the integration of data retrieved from the industrial machines and devices in order to improve the effectiveness of the early warning system.

Acknowledgment

The work was funded by INAIL - Grant PAR/BRIC2016/ID15 "Smart-Bench Project".

References

[1] Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. *Computer networks*, *54*(15), 2787-2805.

[2] Kaiser, W., & Schindler, M. (1999). Precautions against industrial accidents: experience in applying the Seveso II Directive in central and eastern European countries. Journal of hazardous materials, 65(1-2), 59-75.

[3] Duijm, N. J., Fiévez, C., Gerbec, M., Hauptmanns, U., & Konstandinidou, M. (2008). Management of health, safety and environment in process industry. Safety Science, 46(6), 908-920.

[4] Riaz, Z., Arslan, M., Kiani, A. K., & Azhar, S. (2014). CoSMoS: A BIM and wireless sensor based integrated solution for worker safety in confined spaces. *Automation in Construction*, *45*, 96-106.

[5] Da Xu, L., He, W., & Li, S. (2014). Internet of things in industries: A survey. IEEE Transactions on industrial informatics, 10(4), 2233-2243.

[6] Salvatore, C., Bocchino, S., Petracca, M., Pelliccia, R., Ghibaudi, M., & Pagano, P. (2012, September). WSN and RFID integrated solution for advanced safety systems in industrial plants. In *SoftCOM 2012, 20th International Conference on Software, Telecommunications and Computer Networks* (pp. 1-5). IEEE.

[7] Conway, J. (2016). The Industrial Internet of Things: an evolution to a smart manufacturing enterprise. Schneider Electric.

[8] Alqassem, I., & Svetinovic, D. (2014, December). A taxonomy of security and privacy requirements for the Internet of Things (IoT). In Industrial Engineering and Engineering Management (IEEM), 2014 IEEE International Conference on (pp. 1244-1248). IEEE.

[9] Scheuermann, C., Verclas, S., & Bruegge, B. (2015, August). Agile factory-an example of an industry 4.0 manufacturing process. In Cyber-Physical Systems, Networks, and Applications (CPSNA), 2015 IEEE 3rd International Conference on (pp. 43-47). IEEE.

[10] Bragatto, P., Faramondi, L., Failla, F., Gnoni, M.G. (2018) Potential and limits of IOT for hazardous job in process industries Chemical Engineering Transactions, 67, pp. 865-870.

[11] Jang, J. S. R., Sun, C. T., & Mizutani, E. (1997). Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [Book Review]. *IEEE Transactions on automatic control*, *42*(10), 1482-1484.

[12] Cordón, O. (2011). A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems. *International Journal of Approximate Reasoning*, *52*(6), 894-913.

LoRa-SDN: Providing Wireless IoT Edge Network Functions via SDN

F. Holik*, U. Roedig** and N. Race***
* University of Pardubice, Faculty of Electrical Engineering and Informatics, Pardubice, Czech Republic
** University College Cork, School of Computer Science and Information Technology, Cork, Ireland
*** Lancaster University, School of Computing and Communications, Lancaster, United Kingdom
filip.holik@upce.cz

***Abstract* - Large-scale Internet of Things (IoT) deployments such as smart cities and smart grids are becoming a reality. In these topologies, extensive numbers of wireless devices transmit data to gateways that forward the collected data to back-end systems over a fixed network infrastructure – the core network. It is expected that in the near future the core network will utilize software-defined networking (SDN), as has already happened in data centres and networks of service providers. This enables simplified deployment of network functions and dynamic reactions to observed network conditions. This paper explores how SDN mechanisms can be applied beyond the traditional core network to include wireless IoT edge networks as well. The most popular IoT technology – Long Range (LoRa) – was selected as the main use case technology. The paper describes the LoRa integration with SDN and proposes the LoRa-SDN integration architecture.**

Keywords - LoRa; SDN; IoT; wireless edge networks

I. INTRODUCTION

IoT networks are now deployed on a large-scale to support applications such as smart cities, healthcare, smart mobility or the grid. In these deployments a very large number of wireless IoT devices are used, feeding data via gateways into a fixed network backbone infrastructure. In some cases, data also flows, from the backbone to IoT devices to trigger actions.

Currently, IoT edge networks are managed independently from the core network. Functionality of the IoT edge networks is managed via a dedicated set of tools. Furthermore, IoT edge networks are in many cases run by different stakeholders than the core network (a different division in an organization or even a different provider). Finally, in large-scale IoT deployments, multiple types of IoT networks are present and a heterogeneous infrastructure has to be considered. Hence, management decisions regarding operations of the wireless edge networks are made independently from management decisions of the core network. A holistic network view is missing and operational decisions affecting the entire network are generally implemented manually. However, it is desirable to have a holistic network view and the ability to adjust network functions across the entire network, including core network and IoT edge networks.

It is likely that in the near future the core network will be implemented using SDN technology. This enables simplified deployment of network functions and dynamic reaction to observed network conditions. We believe it is sensible to extend SDN capability to IoT edge networks to provide a standardized way of network integration. Many network functions of an IoT edge network can be managed and implemented via existing SDN infrastructures. Rather than building and using a dedicated infrastructure to implement and manage IoT network functions, it is possible to use available SDN infrastructure for this purpose. As a result, a tighter integration of the network infrastructure is achieved which enables us to make better decisions regarding network resource usage. Additionally, such integrated infrastructure allows us to implement configurations across the entire network faster and in an automated fashion.

In this paper we explore how SDN mechanisms can be applied beyond the core network and how they can be applied to wireless IoT edge networks. Specifically, we use LoRa as a popular example of IoT edge network technology and show, how LoRa gateway selection can be mapped onto an SDN infrastructure, instead of leaving gateway selection to the LoRa infrastructure. This approach is especially relevant for the integration of LoRa edge networks into a complex smart city network.

The related work is analysed, and benefits of the LoRa-SDN integration are described in the next section. Section 4 presents basic LoRa and SDN concept details. The integration and architecture are described in section 5.

II. RELATED WORK

To the best of our knowledge, there is currently no work focusing SDN integration with LoRa technology. The existing work in the area focus only on the general IoT technologies. This covers SDN architecture for flexible and efficient IoT management as proposed in [1], Wireless Sensor Networks [2], and generally described review of SDN opportunities and challenges in IoT [3].

On the other hand, LoRa itself is experiencing a rapid increase in interest and many new papers have been published recently. Most of the research is focused on objects localization/tracking [4-5] and benchmarking – performance [6-8], reliability [9], coverage [10-11], interferences [12], signal propagation [13-14] power consumption [15], and scalability [16].

There are also approaches to improve or extend the LoRa functionality. Haghighi et al. [17] proposed an algorithm for optimizing transmission time, receiver sensitivity, packet loss and energy consumption. A replacement of the LoRaWAN MAC protocol by traditional IPv6 network stack was implemented in [18]. Use of this solution, amongst its other benefits, would simplify the SDN integration into the LoRa network by providing standardized OpenFlow match fields.

While there is no exhaustive work on a unified integration system, it is becoming clear that in the near future, such a system will be needed for a unified integration of heterogeneous IoT transmission technologies. This will be more important than ever in the upcoming 5G systems [19] given SDN is expected to be the key enabling technology.

III. Benefits of SDN Integration with General IoT Edge Networks

There are multiple benefits of integrating an SDN with general IoT edge networks. At this point of the discussion we do not assume a particular IoT technology. Recent Low-Power Wide-Area Network (LPWAN) technologies such as LoRaWAN [20], Sigfox [21], RPMA [22] and Weightless [23] might be used. They assume single-hop communication from wireless nodes to gateways. However, other technologies used for IoT edge networks such as Zigbee [24], Z-wave [25], or WirelessHART [26] which use multi-hop forwarding in the wireless network may be used as well. SDN integration concepts and benefits are applicable to IoT edge networks independent of the used technology.

Using an SDN infrastructure to integrate IoT edge networks has the following benefits:

- Reusability: an existing SDN infrastructure provides many elements that would have to be otherwise duplicated for each IoT network (in the case that the functions would be individually implemented for each IoT network).
- Abstraction: the implementation of network functions can be shared among different IoT edge network types. Thus, the SDN infrastructure provides a suitable abstraction layer for IoT network functions.
- Resource usage: tight integration of IoT edge networks and the core network enables network-wide decisions on resource usage. IoT behaviour can be dynamically adjusted to core network conditions.
- Management: when implementing IoT edge network functions and core network functions on the same platform, it is possible to share management principles and tools (for example, QoS setting). Network operation is therefore simplified, leading to cost savings.
- New opportunities: integration of IoT edge network and core network provides new opportunities in network functions spanning these networks.

A. SDN Integration with Smart City Networks

The mentioned benefits of integration of SDN into IoT networks are especially relevant for smart city networks. A developed architecture of a smart city network can be re-used in similar cities. Smart city sensors use heterogeneous communication protocols, but SDN can provide unified abstraction. Resource usage can be lowered as explained in section 5. SDN with centralized monitoring can automatically allow a more efficient management. SDN can also virtually separate different network traffic flows and therefore enable implementation of innovative features. This can result in more dynamic evolving of these networks.

IV. Technological Concepts

A. Long Range

LoRa is a proprietary spread spectrum modulation technique by Semtech [27]. It is a derivative of Chirp Spread Spectrum (CSS). The LoRa physical layer may be used with any Medium Access Control (MAC) layer; however, Long Range Wide Area Network (LoRaWAN) is the currently proposed MAC which operates a network in a simple star topology.

B. LoRaWAN

As LoRa is capable of transmitting over very long distances, it was decided that LoRaWAN only needs to support a star topology. Nodes transmit directly to a gateway which is powered and connected to a backbone infrastructure. Gateways are powerful devices with powerful radios capable to receive and decode multiple concurrent transmissions (up to 10). Three classes of node devices are defined:

- Class A: the node transmits to the gateway when needed. After transmission the node opens a receive window to obtain queued messages from the gateway.
- Class B with scheduled receive slots: the node behaves like a Class A node with additional receive windows at scheduled times. Gateway beacons are used for time synchronization of end-devices.
- Class C with maximal receive slots: nodes are continuously listening which makes them unsuitable for battery powered operations.

C. LoRa Physical Layer

LoRa is a physical layer specification based on CSS with integrated Forward Error Correction (FEC). Transmissions use a wide band to counter interference and to handle frequency offsets due to low cost crystals. A LoRa receiver can decode transmissions 19.5 dB below the noise floor. Thus, very long communication distances can be bridged. LoRa key properties are: long range, high robustness, multipath resistance, Doppler resistance, and low power. LoRa operates in the lower ISM bands (EU: 868 MHz and 433 MHz, USA: 915 MHz and 433 MHz). A LoRa radio has four configuration parameters: carrier frequency, spreading factor, bandwidth and coding rate.

The selection of these parameters determines energy consumption, transmission range and resilience to noise.

D. SDN Concepts

SDN is a concept of separating control and forwarding planes while achieving logical centralization of network control. This control is implemented on a specific device called an SDN controller. The controller instructs forwarding devices via a so-called southbound interface – the most used protocol on this layer is OpenFlow. The controller can also provide a so-called northbound interface for communication with more advanced applications. SDN is a well described topic and more detailed information can be found in [3].

E. OpenFlow

OpenFlow [28] is the first and mostly used southbound interface protocol, allowing control over forwarding layer. This is achieved by manipulation with a data structure called *flow table*, located in an SDN-enabled forwarding device. The flow table contains forwarding rules pushed from the controller. These rules have form of a *match-action* and also contain counters, priority, and timeouts. An action is executed when a datagram matches the rule.

Depending on the OpenFlow version, several match fields are required to be supported, while the support of most of them is only optional. Required fields in the most widespread version of OpenFlow - 1.3 are: incoming port, source/destination MAC address, ethernet frame type, IP protocol, IPv4/IPv6 source/destination address, and TCP/UDP source/destination port number.

The same OpenFlow version requires two instructions: write-action and go to table (for additional processing in a different flow table). The mostly used optional instruction is apply action, which allows forwarding of a packet (to a specific port, controller, or a special data structure – queue or group) and dropping it.

OpenFlow counters are maintained for every flow table, entry, port, and other specific data structures (queue, group, etc.). Depending on the entity, they gather statistics like number of flows, packets, bytes transferred, errors, etc. These statistics can be dynamically requested by the controller and used by an SDN application.

Priority is then used for selection of the most preferred rule, if multiple rules match the datagram. It can be defined in the range of 0 – 65 535. Zero is the lowest priority (typically used for the default rule) and 65 535 is the maximum priority.

The last OpenFlow rule fields are two timeouts defined in seconds: idle and hard. Idle-timeout specifies time, after which the rule is deleted if it is not used. Hard-timeout, on the other hand, delete the rule regardless if it is used or not. Timeouts can be set to infinity (0).

V. LoRa-SDN Integration

Typical LoRaWAN networks consist of nodes, gateways, network server and application server. A typical communication in the network has therefore the following three parts:

- Nodes – gateways: the communication is realized over low frequency radio and uses LoRa physical layer.
- Gateways – network server: the communication is realized over standardized IPv4 / IPv6 in a LAN / MAN environment.
- Network server – application server: the communication is realized over standardized IPv4 / IPv6 in a WAN environment.

Gateways act as bridges between different network types. As such, they are responsible for message encapsulation and de-encapsulation. Typically, after a message from a node is received, the content is processed into the JSON format and put into payload of a newly created IPv4/IPv6 packet. A metadata generated by the gateway like RSSI and channel bandwidth can be also added into the payload in the same JSON format.

SDN based on the OpenFlow protocol can be deployed in any network using traditional TCP/IP stack. As shown in Fig. 1, this covers core network (SDN forwarding devices) as well as LoRa gateways. There are however several points which have to be taken into the account when integrating SDN into a LoRa network:

- OpenFlow support: networking devices have to support the particular version of OpenFlow protocol. This might not be issues on typical core network devices, but might be problematic on experimental gateway devices. On the other hand, these devices often run Linux OS and a software switch supporting OpenFlow (Open vSwitch, OF13SoftSwitch, or Indigo Switch) can be installed on them.
- Modification of LoRaWAN gateway application: OpenFlow protocol can match only standardized header fields of the TCP/IP stack. For this reason, selected data from proprietary LoRa messages must be extracted and inserted into traditional TCP/IP headers during the message encapsulation. Example of an important field is sensor ID.
- Modification of gateway message processing: a typical gateway will encapsulate any received message from a node into an IP packet and send it to an IP network. If an OpenFlow software switch is implemented on the gateway, the encapsulated message must firstly be sent via this switch. Depending on the SDN logic, the message can then be discarded, modified, or sent to the IP network.

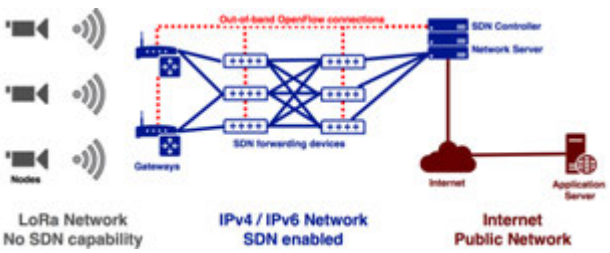


Figure 1. LoRa-SDN deployment scope

- OpenFlow matching process: OpenFlow can match standard and experimenter fields. In the **standard** field, the LoRaWAN gateway application has to firstly encapsulate the received LoRa message into an IP packet, or TCP/UDP segment. Any data from the LoRaWAN message, which will be used by SDN has to be putted into header fields supported by OpenFlow. In a theory, almost any field can be used – either in network layer header (IPv4/IPv6) or in transport layer header (TCP/UDP), but this would be considered a field hijacking and could not be performed outside the gateway boundary. The only exception is the *IPv6 Flow Label*, which is currently not used in most of the networks. This field has length of 20 bits, which is sufficient space only for a limited amount of information. The second, more flexible option is the **experimenter** field, which is an optional feature introduced in OpenFlow 1.2. It provides the flexibility of defining an arbitrary field to be matched. On the other hand, to support such a field within the whole network infrastructure, it must be supported on the controller (that requires a relatively simple software modification) and on all forwarding devices (a significantly more complicated modification, especially if kernel-based switches are used).
- SDN controller features: location of the controller is from the network perspective critical. In a LoRa-SDN integration, the most sensible location is on the same physical server as the LoRa's network server. Most of the applications will require tight cooperation of these two systems, so the performance benefits of their close proximity can be utilized. Data flow of messages received by the controller (with an unknown destination), can be sent straight to the network server instead of returning it to the receiving gateway, therefore saving the network and devices utilization. The controller must be also able to effectively communicate with all SDN devices, either via in-band, or preferably out-of-band connection. Dedicated connections are more expensive but ensure better performance and security.
- Network server modifications: the application has to be modified to allow communication via an API to the SDN controller. Depending on the SDN application functionality, the controller may need to keep records of some data from received LoRa messages (which sensor is transmitting to which gateway, typical RSSI, etc.). This data is already stored on the network server and therefore does not need to be duplicated, if a communication between the network server and the SDN controller is created.

A. *Gateway Selection Example*

This section explores an example of efficient resource usage in a LoRa network integrating the SDN. Fig. 2 shows the described topology with several IoT devices, gateways which relay messages to a core network, the core network and a back-end device used for data storage, analysis and to provide user access.

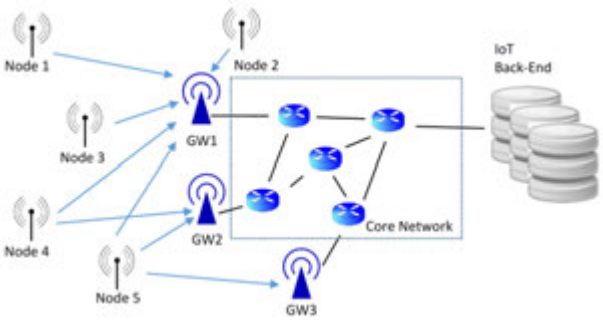


Figure 2. Generic IoT network infrastructure

Wireless devices transmit data via gateways. Redundancy is desirable in order to cope with fluctuating wireless link quality (not all gateways in range of a device will always receive a message). For example, messages from a node can be potentially received by multiple gateways. A simplistic solution (applied in many IoT networks) is to have all gateways to forward received messages to the back-end which then discards any duplicates. This requires the core network to transport duplicates and in a largescale IoT deployment creates an unnecessary burden on the network – as [29] mentioned, high amount of data is one of the biggest problems in current LoRa networks and pre-processing on LoRa gateways is therefore recommended. This applies especially in smart city networks as they transfer large amount of data traffic and a more efficient handling of redundant messages can significantly reduce cost and improve reliability of such networks.

This example proposes a more efficient approach to control the gateway devices and to limit forwarding of duplicates. Depending on the application goals it may be sufficient to use only some gateways as forwarders of received messages. From a set of gateways, the correct ones need to be identified such that application requirements are met. Requirements may be formulated in terms of message delivery reliability or simply path redundancy. Gateway selection strategies (i.e. network functions) can be implemented for the IoT edge network at hand, independently of a present SDN capability in the core network. However, there are benefits of a tight integration.

Gateway selection as a network function can be easily implemented using an SDN infrastructure. Packets received from a node for the first time by a gateway can be forwarded to an SDN controller. The controller keeps track of received packets of gateways and decides which of gateways reporting the same received data packet should be instantiated as forwarders. Gateway selection can be implemented as a service on the SDN controller. To implement this function a simple program on the controller is needed, all other required functionality is provided by the SDN infrastructure. Hence, a function such as gateway selection can be implemented with little effort in an SDN environment.

Besides the reduced implementation effort there are other benefits. The gateway selection is implemented in the same way as other network functions used in the core network. This potentially reduces administration and

operation cost as the same toolsets can be used. Furthermore, gateway selection for different IoT edge network types can be implemented using the same method. Thus, code can be reused, and different network types can easily be configured to run the same gateway selection strategy. Finally, the SDN integration with the IoT edge network enables better coordinated network-wide responses. For example, if the core network is overloaded it would be possible to temporarily reduce forwarding gateways to just one per node. Such integrated responses cannot be implemented if network infrastructures are operated disjoint.

B. *The Integration Architecture*

The LoRa-SDN integration (as described in the above-mentioned example) requires several infrastructure modifications – especially on the LoRa gateways. These modifications are shown in Fig. 3. All necessary changes are software-based, and they require customized order of message processing, message encapsulation and message de-encapsulation.

When a LoRa message is received by a gateway, it has to be firstly forwarded to the *custom application 1*, which parses the message and selects important fields for the SDN operation. These fields have to be encapsulated into standard fields or the experimenter header field, in order to be matched by the OpenFlow protocol. The most interesting field is sensor ID, which can be used to identify the message source. The second step is to encapsulate the LoRa message into the traditional IP (preferably IPv6) packet, which will be used only locally on the gateway. Therefore, not all the header fields have to be filled, or they can be filled only with arbitrary numbers. This encapsulation can be performed by a virtual network device such as a tunnel interface (for example *gre tunnel*).

The encapsulated packet is then passed to the SDN software switch, which has to be implemented in the gateway (depicted by blue colour). As already described, this switch can use arbitrary implementation, as long as it supports the OpenFlow protocol. Such a switch has an OpenFlow table – a data structure containing rules from the SDN controller. The table is searched for a match based on pre-defined header fields set in the first step. If no match is found, the message has to be re-encapsulated and send to the controller for a more detailed processing. Otherwise, the match is found and the corresponding rule either allows or discards the message (or performs any other defined function).

If the message is allowed for forwarding, it is sent to the *custom application 2*. This application takes the original payload and de-encapsulate it from the temporal (IPv6) packet. The payload is then inserted into the final UDP segment (step 3), which is then forwarded to the core network. This application is therefore only a slight modification of the application, which is normally used in typical LoRa gateways. The only difference is, that it takes the custom IPv6 packet instead of the original LoRa frame. The message is then delivered to the LoRa manager / network server.

In case of a new traffic flow from a node, no corresponding OpenFlow rule is found and the message has to be forwarded to the SDN controller. The controller must compare its database of received messages (can be integrated with the network server) and it can then decide, if the message should be allowed or denied (on the particular gateway). The corresponding OpenFlow rule is then created and inserted into the gateway via a *FlowMod* message. All the consequent messages can then be processed immediately on the device, without the need to further contact the controller. This will result in faster and more efficient processing. The controller can then periodically update the rules, or these rules can automatically time-out. This would result in the need to contact the controller again for new message processing rules.

C. *The Integration Architecture Implementation*

The proposed architecture can be implemented in all LoRa networks, which use gateways with open and programable operating system. The architecture provides abstraction for LoRa IoT edge networks, which can then be integrated into a smart city network controlled by SDN. This segment of the smart city network can then be controlled and managed independently on the used communication technology.

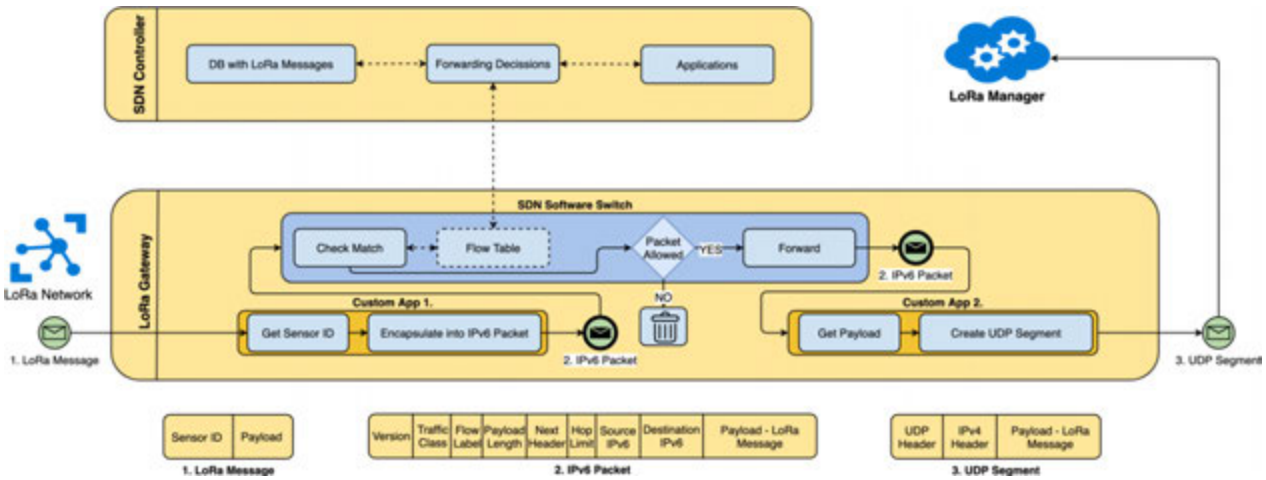


Figure 3. The LoRa-SDN integration architecture: gateway modifications

VI. CONCLUSION

The growth of IoT networks will require innovative traffic handling, which would be able to cope with the strict requirements – mostly on abstraction from several heterogeneous technologies, massive amount of traffic, performance, and security. One of the most promising technologies, which can help with these demands is SDN.

The paper analysed, how SDN can be integrated into LoRa-based IoT networks, which represent the dominating technology used in smart cities and similar large-scale scenarios. The paper proposed the integration architecture which requires only a software modification of LoRa gateways – devices used to interconnect IoT nodes with the core network (which can utilize SDN). The modification requires several steps: to write a custom application, which encapsulates LoRa message frames into IPv6 packets and inserts selected important data into common header fields. The IPv6 packet then has to be forwarded into the installed software switch with support of the OpenFlow protocol. This switch contains rules from the SDN controller and controls the message processing. The last step is the second custom application, which is used for final message processing, if the message should be forwarded. In this case, the IPv6 packet has to be de-encapsulated and re-encapsulated into the final datagram. This is typically achieved using UDP segments with the payload in the JSON format.

In general, the proposed architecture can be – with minor modifications – used in any IoT technology (not only LoRa). SDN integration with any IoT technology requires only usage of either standard or experimenter fields for OpenFlow protocol matching process. This allows abstraction from the used transmission protocol and integration of different IoT technologies under the unified network management. This approach can then help to meet the future IoT demands.

REFERENCES

[1] Z. Qin, G. Denker, C. Giannelli, P. Bellavista, and N. Venkatasubramanian, "A Software Defined Networking architecture for the Internet-of-Things," in 2014 IEEE Network Operations and Management Symposium (NOMS), May 2014, pp. 1–9.

[2] A. El-Mougy, M. Ibnkahla, and L. Hegazy, "Software-defined wireless network architectures for the internet-of-things," in Local Computer Networks Conference Workshops (LCN Workshops), 2015 IEEE 40th, Oct 2015, pp. 804–811.

[3] K. Sood, S. Yu, and Y. Xiang, "Software-Defined Wireless Networking Opportunities and Challenges for Internet-of-Things: A Review," IEEE Internet of Things Journal, vol. 3, no. 4, pp. 453–463, Aug 2016.

[4] D. H. Kim, J. B. Park, J. H. Shin, and J. D. Kim, "Design and implementation of object tracking system based on lora," in 2017 International Conference on Information Networking (ICOIN), Jan 2017, pp. 463–467.

[5] W. San-Um, P. Lekbunyasin, M. Kodyoo, W. Wongsuwan, J. Makfak, and J. Kerdsri, "A long-range low-power wireless sensor network based on u-lora technology for tactical troops tracking systems," in 2017 Third Asian Conference on Defence Technology (ACDT), Jan 2017, pp. 32–35.

[6] K. Mikhaylov, J. Petajajarvi, and T. Haenninen, "Analysis of capacity and scalability of the LoRa low power wide area network technology," in European Wireless 2016; 22th European Wireless Conference, May 2016, pp. 1–6.

[7] T. Petri, M. Goessens, L. Nuaymi, L. Toutain, and A. Pelov, "Measurements, performance and analysis of lora fabian, a real-world implementation of lpwan," in 2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Sept 2016, pp. 1–7.

[8] J. Petajajarvi, K. Mikhaylov, M. Hmlinen, and J. Iinatti, "Evaluation of LoRa LPWAN technology for remote health and wellbeing monitoring," in 2016 10th International Symposium on Medical Information and Communication Technology (ISMICT), March 2016, pp. 1–5.

[9] A. J. Wixted, P. Kinnaird, H. Larijani, A. Tait, A. Ahmadinia, and N. Strachan, "Evaluation of lora and lorawan for wireless sensor networks," in 2016 IEEE SENSORS, Oct 2016, pp. 1–3.

[10] J. Petajajarvi, K. Mikhaylov, A. Roivainen, T. Hanninen, and M. Pettissalo, "On the coverage of lpwans: range evaluation and channel attenuation model for lora technology," in ITS Telecommunications (ITST), 2015 14th International Conference on, Dec 2015, pp. 55–59.

[11] [21] L. Li, J. Ren, and Q. Zhu, "On the application of lora lpwan technology in sailing monitoring system," in 2017 13th Annual Conference on Wireless On-demand Network Systems and Services (WONS), Feb 2017, pp. 77–80.

[12] M. Lauridsen, B. Vejlgaard, I. Z. Kovacs, H. Nguyen, and P. Mogensen, "Interference measurements in the european 868 mhz ism band with focus on lora and sigfox," in 2017 IEEE Wireless Communications and Networking Conference (WCNC), March 2017, pp. 1–6.

[13] L. Gregora, L. Vojtech, and M. Neruda, "Indoor signal propagation of lora technology," in 2016 17th International Conference on Mechatronics - Mechatronika (ME), Dec 2016, pp. 1–4.

[14] D. H. Kim, J. Y. Lim, and J. D. Kim, "Low-power, long-range, high-data transmission using wi-fi and lora," in 2016 6th International Conference on IT Convergence and Security (ICITCS), Sept 2016, pp. 1–3.

[15] F. Orfei, C. B. Mezzetti, and F. Cottone, "Vibrations powered lora sensor: An electromechanical energy harvester working on a real bridge," in 2016 IEEE SENSORS, Oct 2016, pp. 1–3.

[16] O. Georgiou and U. Raza, "Low power wide area network analysis: Can lora scale?" IEEE Wireless Communications Letters, vol. 6, no. 2, pp. 162–165, April 2017.

[17] M. Haghighi, Z. Qin, D. Carboni, U. Adeel, F. Shi, and J. A. McCann, "Game theoretic and auction-based algorithms towards opportunistic communications in lpwa lora networks," in 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), Dec 2016, pp. 735–740.

[18] S. Thielemans, M. Bezunartea, and K. Steenhaut, "Establishing transparent ipv6 communication on lora based low power wide area networks (lpwans)," in 2017 Wireless Telecommunications Symposium (WTS), April 2017, pp. 1–6.

[19] Luis Tello-Oquendo, Ian Akyildiz, Shih-Chun Lin, Vicent Pla, "SDN-Based Architecture for Providing Reliable Internet of Things Connectivity in 5G Systems," 17th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net 2018), Jun 2018, Capri Island, Italy. IFIP eCollection-2, pp.108-115, 2018, Annual Mediterranean Ad Hoc Networking Workshop.

[20] "LoRaWAN," http://lora-alliance.org.

[21] "Sigfox," http://www.sigfox.com.

[22] "Ingenu RPMA," https://www.ingenu.com/technology/rpma/.

[23] "Weightless," http://weightless.org.

[24] "Zigbee alliance", https://www.zigbee.org/.

[25] "Z-Wave", https://www.z-wave.com/.

[26] "FieldComm Group", https://fieldcommgroup.org/.

[27] "Semtech", https://www.semtech.com/lora.

[28] Open Networking Foundation. (2016) Openflow. [Online]. Available: https://www.opennetworking.org/sdn-resources/openflow

[29] J. So, D. Kim, H. Kim, H. Lee, and S. Park, "LoRaCloud: LoRa platform on OpenStack," in 2016 IEEE NetSoft Conference and Workshops (NetSoft), June 2016, pp. 431–434.

Analyses of Ecological and Energy Footprint as indicators of Energy Management in the transition to sustainability using social networks

Bojan Mihajlovski*, Bekim Fetaji**, Lejla Abazi*, Majlinda Fetaji*
* South East European University, Contemporary Sciences and Technologies, Ilindenska bb, 1200 Tetovo, Macedonia
** University Mother Teresa, Informatics, 1000 Skopje, Macedonia
bm34118@seeu.edu.mk, bekim.fetaji@unt.edu.mk, l.abazi@seeu.edu.mk, m.fetaji@seeu.edu.mk

***Abstract*— The aim of this research study was to investigate Energy Management issues (increased energy use, limited raw fuel reserves, global warming, etc.) and then develop and increase awareness and perception of the Ecological and Energy Footprint as indicators of sustainability and Energy Management. The model developed for this purpose is called the Ecological and Energy Footprint Analysis Model (EEFAM) which offers metrics for comparing the Ecological and Energy Footprint with the available Bio-capacity. Case Study has been chosen. As research method used questionnaire, and methodology used was grounded theory where we will collect different data from the Ministry of Ecology and later process them using different Software Tools. Findings, issues and recommendations have been provided, discussed and argumented.**

Keywords*— Ecological and Energy Footprint, sustainability, efficiency, metrics, energy management

I. INTRODUCTION

The research field falls in the multidisciplinary intersection of Energy Management, Environmental dimension and computing social networking. Ecological and Energy Footprint as indicators of sustainability and sustainable development, present a challenge to ensure continuous improvement of the energy management, ecology, and their quality assurance.

Unsustainable trends, such as climate change, energy usage, poverty and social exclusion, demographic pressure and aging, threats to public health, use of natural resources, and loss of bio-diversity, are all processes which call for a high level of urgency as disscused by [1]. The destiny of the future generations is in the hands of the present generation, which has an obligation to preserve the planet's capacity to support life with its complete diversity, and to ensure sustainable development [2].

Sustainable development includes three mutually-dependent and interrelated dimensions:

- Environmental dimension (natural resources, protection, and sustainable use of nature and a fight against pollution);
- Social networking (social resources, social aspects);
- Economic dimension (economic resources, development and growth).

The entire shift of the progress of civilization to "sustainable development" depends to a large extent on the transition from the use of conventional energy sources to alternative ones, and in order to promote alternative sources, it is necessary to emphasize that the basic tool for this endeavor is Energy Management (EM). The main goal of EM is to increase profits by reducing costs in all forms of energy consumption, while one of the most important specific goal is reduction of emissions of harmful gasses and reduction of air pollution.

The sustainable indicators reflect the trends in socio-economic development, sustainable consumption and production, social inclusion, demographic changes, public health, climate change and energy, natural resources, sustainable transport, and trends in global partnership and good governance. The sustainable indicators are defined in accordance with the European Strategy for Sustainable Development, and are calculated according to a methodology harmonized with that of the European Union. Therefore, they are comparable with the sustainable development indicators in European countries, and they are a good platform for monitoring the priorities in national policies for sustainable development.

An Energy Footprint is a measure of land required to absorb the CO2 emissions. This approach focusses on the outcome of energy use that is CO2 emissions and it is following the guidelines from [1]. Ecological and Energy Footprint as indicators of sustainability, and Ecological and Energy Footprint calculation can provide a baseline from which progress is measured toward a more sustainable lifestyle, and help in achiving a smaller Footprint. This measurments and objective is the motivation of the research study.

II. OBJECTIVE AND AIM OF RESEARCH

The aim of this research study is to investigate current situation in Energy Management issues (increased energy use, limited raw fuel reserves, global warming, etc.) and then develop and increase awareness and perception of the Ecological and Energy Footprint as indicators of sustainability and Energy Management. The model to be developed for this

purpose is called the Ecological and Energy Footprint Analysis Model (EEFAM) which offers metrics for comparing the Ecological and Energy Footprint with the available Bio-capacity. Case Study has been chosen.

Social networks will be used to analyze and measure the values of the personal Ecological and Energy (CO_2) Footprint, as an indicators of sustainability. By using social networks important information can be obtained about the attitude of the individual toward nature, and the awareness of their impact on the habitat. This analysis would enable us to undertake personal and collective initiatives directed toward a more rational use of the existing resources, as well as raising the quality of life and sustainability. Survey and Questionnaires as quantitative method of collection of data will be distributed through using social networks like Facebook and Twitter to analyze and measure the respondent inputs. Also, focus groups for capturing the qualitative input will be used within the social networks.

For some time, the global public has undertaken an initiative for an inclusive set of Sustainable Development and Energy Management objectives, and a number of measures have been undertaken in the world. They are aimed toward reducing starvation and impoverishment; improving medical protection, birth prevention and schooling; providing conditions for the creation of a low-emission society; and, improving agricultural productivity, while decreasing habitat deterioration. Today, thousands of administrative and non-administrative corporations, private companies and persons around the globe have accepted the concept of sustainability and energy management, and have started to raise awareness, and funds, for sustainability and energy management plans. Companies have provided metrics to follow the economic, social and environmental effects of their activities. The overall aim of sustainable development is to develop a set of actions and measures that will enable continuous improvement of the quality of life, both for the current and future generations, through the creation of sustainable communities, which will be able to manage and use resources efficiently, and apply the ecological and social innovation potential of the economy, ensuring prosperity, environmental protection, and social cohesion. For this reason, one of the directions of future development should aim at strengthening the whole system for implementing sustainable development and energy management, including a holistic approach, instead of the present, fragmented one. The aim of the Ecological and Energy Footprint, as indicators of sustainability, is to prove that the present way and rhythm of living is not sustainable. Therefore, it is important to take concrete action to inform society about the importance of the Ecological and Energy Footprint. Additionally, the Ecological Footprint can serve as a useful means of education, teaching people about the Earth's limited capacity, and human overconsumption, with the aim of changing the individual lifestyle, arguing that many contemporary lifestyles are not sustainable.

III. RESEARCH METHODOLOGY

Research Methodology to be used is Qualitative and Quantitative methodology.

We will also use the triangulation technique that combines both of the quantitative and qualitative methodology and means using more than one method to collect data on the same topic. This is a way of assuring the validity of research through the use of a variety of methods to collect data on the same topic, which involves different types of samples as well as methods of data collection. For quantitative methodology will use questionnaires while for qualitative will use focus groups.

The model to be developed for this purpose is called the Ecological and Energy Footprint Analysis Model (EEFAM) which offers metrics for comparing the Ecological and Energy Footprint with the available Biocapacity.

As research method will be used questionnaire, and grounded theory where we will collect different data from the Ministry of Ecology and latter process them using different Software Tools. Software to be used are SPSS for Statistical data analyses, online questionnaire, and different social networks software.

Sustainable development is necessary, but without any measurable proof such a claim can hardly attract any attention. Only clear metrics can help this concept turn into an argument based on empirical facts. The first step in achieving sustainable development is to collect information on the extent of mankind's capacity of natural resources; how much we need, and how much of the natural resources we spend. The method developed for this purpose is called the Ecological Footprint Analysis. The Ecological Footprint Indicator gives evidence of human consumption of natural capital, while the Ecological Footprint Analysis offers metrics for comparing the Ecological Footprint with the available Biocapacity.

The Ecological and Energy Footprint can be measured for an individual, city, state, region, and the entire world. Also, an individual Ecological and Energy Footprint can be measured for a particular product, service, or process based on their life cycle.

The Ecological and Energy Footprint of the individual provides insight into how everyday personal activities and habits have a harmful environmental impact and impede sustainable development. The Ecological and Energy Footprint of a person shows by what kind of individual or collective action we can reduce our harmful influence.

The Ecological and Energy Footprint of a person is measured by a personal index calculator. The personal index calculator sets up a series of simple questions and then calculates how much land, in global hectares, is needed to support our lifestyle.

For the purposes of this research, the "Footprint Calculator" survey will be used, which consists of data needed to calculate a future Ecological Footprint. The survey will be applied to a sample of 70 respondents which we have been using as case study. Those are first-year students from the Department of English Language and Literature at the 'Blaze Koneski' Faculty of Philology (UKIM) in Skopje. 35 male and 35 female examinees will be surveyed.

The survey consists of 27 multi-response questions. The answers are related to the consumer habits and lifestyle of the respondents. The survey questions, concerning students' basic needs, are categorized into the following four consumer areas or categories: Food, Housing (an object that includes the use of energy), Products and Services, and Mobility (personal and public transport).

The responses from the survey as data will be entered separately for each respondent, in an internet calculator created by the Global Footprint Network (GFN), the so-called Footprint Calculator.

The answers have been calculated by the calculator in equivalents of the Ecological and Energy Footprint (global hectares, accepted as an international standard), and statistically analyzed by Excel.

IV. HYPOTHESES AND RESEARCH QUESTIONS

The hypothesis:

H1: By raising the awareness and perception for Ecological and Energy Footprint rises the impact they have on sustainability.

By analyzing and measuring the values of the personal Ecological and Energy Footprint, as indicators of sustainability, and by using social networks, important information can be obtained about the attitude of the individual toward nature, and the awareness of their impact on the habitat. This analysis would enable us to undertake personal and collective initiatives directed toward a more rational use of the existing resources, as well as raising the quality of life and sustainability.

According to statistical data and fact-oriented material which will be collected, it is anticipated that the research will reveal a low awareness among respondents about the impact they have on sustainability with their personal lifestyle.

Research questions to be investigated:

How much can social networks help in comparing the Ecological and Energy Footprint with the available Biocapacity?

How much we, as individuals, have an impact on the environment with our ecological and energy footprint?

Is our personal awareness for ecological and energy footprint directly connected with the impact we have on the environment?

V. CALCULATION METHODOLOGY

The calculation methodology of the Ecological Footprint is prepared by the international organization known as the "Global Footprint Network". This organization redefines and corrects Ecological Footprint calculation methodology on a yearly basis. There is no one specific Ecological Footprint calculation methodology; however, present methodologies are similar (Wackernagel, 2006; Calculation, 2008). Generally speaking, the Ecological Footprint is calculated using formulas 1 and 2 (Calculation, 2008):

$$ai = ci / yi \times F \times Ef \quad \ldots \quad (1)$$

where: ai – Ecological Footprint of each element;
ci – Yearly consumption of the element;
yi – Land productivity or output of each element (kg/ha);
F – Yield factor;
Ef – Equivalence factor;
Fp – total Ecological Footprint for a population.

The first formula (1) is used for the calculation of each component of the Ecological Footprint needed for calculation of Ecological Footprint. In other words, the Ecological Footprint of each component (for instance, field of grain) is calculated by dividing the yearly consumption of the element by the productivity of the land (component output kg/ha), and this ratio is multiplied by the yield factor F and equivalence factor Ef. These factors are used in transforming of land into global hectares.

Generally, the Ecological Footprint is a negative ecological indicator, while, the other Ecological Footprint related indicator – Biocapacity (BC) can be perceived as a positive ecological indicator. Biocapacity is the capacity of the land's biosphere to produce renewable resources. Generally, we can say that the Ecological Footprint represents the needed resources to meet the needs of the people (their demand), and the Biocapacity of the land is the ability of the land to create the resources (supply). Biocapacity of the land or certain territory, as well as Ecological Footprint is measured in global hectares (gha). Land's biocapacity is calculated based on the following formula (Calculation, 2008):

$$BC = S \times F \times Ef \quad (2)$$

where: BC – biocapacity of the land (territory); S – area; F – yield factor; Ef – equivalence factor.

VI. RESULTS

The acceptance and application of the Ecological Footprint as one of the indicators of sustainability is widespread.

Several countries (Switzerland, Japan), regions (Wales, South Australia, and Victoria), and cities (Calgary, London, Cardiff, Santa Monica, and The Hague) have accepted the Ecological Footprint as a measure of their own sustainability, and have prepared plans and strategies for reducing it. The United Nations, in the document "State of the World Population 2001", calculated the Ecological Footprint. It can be seen from the below Table that all Balkan countries are in Ecological Deficit.

Country, region	Global hectares per capita		
	Biocapacity	Ecological Footprint	Ecological Deficit (-) or Reserve (+)
Worldwide average	**1.68**	**2.84**	**- 1.16**
High income countries	**2.91**	**6.12**	**- 3.21**
Middle income countries	**1.58**	**2.43**	**- 0.85**
Low income countries	**1.10**	**1.12**	**- 0.02**

Table 1 - Ecological Footprint and Biocapacity of the countries and regions by income (Source: Global Footprint Network)

Balkan Country	Total Biocapacit	Total Ecologic al	Biocapaci ty (Deficit)	Number of Earths required	Number of Countries
Romania	2.7	2.6	0.1	1.5	1.0
Bulgaria	3.1	3.1	0.0	1.8	1.0
Montenegro	3.1	3.6	-0.5	2.1	1.2
Croatia	2.8	3.8	-1.0	2.2	1.4
Albania	1.2	2.3	-1.1	1.3	1.9
Macedonia	1.6	3.1	-1.5	1.8	2.0
Serbia	1.6	3.1	-1.5	1.8	2.0
Bosnia and Herzegovina	1.7	3.2	-1.5	1.9	1.9
Slovenia	2.3	4.7	-2.4	2.8	2.1
Greece	1.5	4.2	-2.7	2.5	2.8

Table 2 - Ecological Footprint and Biocapacity of the Balkan countries

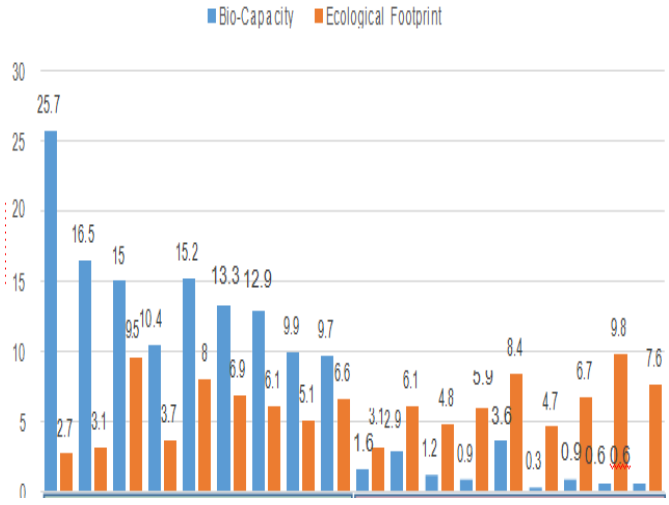


Figure 1 - Ecological Footprint and Biocapacity

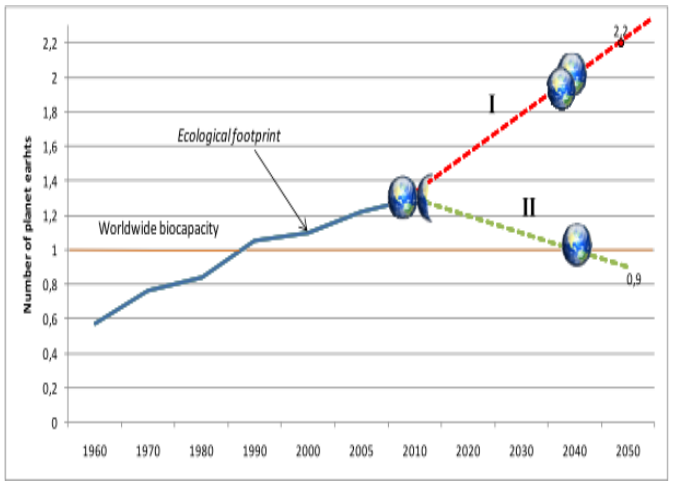


Figure 2 - Possible scenarios of our planet's ecological development

Environmental degradation of our planet and Ecological Footprint indicators can be reduced by the implementation of various governmental and public administration measures. That can be illustrated with the example from EU environmental policies.

Observations from the results

Based on Table below, results reveal that the university students' average consumption Ecological Footprint is two

times bigger (6.45 gha) than the Macedonian national average (3.1 gha).
Furthermore, most of their Ecological Footprint comes from their Carbon Footprint (65%).

	All	Male	Female	National
Total Ecological Footprint (gha)	**6.45**	**6.47**	**6.42**	**3.10**
Carbon Footprint, (CO_2 emissions in tons per year)	**12.26**	**12.24**	**12.28**	**5.10**
Carbon Footprint (% of total Ecological Footprint)	**65%**	**65%**	**65%**	**61%**

As it can be seen from Figure 3, according to the consumer categories, the most important share in the total Ecological Footprint is held by the following three categories: Products and Services (3.53 gha), Transport (1.26 gha) and Food (0.89 gha).

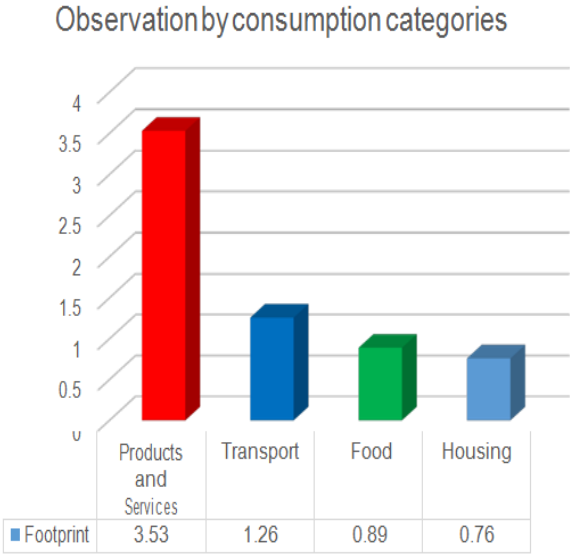


Figure 3 - Ecological Footprint of the respondents by the consumption category

VII. CONCLUSION

The research study is calculating the Ecological and Energy Footprint on a selected sample of respondents as case study. For the purposes of this research, the "Footprint Calculator" questionnaire is used, which consists of data needed to calculate a Personal Ecological and Energy Footprint.

Statistical and descriptive methods are used for data analysis. The statistical method is applied to show the attitude of the respondents, provided in percentage and in number, while the descriptive method is used to explain the results obtained. In an emerging sustainability science, much work has been done on indicators of sustainable development. Yet, to date, there are no indicator sets that are universally accepted, backed by convincing theory, precise data collection and analysis, and influential in policy. Why is this so? We offer three major reasons:

a. the enigma of sustainable development;

b. the large number of purposes in characterizing and measuring sustainable development; and,

c. the confusion of terminology, data, and methods of measurement.

Ecological and Energy Footprint indicators provide a baseline from which to measure progress toward smaller Footprints, and a more sustainable lifestyle. It challenges students to set goals for themselves in each lifestyle category (i.e., to eat less meat, or to spend more time outdoors).

Someone commented very aptly: "There are no passengers on this passenger ship; we are all crew members." This means that all people, as part of the biosphere, have great responsibility in view of the deliberately planned activity that affects its integrity.

The perspectives that we will share in this thesis are meant to help individuals, and especially scientists, be agents of change in the sustainability and energy management transition. We will present a framework for relating the sustainability goal of inter-generational well-being to five basic assets that ultimately determine society's ability to meet those goals. We will suggest that an appreciation is needed of the governance processes through which agreements and decisions are made, allowing communities to achieve together what cannot be achieved alone. And, finally, we will draw attention to the ways that useful knowledge, tools, and approaches can be developed in collaboration with decision-makers and other stakeholders at all levels.

Sustainability and Energy Management is not something to be achieved, but is instead something to be worked toward. The world keeps changing, new challenges continue to emerge and change the game, and some seemingly smart things do not work. Yet, each step forward matters, and the sooner the better. Good intentions are not enough; each of us needs to take the next steps toward sustainability and better energy management. The inevitable conclusion seems to be that less is more: we all need to shrink our Ecological and Energy Footprints in order to make our contribution toward sustainability and better energy management.The importance of the thesis is to highlight, through theoretical considerations

and research of the Ecological and Energy Footprint value for university students, the significance of the sustainability concept, which implies respect for nature and its laws, and improved well-being for present and future inhabitants of the Earth. At the same time, it is also essential to point out to one of the ways by which sustainability and energy management are achieved, that is, the balance between human needs and the bio-capacity of the Earth, which can be accomplished by the Ecological and Energy Footprint Analysis.

The topic of Ecological and Energy Footprint is rapidly gaining importance. Nowadays it has become a hot topic globally. The aim of the Ecological and Energy Footprint is to prove that the present way and rhythm of living is not sustainable. Therefore, it is important to take concrete action to inform society about the importance of the Ecological and Energy Footprint, and, with that, to raise public awareness.

What is the student's motivation to reduce their individual Ecological and Energy Footprint? If there is no target to achieve, then an individual can never do enough; their efforts will most likely tend to increase and decrease over time, but, more importantly, highly-motivated individuals will always feel guilty that they are not doing enough.

The information which will be revealed in the research will be intended to help the Faculty in its future directions for a more sustainable university. Moreover, students need to think on a personal level about how they can reduce their impact on nature. As the study will show, a reduction in carbon intensive activities and consumption, as well as a decrease in the dependence on forest- and cropland-based products may be a practical way to reduce one's personal ecological impact. In light of our present dilemma with the emergence of climate change in which greenhouse gases are the main cause, our prediction is that students at the university will show to have a greater impact in terms of carbon emissions as compared to other types of environmental impacts. This, essentially, will call for a thorough and massive information dissemination with regards to carbon footprint reduction, as well as the promotion of sustainable lifestyles around the campus, such as walking or biking rather than driving, consuming organic foods, energy conservation, and so on. In this context, the Ecological Footprint can be used as an educational tool apart from being a policy tool, especially for climate change mitigation, as well as for addressing natural resource degradation.

The aim of the Ecological Footprint is to prove that the present way and rhythm of living is not sustainable. Therefore, it is important to take concrete action to inform society about the importance of the Ecological Footprint. Nature can renew its resources, however, only at a certain pace. People are constantly consuming more renewable resources than the nature is able to regenerate. The main global aim is to unite the humanity to save the planet. As long as the representatives of the governments and corporations are unfamiliar with the level of natural resources used, and how these resources match with the present stock of resources, the over-usage may remain unnoticed, which will cause the ecological deficit to grow, and the opportunities to satisfy social needs will decrease.

REFERENCES

[1] Bonini, S., Oppenheim, J. (2008). "Cultivating the Green Consumer". Stanford Social Innovation Review, 6, 56-61.

[2] Boyle, G. (Ed). (2012). Renewable energy: power for a sustainable future. 3rd Ed. Oxford University Press. UK

[3] Credit Suisse. (2014). Credit Suisse Global Wealth Databook. Zurich: CreditSuisse Research Institute

[4] Dasgupta, P. (2004). Human Well-Being and Natural Environment. New York: Oxford University Press;

[5] United Nations International Human Dimensions Programme (UNU-IHDP) and United Nations Environment Programme (UNEP). 2012. Inclusive Wealth Report 2012: Measuring Progress toward Sustainability. Cambridge: Cambridge University Press.

[6] Harris, J. Wise, T. Gallagher, K. Goodwin, R.N. (2001). A Survey of Sustainable Development. Island Press.Hart, S., (January, 1997). "Beyond Greening: Strategies for a Sustainable World", Harvard Business Review.

[7] Heinrichs, H., Martens, P., Michelsen, G., Wiek, A. (Eds.). (2016). Sustainability Science. Springer.

[8] Kearney, J.(2010). "Food Consumption Trends and Drivers" Philosophical Transactions of the Royal Society B: Biological Sciences 365(1554): 2793-807.

[9] Kumar De A. (2004). Environmental Education, New Age International (P) Ltd., Publishers, New Delhi.

[10] Matson, P., Clark, W. Anderson, K. (2016). Pursuing sustainability. Princeton University Press. New Jersey. USA.

[11] Ruževičius, J., & Waginger, E. (2007). "Eco-labelling in Austria and Lithuania: a comparative study". Inzinerine Ekonomika-Engineering Economics, 4, 96-102.

[12] Zobaa, A., Bansal, R.C. (2011). "Handbook of Renewable Energy Technology". Brunel University, U.K.

Review of discrete simulation modelling use in the context of smart cities

Mario Jadrić, Maja Ćukušić and Dino Pavlić
University of Split, Faculty of Economics, Business and Tourism, Split, Croatia
{mario.jadric, maja.cukusic, dino.pavlic}efst.hr

***Abstract* - The need to plan integral systems that can process massive amounts of data (for services such as smart waste management, smart traffic control, or citizen safety management); the great availability of different commercial ICT solutions in this domain; and the need for devising new business models and services for citizens – all pose a real challenge for city planners. The described context is an ideal one for using concepts, models and tools from (discrete) simulation domain. The paper systematically reviews relevant studies where discrete event simulation has been used to support the design of the enabling ICT infrastructure for smart cities. Following the literature review, the studies are mapped to eight different categories (engagement and community, economic development and housing, energy, waste, water, mobility, security, and healthcare) followed by the discussion and identification of the promising scenarios for decision making in these specific areas.**

Keywords – discrete simulation modelling; smart cities

I. INTRODUCTION

Managers and planners of smart city initiatives in developing strategies and policies and in evaluating their effect in particular should be aware of the challenges and the potential for the implementation of smart city services/applications in their (local) contexts. The continuous urbanization and technological advancement forces city planners to be more creative and supported by approaches that enable an abstract perspective on the complex reality as recent studies have argued [1]. In an effort to grasp possible directions and the potential for using simulation modelling as one of the approaches for planning and optimizing either the implementation of plethora of smart city applications or some specific aspect of its management, we have attempted to systemize the areas where simulations were used either as a tool to predict the impact on user behaviors for the purpose of improving (key) urban business processes or to devise complex what-if scenarios related to present-day smart city challenges.

In its second part, this review paper focuses on smart city categories/domains that are commonly considered as the most relevant ones followed by the presentation of benefits and possibilities of using various simulation paradigms to plan or optimize processes or scenarios from selected domains. The methodology, the results and conclusions of the study that aims to identify and analyze relevant scientific papers dealing with simulations in smart cities are presented in parts 3,4,5 and 6 respectively.

II. THEORETICAL BACKGROUD: KEY CONCEPTS

A. Categorisation of Smart City Intervention Areas

In defining smart cities, authors [2-5] commonly emphasize that ICT is both a driver and a prerequisite, but also stress the socio-economic, managerial and multi-stakeholder aspects that are important to achieve the main rationale of a smart city – to achieve better/high quality of life for all its citizens. In practice, a city can be labeled as 'smart' if it implements at least one initiative out of six universally acknowledged action areas [2]: Smart Governance, Smart People, Smart Living, Smart Mobility, Smart Economy and Smart Environment. Other authors recognize a wider number of domains or, as they call them – aspects, e.g. [6]: Smart Energy, Smart Buildings, Smart Mobility, Smart Technology, Smart Healthcare, Smart Infrastructure, Smart Governance and Smart Education, Smart Security and Smart Citizens.

For the purposes of another research study [7], a more specific list of eight categories of smart city applications was found very useful. In an exhaustive report [8], the authors looked at dozens of smart applications that will be relevant for cities through 2025 and gathered evidence from 50 cities from around the globe to assess how smart city applications could affect various quality-of-life dimensions. Here, we present a categorized catalogue of the applications (total of 59) as this is the structure we would base the classification of scientific papers later on:

- Engagement and community: Digital citizen services, Local civic engagement applications, Local connection platforms;
- Economic development and housing: Digital business licensing and permitting, Digital business tax filing, Digital land-use and building permitting, Local e-career centers, Online retraining programs, Open cadastral database, Peer-to-peer accommodation platforms, Personalized education;
- Energy: Building automation systems, Distribution automation systems, Dynamic electricity pricing, Home energy automation systems, Home energy consumption tracking, Smart streetlights;
- Waste: Digital tracking and payment for waste disposal, Waste collection route optimization;
- Water: Leakage detection and control, Smart irrigation, Water consumption tracking, Water quality monitoring;

- Mobility: Autonomous vehicles, Bike sharing, Car sharing, Congestion pricing, Demand-based microtransit, Digital public transit payment, E-hailing (private and pooled), Integrated multimodal information, Intelligent traffic signals, Parcel load pooling, Predictive maintenance of transportation infrastructure, Real-time public transit information, Real-time road navigation, Smart parcel lockers, Smart parking;
- Security: Body-worn cameras, Crowd management, Data-driven building inspections, Disaster early-warning systems, Emergency response optimization, Gunshot detection, Home security systems, Personal alert applications, Predictive policing, Real-time crime mapping, Smart surveillance;
- Healthcare: Data-based public health interventions for maternal and child health, Data-based public health interventions for sanitation and hygiene, First aid alerts, Infectious disease surveillance, Integrated patient flow management systems, Lifestyle wearables, Real-time air quality information, Remote patient monitoring, Telemedicine.

Based on their focus, the selected and analyzed scientific papers in this review study would be mapped out to one of the eight different categories by an expert who has experience in digital transformation of public services.

B. Simulation Modelling and Types of Simulations

A simulation model is a representation of a real system and its behavior, and at the same time it is an instrument to perform different analyses. In some cases, the simulated system is real, however, the cases where it has yet to be designed or implemented are more frequent [9]. All the same, simulation is used for the performance evaluation of complex systems. Consequently, issues at operational and tactical level, in particular, are suited for modelling and analyzing using simulation techniques [10] as the concept implies a proper understanding of the environment and supports the development of multiple potential scenarios [11]. In that way, by using process models and simulation, it is possible to experiment with multiple scenarios (e.g. what-if analyses) in order to find the optimal one [12]. Thus, it is possible to avoid experimentation in real life systems as it can be very costly, but sometimes impossible to perform as well [13].

Depending on the type of a problem, i.e. the system that is simulated and the simulation goals, it is necessary to choose the most appropriate type of simulation. The choice is first between the continuous and discrete simulation depending on the purpose and the context of simulation modelling but also taking into account availability and the appropriateness of software tools [14] to support the simulation. By general understanding, the type of simulation that is quite easy to use is the discrete event simulation. It is commonly used for optimization of business processes. This simulation paradigm is presented in detail in a recent study [9] and contrasted with its enhanced versions – Parallel Discrete Event Simulation (PDES) and Parallel and Distributed Simulation (PADS). Also, the use of agent-based simulations and coupling of different types of simulations in hybrid, multi-level simulation frameworks is explained and advocated in the paper.

Acknowledging the differences between simulation paradigms, the selected and analyzed scientific papers in this review study would be categorized as either using discrete event simulations, agent-based simulations, continuous – system dynamic models or as hybrid approach simulations by an expert with keen interest in business process simulations.

C. Motivation for the Study on Simulations for/related to Smart Cities

The authors of this paper have been using simulation tools in their research and teaching efforts. As of recently their interest shifted to investigating the challenges and the potentials for implementing smart technologies in their community with the purpose of improving key urban business processes. Two discrete event simulation models were developed to demonstrate usefulness of simulations in that regard. Proof-of-concept simulation model was built in Arena as a decision support tool for city managers [7]. Specifically, having a large portfolio of all possible/available smart city applications, and based on the internally- and/or externally-set priorities, including citizens' opinions and experts' assessment about whether it is possible or not to implement the application within a five-year timeframe, a city manager can experiment to select the optimal set of applications handled by the designated smart city department. Another model was developed using ARIS simulator to investigate benefits of using smart technologies in airport operations [15] as a part of ongoing smart initiatives and solutions - specifically in the context of internal business processes and organizational unit optimization.

In the process of preparing the two publications, it has become apparent that there are some smart city domains where simulations are used intensively. For example, there are many cases where simulations are used to test new concepts for reducing energy consumption in smart cities [16]. Other prominent cases of simulations include management of transportation systems that have made good use of smart city data-driven concepts coupled with simulations for some time [17]. However, there are some smart city domains where the examples of using simulations as a tool are scarce.

To get a good grasp on the dominant type of simulation models and the less-explored areas where simulations could provide a novel perspective and uncover areas where research is needed, a study of relevant literature was conducted. The effort to provide the "state of the field" is presented in the following section of the paper. The reviewing and revision process is based on the relevant recommendations [18].

III. LITERATURE REVIEW

This part presents the process of identification and review of relevant studies where the design and application of the ICT infrastructure for smart cities was analyzed and supported by (discrete event) simulation. It was decided that only studies that used simulation as a

tool to predict the impact on user behaviors, to improve the core urban business processes or to conduct complex experimental what-if scenarios would be analyzed. Simple studies that only simulated data regardless of the application area were excluded from the analysis as is described later in the paper.

A. Identifying Relevant Literature

A smart city is a multidisciplinary concept. Therefore, for the literature review, as the most appropriate resource, Web of Science database was selected as it contains multidisciplinary content that covers highest impact journals worldwide but also includes conference proceedings. The search was performed using the web interface (webofknowledge.com) and limiting the results to Web of Science Core Collection. The database was searched on January 18th 2019 using the following query:

TOPIC: (SMART CIT) AND TOPIC: (SIMULATION) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article OR Proceedings Paper).*

The search returned a total of 1.029 database records as illustrated in Fig. 1. The records are presented per Web of Science categories. Only top 10 categories are visualized, each containing more than 50 publications or more than 5% of records. Most of the studies are from engineering (42.33%) and telecommunications (29.22%). However, the multidisciplinary nature of the concept is evident as there are studies from computer sciences, transportation sciences, energy and green sustainable science and technology and so on. As for the document types, most of the records are proceedings papers (541, i.e. 52.57%), followed by a similar number of journal papers (500, i.e. 48.59%), and a small number of book chapters (11, i.e. 1.07%).

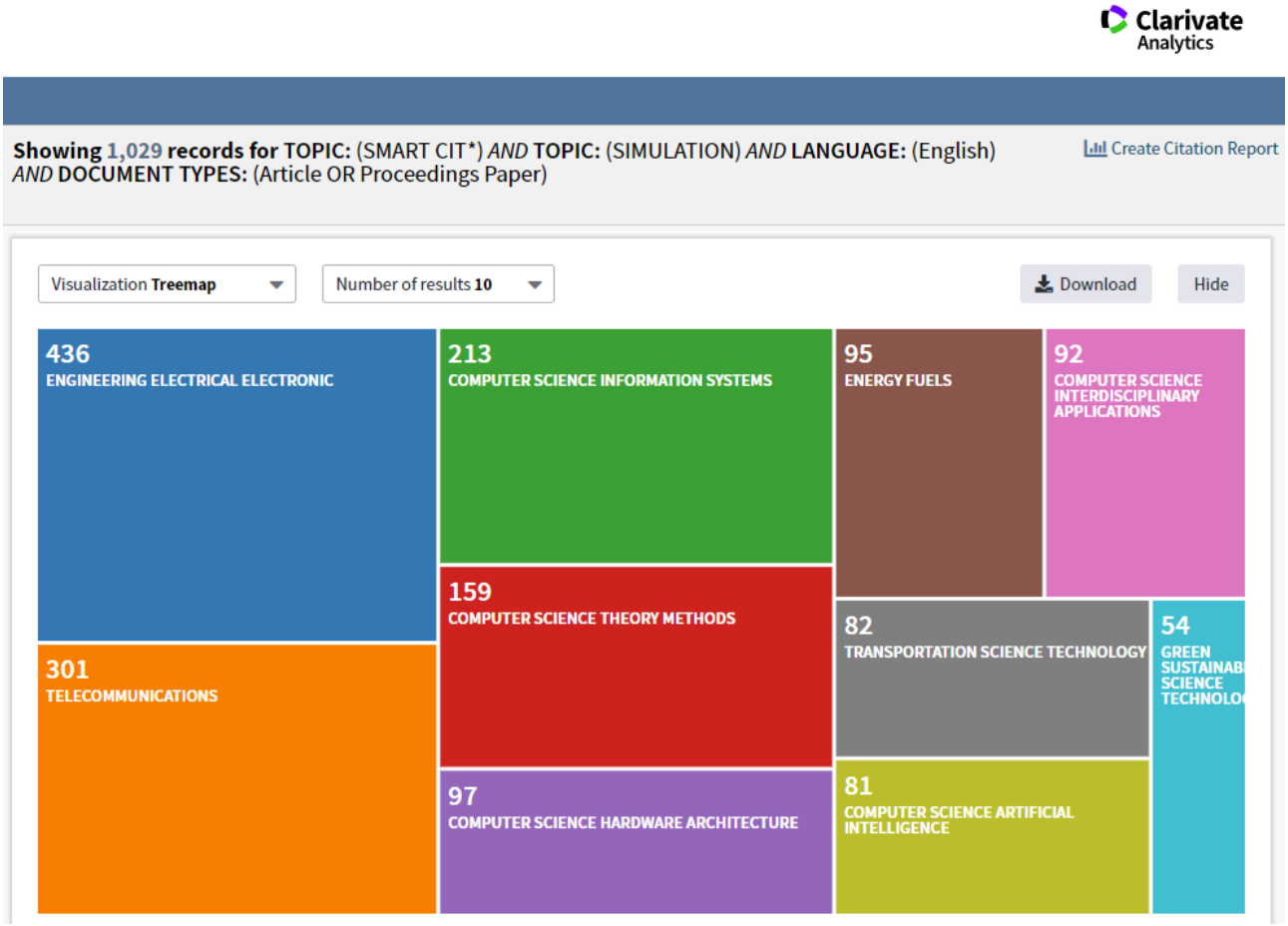


Figure 1. Visualisation of the results per Web of Science category

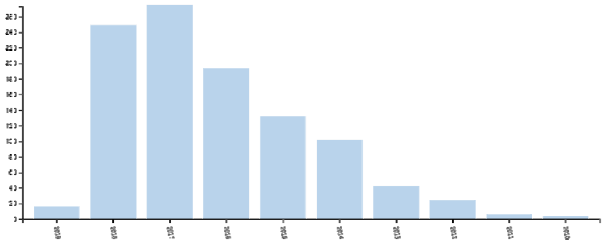


Figure 2. Visualisation of the results per year

With regards to publication years (Fig. 2), over half of the publications are from 2017 (273, i.e. 26.53%) and 2018 (247, i.e. 24%). The growth in the popularity of the two concepts as of 2010 can be observed as follows: in 2010 only 2 (0.19%) papers were published, in 2011 there were 4 (0.39%) papers, in 2012 there were 22 (2.14%) papers, in 2013 there were 40 (3.89%) papers, in 2014 there were 100 (9.72%) papers, in 2015 there were 130 (12.63%) papers, and in 2016 there were there were 192 (18.66%) papers that deal with both smart city and simulation concepts. In early January 2019 there are already 14 (1.36%) papers published on this topic.

B. Selection and Analysis of Publications by Keywords

The list was then refined based on the keywords of the publications. Specifically, only the papers that contained both 'smart cit*' and 'simulation' keywords were left. The reasoning behind this was that in many cases where authors of the paper did not list both keywords, they either focused on smart cities or simulations and simply identified one or the other concept in the paper as emerging or useful. This enabled us to shorten the list from 1.029 to 71 papers that were left for thorough evaluation.

C. Analysis of Abstracts and Intial Categorisation

Based on the analysis of the abstract the papers were categorized by an expert into one of the eight different categories of smart city applications (Engagement and community, Economic development and housing, Energy, Waste, Water, Mobility, Security, and Healthcare). In case a paper would refer to a specific technology, it was categorized as 'other'. In parallel, the papers were categorized by another expert based on the type of simulation modelling that was employed (discrete simulation, agent-based simulation, system dynamics) and in case the type was not made specific in the abstract, the paper was flagged as it needed the full text analysis. Based on the categorization it was clear that further inspection of 35 papers would be needed, as these were either categorized as clearly employing 'discrete simulation modelling' or the type of simulation had not been specified in the abstract.

D. Full Text Analysis of the Papers

After reading all 35 papers, it was evident that only 11 use discrete simulation modelling concepts or tools primarily (table 1). Other papers presented either only a vague smart city concept that simulation modelling can support (at the conceptual level) or were based on one of the two approaches - a hybrid one or an agent-based. As far as the domains are concerned, the simulations in Energy and Mobility are the most frequent ones. Simulation opens up new opportunities for optimizing and reducing the consumption of energy. Optimizations of energy consumption are targeted primarily towards the monitoring and control of energy generation. In terms of the mobility – the focus is on transit systems as it is possible to have an accurate view across the overall infrastructure, and timely information on the present as well as expected future events. Transformation of transit systems relies however on the use of dynamic analytical techniques in order to understand and process the insights from the real-world events.

TABLE I. CATEGORIZATION OF THE PAPERS BY TYPE OF SIMULATION

Smart city category	Discrete sim.	Agent-based sim.	Other	Total
1 - Engagement and community	1		2	3
2 - Economic development and housing	2	1	1	4
3 – Energy	1	1	8	10
6 – Mobility	7	1	7	15
7 – Security			2	2
8 – Healthcare			1	1
Total	**11**	**3**	**21**	**35**

IV. RESULTS: STUDIES ON SIMULATIONS FOR/RELATED TO SMART CITIES

As noted, only 11 studies used discrete simulation as a tool to support planning or optimization of a smart city related concept. These studies are listed in table 2. Only four out of eight categories are represented with at least one study. Most of the studies that employ discrete event simulation are from the Mobility category (7 in total). Overall, the studies concentrate on resource efficiency, e.g., considering energy or 'time' as a scarce resource.

TABLE II. STUDIES USING SIMULATIONS PER SMART CITY DOMAINS

Smart city category	Authors
1 - Engagement and community	Komarevtseva, 2017 [19]
2 - Economic development and Housing	D'Angelo et al., 2018 [9] Casas et al., 2017 [20]
3 – Energy	Lau, et al. 2014 [21]
6 – Mobility	Ahmad et al. 2018 [22] Ahmad et al, 2018 [23] Mihaita et al, 2018 [24] Galan-Garcia et al, 2014 [25] Latorre-Biel et al, 2017 [26] Mangiaracina et al, 2017 [27] Farkas & Lendak, 2015 [28]

In terms of the first category, and related to digital citizen services specifically, a recent study [19] developed a model that manages calculations to determine the readiness of cities to implement smart city technologies with the view to use it in the development of strategic programs based on indicators such as technological supply, innovative development of urban environment, intellectualization, financial stability, energy efficiency of urban environment, and the project potential.

Tackling the (economic) development of regions, in particular of "smart shires" that represent decentralized geographical spaces aiming to manage resources (natural, human, equipment, buildings and infrastructure) in a sustainable way, one of the studies [9] presents a related use case, at the same time bringing forward a strong case for better scalability with respect to traditional simulation using multilevel simulation approaches. On a micro-level, another study [20] presents a well-developed holistic approach to simulating typologies of smart homes thus aiming to optimize comfort and economic criteria.

Under energy category, a study [21] that proposes a new distributed traffic aware lighting scheme management network simulated and analyzed the new lightning scheme compared to the "standard" or always-on lightning networks demonstrating that significant energy savings could be achieved by using the proposed lightning system.

Finally, most of the reviewed studies could be categorized as referring to mobility. One of the studies [22] uses simulation to evaluate different scenarios to achieve desired traffic efficiency in a trust-based environment of vehicular ad-hoc network, a vital component of intelligent transportation systems technology by focusing on security of communication between dynamically connected vehicles and static road side units; the developed simulation-based model aims to evaluate the impact of various attacks in the described context. The model is later augmented [23] to simulate a wide range of trust models where their efficiency is

evaluated against different quality of service and security-related criteria, all of which can be used for planning smart cities and/or by car manufacturers. In another study [24] a simulation-based model was developed with the purpose of reducing congestion during peak hours followed by a multi-objective optimization by using evolutionary algorithms to develop the optimal traffic control plan in one specific use case. Similarly, simulation modelling has proven to be a flexible and easy tool to simulate traffic flow in a city using smart signals in one other study [25] as well. In another example [26], the use of simulation based on a Petri net model in a traffic scenario resulted in the development of a flexible decision support tool for many different decision makers; the model is developed with enough generality to be adapted to any city, town or road network, modelling the behavior of traffic with the level of detail of a mesoscopic simulation. Another mobility aspect, smart parking, was explored [27] using a simulation to reduce time spent by drivers and, consequently, to assess the economic and social benefits of the service. Similarly, crowd-sensing based applications in urban parking scenarios were modelled and analyzed [28] where citizens collect and share parking-related events demonstrating significant benefits, even while having a relatively low user base.

V. Discussion: Widening the Scope of the Review

Major contributions in the field are more likely to be in the leading journals. However, following the recommendations [18], an additional effort was made to investigate other published articles in the field apart from the ones that are indexed in Web of Science and presented already. Proceedings from the conferences with the focus on smart cities and simulations were searched in particular to present either a matching or contrasting view on the proliferation of simulations in this specific application area. Overall, the focus of the additional studies that were found correspond clearly to the ones presented in the previous section. A short overview of the studies is given hereinafter with the view on the already introduced smart city categories – energy, mobility and economic development and housing.

In the energy segment, examples of energy management are presented in terms of the reduction of energy consumption and cost minimization in buildings [16] and within and between houses [29]. Also, energy consumption of the electric vehicles was optimized using discrete simulation to avoid the prohibitive costs in the deployment of electric charging stations [30].

Again, most of the studies deal with an aspect of mobility. Simulation was successfully used to evaluate the effects of a peer-to-peer traffic information system, by using a realistic mobility model applied to a medium or large city [31] and to estimate the impact of vehicle allocation algorithms, battery charging methods, vehicle availability, system limitations and stress loads handling in an automated on-demand personal rapid transit system for intracity transportation needs [32]. Another study [33] presented a model that can forecast behavior within the large traffic scenarios that could be used to represent the traffic of the entire cities. Smart parking solutions are also studied frequently as a part of traffic management systems, especially in cities with high vehicle density; the simulation results [34] help minimize the wait time and increase the probability to successfully find a parking spot.

In terms of economic development, simulations were used to project the future state with regards to urbanization, since the models allow quick change and adjustment of parameters, and enable multiple what-if scenario experiments while planning - for example housing and public facilities. An example is SmartCitySIM [35], used to understand how government policy decisions can influence the transformation of cities, while focusing on the relation to habitat and installation of public equipment. Similarly, the effects of dynamic planning methods in waste management were evaluated using discrete event simulation that reflected in customer satisfaction, profits, and CO2 emissions optimization [36].

Clearly, there are some areas where simulation modelling is used frequently and successfully to develop and operationalize decision support tools in smart city contexts. Specifically, these are – energy (energy management for resource optimization) and mobility (traffic management / traffic control). Correspondingly, these are the areas where simulations were commonly used even before the increase in the popularity of smart city concepts due to availability of data sources and the need to test out different solutions to pressing problems (traffic congestion and so on). However, the new simulation models targeting smart city scenarios (presented in the paper) are enhanced by using data from multiple sources and resulting in useful analyses for smart city managers working in different domains. At the same time, there are little examples of modelling issues from waste, water, security and healthcare domains. As processes and sample services from these categories presented in part 2 are transforming into data-rich and smart environments, these are becoming more appropriate cases for simulation concepts. This gap represents an opportunity for contributions in these areas as it signifies novel application areas and potential for new results and different interpretations of common smart city issues.

VI. Conclusion

Reviewed studies demonstrated that the planning, design and analysis of smart city issues can be effectively supported by (discrete) simulation models. Drawing on relevant worldwide evidence-based studies on smart city services the paper presented a research effort to synthesize prior research and set our direction for future research.

It is important to acknowledge the complexity of assembling a review in this highly interdisciplinary field – this being an important limitation of the study. Another limitation is a relatively narrow focus on discrete simulation in particular.

However, in bring forth an overview of recent research and of promising scenarios in an emerging field followed by short analysis of implications for researchers that use simulations as an instrument, a relevant contribution can be identified. The use of (discrete) simulations in optimizing the implementation of plethora of smart city applications is still not an extensively-researched topic.

ACKNOWLEDGMENT

This work has been supported by the Croatian Science Foundation under the project UIP-2017-05-7625: User-oriented process (re)design and information systems modelling – a case of smart city services.

REFERENCES

[1] D. Bork, H.G. Fill, D. Karagiannis, E.T. Miron, N. Tantouris, M. Walch, "Conceptual Modelling for Smart Cities: A Teaching Case", Interaction Design and Architectures. 27. pp. 10-27, 2015.

[2] C. Manville, G. Cochrane, J. Cave, J. Millard, J. K. Pederson, R. K. Thaarup, A. Liebe, et al. Mapping Smart Cities in the EU: Study. Policy Department: Economic and Scientific Policy, 2014. DOI: 10.1017/CBO9781107415324.004.

[3] European Commission, "Digital agenda for Europe", Luxembourg: Publications Office of the European Union, 2014. DOI: 10.2775/41229.

[4] C. Snow, D.D. Håkonsson, B. Obel, "A Smart City is a Collaborative Community: Lessons from Smart Aarhus", California Management Review, vol. 59, no. 1, pp. 92-108, 2016.

[5] J. Van Den Bergh, S. Viaene, "Unveiling smart city implementation challenges: The case of Ghent", Information Polity, vol. 21, no. 1, pp. 5-19, 2016. DOI: 10.3233/IP-150370.

[6] H. Arasteh, V. Hosseinnezhad, V. Loia, A. Tommasetti, O. Troisi, M. Shafie-khah, P. Siano, "IoT-based Smart Cities: a Survey", 2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC), Florence, pp. 1-6, 2016.

[7] M. Ćukušić, M. Jadrić, T. Mijač, "Identifying challenges and priorities for developing Smart City initiatives and applications", unpublished.

[8] J. Woetzel, J. Remes, B. Boland, K. Lv, S. Sinha, G. Strube et al., "Smart cities: Digital solutions for a more livable future", McKinsey Global Institute, 2018. URL: https://mck.co/2Jq8ccn

[9] G. D'Angelo, S. Ferretti, V. Ghini, "Simulation of the Internet of Things," International Conference on High Performance Computing & Simulation (HPCS), Innsbruck, pp. 1-8, 2016. DOI: 10.1109/HPCSim.2016.7568309.

[10] A.A. Tako, S. Robinson, "The application of discrete event simulation and system dynamics in the logistics and supply chain contex", Decision Support Systems, vol. 52, no. 4, pp. 802-815, 2012. DOI: 10.1016/j.dss.2011.11.015.

[11] A. Borschev, The Big Book of Simulation Modeling, AnyLogic, 2013.

[12] M. Pidd, Computer simulation in management science, 5th ed., Chichester: Wiley, 2004.

[13] A.M. Law, "How to build valid and credible simulation models", Winter Simulation Conference, Monterey, USA, pp. 58-66, 2006.

[14] M. Jadrić, M. Ćukušić, A. Bralić, "Comparison of discrete event simulation tools in an academic environment", Croatian Operational Research Review, vol. 5, no. 2, pp. 203-219, 2014.

[15] D. Pavlić, M. Jadrić, M. Ćukušić, "Discrete simulation modeling of intelligent passenger boarding", 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1272-1277, 2018.

[16] A. Maatoug, G. Belalem, K. Mostefaoui, "Modeling and Simulation of Energy Management System for Smart City with the Formalism DEVS: Towards Reducing the Energy Consumption", International Journal of Computer Applications, vol. 90, no. 18, pp. 38-43, 2014.

[17] R.M. Fujimoto, N. Celik, H. Damgacioglu, M. Hunter, D. Jin, Y-J Son et al., "Dynamic data driven application systems for smart cities and urban infrastructures", Winter Simulation Conference (WSC), Washington, DC, pp. 1143-1157, 2016. DOI: 10.1109/WSC.2016.7822172

[18] J. Webster, R. Watson, "Analyzing the Past to Prepare for the Future: Writing a Literature Review". MIS Quarterly, vol. 26, 2002. DOI: 10.2307/4132319.

[19] O.O. Komarevtseva, "Forming model for dynamic measurement of cities' readiness to implement smart city technologies", 2nd International Ural Conference on Measurements (UralCon), Chelyabinsk, pp. 415-420, 2017. DOI: 10.1109/URALCON.2017.8120745.

[20] P.F. Casas, A.F. Casas, "Using Specification and Description Language for Life Cycle Assesment in Buildings", Sustainability, vol. 9, no. 6, 1004, 2017. DOI: 10.3390/su9061004

[21] S.P. Lau, G.V. Merrett, A.S. Weddell, N.M. White, "A traffic-aware street lighting scheme for Smart Cities using autonomous networked sensors", Computers & Electrical Engineering, vol. 45, pp. 192-207, 2015. DOI: 10.1016/j.compeleceng.2015.06.011.

[22] F. Ahmad, A. Adnane, V. Franqueira, F. Kurugollu, L. Liu, "Man-In-The-Middle Attacks in Vehicular Ad-Hoc Networks: Evaluating the Impact of Attackers' Strategies", Sensors, vol. 18, no. 11, 4040, 2018. DOI: 10.3390/s18114040.

[23] F. Ahmad, V.N.L. Franqueira, A. Adnane, "TEAM: A Trust Evaluation and Management Framework in Context-Enabled Vehicular Ad-Hoc Networks" IEEE Access, vol. 6, pp. 28643-28660, 2018. DOI: 10.1109/ACCESS.2018.2837887.

[24] A.S. Mihăiță, L. Dupont, M. Camargo, "Multi-objective traffic signal optimization using 3D mesoscopic simulation and evolutionary algorithms", Simulation Modelling Practice and Theory, vol. 86, pp. 120-138, 2018. DOI: 10.1016/j.simpat.2018.05.005.

[25] J.L. Galán-García, G. Aguilera-Venegas, P. Rodríguez-Cielos, "An accelerated-time simulation for traffic flow in a smart city", Journal of Computational and Applied Mathematics, vol. 270, pp. 557-563, 2014. DOI: 10.1016/j.cam.2013.11.020.

[26] J.I. Latorre-Biel, J. Faulin, E. Jiménez, A.A. Juan, "Simulation Model of Traffic in Smart Cities for Decision-Making Support: Case Study in Tudela (Navarre, Spain)", Lecture Notes in Computer Science, vol. 10268. Springer, 2017. DOI: 10.1007/978-3-319-59513-9_15.

[27] R. Mangiaracina, A. Tumino, G. Miragliotta, G. Salvadori, A. Perego, "Smart parking management in a smart city: Costs and benefits," International Conference on Service Operations and Logistics, and Informatics (SOLI), Bari, pp. 27-32, 2017. DOI: 10.1109/SOLI.2017.8120964.

[28] K. Farkas, I. Lendák, "Simulation environment for investigating crowd-sensing based urban parking," Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), Budapest, pp. 320-327, 2015. DOI: 10.1109/MTITS.2015.7223274.

[29] A. Molderink, M.G. Bosman, V. Bakker, J.L. Hurink, G.J. Smit, "Simulating the effect on the energy efficiency of smart grid technologies", Winter Sim Conference, pp. 1530-1541, 2009.

[30] M.T. Sebastiani, R. Lüders, K.V. Fonseca, "Allocation of charging stations in an electric vehicle network using simulation optimization", In Proceedings of the 2014 Winter Simulation Conference, pp. 1073-1083, 2014.

[31] M. Picone, M. Amoretti, F. Zanichelli, "Simulating smart cities with DEUS". 5th International ICST Conference on Simulation Tools and Techniques, pp. 172-177, 2012.

[32] K. Mueller, S.P. Sgouridis, "Simulation-based analysis of personal rapid transit systems: service and energy performance assessment of the Masdar City PRT case", Journal of advanced transportation, vol. 45, no. 4, pp. 252-270, 2011. DOI: 10.1002/atr.158.

[33] E.F.Z. Santana, N. Lago, F. Kon, D.S. Milojicic, "Interscsimulator: Large-scale traffic simulation in smart cities using erlang". In International Workshop on Multi-Agent Systems and Agent-Based Simulation, pp. 211-227. Springer, Cham, 2017.

[34] T.N. Pham, M.F. Tsai, D.B. Nguyen, C.R. Dow, D.J. Deng, "A cloud-based smart-parking system based on Internet-of-Things technologies", IEEE Access, vol. 3, pp. 1581-1591, 2015. DOI: 10.1109/ACCESS.2015.2477299.

[35] F. Barramou, M. Addou, "SmartCitySIM: Urban Simulator of Smart Cities", International Journal of Emerging Technology and Advanced Engineering, vol. 5, no. 1, pp. 10-16, 2015.

[36] M. Mes, "Using simulation to assess the opportunities of dynamic waste collection", In Bangsow S. (ed) Use cases of discrete event simulation, Springer, Berlin, Heidelberg, pp. 277-307, 2012. DOI: 10.1007/978-3-642-28777-0_13.

Traffic State Estimation Using Speed Profiles and Convolutional Neural Networks

L. Tišljarić*, T. Carić *, T. Erdelić *, M. Erdelić *,
* Faculty of Transport and Traffic Sciences/Department for Intelligent Transport Systems, University of Zagreb, Croatia
ltisljaric@fpz.hr

Abstract – **Determining the traffic state is one of the most attractive problems for experts in the field of Intelligent Transport Systems (ITS). In this paper, a deep learning model for determining the traffic state is presented. Model is based on Convolutional Neural Networks (CNN) and uses weekly speed profiles as input data. The proposed model consists of input and output layer with an addition to four convolutional layers, two pooling layers and two fully connected layers that are extracting important features and classifying intersections as congested or not congested. We analyze data and predict traffic state for the most relevant road segments in the City of Zagreb which is the capital and largest city in Croatia. Speed profiles from included road segments are represented as one traffic image and used to train CNN. In that way traffic state for all sequentially connected road segments is estimated. The proposed method achieves a classification accuracy of more than 90% on three analyzed types of road topologies. The results show that CNN trained with traffic images can be used as a tool for traffic state estimation.**

Keywords – traffic state estimation, convolutional neural networks, speed profiles, intelligent transport systems

I. INTRODUCTION

Traffic state estimation is commonly referred as a process of the determining traffic state parameters on one or many road segments using collected traffic data. Most often used traffic state parameters are traffic flow, density and speed, but other parameters can be used such as acceleration, spacing and other equivalent variables [1]. High quality estimation of traffic parameters has an important role in many ITS applications. Applications like ramp metering, speed limit control, toll collection and signal plan control at intersections require accurate traffic state estimation in order to reduce congestion and obtain fluent traffic flow.

In this paper we are using GPS speed data aggregated into speed profiles to estimate traffic state. Speed profile is a discrete function which represents the speed change over observed time. Useful information can be extracted from a speed profiles such as: the congested areas, occurrence and duration of rush hours, congestion propagation patterns etc. For that reason, historical speed profiles are chosen for the estimation of a traffic state on an urban road network. CNN was trained with speed profiles represented as images in order to predict the traffic state as congested or not congested. Therefore, traffic state estimation is modeled as a binary classification problem.

The remainder of this paper is organized as follows. In section 2 related work on recent developments related to the estimation and prediction of traffic state parameters based on the deep learning methods is presented. Section 3 describes the method for collecting and processing GPS data to extract speed profiles and represent them as traffic images. In section 4 method for traffic state estimation based on CNN is presented. In section 5 main research results are presented and analyzed. The conclusion of the paper is given in section 6.

II. RELATED WORK

According to [2] two approaches in traffic state estimation and prediction can be identified: non-parametric models and parametric models. Non-parametric models like k-nearest neighbors [3] and support vector regression [4] have been applied to predict traffic speed and traffic flow. The parametric models are represented with the autoregressive integrated moving average model (ARIMA) [5] such as seasonal ARIMA, Kohonen ARIMA and ARIMA with Kalman filter. The Bayesian network [6] and Markov chain models [7] are also applied for traffic state estimation and prediction tasks.

Speed profiles are common input data for the estimation and prediction of traffic state parameters. Authors in [8] are using speed profiles to estimate time dependent traffic states and use them in routing application. Highway capacity manual [9] classifies the level of service on road segments into six groups based on speed data. Classes ranges are estimated by comparing the measured speed value to the Free Flow Speed (FFS). In [10] authors are using speed profiles to predict traffic speed for highway operations.

Recently, a great number of deep learning models based on neural networks were developed for traffic state parameters estimation and prediction. Authors in [11] are using deep learning models to predict traffic flow. Deep learning traffic state prediction model that takes into account spatio-temporal relation is proposed in [12]. Deep learning model is applied using autoencoders as building blocks that represent traffic flow features for prediction. Authors in [2] propose a CNN based method to predict large-scale, network wide traffic speed. As input, the authors used speed profiles computed into two minutes intervals. Speed profiles were then represented as traffic images. Traffic images are constructed as a spatio temporal matrices. Then, traffic images are used to train CNN to

extract and predict traffic parameters. In [13] authors are using CNN and recurrent neural networks to predict and estimate traffic state. The authors classify traffic state as fluent, slow, congested and extremely congested.

III. DATA PROCESSING

For the computation of speed profiles, we used GPS data acquired from the vehicles equipped with the GPS tracking devices. Each record contains a time-stamp, geographical longitude and latitude, speed, heading etc. Due to the storage limitation most of the data are sampled in the following way: sampling rate of 100 m for vehicles in driving mode and every 5 min for turned off vehicles. Raw data are map matched to the road segments in a digital map based on the measured latitude, longitude and heading. Data which could not be matched to appropriate road-segment due to the GPS error caused by tunnels, high building concentration or other causes were filtered out [14], [15]. GPS data for the road network of Croatia were recorded during a five-year period between August 2009 and October 2014 by approximately 4200 tracked vehicles. The tracked vehicle fleet is versatile and consists mostly of delivery vehicles (vans, caddies, small trucks) and taxi cars. The historical tracked GPS data which consists of 6,55 billion GPS records were provided by the company Mireo Inc., as a part of SORDITO project [16].

A. Speed profiles

To construct a speed profile, speed data from GPS records are aggregated into 5 min intervals. In that way, a day is divided into 288 intervals with corresponding speed values. To cope with different speed limits and driving behavior on road segments, relative speed profiles are used. Speed is computed relative to the FFS on the road segment. FFS can be defined as the speed of a single vehicle on an empty road segment or as the speed limit on the observed road segment [3]. Every speed is represented with in interval [0, 100] relative to FFS on the road segment. For FFS value, maximal value of recorded speed was used. Fig. 1 shows a difference between the relative and absolute speed profile for a road segment with two typical rush hour periods during a day. In such way similar driving patterns on different road segments can be observed (grouped).

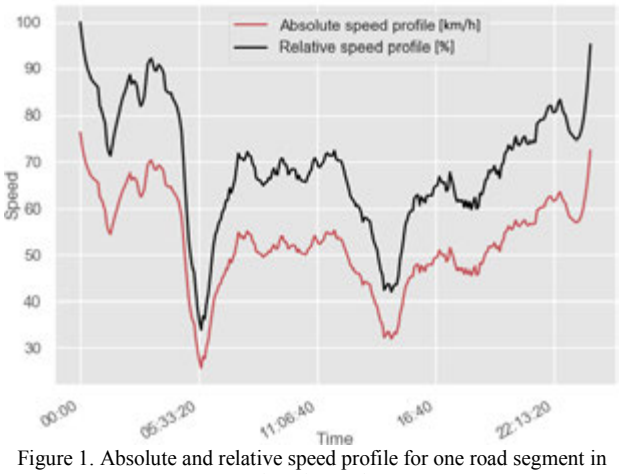


Figure 1. Absolute and relative speed profile for one road segment in city of Zagreb

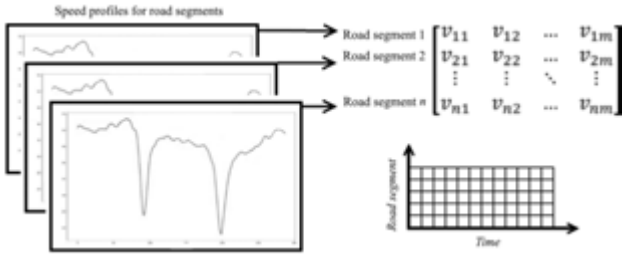


Figure 2. Speed profiles on the road segments represented as space-time matrices

B. Traffic images

To predict traffic speed, time and space dimensions should be jointly considered. We represent speed profiles for road segments in the road network as 2D matrices (Fig. 2). Two dimensions of the matrix are m and n where n represents the number of road segments under consideration (space dimension) and m represents the number of consecutive discrete time intervals (time dimension).

The time dimension is usually represented as intervals that extend throughout the day and are divided into 2 or 5 min intervals. In this paper, 5 min interval was used. Space-time matrix M is presented in (1), where v_{ij} is the average traffic speed recorded on road segment $i = 1, \dots, n$ in the time interval $j = 1, \dots, m$.

$$M = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1m} \\ v_{21} & v_{22} & \dots & v_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \dots & v_{nm} \end{bmatrix} \quad (1)$$

For every week of the year one traffic image was generated. To lower the deviation, summer months, July and August, are excluded from the model, because they significantly influence the results due to the different traffic flows caused by vacations [17]. Considered seasonal differences resulted in 47 weeks of data. Data are further divided into two groups: working days and weekend days. Working days speed profiles from Monday to Friday are different than the weekend speed profiles for Saturday and Sunday mostly due to the daily commuters. To train a binary classification model, we used speed profiles for working and weekend days as a train and test data for classifying road segments as congested and not congested, respectfully.

IV. MODEL

The CNN is a type of a feed forward Artificial Neural Network (ANN). CNN architecture makes the assumption that the input is in form of an image. In comparison to classical neural networks architecture, CNN reduces the number of parameters using convolutional layers which allows a network to be deeper. Common CNN consists of a combination of the three types of layers: convolutional, pooling and fully connected layers.

A. CNN layers

Convolution layers (CL) consists of one or more filters (kernels) which weights are adapted through the learning process. Every filter has a small width and height, but it extends through the full depth (all channels) of the input image. CL takes several feature maps as input and produces n_f feature maps as output, where n_f is a number of filters in the CL. Feature maps are then passed as input data to the next layer of the network.

Train data for CNN can be denoted as $\{x_i, y_i\}_{i=1}^{n_{tr}}$, where x_i is input data described as a matrix with three dimensions $x_i \in \mathbb{R}^{h \times w \times d}$, $h \times w \times d$ represents image height, width and depth, output data $y_i \in \mathbb{R}$ and n_{tr} number of training examples. Then, the convolution process can be described by (2), where fm are a feature maps obtained after convolution operation on image x_i with the set of weights w_j and bias b_j.

$$fm = Conv(x_i) = \sum_{j=1}^{n_f} w_j \times x_i + b_j \qquad (2)$$

Pooling layers are commonly applied after one or more convolution layers. Their main purpose is to downgrade an input image by reducing its spatial dimensions. The most common pooling filter size is 2x2, which reduces dimensions by half with one of pooling methods like pooling by maximal value or average pooling.

Fully connected layers are very similar to hidden layers in the multi-layer perceptron network. The main usage of fully connected layers is to learn weights to classify given image. In contrast to the convolution layer, a fully connected layer has only one dimension. To use it, the output of the last convolution or pooling layer has to be flattened in form of a vector.

B. CNN Hyperparameters

Regarding to architecture of the CNN, three important factors must be considered: (I) determining the hyperparameters of the CNN layers such as kernel size, pooling size, pooling method, activation functions, (II) depth of the network (number of filters used) and (III) number of layers in the network. For determining all mentioned hyperparameters there are no strict rules to follow. Every problem is different, and most of the hyperparameters have to be determined experimentally [2].

TABLE 1. HYPERPARAMETERS OF THE CNN LAYERS

Hyperparameter	***Value***
Kernel size	5x5
Pooling size	2x2
Pooling method	Maximal pooling
Activation function (convolutional layers)	Rectified linear unit
Activation function (classification layer)	Sigmoid

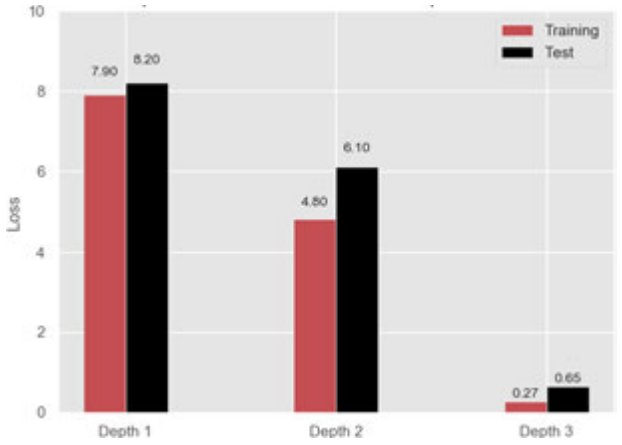


Figure 3. Results for different depths of the CNN

To determine CNN hyperparameters, well known CNN architectures can be referenced. First successful architecture was LeNet, based on the work of Yann LeCun in 1990's where 32x32 pixelated images are used as input. The LeNet has two convolution layers, two pooling layers followed by two fully connected layers [18]. Later on, deeper and larger CNN architectures were developed as a part of popular image classification competition called ImageNet Large Scale Visual Recognition Competition (ILSVRC). In 2010. AlexNet was introduced with the architecture based on five convolutional layers with filter sizes 11x11 and 3x3, three fully connected layers and three maximal pooling methods. For input, the network is using images with the size of 224x224 pixels [19]. VGGNet was introduced in 2014. This network is using 3x3 filters in 13 or 16 convolution layers and three fully connected layers [20]. Similar architecture was used in papers [2] for traffic state estimation and predictions. Table 1 shows chosen hyperparameters based on the AlexNet and VGGNet.

The number of the convolutional filters and the number of layers of the CNN were estimated experimentally. Structure of the CNN was tested on three different depth levels and different number of layers, presented in Table 2. Depth 1 represents only two fully connected layers that perform classification on flattened data. Depth 2 and 3 are using convolutional and pooling layers for feature extraction and then perform classification through fully connected layers. Every two convolutional layers were followed by a pooling layer. Layers were labeled with a notation $h \times w \times d$, where d is a number of filters in a convolution layer (depth), w represents filter width and h filter height. In the experiment, for every traffic state (work week and weekend) 47 images was generated, which resulted with 94 images in total. Each image size is 100x288 pixels. Where the values [0, 100] represents the relative values of the speed in a given interval, and 288 is the number of time intervals used to represent the speed profile. Data were further divided into train and test dataset. To keep the track of the train process and to detect possible overfitting, the 20% of the train dataset is used to validate the training process.

Fig. 3 shows the result of testing the depth of CNN. As expected, depth 1 gives the worst loss results as a model is

TABLE 2. TESTED STRUCTURES OF THE CNN CLASSIFICATION MODEL

	Structure
Depth 1	Two fully connected layers
Depth 2	2(5x5x64) convolutional layers 2x2 maximal pooling
Depth 3	2(5x5x128) convolutional layers 2x2 maximal pooling 2(5x5x64) convolutional layers 2x2 maximal pooling

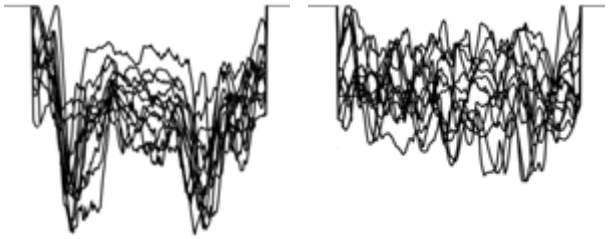
not able to learn features from the presented images. Adding the depth to CNN increases the model capacity and significantly reduces loss on train and test dataset. As a result, depth 3 with two convolution layers 5x5x128, 2x2 max pooling layer, two convolutional layers 5x5x64 and 2x2 max pooling layer gives the best results and is used for further experiments in this paper.

Table 3 represents the hyperparameters of depth 3 CNN used in this paper. The first layer indicates input image size which consists of three dimensions with size 100x288x3. The first dimension is image height, second image width and third dimension is number of image channels. Convolutional layers transform the input image into 128 and 64 channels, without changing the width and height. Pooling layers are downsampling the image to 50x144 and 25x72 pixels. After the last pooling layer with a dimension 25x72x64 image is flattened into a vector with dimension of 115200. Flattening is followed by two fully connected layers with a size of 128. Last layer represents the model output with a size of 1. The last layer in binary classification task gives the result r in $\{0, 1\}$ which shows the class affiliation.

To prevent the model from an overfitting, a dropout layer is applied. Dropout is a method that randomly drops units (neurons) and their weights from a neural network layer during training. This prevents units from co-adapting to feature detectors [21]. It is common to use a dropout layer after a fully connected layer in the CNN structure [22].

TABLE 3. HYPERPARAMETERS OF THE CNN MODEL

Layer number	*Type*	*Parameters*	*Dimension*	*Parameter scale*
Input	Input	-	100x288x3	-
1	Conv.	5x5x128	100x288x128	9728
2	Conv.	5x5x128	100x288x128	409728
3	Pool	2x2	50x144x128	0
4	Conv.	5x5x64	50x144x64	204864
5	Conv.	5x5x64	50x144x64	102464
6	Pool	2x2	25x72x64	0
7	Flat	-	115200	0
8	Full	-	128	14745728
9	Full	-	128	16512
Output	Output	-	1	129



a) Congested road segments b) Not congested road segments

Figure 4. Example images of train and test dataset

V. RESULTS

In this paper, traffic state estimation is considered as a binary classification problem. We classify traffic state as congested or not congested. Congested traffic is represented with weekly computed speed profiles when considering only working day and not congested traffic is represented with speed profiles for weekend days. After filtering, as described in section 3, traffic images were extracted for training and testing the CNN network. Fig 4 shows the example of train and test data for classifying congested and not congested traffic state on the road segments. Image a) represents congested road segments, it can be seen that traffic congestion appears in morning and afternoon rush hours between 7 am and 4 pm. Not congested road segments on image b) do not follow the same pattern.

In order to test the performance of the proposed CNN architecture, Long Short-Term Memory (LSTM) neural network with 128 nodes is chosen for comparison. LSTM is a successor of the recurrent neural network architecture and popular solution because it avoids vanishing gradient problem [23]. LSTM architecture is not developed to use images as an input data. In order to train and test the LSTM, data is represented as 2D matrix with dimensions $m \times n$ described with (1).

CNN model is compared with LSTM model on three different road topologies (Fig. 5): (a) intersection, (b) several consecutive road segments without turnings and (c) several consecutive road segments with turnings. All analyzed road segments have a one-way direction.

Data processing, training and testing of the proposed solution is developed using Python programming language with the addition of deep learning framework TensorFlow [24]. Performance of both networks is shown with Table 4. The results show that CNN architecture achieves better result in every tested topology with an accuracy above 90%.

TABLE 4. PERFORMANCE OF THE CNN AND LSTM

	Topology 1		*Topology 2*		*Topology 3*	
	Loss	Acc[%]	Loss	Acc[%]	Loss	Acc[%]
LSTM	0.71	89	0.83	82	0.69	89
CNN	0.65	92	0.60	93	0.67	90



a) Intersection



b) Sequentially connected road segments without turnings



c) Sequentially connected road segments with turnings

Figure 5. Different test road topologies

VI. CONCLUSION

In this paper deep learning model for traffic state estimation is presented. The proposed model is based on the CNN and traffic state estimation is represented as a binary classification problem as congested or not congested. Model is trained with weekly speed profiles presented as traffic images. The architecture of the model consists of an input and output layer with an addition of four convolutional layers followed by two pooling layers. For the pooling method, pooling with maximal values is chosen. Pooling layers are followed by two fully connected layers that are performing a classification task.

The method is tested on three types of road network topology: the intersection, sequentially connected road segments without turnings and sequentially connected road segments with turnings. Results are compared with the LSTM based neural network. CNN outperforms the LSTM in every test scenario with an accuracy above 90%.

Results are tested on road segments that have a one-way direction. Possible further research would be developing a model that considers two-way direction roads. A similar problem is researched in [25]. The model can be further extended by combining two or more methods like convolutional LSTM networks or diffusion convolutional RNN networks.

ACKNOWLEDGMENT

This research has been supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS).

REFERENCES

[1] F. Kessels, *Traffic Flow Modelling*. Springer International Publishing, 2019.

[2] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction," *Sensors (Switzerland)*, vol. 17, no. 4, pp. 1–16, 2017.

[3] F. G. Habtemichael and M. Cetin, "Short-term traffic flow rate forecasting based on identifying similar traffic patterns," *Transp. Res. Part C Emerg. Technol.*, vol. 66, pp. 61–78, 2016.

[4] Y. S. Jeong, Y. J. Byon, M. M. Castro-Neto, and S. M. Easa, "Supervised weighting-online learning algorithm for short-term traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1700–1707, 2013.

[5] M. S. Ahmed and A. R. Cook, "Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques," *Transp. Res. Rec.*, no. 722, pp. 1–9, 1979.

[6] J. Wang, W. Deng, and Y. Guo, "New Bayesian combination method for short-term traffic flow forecasting," *Transp. Res. Part C Emerg. Technol.*, vol. 43, pp. 79–94, 2014.

[7] Y. Qi and S. Ishak, "A Hidden Markov Model for short term prediction of traffic conditions on freeways," *Transp. Res. Part C Emerg. Technol.*, vol. 43, pp. 95–111, 2014.

[8] L. Rožić, J. Fosin, and T. Carić, "Solving the Time Dependent Vehicle Routing Problem Using Real-world Speed Profiles," *Cent. Eur. Conf. Inf. Intell. Syst.*, pp. 193–200, 2015.

[9] "Highway Capacity Manual 2010." Transportation Research Board, National Research Council, 2010.

[10] L. Linchao, T. Fratrović, Z. Jian, and R. Bin, "Traffic Speed Prediction for Highway Operations Based on a Symbolic Regression Algorithm," *Promet - Traffic Transp.*, vol. 29, pp. 433–441, 2017.

[11] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transp. Res. Part C Emerg. Technol.*, vol. 79, pp. 1–17, 2017.

[12] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Y. Wang, "Traffic Flow Prediction with Big Data: A Deep Learning Approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, 2015.

[13] X. Cheng, R. Zhang, J. Zhou, and W. Xu, "DeepTransport: Learning Spatial-Temporal Dependency for Traffic Condition Forecasting," 2017.

[14] T. Erdelić, S. Vrbančić, and L. Rožić, "A model of speed profiles for urban road networks using G-means clustering," *2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc.*, no. October 2014, pp. 1081–1086, 2015.

[15] Y. C. Zhang, X. Q. Zuo, L. T. Zhang, and Z. T. Chen, "Traffic congestion detection based on GPS floating-car data," *Procedia Eng.*, vol. 15, pp. 5541–5546, 2011.

[16] T. Carić, "SORDITO | Sustav za Optimizaciju Ruta u DInamičkom Transportnom Okruženju." [Online]. Available: http://www.fpz.unizg.hr/sordito/language/en/.

[17] R. Zochowska and K. Grzegorz, "ITS services packages as a tool for managing traffic congestion in cities," in *Intelligent Transportation Systems - Problems and Perspectives*, A. Sladkowski and W. Pamula, Eds. 2015, pp. 81–103.

[18] Y. LeCunn, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proc. IEEE*, 1998.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton., "Imagenet classification with deep convolutional neural networks. In Advances in neural information," *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105, 2012.

[20] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv 1409.1556.*, 2014.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[22] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," pp. 1–18, 2012.

[23] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[24] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015.

[25] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting," no. September 2018, 2017.

Stakeholder Support as Critical Success Factor in Adopting Big Data Technologies for Smart Cities

Jasmina Pivar* and Nikola Vlahović*
*University of Zagreb, Faculty of Economics & Business, Zagreb, Croatia
nvlahovic@efzg.hr, jpivar@efzg.hr

***Abstract* – Smart city is a city seeking to address public issues using innovative information and communication technology, including big data technology, supported by various stakeholders. By using big data technology, city management can get various insights, make conclusions about citizens and cities' infrastructure, and make decisions to improve the quality of life of a city. In context of adopting big data technology for smart cities, stakeholders play an important role. The main goal of this research is to detect which group of stakeholders is crucial for big data technology adoption in cities. The research is based on current literature overview and recommendations for cities regarding stakeholders, with the focus on citizens. In addition, a case study of city of Darmstadt, Germany which is the adopter of big data technology, will be described and analyzed to investigate some of the good practices regarding smart city projects. Based on the Importance-Performance Map Analysis, support from citizens and citizen organization for technology adoption initiatives are recognized as most important. However, their support is perceived very low by city management. Therefore, this paper aims to provide guidelines for city management regarding activities related to citizens as the crucial group of stakeholders for successful implementation of smart city initiatives.**

Keywords – smart city; big data technology; stakeholders; citizens; importance-performance map analysis

I. INTRODUCTION

Smart city is a city seeking to address public issues using recent information and communication technologies (ICT) supported by city management and other stakeholders. Big data technology has contributed to smart city development [1, 2] by enabling rapid generation of large amounts of various types of data, its acquisition, storage and fast analysis. City's infrastructure, objects and residents generate massive quantities of data related to urban living. Big data technology allows city management to monitor city infrastructure and to interact with the community. For instance, real-time monitoring systems upgraded with data acquisition sensors can collect various data from citizens and devices. These data sources may prove vital for increased efficiency, cost effectiveness, and increased customer satisfaction with city management services. Based on the results of big data analysis, city management can get insights, make conclusions about the urban environment, and make decisions to improve the quality of life in the city. In this way, useful content can be created for various stakeholders, including citizens, companies, visitors and others.

There are several challenges associated with the use of big data technology in smart cities. Human dynamics, privacy, security, volume, velocity, variety, veracity and value related to big data in a smart city and the respective solutions to address them were identified by [3]. Human dynamics in smart cities is related to stakeholders. In general, to become smarter, a city must not only apply a wide range of information and communication technology. It also should assure that stakeholders support its adoption. Only this way ICT, as well as big data technology, will enhance the innovation and approaches to problems in a city, encourage people to learn, increase overall knowledge and meet stakeholders' expectations [4 via 5]. In context of big data technology adoption for smart cities, stakeholders play an important role, especially citizens. Citizens should be informed and informed about new technologies so they can adapt it in a fast and efficient way. So, the focus should be on the human capital rather than believing ICT itself can develop cities [6].

Furthermore, the adoption of innovative information technologies, such as big data technologies, increases the social capital of cities, and vice versa. In [4], the author discusses the relationship between technology innovations in the smart city, social capital and participation possibilities to citizens. They recommended values, expectations, and perceptions of citizens to be engaged in the design and implementation of a smart city project. It is vital to assure that citizens, who are intended users of smart city solutions, can participate in the definition and the evaluation process during implementation of these solutions. Furthermore, citizens want active city management, accessible and reliable information, transparency and shared accountability [7]. Beyond considering citizens as being users, testers or consumers of technology, local administrations that can identify, nourish and integrate the emerging citizens' initiatives can contribute to the reinforcement of a smart city reality [3]. In [8] authors explore and develop the ambient urbanization concept in order to shed light on what happens in cities when people become more aware and attuned to smartness and ambience of a city's spaces. Finally, authors [9] describe "people-friendly city" in which cities are designed for people and where planning practitioners can understand citizens' needs, communicate with them and involve them in collaborative processes.

Additionally in [10] authors investigate the role of governance as a factor that influences the participation of citizens in smart cities projects. They used Amsterdam Smart City program as a case study and concluded that

governance characteristics play an essential role for better understanding and management of citizen participation. This is particularly important in smart cities where different stakeholders are interacting.

The perspective of the city as a service system where citizens are essential as co-creators of services is discussed by [11]. This study describes the implementation of the 'one-stop portal' concept. In [12] study showed that 47 leading European smart cities believe that collaborative governance is necessary to be implemented in public sector management. However, less than 50 per cent of them have created citizen participation platforms to promote citizen involvement in public affairs and only a few of them promote online public consultations, discussions and petitions. Also, there are differences in perceptions and e-participation tools used by sample smart cities.

To our best knowledge, no previous research has tested or described a group of stakeholders that are important for big data technology adoption in cities. Main goal of this research is to detect which group of stakeholders is crucial for big data technology adoption in cities. Sub-goals of this research are i) to develop general recommendations for cities regard stakeholders, with the focus on citizens and ii) to describe good practices in implementing smart city projects as well as stakeholders' roles based on City of Darmstadt, Germany which is the adopter of big data technology.

The paper consists of five sections. After the Introduction, there is a Methodology section in which the research hypothesis and data sample of the main research are presented. Additionally, Methodology section contains the description of data preparation for Importance-Performance Map Analysis and Stakeholders' Support indicators. The third section provides Results and discussion of IPMA. The fourth section is the Conclusion.

II. METHODOLOGY

A. Stakeholder Support in Adopting Big Data Technologies for Smart Cities

This research is a part of a broader research model that analyses the adoption of Big data technologies in cities. The endogenous construct of the main research is Big data adoption in cities – BDA.

The focus of this research is the Stakeholders' Support construct, one of the exogenous constructs of the initial model. Additional exogenous constructs of the model are: Absorptive Capacity – AC, City Management Support – CMS, Smart City Strategy – SCS, Technology Readiness – TR and Compatibility – COM.

For the purpose of this paper, additional exogenous constructs are used only as necessary elements for valid conduction of the Importance-Performance Map Analysis procedure and for comparison of their importance and performance with those of Stakeholders' Support.

The hypothesis of the main research related to Stakeholders' Support is stated as follows: *Stakeholders' support (SS) has a positive influence for big data technology adoption in cities (BDA).* Confirmation of that hypothesis is the main prerequisite of the Importance-Performance Map Analysis (IPMA), which is conducted for the purpose of this paper. Since the hypothesis is confirmed (p-value: 0.000, path coefficient 0.271) the main prerequisite for IPMA is fulfiled. Indicators of Stakeholders' Support construct are described in subsection D of this chapter.

B. Data Sample

In order to evaluate theoretical constructs, a survey is conducted with the following definition of the population and sample. The target population of this research are cities of the European Union - EU28 countries with more than 40 thousand inhabitants. According to data published by central national statistical services of EU28 countries, 1481 cities met the size criterion. The highest number of cities included in the target population were from Germany, Spain, France and Italy. The Republic of Malta was not included in this research since it does not have cities with more than 40 thousand inhabitants. The sampling frame was a list containing data on all of the 1481 cities, including potential respondents with contact data. Target respondents of the research were city mayors, managers of the city's IT departments, CIOs or managers of smart city initiatives. The questionnaire was sent by e-mail. All of the questions were in the form of a statement. Respondents were asked to state their level of agreement with the statements by using a 7-point Likert scale.

After data collection, questionnaires with inconsistent responses or less than 5% of missing values per indicator were excluded from further analysis. Mean replacement option was used to treat missing values. Responses from 94 cities were used to evaluate the research model.

C. Data Preparation for The Importance – Performance Map Analysis

Partial Least Squares – Structural Equation Modelling (PLS-SEM) is used to evaluate measurement and structural model. Partial Least Squares – Prediction Oriented Segmentation (PLS-POS) and Finite Mixture Partial Least Squares (FIMIX-PLS) procedures are used to i) uncover unobserved heterogeneity in the research PLS path model, and ii) to detect and estimate specific segments of the cities.

Two segments are detected: Segment 1 that contains 67 cities and Segment 2 that contains 27 cities. PLS-SEM procedure is used to evaluate the model based on Segment 1 data, since Segment, 2 data were insufficient for valid estimation of the research model. Results of procedures mentioned above are the basis for the Importance-Performance Map Analysis (IPMA). All of the procedures are conducted using SmartPL3 software. COM construct is excluded from IPMA because it is shown as insignificant in adoption of big data technology adoption for cities.

D. Stakeholders' Support Indicators & Importance-Performance Map Analysis

Stakeholders Support in the context of big data technology adoption for cities refers to stakeholders that (1) provide technology and technical solutions related to

technology adoption in the city, (2) incorporate technology into their processes or (3) facilitate technological and policy frameworks related to technology adoption. Construct Stakeholders' Support (SS) contains four indicators. For the model estimation, their values are calculated based on the survey questions related to support from the groups of stakeholders as shown in Table 1. Indicator SS1 refers to stakeholders which enable technology and technical solutions related to technology adoption in a city. Such stakeholders are called technology enablers. Indicator SS2 and SS3 are related to technology drivers - stakeholders who are interested in technology adoption initiatives in a city. Finally, the last group of stakeholders are called framework enablers, institutionalised stakeholders who support technology adoption to enhance a city.

TABLE I. DESCRIPTION OF STAKEHOLDERS' SUPPORT CONSTRUCT'S INDICATORS

Construct SS Indicators	***Support from [13]:***
SS1	ICT companies, city services companies, utility providers
SS2	national and regional governments
SS3	citizens and citizen organisations
SS4	academia, research organizations, specialized bodies; standardization bodies; international, regional, multilateral organizations and industry organizations

Source: Authors

The IPMA is the extension of the standard PLS-SEM analysis. It allows comparison of the total effects, the exogenous constructs' and indicators' importance in predicting a specific endogenous construct, with their average latent variable scores indicating their performance [14].

Purpose of IPMA is to identify exogenous constructs and indicators that have relatively high importance for predicting the endogenous construct, but also have a relatively low performance so that improvements can be implemented. Therefore, we can provide recommendations for improvements regarding aspects of such constructs based on the interpretation of the importance-performance map [14].

For example, improvements can be related to specific activities regard stakeholders who can then positively influence the adoption of big data technologies in cities. For the purpose of this paper, IPMA analysis is used to detect if citizens as stakeholders are crucial for big data technology adoption in cities.

Furthermore, IPMA results serve as the basis for the development of general recommendations for cities regarding stakeholders, which is the first sub-goal of this paper. As described in [14], Importance-Performance maps have x-axis, which represents importance, and y-axis showing a performance of a construct or an indicator. Additionally, the Importance-Performance Map contains two additional lines, a vertical line that represents average importance value and a horizontal line that represents average performance value. When it comes to interpretation of Importance-Performance Maps, we focus on constructs and indicators in the lower right area of the map where importance of the construct is above average and performance of the construct is below average. Such constructs and indicators have the highest potential to achieve improvement. Therefore, it is safe to conclude that IPMA guides the prioritisation of managerial or policy activities of the aspects underlying Stakeholders' Support construct.

III. RESULTS AND DISCUSSION

IPMA procedure was run using SmartPLS 3 software. Table II. Contains standardised and unstandardized importance and importance values for constructs SS, AC, CMS, TR and SCS based on the described data sample. These values are necessary for the creation of an Importance-Performance Map of constructs. Average values of performance and importance were calculated (Table II.). These values are used to draw horizontal and vertical lines on the Importance-Performance maps (Fig.1 and Fig. 2).

TABLE II. IMPORTANCE AND IMPORTANCE VALUES FOR CONSTRUCTS, STANDARDIZED AND UNSTANDARDIZED EFFECTS - EU CITIES (SEGMENT, N=67)

Constr-uct	**Importance**		**Performance**	
	Standar-dized	***Unstandar-dized***	***Standar-dized***	***Unstandar-dized***
AC	0.24	0.22	64.03	64.03
SS	0.40	0.27	63.47	63.47
CMS	0.20	0.18	65.73	65.73
SCS	0.32	0.28	72.20	72.20
TR	0.19	0.19	61.83	61.83
Average	0.20	0.17	64.16	64.16

Source: Authors, using SmartPLS3 and Excel software

Figure 1. shows unstandardized effects and Figure 2. shows standardised effects of constructs on Big Data Technology adoption in cities. They both indicate that the Existence of Smart City Strategy (SCS) has the highest performance in regards to Big Data Technology adoption in cities. However, Stakeholders' Support (SS) and Absorptive Capacity (AC) constructs have relatively low performance, but high importance for Big Data Technology adoption. Therefore, maps suggest that SS, followed by AC have high potential for improvements and, consequently, greater positive influence for Big Data Technology adoption in cities.

This means that stakeholders, in general, can significantly support the adoption of big data technology in cities. However, at the construct level of analysis it is not evident which groups of stakeholders are specially important, technology enablers, technology drivers framework enablers, or all of them. To determine the importance of these factors IMPA at indicator level is conducted to answer this question.

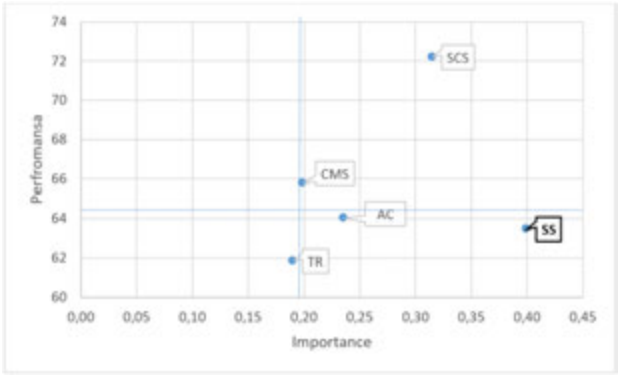


Figure 1. Importance-Performance Map of constructs, standardised effects - EU cities (segment, N=64) (Source: Authors, using SmartPLS3 and Excel software)

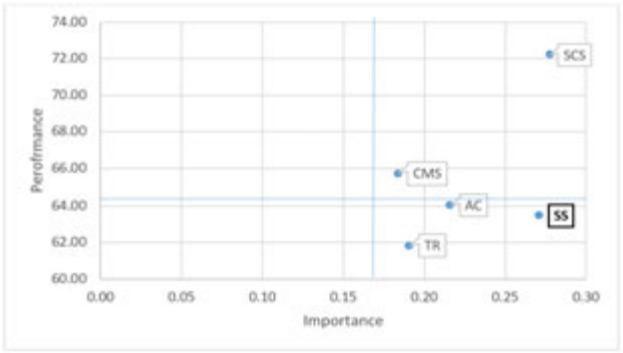


Figure 2. Importance-Performance Map of indicators, unstandardized and standardised effects - EU cities (cluster, N=67) (Source: Authors, using SmartPLS3 and Excel software)

As stated previously, IPMA at indicator level is conducted to determine importance of different groups of stakeholders. Standardized and unstandardized importance and performance values are calculated for indicators of constructs SS, AC, CMS, TR and SCS based on collected from 67 EU28 cities (Table III.)

TABLE III. IMPORTANCE AND IMPORTANCE VALUES FOR CONSTRUCTS, STANDARDIZED AND UNSTANDARDIZED EFFECTS - EU CITIES (SEGMENT, N=67)

Indicator	**Importance**		**Performance**	
	Unstandardized	***Standardized***	***Unstandardized***	***Standardized***
AC1	0.001	0.001	66.92	66.92
AC4	0.137	0.139	61.19	61.19
AC5	0.103	0.104	67.91	67.91
SS1	0.124	0.100	66.42	66.42
SS2	0.075	0.069	65.17	65.17
SS3	0.107	0.097	60.20	60.20
SS4	0.094	0.077	61.94	61.94
CMS1	0.050	0.050	72.89	72.89
CMS2	0.054	0.052	68.66	68.66
CMS3	0.052	0.054	58.96	58.96
CMS4	0.043	0.051	61.94	61.94
SCS1	0.074	0.075	74.88	74.88
SCS2	0.077	0.075	72.14	72.14
SCS3	0.084	0.082	71.89	71.89
SCS4	0.081	0.082	70.15	70.15
TR1	0.059	0.079	55.97	55.97
TR2	0.061	0.073	61.44	61.44
TR4_5	0.070	0.072	67.16	67.16
Average	0.052	0.051	64.56	64.56

Source: authors

Figure 3. shows unstandardized effects and Figure 4. shows standardised effects of all of the indicators on Big Data Technology adoption in cities. They both indicate that ICT companies, city services companies and utility providers have the highest importance of all the stakeholders. They also have good performance which is above average. This indicates that respondents from the analysed cities are satisfied with technology enablers' support regard big data technology adoption, and are perceived as interested in technology adoption initiatives in the city. Specifically, ICT companies research, innovate and provide the standard, compatible, scalable ICT infrastructure and technical solutions that support and integrate city services. City services companies, in some cases, create new service that covers a new or an uncovered city need. Finally, utility providers are responsible for the deployment of some smart city features, such as smart grid and smart water management.

National and regional governments, which are related to the sub-group of stakeholders that are interested in technology adoption initiatives, so-called technology drivers, are perceived as less important in comparison to other stakeholders. However, the performance of these stakeholders is above average. This means that, in general, national and regional governments promote and manage technology adoption initiatives, as well as they define policies and legal frameworks for such initiatives.

Furthermore, citizens and citizen organizations are identified as those of the highest importance of all of the stakeholders and at the same time having the lowest performance. This means that support from citizens and citizens' organizations are recognized as most important group of stakeholders for big data technology adoption initiatives. Their support, on the other hand, is perceived very low. This means that city government should perform practical and managerial activities in order to increase the support of citizens and citizen organisation for big data technology adoption.

Finally, academia, research organizations, specialized bodies, standardization bodies, international, regional, multilateral organizations and industry organizations have high importance and below average performance. These stakeholders are framework enablers. Specifically, this means that academia, research organisations and specialised bodies should somehow be encouraged to drive and support the research and innovation in fields related to technology adoption initiatives. Also, international, regional, multilateral organisations and industry associations should provide funds and make promotion programs to drive technology adoption initiatives. Standardisation bodies should lay the framework and define standards for technology adoption and smart cities. This way big data technology adoption in cities can be significantelly increased.

In order to fulfil the second sub-goal of this paper, good practices regarding smart city projects, and stakeholders are described based on the City of Darmstadt in Germany that is the adopter of big data technology. City od Darmstadt is an example of how important it is to include various stakeholders, especially citizens into the adoption process of various smart city initiatives.

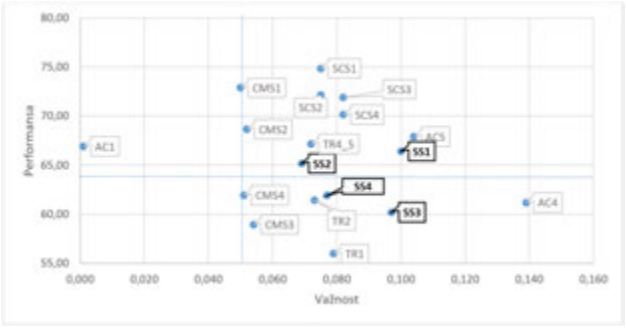


Figure 3. Importance-Performance Map of constructs, unstandardized effects - EU cities (segment, N=67) (Source: Authors, using SmartPLS3 and Excel software)

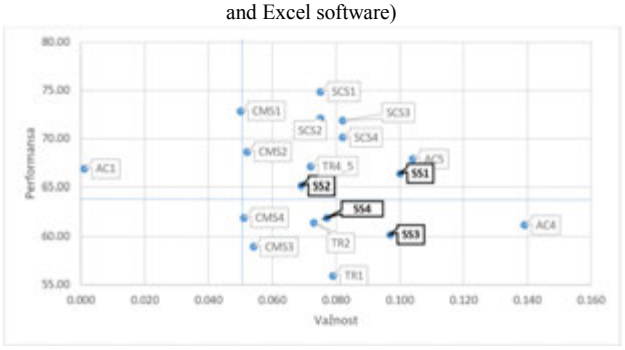


Figure 4. Importance-Performance Map of constructs, unstandardized effects - EU cities (segment, N=67) (Source: Authors, using SmartPLS3 and Excel software)

Smart City does not have to be a big city. City of Darmstadt, Germany with its 161 thousand inhabitants is one of these examples. Darmstadt known as "Science City" is supposed to become a digital model city. It has been called "digital city" since it won a competition sponsored by Bitkom and the Association of Towns and Municipalities. The city management has taken the advantage of state and private funding and used it to build digital infrastructure.

The city is equipped with a LoRaWAN network that is the prerequisite for Internet of Things.

Darmstadt has implemented smart traffic management called The UI! Traffic system where the traffic flow is monitored in real time and the traffic flow control is automated. Citizens can see the current level of traffic online (https://darmstadt.ui-traffic.de/). Darmstadt launched a connected parking app that helps citizens and visitors find and reserve parking spaces. They also experiment with autonomous buses and semi-autonomous trams. The city plans to implement intelligent street and waste bins. By 2025, they plan to implement car sharing options depending on demand traffic in order to enhance the mobility of the citizens. Furthermore, citizens will be able to use digital assistance in emergencies.

Darmstadt is considered a strong IT location with various stakeholders involved in ICT initiatives. For example, communications companies, research institutes and many university graduates in mathematics, computer science, science and technology, as well as citizens are all involved in smart city projects. When it comes to citizens, the city clearly states and informs their citizens about plans. Citizen participation has its digital participation platform. For example, "DA ist was!", is the defect detection application through which citizens can report visible deficiencies in the city, and can also track the processing status.

There is also the Open Data for Darmstadt project (http://www.offenesdarmstadt.de/). The association "Open Darmstadt" connects city management and administration, politics, research institutions, companies and citizens. Their role is to advise the administration on topics related to OpenData and IT, give recommendations on published data, review the resulting projects and promote them.

IV. CONCLUSION

Main purpose of this research is to detect which group of stakeholders is crucial for big data technology adoption in cities. Additionally, the research aims to develop general recommendations for city management regarding its stakeholders with special focus on citizens. Recommendations provided, are based on empirical research and they describe good practices in implementing smart city projects and stakeholders' roles.

Research results showed stakeholders, in general, can significantly support the adoption of big data technology in cities. Furthermore, citizens and citizens' organizations are identified as those of the highest importance of all of the stakeholders while at the same time having the lowest performance. This means that city government should perform practical and managerial activities in order to increase the support of citizens and citizen organization for big data technology adoption. City of Darmstadt, Germany serves as an example of how important it is to include various stakeholders, especially citizens into the adoption process of various smart city initiatives.

Limitation of the research is that some of the acquired data was not included in the Importance-Performance Map analysis. Cluster analysis was used to detect this data which was inadequate for valid evaluation of the data model thus fulfilling the prerequisite for the IPMA.

Future research should include the following. Firstly, to detect what is the role of City Management Support, Absorptive Capacity, Smart City Strategy, and Technology Readiness, and whether they are crucial for big data technology adoption in cities. Secondly, to develop general recommendations for cities based on results of IPMA for these factors.

REFERENCES

[1] Hashem, I. A. T. et al., "The role of big data in smart city", International Journal of Information Management, vol. 36, issue 5, pp.748–758., 2016.

[2] Kitchin, R., "The real-time city? Big data and smart urbanism", GeoJournal, vol. 79, issue 1, pp.1–14., 2014.

[3] Capdevila, I. and Zarlenga, M. I. Zarlenga, "Smart city or smart citizens? The Barcelona case", Journal of Strategy and Management, vol. 8, issue: 3, pp.266-282, 2015.

[4] Albuquerque, C. M., "Cities Really Smart and Inclusive: Possibilities and Limits for Social Inclusion and Participation", In I. Management Association (Ed.), E-Planning and Collaboration: Concepts, Methodologies, Tools, and Applications, pp. 1426-1445, Hershey, PA: IGI Global., 2018.

[5] Komninos, N., Intelligent Cities. London: Spon., 2002.

[6] Hollands, R. G., "Will the real Smart City please stand up?", City: Analysis of Urban Trends, Culture, Theory. Policy and Action, vol. 13, issue 3, pp. 303–320., 2008.

[7] Deakin, M., "Smart Cities: Governing, Modelling and Analysing the Transition",Taylor and Francis Group, 2013.

[8] McKenna, H. P., "Urbanizing the Ambient: Why People Matter So Much in Smart Cities", In I. Management Association (Ed.), Smart Technologies: Breakthroughs in Research and Practice, pp. 527-549), Hershey, PA: IGI Global, 2018.

[9] Melis, G., Masala, E., and Tabasso, M., "From the Smart City to the People-Friendly City: Usability of Tools and Data in Urban Planning". In I. Management Association (Ed.), E-Planning and Collaboration: Concepts, Methodologies, Tools, and Applications, pp. 679-697, Hershey, PA: IGI Global, 2018.

[10] Capra, C. F., "The Smart City and Its Citizens: Governance and Citizen Participation in Amsterdam Smart City", In I. Management Association (Ed.), Smart Cities and Smart Spaces: Concepts, Methodologies, Tools, and Applications, pp. 1407-1427, Hershey, PA: IGI Global, 2018.

[11] Peters, C., Korthaus, A., and Kohlborn, T., "Smart City Portals for Public Service Delivery: Insights From a Comparative Study", In I. Management Association (Ed.), Smart Cities and Smart Spaces: Concepts, Methodologies, Tools, and Applications, pp. 578-598, Hershey, PA: IGI Global, 2019.

[12] Rodríguez Bolívar, M.P., "Creative citizenship: the new wave for collaborative environments in smart cities", Academia Revista Latinoamericana de Administración, vol. 31, issue: 1, pp. 277-302, 2017.

[13] ITU-T Focus Group on Smart Sustainable Cities, "Setting the stage for stakeholders' engagement in smart sustainable cities", Available at: :http://www.itu.int/en/ITU-T/focusgroups/ssc/Pages/default.aspx [10.prosinca, 2015.]

[14] Hair, J.F. jr., Sarstedt, M., Ringle, C.M., Gudergan,S.P., "Advanced issues in partial least squares structural equation modeling", SAGE Publications, Inc., 201

Enhancing Occupants Comfort and Well-being through a Smart Office setup

Ankica Barišić*‡, Vasco Amaral*, Moharram Challenger†
*Universidade NOVA de Lisboa, NOVA LINCS, DI, FCT, Lisboa, Portugal
†Electronics and ICT Department, University of Antwerp, Belgium
‡Center for Informatics and Computing Science, Ruđer Bošković Institute, Zagreb, Croatia
Email: a.barisic@campus.fct.unl.pt, vma@fct.unl.pt, moharram.challenger@uantwerpen.be

***Abstract*—A smart office, a particular case of Building Automation, represents a physical environment enriched with sensing, actuation, communication and computation capabilities aiming at acquiring and exploiting knowledge about the environment and its inhabitants. A fundamental goal of smart offices is to ensure a comfortable working environment as well as adapting to the circumstances related to optimizing specific variables which indicate the safety and energy efficiency of the office, as well as the health and productivity of its occupants.**

In this position paper, we illustrate a representative system by exploring emerging technologies, which presents a unique set of challenges and opportunities to research a new generation of software control systems, by supporting non-intrusive interfaces, that put the occupant's at the center of the system's concerns, by focusing on comfort and well-being while still leveraging the occupants' behaviour.

I. Introduction

Smart buildings can be seen as a category of Cyber-Physical System (CPS) where the smart grid, the external communication networks with heterogeneous sensors and actuators, as well as its environment, interact by exchanging the energy and information while providing comfort services to humans. An effort is being made to establishing a reliable and sustainable technology for deploying green and zero-energy buildings [20]. Moreover, the buildings are being instrumented to automate and show intelligent behaviour.

While there are many projects and studies related to energy efficiency, the methods which consider assurance of the occupant's comfort and well-being are still in their early phase. The existing methods which take into consideration socio-technical aspects can be seen as too intrusive, as they ask the occupants for a constant input or feedback to adapt, which may be time-consuming and disturbing. On another hand, these methods often adopt the system based on post-occupancy data and predetermined parameters of energy exchanges, rather than acting on real-time situations and realistic sensing data (e.g. [19]). There are ample opportunities for improvements of efficient control through continuous evolution of real-life CPSs [9].

The current state-of-art of technology for building automation is a rule-based control which for instance use the current ambient temperature adjusted dynamically after every measurement to define set points [14]. Lack of this approach is that it is not able to dynamically respond to changing conditions, such as updates in temperature predictions, sunshine intensity or the assumptions on the number of occupants and their behaviours. Some researches suggest adopting predictive control model [12] which is supposed to better performance than the current control approach. Target expectations are better energy efficiency of the smart environment and enhancement of the occupant's comfort by incorporating the disturbance predictions (e.g. occupancy profiles, changes in the outside environment) and deployment of distributed control procedures with adjustable parametrisation of the distributed agent-based controllers.

In the context of issues related to smart buildings, we focus on the smart office, which represents a more specific and realistic testing environment that we can control and where is possible to isolate relevant parameters for experimentation. In the case of office setup, the occupant's primary goal is productivity related to their work tasks. Their productivity can be influenced by the number of factors, including their comfort, health and safety. The real-time feedback regarding the occupants can be collected through wearable and office sensors. Facility manager commonly provides new rules to a decision system which support the control of office behaviour by optimising the trade-offs between the energy savings and the occupant's comfort.

The rest of this paper is organised as follows: Section II summarises the regulations related to the offices, workplaces, their occupants and workers. In Section III, the principles and terms used in smart offices are defined. The proposed smart office setup is detailed in Section IV. The challenges are discussed in Section V. Finally, the paper is concluded in Section VI.

II. Occupational safety and health at workplace

According to the European Union (EU) regulations[1], to protect the workers from work-related accidents and diseases, the European Commission has adopted EU Occupational Safety and Health (OSH)[2] Strategic Framework 2014-2020 [10].

This framework identifies major challenges related to health and safety at work by calling for the implementation of existing health and safety rules by placing effective and efficient risk prevention strategies by tackling new and emerging risks without neglecting existing risks.

[1]European Union Law: eur-lex.europa.eu
[2]OSH Wiki: oshwiki.eu

While well-being at work and working conditions have been improved by the introduction of new technologies and innovations in work organisation, effective prevention of work-related diseases requires anticipating potential adverse effects of new technologies on workers' health and safety. These technologies lead to new products and processes, which need to be sufficiently tested to ensure that they do not represent significant hazards for occupants and workers. Some of emerging risks are linked to the development of bio-technologies and green technologies.

Changes in work organisations introduced by development of information technology, especially ones permitting constant connectivity, yield new flexible and interactive work processes solutions. According to recent Euro-barometer survey reported in the OSH Strategic Framework, workers consider stress as one of the leading occupational risks (53%), followed by ergonomic risks (repetitive movements or tiring or painful positions) (28%) and lifting carrying or moving loads on a daily basis (24%). This calls for specific attention to address the impact of changes in work organisation in terms of physical and mental health. One of the critical strategic objectives presented by the OSH Strategic Framework is *improving statistical data collection to have better evidence and developing monitoring tools*. Although, since 2013, the evidence-based policy is implemented to collect work-related accidents data, still remain to collect timely, reliable and comparable statistical data on occupational exposures and work-related ill-health, as well to analyse the costs and benefits in the area of OSH.

A risk assessment regarding OSH at work, as well implementation of prevention measures, must be performed by employers every time when a change is introduced in the workplace (e.g. new work equipment or procedure) or at least every year. The safety control measures to be implemented should be based on updated technical and/or organisational knowledge, and best practices. Prevention measures application aims to reduce the occupational disease occurrence and/or probability of work accident. It is expected that engineering or technical measures act directly on the risk source by removing the risks during the workplace design phase. Another way is to use administrative or organisational measures which promote proper attitudes and behaviours, as well as safety culture, by raising awareness through training, information sharing, and which implement appropriate working procedures and supervision, management and proactive monitoring, routine maintenance and housekeeping procedures and implementation of protection measures [10].

Employers are obligated to meet minimum requirements for workplaces predefined by EU commission. In a case of the smart office it is possible to automatise control over requirements referring to:

- Fire detection and fire fighting
- Ventilation of enclosed workplaces
- Room temperature
- Room lightning (natural and artificial lighting)
- Windows and skylights
- Doors and gates

III. Principles and Methodology

This section describes the terms used in this study for the proposed smart offices. Also, the general software and system techniques which can be used in the development of the proposed smart offices are defined. Finally, the main modules and high-level architecture are presented in this section.

The main perspectives of smart and sustainable office are:

- Well use of resources: Controlling the energy consumption by adapting to occupants presence, market trends and the outside environment, e.g. temperature, sunlight exposure, wind, humidity.
- Occupants Comfort: Supporting the occupants' work habits and health needs, e.g. considering if the person is unusually cold/warm, supporting notifications regarding drinking and eating needs, managing an appropriate light/noise level depending on the work tasks.
- Emergency and Resilience: Monitoring emergencies such as fire and gas leak [13], and help people to evacuate the building in the shortest time possible.

To address the real-time realisation of these perspectives, we suggest the following techniques be exploited in the development of a smart office are: modelling techniques, experimental software engineering, and dew computing.

Modelling techniques: Supporting the design and modelling of the system with domain-specific languages [6] and data-flow modelling [13] which increase the level of abstraction and paves the way for tackling complexity. Modelling techniques such as metamodelling [8] and generative languages [18] are used for the development of different IoT systems. Also, there is an example of smart office design and implementation by using a smart environmental metamodel as a framework for designing smart cyber-physical environments [5]. A model can present functional component interactions and offer and used services, along with a data meta-model that supplies the description of the properties and relationships of the data sources involved in the environment.

Experimental software engineering: To perform validation of design decisions or the evaluation of the smart office it is necessary to set up an experiment [4]. It is valuable to leverage occupants as sensors to quickly capture indoor environmental quality or conditions in a work environment. To understand the occupant's variables such as comfort, satisfaction, health, and performance, it is essential to measure the environmental conditions across all variables and capture records of critical workplace attributes that define the physical environment. As an example we have an integrated approach for building performance evaluation which match post-occupancy evaluation subjective tools by metrics [17], and is connected to database of over 1600 workstations with statistically significant findings about the measured and user-perceived quality of the indoor environment, as well as the technical attributes of building systems that contribute to successful, high performance buildings.

Dew computing: Using the transceiver devices the system can collect data [19] and identify the presence of the occupants [15] and their preferences with minimum disturbance e.g.

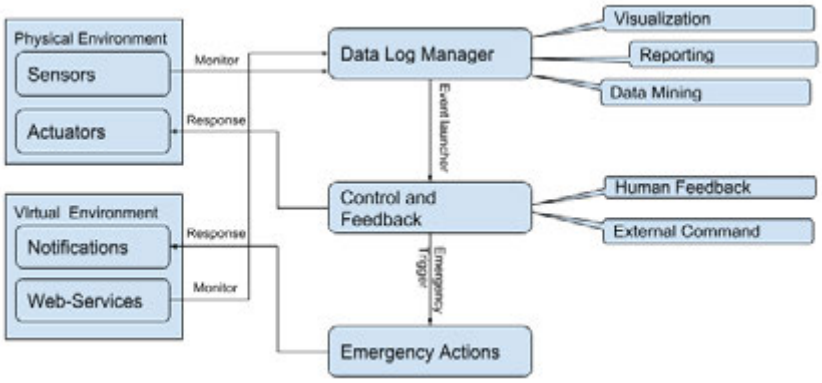


Fig. 1. The high-level architecture of a smart office

interpreting needs based on the bio-signals instead of explicitly asking the human to control the office. Dew Computing paradigm is integrating the lowest level, physical-edge human environment-controlling devices into higher level behavioural systems. Tiny microcontrollers are in direct contact with the human physical environment, controlling a multitude of simple or just slightly complicated environmental processes. These are the micro-controllers in gas boilers, air-conditioning and heating equipment, solar power controllers etc. These devices must possess self-sufficiency and can enable, by cooperation with devices of the same or higher level, interesting possibilities of energy saving, faster and more effective emergency service reaction and health-monitoring, among others. Dew devices, as micro-services, can be seamlessly integrated higher up into the computing hierarchy through Internet gateways into the Fog and Cloud, enabling the coordination of the physical edge control through common strategies like adaptive-learning, predicting and cognitive computing [21]. For example, we can find the design specifications of an ECG mobile application which enables a real-time acquisition and processing using a given wearable biosensor [7].

Context awareness is a critical task for constructing the application of the smart systems and is mainly seen as the process that explicitly acquires the context from underlying sensing devices to support system performance. Suitable context representation model and reasoning approaches should be examined to enable the high-level semantic acquisition from the raw sensing data [3]. It is important to establishing continuous experimentation as a management instrument which result in continuous user satisfaction. In the area of CPSs experimentation is underexplored and often not considered due to safety considerations and their proximity to the hardware level [11]. However, the CPS requirements are changing rapidly and therefore requiring a shift to iterative, human-centered development, in which continuous experimentation provide an excellent potential for facilitating its evolution and incrementally advancing its sustainability (economically, environmentally and in terms of usability). USE-ME framework [2] supports the design of empirical studies which take into consideration technical, physical and social elements by providing context-dependent metrics. This framework has the potential to model continuous experimentation of smart office.

A smart office can be composed of 3 main groups of modules and components, including data collection and monitoring, control and actuation, and visualisation, reporting and data mining components.

Data collection and monitoring:

- It can collect the status of the office elements, e.g. windows/doors are open/closed, a light turned on/off, heating system, air conditioner, power sockets, office appliances (e.g. coffee machine, water heater) are working or not.
- It can retrieve the data from the outside environment, e.g. sunlight, temperature, air pollution, humidity, and wind speed.
- It can monitor the status of the occupant: presence, position, activity type, stress factors, body temperature, and existential needs (food, water).

Control and Actuation:

- To react for the environmental events, such as closing the windows/door, turn the light on/off, open/close the heating/cooling system.
- To notify or warn about the events, such as reminding to drink water or eat food, to warn about the difference of the temperature between the room and outside, and so on.

Visualisation, reporting, and data mining:

- Visualization and Reporting: The data can be visualised and reported in different forms to different people for analysis, building maintenance, and so on.
- The huge amount of the data (big data) collected from the office, the people and their behaviour during the time can be used for extracting meaningful data and finding anomalies, patterns and correlations to predict the future events and behaviours.

These modules and components can be used in a smart office by adopting the architecture presented in Figure 1. The system includes the log manager, control and feedback, physical environment and virtual environment. The log manager collects the data and provides the events according to the pre-defined rules. These events can be used in the control

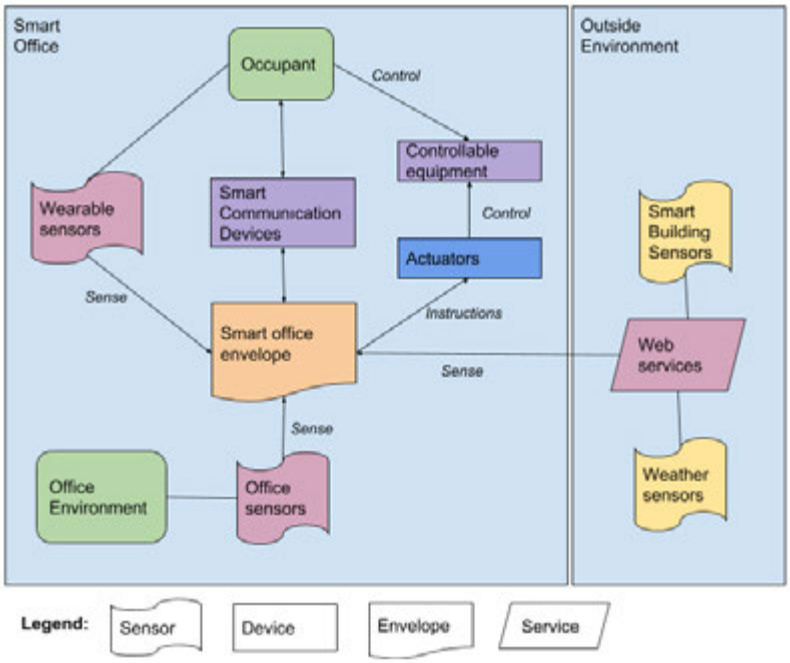


Fig. 2. The smart Office Setup model

and feedback module to send instructions to the actuators in the physical environment and/or messages as notifications to the users' smart devices such as a smartphone in the virtual environment. These are considered as the response to the event. The control section can detect the emergencies, such as fire occurrence or gas leak [1][14], and can trigger related actions in the system.

The system collects data with sensors in the physical environment and forms the web-services in the virtual environment. It can use this processed information to visualise them (for each of user understating), report (for management issues), and do data mining (for example, to find anomalies). Finally, the system can receive the feedback of the user or instructions from admins.

The system can support run-time features such as control of the office equipment, occupants health and safety conditions, and energy consumption. Also, it can support design-time features such as the support for facility manager (planning for resource management).

IV. Smart Office Setup

The smart office, discussed in this paper, is representing a common work-place, not the one which requires special conditions. It is not, for example, a work-place with no air or light. The central part of the office is a Smart Office Envelope (SOE) which performs control and sensing actions over the Inside Environment, Outside Environment and the Occupants (see Figure 2).

SOE via integration of smart technologies is becoming an autonomic system to respond to the environment alterations and user behaviour. It controls the technical services like heating, ventilation, air-conditioning, lights, electrical power, and network. The occupants and technical services are placed in spatial sequence (interior walls, desk, chair, work equipment) of Inside Environment [16].

The system supports interaction between the virtual model and the physical model. The Occupants can communicate with SOE by Smart Communication Devices (such as a smartphone, office computer, and smart-watch) to provide a control input or feedback, as well as to receive a notification from the SOE. Occupants play a proactive role in the smart office, not just as the part of the inside environment, but also as sensors/actuators.

Each occupant is having manual control over the Controllable Equipment in the office, such as windows, doors, lights, heating, air conditioning, ventilation, etc. The SOE can control this equipment through the set of instructions based on the rules which take the real-time data to make a decision. The SOE control over the equipment in most the normal conditions should not override the actions of the occupants. On the other hand, the decision can be made based on the data which is collected from:

- Occupants - through the Wearable Sensors (e.g. biosensors and eye-trackers) from which it is possible to obtain information regarding their body temperature, if they are hot/warm, hungry/thirsty, and in some cases measure a level of stress indicators.
- Office Environment - through the Office Sensors, such as object detector, face recognition system, light sensor, and so on, which can capture the information such as the level of light, noise, temperature, wind, humidity, air pollution (in specific part of the office), and the position of the occupant.
- Outside Environment - through the Web Services which

obtain the information from the web-based sources such as Weather Sensors (forecast) and Smart Building Data. For example, the Weather Sensors can provide information regarding the level of daily light, temperature, wind, humidity and air pollution. And Smart Building Data can give information such as emergency events in the building in which the office is located.

It is worth to note that the Weather Sensors can be in most cases found in the outside environment infrastructure, but if the service which provides this information is not available, it is possible to design outside sensors into the office infrastructure (e.g. from the outside part of the windows).

Inside Environment reacts to the occupant's behaviour and the Outside Environment. The system's objective is to optimise the occupant's comfort, use of energy (such as electricity consumption), as well as hazard and risk prevention , while the occupant's goal is the execution of the work assignments - as they have tasks, schedules, and they use tools and equipment.

To make the application of the system clear, here we present a representative scenario for a smart office. Let's consider a laboratory with some researchers working in it. Upon the occupant's agreement, they wear some biosensors, let say in the form of a wrist-watch or a necklace, to collect some biological information such as heart bit rate, body temperature, blood pressure, breathing rate and so on. On the other hand, the office can have sensors for temperature, humidity, CO2, smoke, and so on. The data can be stored in a database and the control can be done using a data-flow model, using data-flow programming such as NodeRed. This control mechanism can access the building data, such as catastrophic events and announcement. Also, the system is integrated with outside services such as weather forecast system. This control part of the system can handle the environmental equipment such as windows, doors, air condition, light, and so on, using some actuators and control devices such as step motors, remote switches, and so on.

This setup can pave the way to realise the requirements discussed in the previous sections mainly focusing on the occupant's comfort, energy saving, and hazard/risk management. In the realisation of these features, a golden rule is not to disturb the occupants and avoid unnecessary work interruptions. For example, the system can perceive the tiredness, thirstiness, or hungriness of the occupants and give some suggestions and recommendations based on the available facilities recorded in the system, to prevent the unhealthy situation, which will indirectly impact on the performance of the people. This can be also considered for the stressful times of the occupants by checking the body information and analysing the person's schedule and come up with useful suggestions, e.g. having a short walk or grabbing some food.

Also, the system can remind the staff when to leave the office, considering the air-pollution, the traffic, the appointments which he/she may have, and/or, as facilitator, by coordinating with some other colleague who plan to leave at the same time and close vicinity, to go together, resulting less pollution, safer, and healthier transportation. The system even can consider the temperature difference between the office, the building, and outside and based on this input automatically close the window and open the air-condition. For instance, in a high level of air pollution outside the building, a system can consume more energy in favour of the health of the occupants, without annoying them by control instructions.

Finally, the system can detect hazardous events, such as fire ignition, gas leak, and so on, at the earliest time possible and give the alarms and guide the occupants to evacuate the office at the soonest time possible. In this regard, the system can consider the exits, the number of people in each office and immediately direct the people in a way that avoid unnecessary rush and crowd, which usually lead to bottlenecks and even injuries/casualties.

V. Challenges

Successful implementation of the discussed office setup implies the following major challenges to be tackled:

Enhanced decision support systems - which balance rational use of resources, occupants comfort and emergency and resilience. Many works are addressing a rational use of resources by providing the occupants with information on how to reduce their spending. However, there are not many approaches which focus on ensuring occupants safety and comfort. We need accurate models that would allow us to reliably predict and simulate the behaviour of the work office which could be standardised.

User-friendly tools - with multi-modal user interfaces supporting the different stakeholders such as office designers, occupants and facility managers to facilitate appropriate human-computing interaction. For the design phase, we need tools to help the system's prescription. Also, during the commissioning period and the system's run-time, the occupants should be supported with the tools that enable easy feedback processing and individual control. Besides, facility managers, induced by tools that present the controllers' monitoring input, stream data analysis (with learning patterns), and the legal framework should be able to use the feedback to trigger a change in the set up of the office appliances infrastructure, rules and also behaviour. For that, easy to use tools and languages are needed.

Feedback data collection - of smart office may not only interact with humans (occupants and facility managers) but also communicate with the other systems using network infrastructures (e.g. outside environment and smart building). Collecting data from the occupants' trough wearable sensors should be seen as a complementary source of information resulting in the office environment improvements by adopting it to the individual needs. However, the office should be resilient by operating even when this data is not available by relaying o the office sensors.

Continuous experimentation - rely on the data continuously collected by smart office system and is a powerful resource for activities such as run-time diagnosis, early problem identification, improvement of system functionality, and delivery of the new features and products. However, its successful application complicates the organisational processes as

different stakeholders (e.g., hardware engineers, developers, and managers) should be coupled with each other. Finally, it is necessary to have supportive instruments for experimentation which provide virtual test environments for the new software updates and extensions or configuration changes of the office system.

Conformance with legal aspect - is an essential aspect while addressing statistical data collection. Given the variety of situations in terms of company size and diversity of the workforce, the interpretation of the legal issues can be ambiguous and should be often updated. However, while there is a need to design targeted and effective policy measures, OSH strategic framework calls for non-legislative tools to be used (i.e. benchmarks, identifying and exchanging good practices, awareness-raising, among others). Several EU funds are created to support the design and implementation of innovative and more productive ways of the workplace with the objective to increase coherence between environmental policy and worker's protection.

VI. Conclusions and Future Work

In this position paper, we present a smart office setup which exploits the new-age possibilities to improve the monitoring and control of the smart office while ensuring occupant's comfort, well use of resources and the emergency and resilience perspectives. While there is a lack of effective settings and tools which are supporting the design and analysis phases of smart-offices, they can be addressed by exploiting the emerging technologies (i.e. modelling techniques, experimental software engineering and dew computing).

The illustrated office setup collects data continuously through the office and wearable sensors. Although the impact of wearable sensors is still an under-exploration, the data continuously collected from these sensors can contribute to systematic and highly automated virtual testing.

As a future work we plan to adopt the approach in the context of Smart Office called SmartLab which is being implemented in our laboratory. The idea is to develop adaptive mechanisms which can automatically incorporate the measured information into the prediction models.

Acknowledgment

This research was (partly) supported by: i) the Scientific and Technological Research Council of Turkey (TUBITAK) under grant 115E591, and ii) Portuguese grants NOVA LINCS Research Laboratory (Grant: FCT/MCTES PEst UID/ CEC/04516/2013) and DSML4MAS Project (Grant: FCT/MCTES TUBITAK/0008/2014), and (iii) the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS).

References

[1] S. Arslan, M. Challenger, and O. Dagdeviren. Wireless sensor network based fire detection system for libraries. In *2nd International Conference on Computer Science and Engineering, UBMK 2017*, 2017.

[2] A. Barišić, V. Amaral, and M. Goulão. Usability Driven DSL development with USE-ME. *Computer Languages, Systems and Structures (ComLan)*, 51:118–157, 2018.

[3] A. Barišić, D. Blouin, V. Amaral, and M. Goulão. A Requirements Engineering Approach for Usability-Driven DSL Development. In *10th International Conference on Software Language Engineering (SLE)*, Vancouver, British Columbia, Canada, 2017. ACM.

[4] A. Barišić, C. Debreceni, D. Varro, V. Amaral, and M. Goulao. Evaluating the efficiency of using a search-based automated model merge technique. In *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 193–197. IEEE, oct 2018.

[5] F. Cicirelli, G. Fortino, A. Guerrieri, G. Spezzano, A. Vinci, F. Cicirellia, G. Fortino, A. Guerrieri, G. Spezzano, and A. Vinci. Metamodeling of Smart Environments: from design to implementation. *Advanced Engineering Informatics*, 33:274–284, 8 2017.

[6] S. Demirkol, M. Challenger, S. Getir, T. Kosar, G. Kardas, and M. Mernik. Sea_l: A domain-specific language for semantic web enabled multi-agent systems. In *2012 Federated Conference on Computer Science and Information Systems, FedCSIS 2012*, pages 1373–1380. IEEE, 2012.

[7] E. Domazet, M. Gusev, and L. Antovski. Design specification of an ECG mobile application. In *2017 25th Telecommunication Forum (TELFOR)*, pages 1–4. IEEE, 11 2017.

[8] C. Durmaz, M. Challenger, O. Dagdeviren, and G. Kardas. Modelling contiki-based IoT systems. In *OpenAccess Series in Informatics*, volume 56, 2017.

[9] S. Engell, R. Paulen, C. Sonntag, H. Thompson, M. Reniers, S. Klessova, and B. Copigneaux. *Proposal of a European Research and Innovation Agenda on Cyber-physical Systems of Systems - 2016-2025*. www.cpsos.eu, 2016.

[10] C. European. EU Occupational Safety and Health (OSH) Strategic Framework 2014-2020. Technical report, Brussels, 2014.

[11] F. Giaimo, H. Yin, C. Berger, and I. Crnkovic. Continuous Experimentation on Cyber-Physical Systems. In *Proceedings of the Scientific Workshop Proceedings of XP2016 on - XP '16 Workshops*, pages 1–2, New York, New York, USA, 2016. ACM Press.

[12] T. Johnson, A. Kerzhner, C. J. J. Paredis, and R. Burkhart. Integrating Models and Simulations of Continuous Dynamics Into SysML. *Journal of Computing and Information Science in Engineering*, 12(1):011002, 3 2012.

[13] B. Karaduman, T. Asici, M. Challenger, and R. Eslampanah. A Cloud and Contiki based Fire Detection System using Multi-Hop Wireless Sensor Networks. In *Proceedings of the Fourth International Conference on Engineering & MIS 2018*, page 66. ACM, 2018.

[14] B. Karaduman, M. Challenger, and R. Eslampanah. ContikiOS based library fire detection system. In *2018 5th International Conference on Electrical and Electronic Engineering (ICEEE)*, pages 247–251. IEEE, 2018.

[15] N. Karimpour, B. Karaduman, A. Ural, M. Challenger, and O. Degdeviren. IoT based Hand Hygiene Compliance Monitoring. In *The 2019 International Symposium on Networks, Computers and Communications (ISNCC'19)*, Istanbul, 2019. IEEE (To be Published).

[16] D.-Y. Kim and S.-A. Kim. An exploratory model on the usability of a prototyping- for designing of Smart Building Envelopes. *Automation in Construction*, 81:389–400, 9 2017.

[17] V. Loftness, V. Hartkopf, A. Aziz, J.-H. Choi, and J. Park. Critical Frameworks for Building Evaluation: User Satisfaction, Environmental Measurements and the Technical Attributes of Building Systems (POE + M). In *Building Performance Evaluation*, pages 29–48. Springer International Publishing, Cham, 2018.

[18] H. M. Marah, R. Eslampanah, and M. Challenger. DSML4TinyOS: Code Generation for Wireless Devices. In *2nd International Workshop on Model-Driven Engineering for the Internet-of-Things (MDE4IoT), 21st International Conference on Model Driven Engineering Languages and Systems (MODELS2018)*, Copenhagen, Denmark, 2018.

[19] L. Özgür, V. K. Akram, M. Challenger, and O. Dagdeviren. An IoT based smart thermostat. In *2018 5th International Conference on Electrical and Electronic Engineering (ICEEE)*, pages 252–256. IEEE, 2018.

[20] P. H. Shaikh, N. B. M. Nor, P. Nallagownden, I. Elamvazuthi, and T. Ibrahim. A review on optimized control systems for building energy and comfort management of smart sustainable buildings. *Renewable and Sustainable Energy Reviews*, 34:409–429, 6 2014.

[21] Z. Sojaat and K. Skala. The dawn of Dew: Dew Computing for advanced living environment. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 347–352. IEEE, 5 2017.

Innovative predictive model for Smart City security risk assessment

L. Franchina* and A. Socal**
* President of AIIC (Italian Association of Critical Infrastructures' Experts), Rome, Italy
** Member of the Board of Directors of AIIC (Italian Association of Critical Infrastructures' Experts), Rome, Italy

Abstract - **In a Smart City, new technologies such as big data analytics, data fusion and artificial intelligence will increase awareness by measuring many phenomena and storing a huge amount of data. 5G will allow communication of these data among different infrastructures instantaneously. In a Smart City, security aspects are going to be a major concern. Some drawbacks, such as vulnerabilities of a highly integrated system and information overload, must be considered. To overcome these downsides, an innovative predictive model for Smart City security risk assessment has been developed. Risk metrics and indicators are defined by considering data coming from a wide range of sensors. An innovative "what if" algorithm is introduced to identify critical infrastructures functional relationship. Therefore, it is possible to evaluate the effects of an incident that involves one infrastructure over the others.**

Keywords - predictive model; "what if" methodology; domino effect; consequences calculation; risk assessment; smart city; big data analytics.

I. INTRODUCTION

In a modern smart city, everyday life relies on Critical Infrastructures (CI) such as energy distribution systems, transportation systems, hospitals, financial centers and so on. As CI deliver vital services to the citizens, their societal impact has increased significantly in the last century. For these reasons, significant efforts are being produced at national and EU level, to increase CI security both from a governance and a technological point of view.

A. State of the art on the Critical Infrastructures

From a governance point of view, EU legislation produced in the last years two important Directives:

1. The Council Directive 2008/114/EC of 8 December 2008 on the identification and designation of European Critical Infrastructures (the ECI Directive) and the assessment of the need to improve their protection [1].
2. The Council Directive 2016/1148 on security of Network and Information Systems (the NIS Directive), which identifies the perimeter of recognition of the society's vital sectors and moreover that rely heavily on ICTs, such as energy, transport, water, banking, financial market infrastructures, healthcare and digital infrastructure [2].

These Directives are the pillars of EU policy regarding CI identification and protection. Particularly, with the ECI Directive, the term Critical Infrastructure has been defined as: "an asset, system or part thereof located in Member States which is essential for the maintenance of vital societal functions, health, safety, security, economic or social well-being of people, and the disruption or destruction of which would have a significant impact in a Member State as a result of the failure to maintain those functions". Moreover, in the ECI Directive, the cross-cutting criteria for impact definition have been defined as:

1. Casualties criterion (potential number of fatalities or injuries);
2. Economic effects criterion (economic loss, degradation of products or services, environmental effects);
3. Public effects criterion (public confidence, physical suffering and disruption of daily life, loss of essential services).

B. Benchmark of tools for modeling and simulating impacts and interdependencies

From a technological point of view, in the last twenty years, many were the projects undertaken to analyze and face cascading effects of an incident on critical infrastructure complex and interdependent systems.

In the USA, the Department of Homeland Security (DHS) has launched the Enhanced Critical Infrastructure Protection program from the Public-Private-Partnership perspective. Information exchanges and visits to critical infrastructure operators are often followed by a security survey performed through the Infrastructure Survey Tool (IST) [3]. IST is an optional, web-based survey to identify and document the overall safety and resilience of the facility. The tool is used for:

- Identify and evaluate physical security, resources security, management, information exchange, security measures and dependencies related to the preparation, mitigation, response, resilience and recovery of the facility;
- Identify security gaps;
- Create indicators for monitoring protection measures and resilience to compare different infrastructures;
- Monitor progress in security.

Survey data, composed of scores based on a variety of factors for the specific infrastructure, are provided to owners and operators through a dashboard that compares the data with similar structures and generates reports on the

status of the protection measures, resilience planning and resource allocation.

Argonne National Laboratory, one of the largest and most historic national laboratories of the U.S. Department of Energy, has developed a methodology for assessing the risk and resilience of infrastructures with respect to a variety of anthropogenic and natural threats for several sectors such as energy, transport, water treatment, financial institutions and commercial facilities. Under DHS's Enhanced Critical Infrastructure Protection (ECIP) program, they developed the BIRR methodology (Better Infrastructure Risk and Resilience) for operators in 18 critical infrastructure sectors [4]. This method is distinguished by:

- Approach: a sectoral approach that reaches the level of asset detail and gives priority to protection measures applied mainly to terrorist threats;
- Purpose: to provide policy makers with tools for the analysis of various sectors, identification of vulnerabilities and preparation of risk reports;
- Target audience: Policy maker

The methodology focuses on the evaluation of three correlated indices:

- VI, Vulnerability Index
- PMI, Protective Measures Index
- RI, Resilience Index

Developed by the NI2 Center for Infrastructure Expertise, the CARVER method (Criticality Accessibility Recoverability Vulnerability Espyability Redundancy) is a non-technical method for comparing and prioritizing critical infrastructures and key national resources [5]. It is declared to be the only evaluation tool that prioritizes critical infrastructures belonging to different sectors. There is a PC stand-alone version and a server / client version (CARVER2Web). The methodology was designed to cover both terrorist threats and natural disasters by implementing a multi-hazard approach.

The Critical Infrastructure Protection Decision Support System (CIP / DSS) project is developing a risk-informed decision support system that provides information / analysis to make decisions about critical infrastructure protection by considering 17 critical infrastructure categories and their primary interdependencies. Initiated as a proof-of-concept in August 2003, the CIP / DSS project completed a prototype and two case studies in February 2004. It demonstrated how CIP / DSS will help decision makers to make informed choices through:

- a functional representation of all 17 critical infrastructures with their interdependencies,
- a calculation of the impact on human health and safety, economic, national, on trust and the environment,
- summarizing a methodology that is technically valid, defensible, and extensible.

The National Infrastructure Simulation and Analysis Centre (NISAC) developed Fast Analysis Infrastructure Tool (FAIT) a tool that exploits data from infrastructure and expert knowledge. Composed by four main elements, interdependency assessment, co-location of critical infrastructure and information sharing and economic impact, it deals with interdependencies by considering expert knowledge coded in a rule-based system software language used to express the relationships between the various infrastructures.

Sandia National Laboratories developed Modular Dynamic Model whose theoretical background is based on agent-based modelling and dynamic modelling systems. It provides a first screening that can be followed by a more detailed analysis if the identified risks require it.

Based on agent modelling system in a stochastic environment, the Network Security Risk Assessment Modelling (NSRAM) of James Madison University aims at determining how system react and interact to different kind of incidents and attacks. Besides failure events, such a model includes repair capability that models the effects of repair personnel or part scarcity, communication requirements and uncertainty.

The Centre Risque & Performance (CRP) of the École Polytechnique de Montréal developed DOMINO, an early warning tool that through a consequences-based risk management approach, aims at anticipating and modelling the spatial and temporal spread of domino effects. Started in 2008, it continues to be at prototype stage [6].

CIPRNET project (Critical Infrastructure Preparedness and Resilience Research Network) aims at improving the understanding, preparation and mitigation of the consequences of CI disruptions following an all hazards approach. It developed different tools as a Decision Support System (DSS), which includes consequences analysis, threats prediction and visualization and data collection and access. It was designed as a forerunner of the European Infrastructures Simulation & Analysis Centre (EISAC) by 2020 [7] [8].

A DSS called CIPCast has been developed from CIPRNET by ENEA (Italian National agency for new technologies, energy and sustainable economic development) [9]. CIPCast is a GIS-based type DSS which allows the sharing of territorial and environmental information of the area of interest, the monitoring and security of the ICs, as well as the assessment of risk scenarios and the management of emergencies. It is based on the integrated availability of a series of elements:

- geographic data and territorial information, monitoring data, etc.;
- tools and applications for the analysis of such data (both in real-time and in simulation mode);
- analysis systems for modeling and / or simulation (risk forecast / assessment, damage scenarios, etc.);
- geographical interface (WebGIS) as an interactive tool for decision support.

Overall, thanks to the features implemented in CIPCast's DSS system, it is possible to carry out an

assessment of the impacts on the integrated IC system and estimate the possible consequences, both starting from real context data and through the preparation of simulated scenarios (eg., through the CIPCast-ES module it is possible to carry out earthquake simulation and estimate damage to the buildings and the CI) [10].

Started in 2014 CASCEFF aims at improving knowledge about cascading effects during a crisis by identifying a sequence triggered by an initiating event and considering the characteristics of the systems and impacts on the dependent systems and overall system [11].

The three-year project FORTRESS (Foresight Tools for Responding to cascading effects in a crisis) was meant to identify and understand cascading effects in a crisis by processing data based on empirical results coming from a wide range of crises events. Among its main outputs there is the development of a predictive model addressing potential impacts of decisions made in crisis situations, and the development of an incident evolution tool to assist decision-makers in preparing and training for crises with cascading effects [12].

Platform GRASSP (Geospatial Risk and Resilience Assessment Platform) is a World Wide Web oriented architecture bringing together geospatial technologies and computational tools for the analysis and simulation of critical infrastructures. It can be used for the analysis of complex networked systems taking into consideration cross-sectoral and cross-border interdependencies [13].

The Australian tool CIPMA (Critical Infrastructure Program for Modelling and Analysis) is a computer-based capability, which uses a wide range of data and information to model and simulate the behavior and dependency relationships of critical infrastructure systems. The overall aim of the CIPMA Program is to develop the capability to answer questions about critical infrastructure dependencies and interdependencies, and the flow-on consequences of a complex systems failure in one sector [14].

II. "WHAT IF" METHODOLOGY

The presented methodology is the result of a decennial work [15] [16] [17] [18]. Within this paper a contextualization to the smart city environment, along with innovative future developments, is presented.

A. Item definition

The developed predictive model aims to study and quantify the consequences of negative, unexpected and blocking events and to develop a propagation model of failure events (cybernetic and mixed) on the economic and social infrastructures that operate in our society. The analysis aims to provide a concise description of the interdependencies between the various productive and social sectors.

The "What if" methodology is based on a classification of socio-economic activities that characterize the quality of life in modern society; this classification, derived from statistical-economic sources, from sociological studies and from the international literature on critical infrastructures, includes 104 "items". Every item relates a specific resource and is composed by different components. Assets of an item are its functional parts (see figure 1).

Resource	Item	Component	Asset
Transport	Railway Transport	Production Transport Distribution Fruition	Railway Stations
	Air Transport	Production Transport Distribution Fruition	...
	...	...	...

Figure 1. Example of Items related to transport resource with their linked components and assets

For each item, the comprehensive supply chain is considered. Thus, it is possible to both identify the specific level (production, transport, distribution or fruition) which causes an outage of service and the effects of the out of service itself on other item's levels (see figure 2).

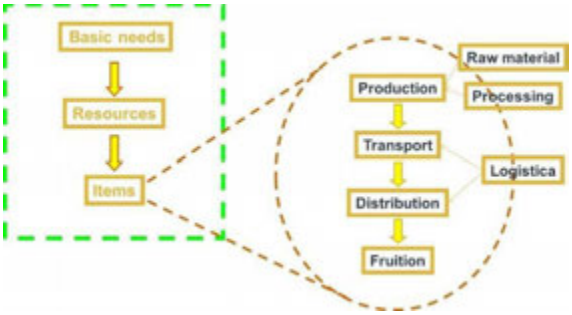


Figure 2. Schema of an Item supply chain

B. Impact assessment and input data mode

The assessment of the impact of the out of service on the population is based on three types of indicators defined by the European Directive 114/08 / EC, related to the economic impact, the number of potential victims of the out of service and the public impact (suffering of the population, distrust of institutions, disruption of daily life). In order to evaluate these indicators, economic losses, fatalities and measurement of public confidence are taken into account.

Economic losses, fatalities and measurement of public confidence represents the input data of the developed methodology. Regarding data input mode, two different ways of feeding input data are possible:

1. data feeding from humint with surveys to the experts of the various service categories with a questionnaire;
2. data feeding from a platform connected to the main critical infrastructures, able to receive and interpret their log and control data from SIEM (Security Event and Information Management) systems.

C. Domino effect forecast model

The "what if" methodology is based on two different analytic models. The first model, called "Domino Effect Forecast model", starting from the list of items and direct dependencies between them, aims to generate for each item the so-called "domino effect tree", that is a graphic representation of the timeline with which, in the event of an out of service of the item itself, other items are no longer able to provide their service.

The first step consists in defining the “QoS (Quality of Service) degradation matrixes” among different items, that are the matrixes that identify which item is influenced by which other one with respect to time count after out of service of first item. As an example, in figure 3 four different degradation matrixes are presented: according to the figure, item 2 is affected by outage of item 3 after 4 hours from the impact, then it goes out of service after 24 hours and other items are affected and so on.

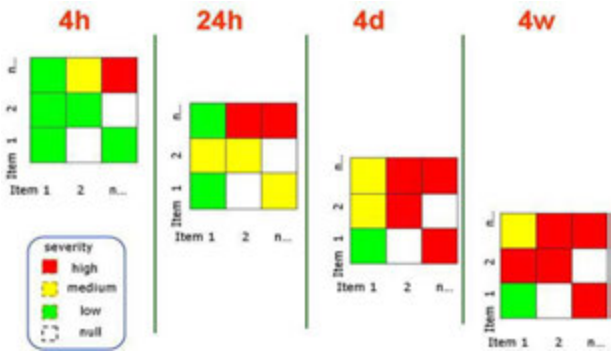


Figure 3. Example of QoS degradation matrixes

The second step consists in defining the proper domino effect tree, starting from the “QoS degradation matrixes”. As an example, in figure 4 a possible tree of the Domino effect of item 1 is represented: according to the figure, if in a given moment of time item 1 goes out of service, after 4 hours, item 4 would not be able to provide its service anymore; after another 4 hours also item 3 would be out of service, followed after 24 by item 6 and so on. Although in reality the domino effect is a continuous time phenomenon, the tree that represents it uses a discrete time scale in which the origin (t = 0) is the instant in which the out of service of a particular occurs item (in the figure, item 1) and the presence of other items (item 4, item 3, item 6, item 2 and item 5) at a particular time band indicates after how much time, compared to at = 0, 1 'reported item will go out of service. In the "a priori" analysis, each item can be associated with several trees of domino effects, depending, for example, on the period of the year or on the geographical location in which the out of service occurs. Also, the quantized time scale can vary according to the cases considered.

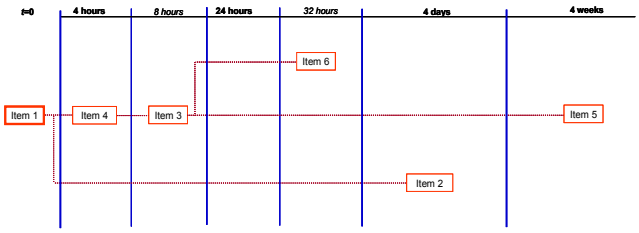


Figure 4. Example of domino effect tree

Each tree represents the decay flow (extinction) of the various sectors that make up the social structure and the labor system of a country following the extinction (outage) of a "root" node. The sectors considered are all those that give rise to the "quality of life" of the average citizen, regardless of whether they generate a money exchange. All sectors have IT and computerized aspects and have been the target of various cyber-attacks. Through the domino effect model, it is possible to draw diagrams of "criticality" by analyzing the timing and the relationships of direct dependence for all the taxonomy sectors, as shown in the example of figure 5.

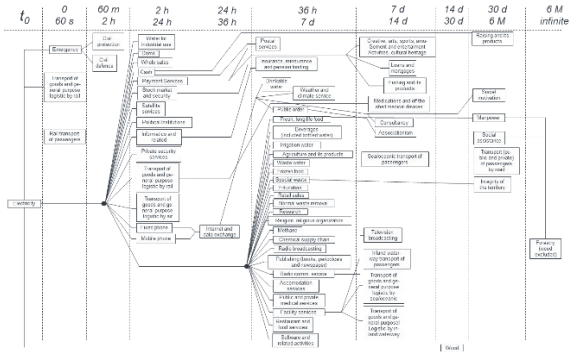


Figure 5. Domino effect for electricity sector

D. *Consequence calculation model*

The second analytical model developed is called "Consequence Calculation Model". It aims to associate each of the items with a calculated impact value in terms of the indicators envisaged by the NIS Directive, such as economic impact, number of potential victims of the off-duty and public impact (see figure 6).

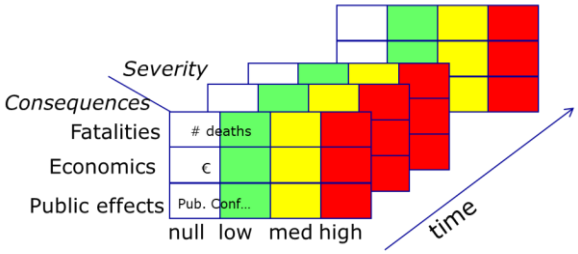


Figure 6. Consequence calculation model

From the combination of the two models it is possible to evaluate the consequences of an out of service (determined by cyber or cyber / kinetic events) of each item on the population, considering the direct impact and that caused by the domino effect.

The “what if” methodology has been developed starting from a national level, considering all items of a typical European country. However, the methodology is characterized to be scalable, in order to be adapted to different national levels such as regions or cities or even to specific service operators (see figure 7).

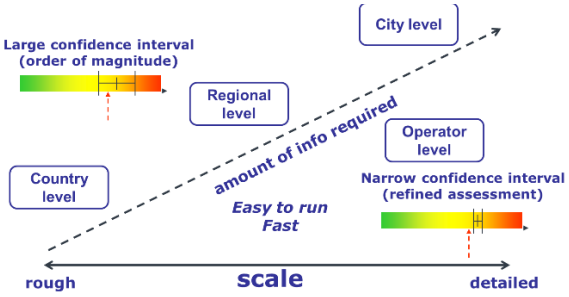


Figure 7. Confidence levels of the developed methodology

E. Model Validation

Both the Domino effect forecast model and the consequence calculation model have been integrated in an intelligence modular platform called Dominio. Dominio is a dashboard composed by six modules (which can be increased in the future), each of which aims to answer to a specific question, thus reducing complexity.

The presented methodology is under continuous validation since it is used in several application at operator level by Dominio platform users. Operator's data are measured in terms of reliability index with reference to input data. The developed mathematical model is deterministic, so, the methodology has a high reliability index. However, input data may be biased or unreliable, thus introducing systematic error in the model.

Authors are available for further details regarding these aspects, according to NDAs signed with Dominio platform operators.

F. Model application to a Smart City

There is no definitive definition of a smart city because of the variety of the technologies that can be incorporated into a city to be considered a smart city. Mark Deakin defines it as a city that utilizes ICT to meet the demands of its citizens, and that community involvement in the processes is a necessity for a smart city. From the definition given by Husam Al Waer and Mark Deakin in their research publication "From Intelligent to Smart Cities" [19] the factors that contribute to a city being classified as smart are:

- The application of a wide variety of digital and electronic technologies to the city and its communities
- The application of ICT to uplift life and the working environments in the region
- The embedding of such ICT within government systems
- The territorialization of practices that bring the people and ICT together in order to foster innovation and enhance the knowledge that they offer

Arkalgud Ramaprasad, Aurora Sánchez-Ortiz and Thant Syn in their publication: "A Unified Definition of a Smart City" [20] even realized an ontology which provides a path to conceptualize systemically and systematically this novel domain including, refining, and extending previous definitions and conceptualizations of smart cities in a simple but powerful way.

However, even if definitions of smart city are still various and different one from each other, they all agree on identifying massive use of ICT in shaping architectures, systems, policies and processes of a smart city. Moreover, because of its nature, smart city widely relies on CI for its sustainability, depending on all the vital services identified by NIS Directive (energy, transport, water, banking, financial market infrastructures, healthcare and digital infrastructure).

Thus, the presented model is perfectly suitable for a smart city, even if some aspects should be adjusted to the specific environment:

- Items should be defined according to economic and social structures which operates within the smart city;
- Input data mode should rely mostly on automatic systems, considering possibilities given by ICT within smart city;
- Sensors integrated in the smart city environment should be used to automatically generate the "consequence calculation model" according to their capability in detecting different types of impact (casualties, economic or public effect).

G. Future developments

A criticality linked to the current approach to the problems present in the world of cyber security lies in the heterogeneity of the technologies and products dedicated to monitoring and protection of IT infrastructures. This phenomenon leads to a lack of vision of the global reference context in which cyber security problems are inserted.

From this consideration arises the need to create a single point of monitoring of multiple IT infrastructures that can collect and correlate data and information produced by the protection technologies present in all the infrastructures to be monitored.

A tool of this type has multiple advantages:

- Have a complete picture of the current IT security problems present in multiple IT infrastructures;
- Being able to evaluate in advance the risks and the impact of specific security problems that have occurred in other infrastructures similar to the one monitored;
- Have visibility and be able to intercept attacks implemented in different infrastructures but operationally interconnected.

If applied to a smart city environment, such a tool would represent an innovative Decision Support System (DSS). Statistical data coming from smart city sensors will build the knowledge basis for the system. The DSS would then be able to support all phases of the risk analysis process, starting from event forecast, prediction of adverse scenarios, impact estimation (with respects to the three cross-cutting criteria defined by NIS) and, finally, cascade effects. DSS outcome should be emergency plans and strategies to be adopted to reduce the impact and manage the recovery phase depending on "domino effect tree" and "consequence calculation" outcomes integrated with statistical data.

To be more specific, starting from the knowledge of the punctual position and the functional dependencies among different IC which guarantee the smart city subsistence, once a failure event is detected the DSS developed by the presented methodology would be able to calculate the outage of service time and the cascade effect of the phenomena. For example, considering a blackout, it would be possible to calculate its duration for any infrastructure

located in the smart city area and to estimate the reduction or loss of any service whose supply chain has been affected by the impact.

Such a DSS could be obtained, for example, by integrating an already existing system (like for example CIPCast), with the developed "what if" methodology. The result would be a visual analytics tool which integrates contextual data on a smart city IC outage with both outage duration data and effects on other IC coming from "domino effect tree" and losses evaluation obtained from "consequence calculation model".

Another aspect of interest concerns the need to have available an environment in which to replicate a real IT infrastructure and the set of processes that the infrastructure manages. The objective is to verify the resistance of the infrastructures to the different types of attack and the operational assessment of the impact of the attacks on the infrastructures. This implies the need for a simulation standard to be used to support rapid training, vulnerability assessment and targeted testing of protection tools. An erroneous assessment of the impact can negatively influence the perception, categorization and classification of risks by the management. This is even more significant in the field of IT risks whose impact is inherently more difficult to evaluate and predict.

"Digital Twin" is emerging as a disruptive technology that will change the paradigm of simulation. "Digital Twin" refers to a computer model which mirrors and simulates an asset or a system of assets and their surrounding environment. "Digital Twin" models can help organize data and pull it into interoperable formats so that it can be used to optimize infrastructure use. "Digital Twin" has been included in both 2018 and 2019 top 10 strategic technology trends by Gartner, the world's leading research and advisory company [21] [22].

Integration of the developed predictive model with "Digital Twin" technology, could bring to an even more realistic model. Characteristics such as robustness of the model, the link to the real world, the application of advanced big data analytics and the ability to evaluate "what if" scenario are perfectly suitable for the purpose to detect future disruptive events and identify the proper measures to be implemented to reduce both risk and impact effects.

III. CONLUSION

The predictive model presented is particularly innovative in the actual scenario for the simplicity of the input system itself. Compared to other model and systems, the presented one still relies significantly on the human expertise, reducing systemic errors which could be produced by false positive or altered data. This way, output results would be more accurate rather than precise. Moreover, flexibility of the "what if" methodology and its independence from specific HW and SW technologies supports the future integration of the model itself in a DSS dashboards and / or with a "Digital Twin" model.

REFERENCES

[1] ECI Directive (Directive 2008/114/EC) on EUR-Lex website: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM%3Ajl0013

[2] NIS Directive (Directive 2016/1148) on European Commission website: https://ec.europa.eu/digital-single-market/en/network-and-information-security-nis-directive

[3] IST tool on US Homeland Security website: https://www.dhs.gov/infrastructure-visualization-platform

[4] BIRR on Argonne National Laboratory website: https://www.anl.gov/article/better-infrastructure-risk-and-resilience

[5] G. Giannopoulos, R. Filippini, M. Schimmer, "Risk assessment methodologies for Critical Infrastructure Protection. Part I: A state of the art", Luxembourg: Publications Office of the European Union, 2012

[6] Polytechnique Montréal website: https://www.polymtl.ca/centre-risque-performance/en/domino

[7] Critical Infrastructure Preparedness and Resilience Research Network on European Commission website: https://cordis.europa.eu/project/rcn/107425/reporting/en

[8] V. Rosato, "Decision Support System for Critical Infrastructure Protection", from ERNCIP Pre-Conference Special Session "2nd ERNCIP Conference", Brussels April 15, 2015.

[9] V. Rosato, "Sistemi innovativi Sistemi innovativi innovativi e gestione gestione gestione delle crisi delle Infrastrutture critiche Infrastrutture critiche critiche: il progetto progetto progetto EU CIPRNet", ENEA Centro Ricerche Casaccia, Roma, http://www.masterhomelandsecurity.eu/wp-content/uploads/2016/12/Vittorio-Rosato_Infrastruttue-Critiche.pdf

[10] C. Moriconi, M. Pollino, V. Rosato, "La protezione delle Infrastrutture Critiche e il controllo del territorio", ENEA Centro Ricerche Casaccia, Roma, published on Energia, ambiente e innovazione, january 2017, pages 52-57

[11] CASCEFF website: http://casceff.eu/

[12] Foresight Tools for Responding to cascading effects in a crisis on European Commission website: https://cordis.europa.eu/project/rcn/185488/factsheet/en

[13] Geospatial Risk and Resilience Assessment Platform on European Commission website: https://ec.europa.eu/jrc/en/grrasp

[14] CIPMA website: https://www.tisn.gov.au/Pages/CIPMA.aspx

[15] L. Franchina, M. Carbonelli, L. Gratta, C. Petricca, D. Perucchini "An effective approach for cascading effects prevision in Critical Infrastructures" Critical Information Infrastructures Security, Springer, pp. 386-393, 2008

[16] L. Franchina, M. Carbonelli, M. Crisci, L. Gratta, D. Perucchini "An impact-based approach for the analysis of cascading effects in Critical infrastructures" Int. J. Critical Infrastructures, Vol. 7, No. 1, 2011, pp. 73-90

[17] L. Franchina, Critical Infrastructure Protection: Public emergency plans and command and control chains, the Italian example Non linear approaches to crisis and conflict, Ionut Purica editor, expert editor, Romania, 2011

[18] L. Franchina, Critical Infrastructure Protection Cyber world, Hoepli editore, AA.VV. a cura dell'Osservatorio per la Sicurezza Nazionale, 2012

[19] M. Deakin, H. A. Waer, "From Intelligent to Smart Cities", 2012

[20] A. Ramaprasad, A. Sánchez-Ortiz, T. Syn, A Unified Definition of a Smart City

[21] Gartner Top 10 Strategic Technology Trends for 2018 on Gartner website: https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2018/

[22] Gartner Top 10 Strategic Technology Trends for 2019 on Gartner website: https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2019/

Digitalization and Smart islands in the Kvarner archipelago

Marko Mimica*, Goran Krajačić *, Damir Medved**, Darko Jardas***
* Faculty of Mechanical Engineering and Naval Architecture, Zagreb. Croatia
** Ericsson Nikola Tesla d.d., Zagreb, Croatia
*** Regionalna energetska agencija Kvarner, Rijeka, Croatia
e-mail adresa: mmimica@fsb.hr

***Abstract* - Currently, in the European Union there are two major initiatives that address the issue of sustainable development of the islands. Top-down initiative coming from the Declaration on Clean energy for EU islands and bottom-up Smart Islands initiative. Both initiatives consider islands as living labs that can host innovative pilot projects and lead Europe's transition into a sustainable and low carbon environment. Following objectives of both initiatives, the Croatian government passes a law on islands that support the development of smart islands. The paper presents a comprehensive review of studies, projects, policy documents and scientific publications related to Smart Islands development in Kvarner archipelago which can be used as a model for implementing new sustainable technology on the islands. It also provides expert recommendations for digitalization in several sectors crucial for the development of islands and an integrated approach to management of island resources.**

Ključne riječi – Smart island; Kvarner archipelago; digitalization; blockchain; information and communication technology

I. INTRODUCTION

Current ongoing digitalization is penetrating every sector of society. Its impacts are visible in energy, transport, social, economy and many more sectors. Even more, digitalization has enabled new communication pathways between these sectors. Implementation of advanced information and communication technology (ICT) created a more interconnected world with the highest possibility of a collaboration of different stakeholders than ever before. As in every technological and industrial revolution, there are technologies which can be first applied to the islands, so they could feel the first impact of these positive changes. Islands can be living labs for implementation of new technology and therefore be leaders of change. Although isolation and poor connection with the mainland create difficulties for the island population, these constraints can be beneficial for the execution of some pilot projects and implementation of new sustainable technologies. If successful on the islands, these projects could most probably be implemented in some rural areas on the mainland. The European Union (EU) acted on this issue and supported two initiatives that will create the framework for future development of the islands. Top-down initiative Clean Energy for EU Islands [1] provides a pathway for islands to create low carbon and sustainable environment with low energy prices. The initiative will result with the implementation of advanced technologies for energy storage, less dependence on import of energy, improvement of air quality and preservation of island natural heritages. Another framework is Smart Islands Initiative [2], which is a bottom-up initiative. Smart Islands Initiative sets ten action points for creating a smart and inclusive society on islands. It seeks to present islands as the test – beds for hosting projects that stress optimal usage of local resources for purpose of creating sustainable communities. With Smart Island Initiative, the role of islands is no longer to follow, but to lead Europe's transition towards a more sustainable community.

Many kinds of research have been made in order to investigate the possibilities for smart technology integration on islands. One of the first methods developed was RenewIslands method presented in [3]. RenewIslands provided a qualitative indicator for evaluation of islands' needs and resources on many different sectors including energy, water, waste, and wastewater needs as well as energy import infrastructure and wind, solar, biomass, geothermal and hydro potential. This method was foundation for modelling energy systems of different islands as for example in [4], where authors modeled energy system of several interconnected islands in tool EnergyPLAN showing that interconnected islands have decline of Critical Excess Electricity Production (CEEP) up to 22% in comparison with scenario when islands are not connected in common system.

The main problem regarding renewable energy sources (RES) integration is balancing between production and consumption as described in [5]. Many researchers suggest that integration of different flows substantially increase system possibility to integrate RES. Dominković et. al [6] combines smart charging, reverse osmosis technology and an ice storage for increasing the possibility for RES integration. In another work, Segurado et al. [7] present analysis of integrated water and energy supply where excess electricity production from the wind power plant is used by desalination units for covering water demand of the population or where

desalinated water is used as energy storage in a pumped hydro system.

These papers made a significant contribution regarding increasing smart technology integration in island systems by combining different flows, but they don't investigate the role of information and communication technology (ICT) in this process. Importance of ICT in RES integration is described at [8] where the author emphasizes that with smart meters and tracking of generation and consumption in real time it is possible to effectively balance the energy system. Bibri S. E. et al. [9] emphasizes the Internet of Things (IoT) concept as one of the key components for implementing ICT in smart systems. According to him, IoT has great potential to improve energy efficiency and, in combination with big data applications, to accelerate sustainable development. In [10] the authors provide a concept for achieving automatic control by using IoT and cloud computing. They suggest that opportunities and challenges that emerge from big data implementation will be the driving force for the development of smart cities. Gubbi et al. [11] present user-centric cloud model based on an interaction of private and public clouds and indicates the importance of integration of IoT and cloud computing in order to provide reliable support to users using services offered by multiple stakeholders.

Several studies and action plans were completed regarding islands in the Adriatic Sea, mostly focused on mapping and suggesting measures for the implementation of sustainable technology for achieving a low carbon and sustainable environment. Most advanced islands in this aspect are Kvarner archipelago islands. The biggest islands of Kvarner archipelago are Krk, Cres, Lošinj, and Rab while some of smaller ones are Unije, Susak, and Plavnik. Pag is in the geographic area of Kvarner archipelago, but administratively is not part of it and therefore is not considered as part of Kvarner archipelago. The available data about these islands from studies [12], [13], [14], [15] is presented in Table I.

II. METHODS

This paper's objective is to offer a path for islands to mitigate towards more sustainable and low carbon environments. Framework for development of Smart islands in Croatia is given in Table II. It is consisted out of five phases (P1 – P5) and for every phase, there is an estimated duration as well as required actions and desirable results. It is worth noting that it is not desirable to try to implement this concept on the whole area, such as the Adriatic Sea, but to rather implement Smart island concept on a single island and then expand it further (Figure 1). Islands that have similar characteristics or that are located in the same area such as Kvarner archipelago have a better chance that technological solutions implemented on one island will be suitable on other islands as well. One of the key aspects in setting up the right framework for the development of Smart islands is to do a quality mapping. RenewIslands is a method that will is mostly used for this purpose. It is used for the creation of several alternative scenarios for energy planning based on islands needs and resources. Mapping of islands needs is assessed with electric energy consumption, heat and cold energy consumption, transport energy consumption, waste and wastewater treatment and water demand, while islands resources are assessed with the wind, solar, hydro, biomass and geothermal potential. In addition to this, there are also parameters for the islands' energy import infrastructure and water infrastructure.

TABLE II. FRAMEWORK FOR DEVELOPMENT OF SMART ISLANDS IN CROATIA

Phase	Framework		
	Duration	***Actions***	***Results***
P1	1991 – 2003	Generating ideas, setting up frameworks	National Island Development Programme, Law on Islands
P2	2004 – 2016	Advocacy, Mobilization, Screening, Mapping	IEE projects STORIES [16], Meshartility [17], BEAST [18], SEAPs, SECAPs
P3	2017 – 2030	Experimentation with top – down and bottom – up frameworks, ICT expansion	Smart Islands Declaration [2], Clean energy for EU islands [1], Horizon 2020, Pilot islands, ICT/ Digitalization, National Legal framework for Islands
P4	2030 – 2040	Society assisted acceptance and full commercialization (self-sustained) of Smart Island approach	Smart energy, Smart transport, Smart water management, Smart waste management, Smart Governance, Smart economy, Climate change adoption
P5	2040 – 2050	New innovation and business model	Innovation, Business models, Resilience

TABLE I. DATA ABOUT KVARNER ARCHIPELAGO ISLANDS FROM SELECTED STUDIES

Island	**Population**	**Area [m^2]**	**Nr. Of tourists**	**Final Energy Consumption [MWh]**	**Energy Consumption [MWh/cap]**
Krk	19 383	405,8	682 200	190 842	9,846
Cres	3 184	405,8	114 800	54 323	17,06
Lošinj	10 141	74,36	293 400	145 320	14,33
Unije	88	16.9	6 783	1 209	13.741

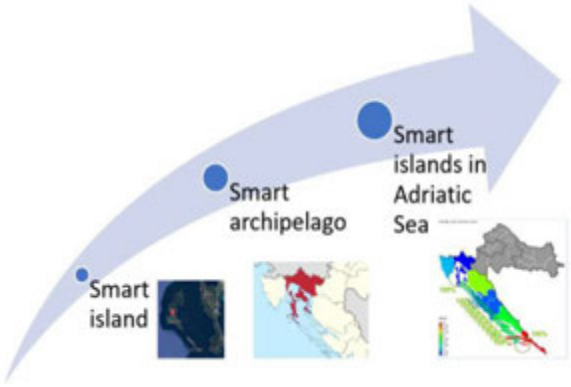


Figure 1. Development of Smart Island in the Adriatic Sea

One of the most important expected results of phase P3 of the presented framework is digitalization. Integration of digital processes is crucial for the development of Smart islands. Key role in this process will have advanced information and communication technology (ICT). Implementation of ICT will provide new management strategies and open new flexibility opportunities. Energy systems are currently going through the digitalization process with a final goal of developing a sustainable and low carbon energy system. This is particularly important for electric power systems since the electrification of different sectors such as transport or heating will play an essential role in achieving this goal. Because of this electric power systems need to become smart grids which will consist of not only smart devices such as smart meters, but that will also have the possibility of automated and remote control. Also, a change of end users' role will be needed. Today's passive consumers will need to be transformed into active components of the energy system or prosumers. Prosumers will be able to consume and produce energy. Ideally, prosumers will be able to satisfy their needs and export the rest of the energy to the electric grid or store it in energy storages. This concept will be beneficial when there will be high demand for electricity in the system and DSO will be able to easily send a signal to activate batteries from prosumers to satisfy the demand. This could lead to a huge amount of distributed energy resources and create concepts known as virtual power plants. Furthermore, this will require that the distribution system operator (DSO) become an active operator like transmission system operator (TSO). Data manipulation will become more important as information collected from smart devices will need to be processed and used properly.

This kind of progress in the energy system will not be possible without adequate communication. Concepts such as the Internet of Things (IoT) where different devices are connected in the network and communicate together will be needed. Such communication will have to be executed through a fast and secure communication network and a 5G network that is already being implemented. 5G network will enable a huge implementation of IoT in every aspect of everyday life providing faster communication able to transfer a huge amount of data in a short time thus creating a more connected world. All data will need to be gathered in a cloud which will process and manipulate the data. Clouds will be organized according to themes defined in [2]:

1. Energy
2. Transport
3. Waste
4. Water
5. Governance
6. Information and communication technologies
7. Economy
8. Education
9. Environment

These clouds will be connected so that they create a synergy between different themes. Special emphasis will be placed on the energy area for an efficient and low carbon energy system is the main issue to deal with to create not only Smart island but also Smart city.

By using currently available top-down studies and also bottom-up studies it is possible to propose a new concept of Smart Island according to Figure 2.

III. RESULTS

According to the presented framework, development of Smart Islands in Croatia is currently in phase P3. During the past twenty years many studies and projects were completed in order to investigate Croatian islands' possibilities for sustainable development. The island that was most detailed investigated is Korčula. In the study [19], the authors presented a case with a 100% intermittent RES for the island of Korčula. The authors investigate which is the best combination of wind and solar power in order the achieve low-cost solution and the combination with the least export and import of electric energy. Figure 3. presents energy planning scenario until 2030. together with energy import and export, RES share in primary energy supply (PES), RES share in total electricity production and CO_2 emissions on the island Korčula.



Figure 2. Combining top-down with a bottom-up approach to develop Smart islands

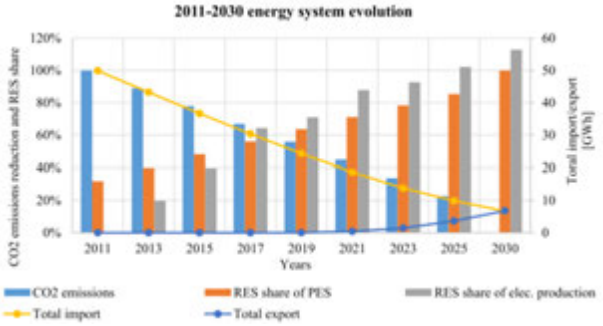


Figure 3. Energy system characteristics of island Korčula by 2030.

The study clearly shows huge potential for achieving low carbon energy system, but in order to balance energy system with 100% intermittent RES, advanced ICT is a necessity that must be implemented or otherwise this kind of energy system would not be possible.

Many ideas and methods were generated during this period such as RenewIslands method. As mentioned, RenewIslands is based on qualitative indicators. During phase P3, this method will be transformed into advanced RenewIslands methods which will be called Smart Islands method. Smart Islands method will consist of more accurate, quantitative indicators for mapping islands need and resources such as energy consumption per capita, load factor, available reserve, and similar indicators. Based on these indicators, Smart Islands method will optimate and suggest several energy planning scenarios that will set the framework for the development of the islands' energy sector.

In addition, some frameworks have been set up and the government passed a law on islands [20]. Nevertheless, real progress is yet expected to be made in this phase of the presented framework. The technological breakthrough which is currently happening will make possible to create the first smart islands in the Adriatic Sea. Islands of Kvarner archipelago are especially interesting to researchers since islands of this archipelago already implemented some advanced sustainable technologies. Chargers for EV are installed on Krk, Cres and Lošinj [21], desalination plant, two photovoltaics systems of 100 kW and a 500 m^2 of solar thermal collectors at a public hospital [22] are installed on Unije and solar power plant was installed on Krk. Krk has also modern public lightning with connected optical cables which open the opportunity for easy implementation of smart monitoring. Further planned projects are solar power plant on Unije and anaerobic digestion plant for biogenic waste, an optical network, and Energy Academy on Krk [22] and the biggest 6,5 MW solar power plant in Croatia on Cres.

Substantial improvement is made in the waste management sector of islands of Kvarner archipelago in comparison with other islands of Adriatic Sea. Waste management possibilities of Cres and Lošinj are presented in [23]. The study presents current waste management problems such as disposal and transfer to the mainland and suggest an integral solution to the problems. The best example of waste management is on island Krk, where 53,6% of waste was separated in 2017. [24]

The number of tourists is especially important for the islands because during their stay need for waste treatment, water, and energy increase substantially. It is possible to notice from Table I. that the number of tourists is several times higher than the islands' population. This can completely exhaust the islands' resources during the summer months. Because of that, islands often have problems with balancing of the electric power system during these periods. Electric vehicles (EVs) are one of the possible solutions to this problem because they can serve as storage solutions. EVs can be in charge or discharge mode depending on power system requirements. Another option for power system balancing improvement is the implementation of demand response technologies. Demand response is ancillary service that enables activation od deactivation of larger loads in the system depending on when the power system requires it. This service can change the daily demand curve and increase the flexibility of the system. Full implementation of demand response will not be possible without advanced technology such as smart meters and advanced ICT. Smart metering systems monitor electricity consumption, measure the quality of electricity and detect faults [25]. Thus, smart meters increase the security of the system and provide prosumers an opportunity to act on market prices in real time. Data collected from smart meters and smart sensors will need to be quickly and securely transferred to the cloud. As mentioned before, the 5G network will play a crucial role in this process together with IoT and cloud computing (Figure 3.).

Advanced technology such as 5G, IoT and cloud computing open new opportunities for integrating different energy flows. A good example of that is the integration of energy, waste and wastewater sector. Similar concept to this is presented in [26] where authors present the methodology for integration of waste, water, and energy sector in Smart city concept. Intermittent RES production brings new insecurities into the energy system. Since waste and wastewater processing consume a substantial amount of energy, by adjusting the time of processing it is possible to increase system flexibility. For example, if the data collected from smart meters and processed in cloud indicates an increase in consumption on the island the system sends the signal to deactivate processing in waste or wastewater plant in order to reduce the total consumption. Another way of integrating these flows is by using residual heat that comes from waste and wastewater plants. Heat from these plants could be used and therefore increase the efficiency of the heating system.

Similar flow integration can be achieved by combining the energy and water sector with desalination plants. Desalination plants are used for separating mineral components from saline water to produce water for irrigation or human consumption. By timing the work of



Figure 3. The synergy between 5G, IoT, and Cloud [27]

desalination plant it is possible to support the energy system same way the waste management plants supports it. Such flow integration was investigated in [28], where the desalination plant was powered by microgrid. Fast communication and transfer of data is achieved with wireless communication and the authors present cost – efficient desalination management with desalination plant participating in demand response market which is possible with the implementation of advanced ICT. Furthermore, since desalination systems increase the possibility of RES integration in energy systems they also have an impact on reducing the production from fossil fuels plants as implicated in [29].

Social acceptance is crucial for implementing changes such as island transition to Smart island. Digitalization opens new opportunities for informing and including people. One of the most influential ways of using digitalization is through social networks which offer a possibility to inform a wide range of people as well as ask them for feedback on any matter. Digitalization enables easier multi-stakeholder collaboration and increases the possibilities for creating a circular economy as emphasized in [30]. It empowers citizens and provides an opportunity for local ownership and participation of citizens in financing projects like solar power plants. Such citizen participation is planned on Krk where local citizens will have the opportunity to become a stakeholder in solar power plant Barbičin. Nominal power of solar power plant Barbičin will be 5 MW and it should be in operation until the end of 2019.

Additionally, with digitalization, it will be possible to make effective education strategies. Islands will be living labs that will secure the knowledge transition and projects implemented on the islands will serve as a starting point for other energy transition projects in cities.

The smart grid will enable the development of new energy market that will be based on peer to peer trading. The technology that will allow this to happen is blockchain and there are already pilot projects that will test this technology such as IMPACT [31]. The blockchain is a technology that stores records or transactions in a growing and continuous list of transactions that are called blocks. Every block is linked to another block and secured with a cryptographic hash key. The electricity trading is done from one peer or producer to another peer or consumer through an encrypted transaction. The encrypted transaction is in the form of a smart contract that is activated when a specific term is satisfied or when there is a signal from the system to execute the transaction. Many benefits can emerge from this kind of electricity markets such as lower market price, better voltage control, improved balancing of a power system and better quality of electric energy. Simplified diagram of this kind of market is shown in Figure 4.

Buyers and sellers submit their offers on blockchain platform while DSO coordinates preservation of power system balance with buyers and sellers and approves transactions on the blockchain platform.

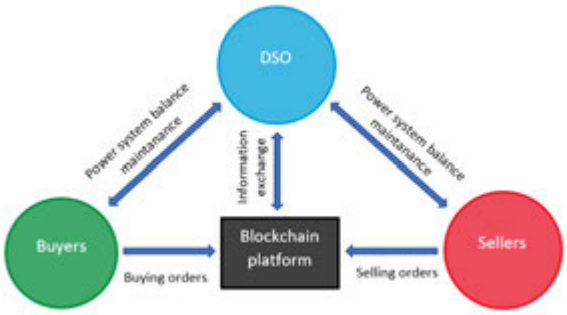


Figure 4. Diagram of the electric energy market based on blockchain

Another blockchain implementation could be insecurity of the data saved in the cloud [32]. As data is becoming a more important factor in decision making and regulation, the significance of data security is one of the most important issues. This is especially true when users don't have access to physical data storage as in cloud computing. Due to encryption and high security, blockchain appears to be a suitable solution to this problem. The problem of blockchain is that it still isn't enough stable technology and therefore, more research will have to be done in order to achieve full implementation of blockchain for the security of the data.

IV. CONCLUSION

Kvarner archipelago represents one of the best places in the Adriatic Sea for the implementation of advance sustainable technology. Framework for developing Smart islands in Croatia suggests that the results of the current phase should be a development of top – down and bottom – up strategies, which is achieved, and implementation of ICT and overall digitalization of existing processes on the islands. Smart Islands Initiative differences nine sectors in which digitalization has to be implemented. The energy sector is the key to developing a low carbon sustainable environment. However, integration of intermittent RES in electric power system brings insecurity in energy system balancing. This paper presents several studies that offer solutions to this problem. Furthermore, this paper suggests several models of digitalization and the integration of different sectors in order to increase energy system flexibility. Combination of IoT with fast and secure 5G network and with cloud computing is crucial for successful digitalization. This combination is applied to all nine sectors with special emphasis on the energy sector. The paper underlines the importance of advanced technologies such as smart meters and sensors and describes the possibilities of

demand response and EVs. These technologies cannot be fully used unless they are connected through advanced ICT. Furthermore, this paper presents a model of future decentralized energy market based on blockchain technology. Decentralized energy market will bring lower electricity prices as well as improved power system control. However, decentralized energy market will not be possible unless advanced ICT such as IoT, 5G, and cloud are implemented because this technology is vital for DSO to preserve the stability of the power system. Islands of Kvarner archipelago are ready for integration of advanced sustainable technology and have already implemented some of the sustainable technology. If implemented, future plans for islands of Kvarner archipelago described in this paper will take these islands significantly closer to completing the task of phase P3 and becoming Smart islands.

Acknowledgments

This work has been financially supported by the Croatian Science Foundation and this support is gratefully acknowledged.

REFERENCES

[1] European Commission, "Clean Energy for EU Islands launch," *European Commission Energy News*, 2017.

[2] Smart Island Initiative, "Smart Islands Declaration," 2017.

[3] N. Duić, G. Krajačić, and M. da Graça Carvalho, "RenewIslands methodology for sustainable energy and resource planning for islands," *Renew. Sustain. Energy Rev.*, vol. 12, no. 4, pp. 1032–1062, 2008.

[4] A. Pfeifer, V. Dobravec, L. Pavlinek, and G. Krajačić, "Integration of renewable energy and demand response technologies in connected island systems – Case study of islands of Vis, Lastovo, Korčula, Mljet and Pelješac."

[5] P. D. Lund, J. Lindgren, J. Mikkola, and J. Salpakari, "Review of energy system flexibility measures to enable high levels of variable renewable electricity," *Renewable and Sustainable Energy Reviews*. 2015.

[6] D. F. Dominkovic, G. Stark, B. M. Hodge, and A. S. Pedersen, "Integrated energy planning with a high share of variable renewable energy sources for a Caribbean Island," *Energies*, 2018.

[7] R. Segurado, M. Costa, N. Duić, and M. G. Carvalho, "Integrated analysis of energy and water supply in islands. Case study of S. Vicente, Cape Verde," *Energy*, vol. 92, pp. 639–648, 2015.

[8] Y. Kuang *et al.*, "A review of renewable energy utilization in islands," *Renewable and Sustainable Energy Reviews*. 2016.

[9] S. E. Bibri, "The IoT for smart sustainable cities of the future: An analytical framework for sensor-based big data applications for environmental sustainability," *Sustain. Cities Soc.*, 2018.

[10] D. Li, J. Cao, and Y. Yao, "Big data in smart cities," *Sci. China Inf. Sci.*, 2015.

[11] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Futur. Gener. Comput. Syst.*, 2013.

[12] igr d.o.o.,igr AG, IInstitut für angewandtes Stoffstrommanagement Umwelt-Campus Birkenfeld, "Interdisciplinarna strategija nulte emisije stakleni kih plinova za integrirani održivi razvoj otoka Krka," Zagreb, 2012.

[13] Sensum, "Akcijski plan energetski održivog razvitka grada Cresa." 2016.

[14] B. Hermetek, "Energy system planning on Island Lošinj with use of RenewIslands metodology," Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, 2007.

[15] S. Bratić, "Planning of development for the energy self-sufficient Island Unije," Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, 2011.

[16] Intelligent Energy Europe - European Commission, "STORIES Addressing barriers to storage technologies for increasing the penetration of intermittent energy sources." [Online]. Available: https://ec.europa.eu/energy/intelligent/projects/en/projects/stories.

[17] "MESHARTILITY - Measure and share data with utilities for the Covenant of Mayors." [Online]. Available: http://www.meshartility.eu/en/about-meshartility.

[18] "BEAST - Beyond Energy Action Strategies." [Online]. Available: http://www.beastproject.eu/.

[19] H. Dorotić, B. Doračić, V. Dobravec, T. Pukšec, G. Krajačić, and N. Duić, "Integration of transport and energy sectors in island communities with 100% intermittent renewable energy sources," *Renewable and Sustainable Energy Reviews*. 2019.

[20] Republic of Croatia, "Law on Islands." 2018.

[21] "www.puni.hr." .

[22] DAFNI, "Smart Island Projects and Strategies." .

[23] H. Grofelnik, "Mogućnosti gospodarenja otpadom na otocima Cresu i Lošinju Possibilities of Waste Management on the Islands of Cres and Lošinj," pp. 118–124, 2012.

[24] Ponikve usluga d.o.o., Ponikve voda d.o.o., and Ponikve eko otok Krk d.o.o., "Yearly report 2017.," 2017.

[25] V. Høynes and J. Berntzen, "The second wave of smart meters: a win–win situation for consumers and energy companies." [Online]. Available: https://perspectives.tieto.com/blog/2016/10/the-second-wave-of-smart-meters-a-winwin-situation-for-consumers-and-energy-companies/.

[26] A. Strzelecka, B. Ulanicki, S. Koop, L. Koetsier, K. Van Leeuwen, and R. Elelman, "Integrating Water, Waste, Energy, Transport and ICT Aspects into the Smart City Concept," in *Procedia Engineering*, 2017.

[27] D. Medved, "5G network." 2019.

[28] S. W. Lee, S. Sarp, D. J. Jeon, and J. H. Kim, "Smart water grid: the future water management platform," *Desalin. Water Treat.*, 2015.

[29] A. Corsini and E. Tortora, "The influence of water desalination systems on load levelling of gen-set in small off-grid islands," in *Energy Procedia*, 2017.

[30] M. Antikainen, T. Uusitalo, and P. Kivikytö-Reponen, "Digitalisation as an Enabler of Circular Economy," in *Procedia CIRP*, 2018.

[31] S. Ilak, Perica; Rajšl, Ivan; Zmijarević, Zlatko; Herenčić, Lin; Krajcar, "Decentralised electricity trading in the microgrid Implementation of decentralized Peer-to-Peer Concept for Electricity Trading (P2PCET)." 2018.

[32] E. Gaetani, L. Aniello, R. Baldoni, F. Lombardi, A. Margheri, and V. Sassone, "Blockchain-based database to ensure data integrity in cloud computing environments," in *CEUR Workshop Proceedings*, 2017.

Autonomous mobility and user perception: A Case of City as a Lab in Slovenia

I. Zajc*, R. Sernec**, G. Lenart *** and A. Pucihar***
* IBM Slovenia, Ljubljana, Slovenia
**AV Living Lab, Ljubljana, Slovenia
*** University of Maribor/Faculty of Organizational Sciences, Kranj, Slovenia
iztok.zajc@si.ibm.com, radovan.sernec@avlibinglab.com, gregor.lenart@fov.uni-mb.si, andreja.pucihar@fov.uni-mb.si

***Abstract* - Autonomous vehicles (AV) have the potential to disrupt the entire transport industry. Technology of AV offers many opportunities as for example reduction of road accidents, fewer congestions on the roads, lower number of vehicles while having better utilization. Beside significant contribution of electric vehicles (EV) to cleaner environment, fully AV also bring new social element as they enable mobility for all. In addition, the use of digital technologies in combination with AV introduces a radical and new paradigm in the transportation, where the lines between car ownership, rental, and lease modes are more and more blurred. To explore the potential of AV in smart city context, AV Living Lab has been established on the premises of BTC City. The AV Living lab was created to test and to learn about real-life solutions for implementation of AV. The underlying concept is BTC City as a Lab ecosystem, where the latest advanced technologies, business models, and services are tested with real users, real cars, on real roads over the real interactions in cross-industry environment. We describe a specific use case of AV shuttle driving and its human experience perception, which was evaluated after a large scale pilot demonstration.**

Keywords – autonomous mobility; living lab; city as a lab; case; Slovenia

I. INTRODUCTION

Nowadays, passengers as well as pedestrians safety, are key concerns in the automobile sector. Advanced technologies enabled the development and implementation of passive and active safety systems, and more recently advanced driver-assistance systems (ADAS) in vehicles [1], [2].

Based on the safety systems implemented in vehicles, Society of Automotive Engineers (SAE) classified automated driving on 5 level scale, with Level 5 (full automation), being the most advanced, requiring no human intervention, even in driving in most demanding environments and under any driving scenarios [3]. Currently available automated driving technology implemented in cars can be labeled as Level 2 - partial automation and Level 3 – conditional automation [4]. Level 2 (partial automation) is available in current serial production cars, while Level 3 (conditional automation) is available in small-scale produced vehicles [5]. However, there are also shuttle buses in operation that are already achieving Level 4 (high automation). For example, Navya company provides autonomous, driverless and electric solutions for sustainable transportation [6].

Automated driving is closely related to broader changes going on in transportation systems, shift towards intelligent road solutions (ITS), active communication between vehicles and infrastructure (V2I), and vehicles to other surroundings in general (V2X). Next generation of cellular communications, 5G, will play a significant role in connected vehicle communication (C-V2X), that already started with the ratification of PC5 mode on existing 4G/LTE mobile networks, but will require substantial infrastructure investments [7].

Although fast technological development drives innovation in the field of autonomous vehicles, user acceptance of automated and future autonomous mobility will play a significant role in forthcoming transportation transformation. Users will be the final judges of business models transformation acceptance, i.e., change from ownership to shared, leased or hailed models, and technological advancements acceptance of autonomous driving. Importance of human feedback in early stages of technology transformation is vital to steering the pace of transformation change and pushing new solutions to the benefit of broader society and inclusion of specific populations that are seen as primary first users of automated mobility, e.g., elderly, kids, disabled people, etc. [8], [9]. Existing surveys [10]–[13] show a positive attitude towards automated vehicles. However, (potential) users have exposed different concerns, primarily related to safety, cybersecurity, availability (high cost). In general, (potential) users agree that autonomous driving is the inevitable future trend, but are still unsure when to expect it.

In April 2018, Autonomous Vehicles - AV Living Lab, which spreads across one of the largest shopping centers in Europe - BTC City Ljubljana, Slovenia, has organized autonomous vehicle demonstration event. During the event, participants could drive in 3 different automated cars (advanced Level 2, Level 3, Level 4) and autonomous shuttle bus in an urban environment. The demonstration consisted not only of automated vehicles driving but included also driving simulator and pedestrian safety demonstration. These experiences directly influenced participants' perception of autonomous driving capabilities. In order to reveal participants attitude towards

autonomous driving, we prepared the questionnaire and conducted the survey during the event. Participants represented wider public and ordinary commuters that currently use different means of transportation. These participants represent the potential user group of future mobility autonomous solutions. The survey covered questions on user perception of automated driving, its market readiness and perceived hurdles before wider adoption. Altogether, we gathered 153 fully filled-in questionnaires. The study provides first insights into human perception and readiness of potential users for automated driving solutions in Slovenia. Insights are useful for providers to promote and further design proper tailored solutions and business models for faster implementation of automated vehicles.

II. AUTHONOMOUS VEHICLES LIVING LAB

In a broader sense, a living lab is a technological, socio-economic approach, which is used as a strategy for the adaptation, innovation, and development of new products and services [14], [15]. From the practical perspective, living lab is an environment in which researchers, developers, and users are working together to develop innovative products, services or solutions, in accordance with users' needs, in the shortest time possible, and test it in a real environment [16].

Recently also cities around the globe adopted these concepts and became platforms for innovation economy. City as a Lab is a new and unique concept that goes beyond an internal/external lab environment and is transferred to a whole city ecosystem. Cities are increasingly becoming living laboratories for rapid prototyping, applying novel digital technologies and testing innovations for complex challenges [17].

Following these innovation concepts, AV Living Lab has been created in BTC City in Ljubljana in Slovenia in 2018. BTC City is one of the largest business, leisure and shopping centers of Europe. The BTC City as a Lab provides the physical and technical infrastructure and ecosystem to experiment, develop proof of concepts, demonstrations and pilot operations of innovative new products, applications, services and business models. It provides the physical and virtual environment for a complete BTC City as a Lab with 21 million yearly visitors, 38,000 cars per day, spread over 475,000 m2 with 450 shops and services, and 3000 businesses [18].

The whole smart city ecosystem is composed of physical infrastructure, businesses, people and overlaid with a multitude of data flows, augmented with sensors and actuators on the main physical assets to monitor and manage the operation securely (**Figure 1**). In this all-in-one place, AV Living Lab offers a commercial service of sandbox for the Autonomous Vehicle driving experience, Artificial Intelligence, Cybersecurity, Blockchain with retail applications, Infotainment and Vehicle-to-Anything Communication [18].

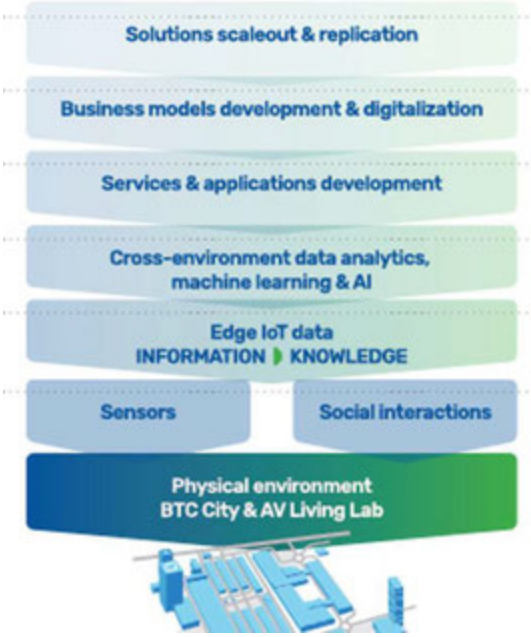


Figure 1. City as a Lab: Layered concept [18]

The AV Living Lab offers infrastructure to move solutions (or prototypes) from technology readiness level, validated in laboratory (TRL4) into relevant real environment validation (TRL5) or demonstration (TRL6) within a smart city. Beyond that, it provides infrastructure for long-term operational demonstrations in interactions with other cross-industries (TRL7). Complete system qualification in a real smart city environment (TRL8) enables testing of solutions for the next step towards the market (TRL9). This one-stop ecosystem environment approach shortens the development and deployment cycle and enables faster entrance to the market with more robust and already tested solutions. It also provides an intelligent showcase for analyzed use cases that can be readily demonstrated to potential customers.

III. METHODOLOGY

In April 2018, TEN-T Days event was co-organized by the European Commission, BTC and AV Living Lab [19]. The main purpose of the event was to provide autonomous vehicle demonstration.

Awareness building campaign started a month before the start of the event itself and was done through official TEN-T Days web page, newspapers and banners in Ljubljana to attract the broadest possible audience. Several vendors demonstrated automated driving capabilities, mobility, and future of transportation. Advanced Level 2 and Level 3 cars were presented: Mercedes Benz, BMW, Magna, Audi, as well as Level 3/4 bus shuttle demonstration was done by Navya (Figure 2). All vendors provided competent technical teams that offered questions and answers sessions to individual visitors and demonstration participants. AMZS organized special pedestrian safety demonstration with Level 3 car and also provided professional technical competence on site [20]. NervTech presented the working demo of the driving simulator [21]. The high technical level and marketing

Identify applicable sponsor/s here. If no sponsors, delete this text box.

exposure of event assured large, diverse group of visitors, which actively participated in demonstrations. Many engaged participants also provided high-quality feedback about using AV.



Figure 2. Automated shuttle bus, Navya driving during survey event

During the event we also conducted the survey with participants. The survey consisted of 41 questions and covered user perception of automated driving, its market readiness and perceived hurdles before wider adoption. Respondents were surveyed immediately after automated driving or demonstration experience. This enabled us to capture realistic perceptions of autonomous vehicles. Respondents represented ordinary commuters that currently use many means of transportation. Therefore, we can assume, they presented a good fit to the future users of autonomous mobility solutions.

Out of 1.100 participants in the event, 153 participated in the survey. Out of 153 participants, 39% were female and 61% were male. The respondents can be divided into four age groups with the following distribution: 26,1% respondents fall in group 18-24 years, 30,1% respondents in the group of 25-40 years, 36,6% in group 41-65 and 7,2% respondent in the group of 65 years or more. Most of the participants are employed, the second largest group of respondents were students, and only 2,6% of respondents were unemployed. Figure 3 presents the demographic data of surveyed participants.

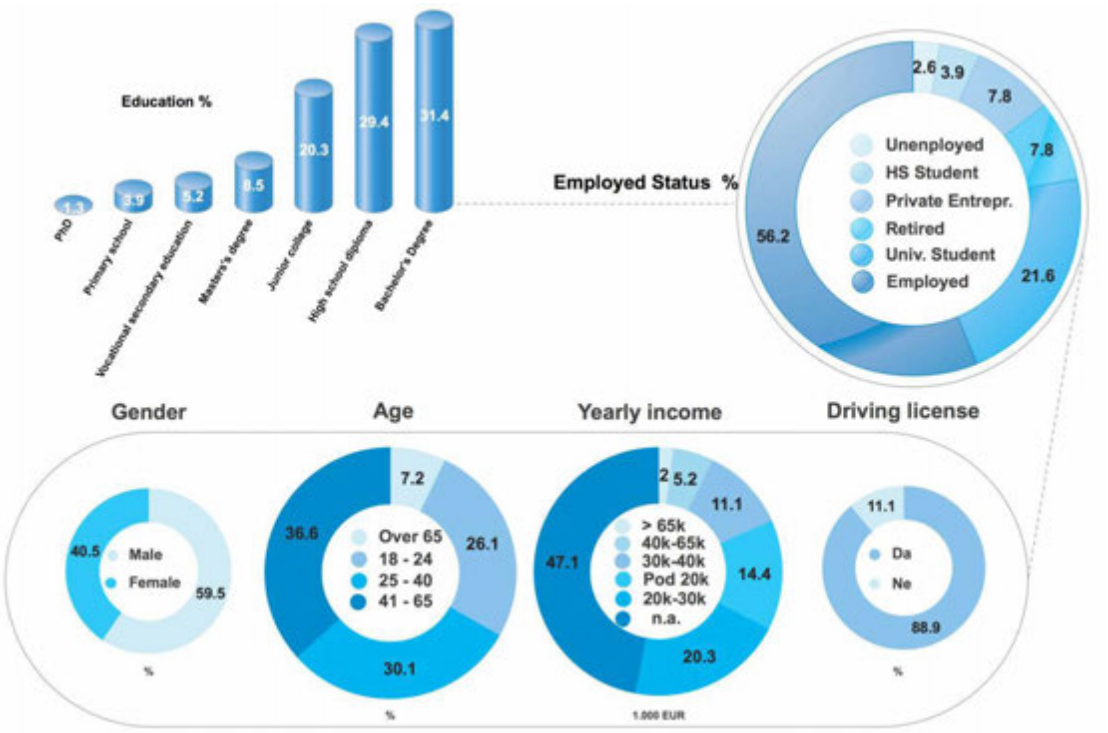


Figure 3. Survey population characteristics

IV. RESULTS

The questionnaire contained three parts. The first and most crucial part was focused on user perception on autonomous mobility acceptance. The second part of the questionnaire was focused on patterns of autonomous vehicle usage and what users would like to do in the time of autonomous driving. Last part of the questionnaire focused on demographic data of respondent.

The first question focused on the means of transportation used by participants in their daily life. There were seven different means of transportation listed in the questionnaire: car, walking, bicycle, public transit, taxi, car renting and car sharing. Participants were asked to mark the frequency of the transportation use on the scale from 1 to 5, where 1=Never, 2=Occasionally, 3=Monthly, 4=Weekly, and 5= Daily.

Results showed that most convenient personal transportation mean of respondents is a private car with the average mark of 4,5 (out of 5), walking is the second

most frequently used means with average of 3,7, bicycle is the 3rd most commonly used option with average of 2,8, public transportation represented with average value of 2,5, and Taxi with 1,9. The least used means of transportation were car sharing with 1,8 and car rental as last marked an average of 1,5 out of 5 (Figure 4).

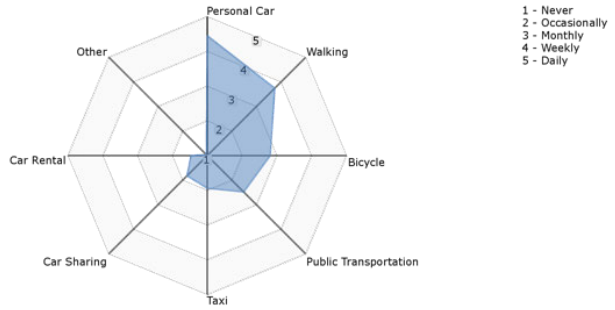


Figure 4. Means of transportation used by participants

Respondents were asked to mark their agreement about the challenges of autonomous vehicles, which need to be solved in advance before possible wider deployment. They were asked to score their agreements on the scale from 1 – strongly disagree to 5 – strongly agree. The biggest perceived challenge among respondents for autonomous vehicle implementation is legislation which was marked with as 3,9 out of 5. Cybersecurity was the second most represented challenge with score 3,7. Maintaining autonomous vehicles maps was marked with 3,2 and Weather conditions with 3,1. Participants are less worried about Artificial Intelligence capabilities. Participants expressed their opinion on readiness of autonomous vehicles for massive use with an agreement level of 2,8 out of 5 (Figure 5).

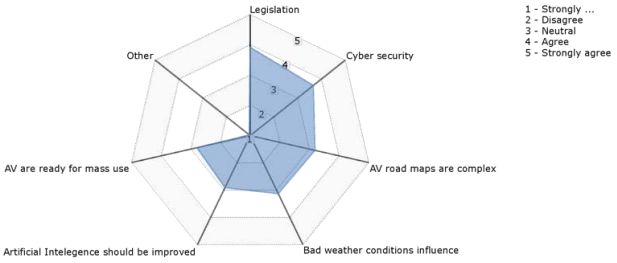


Figure 5. AV issues to solve before massive deployment

Respondents were asked for what kind of daily activities they would use autonomous vehicles. They were asked to express their opinion on the scale 1 (meaning definitely not) to 5 (meaning definitely). Participants would use autonomous vehicles for personal use (3,9), transportation to work (3.8), delivery of goods and business transportation (3,7), for fun (3,6) and pleasure trip (3,5) (Figure 6).

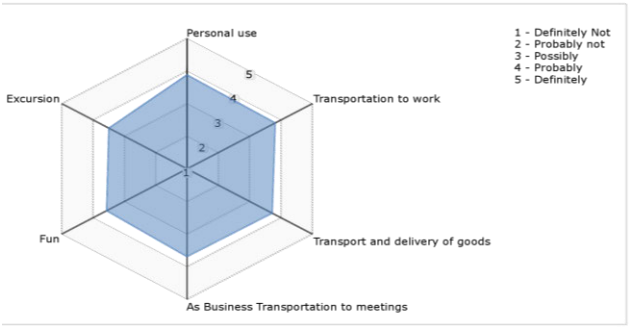


Figure 6. Usage patterns of AV

Participants were also encouraged to express their ideas on what they would be doing during the time of driving in autonomous vehicles. The purpose of this question was to possibly gain new insights for future services, which could be delivered during the time of autonomous driving. Participants quoted the following activities: listening to the music, resting, working, watching movies and education (Figure 7).

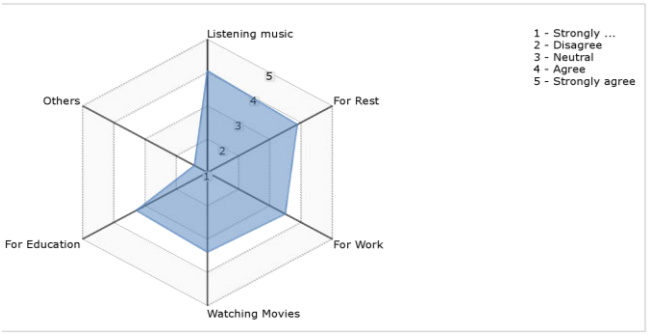


Figure 7. AV commute activities and services

V. DISCUSSION

The paper presents the results of the study on user perception of autonomous driving. The survey was conducted on TEN-T days, where four automated vehicles, driving simulator and pedestrian safety were introduced. During the surveying, a comprehensive technical, Q&A support was offered to all event participants. Survey captured immediate feelings of respondents since it was taken on the event itself after participants were involved in demonstrations.

Results showed that most convenient personal transportation is a private car, following by walking, bicycling and public transit. It seems that participants are not yet ready for new transportation business models as for example car sharing.

As the most significant challenge that needs to be resolved before massive deployment of autonomous driving participants perceive legislation, following by cybersecurity, and maintaining of autonomous vehicles maps. They also perceive weather conditions as an important risk factor. It seems that survey participants are not worried about AV artificial intelligence capabilities, nevertheless some serious accidents already happened on the public roads during AV pilot and experimental operations around the world. These opinions might be related to lack of knowledge and experiences with

autonomous vehicles on the roads. It is interesting that participants with driving licenses are more worried about cybersecurity issues than participants without driving licenses.

Participants expressed that they would use autonomous vehicles for several occasions as for personal use, for transportation to work, delivery of goods, business transportation, for fun and pleasure trip. Participants also marked possible activities during driving in autonomous vehicle. Most of the participants revealed the following activities: listening to the music, resting, work, watching movies, and education.

VI. CONCLUSIONS

This study offered first insights into the opinion of general public (which had a chance to test autonomous vehicles) about autonomous driving in Slovenia. It only reflects general opinions and attitudes towards the use of autonomous vehicles. It seems that participants in the study are relatively keen for autonomous driving; however, they also expose some of the concerns. From the results, it is also evident, that participants are not experienced with autonomous driving and they lack of general knowledge about technologies used for safe autonomous driving – artificial intelligence. The current participants' opinion on readiness of autonomous vehicles for adoption is neutral. For massive and safe adoption of autonomous vehicles, much more has to be done concerning the legislation framework, AV technology development, infrastructure adjustment (i.e. electric grids, road signs, etc.) and service provisioning as well as awareness creation and education of general public about autonomous vehicles. AV Living Lab with its capabilities will play a significant role in all of these activities.

ACKNOWLEDGMENT

Authors would like to thank organizers for comprehensive support during the demonstration event that significantly contributed to the quality of the survey. In particular we express gratitude to teams from BTC d.d., AMZS for building awareness on driving safety, NervTech for delivering driving simulator and Termodron d.o.o. for providing the drone, camera and filming of survey and demonstrations.

The survey was designed during master thesis preparation of 1st author of this paper Iztok Zajc under the supervision of Andreja Pucihar.

REFERENCES

[1] European Road Safety Observatory, "Advance driver assistance systems," 2016.

[2] A. Kumar, "Active and passive automotive safety systems," 2017. [Online]. Available: https://automotive.electronicspecifier.com/safety/active-and-passive-automotive-safety-systems. [Accessed: 12-Feb-2019].

[3] SAE, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," 2018. [Online]. Available: https://www.sae.org/standards/content/j3016_201806/. [Accessed: 21-Feb-2019].

[4] L. Guvenc, A. G. Bilin, and T. E. Mumin, "Connected and autonomous vehicles," in Internet of Things and Data Analytics Handbook, G. Hwaiyu, Ed. 2016, pp. 581–595.

[5] M. Taylor, "The Level 3 Audi A8 Will Almost Be The Most Important Car In The World," 2017. [Online]. Available: https://www.forbes.com/sites/michaeltaylor/2017/09/10/tthe-level-3-audi-a8-will-almost-be-the-most-important-car-in-the-world/#38977d82fb3d. [Accessed: 02-Nov-2019].

[6] Navya, "New mobility is here. Autonomous, shared and electric." [Online]. Available: https://navya.tech/en/. [Accessed: 11-Feb-2019].

[7] Qualcomm, "Let's set the record straight on C-V2X." [Online]. Available: https://www.qualcomm.com/news/onq/2018/04/25/lets-set-record-straight-c-v2x. [Accessed: 11-Feb-2019].

[8] J. Meyer, H. Becker, P. M. Bösch, and K. W. Axhausen, "Autonomous vehicles: The next jump in accessibilities?," Res. Transp. Econ., vol. 62, pp. 80–91, Jun. 2017.

[9] J. M. Anderson, N. Kalra, K. D. Stanley, P. Sorensen, C. Samaras, and O. A. Oluwatola, Autonomous vehicle technology - A guide for policymakers. Santa Monica, CA. RAND Corporation, 2016., 2016.

[10] S. Nordhoff, J. De Winter, M. Kyriakidis, B. Van Arem, and R. Happee, "Acceptance of Driverless Vehicles: Results from a Large Cross-National Questionnaire Study," J. Adv. Transp., vol. 2018, no. April, 2018.

[11] L. M. Hulse, H. Xie, and E. R. Galea, "Perceptions of autonomous vehicles: Relationships with road users, risk, gender and age," Saf. Sci., vol. 102, pp. 1–13, Feb. 2018.

[12] P. Bansal, K. M. Kockelman, and A. Singh, "Assessing public opinions of and interest in new vehicle technologies: An Austin perspective," Transp. Res. Part C Emerg. Technol., vol. 67, pp. 1–14, 2016.

[13] H. Abraham, B. Reimer, B. Seppelt, C. Fitzgerald, B. Mehler, and J. F. Coughlin, "Consumer Interest in Automation: Preliminary Observations Exploring a Year's Change," 2017.

[14] C. Liedtke, M. Jolanta Welfens, H. Rohn, and J. Nordmann, "LIVING LAB: User-driven innovation for sustainability," Int. J. Sustain. High. Educ., 2012.

[15] H. Schaffers, S. Budweg, R. Ruland, and K. Kristensen, "Collaborative environments to support professional communities: A living lab approach," in IFIP Advances in Information and Communication Technology, 2009.

[16] A. Følstad, "Living Labs for Innovation and Development od Information and Communication Technology: A Literature Review," Electron. J. Virtual Organ. Networks, 2008.

[17] B. Cohen, E. Almirall, and H. Chesbrough, "The city as a lab: Open innovation meets the collaborative economy," Calif. Manage. Rev., vol. 59, no. 1, pp. 5–13, 2016.

[18] AV Living Lab, "No Title." [Online]. Available: http://avlivinglab.com. [Accessed: 14-Feb-2019].

[19] European Commission, "TEN-T Days," 2018. [Online]. Available: https://www.tentdays.eu/2018/demo.html. [Accessed: 02-Nov-2019].

[20] AMZS, "Pripravljeni na prihodnost? Zaupate čipom in procesorjem?," 2018. [Online]. Available: https://www.amzs.si/motorevija/mobilnost/reportaze/2018-04-26-pripravljeni-na-prihodnost-zaupate-cipom-in-procesorjem. [Accessed: 02-Nov-2019].

[21] Nervtech, "The future of road safety and autonomous AI." [Online]. Available: https://www.nervtech.com/. [Accessed: 11-Feb-2019].

The European Infrastructure Simulation and Analysis Centre (EISAC) initiative and its technological assets

V. Rosato*, A. Tofani*, A. Di Pietro*, M. Pollino*, S. Giovinazzi*, Luisa Lavalle*, Gregorio D'Agostino*
*Enea, Laboratory for the Analysis and Protection of Critical Infrastructures (APIC), Rome, Italy
vittorio.rosato@enea.it

***Abstract* - The European Infrastructure Simulation and Analysis Centre (EISAC) initiative has been launched in the FP7 CIPRNet project to empower the Critical Infrastructure Protection (CIP) program aiming at supporting Public Authorities and Infrastructure Operators. The Italian node EISAC.it relies on several technological systems: among them, a new Decision Support System called CIPCast enabling risk forecast due to natural events. CIPCast provides Operators and other Emergency Management players of a crisis assessment, from the identification of disrupted assets (e.g., buildings, electrical networks, pipelines, cables, roads) to the measure of the expected impact of services outages on citizens. The CIPCast system is described with its available functionalities and used to analyze synthetic critical scenarios in the city of Roma, related to flooding and earthquake.**

Keywords – critical infrastructure protection; natural events; impact;

I. INTRODUCTION

Critical Infrastructure (CI) are the set the assets, systems, and networks vital for the functioning of a society and insuring citizen's wellbeing and the industrial and economical development. In order to achieve a better CI protection and to enhance their resilience, several international initiatives have been launched, within the H2020 program at EU scale and by other similar national initiatives to support this societal program of increasing a "systemic" protection level of CI recognizing the fact that they constitute a unique, large system of interdependent assets which, moreover, constitute an almost unique "system of systems" covering all the EU area [1].

In US, the National Infrastructure Simulation and Analysis Center (NISAC) is a modeling, simulation, and analysis program within the Department of Homeland Security (DHS) acting as a source of national expertise to address CI protection, providing analysis, research and risk forecast. NISAC focuses mainly on interdependencies, vulnerabilities, and consequences of infrastructure disruptions across all 16 CI sectors at national, regional, and local levels.

With a similar aim, the European Infrastructure Simulation and Analysis Centre (EISAC) initiative has been launched through the FP7 CIPRNet project. The

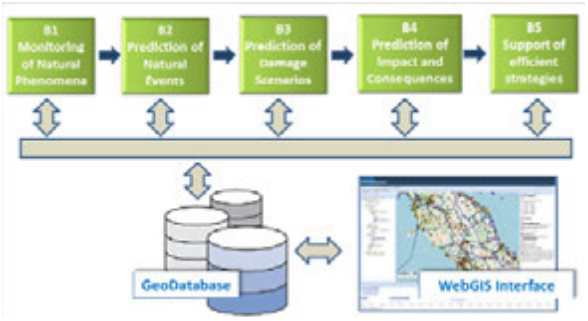


Figure 1. Magnetization as a function of applied field. Note how the caption is centered in the column

Italian node of the European Infrastructure Simulation and Analysis Centre (EISAC.it) is the first implementation of the EISAC initiative. It aims at creating an Italian Competence Center committed in the support of Public Administration and CI Operators in the protection of assets, by deploying advanced technologies for risk analysis and forecast. EISAC groups competences of two main Italian research institutions: ENEA, the national Agency for New Technology, Energy and Sustainable Economic Development and INGV, the national Institute for Geophysics and Volcanology.

EISAC.it is going to deploy its technological systems for supporting the resilience enhancement of strategic national networks, with respect to perturbations coming from different natural hazards (earthquakes, flooding, landslides, which constitute constant threats for the Italian territory). EISAC.it, moreover, is able to deploy its tools also for "scenario analysis" related to man-induced hazards. The tool, belonging to the class of Decision Support Systems (DSS), can also be used to support the design of "educated" Emergency Plans (i.e. designed on the bases of the analysis of synthetic scenarios representative of specific class of expected perturbations) needed to design prevention and mitigation strategies in "good times" (e.g. as part of business as usual maintenance operations), to produce more effective Emergency Management operations and to identify "build-back-better" strategies in the aftermath. A further goal of EISAC.it is to increase the awareness among CI operators, public authorities and the local communities on the importance of investing and acting towards a resilience enhancement of CI.

II. CIPCAST

The DSS CIPCast is the main tool deployed for CI protection. It allows a risk forecast of CI in an area and is devoted, in its operational (24/7) mode of functioning, to support CI Operations. Figure 1 shows the main components (B_i) of CIPCast, which are detailed in the following:

- B_1 Data Acquisition – CIPCast gathers external data from many different sources (e.g., seismic data, nowcasting, weather forecast, real time data from land sensing stations and remote sensing instruments), in order to establish the current "environmental" conditions; the CIPCast DataBase also collects the geo-localized data of all elements of CI networks;
- B_2 Prediction of Natural Events – all collected data in B_1 are used to forecast the incoming natural events (or detect their occurrence when unpredictable, such as earthquakes), their localization and strength, for predicting the perturbation extension;
- B_3 Prediction of Damage Scenario - it correlates the strength of the expected hazard with the vulnerability of the different CI components located in the affected areas. Knowing the localization of CI elements and the localization of the expected perturbation, CIPCast predicts the "Damage Scenario" (i.e. the map of all the CI components having a high probability of being disrupted or severely damaged by the expected event(s):
- B_4 Prediction of Impacts and Consequences – this module combines: (i) the expected damages of the CI components (e.g., the disruption of a transformer in an electric substation); (ii) the resulting topological perturbation to the networks and (iii) the network (inter-) dependencies in order to estimate the degradation of the service(s) level provided by the network(s) (e.g., an electric outage caused by a failed substation);
- B_5 Support to the identification of Efficient Strategies – this module is devoted to support CI Operators in the decision-making process (e.g., to provide multiple strategies to manage crisis scenarios).

CIPCast DSS, along with all its innovative internal technologies, is devoted to provide EISAC.it capabilities to promoting effective behavior change and shared learning and to increasing the awareness of the importance of CI protection and resilience enhancement, by providing, among others:

- a user-friendly and interoperable platform conceived as a combination of free/open source software environments including GIS features (Figure 2);
- an interoperable database (including CIs, hazards, strategical buildings and points of interests for the communities among other geo-localized data) accessible by different kind of users with different level of protection of the data there included;

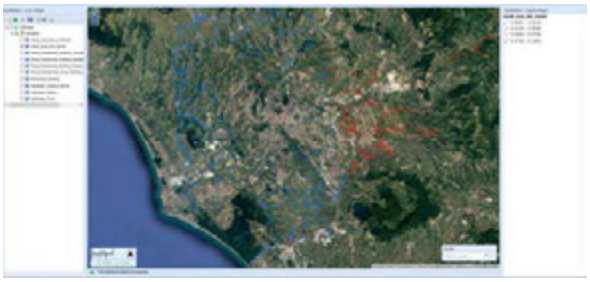


Figure 2. Visualization of some networks elements in a given geographical area. Behing the graphical (webGIS) representation, the CIPCast DB contains a large amount of data of the CI Networks, directly provided by the networks Operators.

- real time and operational (24/7) monitoring and risk analysis of interdependent critical infrastructures for different kind of natural and man-induced hazards;
- simulation capabilities for both real or user-defined critical events towards the evaluation of: the induced damage to buildings, the impacts on the delivery of primary services (i.e. electricity, water, mobility), social and economic impacts on the affected communities. These simulations can be run at different level of detail and can support, among others the planning of mitigation strategies in co-creation mode, the experiential learning, the planning of drills, that of emergency management before and after the event, the selection and testing (with cost-benefit or multi-criteria analysis) of post-disaster "build back better" reconstruction strategies.

A. Earthquake simulation

CIPCast Earthquake Simulation module (called CIPCast-ES) has been developed as the CIPcast element able to assess the earthquake-induced damage at single building level, and the relative expected consequences on citizens in term of prediction of the number of casualties and population to be evacuated in the aftermath of a seismic event. CIPCast-ES works on a deterministic basis, simulating damage and impact scenarios for earthquake whose data are input by the end-user or input by the B_1 automatic data input (in terms of epicenter localization, magnitude, hypocentral depth).

CIPCast-ES estimates the earthquake geographical propagation in terms of local macroseismic intensity *I* providing a qualitative description of the seismicity in relation to the damage observed on the built environment [3]. The model allows the use of different Ground Motion Prediction Equations (GMPE) to describe the damping of the ground motion as a function of the distance from the fault rupture [4][5]. Moreover, CIPCast-ES allows the assessment of the road vulnerability to ground failure [2] by estimating the possible invasion of the debris of collapsed buildings on roads with the subsequent reduction of their available width (Figure 3).

With respect to the damage estimate of CI network elements, to date CIPCast-ES can also estimate the damage to the Distribution electrical network elements



Figure 3. Road functionality assessment due to earthquake.

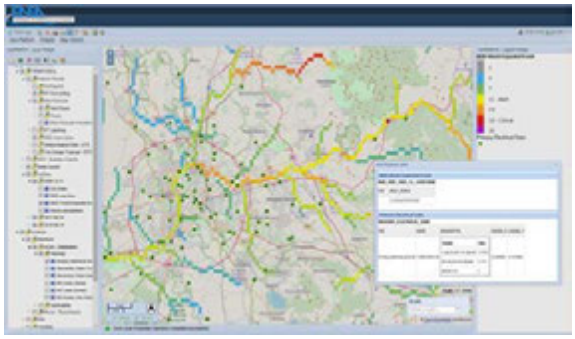


Figure 4. Example of BDD index map (flood alert) related to eletrical substations (distribution network).

(substations, cables) and to the Water distribution networks (pipeline) due to earthquakes [6], and the consequent increase of Repair Rate.

B. Flooding Risk assessment on power networks

CIPCast allows 24/7 risk assessment of the distribution electrical grid of Rome due to meteorological events. This feature relies on hazard data provided by CHyM (CETEMPS Hydrological Model), a distributed grid-based hydrological model for drainage network extraction and rainfall data assimilation [7]. In particular, CIPCast exploits two specific indices: BDD (Best Discharge-based Drainage Alarm Index) and CAI (CHyM Alarm Index), that provide flood alert mapping and hydrological risk expected on the Tiber basin and on the neighboring basins, and more specifically on their effects.

Such indices of hydrological alarm have been "specialized" to provide information on a set of specific sites that the owner of the electrical distribution network has stated to be of major interest, for the purposes of network operation and service continuity (in particular, several High-to-Medium Voltage substations of the distribution network). For each substation has been identified the closest portion of drainage network; then, taking into account the rainfall forecast and the runoff in the upstream basin, CHyM is able to evaluate if such expected event can be critical for the considered substation(s).

Results are automatically acquired by CIPCast and synthetized in a specific GIS layer in which, for each

1	2	3	4	5	6	7	8	9	10	11	12
76	26	32	84	83	40	49	48	62	43	85	98
0.8	0.4	12.5	1.1	0.7	0.3	0.5	0.6	0.8	0.3	0.2	0.4
14516	16953	101	1047	1049	16631	16563	1766	100	813	1576	116
2.6	1.5	7.2	4.1	3.8	1.8	2.0	2.8	4.2	4.7	4.3	6.0
3.2	1.9	8.6	5.2	4.9	2.3	2.5	4.0	5.1	5.6	5.4	6.8
3.7	2.4	9.2	6.4	6.1	2.8	3.0	5.1	5.5	6.4	6.5	7.3
4.3	2.9	9.5	7.6	7.3	3.4	3.7	6.3	6.0	7.4	7.7	7.7
4.8	3.4	9.8	8.9	8.6	4.1	4.4	7.5	5.8	8.6	9.0	7.7
5.3	3.9	9.8	10.2	9.9	4.7	5.1	8.7	5.5	9.7	10.2	7.1
5.8	4.5	9.2	11.3	11.0	5.5	5.8	9.9	5.2	10.2	11.2	6.2
6.2	5.1	8.0	12.1	11.9	6.2	6.6	11.0	4.8	10.6	11.8	5.4
6.5	5.7	6.7	12.1	12.4	6.9	7.3	12.0	4.2	10.9	12.2	4.7
6.7	6.4	5.6	12.6	12.6	7.6	7.9	12.7	3.8	11.1	12.2	4.1
6.9	7.0	4.7	12.6	12.6	8.2	8.5	13.0	3.4	11.2	12.1	3.7
7.0	7.7	4.1	12.4	12.4	8.7	9.0	13.0	2.9	11.2	11.9	3.2
7.1	8.3	3.4	12.2	12.2	9.1	9.3	12.9	2.6	11.2	11.7	2.8
7.2	8.8	2.8	11.9	12.0	9.4	9.5	12.6	2.3	11.1	11.5	2.4
7.3	9.2	2.5	11.6	11.7	9.6	9.6	12.3	2.0	11.1	11.2	2.1
7.3	9.5	1.8	11.4	11.4	9.6	9.6	11.9	1.7	11.1	11.0	1.7
7.4	9.6	1.5	11.1	11.2	9.6	9.6	11.6	1.5	11.0	10.8	1.4
7.5	9.7	1.3	10.8	10.9	9.6	9.6	11.3	1.3	11.1	10.5	1.2
7.6	9.7	1.2	10.4	10.5	9.6	9.5	10.9	1.2	10.9	10.3	1.0
7.7	9.7	1.1	10.1	10.2	9.6	9.5	10.6	1.1	10.5	10.0	0.8
7.9	9.6	1.1	9.8	9.9	9.5	9.5	10.3	1.1	10.1	9.8	0.8
8.0	9.6	1.4	9.6	9.6	9.5	9.5	9.9	1.0	9.6	9.5	0.7
8.2	9.6	1.7	9.4	9.4	9.5	9.5	9.6	1.0	9.0	9.4	0.7
8.4	9.6	1.9	9.3	9.3	9.6	9.6	9.4	1.2	8.3	9.4	0.8

Figure 5. Example of BDD alarm index table, relaed to specific elemats of the electrical disribution network. Alert values (BDD≥6) are highlighted in yellow; critical values (BDD≥11) are highlighted in red

substation, the temporal evolution of hydrological alerts is reported and mapped (Figure 4).

Alert levels were defined and tuned by simulating several past events, occurred during the last ten years within the area of interest (Lazio Region).

Figure 5 shows and example of table (automatically produced by CIPCast) containing BDD index values related to specific elements of the distribution electric network (primary substations) in case of a severe meteo/hydrological event: alert values (BDD≥6) are highlighted in yellow; critical values (BDD≥11) are highlighted in red.

In such a way, in case of severe weather forecasts, CIPCast provides an effective space-temporal mapping of possible flood alerts for specific elements of a critical infrastructure, allowing early adoption of preparedness measures.

III. CONLUSION

The EISAC initiative has been launched in 2016; it has allowed the creation in Italy of the first EISAC Competence Centre (EISAC.it) whose technologies are going to be deployed in favor of CI Operators and the Public Authorities for supporting protection and resilience enhancement of national and European strategic assets. EISAC.it is boosted by an advanced DSS (CIPCast) which is an operational tool, for risk forecast on CI, which also enables to deal with synthetic scenario to produce stress test and to design appropriate emergency plans. The CIPCast-ES simulator, moreover, enables the damage estimate (on buildings and citizens but also on CI elements) after seismic events; it also estimates the road damages (directly produced by the quakes and by building collapses) and it has revealed particularly useful to design appropriate contingency plans which also consider these event for a better identification of safe buildings in areas which have small risks to be disconnected due to main roads unavailability.

ACKNOWLEDGMENT

Authors acknowledge EU FP7 project CIPRNet (www.ciprnet.eu) and the Italian funded project RoMA (SCN_00064) which have allowed to design and realize most of the currently available functionalities of the CIPCast DSS. The EISAC.it initiative belongs to a wider initiative to create EISAC centers all over Europe; this challenge is the foundational goal of the international Association 2E!SAC whose Chair dr. Erich Rome at IAIS at the Fraunhofer Institute in Bonn is also greatly acknowledged for his continuous support to the CIPCast challenge. 2E!SAC Association can be contacted for further information at erich.rome@iais@fraunhofer.de.

REFERENCES

[1] Council Directive 2008/114/EC of 8 December 2008, http://eur-lex.europa.eu/legalcontent/ EN/ALL/?uri=celex:32008L0114

[2] D'Agostino, G & Di Pietro, Antonio & Giovinazzi, Sonia & La Porta, Luigi & Pollino, Maurizio & Rosato, Vittorio & Tofani, Alberto. (2019). Earthquake Simulation on Urban Areas: Improving Contingency Plans by Damage Assessment: 13th International Conference, CRITIS 2018, Kaunas, Lithuania, September 24-26, 2018.

[3] Matassoni L., Giovinazzi S., Pollino M., Fiaschi A., La Porta L., and Rosato V. A geospatial decision support tool for seismic risk management: Florence (Italy) case study. Gervasi O. et al. (eds) Computational Science and Its Applications ICCSA 2017. ICCSA 2017. Lecture Notes in Computer Science, 10405, 2017.

[4] Mak S, Clements RA, Schorlemmer D (2015) Validating intensity prediction equations for Italy by observations. B Seismol Soc Am 105:2942-2954.

[5] Mak S, Schorlemmer D (2016) Erratum to validating intensity prediction equations for Italy by observations. Bull Seism Soc Am 106:823.

[6] S. Giovinazzi, A. Di Pietro, M. Mei, M. Pollino, V. Rosato, "Protection of Critical Infrastructure in the event of Earthquakes: CIPCast-ES", XVII Conference ANIDIS, Pistoia 2017.

[7] Coppola, E., Tomassetti B., Mariotti L., Verdecchia M., Visconti G., "Cellular Automata algorithms for drainage network extraction and rainfall data assimilation". Hydrological Science Journal: 579–591, 2007

Comparative analysis of the selected practices in the field of the urban logistics of Polish cities

M. Zysińska
* **Motor Transport Institute**
ul. Jagiellońska 80, (03-301) Warsaw, Poland

malgorzata.zysinska@its.waw.pl

Abstract: **Migration to the cities is a trend which is observed in Poland and it is expected to continue in the near future. An increasing level of motoring and freight being carried in most Polish cities is causing serious transport problems. Bottlenecks, delivery obstacles, congestion, traffic jams and air pollution require prompt organisational and technological intervention. The article starts a discussion about the condition of the urban logistics in Poland, comparing various practices existing in the selected Polish cities from the point of view of cities residents. It presents the main preferences, observed on the basis of the research conducted in 2018 by the author. This is supported by the cost analysis as well as various statistics at the national, regional and local level. The author outlines the barriers and obstacles in the planning and implementing of the modern urban logistical solutions. The article includes recommendations for Polish cities' authorities. The author emphasises that there is still a lack of long-term investment plans in the field of city logistics. There is also an urgent need to introduce regulations limiting freight traffic in Polish cities. The paper includes recommendations for creating Urban Vehicle Access Regulation Schemes for Polish cities.**

Keywords*: *city/urban logistics, ITS, smart cities, urban freight in Polish cities, mobility strategies

I. INTRODUCTION

This article seeks to present the results of the author's comparative analysis of selected solutions applied in Poland's urban (city) logistics. Part one discusses the premises for the development of city logistics, its definitions and related basic notions. Part two presents urban-logistics solutions and practices based on selected urban centres in Poland, pointing to their residents' preferences and the importance of individual solutions and their cost-intensity. Conclusions regarding the investment needs in the area of urban logistics are presented from the author's own research conducted among residents and members of local authorities (governments). The present paper is concluded with a summary dealing with the current state of urban logistics in Poland as well as the expectations and challenges related to its development prospects.

TABLE I. URBAN LOGISTICS: OBJECTIVES AND TASKS

	URBAN LOGISTICS		
	Specificity of needs	***Objectives***	***Tasks and effects***
Residents	Convenient location High throughput of roads Quality of nodal and accompanying infrastructure	Budget savings Reduced pollution Convenience/ comfort and reliability Facilitated/ mainstreamed traffic Limited cargo tran sports Safety/security	Reduced maintenance/transportation time Comfort Redundant flows eliminated Comprehensive Reliable and fast information
Local government	Investment projects in line with adopted strategy Appropriate location and quality of infrastructure Reduced pollution	Modern Adapted to agglomeration's requirements Availability 'Going green'	Competitive Capital-attracting Rationalisation of spend Better absorption of external funds
Private sector	Location ensuring efficiency and undisturbed course of economic processes High throughput of roads Low-cost urban infrastructure operation	Low operation costs Reliability of means of transportation Low purchase cost of rolling-stock and tooling/ instruments	Redundant transports eliminated Specialisation

a. Author's own compilation, based on the research conducted at the Motor Transport Institute in 2018

II. URBAN LOGISTICS

A. Premises of development

The dynamic development of the motorisation of urban areas implies a number of problems, particularly regarding transportation. Increasing overcrowding is typical. Congestion is presently the main restraint of the development of agglomerations and it implies a number of side effects that have a bearing on the quality of urban life. The progressing contamination and pollution of the

environment is the other powerful threat posed by the development of urban motorisation. While occupying a mere three per cent of the Earth's surface, urban areas are responsible for emissions of as much as 75% of all greenhouse gases. Moreover, they produce more than 80% of the world's GDP, only one-hundred of the largest cities accounts for 25% of its value (UN statistics, 2018). This data testifies to an extremely strong, yet bipolar, influence of cities on the lives of people. The dynamically developing agglomerations form an enormous economic potential that is encumbered with the most considerable risks and side effects on the environment, health and quality of life of their residents. Among the key methods of counteracting these problems are the smart city concepts – particularly, smart city logistics, smart transportation or smart mobility – under development for more than the last twenty years.

However, the statistics shows that urban realities as far as transportation is concerned are still far from what is expected. The Central Statistical Office [GUS] statistics show that as much as 60 per cent of Poland's population lives in towns or cities. Similar to the developed EU countries, demographic urbanisation has reached and passed its apogee. It is said nowadays that Polish towns/cities are extending/redeveloping in topographic and administrative terms. As the process of expansion by absorbing the surrounding areas intensifies, problems specific to large agglomerations tend to further stratify – particularly those related to transportation, cargo transport included. A growing number of users leads to increased demand for transports of finished products and municipal products, along with (though not as frequently) raw materials and semi-finished goods, and for the logistic services associated with these.

B. Definition, components, associations

Satisfaction of the transportation needs of cities ranks among the major challenges as far as the management of modern urban area is concerned. There are numerous problems concerned, mainly rooted in the lack of compatibility between the city's spatial structures, needs of their users and transport systems functioning within the city's area. There is a number of definitions and methods of discerning the components of modern urban logistics. The characteristic of the city logistics concept is not only a considerable diversity in the scope of relevant notions but also in the parameters of evaluation. It is also difficult to propose a concrete reference model of a benchmark system [1]. City logistics requires that streams of goods and services be connected/associated and controlled for the purposes of supplying, purifying and ensuring internal functionality of the city – this in order to reduce losses of time, redundant transports, overcrowdings and bottlenecks as well as to optimise the use of the available resources [2].

Innovations in city logistics can be various – technological, organisational, process- or image/reputation-oriented and, therefore, have an effect on the behaviours, lifestyle or preferences in the use of transport facilities [3].

Urban logistics subsystems include functional areas such as the control of flows of material goods and persons, collective and individual transport, storage of material goods, transport of material goods together with transmission of utilities, transport and storage of city waste [4].

The literature has proposed a number of definitions of urban logistics encompassing the transport of persons and goods (cf. Caragliu 2001; Krawczyk 2004; Sołtysik 2001; Szołtysek 2005; Stabenau 1993; Klatte 1992; Hesse 1992; Ihde 1991; Wittenbrink 1992; Benjelloun and Crainic 2009). There is also a number of definitions that narrow the notion of urban/city logistics to cargo transport, without taking collective or individual passenger transport (cf. Taniguchi 2003; Thompson 2010; Yamada 2011; Benjelloun 2012; Crainic 2009; Taniguchi and Heijden 2000; Würdemann 1992). However, the definitions of urban/city logistics increasingly more frequently take into account passenger transport, alongside cargo transport (Benjelloun 2012; Tadic 2014; Szołtysek 2005; Kiba-Janiak and Witkowski 2014). This approach is implied by the trends associated with the development of motorisation and their consequences. The problem with compiling a clear definition of smart urban logistics in Poland stems from its connotations that are broadly perceived by society. The notions of smart city or smart city logistics have become en vogue and hence tend to be overused. Numerous initiatives suffer from it, perishing in the information overload.

The optimisation of flows and logistic operations in a city is only possible through managerial actions taken in a planned manner, in line with the needs of the given area. In every town/city within Poland, there are numerous cooperative processes and systems or arrangements on various levels of functioning. Hence, it is difficult to specifically identify urban logistic chains (as a traditional concept); instead, identification, analysis and action needs being applied to extended logistic networks in cities based on broad public participation. In the real-life functioning of Polish cities, however, there appear considerable divergence between the intentions of public- and private-sector stakeholders as well as between the residents. This is shown in Table 1 which summarises the outcome of the author's research carried out at the Motor Transport Institute [ITS] 2018 into the expectations and objectives of various groups of urban logistics stakeholders.

C. Results of comparative analysis of urban logistics solutions applied in Poland

Urban transport represents the major part of passenger transport in Poland in terms of passenger numbers. It exists in about 260 cities and municipalities inhabited by 17 million people. Urban transport enterprises carry in the region of 3.7 billion passengers annually, which represents approximately 80% of the total number of passengers in the public transport sector. In towns and cities, bus services are the main form of urban transport. They are provided mainly by municipal operators operating a total of around 11,500 buses. Tram systems exist in 15 cities with a total number of 3,342 trams running along 2,342 km of tracks. Trolleybus systems operate in three cities: Lublin, Gdynia and Tychy. Together, these cities have 215

trolleybuses and approximately 180 km of traction. The only existing metro system in Warsaw consists of two underground lines with 28 stations, and a total length of 29.2 km. In Warsaw, the share of travel by public transport is nearly 60%, which ranks the capital of Poland quite highly among major European cities [5].

According to the research conducted in Poland's seven largest cities, the working population incurred in 2013 congestion-related cost totalling 3.5 bn per annum, this being PLN 2,905 per capita (Rzepnikowska, 2014). As per the most recent European Environment Agency report, Poland infamously ranks among countries with the most polluted cities [6].

This calls for urgent legislative change, which in particularly requires that documents referred to as 'studies of land use conditions and directions' or, in brief, land use plans [Polish abbr. 'SUiKZP'] that are adopted separately and specifically for each particular town or city.

Such actions are already carried out in certain Polish metropolises. Wroclaw (since 2013 – as part of Wroclaw Mobility Policy and Comprehensive Traffic Research project) and Warsaw (since 2016 – Warsaw Mobility Policy) were the pioneering cities, and have been followed by Rzeszów, Poznań, or Krakow [7]. Success of these actions depends, however, not only on the timely delivery of tasks and on-schedule expenditure of funds allocated to such ends. In addition, it calls for broad public involvement. Expeditious and common access to applications facilitating comfortable movement across downtown zones, with use of means other than private cars, is a must. The inhabitants of large European cities are increasingly less interested in having private cars, which implies a reduced population's motorisation rate. Warsaw and most of the other Polish cities still see a reverse trend, which should incite more resolute actions. The dominance of cars has not declined although investments are made in public transport and as more people travel by bike and on foot. Digitisation, automation and access to new types of electric vehicles have the potential to change the way we travel and move throughout the city.

The author carried out in 2018 research in urban logistics practices among inhabitants of Polish cities and members of their local governments. The study extended to 9.5 per cent of the urban population as at 1 January 2018 and reflected complete representation of cities with district (*powiat*) rights. Questionnaire-based research was based on a stratified sample and multi-criteria analysis, taking into account parameters such as size (in population terms) of urban centre, area, frequency of appearance and population density. Multi-criteria analysis was conducted using the CATI method on a sample of 2,800 inhabitants and 123 members of local governments of a total of eighty-nine Polish cities selected pursuant to the above-specified parameters. Table 2 shows a summary of selected results of the comparative analysis of urban logistics practices received for a group of inhabitants. The table also shows the cost-intensity of individual actions based on questionnaire research conducted among the members of local authorities.

The study has led to the following conclusions:

The results for members of local governments and residents have proved different: the observed discrepancies basically boil down to what is specified in Table 1. For the group of residents, the main objectives sequentially include: savings in the municipal budget, reduced air pollution, availability of advanced applications making the use of urban facilities more efficient, improved quality and access to urban services, facilitated traffic, restricted cargo transports in residential zones and improved safety of traffic. Local governments specified as their priorities improved quality and access to city services – improved air quality being low on their agenda.

Urban logistics projects imply, however, diverse budgetary outlays owing to the type and specificity of intervention. Typically, projects are undertaken in the five basic areas: (I) infrastructure; (II) traffic management; (III) area management; (IV) internal regulations; and (V) outreach/dissemination activities.

Detailed preferences of the inhabitants for all the areas and sub-areas of urban logistics action, specifying the cost-intensities based on the questionnaire research conducted among local government members, are specified in Table 2. All the ratios or rates, apart from those expressed in per cent terms, pertain to the range of values from 0 to 5. The urgency of intervention for a given ratio/rate is aggregately based on its percentage indications and the weight allotted. The practices regarded as the most urgent within the confines of an urban policy correspond with the highest values. Cost-intensity rates have been calculated for the specified criteria based on questionnaire research involving members of local governments. Likewise, the value range of 1 to 5 pertains to them, the high values always standing for a barrier to implement intervention areas regarded as urgent.

According to the representatives of local governments, projects implemented in the areas of infrastructure and traffic management require almost threefold higher expenditure compared to the other three categories (area management, regulatory environment, and outreach or promotion/information). All the same, and paradoxically, scarcity or complete lack of the most necessary solutions in the field of urban logistics, implemented at a relatively low cost, is definitely a weak point of Polish cities. Such solutions would include, for instance, internal regulations on restricted traffic in urban zones, time-windows for deliveries of goods, parking of vehicles, and traffic routes of variable usefulness conditioned by the temporal specificity of transport within the urban area. As shown by the author's research, such solutions are most frequently pointed to by inhabitants as the most demanded ones for the optimisation of the city's logistics.

Based on the research conducted, investment projects of high importance and with high preference rates among residents also include area management activities, along with selected traffic management sub-areas – specifically, adaptive traffic-lights control systems, or dynamic passenger information systems. Subsequently, awareness-raising activities fostering environmental protection, shaping behavioural models in view of a smart city concept and, to a lesser extent, actions taken to improve the safety of traffic have been pointed out.

The least amount of response among the residents (62%), compared to the other areas (69%–78%), concerned transportation infrastructure actions, particularly those related to the upgrade of the public transport fleet. With the relatively highest cost-intensity (4.8) of such activities, recent years in city public transport in Poland have seen large investment projects and purchases of bus or tram rolling stocks. All the same, the total number of passengers has slightly dropped, which might undermine the legitimacy of high outlays on infrastructural projects in the area of transport with no parallel support in the form of organisational and technological improvements.

The area of cities' internal regulations has been recognised as the most important one in respect of weight, frequency of indications among inhabitants and relatively low outlays compared to the other categories of intervention. Consequently, regulatory actions taken to optimise urban logistics may be implemented in most centres, in spite of the budgetary restrictions.

An indispensable solution from the standpoint of local authorities, a fairly low cost-intensive one, is charging fees for entering the downtown area depending on the technical parameters and on how long it stays therein. This is conditioned by the extraordinarily growing motorisation rates in Polish cities – especially in respect of individual motorisation. Due to the high unit cost borne by the user, this does not translate into the popularity of such solutions among residents or transport firms. Local authorities perceive such solutions in much more positive terms (urgency of intervention: 3.9). However, both groups appreciate the importance and efficiency of this particular intervention category in optimisation of urban logistics, as attested by the relatively high weights assigned by the inhabitants (3.2 to 4.8), compared to the other categories of intervention.

The other solutions enjoying high preference in Polish cities include marked-out unloading zones in downtown areas (loading bays); reloading spaces delineated within public spaces of (mainly, underground) multilevel parking lots; parcel lockers fixed close to hubs/transfer points; use of public transport rolling stock to handle transports (intervention urgency: 3.6). Residents and local government appreciate, to an almost equally high degree, the importance of construction of urban consolidation centres modelled after those already functioning in Poznań or Wroclaw (weight 4.8 to 5.0). As such projects are cost-intensive, however, they would not appear on a larger scale as yet. This is confirmed by mean preference rates among local government members and inhabitants (52% and 63%, respectively).

In spite of the apparently easily accessible funds for urban logistics purposes, local authorities tend to encounter difficulties in absorbing such funds. First projects meant to support of smart urban logistics are usually implemented too slowly or are not mutually coherent. Such projects are often not followed up – lack of decision-making continuity in local governments posing a barrier. Owing to political changes, implementation of comprehensive action strategies for 5, 10 or 15 years often proves impossible. Hence, most urban logistic investment projects remain a one-off.

Polish towns/cities suffer particularly from budget restrictions and spatial limitations due to administrative borders of metropolises, or, from multiple managers of transport infrastructure; such limitations/restrictions become problematic at the stage of planning solutions for urban logistics. Once an investment-project decision has been made, strong politicisation and term-of-office model of local governments pose serious difficulties at the implementation stage.

As shown by the research, burdensomeness of delivery trucks/vans and heavy-goods vehicles poses a growing problem in the context of urban congestion in Poland, particularly with respect to the inner centres of cities and their historical sections. In combination with economic laissez-faire and fear of restricting business freedoms in any way whatsoever, uncontrolled traffic of such vehicles within unique urban spaces is all-too-often admitted. The specificity of deliveries to small commercial/services outlets never ceases to pile up problems whose solution cannot be based on a single universal approach. In Poland, one more reason for why urban overcrowding caused by traffic of delivery trucks is steadily growing is the lack of regulations that would limit the operation of vehicles whose allowable total weight is 3.5 t. As far as transport operations in urbanised areas are concerned, the share of high-value loads against the total mass of consignments has seen a dynamic increase. Most of the changes taking place in urban logistics comes from promulgation of so-called 24-hour economy and a dynamic development of e-commerce, which in Poland is also fuelled by the recently-imposed ban on trading on Sundays. One of the effects is the growing dispersion of receipts and senders of cargos. The other experienced consequences include diminished unit weight of cargo, change in cargo structure and problems related to unused loading space. In consequence of regulatory abandonment in this respect, growing congestion arises. Yet another problem contributing to overcrowding is the lack of uniform regulations for the collection of municipal waste in intensely built-up areas and/or in narrow streets.

The increasingly frequent actions taken by local governments in favour of ecological innovations in urban transport deserve a positive response. Poland sees an increasing amount of solutions in the field of transport which facilitate market access for green-fuelled vehicles. In the planning of local-government projects favouring urban transport and logistics, electric fleet is gradually becoming a standard; this is followed by charging station infrastructure facilities. However, it is taxi corporations that continuously act as the primary exponents of fleet change. In addition, logistics operators and courier companies have begun to implement their own solutions in this respect. Poczta Polska [Polish Mail], being one of the leaders in the courier delivery sector in Poland's urban areas, resolved to invest in 2018 in its electric vehicle fleet. At present, a key electric-vehicle (EV) charging station operator in the region of Pomerania plans to provide 200 charging stations for EVs by the end of 2020, including 10 ultra-fast and 135 fast stations. Energy storage facilities will be deployed at places where the need

for charging services is considerable; apart from reducing the waiting time, these facilities will allow to diminish the load on electricity grids at peak times. Although the plans envisioning a million of electricity-powered cars in use in Poland by 2025 seem overly ambitious, it seems obvious that in order to increase the share of EVs in urban vehicle traffic an external regulatory environment – such as fees or restrictions imposed on traditional combustion-engine vehicles – will not suffice. A well-thought-over system of developed EV infrastructure, adapted to the needs and specificities of the particular city, is a must. The idea of extending the EV charging infrastructure in Poland's largest agglomerations has recently been joined by the country's leading fuel companies PKN Orlen and Lotos.

The strife over the diminishment of individual motorisation in favour of broader use of collective public transport facilities is becoming increasingly popular and, therefore, the prioritisation of bicycle traffic, pedestrian traffic or public transport. The proportion of these forms of movement in Polish cities' transportation mix is however still too low compared to a number of European countries, notably Scandinavian ones. In spite of the growing support from local governments for the development of bicycle transport or collective transport, Polish cities have not yet been successful in overcoming the stereotypical patterns of behaviour. Even in centres such as Wroclaw or Warsaw, considered as pioneering in urban cycling solutions, their proportion in the overall transport mix is still low, at 4% and 3%, respectively, whereas the rate for Copenhagen is 36%. However, in the last few years Poland has been witnessing a thriving development of urban bike rental stations. This is reflected in municipal policy papers – more specifically, in the transport strategies or mobility policies. Nextbike Polska is the country's major city bike network operator. With over forty cities covered at present, Nextbike runs rental stations offering more than 15,000 bicycles.

The situation is somewhat better in pedestrian traffic: the figure for Wroclaw and Warsaw is, respectively, 19% and 18% (the average-sized West European city has 28%).

III. Conlusion

Most of the barriers inhibiting the implementation of urban logistics solutions in Poland stems from scarce knowledge on the actual needs and methods, and on their cost-intensity – as has been confirmed by the author's research. The projects that have been least frequently identified by the residents, which have the highest cost-intensity, are, paradoxically, those most frequently launched ones. There still are no long-term investment plans in the area of urban logistics or mobility policies. The first-ever studies and official documents regarding city logistics and smart mobility were adopted in Poland in the years 2009–2010.

On the other hand, the awareness is growing of how important urban logistics is, with its key areas. Contributions have been made by projects carried out in a number of cities – Warsaw, Rzeszów, Wroclaw, Bydgoszcz, Gdansk, Katowice, Gdynia, Toruń, Białystok, Lublin, or Krakow among them. Activities of local authorities that foster going green in city transport, particularly as far as upgrade of the fleet is concerned, deserve appreciation. Complementary regulatory action, low-cost and enjoying the highest preference among inhabitants, remains pretty scarce. Smart traffic control systems and numerous organisational solutions, well-fitting the city's needs and specificities, are not of the least importance to urban logistics.

In adapting the 'best practice' solutions, it is necessary to take into account the large group of prospective users of such systems and identify their influential power – not only at the designing stage but, likewise, during the implementation and, later on, monitoring. It is necessary to broadly recognise the exogenous and endogenous factors whose impact on the functioning of a given urban space is decisive, taking into account the limitations or restrictions, those presently existing and those extrapolated into the future, and the urban centre's development potential.

Some of the urban logistics solutions under analysis cannot be disseminated on a larger scale owing to technical, organisational, or financial constraints; some other ones are gaining in importance and it can be expected that they will be broadly applied in the years to come. The latter include restricted traffic zones, fee collection systems, marked-out traffic lanes – including for heavy-duty vehicles, adaptive traffic-lights control systems as well as cargo tram, train and bike facilities.

The awareness of importance of city logistics has been steadily increasing in Poland, gradually translating into the cities' investment-project priorities – as is reflected in the strategy papers.

The research outcome discussed hereinabove demonstrates that the efficiency of the functioning of Poland's urban logistics network does not necessarily imply cost-intensive infrastructural development. The optimisation of logistics activities should be brought about in urbanised areas primarily based on the organisational changes in the transportation system, customs, habits and behavioural patterns of its users as well as the altered functions of selected. Such an approach is the closest to the views expressed by the inhabitants; however, it all too often proves contradictory to the interests of the other stakeholders of the system. Any changes/alterations to the city's transport layouts or systems, or mobility policies (as a broader concept), call for a broad public contribution when it comes to consulting projects and designing the improvements. Organisational and financial support is likewise indispensable, along with the development of an incentives system.

REFERENCES

[1] M. Zysińska, C. Krysiuk B. Zakrzewski, „Koncepcja smart cities w kontekście rozwoju systemów transportowych" [The concept of smart cities in the context of the development of transport systems"], Logistyka 6/2014, pp. 11972–1197, June 2014.

[2] B. Tundys, „Logistyka miejska, [„City logistics"] pp. 163, Difin, 2008.

[3] Final Report, "Use of Information and Communication Technologies", European Commission, pp. 239-242, 2017.

[4] E.Taniguchi, R.G. Thompson, *City Logistics*, Pergamon Press, pp. 189, 2010.

[5] https://www.intelligenttransport.com/transport-articles/18367/public-transport-in-poland-2/, 1.03.2019.

[6] M. Rzepnikowska, M. Mikołajczak, R. Antczak, „Raport o korkach w 7 największych miastach Polski: Warszawa, Łódź, Wrocław, Kraków, Katowice, Poznań, Gdańsk [„The Report of traffic jams in 7 the largest Polish cities Warszawa, Łódź, Wrocław, Kraków, Katowice, Poznań, Gdańsk"], Deloitte Targeo, pp.37-40, 2014.

[7] „Warszawska Polityka Mobilności 2016", http://www.transport.um.warszawa.pl/sites/default/files/RAPORT%20WPM_1.pdf, pp. 49-50, 22.12.2018.

TABLE II. RESIDENTS' PREFERENCES REGARDING THE DEVELOPMENT OF URBAN LOGISTICS IN POLAND

Area of intervention	Weight (0–5)	Indications (%)	Urgency to intervene (0–5)	Expend-iture (0–5)
I. Infrastructure	**4.1**	**62%**	**2.6**	**4.2**
P&R, P&B, P&K systems Bicycle/single-track vehicle rental stations	5.0	73%	3.6	3.5
Minibuses and shorter buses with higher frequencies	3.9	60%	2.3	3.8
Transport/transfer nodes, stops	3.2	68%	2.2	3.9
Alternative-fuelled buses, trolleybuses	3.8	47%	1.8	4.8
Urban consolidation centres	4.8	63%	3.0	5.0
II. Traffic management	**3.5**	**78%**	**2.8**	**3.9**
Adaptive traffic lights control systems	4.1	84%	3.4	3.9
Dynamic passenger information systems	3.5	78%	2.7	3.7
In-zone integration of tariffs and timetables	2.9	73%	2.1	4.2
III. Area management	**3.9**	**73%**	**2.9**	**2.0**
Bus lanes, traffic lanes for heavy-duty and emergency vehicles	3.5	63%	2.2	2.1
Delivery bays, unloading/reloading sites Public collective transport facilities available for cargo transport	4.3	83%	3.6	1.8
IV. Internal regulations	**4.5**	**70%**	**3.3**	**2.2**
Fees charged for entering downtown area	4.7	54%	2.5	2.6
Traffic completely banned inside town centre	3.2	28%	0.9	1.9
Heavy-duty vehicle traffic banned inside town centre	4.8	88%	4.2	2.2
Restricted traffic zones at peak hours	4.8	92%	4.4	1.4
Time windows for heavy-duty vehicles, night-time deliveries	5.0	89%	4.5	0.7
V. Popularisation	**3.4**	**69%**	**2.3**	**1.2**
Environmental protection-oriented actions	3.1	71%	2.2	1.1
Road-traffic safety actions	3.6	67%	2.4	1.3

a. Author's own compilation, based on the research conducted at the Motor Transport Institute in 2018

Unleashing the Power of Urban Living Labs to Make our Cities Humanly Smart

Francesco Molinari and Božidar Kovačić
University of Rijeka, Department of Informatics, Rijeka, Croatia
mail@francescomolinari.it, bkovacic@inf.uniri.hr

***Abstract* - Modern Cities are experiencing three distinct revolutions: technological, demographic and psycho-social, which all concur to increase their initial levels of Smartness. We argue that the prevailing concept of Smart City can fully deliver its promises (maybe) only in large metropolises with more than 10 million inhabitants. However, less than 9% of global population in 2030 will live in Megalopolises, while 26% and more will reside in cities with fewer than 500.000 inhabitants. There is evidence in the EU that Urban Living Labs hold the power of unleashing the collective creativity and innovation capacity of citizens to provide better valued and more inclusive solutions to urban problems, while at the same time engaging people constructively for the common good. We conclude that the Urban Living Lab approach is a reliable companion to the Smart City concept, particularly in the small to medium sized communities.**

Keywords - Smart Cities, Public Service Co-creation, Urban Living Labs

I. INTRODUCTION

Everybody knows what a Smart City is about. Or at least, one should. The term is clearly an umbrella – not by chance, often accompanied by maturity models [1] and international rankings going more deeply into its multiple facets – that has increased its popularity over the past 5-6 years, as shown in the picture below. The concept is about local government – Town Hall and its various agencies, specialized by topic, e.g. waste or energy management, transport or housing – that modernizes itself using data and technology to improve the quality of urban life for the citizens and businesses that inhabit or visit the city. Thus, it is definitely *evolutionary* (as witnessed by the maturity model examples) as well as *outcome related* (as the many analytical dimensions used by the international rankings are there to demonstrate).

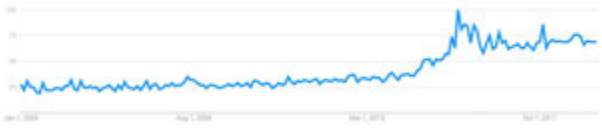


Figure 1. Smart City – Interest over time. Source: *Google Trends*.

Truth to be told, it's not all about data and technology. Or at least, it should not be so. Policy design, stakeholder governance, and particularly citizen related aspects – such as enhancing the creative and innovative capacity of the community and promoting the participation of people (in all their possible roles, from service beneficiaries to social parties) to the realization of Smart City goals – are pretty important too [2]. In fact, two extreme visions confront each other here: the 'technology push', believing in the transformative power of data and IT infrastructures alone, and the 'demand pull', pretending that the only innovation that works is the one that people would love to adopt [3]. A sort of accepted compromise between the two visions is that only by an appropriate merge of 'hard' and 'soft' investments – with the former including sensors, databases and applications and the latter public service reform processes, possibly seeing the active engagement of city dwellers and service users – can local government genuinely think of revitalizing and enhancing the livability of urban settings in an effective and long lasting manner.

Indeed, one can easily agree that the mission of local government includes working to improve the wellbeing of the people inhabiting or using the city. This implies that whatever contributes to fulfilling that mission must be put high on the reform agenda. Unfortunately, the amount of public investment required to implement a Smart City in practice is so huge, and its complexity so high – because of multiplicity, heterogeneity and pervasiveness of target domains, often coupled with an underdeveloped or uneven status of initial automation – that the actual pace at which this (r)evolution can proceed is notably very slow. This is witnessed by the diffuse adoption of the logic of 'pilots': simulated or small-scale environments where the main IT infrastructures are deployed and tested for their potential impacts on public service delivery and urban life quality.

Key advantages of the 'pilot' approach are that it is not binding – the process can be reversed, if anything goes wrong – and it is more compatible with a logic of diffused innovation – both technological and organizational/social – for the involved public service area. In fact, a common feature of most Smart City solutions is their relatively low Technology Readiness Level (TRL) [4] associated with a difficulty to streamline the process of adoption. In other words, the novelty of solutions is so high that some degree of adaptation to local specificities is usually required for successful take-up. This adds further elements of risk to a situation where real people, the users of public service, are affected by these innovations and no major disruption to the continuity of service delivery can be foreseen. In light of the above, it is of little surprise that most Smart City pilots rely on public grants: according to a 2018 research by Deloitte, using European Commission data, only 10% of them were exclusively supported by private finance [5].

Indeed, financial support is an appropriate reaction to a potential market failure: local governments lagging behind

The first author is PhD Scholar at the host University Department.

in adoption of Smart City solutions, due to the risk factors mentioned above, or more generally being reluctant to add data and technology driven innovation to public service delivery, also because of their ailing budgetary conditions, especially manifest over the past decade in the Southern countries of Europe. Thus, we tend to give credit to the results of another research done in 2015 by Cisco Smart Cities Council [6] implying that only 1 local government in 10 moved from the end of a pilot phase to the activation of follow-up projects, not to mention full and permanent deployment of the tested solutions. The other 9 were still engaged in studying and planning (6) or further piloting (3) to verify the possibilities offered by full deployment.

Despite the above caveats, estimates of the Smart City solutions' market size are steadily growing: recently, IDC quantified related investments in $81 billion in 2018 with an expected progress up to $158 billion in 2022. The three largest use cases, attracting nearly one fourth of last year's global spend, were visual surveillance, advanced transport and outdoor lighting. In perspective, application domains such as intelligent traffic management, officer wearables and vehicle-to-everything connectivity are reported to gain momentum and reach critical mass [7].

Little is said about the revenue sources sustaining such growth and expansion; however, considering the low TRL and fast turnover rate of many experimented technologies – some of which compete with each other because of the multiplicity of vendors and consultants – the permanent, 'post-pilot' deployment of Smart City solutions should not be disjoint from extensive ROI calculation and evaluation, which would prove useful even to the local governments having enough financial resources available to invest in their modernization. Prioritizing the investments with the highest financial cashback would allow governments *more of the same* – when deployment can be made modular over time – and/or *more of else*, by reinvesting the returns into other areas of Smart City development.

Unfortunately, assessing the ROI is problematic when investments have a transformative nature and the novelty and context dependence of deployments do not facilitate the retrieval of similar, past use cases to take inspiration and information from. Quite often, benefits are not evenly distributed or not all monetary, touching on such issues as (e.g.) the safety of urban life – see IDC's use case of fixed visual surveillance – or the reduction of air pollution or CO_2 footprint – in transport or lighting cases. Thus, it is not much surprising to read in the aforementioned Cisco research that 1 public official in 4 was concerned about the impossibility of calculating the returns and even the full costs of Smart City investments, that only 1 in 3 had prepared any type of business plan and 1 in 5 some budget allocation for the post-pilot phase.

This paper starts from the situation depicted above to briefly analyze the implications for the Smart City concept of two recently emerged trends in urban population, one demographic and the other psycho-social. Very concisely, the demographic trend shows that despite their worldwide growth, megacities – with 10+ million inhabitants – will still be home to no more than 9% of population by 2030, while 27% will continue to reside in small and medium sized cities and towns, with less than 500.000 inhabitants. Given the big differences between the two typologies and their respective ways of organising urban life and services, it seems quite arbitrary to adopt and recommend a single approach to becoming a Smart City, which ignores this important differentiation.

The second trend is more of psycho-social nature and tells us about the growing appeal to large groups of people of issue-driven and value-laden forms of self organisation, protest, and in many cases active engagement. Channeling this energy into the flow of Smart City investments would add freshness (and more importantly, effectiveness) to the transformations induced by the adoption of technologies. Indeed, capturing people's strength and driving it to the common good is an old dream of many public service innovators. We argue that this is becoming more and more possible by exploiting the methods and tools provided by the Living Lab approach, which has proven particularly appropriate for tackling issues that are geographically well defined and conceptually rooted in the needs of a small sized community.

Our thesis is that these two revolutionary trends should be considered jointly with the growth and proliferation of Smart City innovations reaching the market too slowly, which is where our reflection started from. By so doing, one could possibly give more substance and rationale to the goal of integrating human aspects (alongside policy and data/technology related ones) into the modernization of local government worldwide.

The paper is structured as follows: section II considers the demographic trend, section III the psycho-social trend and section IV introduces the Living Lab approach and its adaptation to urban environments; section V discusses the limits and proposes a revision of the merely 'technology push' vision of Smart City realization. Section VI draws a short conclusion.

II. A DIFFERENT VIEW TO URBAN POPULATION TRENDS

Thinking about the future is not easy and needs special eyes and lenses. When demography comes into play, the United Nations is the acknowledged source of information for everyone willing to know more about future trends and developments. In a 2018 publication, The World's Cities [8], the UN made a projection until 2030 of the expected trends of urban population. While it is not surprising that another 5,1% will move from rural to urban environments, more noteworthy is the forecast that the share of people living in cities with fewer than 500.000 inhabitants will increase by 0,3% - as the following table exhibits – thus keeping the record of being the most populated typology of urban settlement worldwide, with a 26,8% of dwellers, compared to 13,8% in cities with 1 to 5 million inhabitants and a mere 8,8% in large metropolises with 10+ million residents. In absolute terms, the people living in smaller sized cities will still be about 2.3 billion and their increase until the year 2030 is expected to be the highest among all other classes of urban settlement.

Before proceeding, it has to be stressed that when we speak of small sized cities or towns we are not considering rural settlements, but urban agglomerations that – for both administrative and functional reasons – conform to the idea that everyone has of a city, as the UN report clarifies.

TABLE I. VARIATIONS IN WORLD'S POPULATION BY SIZE CLASS OF SETTLEMENT, BETWEEN 2018 AND 2030 (SOURCE: UNITED NATIONS)

Size class	Variations 2018-2030		% of world population in 2030
	Population (mln.)	*% of world population*	
Urban	+947	+5,1	60,4
10 million or more	+223	+1,9	8,8
5 to 10 million	+123	+0,9	5,2
1 to 5 million	+257	+1,7	13,8
500 000 to 1 million	+79	+0,4	5,8
Fewer than 500 000	+266	+0,3	26,8
Rural	-947	-5,1	39,6

Recent academic research, notably within the context of the EU funded ESPON programme [9], has made good progress in conceptualizing the nature, typologies and roles of the so-called 'Small and Medium-sized Towns' to facilitate policy design, planning and innovation for urban growth [10]. Now a similar, yet distinct effort needs to be started in the context of Smart City strategies, to derive options for alternative pathways taking into account the deep diversity of organizational structures, resources and goals that immediately comes to the forefront when trying to compare a city of 10 million with one of 100.000 inhabitants. We will further elaborate on this thesis within section V below.

III. EMERGENT SIGNS OF VITALITY IN INDIVIDUALS AND SOCIETAL GROUPS

Another disruptive trend that characterizes modernity is a growing propensity of people to adhere spontaneously to collective causes and act accordingly. This trend has characterized, with more and more evidence, the new Millennium, but is now acknowledged to be more diffused than within the youngest generations only (the so-called X or Y or Z). Among the fields of engagement where this new attitude is prominent, one can mention the following:

- *Customer engagement*, or an individual's direct or mediated interactive experience with a brand, offering, or organization, materializing into a set of activities that go beyond mere purchase.
- *'Prosumerism'*, or the overlapping roles played by some consumer categories, who also engage with the producer in presiding over critical parts of the value chain, not necessarily to monetize derived benefits, but to complete or refine the delivery in a more satisfactory manner.
- *'Sharing economy'*, or a collaborative/distributed activity holding 'prosumerist' characters, leading to 'beat' the performance of traditional market players (such as taxi services, B&B's, or money lenders) in terms of quality and user experience, usually made possible by dedicated online platforms.
- *'Net-activism'*, or the emergence of new forms of participation that are at the same time, 'virtual' in their being enabled and facilitated by social media and other digital ecosystems, and 'real' in their capacity of attracting, connecting and mobilizing people on very concrete topics, such as gender, biodiversity, or migration.
- *'Social innovation'*, or the development of novel ideas, products, services and business models by the initiative of individual or (more often) groups of people, leading to the fulfilment of societal needs and the creation of links and collaborations that may or may not be supported by Information and Communication Technologies.

At present, the implications of this phenomenon for the future of Smart Cities are neither fully explored by the academic research nor completely understood in practice. On the one hand, a pathbreaking definition of Smart City includes both "investments in human and social capital" and adopting a "participatory government" strategy [11]. On the other hand, we are still way behind a structured inclusion of citizens activism and creativity in the design and implementation of the renewed public services made possible by the data and technology revolution.

However, citizens do not stand still and while public service co-creation remains underdeveloped for the lack of initiative of local governments, they do undertake new and unprecedented actions, complementing public service in a spontaneous and self-organized manner. This is the case of the Social Streets [12] and other comparable examples of urban innovation, which are driven or facilitated by the use of social media and Internet platforms.

IV. URBAN LIVING LABS

A Living Lab is "a working collaboration of Private-Public-People Partnerships, in which stakeholders co-create new products, services, business models or technology applications, within real-life environments, virtual networks, and multi-contextual spheres" [13].

The above is only one of the many definitions that can be retrieved from the academic and industrial design and innovation literature. For instance, [14] characterizes the Living Lab as "user-driven, open innovation environment in real-life settings in which users test and experiment new products or services, in a framework integrating companies, people, research and innovation actors and the public sector" (the so-called PPPP, 'Public-Private-People Partnership'). In turn, [15] define the Living Lab as "an open innovation environment in real-life settings, where user-driven design and experimentation are integrated within a co-creation process of new services, products and societal infrastructures". This sentence points at five key aspects of a Living Lab that are

- An open innovation environment, leading to a full sharing of the IPR (Intellectual Property Rights) of generated results;
- Experimenting in real-life settings, instead of the closed rooms of research laboratories;
- The enhancement of innovation quality and value through adopting co-creation methods and tools

that leverage the creative potential of individual participants in the trials;

- Relying on user engagement, not just involvement as 'guinea pigs', to define the strategic direction and the practical management of the experiments;
- From which new products, services and societal infrastructures are generated.

Ancestors of the Living Lab approach include: the late William J. Mitchell from MIT Medialab, who is said to be the inventor of the term itself, born in an urban planning and architecture design context; Eric Von Hippel, again from MIT, with his elaboration of the 'Lead User'; Henry Chesbrough from Berkeley, recognized by many as the 'father' of Open Innovation; and the Canadian business strategist Don Tapscott, who first introduced the idea of 'Prosumer' as key actor of global markets in the Internet era. However, none of these scientific antecedents alone would suffice to fully grasp the variety and originality of the European Network of Living Labs (ENoLL) – an initiative launched in November 2006 by the Finnish EU Presidency through the so-called Helsinki Manifesto, as a first step towards "a new European R&D and Innovation System, entailing a major paradigm shift for the whole innovation process". In retrospect, and more than 12 years after the ENoLL launch, this step can be safely considered as taken, in view of the hundreds of active members of the Living Labs community, now spanning from Europe to Asia, Africa, Australia, North and South America.

The common trait to all Living Labs – be they part of ENoLL or not – is to embed a Public Private People Partnership (PPPP) that adds one more dimension the familiar Triple Helix scheme, coined years ago to describe the cooperation of Research, Government and Industry within a regional or local innovation system. By adding as fourth stakeholder a (formal or informal) representation of Civil society, Living Labs have moved into a Quadruple Helix scheme, as the picture below displays.

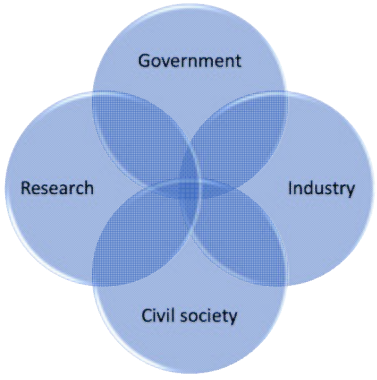


Figure 2. Living Lab – the Quadruple Helix scheme.

Operationally, the 'birth announcement' of a Living Lab coincides with citizens, researchers, policy makers and business persons undertaking some formal or informal agreements, based on which they start to collaborate in user driven, open innovation trials within the thematic domain of election. When such domain points at an urban challenge – such as mobility management, air pollution reduction, or rethinking a green area or built space – and the Quadruple Helix includes stakeholders that belong to a same city or region, it has become commonplace to speak of an Urban Living Lab. This can be compared to a multi-stakeholder governance platform, the members of which congregate to hold conversations and negotiations aimed to achieve goals in terms of e.g. conflict resolution, collective decision-making or concerted action [16].

Although a systematic collection of evidence has not been realized yet, a number of studies highlight both the popularity and the thematic spread of Urban Living Labs across Europe and beyond [17, 18]. However, it is fair to admit that evidence of impacts remains anecdotal [19]. Authors correctly relate these impacts to value generation for trial participants, a better propensity to adopt the solutions that have been co-produced by the citizens, and the possibility for esp. local government to reduce the costs of service (re)design and/or implementation. In most reports, however, success of trial coincides with the delivery of immediate outputs rather than its longer term outcomes, of which little is said. This may reflect the fact that most trials are actually co-funded by the European Union, in coherence with the provisions of the Helsinki Manifesto. Until now, the pragmatic, goal-oriented nature of the various initiatives has impeded the abstraction of a commonly accepted definition of what a Living Lab is – a 2015 research found about 70 of them in literature – and prevented the adoption of standardized or at least more uniform methodological guidelines for Living Lab owners and facilitators (though some of these exist).

Thus, the adoption of Urban Living Labs as innovation governance platforms and ecosystems for faster and more sustainable take-up of Smart City solutions has indeed been proposed [20], but, due to the above and probably other concurrent reasons, has never proved concretely its added value and transformative power for the hosting cities and communities.

V. Discussion and Lessons Learnt

To summarize, the current experience of Smart Cities worldwide presents itself with a number of open issues:

- Public budget limitations, recommending modular or stepwise investments, although with difficulties in estimating the ROI and finding real alternatives to self- or public grant-based financing models;
- Fast pace of technological progress, keeping the TRL of many solutions low and thus forcing the adoption of a narrow 'pilot' logic – never or very seldom reaching full scale;
- A prevalent 'data and technology push' vision, relatively detached from what the citizens could contribute to in terms of public service (re)design and co-production;
- Uneven within-City developments dictated by the 'silos-like' structure of local government, with all-too-technical priorities such as video surveillance, public transport, and street lighting;
- A 'one size fits all' maturity model, making no significant distinction between the Megalopolis and the Small and Medium sized Town, which is

however expected to continue being host to large shares (more than a quarter) of global population.

Truth to be told, academic research has only recently started to explore the structural differences between cities that are determined by or associated with their size. Size is not all, however. Density of population is another relevant variable, as well as the territorial extension a population is distributed on. An empirical study carried out on 70 cities worldwide [21] has proven that the particular combination of priority domains is influenced by a number of context related factors, including demography, environmental and social conditions, as well as the degree of modernization of existing urban infrastructure. Another piece of evidence is the negative correlation between 'harder' and 'softer' investments, confirming the existence of conflicts between 'technology push' and 'demand pull' visions, and thus the difficulty of legitimizing the in-between compromise we were speaking about in the Introduction.

TABLE II. LIKELY EFFECTS OF URBAN CONDITIONS ON THE SUCCESS OF SMART CITY PILOTS (AUTHORS' ELABORATION)

Contextual factor potentially influencing SC take-up	**Differential presence in**		**Expected correlation with SC take-up rates**
	Small and Medium Town/City	***Large Metropolis/ Megalopolis***	
Number of residents in the city	Lower	Higher	Positive
Density of urban population	Lower	Higher	Positive
Extension of the adminstrative area	Lower	Higher	Positive
Critical status of urban environment	Lower	Higher	Positive
Budget autonomy of local government	Lower	Higher	Positive
Global cost of local government	Lower	Higher	Positive
Level of social capital in the city	Higher	Lower	Positive
Inclusiveness of urban governance	Higher	Lower	Positive
Diffusion of social innovators	Higher	Lower	Positive
Sense of belonging to the community	Higher	Lower	Positive

Following this train of logic, the table above tries to differentiate the profile of the Metropolis from the Small and Medium Town according to a number of contextual factors, which can make more or less likely the transition from a Smart City 'pilot' to the permanent deployment of a certain new infrastructure. An important warning to the reader is that taken as such, the two urban profiles are too abstract and therefore unrealistic. Particularly it is only for instrumental reasons – the two size classes of settlement in the United Nations statistics of Table I – that we continue to focus on the segments above 10 million and below half a million inhabitants. When we turn to academic research, the ESPON programme project entitled TOWN has given prominence to the size class between 5.000 and 50.000, with a density between 300 and 1.500 residents/sq.km and that hosts about 27% of the EU population.

Despite the fact that its contents are only exemplary and non-exhaustive, the table sends a clear message in terms of differences between the two city profiles and in broad terms contributes to shaping two distinct models of Smart City development: one based on large sized and IT oriented investments, which is typical of the Megalopolis, and the other being more dependent on user integration and collaboration, thus probably related to public service redesign and co-production, which is more likely to occur in the Small and Medium sized Town or City. We could paraphrase the two models by speaking of a Smart City that grows 'with' and 'without' its citizens.

Of course, reality is far more complex than what is depicted, for at least three reasons:

- Quite often, Smart City 'pilots' are set up in small urban locations such as neighborhoods or districts, which often behave as stand-alone communities in the larger urban community. This contributes to blur the difference between small and large sized population segments and may create situations where some of the contextual factors highlighted in the above table neutralize each other;
- Cities having a number of residents close to half a million still belong to the Small and Medium but hold characteristics that bring them closer to the Large sized ones – particularly in EU countries, where urbanization has historically been far more polycentric than anywhere else in the world. This puts them in the middle of a continuum and makes less easy to predict which set of context factors may actually prevail in them;
- Despite the frequent praise of holistic approaches to shaping a Smart City – instead of progressing at different speeds and with heterogeneous levels of achievement in the various possible specialization areas, such as waste management, urban mobility or housing – even megalopolises do not display a linear evolution pathway across time. This creates gaps and leaps and makes it so that the contextual factors above gain more or less prominence in the specific domain of the 'pilot'.

As a result of these and other complexities, a 2014 study on 240 European cities [22] was unable to identify a clear pattern of correlation between maturity of the Smart City progress model and size of population. However, a piece of evidence was that the largest cities (above half a million inhabitants) tend to be engaged on more fronts at the same time, while the smallest ones focus on fewer – in particular, environment and mobility.

Having given account of its limitations, we continue to think that the proposed duality of Smart City models holds some elements of truth and has some lessons to deliver to urban policy makers, which we summarize here below:

1) Population size matters, though it's not all. How big or small is the community – including neighborhood or district – where a Smart City solution is being piloted, is a relevant factor to consider for the outcome of that experimentation. Data and technology are not the panacea of urban issues, which present themselves differently also in dependence of the number of people involved. Thus, it is no less relevant to think – really holistically this time - of a community as part of the solution, not just of the problem. On the other hand, population is only one aspect

among those listed in the above table, which may have a critical role in determining the transformative power of a Smart City 'pilot' for public service in the community as a whole.

2) People matter, that's all. Though not structured in a systematic fashion, a plethora of Smart City pilots exist showing the importance of 'the human factor' in ensuring both their outputs and outcomes. This importance is just as unquestioned in principle as difficult to document and especially replicate in practice. What is missing in action is not a reusable set of guidelines – which already exist – for profitable user engagement in government reform, but a more nuanced analysis showing not only the advantages but also the risks and the limitations of such an approach in a constructive manner. For instance: everybody agrees in principle on the positivity of a diffuse participation in the political process, but the problems and perils for good decision making of an all too frequent and uncontrolled proliferation of inclusive spaces and opportunities are to a great extent known and considered unmanageable by the policy makers. The result is a limitation in size and scope that does not allow to grasp all potential benefits in terms of quality and inclusiveness of decisions. However, many experiences exist of profitable integration of participation in those stages of decision making where it makes more sense to have it in action. Unfortunately, the reports of these experiences are mixed up with the many others a policy maker would consider inappropriate to replicate.

3) Urban Living Labs as user driven, open innovation governance platforms and ecosystems *are demonstrating their potential for faster and more sustainable take-up of Smart City solutions* in a sizable number of recorded cases. However, their further diffusion and establishment in the European public sector are somehow prevented by a lack of critical assessment studies that leads to similar conclusions as the previously commented example of political participation. Despite that, we think evidence is already enough to revise and upgrade the 'commonly accepted compromise' between the two visions we spoke about in the Introduction, making stable room for Living Lab methods and tools within the 'soft' investments that the 'hard' Smart City vision should bring with it.

VI. CONCLUSION

"If you don't know where you are going, any road will get you there". This quote from Lewis Carroll summarizes the rationale of this paper. With more than one quarter of global population still living in small and medium sized cities and less than 10% in large metropolises, we see the need for a more nuanced approach to increasing the level of smartness of a community, building more on the recent trends, of psycho-social nature, that are characterizing the humankind in this new Millennium. There is room for at least two distinct roads, and the one that integrates Urban Living Labs as structured ways of user and stakeholder engagement in open innovation promises to be effective in solving some of the issues brought about by the pure data and technology driven vision of a Smart City. What we miss is a structured collection of evidence, methods and tools, which would demonstrate not only the potential, but also the capacity of Urban Living Labs to channel the new energies of civil society in a constructive manner.

ACKNOWLEDGMENT

This paper evolves from a keynote delivered by the first author to the 1st International Conference on Future Smart Cities held in Cairo, Egypt, on 28-29 September 2018.

REFERENCES

[1] P. Torrinha and R.J. Machado, "Assessment of maturity models for smart cities supported by maturity model design principles," 2007 IEEE International Conference on Smart Grid and Smart Cities (ICSGSC), Singapore, 2017, pp. 252-256.

[2] L. Berntzen and M.R. Johannessen, "The role of citizens in smart cities," Proceedings of International Conference on Information Management, University of Presov, 2016.

[3] E. M. Rogers, Diffusion of innovations, 5th edition, 2003, New York: Free Press.

[4] M. Heder, "From NASA to EU: the evolution of the TRL scale in public sector innovation," The Innovation Journal, 22 (September 2017), pp. 1–23.

[5] Deloitte Australia, The challenge of paying for smart cities projects, 2018, online: https://content.deloitte.com.au/20180928-cit-inbound-challenge-of-paying-for-smart-cities-2018-registration

[6] Cisco Smart Cities Council, Smart cities financing guide, 2015, online: https://smartcitiescouncil.com/resources/smart-cities-financing-guide

[7] International Data Corporation (IDC) Media Center, 23 July 2018, https://www.idc.com/getdoc.jsp?containerId=prUS44159418

[8] United Nations, The World's Cities in 2018. Data Booklet, 2018.

[9] ESPON Programme, Inspire Policy Making with Territorial Evidence, https://www.espon.eu

[10] L. Servillo, R. Atkinson and A. Hamdouch. "Small and medium-sized towns in Europe: conceptual, methodological and policy issues," Tijdschrift voor Economische en Sociale Geografie, 2017, pp. 1-15.

[11] A. Caragliu, C. Del Bo and P. Nijkamo, "Smart cities in Europe". Procedings of the 3rd Central European Conference in Regional Science (CERS) 2009, Kosice, pp. 49-59.

[12] M. Akhavan, I. Mariotti, L Astolfi and A. Canevari, "Coworking spaces and new social relations: a focus on the Social Streets in Italy," Urban Science, 2019, Vol. 3, No.2, pp. 1–11.

[13] CoreLabs project, "Living labs roadmap 2007-2010: recommendations on networked systems for open user-driven research, development and innovation", 2006.

[14] M. Pallot, "The living lab approach: a user centred open innovation ecosystem", 2009. Online: http://www.cwe-projects.eu/pub/bscw.cgi/715404

[15] R. Santoro and M. Conte, "Living labs in open innovation functional regions," Proceedings of the ICE 1999 Conference, The Hague, 2009.

[16] F. Molinari, "Living labs as multi-stakeholder platforms for the eGovernance of innovation," Proceedings of the ICEGOV11 Conference, Tallinn, 2011.

[17] T. Santonen, L. Creazzo, A. Griffon, Z Bodi and P. Aversano, Cities as living labs – Increasing the impact of investment in the circular economy for sustainable cities, European Commission, DG RTD, 2017.

[18] S. Marvin, H. Bulkeley, L. Mai, K. McCormick and Y.V. Palgan, Urban Living Labs, London: Routledge, 2018.

[19] P. Ballon, M. Van Hoed and D. Schuurman, "The effectiveness of involving users in digital innovation: Measuring the impact of living labs," Telematics and Informatics Vol. 35, 2018, pp. 1201-1214.

[20] F. Bifulco, M. Tregua and C.C. Amitrano, "Co-governing smart cities through living labs. Top evidences from EU," Transylvanian Review of Administrative Sciences, Vol. 50, 2017, pp. 21-37.

[21] P. Neirotti, A. De Marco, A.C. Cagliano, G. Mangano and F. Scorrano, "Current trends in smart city initiatives: Some stylised facts," Cities Vol. 38, 2014, pp. 25-26.

[22] European Parliament, Mapping Smart Cities in the EU, 2014.

The Potential of Mobile Energy Storage in Microgrids

H. H. Abdeltawab, *member, IEEE*, Y. A. I. Mohamed, Senior *Member, IEEE*

***Abstract*—A mobile (transportable) energy storage system (MESS) can provide various services in distribution systems including load levelling, peak shaving, reactive power support, renewable energy integration and transmission deferral. Unlike stationary energy storage units, mobile energy storage can move between different buses by truck to provide different local services within the distribution feeder. This work proposes a day-ahead energy management system (EMS) for a MESS that aims to minimize the cost of the power imported from the grid. The MESS does not only shift renewable energy power to load peak-hours but also can provide localized reactive power support. Given the day-ahead predictions, the EMS decided the optimal MESS stations in the feeder and the operating power. The applicability of the proposed scheduling and operation algorithms is tested on a typical 41-bus radial feeder.**

I. INTRODUCTION

TOMORROW's distribution system will witness an increasing penetration level of renewable energy resources (RESs). Because RESs are not dispatchable, the distribution network operator (DNO) faces system security and power quality problems [1]. The DNO can add energy storage systems (ESSs) to compensate for the intermittent nature of RESs and improve his profitability as well. ESSs have many other power applications including load shifting, energy arbitrage and peak shaving [2], [3]. Besides, ESSs can participate in ancillary services markets to provide online or offline reserve [4]. Also, ESSs can improve the voltage quality by supplying the grid reactively (e.g., in long radial distribution feeder) [5]. Other DNOs use ESSs to supply the load growth instead of upgrading the system (transmission and distribution (T&D) upgrade deferral).

For ESS previous wide applications; control of energy storage has gained the research interest either on the low-level side (charging level) [6] or energy management level [7], [8]. Energy management of multi-storage is formulated as a model predictive problem that maximizes an economic cost function such that the system operational constraints are respected as in [9], [10] and [11]. Further, the current development in electrical vehicles industry [7], [12] has boosted this interest. Although ESSs have the potential to mitigate several problems in modern distribution systems with renewable resources, the cost is still a huge barrier to implement multi-ESSs in typical distribution systems. Recently, mobile ESS (MESS) has been investigated by research centers such as the Electric Power Research Institute (EPRI) [13], [14]. The MESS is a utility-scale storage bank (e.g., a lithium-ion battery) carried on a truck that connects it to different buses of the feeder. The advantage of transportability is the ability to deliver localized reactive power support, power losses reduction, voltage regulation, dispersed RESs integration, and T&D upgrade deferral. MESS is available commercially in 100, 1000 and 5000 kW ratings [15]. An example of a MESS prototype project is a 500 kW/1000 kWh project for tea industry peak shaving in China [16]. Another project under investigation is a 500 kW/776 kWh SCiBTM lithium-ion battery for peak shaving and voltage regulation in a distribution system. The project is located in Spain and conducted by Toshiba Cooperation [17].

Indeed, the MESS is a promising storage technology that will contribute to solving many issues in active distribution systems. Optimal scheduling and energy management algorithms for a MESS in an active distribution system are not developed in the current literature. This work proposes a day-ahead energy management system (EMS) for a MESS owned by a DNO. The DNO uses the MESS for minimizing the day-ahead cost of power imported from the grid. Further, MESS provides reactive power support for the system for voltage regulation at critical loads. The main contributions of this work to the research field are as follows:

1- Developing an EMS system for a MESS that 1) dispatches the MESS, and 2) schedules the truck positions.
2- Developing a MESS transition delay model.
3- Converting the nonconvex optimization problem into a convex one via the proposed two-stage technique.

This paper is organized as follows. Firstly, the system description is presented in Section II. Section III presents the proposed EMS and scheduling framework for the MESS. Simulation results are presented in Section IV. Finally, the conclusions are drawn in Section V.

I. PROBLEM DESCRIPTION

Modern distribution systems face several challenges. The intermittent nature of RESs is a serious concern. Furthermore, with high penetration of RESs and expected load growth, the DNO is required to upgrade the system to guarantee system security and power quality. This is possible by upgrading the thermal capacity for the transmission lines to increase its ampacity or adding capacitor banks to provide local reactive power support at the far end of the feeder during peak-hour. Such upgrades can be extremely expensive; an ESS can provide a cheaper solution [18]. The DNO is also required to curtail any extra

wind or PV power that may threaten the voltage level or leads to power quality issues. This is a serious issue as the wind power is known to be high during late-night while PV is maximum at noon, both do not align with peak-hours. Thus, a big portion of RESs power is curtailed, and the green advantages of distributed generation are not achieved. A MESS can participate in mitigating these problems through time-shifting [2], where the MESS can storage surplus RES energy and inject it back to the grid during peak hours. Finally, because the MESS is a battery bank connected to a DC/DC/AC voltage-source converter, the dc-link capacitor bank provides reactive power regardless of the capacity.

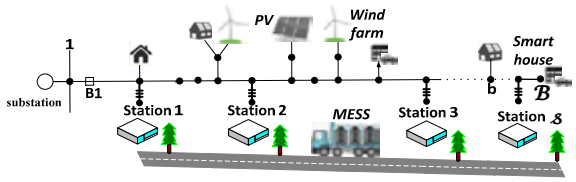


Fig. 1. Radial feeder with multi-MESS stations.

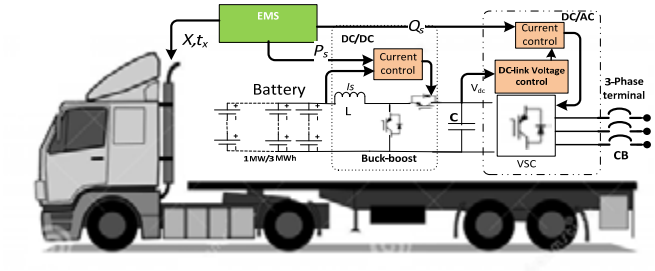


Fig. 2. Mobil energy storage system description.

The aforementioned applications of the MESS make it an attractive solution to increase the DNO profit. Fig. 1 shows an example of a radial feeder. The feeder has RESs set $\mathcal{N}_r$ at different locations of the feeder, such as wind energy conversion systems (WECS) and photovoltaic (PV) power generators. Furthermore, other dispatchable resources (such as microturbines) exist at different buses and defined by the set $\mathcal{N}_g$. Moreover, the feeder has a load set $\mathcal{N}_l$ with some smart houses (net-zero houses) that may use a rooftop PV, heat pumps, electric storage, and electric vehicle. RESs have different profiles that can be forecasted efficiently using both numerical and physical techniques (e.g., [19]). Moreover, electrical loads can be accurately predicted for the day-ahead operation planning [20], [21]. With such predictions, the DNO can schedule a MESS to move between different positions for multi-services. Different MESS positions ($\mathcal{N}_s = \{1,2,..,\mathcal{s}\}$) define the buses at which the MESS can be connected to the system which will be designated as the *MESS-stations*.

Fig. 2 depicts the structure of a MESS which consists of two parts: the ESS and the truck. The ESS consists of an array of battery cells (e.g., lithium-ion). While the battery bank terminal voltage is determined by the number of series cells in the same string, the total current is the sum of the parallel strings currents. The reader is referred to IEEE standards 485 [22] for further sizing details. The ESS is connected to the grid via a DC/DC/AC bidirectional VSC. The DC/DC converter is a current-controlled buck-boost controller. The current controller regulates the charging (discharging) power according to the set point provided by the EMS. When the EMS sends the active power set point (P_s) to the DC/DC current controller, the controller calculates the current set point I_{sref} by dividing the power on the battery bank voltage (V_s); $I_{sref} = \frac{P_s}{V_s}$.

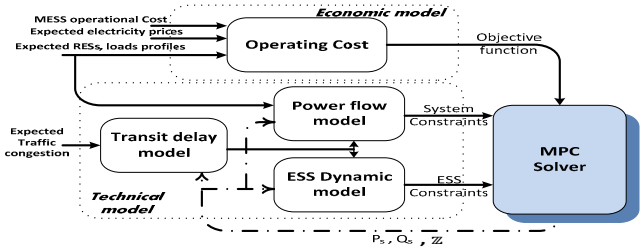


Fig. 3. System techno-economic model structure.

The current controller is a proportional-integral (PI) controller that regulates the battery current by manipulating the modulation index of the buck-boost converter [23]. On the other hand, the DC/AC converter transfers the power to (from) the grid via regulating the dc-link voltage [23]. In the dq-frame, the dc-link voltage controller decides the VSC-direct current set point whereas the EMS controls the reactive power Q_s via changing the quadrature current set point. Further details on the design of the current and dc-link voltage controllers can be found in [23]. A fast charging module is also available in some commercial versions [15].

The truck path is also controlled via the EMS. The driver requires a schedule for the MESS- stations that include the station number s and the transition time to this station t_s.

To sum up, the DNO needs to design an EMS system for a MESS that generates two sets of outputs. Firstly, the EMS calculates the ESS active and reactive power at each sample. Second, the EMS defines the desired station at each time and the transition time to this station. Next, the EMS design is explained in details.

II. Energy Management System

To design the EMS for a MESS, a complete techno-economic model, as depicted in Fig. 3, is required. The technical model is composed of three subsystems: 1) the transit delay model that describes the commute time between two stations; 2) the power flow model that relates (the buses voltage, line current and power losses) to the MESS power and positions; and 3) the ESS dynamic model that represents the battery dynamics. The economic model defines different costs of all resources that are reflected in the EMS objective function.

A. Transit Delay Model

The delay model depends on both time and stations positions. For a set of stations $\mathcal{N}_s$, the distance between stations is defined by the distance matrix $Dist$ with a zero diagonal, any element $d_{ij} \forall i,j \in \mathcal{N}_s$ define the distance between stations (i,j). In case $d_{ij} = d_{ji}$, then the MESS uses the same road between two stations (i,j).

$$Dist = \begin{bmatrix} 0 & \cdots & d_{1s} \\ \vdots & \ddots & \vdots \\ d_{s1} & \cdots & 0 \end{bmatrix} \quad (1)$$

The traffic congestion delay tc_{ijk} is a time and location dependent index can give an indicator [24] for the traffic delay. τ_{ijk} is the commute time (in samples) from a station i to j at time $k \in \mathcal{N}_k$ which is composed from three parts: the commute time , traffic congestion delay and installation time.

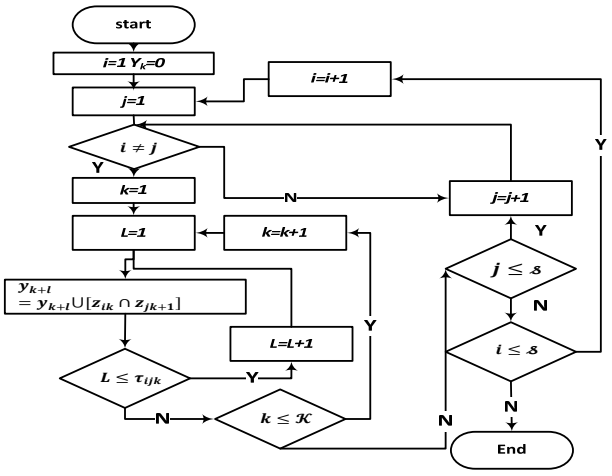


Fig. 4. Transit delay model generation.

$$\tau_{ijk} = round(\frac{tc_{ijk} + d_{ij}V_{avg} + t_{ins}}{T_s}) \forall i,j \in \mathcal{N}_s, k \in \mathcal{N}_k, i \neq j \quad (2)$$

In (2), V_{avg} is the average truck speed, t_{ins} is the MESS installation time (the connection of MESS to the grid at the station). T_s is the sample time in minutes. For the sake of the modeling, let us define the Boolean matrix $\mathbb{z} \in \mathcal{R}^{s \times \mathcal{K}}$, where each element z_{ik} equals one if the MESS exists at station i at sample time k. Next, let us define the Boolean vector $\mathbb{Y} \in \mathcal{R}^{\mathcal{K}}$ that represents the delay model. If the MESS is in transit at a sample T due to moving from a station i to another one j: $z_{iT+1} = z_{jT} = 1, i \neq j \; \forall i,j \in \mathcal{N}_s$. Then $y_k = 1$ if this transition occurs at time T such that $k > T \geq k - \tau_{ijT}$. In order to define y_k, all previous transit conditions that will lead to a future transit delay are modelled. By this way, y_k is expressed by the sum of minimum terms that represent all possible previous transits as shown in Fig. 4.The delay model is described as follows:

$$\forall i \epsilon \mathcal{N}_s, k \epsilon \mathcal{N}_k \begin{cases} \sum_i z_{ik} = 1 & (3) \\ \frac{\sum_i \sum_k |z_{ik} - z_{ik-1}|}{2} \leq N_{trips} & (4) \\ \mathbb{Y} = \{y_1, y_2, y_k \dots, y_{\mathcal{K}}\} & (5) \end{cases}$$

Equation (3) states that the MESS truck can only be at one station i at any sample k. In the case of multiple MESS units, the right-hand side of (3) is replaced by N_{mess} which is the total number of MESS trucks; this case will be investigated in future studies. Equation (4) constraints the maximum number of trips that MESS can make per day (N_{trips}) which equals half the number transitions. In case the MESS moved at time k from a station i to be at station j, then $z_{ik} = 0, z_{ik-1} = 1, z_{jk} = 1, z_{jk-1} = 0$ and the sum of the difference at sample k is calculated as $\sum_i \frac{|z_{ik} - z_{ik-1}|}{2} = 1$, which means one trip is made. However if the MESS stays at station i at samples k and at k-1, then $\sum_i \frac{|z_{ik} - z_{ik-1}|}{2} = 0$ which means no trips are made. Finally, (5) is a set of equalities that represent the delay Boolean flag y_k that indicates that the MESS is at transit in this time. y_k is a sum of minimum terms calculated from Fig. 4.

B. Power Flow Model

Second, the system power flow is a model that links a set of buses $b \epsilon \mathcal{N}_b$ by a set of branches $t \in \mathcal{N}_t$. Buses active and reactive power controls buses angles and voltages. Apparent power at each bus is represented as the sum of RES, load, generators, and MESS powers injected at the bus (if any exists). The apparent power is decomposed into active and reactive power, expressed at (6) and (7).

$$p_{bk} = p_{lk} + p_{sk} - p_{gk} - p_{rk} \quad (6)$$

$$q_{bk} = q_{lk} + q_{sk} + q_{gk} + q_{rk} \quad (7)$$

$$\forall b = s = g = l = r, b \epsilon \mathcal{N}_b, s \epsilon \mathcal{N}_s, k \epsilon \mathcal{N}_k, l \epsilon \mathcal{N}_l, r \epsilon \mathcal{N}_r, g \epsilon \mathcal{N}_g$$

For each branch t (line or transformer) that links two buses (a,b) or $t \epsilon \mathcal{N}_t, t = (a,b)$, the branch complex impedance is calculated as $z_t = r_t + jx_t$, where r_t, x_t is the branch resistance and reactance, respectively. The power flow for each branch is modelled in (8)-(13) as given in [25]. These equations do not include the angle of voltage with the assumption that no big change in voltage angle occurs in the distribution system [26]. It is worth mentioning that the voltage equation in (10) is not applicable at the slack bus ($b=0$) because it is a voltage controlled bus with a reference angle.

$$\forall t \epsilon \mathcal{N}_t, k \epsilon \mathcal{N}_k \begin{cases} p_{bk} = \mathcal{p}_{tk} - r_t \ell_{tk} - \sum_{\forall h \epsilon \mathcal{N}_t, h=(a,b), a \epsilon \mathcal{N}_b} \mathcal{p}_{hk} & (8) \\ q_{bk} = q_{tk} - x_t \ell_{tk} - \sum_{\forall h \epsilon \mathcal{N}_t, h=(a,b), a \epsilon \mathcal{N}_b} q_{hk} & (9) \\ v_{bk} = v_{ak} - 2(r_t \mathcal{p}_{tk} + x_t q_{tk}) + (r_t^2 + x_t^2)\ell_{tk} & (10) \\ \ell_{tk} = \frac{\mathcal{p}_{tk}^2 + q_{tk}^2}{v_a} & (11) \\ \ell_{tk} = |I_{tk}|^2 \quad , t = (a,b) & (12) \\ v_{bk} = |V_{bk}|^2 & (13) \\ loss_k = \sum_{\forall t \epsilon \mathcal{N}_t} r_t \ell_{tk} & (14) \end{cases}$$

In (8)-(14), $\mathcal{p}_{tk}, q_{tk}, I_{tk}, \ell_{tk}, V_{bk}, v_{bk}, loss_k$ are the branch t active and reactive power; the magnitude of the line current; the second norm of the line current; the bus voltage; the second norm of bus voltage; and the total power losses, respectively. All these variables are measured at time sample k. Equations (8) and (9) express the active, reactive power balance equation for each branch t at time k.

Equation (10) represents the Ohm's law used to calculate the voltage drop in each branch. Equation (11) is the power flow equation that calculates each branch current from its apparent power. Equations (12) and (13) show that the used values for the power flow are the second norm for the voltage and branch currents. Finally, the power loss is calculated in (14). The main power flow constraints are the branches thermal limit expressed in (15) and the voltage level limit expressed in (16).

$$\forall b\epsilon\mathcal{N}_b, t\epsilon\mathcal{N}_t, k\epsilon\mathcal{N}_k \begin{cases} \ell_{tk} \le \overline{\ell}_t & (15) \\ \underline{v} \le v_{bk} \le \overline{v} & (16) \end{cases}$$

C. ESS Dynamic Model

The final part of the technical model is the ESS dynamics including the ESS power, the total number of cycles *N*, and the state of charge (*SOC*) which are expressed in (17)-(19).

In this model, we assume that anytime, there is always an ESS at each station *s*; however, the constraint in (3) states that the MESS can be located at one unique station. Thus, the condition for injecting discharge ($p_{dh_{sk}}$) or importing charge power($p_{ch_{sk}}$) at a certain bus *s* at time *k* is that the MESS exists at this station ($z_{sk} = 1$), and the MESS is not at transit state ($y_k = 0$). These conditions are modelled by (17). Equation (18) expresses the SOC as an integrator. The SOC equals one if the accumulated charge reaches the rated energy of the storage $\overline{E}_s$. The charge and discharge efficiency is considered as η_{ch}, η_{dh}, respectively. As one MESS is assumed at each bus, the total SOC is the sum of all stations ESS SOC at a certain time. A similar technique is used for calculating the number of cycles N_k that is assumed as an integrator as well. The number of cycles is incremented by one each time a full charge is absorbed by the MESS and injected again to the grid as expressed in (19).

$$p_{sk} = (p_{ch_{sk}} + p_{dh_{sk}}) z_{sk}(1 - y_k) \quad (17)$$

$$SOC_{k+1} = \sum_s SOC_{sk} + \frac{T_s}{\overline{E}_s}(\eta_{ch} p_{ch_{sk}} + \eta_{dh} p_{dh_{sk}}) z_{sk}(1 - y_k) \quad (18)$$

$$N_{k+1} = \sum_s N_{sk} + \frac{T_s(p_{ch_{sk}} - p_{dh_{sk}})}{2\overline{E}_s} z_{sk}(1 - y_k) \quad (19)$$

$$\forall k\epsilon\mathcal{N}_k, s\epsilon\mathcal{N}_s, \eta_{ch} < 1, \eta_{dh} > 1$$

The dynamic ESS constraints are stated in (20)-(24). While constraints (20) and (21) define the charge and discharge powers as positive and negative power limited by the thermal limit of the MESS, constraint (22) limits the apparent power of the MESS to be upper bounded by the rated MESS apparent power ($\overline{S}$) at any time. Finally, the SOC is limited in (23) between the desired minimum SOC $\underline{SOC}$ and one to avoid deep discharge and overcharging that may reduce the battery life dramatically. Further, the number of cycles at the end of the day is limited by (24) as the battery life is a countable number of cycles.

$$0 \le p_{ch_{sk}} \le \overline{p}_s \quad (20)$$

$$-\overline{p}_s \le p_{dh_{sk}} \le 0 \quad (21)$$

$$p_{sk}^2 + q_{sk}^2 \le \overline{S} \quad (22)$$

$$\underline{SOC} \le SOC_k \le 1 \quad (23)$$

$$N_{\mathcal{K}} \le \overline{N} \quad (24)$$

$$\forall\, b\epsilon\mathcal{N}_b, t\epsilon\mathcal{N}_t, k\epsilon\mathcal{N}_k$$

D. Original EMS Problem

The EMS objective is maximizing the day-ahead profit for the DNO by minimizing the imported energy cost from the grid (energy arbitrage). It is assumed that the MESS is an asset owned by the DNO to provide both reactive power support and maximize net profit. RESs are owned by different parties, and they are paid a fixed feed-in-tariff (FIT) as a motivation by the government for clean energy, such a case exists in several markets, such as Ontario, Canada [27]. RESs are operated according to the maximum power point tracking strategy; the DNO buys this energy from RESs owners with the regulated FIT (C_{FIT}) . The day-ahead cost of buying RESs energy is given in (26). On the other hand, conventional distributed generators (DGs) owned by the DNO, such as microturbines or diesel engines, have an operating cost (C_{DG}). A quadratic cost function, as defined in (25), precisely represents the operation costs of DG units [25]. In (25), $\phi_g\$/(kWh)^2, \beta_g(\$/kWh), \gamma_g(\$)$ are constants that depend on the generator type.

$$C_{DG} = \sum_{k=1}^{\mathcal{K}} \sum_{g=1}^{\mathcal{G}} \phi_g(p_{gk}T_s)^2 + \beta_g(p_{gk}T_s) + \gamma_g \quad (25)$$

$$C_{res} = \sum_{k=1}^{\mathcal{K}} \sum_{r=1}^{\mathcal{R}} p_{rk} T_s C_{FIT} \quad (26)$$

The MESS cost C_{mess} includes both the truck operating cost C_{truck} and the ESS cost C_{ess} as in (27). Firstly, the truck operating cost (28) includes a fuel cost and labor cost. While the fuel cost is modelled as a worst case cost (the multiplication of maximum distance (in km) by the number of the trips by the fuel consumption cost FC ($\$/km$)), the truck labor cost tlc includes the daily stipend of the truck driver and technicians. Secondly, (29) models the ESS cost as a levelised cost ($C_{kwh}\$/kWh\ charged$). This cost is adopted before for ESS in [28]. However, the capital cost is modified to include the truck capital cost and ESS maintenance cost that includes the truck maintenance cost; further details on MESS cost calculations are available but omitted here for space limitations.

$$C_{mess} = C_{truck} + C_{ess} \quad (27)$$

$$C_{truck} = FC \times N_{trip} \times \max(Dist) + tlc \quad (28)$$

$$C_{ess} = \sum_{k=1}^{\mathcal{K}} \sum_{s=1}^{s} P_{ch_{sk}} T_s C_{kwh} \quad (29)$$

The final cost item is the grid power cost C_{grid} that is the multiplication of the expected pool price (buy price) $BP_k(\$/kWh)$ and the grid energy $P_{gridk}T_s$ as shown in (30). This cost has a negative value (becomes income) in case the distribution system exports surplus energy to the grid. The income of the DNO is generated from daily selling energy to the loads with the sell price $SP_k(\$/kWh)$ as given in (31). Finally, the profit is the difference between the income and the costs (MESS operating cost, DG running cost plus RES energy purchasing cost) as shown in (32).

$$C_{grid} = \sum_{k=1}^{\mathcal{K}} T_s BP_k P_{gridk}, P_{gridk} = P_{bk}, b = 0 \quad (30)$$

$$inc = \sum_{k=1}^{\mathcal{K}} \sum_{l=1}^{\mathcal{L}} T_s SP_k P_{lk} \quad (31)$$

$$pro = inc - C_{grid} - c_{mess} - C_{DG} - C_{res} \quad (32)$$

The final optimization problem is expressed as follows:

$$\max_{p_{ch_{sk}}, p_{dh_{sk}}, \mathbb{Z}, q_{sk}} (pro) \quad S.t. \ (3)-(11), (15)-(24) \quad (33)$$

The problem in (33) is a mixed integer nonconvex problem that faces a feasibility problem. In Section IV, this problem will be converted to a mixed integer second-order cone problem (MISOCP) using the proposed two-stage EMS.

III. EMS

To convert the problem in (33) into a mixed integer convex problem, the following suboptimal two-stage solution is proposed. Firstly, the problem is solved assuming instantaneous MESS transit; this problem calculates the MESS designated stations and its transit times (given the previous assumption). Secondly, a particle swarm algorithm optimizes the transit times of the first stage without changing the MESS stations.

The delay model $\mathbb{Y}$ given in (5) is a nonlinear sophisticated Boolean expression that depends on both the truck transit time and stations $\mathbb{Z}$. Assuming that $\mathbb{Y} = 0$ or MESS can move instantaneously from a station to another will reduce the problem complexity on the price of sub-optimality, this part will be discussed in Section IV.B. Even with the assumption $\mathbb{Y} = 0$, the problem (33) is still a nonconvex problem due to nonconvex equality constraints in the power system (11) and ESS model (17)-(19).

For the power system power flow model (11), according to [25], it can be relaxed to the inequality (34) given the following two sufficient assumptions. Firstly, the bus voltage deviation is not too large. Secondly, the buses generated powers are not too large. These conditions are usually satisfied in typical distribution systems. The quadratic constraint (34) is equivalent to the second-order cone constraint in (35) [29].

$$\ell_{tk} v_a \geq \mathcal{P}_{tk}^2 + q_{tk}^2 \forall t \epsilon \mathcal{N}_t, k \epsilon \mathcal{N}_k, t = (a, b) \quad (34)$$

$$\left\| \begin{matrix} 2\mathcal{P}_{tk} \\ 2q_{tk} \\ \ell_{tk} - v_a \end{matrix} \right\|_2 \leq \ell_{tk} + v_a \quad (35)$$

The ESS model (17)-(19) is expressed as a linear set of equalities as in (36)-(38). This is possible after modifying the power limits constraints from (20)-(21) into (39)-(40) and adding a reactive power constraint (41). This simple modification states that the MESS can charge or discharge at any station and time (*s, k*) if and only if it exists at this station or $z_{sk} = 1$. Otherwise, the power limits is set to zero. These limits affect all ESS dynamics as given in (36)-(38).

$$\begin{cases} p_{sk} = p_{ch_{sk}} + p_{dh_{sk}} & (36) \\ SOC_{k+1} = \sum_s SOC_{sk} + \frac{T_s}{\bar{E}_s} (\eta_{ch} p_{ch_{sk}} + \eta_{dh} p_{dh_{sk}}) & (37) \\ N_{k+1} = \sum_s N_{sk} + \frac{T_s (\eta_{ch} p_{ch_{sk}} - \eta_{dh} p_{dh_{sk}})}{2\bar{E}_s} & (38) \\ 0 \leq p_{ch_{sk}} \leq \bar{p}_s z_{sk} & (39) \\ -\bar{p}_s z_{sk} \leq p_{dh_{sk}} \leq 0 & (40) \\ -\bar{s} z_{sk} \leq q_{sk} \leq \bar{s} z_{sk} & (41) \end{cases}$$

Now, the instantaneous EMS problem is expressed as:

$$\max_{p_{ch_{sk}}, p_{dh_{sk}}, \mathbb{Z}, q_{sk}} (pro) \quad (42)$$

$$S.t. \begin{cases} instaneous\ transit\ model: (3), (4) \\ power\ flow\ model: (6)-(10), (15)-(16), (35) \\ ESS\ model: (22)-(24)\&(36)-(41) \end{cases}$$

Problem (42) is a mixed-integer SOCP problem that is solved efficiently using the available commercial solvers, such as GUROBI [30].

IV. CASE STUDY

To validate the proposed EMS, a detailed simulation study is conducted on a 41-bus real radial feeder in Ontario, Canada [31]. The distribution feeder is rated at 16.185 MVA and has a medium voltage 16 kV.

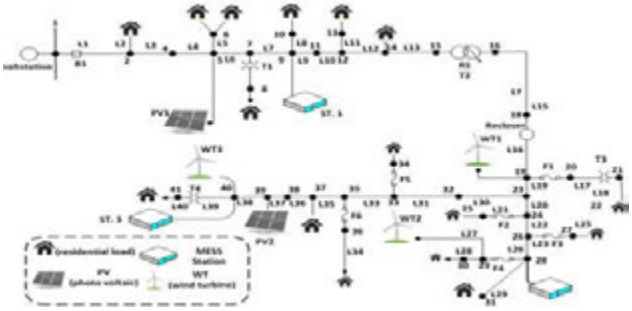


Fig. 5. Radial feeder under study.

The MESS efficiency and cost parameters are stated in Table I. For the MESS delay model, the traffic congestion delay is taken to emulate the traffic congestion pattern in freeways of Los Angeles, USA as in [32]. To sum up the results, Table II compares the techno-economic aspects after and before using the MESS services (system with no storage at all and the system that invests in MESS).

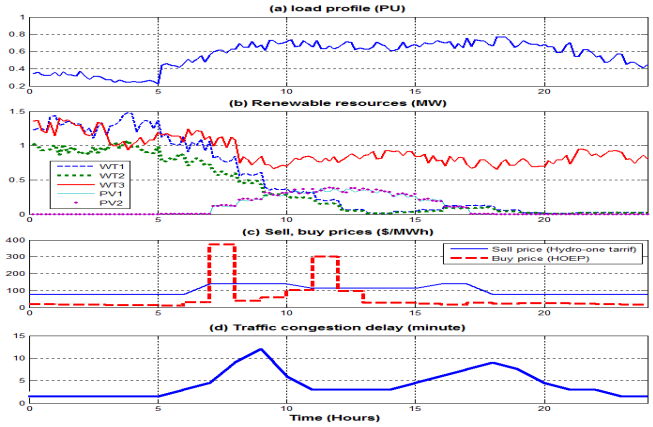


Fig. 6. Day-ahead input data. (a) Load profile. (b) RESs profiles. (c) Wholesale market and sell prices. (d) Traffic congestion delay.

Table I. Input data for the EMS.

RES Ratings	$p_{WT1} = 2MW, p_{WT2} = 1.5MW, p_{WT3} = 1.5MW,$ $p_{pv1} = 0.5MW, p_{pv2} = 0.5MW, c_{FIT} = 128\ \$/MWh$
Delay model	$d_{12} = 16.87\ Km, d_{13} = 19.92\ Km, d_{23} = 16.83\ Km$ $V_{avg} = 40\ {}^{Km}/_{hr}, t_{ins} = 5\ mins., T_s = 10\ mins.$
EMS	$SOC_0 = 0.5, \underline{SOC} = 0.2, N_{trips} = 3, \overline{N} = 1$ $\underline{v} = 0.95, \overline{v} = 1.05, N_0 = 0, \eta_{ch} = 0.75, \eta_{dh} = 1.33$ $FC = 0.4\ {}^{\$}/_{Km}, C_{kwh} = 221\ {}^{\$}/_{MWh}, tlc = 25\ {}^{\$}/_{hr}$
PSO	$M = 30\ mins., \omega_{max} = 0.9, \omega_{min} = 0.1, \overline{\alpha} = 2000,$ $C_1 = 2, C_2 = 2, n = 10$

Firstly, the MESS managed to provide a logistic localized reactive power support to keep the voltage level within allowable limits which improves the power quality. Secondly, regarding losses, the losses have been decreased by 5.5%. It is worth mentioning that the objective function is the difference between the energy bought from the grid and that sold to loads. Thus the power losses is inherently minimized in the objective function with a weight depending the energy prices at each instant. Thirdly, the net profit has increased by 3.1% as compared to that without including the profit from the localized reactive power support that defers adding one or two capacitor banks to this feeder. Finally, it worth noticing that these results mainly depend on the market prices, the load and RES profiles. Thus, in other days with no big market price difference, it is not economic to operate the MESS [33]. Finally, Using MESS in Voltage support is optional in case of long feeders where the transformer tap-changer setting can't guarantee an acceptable voltage at the far end.

V. Conclusion

A day-ahead EMS for an MESS has been presented in this paper. The EMS objective function aims at maximizing the DNO day-ahead profit and regulating the voltage level. The proposed EMS dispatches the MESS and allocates its station in the system. A traffic delay model has been developed between different stations. the problem is solved as a mixed-integer convex problem. A case study on a real 41-buses radial feeder with real data of RESs and loads is used to validate the results. The proposed EMS has successfully fulfilled all the optimization process objectives in an adaptive manner that fits the time-varying nature of renewable resources in modern active distribution grids.

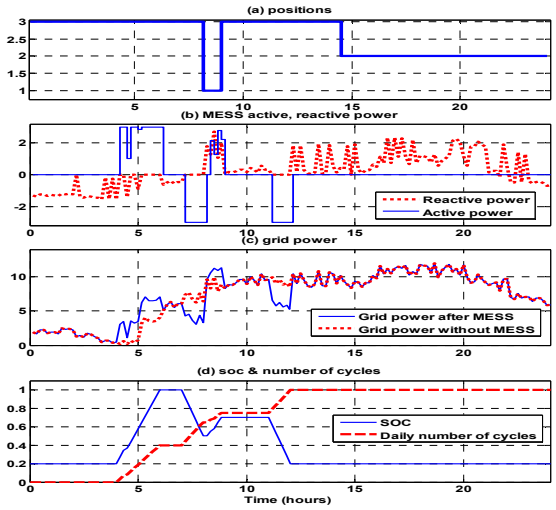


Fig. 7. EMS outputs (a) Output MESS stations (no delay considered) (b) MESS active power set point (c) MESS reactive power set point (d) MESS SOC (e) Purchased grid power.

Table II. System performance comparison.

	No MESS	With MESS
Voltage level	$V_{max} = 1.062, V_{min} = 0.89$	$V_{max} = 1.04, V_{min} = 0.956$
Power losses	$8.1863\ MWh$	$7.7277\ MWh$
Profit	14,320$	14,760$

VI. References

[1] J. Sexauer and S. Mohagheghi, "Voltage Quality Assessment in a Distribution System With Distributed Generation—A Probabilistic Load Flow Approach," *IEEE Trans. Power Del.*, pp. 1652 - 1662, July 2013.

[2] G. C. Jim Eyer, "Energy Storage for the Electricity Grid: Benefits and Market Potential Assessment Guide," SANDIA , Albuquerque, New Mexico, 2010.

[3] J. A. Rahul Walawalkar, "Market Analysis of Emerging Electric Energy Storage Systems," NETL, 2008.

[4] Y. Rebours, D. Kirschen, M. Trotignon and S. Rossignol, "A Survey of Frequency and Voltage Control Ancillary Services—Part I: Technical Features," *IEEE Tran. Power Syst.*, vol. 22, no. 1, pp. 250-357, 2007.

[5] J. Zhong and K. Bhattacharya, "Toward a competitive market for reactive power," *IEEE Trans. Power Syst.*, vol. 17, no. 4, pp. 1206-1215, 2002.

[6] S.-C. Wang and Y.-H. Liu, "A PSO-Based Fuzzy-Controlled Searching for the Optimal Charge Pattern of Li-Ion Batteries," *IEEE Trans. Indust. Electron.*, vol. 62, no. 5, pp. 2983- 2993, 2015.

[7] J. Trovao, V. Santos, C. Henggeler Antunes, P. Pereirinha and H. Jorge, "A Real-Time Energy Management Architecture for Multisource Electric Vehicles," *IEEE Trans. Indust. Electron.*, vol. 62, no. 5, pp. 3223-3233, 2015.

[8] X. Lu, K. Sun, J. Guerrero, J. Vasquez and L. Huang, "State-of-Charge Balance Using Adaptive Droop Control for Distributed Energy Storage Systems in DC Microgrid Applications," *IEEE Trans. Indust. Electron.*, vol. 61, no. 6, pp. 2804- 2815, 2014.

[9] F. G.-T. a. C. Bordons, "Optimal Economical Schedule of Hydrogen-Based Microgrids With Hybrid Storage Using Model Predictive Control," *IEEE TRANS. INDUST ELECTRON.*, vol. 62, no. 8, pp. 5195 - 5207, AUGUST 2015.

[10] R. M. K. R. A. a. K. N. Aymen Chaouachi, "Multiobjective Intelligent Energy Management for a Microgrid," *IEEE TRANS INDUST ELECTRON*, vol. 60, no. 4, pp. 1688-1699, April 2013.

[11] W. Tushar, B. Chai, C. Yuen, D. Smith, K. Wood, Z. Yang and H. Poor, "Three-Party Energy Management With Distributed Energy Resources in Smart Grid," *IEEE Trans. Indust. Electron.*, vol. 62, no. 4, pp. 2487-2498, 2015.

[12] J. Talla, L. Streit, Z. Peroutka and P. Drabek, "Position-Based T-S Fuzzy Power Management for Tram With Energy Storage System," *IEEE Trans. Indust. Electron.*, vol. 62, no. 5, pp. 3061- 3071, 2015.

[13] D. Rastler, "Technical Specification for a Transportable Lithium-Ion Energy Storage System for Grid Support Using Commercially Available Lithium-Ion Technology," EPRI, July 2012.

[14] EPRI, "Transportable Energy Storage Systems Project," EPRI, Oct 2009.

[15] "Winston Energy Group Limited," [Online]. Available: http://en.winston-battery.com/index.php/products/mobile-power. [Accessed 30 July 2015].

[16] "Department of Energy," [Online]. Available: http://www.energystorageexchange.org/projects/153. [Accessed 30 July 2015].

[17] "Toshiba," [Online]. Available: https://www.toshiba.co.jp/about/press/2014_01/pr0704.htm. [Accessed 30 July 2015].

[18] J. Eyer, "Electric Utility Transmission and Distribution Upgrade Deferral Benefits from Modular Electricity Storage," Sandia, New Mexico, June 2009.

[19] N. Chen, Z. Qian, I. Nabney and X. Meng, "Wind Power Forecasts Using Gaussian Processes and Numerical Weather Prediction," *IEEE Trans. Power Syst.,* vol. 29, no. 2, pp. 656-665, 2014.

[20] Y. Chakhchoukh, P. Panciatici and L. Mili, "Electric Load Forecasting Based on Statistical Robust Methods," *IEEE Trans. Power Syst.,* vol. 26, no. 3, pp. 982-991, 2011.

[21] D. Saez, F. Avila, D. Olivares, C. Canizares and L. Marin, "Fuzzy Prediction Interval Models for Forecasting Renewable Resources and Loads in Microgrids," *IEEE Trans. Smart Grid,* vol. 6, no. 2, pp. 548-556, 2015.

[22] I. S. C. C. 2. o. S. Batteries, IEEE Recommended Practice for Sizing Lead-Acid Batteries for Stationary Applications, IEEE, 1997.

[23] R. I. Amirnasr Yazdani, Voltage-Sourced Converters in Power Systems, Wiley, 2010.

[24] I. Cambridge Systematics, "Traffic Congestion and Reliability (Trends and Advanced Strategies for Congestion Mitigation)," Federal Highway Administration , Sept. 2005.

[25] X. X. C.-C. C. R. G. Wenbo Shi, "Distributed Optimal Energy Management in Microgrids," *IEEE Trans. Smart Grid,* vol. 6, no. 3, pp. 1137-1146, May, 2015.

[26] S. H. Low, "Convex relaxation of optimal power flow—Part II: Exactness," *IEEE Trans. Control Netw. Syst.,* vol. 1, no. 2, pp. 177-189, Jun. 2014.

[27] IESO, "feed in tariff," [Online]. Available: http://fit.powerauthority.on.ca/. [Accessed 8 Sept. 2015].

[28] W. T. J. Piyasak Poonpun, "Analysis of the Cost per Kilowatt Hour to Store Electricity," *IEEE Trans. Energy Conversion,* vol. 23, no. 2, pp. 529-534, June 2008.

[29] M. Nick, Distrib., R. Cherkaoui and M. Paolone, "Optimal Allocation of Dispersed Energy Storage Systems in Active Distribution Networks for Energy Balance and Grid Support," *IEEE Trans. Power Syst.,* vol. 29, no. 5, pp. 2300 - 2310, Sept. 2014.

[30] "Gurobi Optimization," Gurobi, [Online]. Available: http://www.gurobi.com/. [Accessed 18 Aug 2015].

[31] Y. M. Atwa, "Distribution System Planning and Reliability Assessment under High DG Penetration," University of Waterloo , Waterloo, Ontario, Canada, 2010.

[32] "L.A traffic," [Online]. Available: http://home.znet.com/schester/calculations/traffic/la/plots/all/index.html. [Accessed 23 Aug. 2015].

[33] H. Abdeltawab and Y.-R. Mohamed, "Market-Oriented Energy Management of a Hybrid Wind-Battery Energy Storage System Via Model Predictive Control With Constraint Optimizer," *IEEE Trans Indust Electron,* vol. 62, no. 11, pp. 6658 - 6670, Nov. 2015.

OPHO

Special Session on

OPTOELECTRONICS AND PHOTONICS

Steering Committee

Chairs: Vera Gradišnik, University of Rijeka, Croatia
Irena Jurdana, University of Rijeka, Croatia
Duško Čakara, University of Rijeka, Croatia

PAPERS

Analysis of a-Si:H p-i-n photodiode detection of HeLa cells luminescence

V.Gradišnik
University of Rijeka/Faculty of Engineering, Rijeka, Croatia
e-mail: Vera.Gradisnik@riteh.uniri.hr

***Abstract* - Today, optical biosensors are the most promising candidates for Lab-on-a-Chip (LoC) with regards to their sensitivity and miniaturization. The label-free optical detection method has a low negative influence on tissue and is therefore suitable for the detection of mammalian cells and the molecule itself. The attractive optical methods for analyte detection, chemiluminescence and fluorescence, require a highly sensitive photodetector as biotransducer. In addition, the optoelectronic technology provides the thin film hydrogenated amorphous silicon (a-Si:H) photodiodes integration in microfluidic systems. In our previous work we have observed the influence of the density of native and light-induced localized states, distributed in the mobility gap of a-Si:H at different bias voltages, on the detection of luminescence of HeLa cells. As in the optical bias dependence of the modulated photocurrent method (OBMPC), the sample is excited by a blue LED light bias beam of constant intensity, to activates the dangling bonds in a-Si:H and induces a second light beam carried out from the chemiluminescent emission reaction in HeLa cells. The analysis of recombination rate and space charge density in a-Si:H is performed to better understand the underlying physics for luminescence detection at applied different bias voltages on the photodiode.**

Keywords - biosensor; a-Si:H p-i-n photodiode; recobination rate; space charege density; chemiluminescence

I. INTRODUCTION

The integration of multiplex laboratory analysis on a single chip for the detection of analytes is today crucial for the development of a Point-of-care testing (POCT). With the development of microfluidic Lab-on-a-chip (LOC) technology to monitor small sample and reagent volumes, medical tests can be performed at the right place and in real time [1]. The LOC devices integrate fluid handling, sample preparation: filtration - homogenization - dilution, target detection, transducer readout and signal processing [2]. The integrated detector should be highly sensitive, have low power consumption and short response time, be inexpensive to manufacture, compact and easy to use, portable, have a network connection to the central database and enable reproducible detection [3-7].

Despite the good properties of mechanical and electrochemical sensors for their application in microfluidic devices, optical biosensors are more suitable for LOC, due to the advanced technology of integrating optical detection methods into a microfluidic device and the increasing use of optical methods in laboratories [2]. The biosensor consists of a bioreceptor and a transducer. Bioreceptors, such as DNA, antibodies, proteins, cells etc.

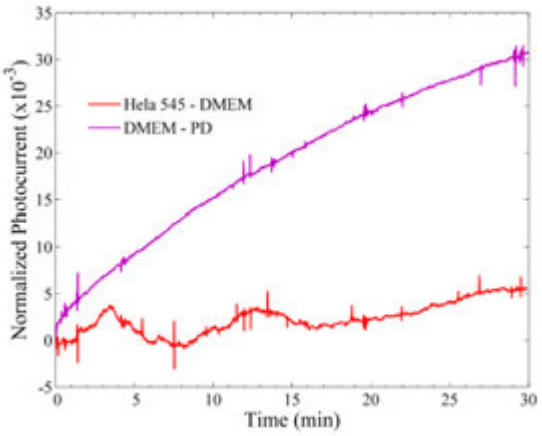


Figure 1. The normalized photocurrent time curve measured by sample with DMEM and from it substracted the a-Si:H p-i-n photodiode current (DMEM-PD) and by sample with HeLa cells concentration, N = 6.25x10^5 cells/mmol in DMEM and from it substracted the current measured by sample with DMEM at reverse bias voltage of V_r= 1.3 V.

capture the analyte and a transducer converts the detected biomolecular signal into a physical signal. In our case, light is converted into an electrical signal. To ensure sensitivity, stability and selectivity, the bioreceptors are in most cases immobilized on the surface of the transducer. The new transducers and detection methods for specific biological interactions and biomolecules have been gradually developed as nanotechnology and microarrays mature.

Optical detection methods include absorbance, fluorescence (FL), chemiluminescence (CL) and surface plasmon resonance (SPR) [8]. The fluorescence detection method mainly uses label-based techniques that lead to a loss of functionality and a laser source that damages the tissue. The label-free chemiluminescence techniques require simple and highly sensitive instruments. The p-i-n a-Si:H photodiode (PD) is a promising, highly sensitive transducer into LoC devices [3-6]. The CL detection of HeLa cervical cancer cells is reported in [9], the label-free CL method based on optical absorption in [10].

In our previous work [11, 12] we observed the influence of the density of native and light-induced localized states distributed in the mobility gap of a-Si:H at different bias voltages on the detection of luminescence of HeLa cells. The blue LED light activates the dangling bonds in a-Si:H and induces the chemical reaction in HeLa cells, which is detected as CL signal, as in the optical bias dependence of the modulated photocurrent method (OBMPC) [13, 14].

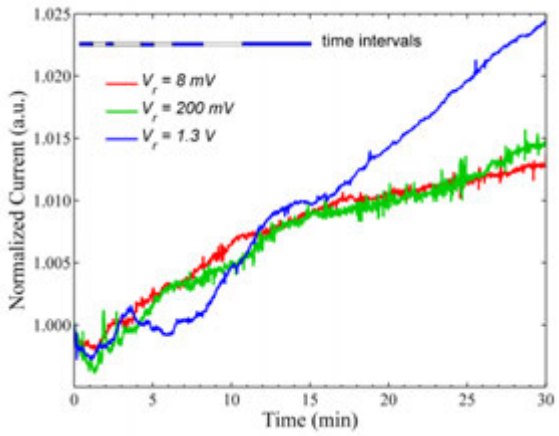


Figure 2. The normalized measured current at HeLa cells concentration and reverse bias voltage: (a) N = 6.5x10^5 cells/mmol, V_r= 8 mV; (b) N = 8.25x10^5 cells/mmol, V_r= 200 mV; and (c) N = 6.25x10^5 cells/mmol, V_r= 1.3 V. The time intervals (blue white line) in case of (c).

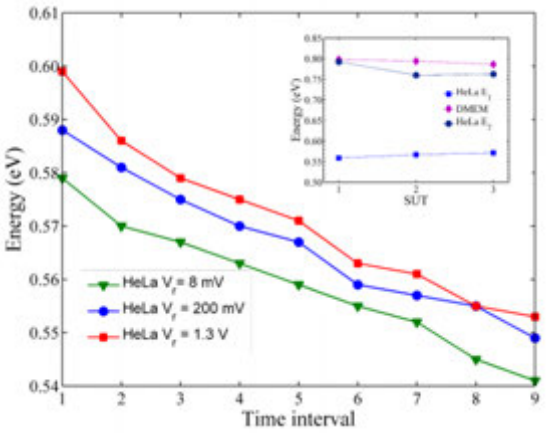


Figure 3 The calculated shallow trap energies for samples with HeLa cells vs. time interval and inside the average of two energy levels by sample with HeLa cell in DMEM and one energy level by sample with DMEM, at applied reverse bias voltage on SUT 1: 8 mV, SUT 2: 200 mV and SUT 3: 1.3V.

The product of oxygenation of HeLa cells is singlet oxygen, $[^1O_2]$, which emits light in the visible and IR part of the spectrum. The singlet oxygen is highly reactive and act with mitochondria, DNA and lipid membranes, which leads to the degradation of cellular microstructures and finally to cell death [15].

The objective of this study is to investigate a key difference in the detection of the CL and FL in HeLa cells in Dulbecco's Modified Eagle's Medium (DMEM, the base medium used for the HeLa cell line) and DMEM, respectively. Here is an enhanced analysis of our own experimental data previously published in [11, 12]. The method is based on the analysis of the role of native and light-induced localized states at specific trap energy depths in recombination and trapping-detrapping processes of free charge carriers, and related space charge density [16, 17] inside i-layer of a-Si:H p-i-n photodiode. Here presented results are obtained with previous [11] observed optimal photodiode bias voltage for label-free luminescence detection and without immobilization of cells on the surface of plastic well.

II. THEORETICAL MODEL

The present procedure and analysis is based on the following steps. The sample under test (SUT) is illuminated with a monochromatic blue LED light "bias" beam of constant intensity (DC) and with a second light "probe" beam (OBMPC) of unknown intensity, wavelength and modulation frequency emitted by a HeLa cells. The a-Si:H p-i-n response on probe light beam is shown in current transients (Fig.1), which can be observed in the time intervals of the measured current curves shown in Fig.2. The current measured by the DMEM sample shows no transient responses (Fig. 1).

The measured current is normalized (Fig. 2), $PCR = I_{ph}(t)/I_{ph}(t_0)$, where $I_{ph}(t_0)$ is the first peak value immediately after applying the reverse bias voltage, V_r, to the PD, for further analysis with the MATLAB program [11, 12]. The PD dark current is low [3-6], [18] and is neglected in the analysis.

In our experiment HeLa cells do not emit light in the dark. The blue light induces the formation of hydrogen peroxide (H_2O_2) and nitric oxide (NO) in HeLa cells, and the CL reaction takes place with formation of singlet oxygen ($^1O_2$) and emission of light. The luminescence kinetics of singlet oxygen describes the two-exponential model [15] well, given by

$$I_n(\mathrm{t}) = \mathrm{A}\frac{\tau_\Delta}{\tau_T - \tau_\Delta}\left(\mathrm{e}^{-\mathrm{t}/\tau_T} - \mathrm{e}^{-\mathrm{t}/\tau_\Delta}\right), \qquad (1)$$

where τ_T is the triplet decay time and τ_Δ is the singlet oxygen decay time, A is the current at t=0 proportional to the initial concentration and quantum yield of singlet oxygen [11, 15].

In each time interval the current time curve is normalized and shifted to the initial value zero and fitted to the model (1) with τ_T, τ_Δ and A as free parameters.

The normalized current time curve in the case of a DMEM sample illuminated with the blue LED (Fig. 1) is fitted to the Gomperz function (DMEM take FL reaction), which is model normally used in biological analysis [19]

$$I_n(\mathrm{t}) = I_{n0}\left(\mathrm{e}^{-e^{-(t-t_c)/\tau_{fl}}}\right), \qquad (2)$$

where I_{n0} is the final asymptotic current value; and the time constants, τ_{fl} and t_c are the fitting parameters representing the FL reaction rate and the FL range, respectively.

In an a-Si:H p-i-n photodiode, the time constants extracted from measured current time curves by samples with HeLa cells in DMEM (1) and DMEM (2), respectively, results from the interaction of electrons as majority carriers with the gap states $D^i(E)$ through the trapping-detrapping processes. Electron trapping-detrapping takes place in different types of gap states, in $D^i(E)$ at the respective energy levels $E_{ti} = kT \cdot \ln(\nu_0 \tau_{ri})$, where ν_o is attempt-to-escape frequency in the order of $2 \cdot 10^{13} s^{-1}$ and $1/\tau_{ri}$, trapping term. The dangling bond states could be neutral, D^0, positive, D^+ and negative, D^- charged states.

The energy and spatial distribution, the types and the temporal activity of the localized states in the mobility gap

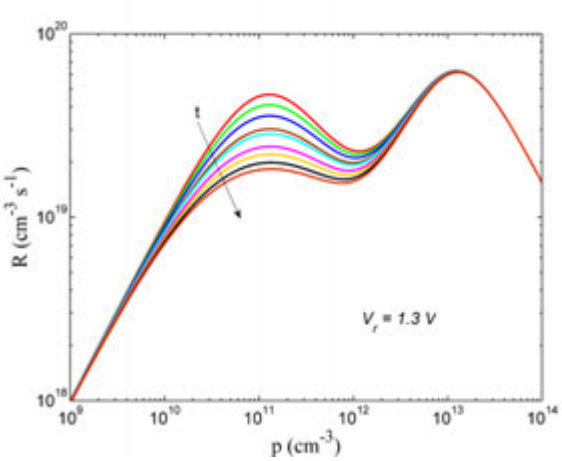


Figure 4 Recombination rates against hole concentration, p by sample with HeLa cell in DMEM at reverse bias voltage of 1.3 V and blue light LED illumination in time intervals.

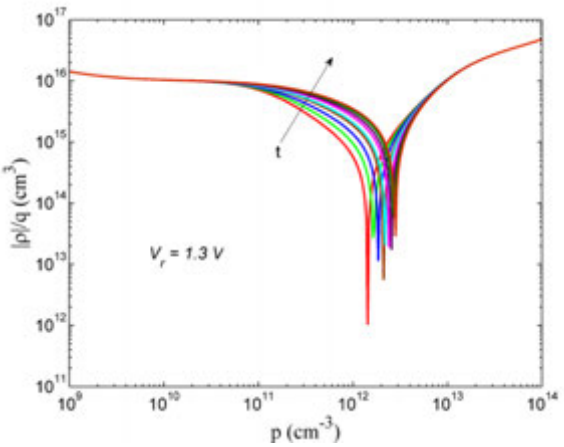


Figure 5 Space charge density against hole concentration, p by sample with HeLa cell in DMEM at reverse bias voltage of 1.3 V and at blue light LED illumination in time intervals.

of a-Si:H i-layer, which act as trapping-detrapping and recombination centers, are crucial for the detection of CL or FL, and are closely related to the characteristic transient response shapes.

The energy levels of the localized states were calculated with the help of our experimental data in [11] and are shown in Fig. 3. Those from measured current by samples with HeLa cells in the DMEM result in two trap energy levels; the shallow trap and the deep trap. The samples with DMEM result in one trap energy level: the deep trap. The non-radiative transitions of free carriers from band tails to localized states (shallow and/or deep) are considered in the trapping-detrapping model. The transitions between the shallow and deep states are not considered in the model. The model of the recombination rate through localized states with two energy levels in the mobility gap and the related space charge region is proposed by Dhariwal et al. [16]. In this paper the kinetics of trapping and detrapping are analyzed. In each time interval in which the transient response is observed, the quasi-equilibrium between the trapped and free charge carriers in the conduction and valence band is assumed, respectively. At low bias voltages the recombination through the tail states is neglected.

The variation of the np product from the thermal equilibrium value (n_i^2) is used in the model to validate the semiconductor excitation. The photo-induced excitation is assigned to the constant generation rate, which is equal to its mean value as a function of depth. Uniform illumination is assumed on the active surface on the p-type side. The n^+ and p^+ layers are thin and photoinactive. The simulation with assumed uniform electric field inside i-layer is performed.

The model predicts transitions that occur at correlated energy states with energies E_1 and E_2 from the conduction band edge, E_C. The energies E_1 (D^+ and D^0) and E_2 (D^0 and D^-) are calculated by fitting the current transient response in each time interval to the two-exponential model in the case of HeLa cells in the DMEM. The energy level E_2 is calculated by fitting to the Gomperz function (Fig. 3) and $E_1 = 0.9\,eV$ in the case of DMEM is assumed in accordance with the defect-pool model.

The recombination rate [16] is modelled during each time interval by

$$R = N_{db} \cdot f^0 \left[nc_{n1} \left(\frac{pc_{p1} + n_1 c_{n1}}{nc_{n1} + p_1 c_{p1}} \right) + nc_{n2} - n_1 c_{n1} - n_2 c_{n2} \left(\frac{nc_{n2} + p_2 c_{p2}}{pc_{p2} + n_2 c_{n2}} \right) \right] \quad (3)$$

where are the net capture rates for electrons, holes: c_{n2}, c_{p1} and c_{n1}, c_{p2} at two correlated bond states, neutral traps and Coulomb attractive charged traps, respectively; N_{db} the concentration of the dangling bond; n, p the free carrier concentrations of electrons and holes estimated for given illumination and reverse bias voltage; n_1, p_1, n_2, p_2 the concentrations of electrons and holes at energy levels E_1 and E_2; f^0 the occupation probability for the neutral charge states in the steady state, which values are given in Table I. The light intensity of the HeLa cells is estimated.

The space charge density inside i-layer changes in five regions: (i) positive valence band tails dominates close to the p^+-i junction, (ii) positive dangling bonds, D^+, dominate the next region, (iii) the central part of the i-layer is a neutral semiconductor, (iv) negatively charged dangling bonds, D^- dominates near the n^+-i junction and (v) negatively charged conduction band tail dominates close to the n^+-i junction.

TABLE I. VALUES OF PARAMETERS [12-14], [16]

No	Parameter	Value	Explanation
1	v_o	$2 \cdot 10^{13} s^{-1}$	Attempt-to-escape frequency
2	n_i^2	1.44 ·10^9 cm^{-6}	Square of intrinsic carrier concentration
3	E_G	1.72 eV	Mobility gap
4	N_{db}	10^{16} cm^{-3}	The concentration of dangling bond
5	c_{n2}, c_{p1}	10^{-9} cm^{3}s^{-1}	Cpture rates for neutral traps
6	c_{n1}, c_{p2}	10^{-7} cm^{3}s^{-1}	Cpture rates for Coulomb attractive charged traps

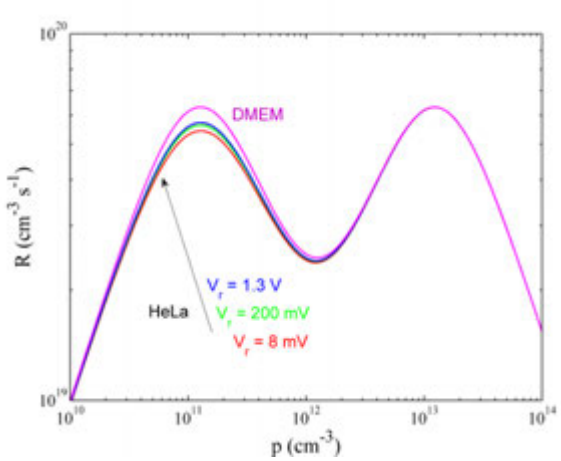


Figure 6 The recombination rate against hole concentration, p by sample with HeLa cell in DMEM and DMEM at reverse bias voltage of 8 mV, 200 mV and 1.3 V and blue light LED illumination.

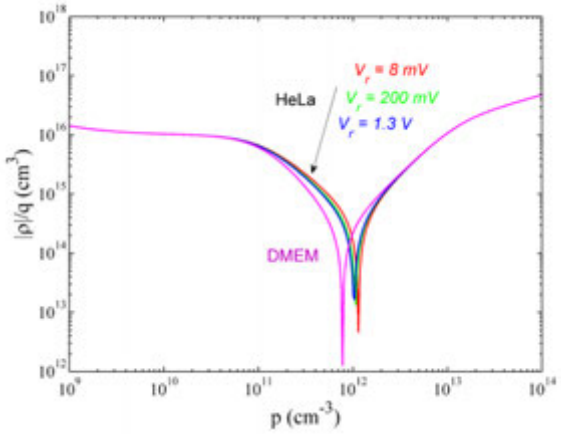


Figure 7 Space charge density against hole concentration, p by sample with HeLa cell in DMEM and DMEM at reverse bias voltage of 8 mV, 200 mV and 1.3 V and blue light LED illumination.

The trapping and detrapping processes take place during the transient response when positive D^+ and negative D^- dangling bonds near p^+-i and n^+-i junctions, respectively are charged and discharged. The space charge density [16] in the i-layer under constant blue light LED illumination, assumed low frequencies HeLa cells emission and at low bias voltages is

$$\rho = q\left[p + p_t - n - n_t + N_{db}\left(f^+ - f^- \right)\right] \qquad (4)$$

where N_{db} is the density of the dangling bond, $p \cong G_{opt}\tau_p$, the free hole, and $n \cong G_{opt}\tau_n$, the electron concentrations, p_t, the trapped charge density of holes in the valence, and n_t, electrons in the conduction band tail states, respectively and f^+, f^-, the occupation probability of a state D^+, D^- at specific trap energy, E_{ti}.

III. Experimental Setup

The array of 5x6 a-Si:H p-i-n photodiodes [18] (Department of Information Engineering, Electronics and Telecommunications, Sapienza Università di Roma, Italy) is deposited by plasma enhanced chemical vapor deposition (PECVD) in a three-chamber high vacuum system on a 50 x 50 x 1,1 mm^3 glass substrate. The area of each photodiode is 2 x 2 mm^2. The a-Si:H-p-i-n photodiode has spectral responsivity of about 0.2 A/W for blue light (430 nm). A blue LED diode (RGB LED Lamp Kingbright) emitting at 430 nm was used as a constant monochromatic light source. All measurements were performed at room temperature.

The measurement set-up [11, 12] is as follows: In a dark metal box there is a plastic well on the surface of the a-Si:H PD array. The PD is connected in series with a load resistor, R_L, of 10 kΩ, voltage source (Agilent Technologies E3631A DC voltage source), and digital multimeters, DMM, (Agilent Technologies 34450A meter). Before starting the assay, the a-Si:H PD is first illuminated with withe light to neutralize the defects induced by the previous blue light illumination. Then the open-circuit voltage under blue LED light illumination is measured, and then the reverse bias voltage is applied to the photodiode. In this way the starting point of the measured current transient is defined and calibration is performed. In repeated experiments where the luminescence reaction is measured, the bias voltage was 8 mV, 200 mV and 1.3 V, respectively. Each measurement is performed for one hour under constant blue LED ($\lambda = 430\,nm$, $I_{LED} = 20\,mA$) illumination. The PD current is monitored with the Keysight BenchVue software. The same procedure is repeated with: a) a plastic well on the surface of PD; b) 3 ml DMEM introduced into the plastic well with a pipette and c) the 3 ml solution containing the corresponding concentration of HeLa cells in the DMEM. The cells were cultivated as a monolayers and held in DMEM [11]. The HeLa cells were label-free and not immobilized on the surface of the plastic well.

IV. Results and Discussion

In our experiment we consider the signal of the HeLa cells as an experimental signal. To separate the quasi-Fermi levels for electrons and holes, each by much more than k_BT from the equilibrium Fermi level and to take the electrons as majority carriers, and to make the quasi-Fermi level of the trapped electrons coincide with that of the free electrons [13, 14] the DC generation rate is provided by constant blue LED light.

Knowing the extracted trap activation energy levels (Fig. 3) from experimental data, the recombination rate (Fig. 4) and the related space charge density (Fig. 5) are numerically simulated with the models proposed in [16].

The simulated recombination rate shows two peaks (Fig. 4), one near the p^+-i junction related to the deep trap $D^{+/0}$ states, and another near the n^+-i junction related to the shallow trap $D^{-/0}$ states. The same behavior is observed for each applied reverse bias voltage and Hela cells concentration. The recombination rate around the p^+-i junction through the deep traps shows no changes over time. Near the n^+-i junction, the recombination rate shows a smaller peak that decreases with time. This shows that the light detected from CL in Hela cells is responsible for the detrapping process within time intervals with characteristic current transient shapes.

The peak of the recombination rate closer to the n^+-i junction decreases with time as the activation energies of

the dangling bond (D^-) become shallower with time (Fig. 3). At the same time, the space charge density dip shifts towards the p^+-i junction, as shown in Fig. 5. This shift indicates that the illumination by a monochromatic blue light LED "bias" beam of constant intensity (DC) contributes more and more to the PD current, as in [17], and the emission of the HeLa cells decreases.

In case of a sample with DMEM the fitting with the Gomperz function results in only one energy level. The second energy level is assumed to be 0.9 eV according to the given model in [16], assuming the energy levels in the a-Si:H mobility gap. The recombination rate shows two peaks that are quasi-independent of the applied voltage, as shown in Fig. 6, since the extracted energy levels are close together and close to the reference energy level of 0.9 eV. The space charge density (Fig. 7) does not show the significant changes even with applied reverse bias voltage during the observation period.

As shown in Fig. 6 and Fig. 7, the increased applied reverse bias voltage has no effect on the recombination rate and space charge density in the case of the sample with DMEM. The recombination rate near the n^+-i junction increases with increasing applied reverse bias voltage for the sample with HeLa cells in DMEM and the region of the quasi-neutral space charge density enlarge through i-layer towards n^+-i junction.

V. CONCLUSION

This paper describes the novel chemiluminescence-based detection method in Hela cell using a p-i-n a-Si:H photodiode, in which HeLa cells were not immobilized on the surface of the plastic well and excited with blue LED light. The current transients, ascribed to detected chemiluminescence, were observed by samples with HeLa cells, in contrast to the current increase measured by samples with DMEM and ascribed to fluorescence. From the extracted energy levels of distributed excited localized states in the case of samples with Hela cells or in the case of DMEM and their influence on the recombination rate and space charge density in time, it follows that the $D^{-/0}$ near the n^+-i junction are involved in the detection of chemiluminescence in HeLa cells. This method provides a novel way of HeLa cell detection and destruction.

ACKNOWLEDGMENT

We greatly thank Prof. D. Caputo and Prof. G. de Cesare (Sapienza Università di Roma, Italy) for the helpful discussions, design, fabrication and providing a-Si:H p-i-n PD samples. We thank Prof. S. Kraljević-Pavelić (University of Rijeka) for useful discussion and HeLa cells providing and finally Darko Gumbarević for assistance in experiment and useful discussion.

REFERENCES

[1] N. M. M. Pires, T. Dong, U. Hanke and N. Hoivik, "Recent Developments in Optical Detection Technologies in Lab-on-a-Chip Devices for Biosensing Applications," *Sensors*,Vol. *14*, pp. 15458-15479, 2014; doi:10.3390/s140815458.

[2] D. Duval and L. M. Lechuga, "Breakthroughs in Photonics 2012: 2012 Breakthroughs in Lab-on-a-Chip and Optical Biosensors," IEEE Photonics J., Vol. 5, No. 2, 0700906, April 2013. DOI: 10.1109/JPHOT.2013.2250943.

[3] D. Caputo, G. de Cesare, C. Fanelli, A. Nascetti, A. Ricelli, and R. Scipinotti, "Amorphous silicon photosensors for detection of Ochratoxin A in wine," *IEEE Sens. J.*, Vol. 12, pp. 2674-2679, 2006.

[4] M. Mirasoli, A. Nascetti, D. Caputo, M. Zangheri, R. Scipinotti, L. Cevenini, G. de Cesare, and A. Roda, "Multiwel cartride with integrated array of amorphous silicon photosensors for chemiluminescence detection: development, characterization and comparison with cooled-CCD luminograph," *Anal. Bioanal. Chem*., Vol. 406, pp. 5645-5656, 2014.

[5] A.C.Pimentel, A.T.Pereira, V.Chu, D.M.F.Prazeres, J.P.Conde, "Detection of Chemiluminescence Using an Silicon Photodiode," IEEE Sensors J., vol.7, no.3-4, pp. 415-416, Mar.-Apr. 2007.

[6] T.Pereira, C.Pimentel, V.Chu, D.M.F.Prazeres, J.P.Conde, "Detection of Chemiluminescence Detection of Horseradish Peroxidase Using an Integrated Amorphous Silicon Thin-Film Photosensor," IEEE Sensors J., vol.9, no.10, pp. 1282-1290, Oct 2009.

[7] T. Kamei, S. Ito, T. Kobayashi, and R. Maeda, "Towards a fully integrated laser-induced fluorescence detection device for point-of-care bioanalysis," in Proc. Micro Electro Mechanical Systems (MEMS), IEEE 9th Int. Conf., 2012, pp. 890-893, 2012.

[8] A. Waseema, M. Yaqoob and A. Nabib, " Analytical Applications of Flow Injection Chemiluminescence for the Determination of Pharmaceuticals–A Review," Current Pharmaceutical Analysis, No 9, pp. 363-395, 2013.

[9] L. L. Bronsart, C. Stokes, C. H. Contag, „Chemiluminescence Imaging of Superoxide Anion Detects Beta-Cell Function and Mass," PLoS ONE, Vol 11, No. 1, e0146601. 2016. doi:10.1371/journal.pone.0146601.

[10] M. A. Ahmad, A. Najar, A. El Moutaouakil, N. Nasir, M. Hussein, S. Raji, and A. Hilal-Alnaqbi, „Label-Free Cancer Cells Detection Using Optical Sensors," *IEEE Access*, Vol. 6, pp. 55807-55814, 2018. doi: 10.1109/ACCESS.2018.2872768.

[11] V. Gradišnik, D. Gumbarević, "The Blue Light Defects Activation in a-Si:H Pin Photodiode as a Biosensor," *Key Engineering Materials*, Vol. 843, pp. 64-69, May 2020. doi:10.4028/www.scientific.net/kem.843.64.

[12] V. Gradišnik and D. Gumbarević (November 5th 2018). a-Si:H p-i-n Photodiode as a Biosensor, Advances in Photodetectors - Research and Applications, Kuan Chee, IntechOpen, DOI: 10.5772/intechopen.80503, https://www.intechopen.com/books/advances-in-photodetectors-research-and-applications/a-si-h-p-i-n-photodiode-as-a-biosensor.

[13] M. Pomoni, and P. Kounavis "Determination of trapping-detrapping events, recombination processes and gap-state parameters by modulated photocurrent measurements on amorphous silicon," Philos. Mag., Vol 94, No. 21, pp. 2447-2471, 2014.

[14] J.-P. Kleider, C. Longeaud, and M.-E. Gueunier, "The modulated photocurrent technique: a powerful tool to investigate band gap states in silicon based thin films," Phys. Stat. Sol. (*C*), Vol1, No. 5, pp. 1208-1226, 2004.

[15] M. Niedre, M. S. Patterson, B. C. Wilson, Direct Near-infrared Luminescence Detection of Singlet Oxygen Generated by Photodynamic Therapy in Cells In Vitro and Tissues In Vivo, Photochem. and Photobiol. Vol. 75, No. 4, 2002, pp. 382–391.

[16] S.R. Dhariwal, S.Rajvanshi, Theory of amorphous silicon solar cell (a): numerical analysis, Sol. Energy Mater. Sol. Cells, no. 79, pp. 199-213, 2003.

[17] V. Gradišnik, and A. Linić, "Defects characterization in p-i-n photodiode i-layer," J. Non-Cryst. Solids, Vol. 363, pp. 193-198, 2013.

[18] F. Constantini, C.Sberna, G.Petrucci, M.Reverberi, F.Domenici, C.Fanelli, C.Manetti, G. de Cesare, DeRosa, A.Nascetti, D.Caputo, "Aptamer-based sandwich assay for on chip detection of Ochratoxin A by an array of amorphous silicon photosensors", Sensors and Actuators B: Chemical, Vol. 230, pp. 31-39, 2016.

[19] P. Lu, D. Obata, T. Watanabe, K. Mitsutake, K. Abe, S. Katsuki, H. Akiyama, and C. H. Zhang, "Influence of Intense Pulsed UV Irradiation on the Viability and Proliferation of HeLa Cells," IEEE Trans. Plasma Sci., Vol. 40, No. 8, pp. 2020-2027, Aug. 2012.

Plasmonic Enhanced Photodetectors for Near Infra-red Light Detection

D. Giubertoni[1*], G. Paternoster[1], F. Acerbi[1], X. Borrisé[2], A. Cian[1], A. Filippi[1,3], A. Gola[1], A. Guerrero[4], F. Perez Murano[4], F. Romanato[3], E. Scattolo[1,5] and P. Bellutti[1]

[1] Center for Materials and Microsystems, Fondazione Bruno Kessler, Trento, Italy
[2] Institut Català de Nanociència i Nanotecnologia, Bellaterra (Barcelona), Spain
[3] Department of Physics and Astronomy "G. Galilei", University of Padua, Padua, Italy
[4] Instituto de Microelectrónica de Barcelona – (IMB-CNM, CSIC), Bellaterra (Barcelona), Spain
[5] Libera Università Di Bolzano, Bozen, Italy

* Corresponding author: giuberto@fbk.eu

Abstract - **Silicon based single photon avalanche diodes (SPAD) are able to detect single photons in the visible part of the spectrum with high detection efficiency and high timing resolution. They also provide both single-photon sensitivity and fast responsivity in large-area detectors if arranged in extended arrays as Silicon Photomultipliers (SiPM). However, in applications exploiting near infrared light like light detection and ranging (LiDAR), the detector performance is hindered by the limited Si absorption coefficient. The latter implies absorption depths much larger than the typical active thickness of these devices (10-100 µm against few micrometers) resulting in a quantum efficiency (QE) too low for most of the previous applications. The exploitation of Surface Plasmon Polaritrons (SPP) can convert light in highly-confined modes and enhance the absorption of NIR photons. In this contribution, the first results on the integration of plasmonics nanostructures on thin silicon photodiodes are reported. Electro-optical measurements were carried out and the QE has been measured in the full 400-1100 nm spectrum. The resulting QE on the first prototypes is higher than 7% at 950 nm, an enhancement of about 45% with respect to the reference structure, paving the way for the application of metallic nanograting to SPADs and SiPMs devices.**

Keywords - SPAD; SiPM; surface plasmon polaritrons; Si photodetector; plasmonics.

I. INTRODUCTION

The possibility to detect and accurately measure single photons is a crucial step in a wide range of recent technological applications and scientific experiments. Silicon based single photon avalanche diodes (SPADs) are able to provide both high sensitivity and extremely fast single-photon detection. Furthermore, if arranged in extended arrays they can be the building block of Silicon Photomultipliers (SiPM), providing single-photon sensitivity in large-area detectors [1]. Some emerging applications rely on the photon detection in the red and NIR spectral range, e.g. LiDAR (at 850 - 950 nm wavelengths), NIR spectroscopy, quantum computation, and the detection of light from NIR emitting scintillators. Nowadays, the Photon Detection Efficiency (PDE) of standard SiPMs is strongly limited in the NIR spectral range due to the relatively low Si absorption coefficient [2]. The latter implies absorption depths much larger than the typical active thickness of ("thin") Si SPADs, i.e. few micrometers for the latter against ~18 µm absorption depth at 850 nm and ~150 µm at 1000 nm. Therefore, the detection efficiency is still considered low for most of the previously cited applications.

Latest developments in nanophotonics and plasmonics showed that nanostructured metallic surfaces supporting Surface Plasmon Polaritrons (SPP) could be effectively used to shape light in highly-confined modes and enhance the absorption of NIR photons in thin or poor-absorbing materials like Silicon [3]. Some of these interesting surfaces are: i) 1- and 2-dimensional gratings; ii) bullseye structures; iii) nano-pillars and nano-holes arrays. Nanogratings made of noble metals are very promising structures, also considering the feasibility and the possible integration with SPAD and SiPM technologies, i.e. preserving all the advantages of a Si-based IC technology.

With the aim to investigate and validate the application of metallic nanostructures to SPAD and SiPM devices, in this work, we report on the integration of plasmonic nanostructures on state of art photodetectors (PDs). At the current stage of development, the test devices consist of conventional Silicon photodiodes instead of a proper SPAD, due to the simpler production and characterization procedures with respect to Geiger-mode operations. The photodiodes have been produced at the facility of Fondazione Bruno Kessler (Trento, Italy) by exploiting a custom CMOS-like microfabrication process similar to the one used for FBK-SiPM technology [4].

The produced PDs have been integrated with metal nanograting structures, fabricated directly on the top of the photodiodes by means of Electron Beam Lithography (EBL) followed by Silver evaporation and lift-off. Afterwards, the produced samples have been characterized in terms of Quantum Efficiency (QE) in the full 400-1100 nm spectrum. The paper is organized as follows: in the section II, the device structure and layout are presented

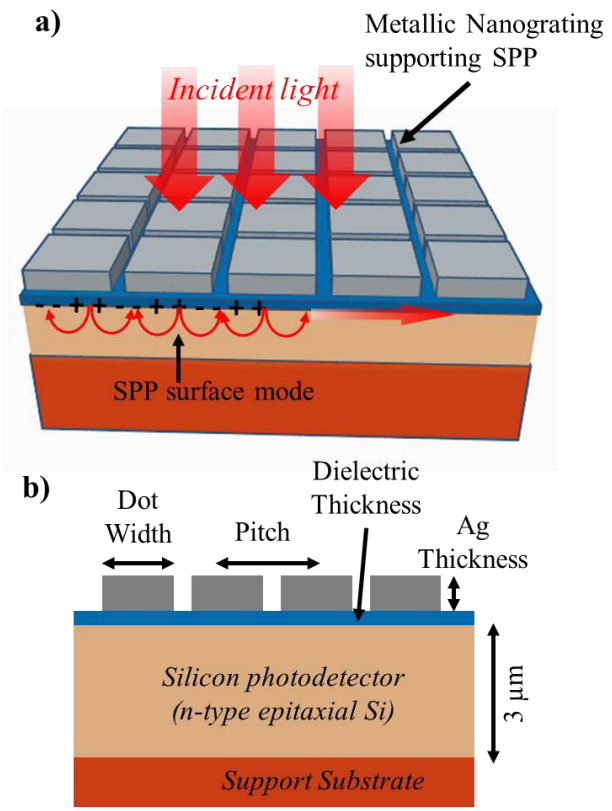


Figure 1. Proposed structure of Metal nanograting fabricated on the top of a Silicon photodiode passivated with a thin dielectric film (a); cross section of the same structure where the parameters optimized in the simulations are indicated (b).

together with the numerical methods used to simulate the optical response of the device; in section III, experimental methods and results from the micro- and nano-fabrication process are reported and discussed; section IV reports the electro-optical characterization of the first prototypes; the conclusions are included in the last section.

II. Device Design and Numerical Simulations

The proposed detector structure is presented in Fig.1, where a not-in-scale 3D sketch and a cross-section are shown. In this design, the silicon photodiode consists of a thin silicon slab (3 µm thick), which is the typical thickness used for SiPMs and SPAD production. The top detector surface is protected by a thin dielectric film (silicon nitride) with a twofold aim: i) to passivate the detector surface and reduce the surface recombination velocity; ii) to act as an optical coupling layer between the detector and the plasmonic nanograting. The latter requires that the dielectric layer thickness has to be precisely tuned, as its thickness and dielectric function define the SPP resonance wavelength. The metal grating supporting SPP is being fabricated directly on the top of the detector in order to lie very close to the active part of the sensor. The grating can be 2-dimensional (periodic "open" strips) or 3-dimensional (dots array, as represented in Fig. 1). In the first case, TM polarized light only excites

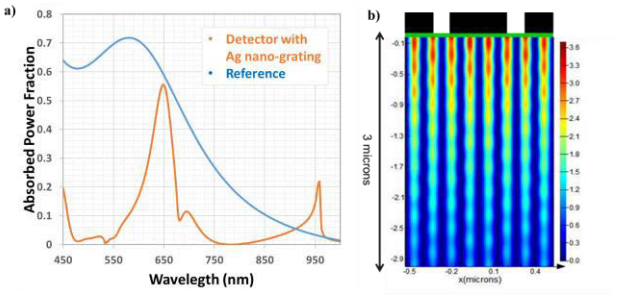


Figure 2. a) Simulated absorption in the 3 µm thick silicon detector with Ag nano-grating (orange solid line) and without (blue solid line); b) absorbed power density in the detector cross-section (a.u) at 952 nm.

SPP resonances as TE modes does not support surface plasmons. In the second case, 3-dimensional numerical simulations show that the same optical response of the 2-d grating appears for both TM and TE polarization, actually leading to a polarization-independent optical response.

In order to finely tune the detector design and estimate the grating optical performance, finite-difference time-domain simulations have been exploited, by using a commercial simulation engine (FDTD Lumerical software [5]). For simplicity, 2-dimensional gratings and TM polarized light have been simulated, while the simulation domain corresponds to a single strip with periodic boundary conditions at the edges. This model setup provides a low simulation time as well as the same level of accuracy obtained with 3-dimensional grating and not-polarized light source. The fraction of the absorbed power in a 3 µm thick Silicon sensitive region has been simulated in the range 450-1100 nm and used as optimization parameters. With Reference to Fig.1b the following parameters have been tuned with the aim of maximizing the absorbed power fraction in the silicon substrate at 952 nm: i) grating pitch; ii) Si_3N_4 thickness; iii) Ag thickness; iv) duty cycle (defined as the ratio between dot width and pitch). The most promising structure features a pitch of 530 nm, duty cycle of about 80%, ~100 nm thick Ag and a thin Si_3N_4 film < 15 nm. The absorbed power fraction as a function of wavelength with this set of parameters is represented in Fig. 2a (orange solid line) together with the absorption curve of 3 µm thick Silicon without Ag grating (blue solid line), used as reference. The simulated spectrum shows multiple absorption peaks that can be related to both plasmonic and hybrid opto-plasmonic resonances. In particular, at 950 nm, a narrow and asymmetrical absorption peak "Fano-like" is present, in correspondence of the Rayleigh singularity wavelength [6]. At that wavelength the absorption reaches the 25%, an eight-fold enhancement with respect to the reference. The physical interpretation of these resonance peaks goes beyond the scope of this paper and can be found in [7]. The map of the absorbed power density at 952 nm, corresponding to maximum of that absorption peak, is reported in Fig. 2b, where it is clearly visible the light confinement in the first micron due to excitation of the hybrid opto-plasmonic mode.

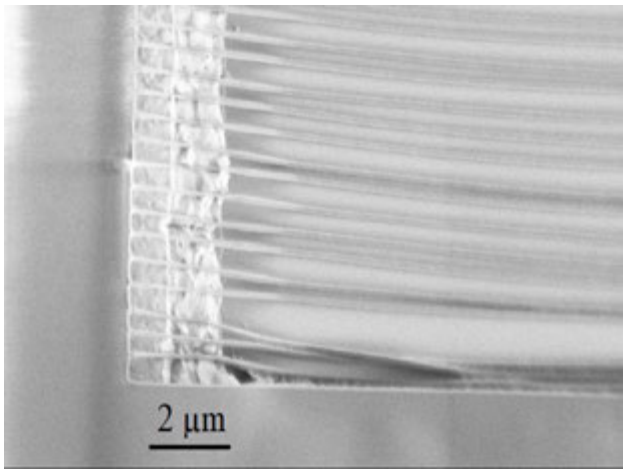


Figure 3. SEM planview of EBL exposed and developed stripe array on PMMA resist, near the edge of the PD. Design linewidth was 340 nm with 535 nm period.

III. MICRO- AND NANO-FABRICATION TECHNOLOGY

A. *Micro- fabrication of photodiodes*

The integration of Ag nanodots in direct contact with the detector top surface poses new challenges in both the micro- and nano-fabrication process, due to the presence of morphology and roughness in the photodiode surface, which hosts electrical contacts and metals for the diode biasing.

The manufacturing technology and the layout of the photodiode have been modified and optimized to host the nanostructure. The most critical parameter of the new diode design is the ultra-thin (< 15 nm) dielectric passivation layer, required in between the nanograting and the Silicon. Small diodes with different side (100 - 500 µm) were produced on a (3.5 ± 0.5) µm thick n-type epitaxial-Silicon wafers, while an ultra-shallow p-n junction has been formed by means of ion implantation. Both the substrate and the junction technology are similar to the one used in the FBK Near Ultra Violet, high-density (NUV-HD) SiPM technology [4], which features detection efficiency peaked at about 400 nm. Despite the target of NIR-wavelength applications, a shallow junction optimized for NUV light detection has been used because most of the plasmonic modes are typically confined in the first hundreds of nanometers from the detector surface.

B. *Nano-Fabrication of the plasmonic arrays*

The process chosen for the nano-fabrication of the dot array is based on positive-tone resist patterning, ultra-high vacuum metal evaporation and lift-off. The resist patterning was carried out by electron beam lithography (EBL) using a Raith150^{TWO} instrument. EBL provides adequate lateral nanometric resolution and flexibility on design parameters. Furthermore, the designed grating can be easily drawn on the surface of the PDs. In order to ensure a successful lift-off, a resist ~3 times thicker than the Ag layer is required. The final choice was to deposit PMMA 950k dilute in anisole (A7) by spin coating (4,500 rpm, 1 minute, curing at 180°C/ 1 minute), then developed with a 30 s dip in a MIBK:IPA 1:3 solution, stopping 30 s in IPA. EBL parameters (electron energy, intensity, dose of exposure) will be discussed in the next paragraphs.

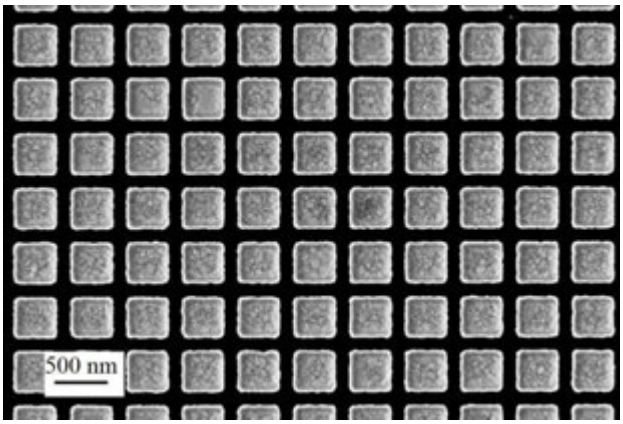


Figure 4. SEM planview of a nanosquare array produced on the surface of a PD with 11.0 nm of Si_3N_4 passivation. Design square size 400x400 nm with 535 nm period, Ag thickness 115 nm. EBL was carried out using a 10 kV/ ~20 pA beam, dose 120 µC/ cm^2.

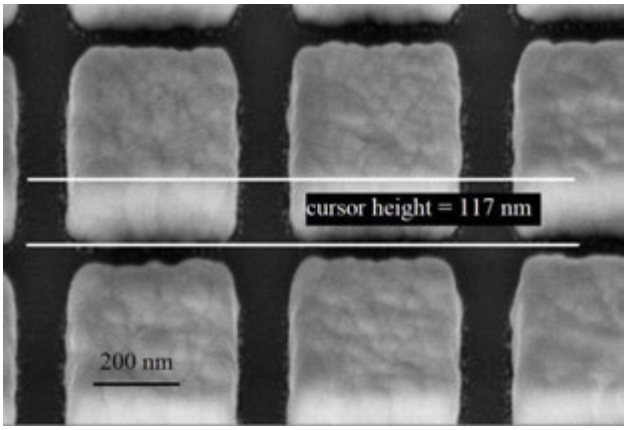


Figure 5. SEM image of the sample of Figure 3 with larger magnification and 45 ° tilt. The measured Ag thickness is 117 nm.

Ag evaporation was carried out in an e-beam evaporator. No adhesion layer was deposited before Ag evaporation to preserve the optical properties of the structures. Lift-off was carried out in acetone at 40 °C without any stirring given the absence of adhesion layers. In order to prevent the oxidation of the metallic pattern the devices were capped by spin coating a thin layer of PMMA resist (4% in Anisole, 4,500 rpm, 1 minute, curing at 180°C/ 1 minute) right after the lift-off process. This passivation allows also an easy opening of the electric contacts by EBL to connect the chip to the electro-optical characterization setup.

The required large duty cycle (up to 80%) implies that structures as wide as 430 nm need to be written by EBL with a period of 535 nm, i.e. leaving unexposed resist areas of only ~100 nm width. This patterning is hindered in this thick resist by the proximity effects, i.e. the unintentional exposure of resist by the electrons backscattered by the substrate. Proximity effects can produce a total pattern overexposure or also only a full exposure under the resist lines among stripes and dots, with possible collapse of the lines. Several combinations of EBL writing conditions were tested and best results were obtained using a 10 kV electron beam. Higher beam energy provided better lateral resolution but did not produce adequate undercuts for lift-off.

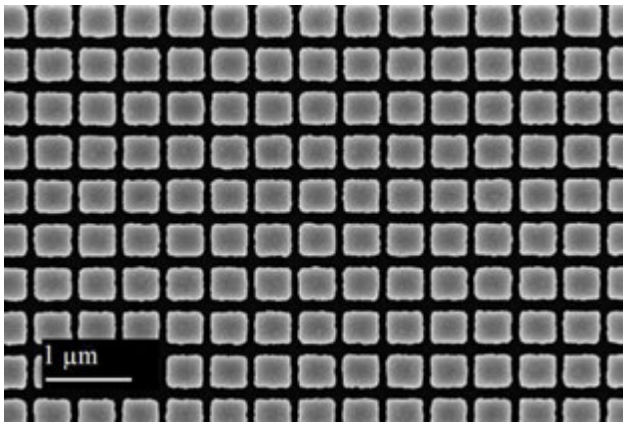


Figure 8. SEM planview of a nanosquare array produced on the surface of a PD with 12.5 nm of Si_3N_4 passivation. Design square size 350x350 nm with 535 nm period, Ag thickness 115 nm. EBL was carried out using a 10 kV/ ~200 pA beam, total dose 120 µC/ cm^2.

Two nano-arrays were tested: stripes and orthogonal array of squares, respectively. For the square arrays, period and duty cycle were always identical in x- and y-directions. Stripe arrays resulted impossible to integrate on the photodiode surface. An example of typical EBL outcome is reported in Figure 3, where the edge of a stripe array is shown. Stripe linewidth was 340 nm with 535 nm periodicity from design. EBL was carried out at low dose (70 µC/ cm^2) with a 10 kV/ ~200 pA electron beam. Even so, the unexposed resist lines among the stripes collapsed as clearly observed from the edge of the array. In general, the ~100 nm unexposed lines and their length (either 100 or 200 µm depending on the chosen PD) did not allow the stability of the resist wall. Therefore, lift-off could not succeed for this kind of design.

A different case was the nanosquare array design. Here the resist structure was somehow self-sustaining thanks to the cross-structure of the lines, and large duty cycles were achieved as reported in Figures 4 and 5 for a PD surface with 11 nm Si_3N_4 (design: 400x400 nm squares, 540 nm period; EBL dose 120 µC/cm^2). It is interesting to notice that these results were achievable only with low current electron beams (~20 pA). In fact, larger intensities of electron beam (~200 pA) resulted in rectangular instead of square shapes like observed in Figure 8, presumably because of some dynamical charging effects. For these reasons, all the samples for the electro-optically characterization were produced with the 10 kV/ 20 pA electron beam condition.

It is interesting to notice that no Ag dots were lost after lift-off despite the lack of an adhesion layer. Furthermore, the large thickness of evaporated Ag resulted in rather rough nanosquares in both planar and vertical directions, due to the polycrystalline nature of the evaporated Ag.

IV. Electro-Optical Characterization

A. Characterization Methods

In order to characterize the first detector prototypes, a custom setup was used. The light source is an unpolarized broadband halogen lamp connected to a monochromator that cover the visible and NIR light spectrum. The monochromator is provided with a variable slit which select the output spectrum width (± 3 nm 3σ). The device under test (DUT) is being mounted to a printed circuit

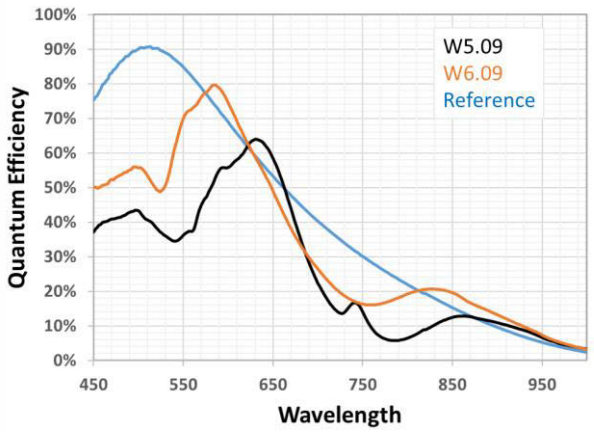


Figure 6. QE as a function of wavelentgh of the two PD with nanogratings (orange and black lines) and of the reference PD without nanograting (blue line).

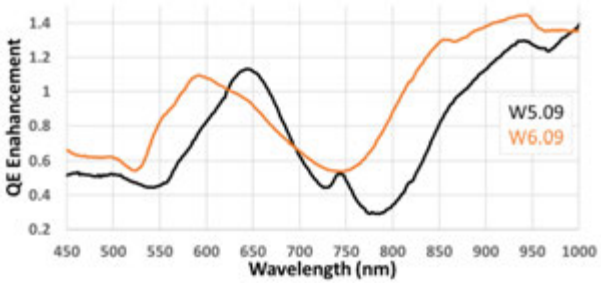


Figure 7. QE enahncement of the two samples with Ag nanograting with respect to the reference PD as a function of wavelentgh.

board (*PCB*) and wire-bonded in order to provide bias to PD. The packaged device is then mounted on a 3-axis stage for detector-beam centering. The actual incident power has been measured by means of a calibrated detector which was moved in place of the DUT during calibration. The PD is biased at 0 V and the photo-generated current is read by means of a semiconductor analyzer with a sensitivity of 1 pA. The measured photo-generated current, corrected for the dark current of the detector is used to calculate the external Quantum Efficiency (QE) of the devices.

B. Characterization Results

Two PD samples with different array geometries have been characterized together with a reference diode (without nano-grating) fabricated on the same silicon chip. The active area of PDs (fully covered by the nanodot array) is 200 x 200 µm^2. The two samples differ in terms of duty cycle of the nano-array and Si_3N_4 layer thickness, as shown in the Table I, which reports the nominal values of the main geometrical parameters of the samples.

TABLE I. Geometrical Parameters of the Samples

Sample ID	W5.09	W6.09	Reference
Grating Pitch	535 nm	536 nm	-
Duty Cycle	76 %	66 %	-
Si_3N_4 Thickness	11 nm	6.6 nm	11 nm
Ag Thickness	115 nm	115 nm	-

The experimental QE of both the PDs provided with dots-arrays (black and orange solid lines for W5.09 and W6.09, respectively) together with the reference diode (blue solid line) are reported in Fig. 6, while the ratio between the QE of the two PDs with dots-arrays and the QE of the reference diode are reported in Fig. 7. It is worth noting that the experimental QE reproduces the main features of the simulated absorption spectrum (orange curve in Fig. 2a). However, no sign of the narrow "Fano-like" peak predicted by the simulations at 950 nm can be found in the measurements, but a broader peak is present at shorter wavelengths (about 850 nm). Probably, the difference between simulated and measured QE in the NIR spectral region, seems to be mostly due to the difference of dielectric and metal thicknesses of the samples with respect to the optimal values suggested by numerical simulations. Deeper investigations are still ongoing and new samples will be soon available.

Even if the performance of the first prototypes are still far from the ideal case suggested by numerical simulation, the QE enhancement in the NIR region reaches the 45% at 950 nm in the case of the sample W6.09, a very remarkable result that pave the way for further optimizations.

V. CONCLUSION

The integration of silver nanosquare arrays supporting Surface Plasmon Polaritrons (SPP) increases the light absorption of NIR photons on Si photodiodes. In particular the plasmonic arrays can convert light in highly-confined modes reducing the NIR photons absorption depth to values comparable to the typical active regions of Si photodetectors. The first promising results of this approach have been here reported, showing a resulting QE higher than 7% at 950 nm, i.e. an enhancement of about 45% with respect to the reference structure, paving the way for the application of metallic nanograting to SPADs and SiPMs devices.

ACKNOWLEDGMENT

This project has received funding from the EU-H2020 research and innovation programme under Grant Agreement 777222 ATTRACT project "PlaSiPM" and under grant agreement No 654360 NFFA-Europe"

REFERENCES

[1] P. Buzhan et al., "Large area silicon photomultipliers: performance and applications," Nucl. Instrum. Methods Phys. Res. A,, vol. 567.1, pp. 78-82, 2006.

[2] F. Acerbi, G. Paternoster, A. Gola, N. Zorzi and C. Piemonte, "Silicon photomultipliers and single-photon avalanche diodes with enhanced NIR detection efficiency at FBK," Nucl. Instrum. Methods Phys. Res. A,vol. 912, pp. 309-314, 2018.

[3] P. Berini, Laser and Photonics Reviews, vol. 8.2, pp. 197-220, 2014, DOI: 10.1002/lpor.201300019

[4] C. Piemonte et al., "Performance of NUV-HD silicon photomultiplier technology, " IEEE Transactions on Electron Devices, vol. 63.3, pp. 1111-1116, 2016.

[5] Lumerical Inc. https://www.lumerical.com/products/.

[6] A. Hessel and A. A. Oliner,"A New Theory of Wood's Anomalies on Optical Gratings," Appl. Opt., vol. 4.10, pp. 1275-1297, 1965.

[7] A. Filippi, "Improving Silicon Photodetectors NIR Responsivity via Hybrid Opto-Plasmonic Resonances" PhD diss., Universita' degli Studi di Padova, 2020.

Variable Angle Spectroscopic Ellipsometry Study of Poly(3,4-ethylenedioxythiophene):Polystyrene Sulfonate Thin Films in Contact with Air

Gaurav Pathak[†§], Duško Čakara[†‡*]
[†]Centre for Micro- and Nanosciences and Technologies
[‡]Department of Biotechnology
University of Rijeka, Radmile Matejčić 2, HR-51000 Rijeka, Croatia

[§]Current address: Centre de Microélectronique de Provence, Ecole Nationale Supérieure des Mines, 13541 Gardanne, France
[*]Corresponding author: dcakara@uniri.hr

***Abstract*—Thin films of conducting polymer complex poly(3,4-ethylenedioxythiophene) : polystyrene sulfonate (PEDOT:PSS) are studied by means of varied angle spectroscopic ellipsometry (VASE) in the wavelength range 370-880 nm, at five different angles of incidence (AOI). The goal is to simultaneously determine the film thicknesses and its dielectric function. Samples of varied film thickness are prepared by spin coating the aqueous suspension onto Si/SiO_2 wafers at different spin rates, and subsequent annealing at 170 °C. The measured spectra at constant angles of incidence are fitted by applying Fresnel's equations for a stratified optical layers model (SLM) which postulates several optical layers with well defined optical interfaces, in combination with the dielectric function $\tilde{\epsilon}(\lambda)$. Different possibilities for data fitting are critically assessed, pointing to some pitfalls in the VASE data analysis and interpretation. Two methods which involve the assessment of a theoretical Lorentz-Drude (LD) dielectric function were compared: one in which a unique function is assumed for all of the prepared samples (i.e. film thicknesses), while in the other, a variation of the dielectric function was permitted. The validation criteria include comparisons with the experimental film thickness measured by means of the atomic force microscope, the quasi-experimental dielectric function fitted independently for each wavelength, and the literature value of the main optical transition energy calculated by quantum mechanical modelling. In the limit of the studied system and model, all of these comparisons point to the necessity of simultaneously fitting the VASE data measured for films of varied thickness, in which case both the thickness and the dielectric function may be determined.**

***Index Terms*—ellipsometry, spectroscopic, PEDOT, conducting, polymer, film**

I. Introduction

The conducting polymer complex poly(3,4-ethylenedioxythiophene):polystyrene sulfonate (PEDOT:PSS) is a widely used thin film material in the field of organic electronics [1], [2], as well as in several other fields [3]. The conduction mechanism of PEDOT in PEDOT:PSS films relies on the fact that it is an electron-deficient (oxidized) π-conjugated polymer. Thus, in its conductive state, PEDOT is doped with holes and behaves as a p-type semiconductor. Meanwhile, the negatively charged PSS electrostatically stabilizes the doping state of PEDOT and makes it dispersable in aqueous medium [4]. Such PEDOT:PSS suspensions are technologically easily processable and mostly used in the form of thin (semi)conducting films, deposited on surfaces by various techniques (casting, spin-coating, ink-jet and roll-to-roll printing etc.). The conductivity of those films as function of chemical environment and in relation with different applications, presents their mostly studied property [1], [5]–[7]. Another attractive field of implementation is that of light emitting and harvesting devices, which rely on the optoelectronic properties of the film, and where there is still an insufficient level of both theoretical understanding and experimental information. Several available experimental studies focused on the electrochromism of electropoymerized PEDOT [8]–[10], and some on the dielectric function of a PEDOT:PSS film [6], [7]. Since the latter function presents a very important feature for the optoelectronic porperties of a film material, the methods for its experimental assessment deserve special attention.

Optoelectronic properties of a conducting or semiconducting material, including conductivity and the intensity of optical absorption and reflection, are derived from the *dielectric function* [11], [12]. The latter is a complex function for semitransparent materials, where the real part is related to the speed of light, while the imaginary part is related to the intensity of light absorption within a material. The electronic conductivity of a material is directly proportional to the free charge carrier density (either electrons or holes) which is detectable in the near infrared region of the dielectric function as the Drude "tail" [13]. Meanwhile, the density and symmetry of the electronic states entering the transition moment (subject

This work was accomplished within the project FP7-PEOLPLE-ITN-2013 (PITN-GA-2013-607896) "Organic bioelectronics", and with support of the European Fund for Regional Development and the Ministry of Science, Education and Sports of the Republic of Croatia, project Research Infrastructure for Campus-based Laboratories at the University of Rijeka (grant number RC.2.2.06-0001).

to selection rules), determine the optical absorption detectable in the ultraviolet and visible (UV-VIS) region [14]. In the case of PEDOT, the electronic states include the valence band (VB), conduction band (CB) and two types of intra-gap phonon-electron coupled states known as polarons and bipolarons [15]. The latter states are the most important for an ample conductivity of PEDOT, and their concentration depends on the doping state which is mainly regulated by the oxidation state.

In general, for most conducting and semi-conducting materials, the dielectric constant is actually not a constant, but a function of the wavelength of the incident electromagnetic radiation to which the material is exposed. This gives rise to different absorption and reflection of light depending on the wavelength, which may be measured with spectroscopic techniques. While the standard transmission spectrophotometry may be implemented for measuring the imaginary part of the dielectric function, the experimental determination of both the real and imaginary parts is possible by applying the spectroscopic ellipsometry techniques. Once the dielectric function is determined, it can be related to the conductivity of the material by invoking the laws of electrodynamics [16].

In the present paper, the dielectric function of a PEDOT:PSS film is experimentally determined from the spectroscopic ellipsometry measurements at varied angles of incidence (i.e. Variable angle spectroscopic ellipsometry, VASE). In addition to the dielectric function, this technique is sensitive to the thickness of the film, which on one hand suggests a possibility of simultaneous film thickness determination. However, since the interpretation of the spectroscopic ellipsometry data depends on the optical model of the sample and calibration with respect to film thickness is not possible (in contrast to standard spectrophotometry), caution is necessary. The optical model often involves a relatively large number of parameters, whereby the determination of both thickness and dielectric function becomes questionable. The aim of the present work is to test several approaches for fitting VASE data measured for a PEDOT:PSS thin film, in order to assess their applicability for simultaneous determination of the dielectric function and film thickness. For this task, it was necessary to perform VASE measurements for samples of varied thickness, which was independently determined from the atomic force microscope (AFM) scans. This work presents a part of a wider effort to establish and validate methods for studying the electrodynamic behavior of such films in the working conditions for organic bioelectronic devices [17]–[19].

Ellipsometry

In standard ellipsometry and reflectometry, the change in the polarization state of monochromatic light upon reflection from a thin semi-transparent film, expressed as the ellipsometric angles Ψ and Δ, is measured [11], [12], [20]. The ellipsometric angles are related to the film thickness via the Fresnel's equations, which involve the complex refractive index ($\tilde{n} = n + ik$) or equivalently the complex dielectric constant ($\tilde{\epsilon} = \tilde{n}^2 = \epsilon_1 + i\epsilon_2$), the thickness d and the

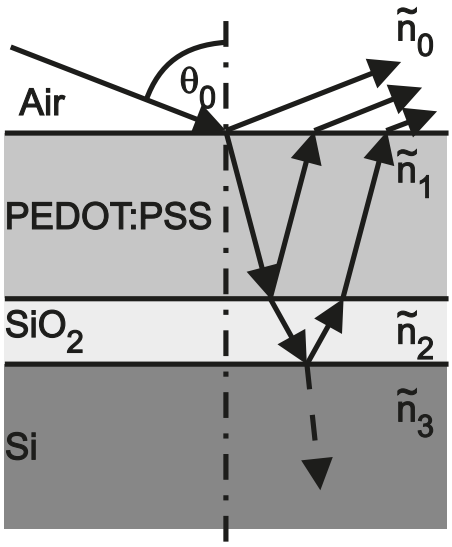


Fig. 1. Stratified layers model for the Si/SiO_2 substrate with the deposited PEDOT:PSS film

angle of incidence (AOI) θ_0. The optical model consists of a stack of m optically homogeneous layers with sharp phase boundaries identified with an abrupt change of the refractive index (stratified layers model, SLM), as depicted in Fig. 1. The Fresnel's reflection coefficients for an optical interface between two infinite homogeneous optical materials, are defined as:

$$\rho_{\sigma,\tilde{n}_j,\tilde{n}_{j+1}} = \frac{\tilde{n}_j \cos\theta_j - \tilde{n}_{j+1}\cos\theta_{j+1}}{\tilde{n}_j \cos\theta_j + \tilde{n}_{j+1}\cos\theta_{j+1}} \tag{1a}$$

$$\rho_{\pi,\tilde{n}_j,\tilde{n}_{j+1}} = \frac{\tilde{n}_{j+1} \cos\theta_j - \tilde{n}_j\cos\theta_{j+1}}{\tilde{n}_{j+1} \cos\theta_j + \tilde{n}_j\cos\theta_{j+1}} \tag{1b}$$

where subscripts σ and π denote Fresnel's reflection coefficients for the σ- and π-components of the elliptically polarized light, reflecting from an interface between the two optical media with refractive indexes $\tilde{n}_j$ and $\tilde{n}_{j+1}$, while θ_j and θ_{j+1} denote the angles of the incident and refracted beam, respectively, to the surface normal. Thus, θ_j and θ_{j+1} are related through Snell's law

$$\tilde{n}_j \sin\theta_j = \tilde{n}_{j+1}\sin\theta_{j+1}. \tag{2}$$

For a semi-transparent thin film with a refractive index $\tilde{n}_1$, deposited on a semi-transparent but infinite substrate with a refractive index $\tilde{n}_s$, and in contact with a medium with a refractive index $\tilde{n}_0$ (with reference to Fig. 1, $\tilde{n}_s$ would replace $\tilde{n}_2$ and $\tilde{n}_3$ as the refractive index of a single "effective" optical medium), Fresnel's coefficients are a property of the studied materials (medium, thin film, substrate). Their relationship with the ellipsometric angles Ψ and Δ, as a property of the measured EM radiation and the angle of incidence, is defined through the total reflection ratios for the σ- and π-parallel components:

$$P_\sigma = \frac{\rho_{\sigma,\tilde{n}_0,\tilde{n}_1} + \rho_{\sigma,\tilde{n}_1,\tilde{n}_s} e^{-i2\beta}}{1 + \rho_{\sigma,\tilde{n}_0,\tilde{n}_1}\rho_{\sigma,\tilde{n}_1,\tilde{n}_s} e^{-i2\beta}} \tag{3a}$$

$$P_\pi = \frac{\rho_{\pi,\tilde{n}_0,\tilde{n}_1} + \rho_{\pi,\tilde{n}_1,\tilde{n}_s} e^{-i2\beta}}{1 + \rho_{\pi,\tilde{n}_0,\tilde{n}_1}\rho_{\pi,\tilde{n}_1,\tilde{n}_s} e^{-i2\beta}} \tag{3b}$$

where $\beta = 2\pi d/\lambda$. The ellipsometric angles Ψ and Δ are related to P_σ and P_π through the general ellipsometry equation:

$$P = \frac{P_\pi}{P_\sigma} = \tan(\Psi) e^{i\Delta}. \tag{4}$$

Eqns. 1 - 4 form a closed set which, for a system composed of three phases separated by two optical interfaces, can be solved analytically (for instance, see [20] p. 35). The general solution for a larger number of layers (four in the present case, cf. Fig. 1) is more complicated and involves formulation of the total reflection ratio in terms of transfer matrix expansion (for a very concise explanation, see for instance Appendix in [21]), which can be solved computationally.

Thus, for calculating Ψ and Δ starting from an optical model, a certain stack of semi-transparent optical layers deposited on a substrate first needs to be postulated. If the thicknesses and complex refractive indexes for all the layers except for the film are known, the thickness and the complex refractive index of the film can be numerically fitted from the experimental VASE spectra $\Psi(\theta_0, \lambda)$ and $\Delta(\theta_0, \lambda)$.

Spectroscopic ellipsometry - the dielectric function

Both the optical and electronic properties of a material depend on the wavelength of the incident EM radiation, and are mutually related through the complex refractive index dispersion [$\tilde{n}(\lambda) = n(\lambda) + ik(\lambda)$] or equivalently complex dielectric function [$\tilde{\epsilon}(\lambda) = \tilde{n}^2(\lambda) = \epsilon_1(\lambda) + i\epsilon_2(\lambda)$]. For materials which absorb light in addition to reflection and refraction, the dielectric function is complex ($\epsilon_2 > 0$). Then, according to classical electrodynamics, the electromagnetic radiation can interact with electrons in the material according to a forced and damped harmonic oscillator model, also known as the Lorentz oscillator [11], [12], [20]. If there is more than one type of electronic transitions present in the material, the dielectric function according to this model can be written as

$$\tilde{\epsilon} = \epsilon_\infty + \sum_j \frac{A_j}{E_{0,j}^2 - E^2 - i\Gamma_j E} \tag{5}$$

where E is the energy of the incoming photons, $E_{0,j}$ is the resonant energy of the electrons, A_j is called the oscillator strength and is proportional to the density of electrons in the state j, while Γ_j is the damping frequency which can be identified with the inverse mean free time for scattered electrons, $\Gamma \equiv 1/\tau$. By applying Planck's law, all the energies in the above expression can be substituted by frequencies. Thus for example $E_{0,j}$, which represents the elastic potential energy of electrons bound in chemical bonds of a material as defined by Hooke's law, can be substituted by the resonant frequency ($\omega_{0,j} = E_{0,j}/\hbar$). A special case of the general Lorentz oscillator model equation (cf. eq. 5) is the *Lorentz-Drude* (LD) model, where the resonant energy is set to zero for one of the oscillators, and which represents the dielectric function of free electrons modelled as electron (Fermi) gas. The strength of a Lorentz oscillator is given as $A = \hbar^2 e^2 N_\mathrm{e}/\epsilon_0 m_\mathrm{e}$, where N_e denotes the number of electrons involved in the optical transition. For the Drude oscillator, the resting mass of electron m_e is replaced by the effective mass m^* of a hole or electron, and the latter is empirical (for some conducting polymers, a ratio $m^*/m_\mathrm{e} \approx 10^{-2}$ was estimated from quantum mechanical modelling [22]). Eq. 5 is written for an optically isotropic medium (film), while in the case of anisotropy, $\tilde{\epsilon}(\lambda)$ has to be separately written for each optical axis. In the latter case, the Fresnel's reflection coefficients also depend on the direction in the laboratory coordinate system. It is worth noting that a generalization and extension of the Drude term in above theoretical dielectric function were proposed for fitting the experimental spectroscopic and conductivity data of the solid state semiconductors and (semi)conducting polymers (particularly in the THz frequency range) [23]–[25], however the discussion in the present case is narrowed to the above simple form which applies to the wavelength range studied in this work.

Variable angle spectroscopic ellipsometry (VASE)

The dielectric function for a thin film can be determined from spectroscopic ellipsometry data [11], [12], by numerically fitting Ψ and Δ measured at varied wavelength, with the values calculated by introducing a model dielectric function for the film, into Fresnel's equations (1). Another possibility is to fit $\epsilon_1(\lambda)$ and $\epsilon_2(\lambda)$ as parameters, from the data measured at varied AOI. The latter is known as the "wavelength-by-wavelength" (WbW) fitting. If the dielectric function is *a priori* known, the film thickness can be fitted from Fresnel's equations, which is frequently used in the case of thin solid state films. On the other hand, if the film thickness is reliably known, the parameters of the dielectric function can be obtained by fitting. However, if both are not known *a priori*, the fitting of both $\tilde{\epsilon}(\lambda)$ and d will rarely be accurate. In that case, it can be helpful to perform the spectroscopic ellipsometry measurements at varied angles of incidence (VASE), and simultaneously fit the Ψ and Δ spectra measured at different AOI.

EXPERIMENTAL PART

The high-conductivity grade PEDOT:PSS (Orgacon HIL-1005) was purchased as aqueous suspension (1.1 % w/w) from Sigma Aldrich, filtered (Chromafil Xtra CA-20/25 cellulose aceteate filters, 0.2 μm) and spin-coated onto silicon substrates (1 cm x 1 cm) at 2000, 3000, 4000, 5000 rpm, by using a computer-controlled spin-coater (SC100, Metallic Nova, Croatia). The sheet resistance of the product, as reported by the manufacturer, is 50-120 Ω/sq, and the pH around 2. After spin-coating, the films were dried and and annealed for 20 min. at 170 °C on a hot plate.

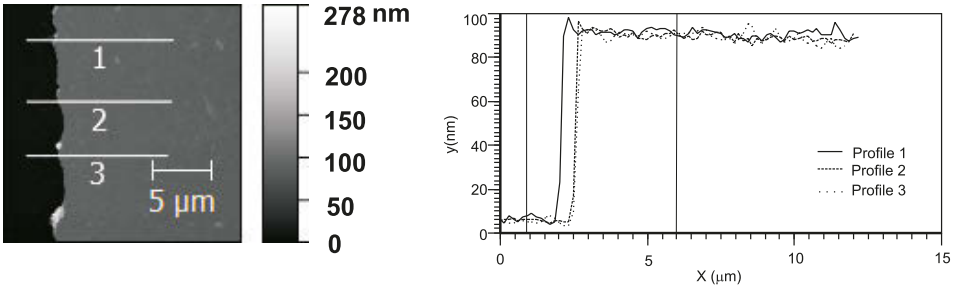


Fig. 2. Scanning atomic force micrograph of the PEDOT:PSS film spin-coated at the Si/SiO_2 wafer at spinning rate of 2000 rpm (left), and the corresponding height profile diagram, determined along the three lines indicated in the micrograph.

AFM measurements were performed with a Bruker Dimension Icon instrument. A step was created in all the samples by scratching the film with a sharp tweezer. Tapping mode AFM scans were performed across the step, at three different spots on the film surface, with the scanned area of 20 x 20 μm^2. The height profile images obtained from the scan were processed by using the Gwyddion 2.44 software. A flattening procedure was performed to remove the so called "bow/tilt" artefact in the height profile image. Three line profiles were extracted per scan, from which an average thickness was calculated.

The VASE measurements were performed with the Accurion EP4 spectroscopic imaging ellipsometer, at room conditions (297.15 K, in absence of humidity control). The computer-controlled instrument applies the nulling method for determining the ellipsometric angles Ψ and Δ (nulling ellipsometry). The sample was first aligned perpendicular to the goniometer axis and the optical plane defined by the incoming light beam and the axis of the reflected beam, by using an in-built alignment sensor. For all the measurements, the 5x optical objective was used at the detector arm. After a perfect alignment, a spectroscopic ellipsometry experiment was performed, in which the spectra were measured in the wavelength range of 360-880 nm. Twenty points were accumulated for each spectrum, at five angles of incidence (AOI = 45°, 50°, 55°, 60°, 65°). The measurement was performed in two zones of polarizer and analyzer positions separated by 180°, from which Ψ and Δ were calculated as average. All the model fittings were performed by using the Accurion EP4 Model software.

Another set of experiments was performed with the same samples as above, in which AOI was varied in the range $45° - 65°$ with a step of $1°$. The acquisition of Ψ and Δ was done for 20 different wavelengths in the same range as above. This was necessary in order to collect sufficient data for comparing the theoretical dielectric function fitted from $\Psi(\lambda)$ and $\Delta(\lambda)$ spectra, with the quasi-experimental dielectric function obtained by applying the wavelength-by-wavelength (WbW) fitting method (see below).

Data fitting methods

The Ψ and Δ spectra, measured by VASE for PEDOT:PSS thin films of varied thickness at a Si/SiO_2 substrate, in the visible range of incident radiation, are fitted and analyzed in terms of the stratified optical layers model (cf. Fig. 1). For the substrate, the dielectric functions of crystalline Si(1,0,0) and amorphous SiO_2 were used as cited in [26]. For the film, both the parameters of the Lorentz-Drude model dielectric function, as well as the film thickness, were fitted as unknown parameters. The film thickness obtained from fitting was compared to the values measured with the scanning atomic force microscope (AFM), as one of the criteria for evaluating the quality of the fitted dielectric function.

Fitting a model dielectric function and the film thickness from the raw VASE data $[\Psi(\lambda,\theta_0),\ \Delta(\lambda,\theta_0)]$ is in principle ambiguous. As will be discussed, if all of the parameters involved in the Fresnel's equations are fitted (the dielectric function parameters and film thickness), their mutual dependence will cause poor statistical determination. This is a very common problem of parameter cross-correlations when fitting functions involving (too) many parameters (also known as *overfitting*). In this case a reduction of the model, fixing some of the parameters, or simultaneous data fitting for experiments selectively designed in order to better express some of the parameters (as for example film thickness in the present case), is necessary. This also invokes validation of the data fitting methods prior to presenting and discussing the results.

The experimental ellipsometric spectra measured at varied angle of incidence (VASE spectra) of the spin-coated PEDOT:PSS films, were fitted with the stratified layers model as described in Fig. 1, in combination either with an isotropic LD dielectric function for the film, or a "quasi-experimental" isotropic dielectric function obtained by applying the WbW method. Thus in the first approach (LD1 and LD2 methods), the complex dielectric function for the film is calculated from the LD model as a sum of a Lorentz oscillator, a Drude oscillator, and a constant term:

$$\tilde{\epsilon} = \epsilon_\infty + \frac{A_\mathrm{L}}{E_0^2 - E^2 - i\Gamma_\mathrm{L}E} + \frac{A_\mathrm{D}}{-E^2 - i\Gamma_\mathrm{D}E} \qquad (6)$$

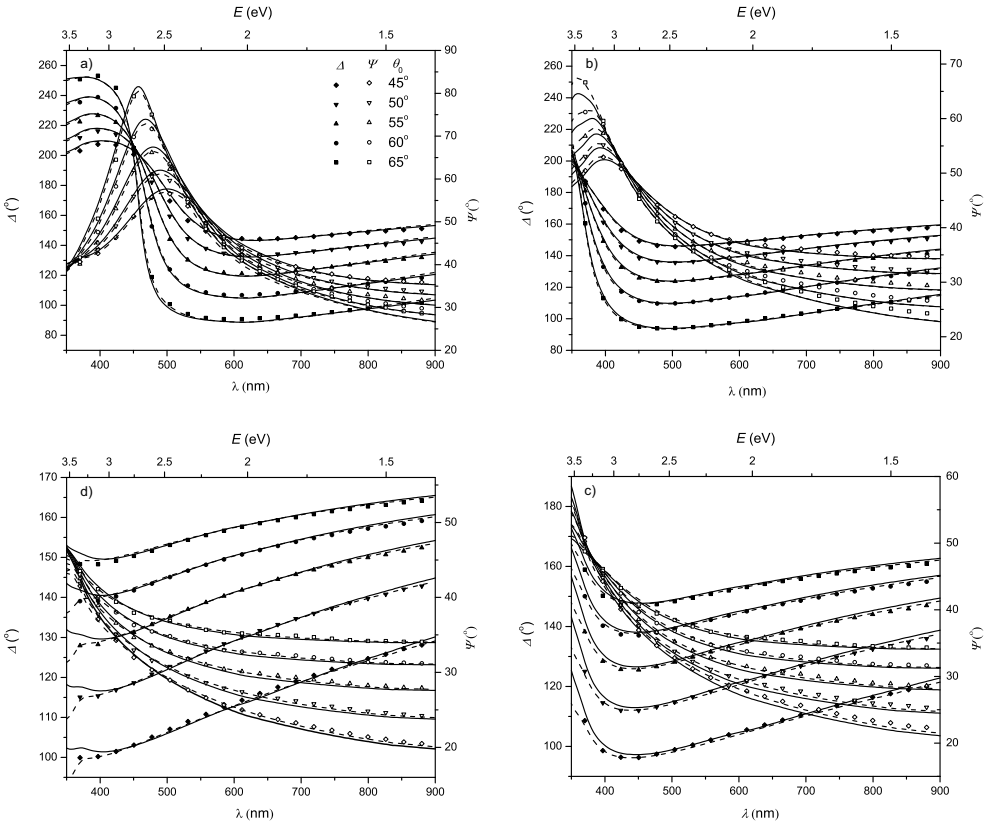


Fig. 3. Experimental VASE spectra ($T = 297.15$K, symbols as indicated in the legend in a), for the studied PEDOT:PSS films on Si/SiO_2, spin-coated at different spin rates: a) 5000 rpm; b) 4000 rpm; c) 3000 rpm; d) 2000 rpm. Solid and dashed lines represent the best fit functions obtained with the Lorentz-Drude model dielectric function 6, by applying methods LD1 and LD2, respectively.

At the right hand side, $E = hc/\lambda$ is the independent variable, while all the rest are treated as parameters. Indices L and D denote the parameters for the Lorentz and Drude oscillators, respectively, A and Γ describe the oscillator strength and damping, respectively, E_0 is the Lorentz oscillator resonant frequency and ϵ_∞ is the purely real high-frequency dielectric constant. Since the dielectric function is treated as continuous over the whole wavelength domain, this approach is suitable for fitting $\Psi(\lambda)$ and $\Delta(\lambda)$ at fixed AOI. The two approaches denoted as LD1 and LD2 differ in the following: in LD1, a common dielectric function was fitted for all the measured ellipsometric spectra at varied AOI and film thicknesses. This results with a common set of the dielectric function parameters for all the film thicknesses. In another method denoted as LD2, a common dielectric function was fitted for the ellipsometric spectra at varied AOI, however spectra for varied film thickness were fitted separately. This allowed fitting different LD dielectric function parameters for different film thicknesses. Since the described optical models already involve a significant number of parameters (six of them stemming from the LD dielectric function and the thickness), these were tested only as isotropic. As will be shown in the results, the choice of only one Lorentz, and the simplest form of the Drude oscillator in the theoretical dielectric function (cf. eq. 6) is in the present case justified, since the introduction of additional oscillators or terms would certainly result in overfitting of the ellipsometric data within the available experimental wavelength range.

The WbW method was applied for fitting the $\Psi(\theta_0)$ and $\Delta(\theta_0)$ at each measured wavelength. In this method, the film thickness was fixed to the value measured with AFM, while $\epsilon_1(\lambda)$ and $\epsilon_2(\lambda)$ were fitted. In this way, best fit values for $\tilde{\epsilon}$ are obtained independently at each wavelength. The resulting points $[\epsilon_1, \lambda]$, and $[\epsilon_2, \lambda]$ may be regarded as "quasi-experimental". Although not free from assumptions laid down in the SLM, WbW presents the method for fitting VASE data in the case when the dielectric function of the film is not known *a priori*, which is least dependent on the model. As such, it may be used as reference for validation of the LD1 and LD2 model dielectric functions.

RESULTS AND DISCUSSION

A typical AFM image obtained for the studied samples is presented in Fig. 2, along with the calculated height profile. The experimental values of the film thicknesses determined from the AFM height profiles measured for different applied spin-coating speeds, are tabulated in Table I, right column.

The VASE spectra for the four studied PEDOT:PSS films with different thicknesses are presented in Fig. 3. The solid

TABLE I
PEDOT:PSS FILM THICKNESSES OBTAINED BY FITTING THE LORENTZ-DRUDE MODEL DIELECTRIC FUNCTION TO THE ELLIPSOMETRIC SPECTRA PRESENTED IN FIG. 3, COMPARED WITH THE THICKNESS MEASURED WITH AFM. IN LD1 METHOD, THE ELLIPSOMETRIC SPECTRA MEASURED FOR DIFFERENT FILM THICKNESSES WERE FITTED BY A COMMON DIELECTRIC FUNCTION. IN LD2, SPECTRA AT VARIED AOI WERE FITTED BY A COMMON DIELECTRIC FUNCTION, SEPARATELY FOR EACH THICKNESS.

Spin speed (rpm)	d (LD1) (nm)	d (LD2) (nm)	d (AFM) (nm)
5000	45.3	48.7	42.50 ± 1.93
4000	54.5	58.0	52.36 ± 2.82
3000	65.0	65.1	59.50 ± 2.58
2000	86.6	89.2	87.88 ± 4.20

lines represent the best fit by the LD1 method, where the data for all the studied samples were fitted simultaneously. The thereby obtained film thickness is presented in Table I and compared with the values obtained from the AFM measurements. A relatively good agreement with these values may be observed, corroborating the applicability of the LD1 fitting method. However in overall, the thickness measured with AFM is by 1-5 % smaller than the values determined from the LD1 fits. The fitted LD1 dielectric function parameters (cf. eq. 6) are listed in the top row of Table II. The RMSE calculated for all 800 data points (for each sample and each of the five angles of incidence, a spectrum contains twenty of both Ψ and Δ measurements) equals 1.45 degrees. From this value it is observed that the fits obtained by the LD1 method are indeed good, in spite of a moderate number of fitted parameters. A deviation of the LD1 curves from the measured data points is observed for the two thinner films (cf. Figs. 3 c) and d). This deviation is more pronounced at lower wavelengths for Δ, while at higher wavelengths for Ψ.

The best fit common dielectric function calculated from the fitted parameters of the Lorentz-Drude model (LD1 method) is presented in Fig. 4. The magnitude of the real part of the dielectric function (ϵ_1) is predominately determined by ϵ_∞, but a mild monotonically decreasing trend with increasing wavelength can also be detected. The latter trend occurs due to the Drude oscillator pull. As can be inferred from the comparison of the two oscillator strengths (A_L vs. A_D, first row in Table I), the Drude oscillator is much stronger than Lorentz. Thus the imaginary part (ϵ_2) of the LD1 dielectric function exhibits a very weak absorption peak around 2 eV.

The dashed lines in Fig. 3 represent the spectra modelled by applying the LD2 method, thus fitting the dielectric function parameters and thickness, both as characteristic for each sample, and keeping those parameters common for all AOI. The fitted curves almost perfectly coincide with the measured data in the whole range of the studied wavelength (except for the sample prepared at 2000 rpm). This may be also evidenced as low RMSE values for the LD2 fits (cf. Table II lower part), which are lower than in the case of LD1 fit. The fitted film thickness is somewhat greater than the LD1 value, and is by ca 10 % greater than the thickness measured with AFM.

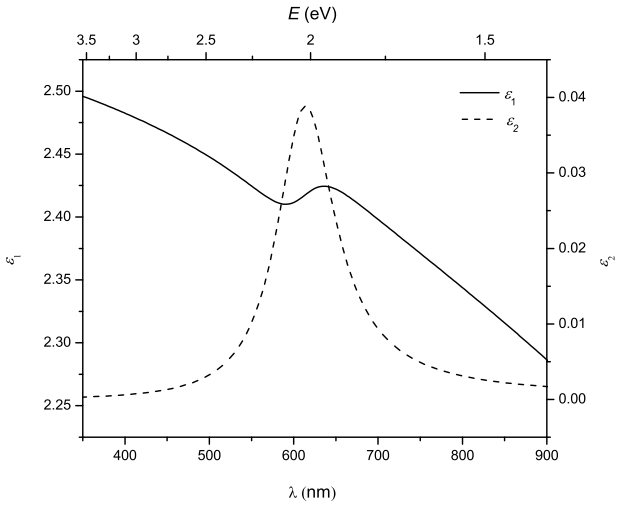


Fig. 4. Best fit common Lorentz-Drude dielectric function (cf. eq. 6) for all the studied films at 297.15 K, obtained by simultaneously fitting all of the measured ellipsometric spectra (method LD1). The corresponding film thicknesses fitted by this method are listed in Table I.

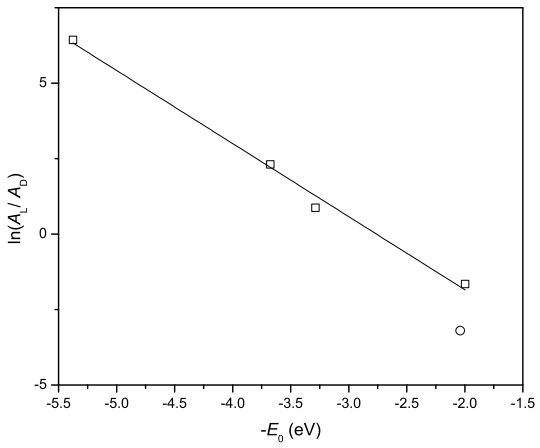


Fig. 5. Correlation between the $\ln(A_\mathrm{L}/A_\mathrm{D})$ and E_0, obtained from LD2 (squares) and LD1 (circle) methods for samples of varied film thicknesses. The fitted straight line $y = -2.4212x - 6.6853$; $R^2 = 0.9936$ was calculated only from the LD2 parameters (squares).

When considering the latter as a reference, one has to keep in mind that the AFM micrographs were scanned in a local area of 400 $\mu\mathrm{m}^2$, while the ellipsometric signal was collected over a much larger area, such that the discrepancies could be caused by a film thickness variation in lateral directions. The values of A_L and A_D parameters obtained from these fits (cf. Table II, lower part) are not coherent, whereby a significant variance in the strengths of the two oscillators can be detected. Thus, while the Lorentz oscillator is dominant for the samples

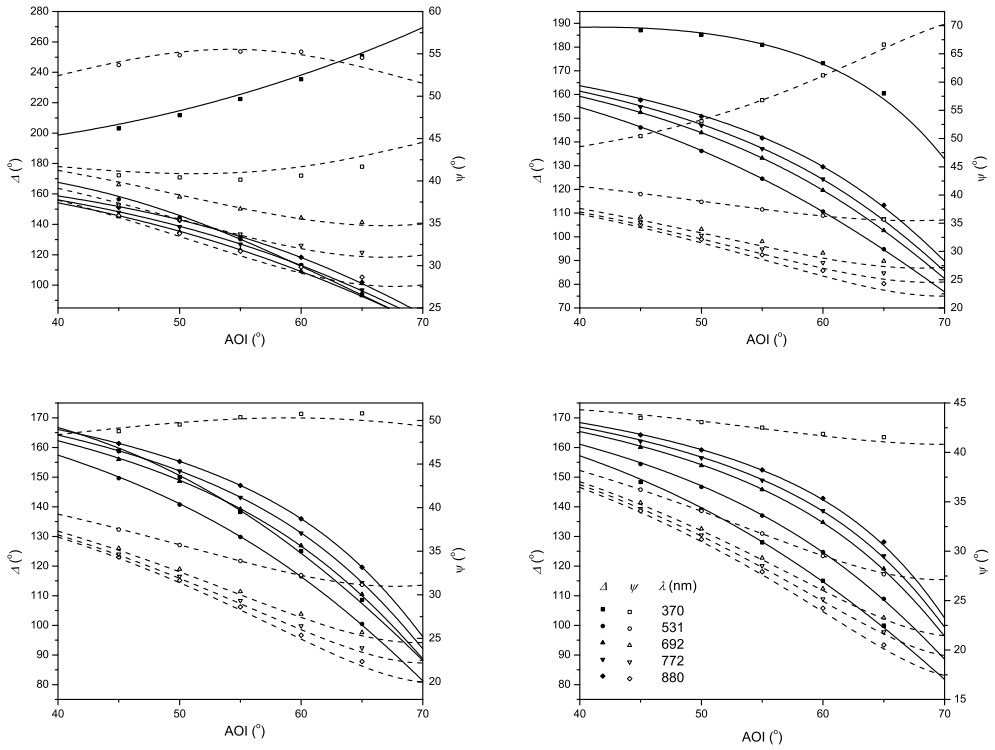


Fig. 6. VASE spectra for five selected wavelengths (see legend), measured at $T = 297.15$K for the PEDOT:PSS films prepared by spin-coating at different spin rates: a) 5000 rpm; b) 4000 rpm; c) 3000 rpm; d) 2000 rpm. The lines present the functions fitted by applying the WbW method (see text).

spun at 5000, 4000, 3000 rpm, the Drude term is stronger for the film spun at 2000 rpm. The values of the Lorentz resonance energy also range from 2.00 eV to 5.37 eV, while the broadening Γ_{L} is a bit more coherent around 1 eV. By keeping in mind that the fitted dielectric function parameters should exclusively represent properties of the film material and not its thickness, such incoherences should in principle not occur. VASE measurements were repeated for samples spun at 2000 and 5000 rpm, and LD2 fitting of those ellipsometric spectra resulted with paratmeters that almost perfectly coincide with those presented in Table II. This points to the fact that the observed incoherence in the LD2 parameters is not caused by an unreliable fit, but might represent either a property of the studied films, or a systematic trend inherent to the fitting method or model.

TABLE II
THE PARAMETERS OF THE LORENTZ-DRUDE DIELECTRIC FUNCTION, OBTAINED FROM LD1 AND LD2 FITS OF THE ELLIPSOMETRIC SPECTRA PRESENTED IN FIG. 3. FOR ALL ENTRIES, $\Gamma_{\mathrm{D}} = 0.00$.

Fit type	Spin rate (rpm)	A_{L} (eV^2)	E_0 (eV)	Γ_{L} (eV)	ϵ_∞	A_{D} (eV^2)	RMSE (°)
LD1	-	0.02	2.04	0.27	2.54	0.49	1.45
LD2	5000	1.46	3.68	1.04	2.11	0.15	0.43
	4000	0.77	3.29	1.24	2.31	0.32	0.66
	3000	5.61	5.37	0.57	2.19	0.01	0.98
	2000	0.16	2.00	0.71	2.53	0.81	1.74

The graph presented in Fig. 5 demonstrates a linear correlation between $\ln(A_{\mathrm{L}}/A_{\mathrm{D}})$ and E_0, observed by applying the LD2 fitting method. Thus the observed cross-correlation of the parameters might lead to a conclusion that for a given film thickness, A_{L}, A_{D} and E_0 may mutually adjust such that either Lorentz or Drude oscillator gains more strength, without much affecting the RMSE. Although in the present case, A_{L} and A_{D} may not be well determined from VASE data fitting, their ratio is determined by E_0, which is purely a property of the Lorentz term in (6). This provides the LD model parameters mathematical determinability, meaning that by including more data sets measured under different conditions (for example thickness, as is the case in LD1), LD dielectric function becomes progressively better determined in the VASE data. Parameter cross-correlations may not be completely excluded even in the case of LD1 (the correlation matrix for LD1 parameters was inspected and shows cross-correlations up to even 90 % as compared to 99 % in the case of LD2). Thus, the RMSE obtained from fitting should not be considered as a main criterion in model optimization and the dielectric function will be significantly better determined when there is a larger number of VASE data points measured for varied film thickness.

The so far presented results clearly demonstrate an ambiguity in simultaneous fitting the LD dielectric function parameters and film thickness, whereupon the LD1 approach apparently leads to a more reliable fit. This is also evidenced

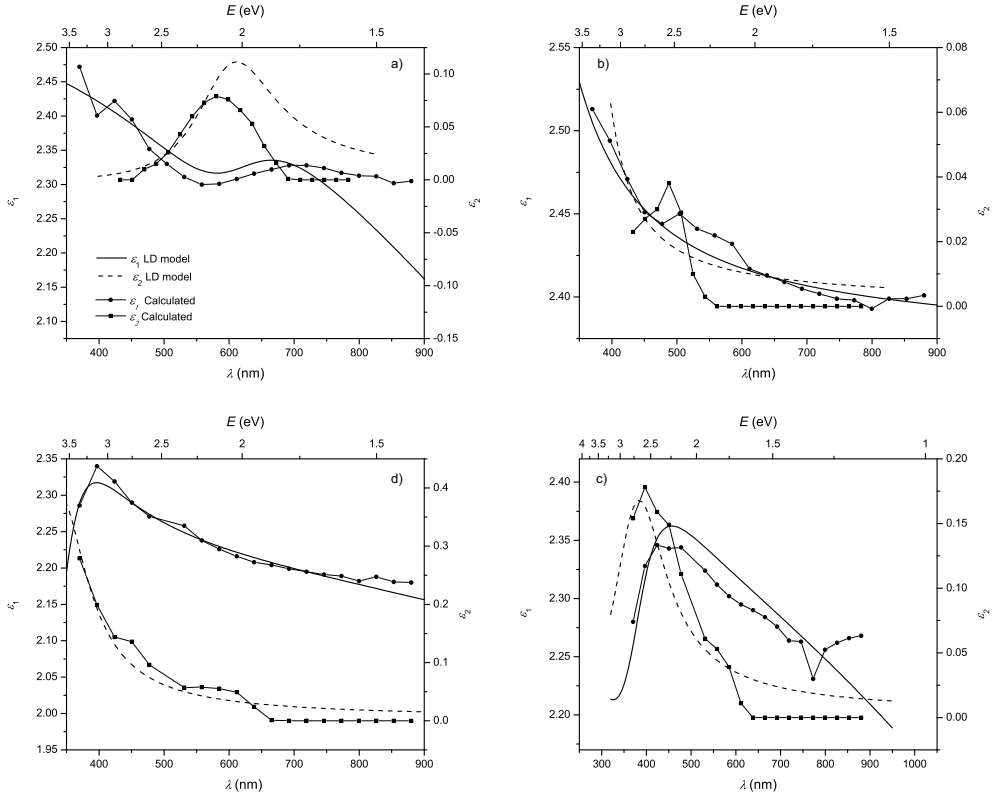


Fig. 7. Dielectric constant fitted independently for constant wavelengths in the VASE spectra (WbW method, symbols connected with lines), compared with the Lorentz-Drude dielectric functions fitted by the LD2 method (curves).

when the fitted E_0 parameter is compared with the expectation value based on theoretical electronic structure of PEDOT [22]. It was found by computational modelling that the main optical transition in oligo- and polythiophenes depends on the chain length (polymerization index) and the oxidation state, and may range from ~ 3.0 eV in the case of a trimer in neutral (unoxidized) state to as low as ~ 0.5 eV for the case of singly oxidized nonamer [14], [27]. When comparing the parameters presented in Table II with these theoretical values, it has to be noted that in the studied samples, the chain length is not known and probably not strictly defined. Moreover the oxidation state of PEDOT was not controlled and the influence of PSS on the electronic density of states of PEDOT should also not be neglected. Nevertheless, we may detect that some of the E_0 values obtained in LD2 fitting are by far too large to be physically meaningful. On the other hand, the value of 2.04 eV obtained from the LD1 fit seems reasonable for the $\pi \rightarrow \pi*$ transition in the neutral state. It is interesting to note that a similar value was obtained by the LD2 fitting method only for the thinnest film spun at 2000 rpm. In this case, the ratio of electron to hole density, which may be calculated from the $A_{\mathrm{L}}/A_{\mathrm{D}}$ ratio, amounts to 0.4 from the LD1 fit, and 2.1 from LD2. This large difference demonstrates the sensitivity of the LD parameters to the density of electronic states in the studied film material.

The VASE ellipsometric angles as function of AOI at a constant wavelength are presented in Fig. 6. For sake of clarity, these graphs include data for only five out of twenty studied wavelengths. A nearly perfect fit of the curves obtained by implementing the WbW fitting method may be detected. The set of dielectric constants obtained in this manner, presented by connected points in Fig. 7, may be considered as the quasi-experimental VASE dielectric function, since ϵ_1 and ϵ_2 were fitted as parameters of the SLM, independently for each wavelength. The thickness of the film was kept fixed at the values obtained from the AFM measurements, while ϵ_2 was constrained to positive values. Under these conditions, some of the ϵ_2 points coincide with zero, which is a consequence of a low absorbed intensity and points to underdetermination of ϵ_2 with respect to ϵ_1. Surprisingly, the LD2 dielectric functions presented in Fig. 7 show good consistency with the WbW quasi-experimental dielectric functions. This points to a possibility that the LD2 functions do not result from overfitting but are correctly determined as a property of the studied films. In that case a large variation in these functions for the studied samples of varied thicknesses (and possibly oxidation state), deserves further investigation.

CONCLUSION

The presented results demonstrate successful introduction of a method for simultaneous determination of PEDOT:PSS

film thickness and the dielectric function, by using variable angle spectroscopic ellipsometry (VASE) in the visible spectral range. The method requires postulation of an optical model for the surface (starified layers model), in combination with an appropriate model for the dielectric function of the studied film. In this respect, an isotropic Lorentz-Drude model function proved to be applicable. The simultaneous determination of the dielectric function parameters and film thickness requires a rigorous and methodical data fitting approach. In order to avoid cross-correlations of the fitted parameters (overfitting), it is recommended to simultaneously fit VASE data measured for samples of varied film thickness, whereupon the RMSE obtained from fitting should not be taken as the sole criterion for selecting the optimal model.

ACKNOWLEDGEMENTS

This study was accomplished with support of the European Fund for Regional Development and the Ministry of Science, Education and Sports of the Republic of Croatia, project Research Infrastructure for Campus-based Laboratories at the University of Rijeka (grant number RC.2.2.06-0001.

GP would like to thank the EU Marie Curie fellowship program (project FP7-PEOPLE-2013-ITN OrgBio)

REFERENCES

[1] O. Bubnova, Z. U. Khan, H. Wang, S. Braun, D. R. Evans, M. Fabretto, P. Hojati-Talemi, D. Dagnelund, J.-B. Arlin, and Y. H. Geerts, "Semi-metallic polymers," *Nature materials*, vol. 13, no. 2, pp. 190–194, 2014.

[2] G. Tarabella, F. M. Mohammadi, N. Coppedè, F. Barbero, S. Iannotta, C. Santato, and F. Cicoira, "New opportunities for organic electronics and bioelectronics: ions in action," *Chemical Science*, vol. 4, no. 4, pp. 1395–1409, 2013.

[3] T. A. Skotheim and J. Reynolds, *Conjugated polymers: processing and applications*. CRC press, 2006.

[4] ——, *Conjugated polymers: theory, synthesis, properties, and characterization*. CRC press, 2006.

[5] D. Khodagholy, J. Rivnay, M. Sessolo, M. Gurfinkel, P. Leleux, L. H. Jimison, E. Stavrinidou, T. Herve, S. Sanaur, and R. M. Owens, "High transconductance organic electrochemical transistors," *Nature communications*, vol. 4, 2013.

[6] C. Gravalidis, A. Laskarakis, and S. Logothetidis, "Fine tuning of pedot electronic properties using solvents," *The European Physical Journal Applied Physics*, vol. 46, no. 01, p. 12505, 2009.

[7] A. Laskarakis, P. G. Karagiannidis, D. Georgiou, D. M. Nikolaidou, and S. Logothetidis, "Optical investigations of the effect of solvent and thermal annealing on the optoelectronic properties of poly (3, 4-ethylenedioxythiophene): poly (styrenesulfonate) films," *Thin Solid Films*, vol. 541, pp. 102–106, 2013.

[8] J. Gustafsson, B. Liedberg, and O. Inganäs, "In situ spectroscopic investigations of electrochromism and ion transport in a poly (3, 4-ethylenedioxythiophene) electrode in a solid state electrochemical cell," *Solid state ionics*, vol. 69, no. 2, pp. 145–152, 1994.

[9] X. Chen and O. Inganäs, "Three-step redox in polythiophenes: Evidence from electrochemistry at an ultramicroelectrode," *The Journal of Physical Chemistry*, vol. 100, no. 37, pp. 15 202–15 206, 1996.

[10] L. Groenendaal, G. Zotti, P.-H. Aubert, S. M. Waybright, and J. R. Reynolds, "Electrochemistry of poly (3, 4-alkylenedioxythiophene) derivatives," *Advanced Materials*, vol. 15, no. 11, pp. 855–879, 2003.

[11] R. M. A. Azzam and N. M. Bashara, "Ellipsometry and polarized light, 1987," *Amsterdam, North-Holland.*, 1981.

[12] H. Fujiwara, *Spectroscopic ellipsometry: principles and applications*. John Wiley & Sons, 2007.

[13] J. Humlıček, A. Nebojsa, J. Hora, M. Stráskỳ, J. Spousta, and T. Šikola, "Ellipsometry and transport studies of thin-film metal nitrides," *Thin Solid Films*, vol. 332, no. 1-2, pp. 25–29, 1998.

[14] J. Cornil, D. Beljonne, and J. Brédas, "Nature of optical transitions in conjugated oligomers. ii. theoretical characterization of neutral and doped oligothiophenes," *The Journal of chemical physics*, vol. 103, no. 2, pp. 842–849, 1995.

[15] A. Patil, A. Heeger, and F. Wudl, "Optical properties of conducting polymers," *Chemical Reviews*, vol. 88, no. 1, pp. 183–200, 1988.

[16] Y. Peter and M. Cardona, *Fundamentals of semiconductors: physics and materials properties*. Springer Science & Business Media, 2010.

[17] G. Pathak, Z. Krasińska-Krawet, L. Szyk-Warszyńska, and D. Čakara, "Doping of poly (3, 4-ethylenedioxythiophene): poly (styrenesulfonate) films studied by means of electrochemical variable angle spectroscopic ellipsometry," *Thin Solid Films*, vol. 651, pp. 31–38, 2018.

[18] F. Cicoira and C. Santato, *Organic electronics: emerging concepts and technologies*. John Wiley Sons, 2013.

[19] J. Rivnay, R. M. Owens, and G. G. Malliaras, "The rise of organic bioelectronics," *Chemistry of Materials*, vol. 26, no. 1, pp. 679–685, 2013.

[20] J. Jung, J. Bork, T. Holmgaard, N. A. Kortbek, and K. Pedersen, "Ellipsometry," Aalborg University, Institute of Physics and Nanotechnology, 7th semester project on Applied Physics, 1. Sep - 21. Dec 2004.

[21] J. Kleimann, G. Lecoultre, G. Papastavrou, S. Jeanneret, P. Galletto, G. J. M. Koper, and M. Borkovec, "Deposition of nanosized latex particles onto silica and cellulose surfaces studied by optical reflectometry," *Journal of colloid and interface science*, vol. 303, no. 2, pp. 460–471, 2006.

[22] E.-G. Kim and J.-L. Brédas, "Electronic evolution of poly (3, 4-ethylenedioxythiophene)(pedot): From the isolated chain to the pristine and heavily doped crystals," *Journal of the American Chemical Society*, vol. 130, no. 50, pp. 16 880–16 889, 2008.

[23] K. Lee, R. Menon, C. Yoon, and A. Heeger, "Reflectance of conducting polypyrrole: Observation of the metal-insulator transition driven by disorder," *Physical Review B*, vol. 52, no. 7, p. 4779, 1995.

[24] N. Smith, "Classical generalization of the drude formula for the optical conductivity," *Physical Review B*, vol. 64, no. 15, p. 155106, 2001.

[25] T. L. Cocker, D. Baillie, M. Buruma, L. V. Titova, R. D. Sydora, F. Marsiglio, and F. A. Hegmann, "Microscopic origin of the drude-smith model," *Physical Review B*, vol. 96, no. 20, p. 205439, 2017.

[26] Sopra, "Public database, sopra s.a., url: http://www.sspectra.com/sopra.html."

[27] J. Cornil, D. Beljonne, and J. Brédas, "Nature of optical transitions in conjugated oligomers. i. theoretical characterization of neutral and doped oligo (phenylenevinylene) s," *The Journal of chemical physics*, vol. 103, no. 2, pp. 834–841, 1995.

Radiation Pressure Sensor

M. Karuza[*◦a], D. Božičević [*◦], G. Cantatore[**◦] and M.Vretenar[***]
[*] University of Rijeka, Department of Physics, Rijeka, Croatia
[**] University of Trieste, Department of Physics, Trieste, Italy
[***] University of Twente, Faculty of Science and Technology Complex Photonic Systems (COPS), MESA+ Institute for Nanotechnology, University of Twente, Enschede, Netherlands
[◦]INFN Sezione di Trieste, Trieste, Italy
[a] University of Rijeka, Centre of excellence for advanced sensors and materials and Centre for micro- and nanosciences and technologies, Rijeka, Croatia
mkaruza@uniri.hr

***Abstract* - Mechanical elements with dimensions in the nanometer range, at least in one direction, have been successfully employed as sensors in various devices. Their mechanical properties must be known with maximum precision in order to quantify the sensor response to external excitation. This often poses a significant challenge due to the mechanical fragility of the sensor elements. Here we present a measurement of the mechanical response of a 100 nm thick silicon nitride membrane. The external excitation force is provided by a laser beam modulated in amplitude, while the displacement of the membrane is measured by a Michelson interferometer with a homodyne readout.**

Keywords – radiation pressure, sensor, mechanical resonator, mechanical impedance

I. INTRODUCTION

The development of production methods and technologies has led to a realization of smaller and ever more sensitive sensors. One example are the thin high-stress membranes made of silicon nitride produced by the Canadian company Norcada. The membranes are basically a thin (~ 50 nm) silicon nitride window stretched over a 200 µm thick silicon frame. Their high mechanical quality factor ($Q \sim 10^5$) and high resonance frequency (~ 10^5 Hz) has made them a valuable tool in various types of experiments. Their use ranges from quantum optics experiments, where light is controlled by light [1] or, more recently, as a force sensor in an experiment searching for the dark energy [2].

Dark energy, which makes up to 70% of the Universe, was introduced in order to provide an explanation for the accelerated expansion of the Universe. One of the most popular models includes a scalar field with a screening "chameleon" mechanism, where the mass of the field depends on the local matter density [3]. This allows for the accelerated expansion, while evading conflict with laboratory experimental results [4]. Since the mass of the field depends on the local matter density, the associated particle must reflect off a density boundary when its energy before the boundary is lower than after, thereby transferring momentum in this process. The transferred momentum can be detected as a displacement of, or force applied to, the boundary. This makes the detection of the chameleon field possible with a suitable sensor which must be calibrated. A calibration scheme where a known force is applied on the membrane has been proposed [5]. This scheme, where the force is provided by the radiation pressure of a laser light beam is used also in gravitational wave detectors where it simulates the expected signal [6].

In this paper we present a study of the mechanical response of a membrane to a sinusoidal varying force. The frequency of modulation was varied in a range from 10 kHz to 100 kHz and the response of the sensor was studied.

II. THEORETICAL BACKGROUND

The membranes are two-dimensional mechanical resonators whose motion is governed by the partial differential equation [7]

$$\frac{\partial^2\psi}{\partial t^2} = v^2\left(\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2}\right), \tag{1}$$

where v is the speed of wave propagation. This equation has a solution depending on the indices *m*, *n* and is given by

$$\psi_{mn} = [A_{mn}\cos(\omega_{mn}t) + B_{mn}\sin(\omega_{mn}t)]\sin\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right). \tag{2}$$

The indices *m* and *n* are integer numbers, and a and b represent membrane dimensions in the x and y direction, respectively. The coefficients A_{mn} and B_{mn} depend on the initial conditions and can be calculated, but they are not interesting for this work. The frequencies of the membrane oscillation modes are given by

$$\omega_{mn} = \pi v\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2}. \tag{3}$$

Eq. (3) is usually used to identify the oscillation modes in the displacement spectrum of the membrane. An example of such spectrum is shown in Figure 1.

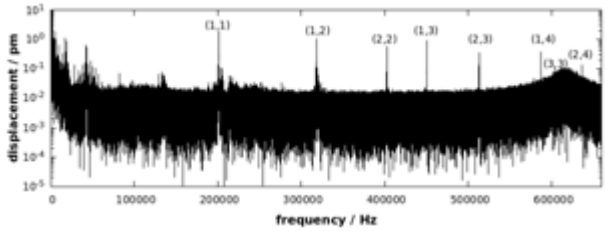


Figure 1. Calibrated frequency spectrum of the oscillations of 2x2 mm^2, 50 nm thick, membrane. The mode indices *m,n* are shown in parenthesis. The frequencies of higher order modes are related to the frequency of the fundamental mode by Eq. (3).

The oscillations are thermally excited, i.e. by Brownian stochastic force with zero mean value, and they are damped by a viscous force with damping rate D [8]. The equation of motion of a viscously damped mechanical oscillator with a sinusoidally driving force is

$$m\ddot{x} + D\dot{x} + kx = F\ \sin(\omega t) \tag{4}$$

Where m is the mass, k is the spring constant of the undriven harmonic oscillator without damping and D is the damping coefficient. The constants k and m are related to the natural frequency of the harmonic oscillator by the usual relation $\omega_0 = \sqrt{\frac{k}{m}}$.

The equation of motion has a general solution which includes the transient solution and steady state response. We are interested only in the latter since we assume that the system has reached equilibrium. The steady state solution becomes

$$x(t) = \frac{F(\sin \omega t - \varphi)}{\sqrt{m^2(\omega^2-\omega_0^2)^2+(D\omega)^2}} \tag{5}$$

where φ is the phase between the driving force and displacement.

The amplitude of the steady state is the value of interest since it is directly measured and is given by

$$A = \frac{F}{\sqrt{m^2(\omega^2-\omega_0^2)^2+(D\omega)^2}} \tag{6}$$

The spectrum from Fig. (1) could be used to obtain the mechanical impedance, however due to the feeble signal the procedure is not straightforward. To verify the response of the sensor to a driving force F in (4), a calibration was performed using the radiation pressure of a laser beam with sinusoidally modulated intensity.

III. Experimental Setup

The measurements were carried out with an experimental setup designed and constructed in the Laboratory for Quantum and Nonlinear Optics at Department of Physics and Centre for micro- and nanosciences and technologies (NANORI) at University of Rijeka. The heart of the setup is a Michelson interferometer where the membrane is placed at the end of the so-called signal (SIG) arm, while in the other arm of the interferometer, called local oscillator (LO), a piezo-actuated mirror is used to maintain a constant phase relation between the two arms. The beams propagating in the two arms are also labeled SIG and LO. The experimental scheme can be seen in Fig 2.

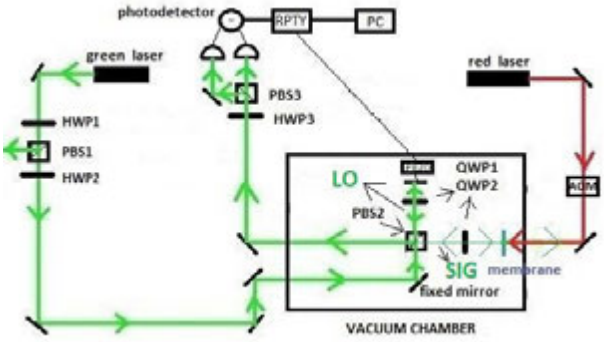


Figure 2. Experimental setup for mechanical impedance measurements. The Michelson interferometer is placed in a vacuum chamber. The green laser is used for the measurement while the red laser is used for excitation. More details are given in the text.

The information on the phase between the SIG and LO beams is read out by a balanced homodyne detector [9] and passed to a single board computer, called Red Pitaya (Rptya) [10] used for control and data acquisition. A custom software has been developed for this purpose.

The measurement beam (green) is produced by a laser where an infrared beam is duplicated in frequency. The green light during the propagation encounters a half-wave plate (HWP1) used to determine the intensity of the light used for measurement. This is done together with a polarizing beam splitter (PBS1) where a fraction of light is passing unperturbed while the rest is exiting on the side and does not contribute to the measurement. After the first optical elements the light encounters the HWP2 which together with PBS2 determines the intensity distribution between LO and SIG beams. Most of the intensity circulates in the LO arm where a quarter wave plate (QWP1) is placed. It transforms linearly polarized light to elliptical polarization, and when the light beam is reflected by a mirror going through QWP1 again, its polarization is changed from elliptical to linear but at right angle with respect to the initial one. This results in light coming out of the other port of PBS2 and in separating the ingoing from outgoing light. The same process takes place in the SIG arm, where QWP2 is placed and the light again comes out at the exit port of the PBS2. The two beams with orthogonal linear polarizations propagate to HWP3, which then rotates the polarization of the incoming light by 45 degrees resulting in two effects. In the PBS3 the SIG and LO beams are mixed and can interfere, and the outputs of the PBS3become of equal intensity. The intensities of two beams are read by a balanced photodiode detector, which subtracts the resulting photocurrents, then converts the output to a voltage and amplifies it. This signal is digitized by Rptya and used to maintain the phase between the LO and SIG beam. This is done by applying the voltage provided by a proportional-integral-derivative (PID) loop to a piezo actuator moving the mirror in the LO arm of the interferometer. The measurements were done by locking the phase to the gray fringe of the interferometer. The data digitized by Rptya is also transferred to a personal computer (PC) where it is stored for off-line analysis. The digitized waveforms are analysed in frequency by a Fourier transform, which gives displacement amplitudes as functions of frequency.

The excitation (red) beam is produced by a HeNe laser. It is modulated in intensity and frequency by an acousto-

optical modulator (AOM). Here only the modulation in intensity is relevant. A sinusoidal signal at an arbitrary frequency is provided to the intensity modulation input of an AOM driving circuit. The result is a sinusoidally varying intensity at a given frequency. The corresponding radiation pressure that is changing in time provides the driving force.

IV. RESULTS

The measurements were done with a *5x5 mm*^{2}, *100 nm* thick membrane that was also used as sensor at CERN during data taking at the CAST experiment. The request for the sensor calibration comes from the requirement to give a limit to the chameleon radiation pressure force obtained during the measurement campaign reported in [5]. While the sensor is self-calibrating in terms of displacement, as it is sufficient to measure the distance between the fringes during mirror movement in the LO interferometer arm, the force calibration is more involved. To this end, a known external driving force must be provided. Here it was accomplished with the photon radiation pressure force as previously described. The radiation pressure in case of electromagnetic waves is related to the beam intensity I by $P = \frac{2I}{c}$. To obtain the force, the pressure P must be multiplied by an effective area S. Since the intensity I is defined as the ratio of incident power p and the area S, $I = \frac{p}{S}$, the radiation pressure force becomes $F = P \cdot S = 2\frac{p}{c}$.

The force then depends only on the light power p, which can be easily measured by a calibrated optical power meter. The power of the excitation beam used was $p = 1.5$ mW while only a fraction, $p_{ref} = 300$ μW, was reflected from the membrane. The displacement of the membrane at the modulation frequency was measured and the result can be seen in the Fig. 3.

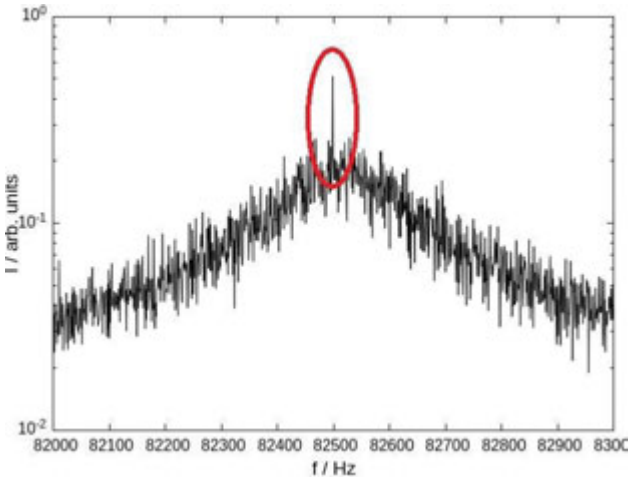


Figure 3. Frequency spectrum of the fundamental mode oscillations of a 5x5 mm^2, 100 nm thick, membrane. The wide peak is thermally excited, while the narrow peak in the red ellipse is excited by external driving force. The displacement is given in arbitrary units. The calibration constant was separately determined and it is $k = 0.7$ pm/a.u.

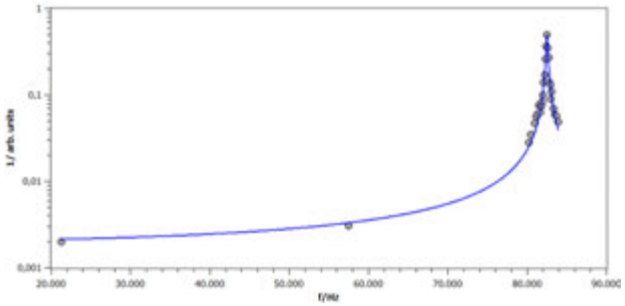


Figure 4. Displacement amplitudes of a membrane for different driving force frequencies. Each gray dot represents one mesurement. It can be seen that the amplitude depends on the driving force frequency. The blue line is result of a least squares fit.

The peak due to the driving force at f = 82500 Hz was observed and the amplitude of the displacement was written down. The procedure was repeated for different values of frequency f of the driving force and the displacement values in arbitrary units were recorded for every case.

The parameters obtained by fitting the Eq. (6) are $\frac{F}{m} = (9.2 \pm 0.3) \cdot 10^6$ in arbitrary units, $\omega_0 = 82495 \pm 1$ Hz and $D = (7 \pm 1) \cdot 10^{-5}$ a.u.

Since the driving force amplitude $F = 1$ pN is known, the mass m can be extracted from the fit parameter A. With all the parameters known, it is possible to reverse the logic and to obtain a force from measured displacement amplitude. For example, for measured displacement $d = 40 \pm 5$ fm, the force on the membrane is

$$F_m = d\sqrt{m^2(\omega^2 - \omega_0^2)^2 + (D\omega)^2} \qquad (7)$$

Which gives a force amplitude of $F = 80 \pm 15$ pN acting on the membrane at 1 kHz frequency.

V. CONCLUSION

We have developed a method for measuring the properties of a membrane mechanical oscillator where the excitation was provided by the electromagnetic radiation force. Displacement peaks due to the external excitation were observed in a range of frequencies around the fundamental mode of the membrane. This method provides a novel way in the determination of properties of mechanical oscillators.

ACKNOWLEDGMENT

We acknowledge support from University of Rijeka with grant no 13.12.2.2.09, 18-126, EU fund projects KK.01.1.1.01.0001 and RC.2.2.06.-0001. The authors express their gratitude to the colleagues of the INFN Sezione di Trieste mechanical workshop where parts of the detector were built.

[1] M. Karuza et al., "Optomechanically induced transparency in membrane-in-the-middle setup at room temperature," Phys. Rev. A 88, 013804 (2013)

[2] S. Arguedas Cuendis et al., " First results on the search for chameleons with the KWISP detector at CAST," Physics of the Dark Universe 26 (2019) 100367

[3] J. Khoury and A. Weltman, “Chameleon cosmology, ” Phys. Rev. D 69, 044026

[4] A. Upadhye, “Dark energy fifth forces in torsion pendulum experiments,” Phys. Rev. D 86, 102003 – Published 7 November 2012

[5] M. Karuza. et al., “KWISP: an ultra-sensitive force sensor for the Dark Energy sector,” Phys. Dark Univ. 12 (2016) 100-104

[6] S. Karki et al., “The Advanced LIGO photon calibrators,” Review of Scientific Instruments 87, 114503 (2016)

[7] E. Kreyszig, Advanced Engineering Mathematics, 10th ed, John Wiley and Sons Inc., pp.575-581

[8] M. Karuza et al., “Optomechanical sideband cooling of a thin membrane within a cavity,” New Journal of Physics 14 (2012) 095015 - Published 17 September 2012

[9] H.A. Bachor and T.C. Ralph, A Guide to Experiments in Quantum Optics, Wiley-VCH, 2004

[10] M. Vretenar, N. Erceg, and M. Karuza, “Energy-resolved coincidence counting using an FPGA for nuclear lifetime experiments,” American Journal of Physics 87 (12):997-1003

Measurement of the human cadaver ossicle vibration amplitude by fiber-optic interferometry

Z. Djinovic*, M.Tomic **, R. Pavelka***, G. Sprinzl****, H. Traxler*****
* ACMIT Gmbh, Wiener Neustadt, Austria
** Institute of Technical Sciences of SASA, Belgrade, Serbia
***ENT specialist, Wiener Neustadt, Austria
****University Hospital St. Pölten, Austria
*****Medical University of Vienna, Centre for Anatomy and Cell Biology, Vienna, Austria
zoran.djinovic@acmit.at

Abstract – **In this paper we present the results obtained by contactless measurement of vibration of the incus inside the middle ear of a human cadaver, exposed to an external calibrated sound source. The measurement configuration was a Michelson fiber-optic interferometer based on a 3x3 fiber optic coupler, with a VCSEL source. The interferometric signal fading was solved using two quasi-quadrature signals and specialized signal processing. The ossicle reflection was augmented by a very light piece of retroreflective material attached to the bone. We measured the amplitude of vibration at different sound pressure levels, independently monitored by a decibel meter placed in the proximity of the cadaver ear. The accuracy of the measurement technique was proved by using a calibrated piezo transducer exposed to vibrate with amplitude of 7nm at 1kHz. We performed the experiments on 8 cadavers in total. The measured amplitudes were found to be in the range between 1 pm and 1 nm, under the sound excitation of 40-90 dB SPL @ 1kHz, respectively.**

Keywords - fiber-optic sensors; vibration measurement; implantable microphone; hearing aid;

I. INTRODUCTION

Two main hearing insufficiencies, middle ear otosclerosis and sensorineural impairment, lead to partial or total hearing loss in millions of people all around the world [1-2]. Very often, they can be overcome by common external hearing aids. However, some patients, having sensorineural malfunction, need to have an implantable hearing device such as cochlear implant (CI) [3-4]. Most implantable cochlear devices today can operate thanks to external microphone that receive the ambiental sound pressure and transmit the processed audio signal through the skull bone to the CI. Basically, this help a lot to the patents, especially to children to develop the speech in the early age [5]. However, the external microphone is always associated with some drawbacks such as difficulties to perform common life activities including sleeping and swimming as well as social stigmatization.

This work was sponsored by the NÖ Forschung & Bildung (NFB), Austria and ACMIT Gmbh in the frame of the project: "Totally implantable fiber-optic sound sensing system for cochlear- and middle-ear hearing aids", Life Science grant: LS14-026.

Many groups in the art [6-10] traying to develop a totally implantable hearing aid having implantable microphone within the device. It is a challenge task, since several requirements have to be fulfilled including high sensitivity and robustness in terms of long-term application in biologically active environment [11]. In this scope it is very important to meet a few clinical requirements including biocompatibility, hermeticity, corrosion resistance against body fluids, overall size and energy consumption and safety application [12].

It appears that fiber-optic sensing technology can fulfill many of aforementioned demands. We already demonstrated capability of this technology [13-14] to be applied for the design and manufacturing of implantable microphone. However, we also met some issues, which require improvements in order to claim a list of technical demands of an implantable microphone. For this reason, we investigated sensitivity and dynamic range of a fiber-optic interferometric technique aimed to be used for manufacturing an implantable microphone as a part of a totally implantable hearing aid.

In this paper we present the results obtained by contactless measurement of vibration of the incus inside the middle ear of a human cadaver, exposed to an external calibrated sound source. We measured the amplitude of vibration at different sound pressure levels (SPL), independently monitored by a decibel meter placed in the proximity of the cadaver ear. The accuracy of the measurement technique was proved by using a calibrated piezo transducer (PZT). We performed the experiments on eight cadaver heads in total. The amplitudes of vibration at 1 kHz were found to be in the range between 1 pm and 1 nm, under the common speech range of sound pressure level.

II. PRINCIPLE OF OPERATION

The sensing configuration, based on modified Michelson interferometer using a single mode 3x3 fiber-optic coupler (FOC3x3), is presented in Fig.1. The middle input arm of the FOC3x3 is connected to a pigtailed laser diode (LAS).

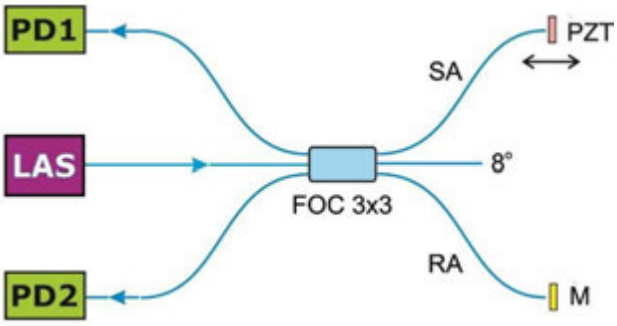


Figure 1. Principle of operation: the fiber-optic Michelson interferometer based on a 3x3 coupler. LAS – laser source, PD1, PD2 – photodiodes, FOC3x3 – single mode 3x3 fiber optic coupler. SA – sensing arm, RA – reference arm, M - mirror, PZT – vibrating target.

The third arm, which is not necessary, is obliquely cleaved, in order to suppress back reflections. The other two input arms, acting in this case as the outputs, are connected to photodiodes (PD1, PD2), providing two quasi-quadrature interferometric signals [15].

The probe beam in SA is directed to the target, a vibrating PTZ, which is independently calibrated. The RA, being approximately equal in the optical length to the SA, is directed to the fixed mirror (M). A vibration of the target changes the length of the optical path in the SA and, consequently, the optical phase difference (OPD) between the two coupler arms. This OPD's changes generate the raw interferometric signals in both photodetectors, which can be represented by:

$$V_{PD1}(t) = V_{DC1} + V_{A1}\cos\big(k \cdot A \sin\omega t + L(t)\big) \qquad (1)$$

$$V_{PD2}(t) = V_{DC2} + V_{A2}\cos(k \cdot A \sin\omega t + L(t) + \Psi)$$

where A and ω are the amplitude and the frequency of the target vibration; ψ is the phase shift (~120˚) introduced by the 3x3 coupler; V_{DC1} and V_{DC2} are DC signal levels at PD1 and PD2; V_{A1} and V_{A2} are interferometric amplitudes, which depend on the target reflection, the coherence degree between the beams and the alignment and state of polarization of the probing beam. L(t) is a random motion of the target, mainly caused by external vibrations. Since L(t) causes the change of the working point in the interferometric sinusoidal transfer function, neither photodetector signal alone can provide a stable amplitude detection. However, the constant phase shift ψ between coupler arms allows elimination the influence of L(t), by the suitable signal processing algorithm, based on the phase extraction from two photodiode signals [13]. The dependence on variable amplitudes V_{A1} and V_{A2} is also removed in this phase demodulation process. Two signals and its processing are absolutely necessary if the amplitude of measured vibration is smaller than half of the wavelength, which is certainly the case of human middle ear bone vibrations.

III. EXPERIMENT

We employed the sensing configuration presented in Fig. 1 in the measuring set-up (Fig. 2), used for determination of amplitude of vibration of middle ear ossicles, examined on human cadavers.

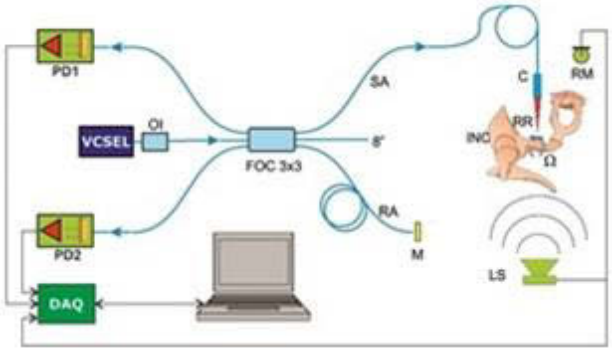


Figure 2. Setup for investigation of fiber-optic sensing configuration based on the miniaturized 3x3 coupler; VCSEL- Vertical Cavity Surface Emitting Laser, OI-optical isolator, SA, RA-sensing and reference arm, C- collimator, RR-retroreflector, M-mirror, PD-photodiode, DAQ-acquisition system, LS- loudspeaker, INC-incus, RM-reference microphone/decibelmeter

Experiments have been performed in a dissection room under the relatively quiet conditions at the Anatomy Division of the Medical University Vienna.

A pigtailed VCSEL diode, emitting at 1310nm, was used as a high-coherence light source. The laser pigtail was connected to the input arm of a single mode 3x3 fiber optic coupler via an optical isolator (OI), in order to suppress backreflections. The VCSEL was driven by low noise driver, with a current of 3mA, giving an optical power of about 240μW out of the pigtail. Two lateral arms were connected to two InGaAs photodiodes (PD1 and PD2), which were further connected to the accompanied low-noise transimpedance amplifiers. Sensing and reference arms were very short, just about 100 mm in length, that was an acceptable length for the implantation. On the other hand, such a short length significantly contributed to decreasing the polarization effects. The sensing fiber was terminated by a spherical collimator (C), with 450μm in diameter, aimed to the target. The reference arm was terminated by a 100% reflective golden surface layer (M).

The surgical procedure was performed in order to assure access to the incus of a human cadaver skull. It included a mastoidectomy and posterior epitympanotomy, with exposition of the short processus and the body of the incus, which was the target in these experiments. The incus was equipped with a small piece of the retro-reflecting tape (RR), attached to the bone by a specially designed clip (Ω). The fiber-optic probe composed of reference and sensing arms, packed in a medical steel guiding needle, was directed and adjusted versus target in order to obtain the maximal signals on the PD1 and PD2. The use of the RR has been necessary to provide sufficiently amount of the reflected radiation [16].

The acoustical measurements have been performed by Affinity audiometer of Interacoustics Co. The calibrated system for generation of acoustic signals use the loudspeaker from JBL Incorporated®, Northridge, CA, control® CM52, set to be of about 70-80cm far away from the skull, capable to produce a pure acoustical stimuli in the range from 40 dB to 90 dB SPL re 20 μPa. The reference microphone signal and two raw photodiode signals have typically been acquired with 32kS/Ch/s, in time duration of about 31 s, that makes 1 million of measuring points.

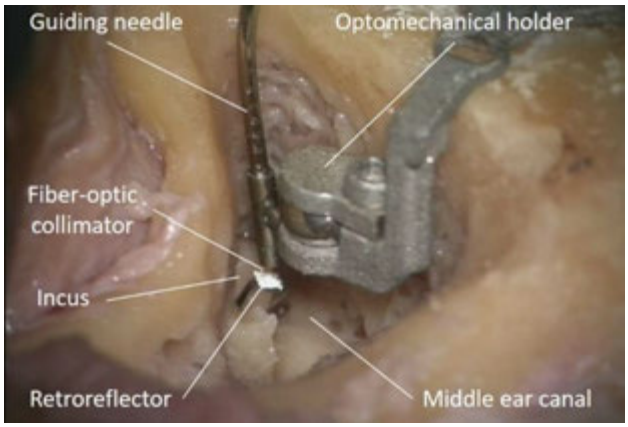


Figure 3. Close look at the middle ear canal and retroreflector firmly fixed over incus; fiber-optic probe set into the guiding needle

The decibel meter values have been recorded manually as an average data. The raw signals have been later offline processed by arctan algorithm, using the specially made software packet.

Before any measurement, the system was tested using a calibrated PZT actuator (Fig. 4), placed at the same distance from fiber-optic probe where the incus will be, about 2.5 mm from the fiber end. An identical piece of retroreflective material was adhesively bonded to the PZT actuator. The PZT was driven by a sinusoidal signal at 1kHz@1Vp-p, producing an amplitude of 7.4 nm.

IV. Results and Discussion

The acquired photodiode signals by DAQ system obtained in the prior checking experiments, with the calibrated PZT, can be seen in Fig. 5, graph down right, in the form of Lissajous figure, after the first stage of processing. The PZT was driven by sinusoidal voltage 1V @1kHz, in the time interval from 1.5 s to 5.5 s. The very beginning of the period of acquisition is used to calculate the noise. Top left graph in Fig. 5 shows the demodulated and stabilized interference signal, i.e. the interferometer phase, obtained by the offline signal processing while the pilot signal is on, using the arcus tangent algorithm.

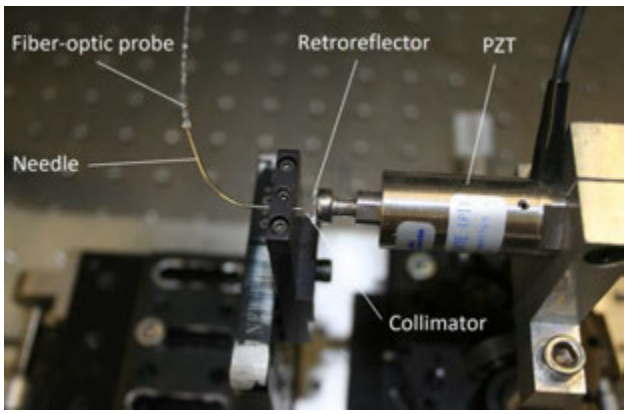


Figure 4. Close look at the fiber-optic sensing probe directed to the retroreflector bonded onto a calibrated PZT driven at 1 kHz, 7.4 nm amplitude

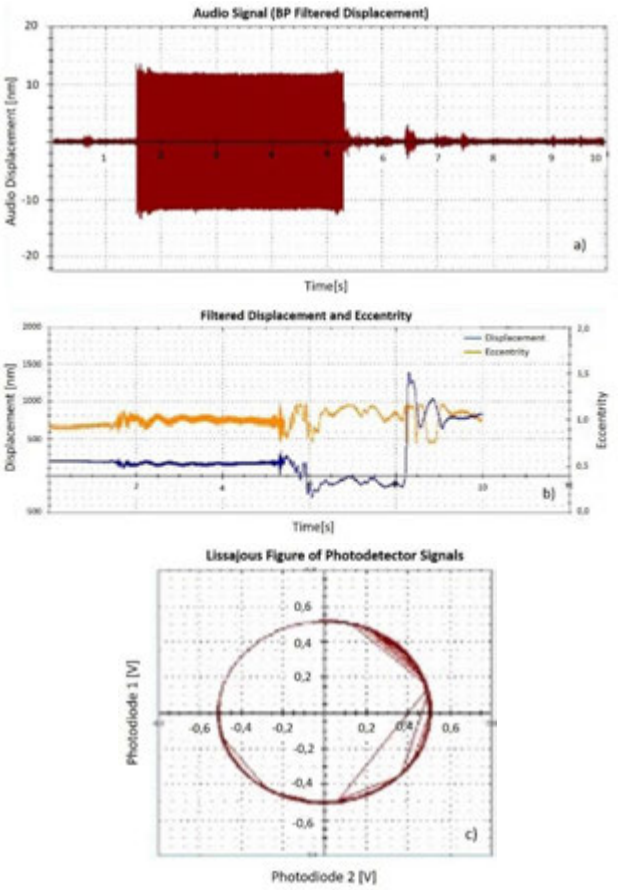


Figure 5. Processed data of PZT vibrations at 1Vp-p, 1kHz, representing: a) overall displacement, b) overall change of OPD of interference arms (blue line) and eccentricity of circle after processing (orange line) and c) Lissajous presentation of the processed signals

The processed signal reveals an RMS amplitude of PZT vibration of about 8 nm, that is in a good agreement with calibration data for 1kHz@1V of 7.4 nm. The accuracy of the measurement technique can be estimated to be better than 10%.

Figure 6 shows the signal and noise spectral density expressed in (nm/√Hz), obtained by FFT analysis of processed signal over the acoustic frequency range. A sharp peak at 1kHz relates to the PZT vibration at this frequency. Several curves (gray and purple) correspond to the noise and pilot signal (green and magenta) parts of acquired data, with and without A-weighting. Based on these results, we calculated the minimum detectable amplitude of vibration of about 1,2 pm/ √ Hz.

Fig. 7 presents the measured amplituds of incus vibration of eight human cadavers, measured at 1kHz for SPL over 40 to 90 SPL. We can see that amplitude of incus vibrations at 40 SPL span from about 1pm to about 10pm depending on the cadaver specimen. Similarly, for 90 SPL the amplitude of incus vibrations take values from about 150pm to about 3nm depending on the subjected cadaver specimen.

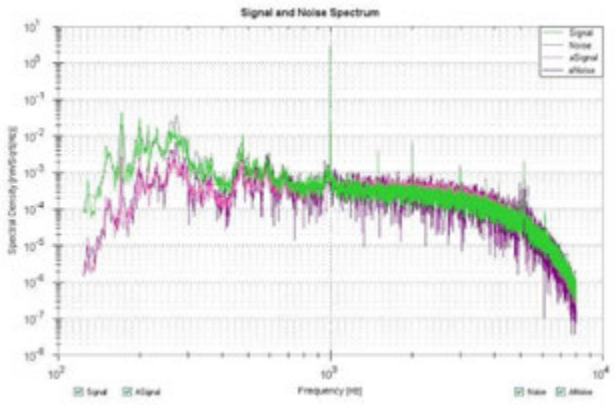


Figure 6. Signal and noise spectrum of PZT vibration at 1V@1kHz, with and without A-weighting

These results show a difference between the measured amplitude of vibrations that is in the range of one order of magnitude among the subjected cadaver specimens. It can be explained by various anatomic characteristic of the middle ear, age and previous health history of human cadavers as well as method of preparing of specimens (e.g. if they were deep forsen or not).

Generally, the obtained results are in relatively good agreement with literature data, which are obtained by amplitude measurement on temporal bones utilizing the Laser Doppler Vibrometry [17]. However, there is a discrepancy between the literature and our data concerning the amplitude measurement at 90 dB SPL and 1kHz. We obtained for about 10 to 20nm smaller amplitude of vibrations. Although, it is difficult to explain this difference without knowledge of the exact measurement and analysis techniques applied, we believe that the main reason is different hitting point of laser light. The literature data relate to the umbo instead incus displacement. The presented fiber-optic sensing technique demonstrated a high sensitivity. The minimum detectable amplitude of vibrations is about 1pm i.e. slightly below of 40 dB SPL.

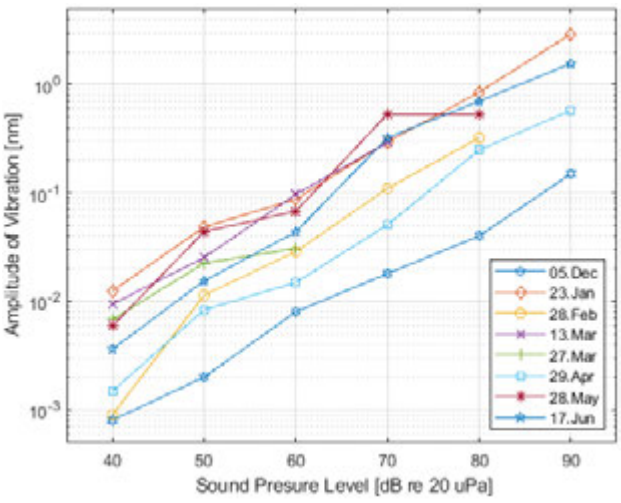


Figure 7. Amplitudes of eight human cadaver incus vibrations measured by fiber-optic interfereometry at 1kHz

V. CONLUSION

The presented fiber-optic sensing technique, investigated by the human cadavers, demonstrated some promising features necessary for a prospective implantable microphone in terms of small dimensions and high sensitivity. The minimum detectable amplitude of incus vibrations of about 1pm corresponds with required technical demand of a totally implantable hearing aid [12]. It was experimentally found that amplitude of incus vibrations at 40 SPL spans from about 1pm to about 10pm, depending on the cadaver specimen. Similarly, for 90 SPL the amplitude of incus vibrations takes values from about 150pm to about 3 nm. The incus response shows a high linearity to acoustical excitation, with a sensitivity of 1-5 pm/mPa @1kHz.

ACKNOWLEDGMENT

The authors thank to the ACMIT Gmbh and NÖ Forschung & Bildung (NFB), Austria, which sponsored this work in the frame of the project: "Totally implantable fiber-optic sound sensing system for cochlear- and middle-ear hearing aids", Life Science grant: LS14-026. Many thanks to Miss Nina Tomic for creating some of the figures.

REFERENCES

[1] D. Young, and P. Cong, "Wireless implantable sensors: from lab to technology breakthrough ambitions". Sensors and Actuators A, Vol. 294, pp. 81–90, August 2019.

[2] S.Park, X. Guan, Y. Kim, F. P. X. Creighton, E. Wei, I. J. Kymissis, H. H. Nakajima, and E. S. Olson, "PVDF-Based Piezoelectric Microphone for Sound Detection Inside the Cochlea: Toward Totally Implantable Cochlear Implants", Trends in Hearing, Vol. 22, pp. 1–11, April 2018.

[3] A. M. Innes, R. P. Morse, Irving and P. Begg, "Implantable microphones as an alternative to external microphones for cochlear implants", Cochlear Implants International, Vol. 18, pp. 304-313, September 2017.

[4] J. M. Ge´rard, L. Demanez, C. Salmon, F. Vanpoucke, J. Walraevens, A. Plasmans, D. De Siati, and P. Lefe`bvre, „Feasibility of an implanted microphone for cochlear implant Listening", Eur Arch Otorhinolaryngol, Vol. 274, pp. 1383–1390, November 2017.

[5] H. A. Jenkins and K. Uhler, "Speech Perception Comparisons Using an Implanted and an External Microphone in Existing Cochlear Implant Users", Otology & Neurotology, Vol. 33, pp. 13-19, January 2012.

[6] S. H. A. Woo, S. T. Woo, B. S. Song, and J. H. Cho, "Feedback characteristics between implantable microphone and transducer in middle ear cavity", Biomed Microdevices, Vol. 15, pp. 867–877, October 2013.

[7] A. G. Bittencourt, P. R. Burke, I. S. Jardim, R. Brito, R. K. Tsuji, A. C. O. Fonseca, and R. F. Bento, "Implantable and Semi-Implantable Hearing Aids: A Review of History, Indications, and Surgery", Int Arch Otorhinolaryngol, Vol.18, pp. 303–310, April 2014.

[8] F. Pfiffner, L. Prochazka, D. P´eus, I. Dobrev, A. Dalbert, J. H. Sim, R. Kesterke, J. Walraevens, F. Harris, C. Röösli, D. Obrist, and A. Huber, "MEMS Condenser Microphone-Based Intracochlear Acoustic Receiver", IEEE Transactions on Biomedical Engineering, Vol. 64, pp. 2431 – 2438, December 2017.

[9] S. T. Woo, D. H. Shin, H. G. Lim, K. W. Seong, P. Gottlieb, S. Puria, K. Y. Lee and J. H. Cho, "A New Trans-Tympanic Microphone Approach for Fully Implantable Hearing Devices", Sensors, Vol. 15, pp. 22798-22810, September 2015.

[10] D. Calero, S. Paul, A. Gesing, F. Alves and J. A. Cordioli, "A technical review and evaluation of implantable sensors for hearing

devices", BioMed Eng OnLine, Vol. 17, pp. 1-26, February 2018.

[11] J. W. Zwartenkot, "Auditory Implants in Otology-Active Middle Ear Implants and Direct Acoustic Cochlear Stimulation: Indications and Outcome", PhD Thesis, Radboud University, 2017.

[12] D. Calero, S. Paul, and J. A. Cordioli, "On implantable sensors for totally implantable hearing aids, 5th Joint Meeting of the ASA and ASJ, Honolulu - November 29, 2016.

[13] M. C. Tomic, Z. V. Djinovic, and S. J. Petricevic, "Demodulation of quasi-quadrature interferometric signals for use in the totally implantable hearing aids", Biomed. Opt. Express, Vol. 8, pp. 3404–3409, July 2017.

[14] Z. Djinović, R. Pavelka, M. Tomić, G. Sprinzl, H. Plenke, U. Losert, H. Bergmeister, R. Plasenzotti, "In-vitro and in-vivo measurement of the animal's middle ear acoustical response by partially implantable fiber-optic sensing system", Biosensors and Bioelectronics, Vol. 103, pp. 176–181, October 2018.

[15] K. P. Koo, A. B. Tveten, and A. Dandridge, "Passive stabilization scheme for fiber interferometers using (3x3) fiber directional couplers," Appl. Phys. Lett., Vol. 41, pp. 616–618, July 1982

[16] M. Koch, T. M. Essinger, T. Stoppe, N. Lasurashvili, M. Bornitz, and T. Zahnert, "Fully implantable hearing aid in the incudostapedial joint gap", Hearing Research vol. 340, pp. 169-178, March 2016.

[17] U. B. Willi, "Middle-ear Mechanics: The Dynamic Behavior of the Incudo-Malleolar Joint and its Role During the Transmission of Sound," PhD Thesis, Uni. Zürich, 2003.

Marine Fiber Optic and Spinning Mass Gyrocompasses

A. Škrobonja, I. Jurdana , I. Panić, N. Wakabayashi*
University of Rijeka, Faculty of Maritime Studies, Rijeka, Croatia
*Graduate School of Maritime Sciences, Kobe University, Kobe, Japan
skrobo@pfri.hr, jurdana@pfri.hr, panic@pfri.hr, waka@kobe-u.ac.jp

***Abstract* - One of the key aspects of modern navigation fundamentally depends on precise and current information that describes ships position and heading. In order to provide this information, different technologies have been implemented and used onboard ships throughout the maritime history such as celestial and terrestrial navigation or magnetic compass systems. First recent major advance has been achieved with implementation of first sufficiently sophisticated mechanical gyrocompass which was on the trail of today's modern mechanical gyrocompasses by Hermann Anschutz-Kaempfe in 1906. This had further influenced greater development and research with respect to gyrocompass manufacturing technology and measurement techniques. This knowledge led to the development and wide scale implementation of stable and precise modern gyrocompasses as they are today, however it has been realized that fundamental errors related to gyrocompass technology can never be fully mitigated. Fiber optic compass technology can eliminate or mitigate errors related to gyroscope-based technology and has been frequently researched and developed as a comparable solution to the commonly used gyrocompass. This paper presents historical overview of gyrocompass research and development, principles of operation, associated errors in measurement and comparative analysis both mechanical and fiber optic gyrocompass.**

Keywords – maritime; gyrocompass technology; fiber optic; navigation sensors

I. INTRODUCTION

First generation of gyrocompasses was based on fast spinning discs with complex mechanism and follow-up system. Due to aforementioned complexity, accurate and stable operation of gyrocompass could only be accomplished with rigorous and continuous maintenance of the gyroscopic system. In that time mechanical gyrocompass technology was the only technology that could be used to manufacture and implement commercially viable product of this type available for ships on the market. In the following years focus was placed on progressive development of gyrocompasses and peripheral navigation equipment and mitigation of known system issues. Newer gyrocompass generation simply offered less complex follow-up mechanisms that were easier to operate and regularly maintain. Development of fiber optic gyrocompass systems aimed to supplement and replace conventional gyroscopic systems. Fiber optic approach uses completely different technology in contrast to conventional gyrocompass. It does not have moving parts, it has lower power consumption and is less limited during its nominal operation [1]. In this paper conventional and fiber optic compass are analyzed, and overview and comparison between these technologies is presented.

II. MECHANICAL VS FIBER OPTIC GYROCOMPASS

Precise information about the orientation and position in space is one of the key elements of information needed for successful navigation. Magnetic compasses in various forms have been used aboard the ship for centuries as a primary method for determining ships heading [2]. There were some inherent problems with precise heading measurements associated with magnetic compass. Errors that were caused by magnetic variation and deviation have been known to exist as early as 15th and 16th century, respectively but scientific research about the topic hasn't been published until 1843, and was done by Archibald Smith [3]. Construction of metal hull ships had increased magnetic compass issues regarding magnetic deviation even more, and with it, erratic compass behavior.

A. Historical overview

Gyrocompass was invented as an alternative to the traditionally used magnetic compass. At the beginning phases of research in gyroscopic effects, there was no rigorous scientific and theoretical background to the experiments done, just trial and error method were used [4]. Idea behind gyroscope began with construction of demonstration models by Johann Bohnenberger in 1817 and Leon Foucault in 1852, where the only purpose was demonstration of physical laws of the Earth's rotation. Demonstration presented by Focault was based on a swinging pendulum referenced to the horizontal plane. It demonstrated gyroscopic tendency of parallelism, where it was postulated that horizontally mounted revolving body would align itself parallel to the axis of Earth's rotation and point true north [5]. Although, the device as such couldn't be implemented on board a ship, next step in development came with gyrocompass patent by Barend Janse and Marinus Gerardus van den Bos. Their patent was based on applying Focault's demonstration principles and adapting them to the already available liquid magnetic compass devices. The rotating wheel axis would serve to determine true north and would be used instead of the

needle of the magnetic compass [6]. Independent development of the first electrically driven gyroscopes by Dumoulin-Froment in 1889 and by G.M. Hopkins in 1890 was innovation that enabled future advances in the use of a fast rotating wheel as a gyroscopic system [7]. In 1889 Arthur Krebs of the French Navy had instructed Dumoulin-Froment to develop and build an electric type gyroscope which was to be used in the French Navy vessels. An electric motor was added to the Dumoulin-Froment conventional marine gyroscope and implemented in the French experimental submarine Gymnote [8],[9]. Gyroscopic wheel had to be driven using an electric motor and its rotation had to be maintained at relatively high speeds. Electrical systems weren't commonly used aboard marine vessels at the time. New and yet unproven electrical and gyrocompass technologies combined with complexity of integration with existing marine systems discouraged the implementation of this technology on ships. Anschutz-Kaempfe continued research on the gyroscope in Germany. In 1904 he presented his first iteration of the gyroscope and in 1905 he presented its improved version. Both times he was rejected due to concerns with respect to the fragility of the device and specific delicate operating conditions. It was again improved and implemented in German Imperial Navy in 1908 [10]. Anschutz's gyrocompass didn't use topology presented by Focault. It was based on a spinning top with three degrees of freedom effected by gravity. Spinning axis was driven at 20000 revolutions per minute (RPM) and was suspended by the horizontal plane in mercury. Steel top center of gravity is either on the figure axis or on the equatorial top plane. Precession forces shift top axis in line with axis of Earth's rotation. Gyroscopic inertial oscillations around the primary axis are reduced by oscillation dampers [11]. Anschutz-Kaempfe's device was mounted pendulously and as such suffered ballistic deflection errors due to ship's acceleration. These errors would later be corrected by Schuler tuning method as was documented by Max Schuler. He described that if a gyrocompass is tuned with undamped period of 84,4 minutes which is the period that equals the pendulum of Earth radius, gyrocompass will self-correct for the effects caused by ship's acceleration. In that way gyroscope sensitivity to lateral motion could be significantly reduced or eliminated, and directional stability could be maintained [12], [13]. Elmer Sperry wanted to advance upon known gyrocompass designs. In 1908 he constructed first ballistic compass with vertical damping system which was later patented in 1911 and implemented on US Navy ships. With Sperry and Anschutz's gyrocompass navigational systems used in the US and German Navy, respectively, magnetic compass was ready to be complemented, if not completely substituted by gyrocompass as a primary device for finding and determining ship's heading [7].

B. Mechanical gyrocompass errors

Mechanical gyrocompass assembly consists from gyrosphere as the primary moving part, container for the gyrosphere as well as outer housing combined with local/remote control unit. Gyrosphere, which is the central

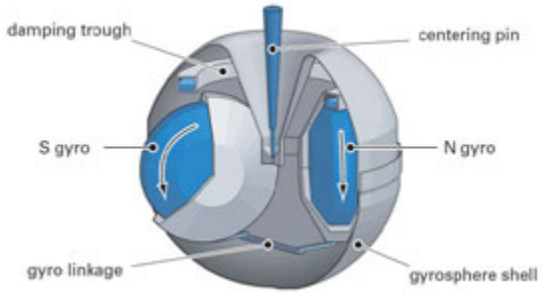


Fig. 1. Sperry Marine gyrosphere overview; Source: [14]

part and North-seeking element of the gyrocompass is comprised of two mechanically linked gyroscopes with spin axes horizontal with respect to the carrying frame as shown on Fig 3. Gyrosphere is placed within the container filled with liquid. This serves as a supporting medium for the gyrosphere. Complete gyrocompass system is comprised of housing, electronic mainboard, container and follow-up system as presented on Fig 4. A follow-up control system keeps the container aligned with respect to the gyrosphere at all time and as a result, heading is derived from container orientation [14]. Errors in the gyrocompasses can be separated into static and dynamic errors. Static errors can be further divided in alignment and transmission errors. Alignment error is an error which can be present if there is deviation between ship's heading and longitudinal line, but it can be successfully solved by alignment error correction algorithm. Transmission error is the error indicated between gyrocompass main unit and any of gyrocompass repeater units. Presently, transmission errors are kept to a minimum, by the use of modern efficient protocols with improved transmission speed.


Fig. 2. Gyrocompass system; Source: [14]

Dynamic errors are comprised from rolling, maneuvering and latitude errors. Mentioned errors happen primarily due to vessel movements. As the name implies, rolling error, occurs when the vessel is subject to rolling. Rolling error greatly depends on ships heading or sailing direction. On north/south oriented routes effect of rolling error is minimized. This effect is however maximally emphasized on the east/west routes. This happens due to maximal lateral acceleration that occurs in north or south direction causing the precession of the gyrocompass. In Sperry Marine modern electronically controlled gyrocompass rolling error impact is minimized by two mechanically linked gyroscopes as shown on figure 3. Mechanical linkage in combination with damping trough half filled with viscous fluid ensures that resultant spin vector of gyroscopes remains stationary relative to the gyrosphere, thus eliminating intercardinal roll error [15], [16]. Speed/latitude errors come from combination of vessel speed and Earth rotation. These errors are automatically corrected if gyrocompass is fed with position and speed sensor. It is important to understand that traditional mechanical gyrocompasses have some functional constraints. Every manufacturer describes their product constraints in installation manuals by exact numbers. For example, Sperry Marine gyrocompasses according to manufacturer notice is reliable between latitudes 70° N and 70° S and in this range heading error is nearly 0.4 x secant latitude. Latitudes over or under 70° are not covered by reliable operation for these gyrocompasses. Operational characteristics present that setting time of gyrocompass is less than three hours, maximal follow up speed 100°/sec and freedom of roll and pitch in real operation can be up to +/-40° and +/-90°. Mean time between failure (MTBF) is 40000 hours. Maximal ambient temperature range in operation is limited to -10°C to +55°C. On Fig 5. is presented correlation between heading error and latitude. Because of physical principles, similar correlation between heading error and latitude is present in any spinning mass gyrocompass. As a result of this correlation, heading information of any spinning mass gyrocompass system becomes unstable at high latitudes or unreliable. It is recommended that in polar navigation, as a heading reference fiber optic gyrocompass or should be used, in order to receive stable and reliable heading reference. Every spinning mass gyrocompass system requires periodical maintenance of internal parts and changing of supporting liquid and cooling fans every 1-1.5 year depends on manufacturer advice [14].

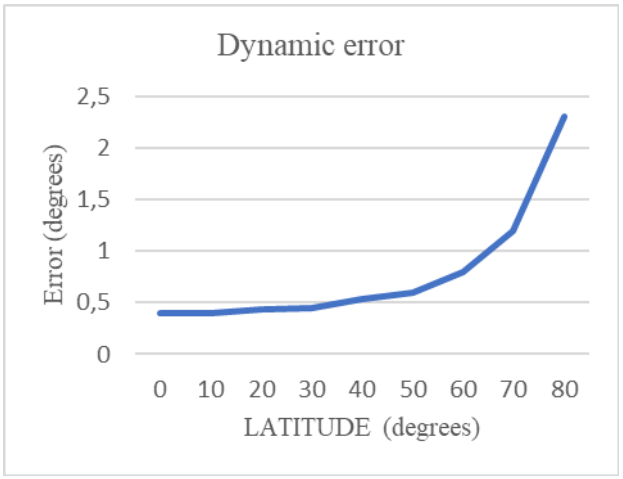


Fig. 3. Heading error and latitude correlation; Source: Author

C. *Fiber Optic Gyrocompass*

Operating principle of all fiber optic gyrocompasses (FOG) is based on Sagnac effect or propagated electromagnetic wave in a rotating medium which accumulates phase shift [17]. North-seeking element in fiber optic gyrocompass or FOG is fiber optic sensor. System overall has no moving parts in itself, and no gimbal system as a conventional gyrocompass with sensitive element or gyrosphere in container. For example, Sperry Marine FOG is constructed from three fiber optic gyros, dual axis electrolytic level sensor, navigation processor unit and a power supply. On Fig. 6. construction of Sperry Marine FOG is presented [18]. 3-rate gyros that are placed in X, Y and Z axes measure vessel rotation. Sperry Marine manufacturer calls this "analytical platform". From external speed and position sensor using Kalman filter algorithm, navigation processor calculating true North direction. Kalman filter basic operation is described in next section of this paper. Kalman filter algorithm activation is done by system startup. In other words, after power-up procedure system establishes a virtual plane or inertial plane. Electrolytic level sensor determines tilt position in order to place the plane horizontally at vessel's location. By making transformation of mentioned measurements system determines how the vessel's movements and earth's rotation act on mentioned virtual horizontal plane as presented on Fig.7.

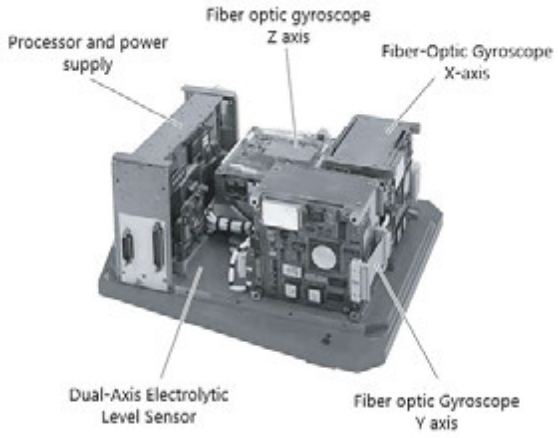


Fig. 4. Fiber optic gyrocompass "Sperry Marine"; Source: [18]

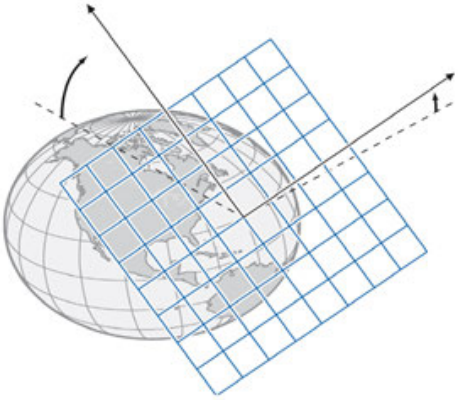


Fig. 5. Virtual horizontal plane; Source [18]

By evaluation of external speed and position and virtual horizontal plane tilt, system determines direction of true North. This is done by rotating the virtual plane in mathematical means until North-South axis of the plane will not tilt due to earth's rotation. Additionally, another FOG manufacturer is presented on Fig. 8., in order to present similarities of both brands. OCTANS is fiber optic gyrocompass and a motion reference unit for marine applications and provides true heading, surge, sway, roll, pitch, heave, rate of turn and accelerations. Basic system of OCTANS is inertial measurement unit equipped with three accelerometers, three fiber optic gyroscopes and a processor part. Processor part is digital signal processor (DSP) which enables real time calculations. Calculations are based on measurement of shifts in local gravity as the Earth rotates and consisting of angle integration using quaternion algebra, heading search algorithm and Coriolis force correction for vessel speed. All gyrocompasses including OCTANS are sensitive to vessel speed and sailing latitude, but these errors are small. Heading error depends on the secant of latitude according to general physical rule that heading cannot be defined at geographical poles. If the latitude information to gyrocompass is incorrect or inaccurate gyrocompass will generate heading error. However, these errors are very small. On figure 9 is presented amplitude of error due to inaccurate latitude data for 1 degree [19].

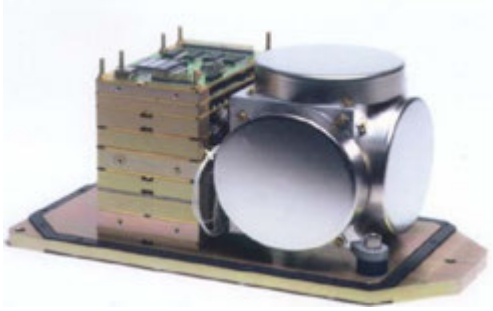


Fig. 6. Fiber optic gyrocompass "IXSEA OCTANS"; Source: [19]

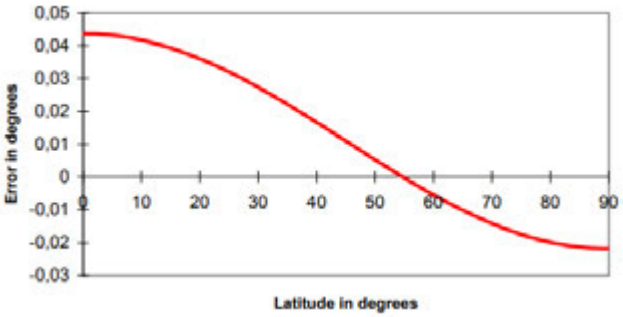


Fig. 7. Heading error in degrees by secant of latitude; Source: [19]

For example, at 40° latitude, an error of 3° in the latitude will cause 0,06 Sec. Lat. heading error. It is crucial to emphasize that fiber optic gyrocompasses such is OCTANS can be used as inertial navigation system. By this possibility FOG is made irreplaceable when compared to conventional gyrocompass. Basic concept of inertial navigation system is comprised of cluster of gyroscopes and cluster of accelerometers. Gyroscopes are sensors for rotation measurement, while accelerometers are sensors for acceleration measurement. Mentioned sensors represent inertial measurement system or inertial reference unit. In order to get all measurements from inertial navigation system, calculation of raw data is needed. Navigation processor or DSP calculates gravitational acceleration and doubly integrates net acceleration for estimation of position of object. FOG represents inertial navigational system produced in strapdown technology. Strapdown technology means that inertial navigation sensor is "strapped down" to the frame of host object, without using any moving parts like gimbals for rotational isolation. Possibility of FOG usage as inertial navigation system is important in areas or situations where Global Navigation Satellite System (GNSS) is jammed or spoofed, especially in military applications. It is unique system where heading and position information is all the time available, accurate and reliable. It is clear that fiber optic technology represents reliable solution in all external conditions and at same time technology which is immune on various threats in modern navigation [20].

D. Mechanical vs FOG

As presented in this paper, accuracy and operation of both conventional and FOGs is high enough for navigation purposes on modern vessels. Still today most common gyrocompasses on modern vessels are mechanical electronically controlled gyrocompasses. This is mainly the case due to lower price of aforementioned gyrocompasses, despite their limitations in operation.

Table I. Gyrocompass operation features comparison

	Mechanical Gyrocompass	Fiber Optic Gyrocompass (FOG)
Dynamic Accuracy	+/- 0.4 x Sec. Lat.	+/- 0.2 x Sec. Lat.
Settle Point Error	+/- 0.1 x Sec. Lat	+/- 0.1 x Sec. Lat
Settling Time	4 hours	3 minutes
Max Follow-Up Speed	100°/s	500˚/s
Operating Temperature	-10°C to +55˚C	-40˚C to +60˚C
Operation Range (Roll, Pitch)	+/- 40˚	no limitation
Power Consumption	80W (max)	12W (max)
Latitude Limitation	reduced accuracy above 60˚N or 60˚S	no limitation
Maintenance Requirement	every year	no

Conventional gyrocompasses have bigger settling time which is about 4 hours, while FOGs have settling time less than 10 minutes. FOGs are produced in strapdown technology without moving parts, while conventional gyrocompasses have moving parts and require preventive maintenance. FOGs are completely maintenance free with MTBF 10 times bigger than conventional gyrocompasses. It is very important to emphasize that in case of fatal failure of FOG it could be very hard to get immediate support, and this is the reason for having backup solution or doubled FOG onboard. FOGs getting more used specially on research and polar vessels where conventional gyrocompasses have excessive limitations due their construction and working principle [18, 19, 20]. In table 1 is presented comparison of all critical working points of FOGs and mechanical gyrocompasses.

III. CONCLUSION

Nowadays, gyrocompass represents a complete subgroup within navigation system topology. In this paper two kinds of gyrocompasses that are widely and commonly used in modern merchant navy are presented. All-important operation features are discussed form design and economical point of view. Conventional mechanical gyrocompasses are still most widely installed on merchant navy and the reason for this is economic affordability. With continual improvements and research. Strapdown technology is becoming more present due to increases in of reliability and accuracy of strapdown gyrocompasses. FOGs still represent future of the gyrocompass systems onboard ships. It is clear that both gyrocompasses are accurate and reliable for navigation purposes. Author's thesis is that future technologies with lower power consumption and space requirements with swappable technology without overcomplex preventive maintenance is coming to the focus.

REFERENCES

[1] Jurdana Irena, Sušanj Josip, "Sustavi elektroničke navigacije", Sveučilište u Rijeci, Pomorski fakultet u Rijeci, 2013.

[2] Universidad de Alicante, "Compass." [Online]. Available: https://web.ua.es/docivis/magnet/compass.html

[3] R. Doerfler, "Dead Reckonings » Magnetic Deviation: Comprehension, Compensation and Computation (Part I)." [Online]. Available: http://myreckonings.com/wordpress/2009/04/18/magnetic-deviation-comprehension-compensation-and-computation-part-i/

[4] J. Broelmann, "The Development of the Gyrocompass - Inventors as Navigators," J. Navig., vol. 51, no. 2, pp. 267–273, 1998

[5] J. Sommeria, "Foucault et la rotation de la Terre," *Comptes Rendus Phys.*, vol. 18, no. 9–10, pp. 520–525, 2017.

[6] P. L. Galison, *How Experiments End*, 1st ed. Chicago: University of Chicago Press, 1992.

[7] Sperry Marine, "History of Sperry Marine | Sperry Marine." [Online]. Available: https://www.sperrymarine.com/corporate-history/sperry-marine.

[8] "KREBS - Le gyroscope électrique du GYMNOTE 1889." [Online]. Available: http://rbmn.free.fr/KREBS_Gyroscope.html.

[9] "French submarine Gymnote (Q1) - Alchetron, the free social encyclopedia." [Online]. Available: https://alchetron.com/French-submarine-Gymnote-(Q1).

[10] "Gyrocompass - History, Construction, Working, and Usage." [Online]. Available: http://www.historyofcompass.com/compass-facts/gyrocompass/.

[11] F. Klein and A. Sommerfeld, *The Theory of the Top. Volume IV*, vol. IV. 2014.

[12] H. W. Sorg, "From Serson to Draper—Two Centuries of Gyroscopic Development," *J. Inst. Navig.*, vol. 23, no. 4, pp. 313–324, 1976.

[13] A. D. King, "Inertial Navigation – Forty Years of Evolution," *GEC Rev.*, vol. 13, no. 3, pp. 140–149, 2009.

[14] Northrop Grumman Sperry Marine, Operation, Installation and Service Manual, Navigat X MK1, November, 2013.

[15] CHENG Jian-hua. Study on the Crucial Technology of Platform Inertial navigation System. Doctor paper of Harbin Engineering University, 2006.

[16] Huang, Weiquan, Hao, Yanling, Cheng, Jianhua, Li, Gang. Research of the inertial navigation system with variable damping coefficients horizontal damping networks. Ocean '04 - MTS/IEEE Techno-Ocean '04, 2004.

[17] E. J. Post, "Sagnac effect,"Rev. Mod. Phys., vol. 39, no. 2, pp. 475–493,1967

[18] Northrop Grumman Sperry Marine, Operation, Installation and Service Manual, Navigat 3000, November, 2013.

[19] Ixsea, Octans III, User guide, 2004.

[20] Mohinder S., Lawrence R., Angus P., Global Positioning Systems, Inertial Navigation, and Integration, 2001., John Wiley & Sons, Inc.

[21] Grewal M.S., Weill L.R., Andrews A.P., "Global Positioning Systems, Inertial Navigation and Integration", John Wiley & Sons, 2001.

[22] Lawrence A., "Modern Inertial Technology. Navigation, Guidance and Control", Springer 2001.

DewCom

Special Session on

DEW COMPUTING

Steering Committee

Chairs: Karolj Skala, Ruđer Bošković Institute, Zagreb, Croatia
Marjan Gushev, Ss. Cyril and Methodius University, Skopje, North Macedonia

Members: Ruppa Thulasiram, University of Manitoba, Winnipeg, Canada
Yingwei Wang, University of Prince Edward Island, Charlottetown, Canada

Program Committee

Members: Dana Petcu, West University of Timisoara, Romania
Davor Davidović, Ruđer Bošković Institute, Zagreb, Croatia
Enis Afgan, Ruđer Bošković Institute, Zagreb, Croatia
Karolj Skala, Ruđer Bošković Institute, Zagreb, Croatia
Marjan Gushev, Ss. Cyril and Methodius University, Skopje, North Macedonia
Parimala Thulasiram, University of Manitoba, Winnipeg, Canada
Partha Pratim Ray, Sikkim University, Gangtok, India
Ralph Deters, University of Saskatechewan, Canada
Ruppa Thulasiram, University of Manitoba, Winnipeg, Canada
Shuhui Yang, Purdue University Northwest, Lafayette, United States
Sven Groppe, Institute of Information Systems, Lübeck, Germany
Yingwei Wang, University of Prince Edward Island, Charlotteton, Canada

PAPERS

Microservice approach to the qualitative study of attractors of binary dynamic systems based on the Boolean constraint method

G.A. Oparin, V.G. Bogdanova, and A.A. Pashinin
Matrosov Institute for System Dynamics and Control Theory of SB RAS, Irkutsk, Russia
prn51@mail.ru, bvg@mail.ru, apcrol@gmail.com

***Abstract* - In recent years, due to the active development of distributed computing technologies, the automation of computations based on these technologies became very important for solving the problems of qualitative research of binary dynamic systems, which are widely used as mathematical models in cryptography, bioinformatics, the study of fault tolerance of computer networks, and some other domains. In this regard, the issues are actualized of developing tools and providing infrastructures for the microservice-based implementation of scientific applications, which are characterized by the dynamism of resource requirements and oriented to use in a hybrid cloud environment. We offer microservice-oriented tools for the automation of solving computationally complex problems of a qualitative study of attractors of binary dynamic systems in a hybrid computing environment. This study is based on the Boolean constraint method and allows the finding of evolutionary parameters of system states (such as, for example, radius, period, and branching) without modeling system dynamics. These tools provide the agents implemented as web-services for the integration of resources of on-premises computers with cloud resources using Dew computing. An example of the application of the developed tools is given.**

Keywords - microservice; microservice-oriented tools; qualitative analysis; hybrid computations

I. INTRODUCTION

Binary dynamic systems (BDS) are widely used as basic mathematical models in many subject areas, in particular in cryptography, bioinformatics, economics, and sociology. We apply the declarative approach to a qualitative analysis of BDS based on the Boolean Constraint Method (BCM) [1]. The fundamental point of BCM is the use of formal definitions of dynamic properties, presented in the language of applied logic of predicates with limited quantifiers of existence and universality, understandable to a specialist in dynamic analysis of systems. Such specification of property (instead of its verbal description) disambiguates the interpretation of the meaning of the property. Based on this approach, a Boolean model of the dynamical property for BDS, whose functioning is considered at a finite time interval, is constructed, and its feasibility is checked. The uniqueness of the constructed Boolean model lies in the fact that it contains restrictions describing both the dynamics of the BDS and the specification of its required property.

A BCM-based qualitative study of BDS requires the software to build Boolean models and verify their feasibility. The computational complexity of some problems of this study leads to the necessity of the use of a high-performance computing environment to solve them. We developed the above software as Applied Microservices Package (AMP) for BCM-based Qualitative Analysis of BDS (BCM-QABDS). The advantages of this AMP are the reduction of computational complexity by using small, loosely coupled, reusable, autonomous microservices [3], and simplicity of its scaling, deploying, and updating. Applied software is implemented as microservices interacting by using lightweight communication mechanisms. This implementation is based on the authors' technology HPCATAMP (High-Performance Computing Automated Technology for implementation of Applied Microservices Package) [2]. The AMP, based on this technology, is an integrated intelligent system for managing distributed computations in a specific subject domain. This system automates these computations using a semantic peer-to-peer (P2P) network of agents deployed on nodes of the computational field (CF). Agents installed on these nodes control the launch of computational microservices.

Further development of the HPCATAMP uses advantages of the integration of P2P network and agents technology for providing additional cloud resources. This possibility is provided due to the openness of this resulted computational network, mobility of AMP agents, and new cloud version HPCSOMAS-MSC of the previous basis software platform [2] of this technology. An agent-manager for processing the request of cloud resources was developed in the new version. The technological level of knowledge used by agents for organizing hybrid computing, allocating additional cloud resources, and ensuring synchronization of on-premises and user cloud data was added to the conceptual model of the subject domain.

A conceptual model of knowledge representation using this technology in the subject domain of a BCM-based qualitative study of BDS is presented. An illustrative example of the usage of the BCM-QABDS for the study of attractors of BDS is considered.

II. Related Work

Currently, one of the software development trends is an integration of distributed and cloud computing technologies in solving resource-intensive complex scientific and applied problems. Some publications (for example, [4-5]) discuss the development of tools and infrastructures for implementation of resource-require scientific applications based on microservice technology, including applications oriented to use in a hybrid cloud environment [6, 7]. Hybrid cloud infrastructure combines the reliability and availability of software running on on-premises computers with the ability to scale computing to the cloud when peak loads occur [4]. Cloud integration models in a hybrid environment are considered in [8, 9]. In [10], a taxonomy of cloud deployment patterns is presented. The template selection is determined by the aims of forming hybrid infrastructure and the requirements for developing service-oriented applications. In our research, a hybrid infrastructure based on a template similar to the "on-premises pattern" [9] is used to solve the problems of studying dynamics and structural-parametric synthesis of BDS based on BCM [1]. The active development of new and improvement of existing methods for the qualitative study of BDS (mainly models of genetic networks) observed in the last decade. However, only a small number of publications are devoted to the direction of the application of web and distributed technologies in creating tools for the qualitative study of BDS. Examples of it are the web-based tool Analysis of Dynamic Algebraic Models (ADAM) [11] and CoLoMoTo [12]. ADAM automatically converts several types of discrete models into polynomial dynamical systems and analyses their dynamics using computer algebra tools. The CoLoMoTo provides access to the range of command-line tools for the qualitative study of the Boolean network (BN). The BoolSi [13] is intended for distributed simulation of synchronous BN. This tool is oriented for use in a high-performance environment. However, it has a command-line interface. We develop an AMP BCM-QABDS intended to automate distributed and parallel solving the problems of a qualitative study of BN in a high-performance hybrid environment on dedicated CF resources. This tool is cross-platform, has web and API interfaces. CF can include both cloud and on-premises resources, in particular clusters, dedicated servers, and personal computers (PC). Unlike the listed above approaches, a BCM-based declarative approach for the qualitative study of BDS allows data parallelism by splitting the Boolean model of the verifying BDS property. This approach provides significantly increasing the dimension of the BDS and the time interval of its functioning.

III. Conceptual Model of BCM-Based Qualitative Analysis Using HPCATAMP

A. Problem Formulation

We consider a synchronous autonomous BDS, which has the form

$$x^t = F(x^{t-1}), \quad (1)$$

where x is the state vector of the dimension n ($x \in B^n$, $B = \{0,1\}$), $t \in T = \{1,2,\ldots,k\}$ is the discrete time, and $F(x)$ is the vector function of logic algebra called the transition function ($F : B^n \to B^n$). Let us define the trajectory $x(t, x^0)$ of (1) for each initial state $x^0 \in B^n$ as the finite sequence of states $x^0, x^1, \ldots, x^k$ from the set B^n. The state x^i is a successor of the state x^{i-1}, and the state x^{i-1} is a predecessor of the state x^i. In an autonomous BDS, each state has only one successor, and the number of predecessors of this state can vary from zero to 2^n-1. A cycle of length k is a closed trajectory ($x^0 = x^k$) in which other states are pairwise distinct. An equilibrium state is the cycle of the length k=1. A cycle without predecessors is an isolated one. The non-isolated cycle is called an attractor. Let X^* is the attractor. The region of attraction (basin) of an attractor of the radius k is the set of all states, from which the set X^* is reachable in $i \le k$ time steps.

A qualitative analysis of BDS, based on the study of the structure of the state space, requires solving the following problems:

- Searching equilibrium states and cycles including cycles of the given length k, and verifying their isolation property
- Description of attractors (their number, type, length)
- Determining the attraction region (basin) of the attractor and its radius, or constructing the attraction region of the given radius k
- Checking the property of belonging the state to the attractor and determining the distance (number of steps) to this attractor
- Determining immediate predecessors for given states
- Checking the reachability property of the target set X^* from the initial values set X^0.

B. BCM-based Qualitative Study of BDS

Solving the above problems based on BCM includes three stages:

- Building a property model in the form of Boolean constraints based on the description of the dynamics of the system (1) and property specifications [1]
- Verification of satisfiability of this model
- Post-processing the results.

We present the equations from [1], which are required for solving these problems.

The system (1) with the initial state $x^0 \in X$ for $t = \{1,2,\ldots,k\}$ is equivalent the one Boolean equation ([1])

$$\Phi(x^0, x^1, \ldots, x^k) = \vee_{t=1}^{k} \vee_{i=1}^{n} (x_i^t \oplus F_i(x^{t-1})) = 0. \quad (2)$$

Equation (2) for k=1 (one-step transitions is considered only) has the following form:

$$L(x^0,x^1)=\vee_{i=1}^{n}(x_i^1\oplus F_i(x^0))=0. \quad (3)$$

The state x^1 is the successor of x^0, and the state x^0 is the predecessor of x^1 in (3). All immediate predecessors x^0 of the state $s\in X$ are solutions of the next Boolean equation:

$$L(x^0,x^1)\big|_{x^1=s}=0. \quad (4)$$

Predecessors of the state s are absent if there are no solutions to the equation (4). A cycle of length k (if it exists) is a solution to the Boolean equation

$$\Phi(x^0,x^1,\dots,x^k)\big|_{x^0=x^k}\vee R(x^0,x^1,\dots,x^{k-1})=0, \quad (5)$$

where the condition for pairwise differences of all states of the set C of a cycle of length k is given by the expression

$$R(x^0,x^1,\dots,x^{k-1})=\vee_{\substack{1\le q\le k-1,\\ k \bmod q=0}}\wedge_{i=1}^{n}y_i^q=0,$$

$$y_i^q=(x_i^0\cdot x_i^q\vee\overline{x_i^0}\cdot\overline{x_i^q}).$$

Let the elements s of the set C be determined by the solution of the Boolean equation $G^c(s)=0$. Then the condition of isolation of the cycle is equivalent to the absence of solutions to the following Boolean equation [1]:

$$G^c(s)\vee L(x^0,x^1)\big|_{x^1=s}\vee\overline{G^c}(x^0)=0. \quad (6)$$

The attraction region of the attractor X^* of the radius k is determined by the set of solutions of the Boolean equation

$$\Phi(x^0,x^1,\dots,x^k)\vee G^*(x^k)=0 \quad (7)$$

concerning variables $x^{k-1},x^{k-2},\dots,x^0$.

The set of parameters AMP BCM-QABDS consists of Boolean models and variables used for their building, sets of states, descriptions of functions, and description formats. The set of modules includes programs for building Boolean models, converting formats from one to another, verifying the truth of models, and post-processing the results. Both of these sets represent algorithmic knowledge for a qualitative study of BDS based on BCM.

C. Knowledge Representation

Algorithmic knowledge about mathematical models, methods, and techniques for their research have a complex hierarchical structure in which there are three conceptually isolated layers: computational, schematic, and production knowledge [14]. In this study, the development of HPCATAMP technology in the direction of hybrid computing automation with Dew-computing support [15] required the inclusion of a new layer of knowledge - a technology layer. This layer provides information support for decision-making by managing agents, in particular, about the need for synchronization, about the possibility of connecting cloud resources. The conceptual model of HPCATAMP-based AMP usage has the form:

$$\text{CM}=<C, A, P, G, T, P^F, M, N, U^P, U^T, R>,$$

where C is the set of *complexes*, which are isolated (structurally) sets of knowledge. These complexes combine objects of the subject domain, corresponding to a family of research methods that intersect in sets of parameters and operations. In a hybrid infrastructure, set C includes three subsets, $C^P\subset C$, $C^D\subset C$, and $C^C\subset C$. The first subset contains agents PSA (Problem Statement Agent), the second – web services installed on the on-premises resources (Dew PSA), and the third – newly developed agents CRA (Cloud Resource Agent), designed to allocate resources in the cloud at the request of the PSA. The set A is divided into four subsets of agents: A^C, A^P, A^I, and A^G. Elements of the subset A^C are DSA (Distributed Solver Agent) that control the launch of the Computational Microservice (CM). Subsets A^P, A^I and A^G consist of, respectively DSAProd agents (DSA Production) using the productional knowledge for the CM launch; the initial (DSAInit), and target (DSAGoal) agents, which are created by the PSA and DPSA when the user formulates the problem [16]. The set of parameters P includes input and output parameters stored in the local knowledge base (KB) of agents from the set A. The set of parameters $P=\bigcup_{i=1}^{n_A}P_i^L$, where n_A is a number of all agents of AMP, P_i^L are the set of input and output parameters from the local (distributed on CF nodes) knowledge base (KB) of the i-th agent. Elements of the set G are active groups of agents formed for the non-procedural (declarative) statement (NPS). Each new NPS is added in the set T, initially empty. The structure T_i ($T_i\in T$) includes an input (T_i^I) and output (T_i^O) parameters, names of which are selected from the parameter dictionary P^F, and parameter values (T_i^D). The set P^A includes values of input parameters by which the output ones are calculated.

The sets M and N contain computational microservices and CF nodes, and the sets U^P and U^T include references to user profiles and their data updating tables. The set R = {*In*, *Out*, *Pr*, *Inc*, *Imc*, *Isc*, *Icv*} (Fig. 1) consists of relationships of the following types: many-to-many (pairs of the Cartesian product of corresponding sets), one-to-one, and one-to-many (functional relationships, the semantic of which are presented in [17]). For example, *In* has many-to-many type:

$$In(P^L,A^C)\subset P^L\times A^C:(P_i^L,A_i^C)\big|P_i^L\in P^L,A_i^C\in A^C.$$

D. Architecture and Functioning of AMP

The scheme of interaction of agents (previously deployed to the nodes of the CF) is presented in Fig. 2. The active group G_i is formed for the NPS $T_i=(T^I;T^O;T^D)$ or briefly, $T_i=(T^I;T^O)$, by the logical inference on the distributed KB of DSA. Fragments of relationships *In* (P^L,

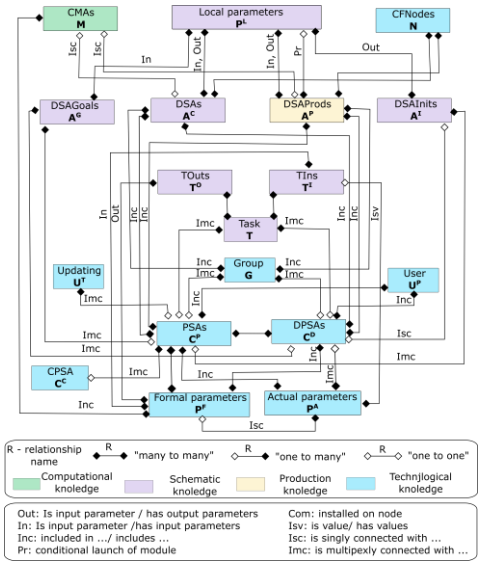


Figure 1. Conceptual model of HPCATAMP-based AMP

A^C) and *Out* (A^C, P^L) ($In \in R$, $Out \in R$) of this KB determine pair interactions of DSA in the process of group formation and solving a problem. The KB contains relationships $Com \subset A^C \times N$ also. The PSA sets the attribute IP of $N_i \in N$ in KB of DSA during its initialization after a preliminary availability check. PSA needs technological knowledge when executing a user request to organize backup processes, synchronize local and cloud knowledge bases, handle CF node failure situations, conduct testing when updating a microservice by a developer, and attract additional cloud resources. Backup and data synchronization is required to save the results and the ability to continue to work when the Internet is turned off. The user installs Dew PSA on his computer and configures it to synchronize the results of computations and input files from available PSA agents upon request, schedule, or following the options specified in the profile.

Scaling cloud computing at the top level of abstraction is as follows. DSAs manage the message queue, and CMAs control the task queue. In the case of queue congestion, the CMA informs DSA about resource exhaustion. DSA sends a request for an additional resource to the PSA, where it is registered. PSA contacts

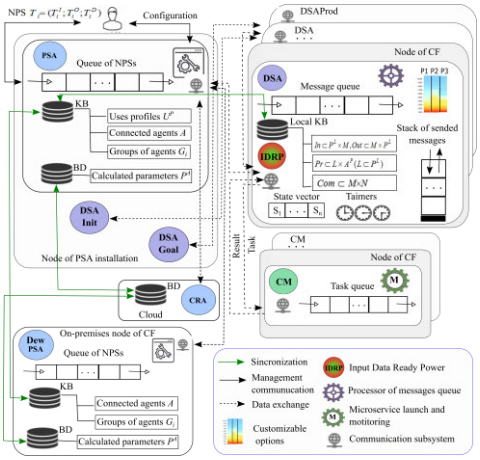


Figure 2. The interaction of the AMP agents

TABLE I. HYBRID INFRASTRUCTURE

Resource	Resource type	Installed software	
		Agents	*Additional software*
Nodes of the cluster «Akademik V.M. Matrosov» [18]	On-premises	PSA, Ag_6, Ag_7, Ag_{11}, Ag_{13}	Tomcat Server
VDS [19]	Private Cloud	PSA Ag_1-Ag_5 Ag_8-Ag_{10}, Ag_{12}, Ag_{13}	SageMath VirtualBox, Docker Server, Docker Container
	Public Cloud	CRA Ag_{13}	Kernel Virtual Machine Tomcat Server
PC	On-premises	DPSA, Ag_1-Ag_3, Ag_{12}, Ag_{14}-Ag_{16}	Tomcat Server

with the special CPSA agent if the option to use a cloud resource is set in the user profile. In this case, the CPSA launches the Docker container with the required CMA, returning a message about the allocation of the resource and the network address for communication with it. Otherwise, a request-rejection message is sent, and tasks are executed in the order of the task queue.

IV. ILLUSTRATIVE EXAMPLE

We used the AMP BCM-QABDS for the study of attractors of BDS. This problem has a practical application in bioinformatics. Attractors correspond to a combination of gene expressions, which specifies a particular cell type or cell fate of an organism [21, 22]. The sizes of the basins of attraction may indicate which attractor is more stable [21]. The computing resources on which the AMP BCM-QABDS microservices are deployed are shown in Table 1. We study the attractors using the following scheme for verifying BDS properties. A search for equilibrium states is carried out according to (4). The isolation property of the found equilibrium states (*ES*) is checked using (6). The equilibrium state is an attractor if the isolation property is not satisfied (there are no immediate predecessors in the array *A_IP*). The basin of attraction (*AB*) is founded as solutions of equation (7). The fragment of the computational model for this scheme is presented in Fig. 3. Problems of verifying these properties are solved both separately and jointly (as task flow).

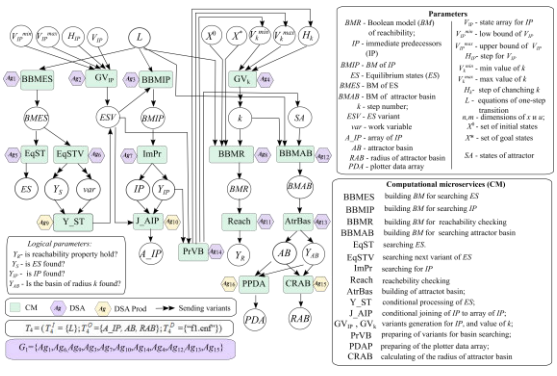


Figure 3. The fragment of the computational model

For example, in the first case, the following NPSs for these problems are used: $T_1 = (T_1^I = \{L\};\ T_1^O = \{ES\}; T_1^D = \{\text{“f1.cnf”}\});\ T_2 = (T_2^I = \{L,\ ESV\}; T_2^O = \{A_IP\});\ T_3 = (T_3^I = \{L,\ SA\};\ T_3^O = \{AB,\ RAB\})$. In the second case, the following NPS is formulated:

$T_4 = (T_4^I = \{L\}; T_4^O = \{A_IP,\ AB,\ RAB\}; T_4^D = \{\text{“f1.cnf”}\})$.

In the above NPSs, the text “f1.cnf” from the set T_i^D is the file name of the description of function L (4), which is built preliminary [16]. All NPSs are formulated using the web-interface of PSA (Dew PSA). Sets T_i^I and T_i^O are selected from the P^F. For the NPS T4, the active group G4 = {Ag1, Ag6, Ag9, Ag3, Ag7, Ag10, Ag14, Ag4, Ag12, Ag13, Ag15} is formed (Fig.3). The group of DSAs performs decentralized control for solving the problem based on direct pair interactions. Agents behavior is coordinated by the input data readiness (the event control is used) [20]. In the case of task flow (NPS T4), computations are carried out by parallel-pipeline way [14]. The next equilibrium state is processed according to the above scheme. For testing developed AMP, the BDS was used from [21] and [23]. The NPS T_4 was used to study a phase portrait for BDS with state vector dimensions n=15 [21] and n=11 [23], dynamic of which is described correspondently by the following equations:

$$
\begin{aligned}
&x_1^t = x_{10}^{t-1}\cdot x_{14}^{t-1}\cdot x_{15}^{t-1}\cdot x_1^{t-1} \vee x_5^{t-1}\cdot x_{14}^{t-1}\cdot x_{15}^{t-1}\cdot x_1^{t-1} \vee \\
&\vee x_7^{t-1}\cdot x_2^{t-1}, x_2^t = x_2^{t-1}, x_3^t = \overline{x_5^{t-1}}\cdot\overline{x_{13}^{t-1}}, x_4^t = \overline{x_6^{t-1}}, \\
&x_5^t = \overline{x_{10}^{t-1}}\cdot\overline{x_{13}^{t-1}} \vee x_7^{t-1}\cdot\overline{x_{10}^{t-1}} \vee x_4^{t-1}\cdot\overline{x_{10}^{t-1}}, \\
&x_6^t = \overline{x_7^{t-1}}, x_7^t = \overline{x_{13}^{t-1}} \vee \overline{x_6^{t-1}}, x_8^t = \overline{x_{13}^{t-1}}, \\
&x_9^t = x_9^{t-1}\cdot\overline{x_{15}^{t-1}} \vee x_9^{t-1}\cdot\overline{x_{10}^{t-1}}, \\
&x_{10}^t = x_7^{t-1}\cdot x_{10}^{t-1}\cdot x_{15}^{t-1} \vee \overline{x_8^{t-1}}\cdot\overline{x_{13}^{t-1}} \vee x_7^{t-1}\cdot\overline{x_{12}^{t-1}} \vee \\
&\vee x_7^{t-1}\cdot\overline{x_{11}^{t-1}} \vee x_7^{t-1}\cdot x_9^{t-1} \vee \overline{x_5^{t-1}}\cdot x_7^{t-1} \vee x_7^{t-1}\cdot\overline{x_8^{t-1}}, \\
&x_{11}^t = 1, x_{12}^t = 1, x_{13}^t = \overline{x_5^{t-1}}\cdot x_6^{t-1}\cdot\overline{x_7^{t-1}}, \\
&x_{14}^t = x_5^{t-1}\cdot x_{14}^{t-1}\cdot x_{15}^{t-1}\cdot x'^{t-1}_1 \vee x_{10}^{t-1}\cdot x_{14}^{t-1}\cdot x_{15}^{t-1}\cdot x'^{t-1}_1 \vee \\
&\vee x_7^{t-1}\cdot x'^{t-1}_{10} \vee x_7^{t-1}\cdot x'^{t-1}_1, x_{15}^t = x_7^{t-1}.
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
&x_1^t = x_2^{t-1}, x_2^t = x_3^{t-1}\cdot x_{10}^{t-1}, x_3^t = x_4^{t-1}\cdot x_7^{t-1}, \\
&x_4^t = x_5^{t-1}\cdot x_7^{t-1}, x_5^t = x_8^{t-1}\cdot x_9^{t-1}, x_6^t = x_3^{t-1}, \\
&x_7^t = x_6^{t-1}\cdot x_8^{t-1}, x_8^t = x_5^{t-1}\cdot x_9^{t-1}, x_9^t = x_8^{t-1}, \\
&x_{10}^t = x_8^{t-1}\cdot x_{11}^{t-1}, x_{11}^t = 0.
\end{aligned}
\tag{9}
$$

Table II shows the results of solving NPS T_4 for these systems. The speedup of distributed parallel-pipeline computations was achieved 7.5 and 2.8 times in comparison with the BCM-based sequential execution. The *AB* solution result (the attractor basin), is visualized using the NPS $T_5 = (T_5^I = \{AB,\ Y_AB\}; T_5^O = \{PDA\})$ (Fig. 4). The decimal state codes of the attractor are calculated by the formula $\sum_{i=1}^{n} a_i * 2^{i-1}$. The Dew PSA

TABLE II. RESULTS OF STUDY

DD	No	*Attractor*	Attractor Type		
			Length	***Basin size***	***Max radius***
(8)	1	100110110011011	1	8	2
(8)	2	000110110011011	1	440	3
(8)	3	110110110011011	1	448	3
(8)	4	000001000011100	1	512	2
(8)	5	010001000011100	1	512	2
(8)	6	010001001011100	1	256	2
(8)	7	000001001011100	1	256	2
(8)	8	111100110111011	1	15168	9
(8)	9	101100110111011	1	568	4
(8)	10	001100110111011	1	14600	9
(9)	1	00000000000	1	1792	4
(9)	2	00001001100	1	240	10
(9)	3	00111111100	1	16	4

with microservice agents less demanding on resources (for example, the builder of the Boolean model, sequential SAT and 2QBF solvers, and post-processor of results) can be installed on on-premises computers. Parallel solvers SAT and 2QBF are installed on on-premises high-performance clusters. BCM-based sequential execution has the advantage in comparison with the sequential BNS-solver [21, 24] for BDS of both the small and the large dimensions n of the state vector (Fig. 5). The BNS achieved the time limit in twelve hours in the searching of cycles of the given length on the value k=30 of BDS, which is the model of stream cipher based on shift registers (Fig. 6). The BCM-based approach allows increasing the interval k of functioning this BDS by splitting constructed Boolean model for searching cycles and subsequent parallel solving subtasks (Fig. 6). The previously available number of resources on NPS startup is allocated, and the dynamical request, according to the need for additional cloud resources, is sent for providing more effective the execution of these subtasks

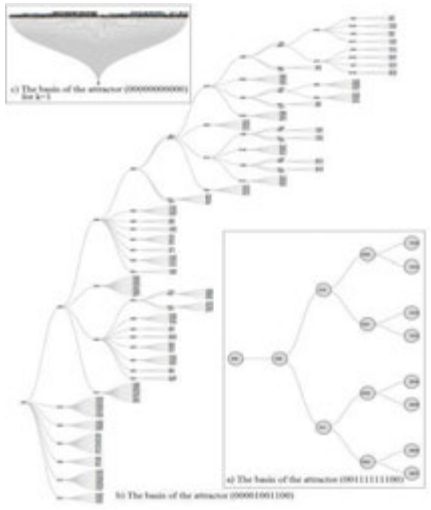


Figure 4. Basins of attractors of BDS (9)

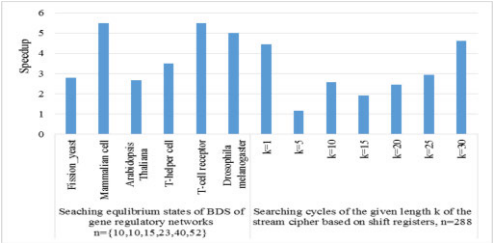


Figure 5. The speedup (S=T_{BCM} / T_{BNS}) of BCM- approach in comparison with the BNS solver, based on algorithmic approach, in sequential solving qualitative analysis problems

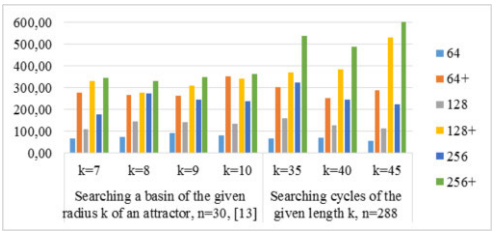


Figure 6. The speedup S=T / T^+, where T and T^+ are runtimes correspondently on the time-independent number of nodes (e.g., 64) and number nodes increasing by cloud node temporarily (e.g., 64+)

in the case of queuing delay of CMA (Fig. 3). The resource is freed after completing the tasks sent on it.

V. CONCLUSION

We proposed further development of HPCATAMP technology for automation of hybrid computing, allocating additional cloud resources, and ensuring synchronization of on-premises and user cloud data. We developed a new version of the program platform based on which this technology is implemented. Using this technology, we developed a package of applied microservices BCM-QABDS for the qualitative study of BDS based on the Boolean constraint method, oriented on an application in the different subject domains, where BDS are used as models of research objects. The conducted experiments confirmed the effectiveness of the offered approach and developed software for solving practically significant scientific problems of the qualitative study of BDS.

ACKNOWLEDGMENT

The present investigation was supported by Russian Foundation of Basic Research, projects no. 18-07-00596.

REFERENCES

[1] G. Oparin, V. Bogdanova, and A.Pashinin, “Qualitative analysis of autonomous synchronous binary dynamic systems,” in Mathematics in engineering, science and aerospace (MESA), vol. 10, no. 3, 2019 pp. 407-419.

[2] G.A. Oparin, V.G. Bogdanova, A.A. Pashinin, and S.A. Gorsky, “Microservice-oriented Approach to Automation of Distributed Scientific Computations,” in Proceedings of the 42nd International Convention on Information and Communication Technology, Elec-tronics and Microelectronics (MIPRO), IEEE, 2019, P. 236-241.

[3] S. Newman, “Building Microservices”. O’Reilly, 2015.

[4] V. Navale and P.E. Bourne, “Cloud computing applications for biomedical science: A perspective,” PLOS Comp Biol, vol. 14(6):e1006144, 2018.

[5] M.A. Netto, R.N. Calheiros, E.R. Rodrigues, R.L. Cunha, and R. Buyya, “HPC Cloud for Scientific and Business Applications: Taxonomy, Vision, and Research Challenges,” ACM CSUR, vol. 51(1), 2018, pp. 8:1-8:29.

[6] P. Mell and T. Grance, “The NIST Definition of Cloud Computing,” 2011, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication 800-145.pdf [online, accessed: 24-Jan-2020].

[7] K. Bakshi, “Secure hybrid cloud computing: Approaches and use cases,” IEEE Aerospace Conference, Big Sky, MT, 2014, pp.1-8.

[8] D. Merkel, F. Santas, A. Heberle, and T. Ploom, “Cloud Integration Patterns,” 4th European Conference on Service-Oriented and Cloud Computing (ESOCC), Taormina, Italy, 2015, pp.199-213.

[9] “CSCC: Practical Guide to Cloud Computing,” https://www.omg.org/cloud/deliverables/CSCC-Practical-Guide-to-Hybrid-Cloud-Computing.pdf [online, accessed: 19.10.2019].

[10] L.C. Ochei, A. Petrovski, and J.M. Bass, “A Novel Taxonomy of Deployment Patterns for Cloud-hosted Applications: A Case Study of Global Software Development (GSD) Tools and Processes,” in International Journal on Advances in Software, vol. 8(3 & 4), 2015, pp. 420-434.

[11] F. Hinkelmann, M. Brandon, B. Guang, R. McNeill, G. Blekherman, A. Veliz-Cuba, and R. Laubenbacher, “ADAM: Analysis of Discrete Models of Biological Systems Using Computer Algebra,” in BMC Bioinformatics, vol. 12:295, 2011.

[12] A. Naldi, C. Hernandez, N. Levy, G. Stoll, P.T. Monteiro, C. Chaouiya, T. Helikar, A. Zinovyev, L. Calzone, S. Cohen-Boulakia, D. Thieffry, and L. Paulevé, “The CoLoMoTo Interactive Notebook: Accessible and Reproducible Computational Analyses for Qualitative Biological Networks,” Frontiers in Physiology, vol. 9, 2018, article 680.

[13] V. Oles and A. Kukushkin, “BoolSi: a tool for distributed simulations and analysis of Boolean networks,”, arXiv:1910.03736v2.

[14] I.V. Bychkov, G.A. Oparin, V.G. Bogdanova, and A.A. Pashinin, “Intellectual Technology for Computation Control in the Package of Applied Microservices.” in Proceedings of the 1st International Workshop on Information, Computation, and Control Systems for Distributed Environments, ICCS-DE 2019, CEUR-WS Proceedings, vol. 2430, 2019, pp. 15-28.

[15] Y. Wang, “Definition and Categorization of Dew Computing,” OJCC, vol. 3(1), 2016.

[16] G.A. Oparin, V.G. Bogdanova, and A.A. Pashinin, “Automation of Microservices Creation for Qualitative Analysis of Binary Dynamic Systems,” in Proceedings of the 1st International Workshop on Information, Computation, and Control Systems for Distributed Environments, ICCS-DE 2019, CEUR-WS Proceedings, vol. 2430, 2019, pp. 88-98.

[17] D.C. Tsichritzis and F.H. Lochovsky, Data Models. Prentice Hall Professional Technical Reference, 1982.

[18] “Irkutsk Supercomputer Center of SB RAS,” http://hpc.icc.ru/ (accessed: 24-Jan-2020).

[19] https://firstvds.ru/ (accessed: 24-Jan-2020).

[20] I.V. Bychkov, G.A. Oparin, V.G. Bogdanova, and A.A. Pashinin, “Service-oriented technology for development and application of decentralized multiagent solvers for applied problems,” Herald of computer and information technologies, vol. 12. pp. 36-44, 2018. [In Russian].

[21] E. Dubrova, “Self-Organization for Fault-Tolerance,” In Self-Organizing Systems, K.A. Hummel and J.P.G. Sterbenz, Eds. IWSOS 2008, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, vol. 5343, 2008, pp. 145-157.

[22] S.A. Kauffman, “The Origins of Order: Self-Organization and Selection of Evolution”. Oxford University Press, Oxford, 1993.

[23] O. Reyes, R. C. Laubenbacher, and B. Pareigis, “Boolean Monomial Dynamical Systems,” in Annals of Combinatorics, vol. 8, 2004, pp.425-439.

[24] E. Dubrova and M. Teslenko, “A SAT-Based Algorithm for Finding Short Cycles in Shift Register Based Stream Ciphers,” in IACR Cryptology ePrint Archive, 2016. https://eprint.iacr.org/2016/1068 [online, accessed: 24-Jan-2020].

Impact of Dew Computing on Cyber-Physical Systems and IoT

M. Gusev
Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering, Skopje, North Macedonia
E-mail: marjan.gushev@finki.ukim.mk

***Abstract*—Cyber-physical systems and IoT devices are part of today's interconnected networked world and participate in huge data exchange transfer. Owners of related devices and systems tend to use them independently of external systems and in addition, to be autonomous devices and systems that can perform their tasks without any external support. Edge computing refers to devices on the edge of the Internet, as they integrate into a more complex cyber-physical system. The concept of edge computing means that the corresponding cloud and edge servers are always connected to the edge devices and IoT. In the analyzed case, although these devices and systems can work in an integrated environment, they are rather classified into the dew computing concept, giving them an environment for autonomous performance, independent of surrounding devices and systems. This paper elaborates the underlying concept, benefits, market trends and impact of integrating the dew computing concept to the cyber-physical systems and IoT devices.**

Index Terms*—*Edge computing*; *post cloud acrhitectures*; *CPSoS

I. Introduction

Cyber-physical systems (CPS) are automated systems that operate with humans in the physical reality with computing and communication infrastructures (cyber world) [1], [2]. The underlying distributed processing in a network world integrates interdisciplinary knowledge to establish the perception of the physical world and realize cognition and control using cyber systems. Organizing a CPS as a system of systems defines a cyber-physical system of systems (CPSoS), referring to connectivity for better monitoring, management, and control of the overall systems [3].

Connecting various devices (everything) on the Internet creates the Internet of Things envisioned as a global network of machines and devices capable of interacting with each other [4]. The kernel functionality (computation, communication, and data management) of CPSoS and IoT devices has been typically provided by post-cloud technologies [5] based on Edge/Fog computing models. Although, cloud computing servers were associated with the IoT devices to enable the realization of more complex systems, information exchange, and user interfaces, the concept of *edge computing* [6] implementations is mostly referred as an implementation design via *fog computing* [7], [8] and *cloudlets* [9], [10].

Recently, research activities, also tend to serverless architectures where the processing is shared among edge devices instead of using only the servers [11], or even towards the deviceless approach [12].

However, these approaches do not solve the CPSoS resilience problem initiated when a smaller CPS will stop performing in case of a drop of connectivity to a more complex CPSoS, which may lead to severe consequences. Imagine an autonomous car when driving on a road where the white marking lines are erased and not visible. Should an autonomous car stop or continue to drive analyzing where is the road using its own algorithms. In addition to this, the problem of consistent and highly productive design of new CPSoS-s is another obstacle,

Dew computing [13], [14] is represented as an additional layer between end-user devices, such as smart modules that convert physical parameters into digital information and vice-versa, and processing and coordination in the Edge/Fog/Cloud layers, offering autonomy, independence, and collaboration features. The dew computing concept can be applied successfully on IoT [15].

The main research trends in CPS and IoT devices has set focus on developing new methods that result with high-confidence cyber systems compatible, synergic and integrated at all scales to the physical requirements. In this paper, we aim at presenting the impact of applying the dew computing concept to the design of CPS of systems and IoT devices.

The rest of the paper is organized as follows. Section II presents the computing and communication aspects of trending CSPoS requirements and how these will be matched with the Dew computing concept. Market trends are elaborated in Section IV with details on the market conditions, market size, growth rate, trends, market barriers and obstacles. Section V discusses the application domains, and benefits of integration of the dew computing concept into CPSoS and IoT solutions, along with a special focus on the relevant socio-economic impact, and technology drivers. Finally, relevant conclusions and future work is specified in Section VI.

II. Computing and communication architecture

Industry 4.0 refers to full process and value chain automation through digitalization and inter-connectivity, including the IoT devices, CPSoS, clouds, and artificial intelligence [16]. Cyber-physical systems integrate physical processes with computation, and can be found around us, including cars, homes, and machines from one side and humans from other side. Their interaction is of primary interest when several complex systems interact among themselves. In addition, physical

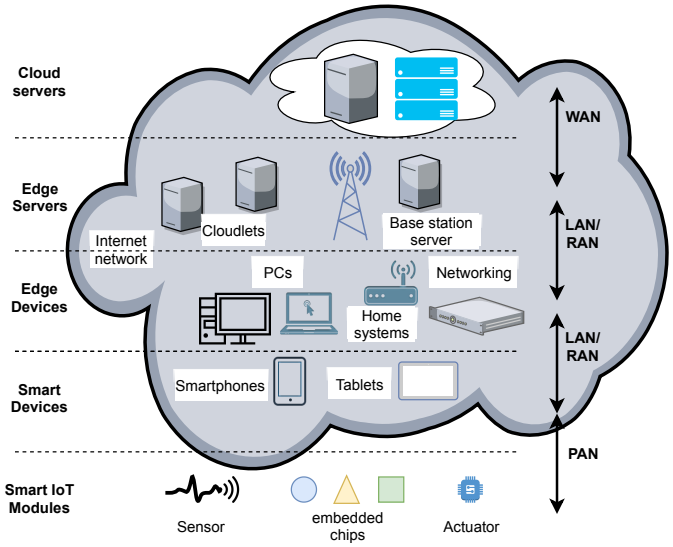


Fig. 1. Architecture of dew computing solutions for CPS and IoT [17]

safety of people can be jeopardized by use of transport or medical systems.

Fig. 1 [17] presents the architectural layers, where the dew computing concept is integrated in the implementation of CPS and IoT.

A typical dew computing architecture defines a clear path from devices to service including several layers:

- smart devices (sensors and multiservice gateways)
- smart dew devices (autonomous physical or computing devices)
- IoT oriented edge framework (edge devices with network connectivity);
- IoT oriented cloud integration platform (edge and cloud servers);
- a service platform (applications and services);
- an intelligence system with business logic.

The corresponding communication between layers is defined by appropriate technologies: smart devices are usually battery-operated devices and use personal area networking (PAN), such as Bluetooth or similar radio frequency technology; smart dew devices are also battery-operated wirelessly connected devices and can use mobile operator networks 3G/4G, WiFI or similar radio technologies; IoT edge devices can use local area network (LAN); edge and cloud servers use LAN and wide area network (WAN). The service platform and intelligence systems communicate using web service protocols.

A complimentary paper [17] gives a more detailed presentation of dew computing architecture for CPSoS and IoT, comparing it to other approaches and architectural designs, including a prototype CPS architecture [18], holistic view [19], five-layer CPS architecture for Industry 4.0 [20], networked mathematical model [21], three-layered (user, cyber and physical) system [22], fog-computing solution [23], adding a hybrid cross architectural layer [24], transparent computing-based IoT architecture [25], etc.

The analyzed architectural models [17] are based on various criteria to build an architecture, such as analyzing the functional requirements, software complexity, computing requirements, user location and perspective, application requirements, and communication approach, among others.

III. Technology Trends and Drivers

A. Human user-related and computing challenges

CPSoS and IoT devices design and implementation face the issues related to the networked control systems, including the event-driven computing, software architecture, communication protocols, variable time responses, and delays, failures and reconfiguration, and distributed architecture approach to support sufficient real-time quality-of-service. From one side, the requirements to distribute the control, processing, and data storage to the lowest architectural level pushes increased complexity, and from the other side, these end-user devices need to be energy efficient and support wireless communication, mobility and low energy consumption to be battery-operated devices.

A lot of design architectures and software models have been applied to trade-off high reliability and quality of service over complexity, and close the gap between the robust system and distributed approach. The implementation of the dew computing concept challenges the speed, reliability, and efficiency as most exposed limitations of today's computing systems.

The dew computing concept implementation in CPSoS and IoT devices targets the human users of complex CPSoS and wearable IoT systems. Identified human user challenges addressed with the dew computing approach are presented in Table I along with a summary of computing requirements for smart devices, including IoT and wireless edge devices.

B. Technology drivers

The following primary drivers are identified as primary for CPSoS and IoT:

- Sensing (transformation from physical to digital, including wireless sensing networks, sensors based on chemical and physical characteristics),
- Computing (processing on the edge, big data analytics, high-performance and distributed computing, etc.)
- Connectivity (communication technologies and protocols, and other IoT and similar radio networking)
- Dependable computing and communication systems (interoperability to exchange data with other systems, networking and transport control protocols, providing fault tolerance, substitution of broken parts, takeover of functions for increased workload, etc.)
- Management (including communication and exploitation policy, strategy to overcome barriers and associated risks, maintenance, etc.)
- Performance and security (real-time response, reliability, data protection, privacy, network security, etc.)

TABLE I
HUMAN USER-RELATED AND COMPUTING CHALLENGES FOR THE DEW COMPUTING CONCEPT IN CPSOS AND IOT

Challenge	Description	Identified needs
Environment	Pollution increased by energy consumption	Dew computing concept consumes energy as little as possible by distributing the processing closer to where data is generated, eliminating a lot of data transfer and redundant energy consumption.
Performance	Small (real-time) response time	Applications need to respond as fast as possible (ex. autonomous car driving; reaction on health threatening arrhythmia)
Resistance	Solutions foreseeing the obstacles and related issues	Surrounding hardware and software systems generate various obstacles that need to be resolved by corresponding applications
Autonomy	Independence of Internet or other surrounding systems	The essence of the dew computing concept is based on the performance of smart devices that can still continue to perform if the other surrounding systems do not exchange valuable information
Data analytics	Need to process large quantities of data	Data transfer is optimized and computing is distributed at the level of end-user devices
Efficient management	Controlling the processing and data transfer	Distributed management closer to the end-user by autonomous devices is more reliable and efficient than centralized control in servers and task delegation and monitoring of CPSoS and IoT devices
Energy efficiency	Need to operate for more than a few days without recharging	Larger batteries are not solution, the smart devices need to be wearable, light and energy supply-independent devices, which requires completely new architectures, technology and software design
Security	Protect information from cyber attacks	Distributing the processing on the lowest level of devices closer to the source is more resistant on cyber attacks than in systems that transfer data to other more powerful systems and servers
Software complexity	Managing interaction between different software layers	Processing at levels closer to the source and transferring only results to higher architectural levels needs less complex synchronization between different software layers developed by different teams

Sensors and networked systems generate huge data volumes with high speed, and, thus, require big data processing capabilities. In addition, this is enhanced by the management of distributed networked devices to reach the state of an integrated system that is intelligent, resilient and self-adaptable machine.

Networking among various devices and CPS parts, including Internet connectivity is a huge potential for vulnerability to technical systems with safety consequences and damages. Therefore, data privacy, protection, and network security are important major concerns and challenges in the design of CPSoS and IoT devices.

TABLE II
GROWING MARKET

Technology	Product	Size	CAGR
IoT/CPS	IoT sensors	18 B	5.5%
40 Billion	IoT connectivity	12.5 B	11.5%
semiconductors	Processors	7 B	14.5%
Computing	IaaS services	21%	
cloud	PaaS services	28%	
services	SaaS services	39%	
	other	12%	
Connectivity	Industrial	5.5 B	25%
30 Billion	Consumer	6 B	14%
devices	Communications	17 B	8%
	Computers	2 B	-2%
	Transportation	1 B	21%
	Healthcare	0.5 B	20%

IV. MARKET DRIVERS

Our market research shows that this technology has a great potential. It will increase the use of the existing market, and mainly attract all those who have experience with building CPSoS solutions. CPS/IoT market is far from being a stable reality, it shows huge revenues in some domains, few revenues in others, customers don't know what CPS/IoT is, what are the advantages, they are scared and don't trust technology, platforms are not mature, there is no interoperability. The reality does not satisfy the expectations.

A. Market size

Various sources that predict 1 trillion connected IoT devices [26] resulted in large overestimation. The later predictions in 2016 [27] are more realistic setting the target of 20-30 billion connected devices (IoT and smartphones) for 2020, or 400 Mil with cellular connections [26].

New data from Juniper Research has revealed that the number of IoT (Internet of Things) connected devices is expected to be 38.5 billion in 2020, up from 13.4 billion in 2015: a rise of over 285% [28].

The number of devices connected to the Internet, including the machines, sensors, and cameras that make up the Internet of Things (IoT), continues to grow at a steady pace [29]. with estimates that there will be 41.6 billion connected IoT devices, or "things," generating 79.4 zettabytes (ZB) of data in 2025.

The *global key players* [30] include IBM Corporation, Microsoft Corporation, Dell EMC, VMware Inc., Hewlett Packard Enterprise, Schneider Electric, SAP SE, Honeywell International Inc., Hitachi Vantara, Oracle.

The market can be segmented by *application areas* (aerospace and defense, automotive, energy and utility, healthcare, manufacturing, consumer electronics etc.) and the market *product types* are usually classified as IoT sensors, connectivity, server hardware, cloud platforms, server software and services.

B. Market trends and growth

Given that most developed nations already use CPSoS, it is clear that the market is growing with distribution presented in Table II.

TABLE III
MARKET BARRIERS

Barrier	Typical case	Solution
Market and research gap	Companies are still far away to match the CPS/IoT expectations	Research has left too many open issues behind, and there is need for more pragmatic research to
Absence of IoT ecosystem	Stakeholders cooperate to CPS/IoT innovation and market development.	Support by private initiatives, policymaking, complementary expertise, infrastructure, funding, academic networks, creation of value network, and develop new interoperable technologies and platforms
Lack of trust	Trust the security, safety, integrity and privacy of the massive transformation that CPS/IoT is generating and will generate.	Key acceptance barrier to overcome by effective marketing and communication to raise awareness, professional training, in several areas: privacy, information security, cybersecurity and security
Lack of interoperability	Key element to inherently control diversity and avoid fragmentation, and is essential to support CPS cooperation and SoS existence	Diversity means richness and added-value; diversity is an indicator of innovation; but fragmentation is a CPS/IoT enemy.
Lack of solid platforms	A CPS/IoT ecosystem requires a secure and efficient platform capable to orchestrate and manage it for the entire lifecycle.	Provide a core set of functionalities: remote man., fleet man., provisioning, device abstraction, APIs; Create the value network, being a reference architecture that allows the integration of SoS, across multiple companies, firms, technologies, standards, vertical domains; Promote CPS/IoT adoption, specifically for starters.

A nice overview of market trends and growth for IoT and CPS is elaborated in [31].

The total installed base of Internet of Things (IoT) connected devices is projected to amount to 75.44 billion worldwide by 2025, a fivefold increase in ten years [32].

It is expected that the commercial, industrial and consumer sectors will account for nearly 50% of new connected devices up to 2030 ([33]). Expectations for 2030 address commercial and industrial electronics to grow up to 5.4Billion devices with 24.4% CAGR, then consumer devices up to 5.9 Billion with 13.8% CAGR, communication devices up to 16.8 Billion with 8% CAGR etc.

Recently, a study has revealed that in 2018 mobile phones are surpassed in numbers by IoT devices, which include connected cars, machines, meters, wearables and other consumer' electronics; and expectations are that between 2016 and 2022, IoT devices will increase at a CAGR of 21 percent, driven by new use cases [26].

C. Market barriers

Table III identifies several barriers and ways to overcome, such as regulatory and legal barriers.

Table IV illustrates some of the identified gaps. These are challenges of the industry to find marketing and sales solutions and persuade customers for use of available technology that will make their lives easier.

TABLE IV
MARKET GAPS

Gaps	Description
Market readiness	Customers still discuss if their future will be with microcontrollers or microprocessors; researchers talk about micro-services, digital twins, embedding artificial intelligence
Lack of trust	Customers don't trust the distributed nature of IoT, especially the security and data protection
Business unawareness	Customers still focus on the cost of the device, not on operative and maintenance costs
Inadequate solutions for connectivity	Adoption of LTE and related technologies slowed down, the mobile operators already stopped further sim activation for IoT devices due to terrorism attacks; The future of 5G is still uncertain and the actual 4G network is overloaded; M1 and Narrowband IoT still don't take off
Customer expectations are not met	Customer requirements are higher and different than the offered technology and solutions developed by research

V. DISCUSSION

The dew computing concept brings the computing closer to the user and/or end-user device. The supported engineering techniques of dynamic CPSoS exploit emerging technologies such as Big Data and machine learning, especially for those dependable CPSoS that do not have a permanent Internet connection and need to perform autonomously.

A. Impact

Several stakeholders are involved in the exploitation of CPSoS and IoT solutions, including the users; system, software and service providers; and companies with complementary competences and businesses. The stakeholders are integrated both vertically and horizontally, coordinated flexibly to create and use a product.

Companies that design CPSoS and IoT solutions implemented a business model starting from their closer environment (direct suppliers and clients) and including available products or services.

The *vertical integration* addresses the IT systems of all companies from logistics, resource planning and usage to production and sales, assuming that the interaction is performed seamlessly through a standardized architecture. On the other side, the *horizontal integration* integrates multiple verticals and connected processes beyond the boundaries of one single entity.

The complexity of CPS/IoT pushes companies to bring all competencies under a single umbrella, providing an open platform where a multiple number of vendors will mix their roles and responsibilities with customers as designers of the new CPS and IoT solutions leaving a lot of space for emerging new service and solution providers.

An overview of application domains of dew computing in CPSoS and IoT is elaborated in a complementary paper [17], including energy efficiency, autonomy, independence, collaboration, interoperability, elasticity, resilience, etc.

TABLE V
SOCIO-ECONOMIC BENEFITS OF IMPLEMENTING THE DEW COMPUTING CONCEPT FOR CPS AND IOT SOLUTIONS

Benefit	Description
Productivity	Effort spent from the design to the development, testing and deployment of the CPSoS and IoT devices is distributed to different vendors which increases the design and development efforts and decreases the delivery time
Open platform	Availability of a multiple vendors to produce various CPS and IoT devices integrating by a unified post-cloud architecure
Business aspects	New system integrating multiple vendors open new opportunities and can improve sales and end-user experience as an environment with focus on vertical business logic

B. Benefits

Even a single sensor, could become a high-level service, either offering an autonomous function (as a dew device) or work as a connected device in the IoT world. These services become the components of a newly added value application and system.

The main expected technology benefits of implementing the dew computing concept for CPS and IoT solutions are:

- Systems can continue performing autonomously and independently even without an Internet connection and awareness of other neighboring or surrounding systems
- Processing is distributed with real-time response and alerting
- Systems will consume less energy and be energy efficient
- Systems will be affordable for a wide range of customers

The introduction of new services and applications implementing the dew computing concept for CPSoS and IoT devices result with socio-economic benefits for users and society et al., as specified in Table V.

An analysis of technology-based benefits of implementing dew computing solutions in CPSoS and IoT can be found in [17].

VI. CONCLUSION

The impact of dew computing on CPSoS and IoT was analyzed in this paper.

Computing and communication solutions have been advancing towards Industry 4.0 and enable the introduction of new concepts in realization of CPS and IoT. The dew computing concept enables the end-user devices with functions and properties for "always connected" devices although they are not. This means that the edge devices can communicate with smart IoT devices via dew servers and dew devices even when there is no Internet or other network connection.

Realizing software for those devices to become smart devices even without connectivity means that they become autonomous, independent and still collaborative devices, which represents the dew computing essence. Their impact is huge, not measured only in technology benefits, but, also in a large number of socio-economic benefits. This paper contains an analysis of market drivers, market size, trends and growth, barriers and identified gaps that need solutions to overcome to enable greater penetration of dew computing solutions.

REFERENCES

[1] E. A. Lee, "Cyber physical systems: Design challenges," in *2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC)*. IEEE, 2008, pp. 363–369.

[2] R. Baheti and H. Gill, "Cyber-physical systems," *The impact of control technology*, vol. 12, no. 1, pp. 161–166, 2011.

[3] S. Engell, "Cyber-physical systems of systems–definition and core research and development areas," *CPSoS*, 2014.

[4] I. Lee and K. Lee, "The internet of things (iot): Applications, investments, and challenges for enterprises," *Business Horizons*, vol. 58, no. 4, pp. 431–440, 2015.

[5] Y. Zhou, D. Zhang, and N. Xiong, "Post-cloud computing paradigms: a survey and comparison," *Tsinghua Science and Technology*, vol. 22, no. 6, pp. 714–732, 2017.

[6] M. Satyanarayanan, P. Simoens, Y. Xiao, P. Pillai, Z. Chen, K. Ha, W. Hu, and B. Amos, "Edge analytics in the internet of things," *IEEE Pervasive Computing*, vol. 14, no. 2, pp. 24–31, 2015.

[7] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*. ACM, 2012, pp. 13–16.

[8] I. Stojmenovic and S. Wen, "The fog computing paradigm: Scenarios and security issues," in *2014 Federated Conference on Computer Science and Information Systems*. IEEE, 2014, pp. 1–8.

[9] M. Satyanarayanan, V. Bahl, R. Caceres, and N. Davies, "The case for vm-based cloudlets in mobile computing," *IEEE pervasive Computing*, 2009.

[10] A. Bahtovski and M. Gusev, "Cloudlet challenges," *Procedia Engineering*, vol. 69, pp. 704–711, 2014.

[11] S. Nastic, T. Rausch, O. Scekic, S. Dustdar, M. Gusev, B. Koteska, M. Kostoska, B. Jakimovski, S. Ristov, and R. Prodan, "A serverless real-time data analytics platform for edge computing," *IEEE Internet Computing*, vol. 21, no. 4, pp. 64–71, 2017.

[12] M. Gusev, B. Koteska, M. Kostoska, B. Jakimovski, S. Dustdar, O. Scekic, T. Rausch, S. Nastic, S. Ristov, and T. Fahringer, "A deviceless edge computing approach for streaming iot applications," *IEEE Internet Computing*, vol. 23, no. 1, pp. 37–45, 2019.

[13] K. Skala, D. Davidovic, E. Afgan, I. Sovic, and Z. Sojat, "Scalable distributed computing hierarchy: Cloud, fog and dew computing," *Open Journal of Cloud Computing (OJCC)*, vol. 2, no. 1, pp. 16–24, 2015.

[14] Y. Wang, "Definition and categorization of dew computing," *Open Journal of Cloud Computing (OJCC)*, vol. 3, no. 1, pp. 1–7, 2016.

[15] M. Gusev, "A dew computing solution for IoT streaming devices," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2017, pp. 387–392.

[16] Y. Lu, "Industry 4.0: A survey on technologies, applications and open research issues," *Journal of Industrial Information Integration*, vol. 6, pp. 1–10, 2017.

[17] M. Gusev, "Dew computing architecture for cyber-physical systems and IoT," *Journal Internet of Things*, vol. 11, no. 100186, pp. 1–8, 2020.

[18] Y. Tan, S. Goddard, and L. C. Pérez, "A prototype architecture for cyber-physical systems," *ACM Sigbed Review*, vol. 5, no. 1, p. 26, 2008.

[19] V. Gunes, S. Peter, T. Givargis, and F. Vahid, "A survey on concepts, applications, and challenges in cyber-physical systems." *KSII Transactions on Internet & Information Systems*, vol. 8, no. 12, 2014.

[20] J. Lee, B. Bagheri, and H.-A. Kao, "A cyber-physical systems architecture for industry 4.0-based manufacturing systems," *Manufacturing letters*, vol. 3, pp. 18–23, 2015.

[21] K.-D. Kim and P. R. Kumar, "Cyber–physical systems: A perspective at the centennial," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1287–1308, 2012.

[22] Y. Liu, Y. Peng, B. Wang, S. Yao, and Z. Liu, "Review on cyber-physical systems," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 1, pp. 27–40, 2017.

[23] A. V. Dastjerdi and R. Buyya, "Fog computing: Helping the internet of things realize its potential," *Computer*, vol. 49, no. 8, pp. 112–116, 2016.

[24] M. Frincu, "Architecting a hybrid cross layer dew-fog-cloud stack for future data-driven cyber-physical systems," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2017, pp. 399–403.

[25] J. Ren, H. Guo, C. Xu, and Y. Zhang, "Serving at the edge: A scalable iot architecture based on transparent computing," *IEEE Network*, vol. 31, no. 5, pp. 96–105, 2017.
[26] Ericcson, "Internet of things forecast," 2019, accessed: 2019-12-15. [Online]. Available: https://www.ericsson.com/en/mobility-report/internet-of-things-forecast
[27] Mark Hung, "Leading the IoT," Gartner Research, 2019, accessed: 2019-04-15. [Online]. Available: https://www.gartner.com/imagesrv/books/iot/iotEbook_digital.pdf
[28] Juniper, "Internet of things connected devices to almost triple to over 38 billion units by 2020," 2019, accessed: 2019-12-15. [Online]. Available: https://www.juniperresearch.com/press/press-releases/iot-connected-devices-to-triple-to-38-bn-by-2020
[29] IDC, "The growth in connected IoT devices is expected to generate 79.4zb of data in 2025," 2019, accessed: 2019-12-15. [Online]. Available: https://ihsmarkit.com/products/consumer-electronics-market.html
[30] MarketWatch, "Internet of things (iot) market 2019 - global industry analysis, size, share, growth, trends, key players and forecast 2019 - 2025," 2019, accessed: 2019-12-15. [Online]. Available: https://www.marketwatch.com/press-release
[31] Paolo Azzoni, "Cps market trends," Eurotech, 2019, accessed: 2019-04-15. [Online]. Available: http://www.cpsschool.eu/wp-content/uploads/2018/09/CPS-Market-Trends.for-PDF.pdf
[32] Statista Research Department, "Internet of things (iot) connected devices installed base worldwide from 2015 to 2025," 2019, accessed: 2019-12-15. [Online]. Available: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/
[33] IHS Markit, "Mobile, consumer & connected devices market research and analysis," 2019, accessed: 2019-12-15. [Online]. Available: https://ihsmarkit.com/products/consumer-electronics-market.html

The Rainbow through the Lens of Dew

Zorislav Šojat*, Karolj Skala*
* Centre for Informatics and Computing, Ruđer Bošković Institute, Zagreb, Croatia
sojat@irb.hr, skala@irb.hr

Abstract - **Ecological Responsibility of Computer Science is essential in further development of the science and application of computer technology.**

The rapid development of all kinds of computer controlled things and mobile devices, without and with Internet connectivity, represents presently the most prominent information technology application area. On the lowest level, below the Edge of Internet, we have a huge amount of individual devices ("things") which do actually the most significant amount of information processing. Those are computer controlled utensils of all kinds – cars, boilers, production machines, traffic lights, plant carers etc., directly intertwined with our natural and human physical environment and our immediate surroundings. These devices, which are neither at the Cloud or Fog edge, nor even at the mobile edge, are the basis of the Dew Computing Paradigm. Further exploration and usage of the realm of possibilities of Dew Computing will solve the basic problems of integration of the "dew" level with the higher service level(s) in the Dew-Fog-Cloud hierarchy. When the distributed Cloud-Dew service architecture is developed and adopted, a Cloud server and many Dew servers will be able to cooperate as a distributed application to provide controls or services.

In view of that it is absolutely necessary to see the present (and specifically future) development through the eyes of Philosophy of Computing and Ethics, and further the development through application of Ethical and Ecological principles.

The Rainbow Ecosystem is a pragmatical hierarchical abstract cybernetic representation of an ecologically viable system, consisting of 9 essential areas, and applicable recursively throughout the integrated hierarchy. The ecological viability of the abstract system will enable proper ecologically responsible development of Rainbow Computing Services, integrating our living Nature-Human-Machine ecosystem.

Keywords: Ecological Responsibility of Computer Science, Philosophy of Computing, Ethics, Ecology, Viable System, Requisite Variety, Rainbow Ecosystem, Rainbow-Computing Services, Dew-Computing, Fog-Computing, Cloud-Computing

I. A Viable System

"A viable system is a system capable of independent existence", but "no ... being is absolutely independent" (Stafford Beer, [1]). Therefore a viable system is a system capable of surviving in its own environment, as self-contained, self-sufficient, self-supporting and self-reliant as necessary to survive in that environment. An ecosystem is a dynamic system of all its sub-systems. As it is a dynamic system (or cybernetically speaking a machine), it includes all internal communication channels and inter-influences, the inter-relationships of the individual sub-system behavioural patterns and unexpected non-patterned comportment. However it does not include its own environment, but is viable only in the sense of continuation of existence (survival) in constant self-organisation with its environment (the external part of the supra-system). Therefore both the sub-systems and the supra-system are essential for its survival (viability). Furthermore, for it to be a viable system it has to be recursively viable throughout, i.e. all of its sub-systems have to be recursively viable, too [2].

There are many approaches towards modeling a viable system, depending on the main aim of the model usage. As presently we have to take into account the enormous destruction humankind has done to its own living environment (well, the Earth will survive with or without us), a model of a viable system applicable to the area of Ecology, and than, consequently, to the Responsibility of Computer Science, is to be used. Regarding the present-day paradigmatic hierarchisation of computing into a three-layered architectural system, Cloud Computing, Fog Computing and Dew Computing [3][4], all of these three layers have to (recursively) be viable. However, it is essential to understand that even now, and much more so in the future, Computer Science is being responsible for a huge area of everyday human living, and for an enormous amount of ecological problems, both by making them (like energy consumption, not only by the use of computing equipment, but even more by its extremely short life-cycle)[1], and by solving them (which must be the prime target for the future)[2].

As obviously it is not any more possible to develop Computer Science without the regard to the whole social, political, economical and environmental spectrum of aspects, the Rainbow Ecosystem model of a recursive viable system is proposed [5]. This model includes all three main aspects of our Global Ecosystem – the Nature, the Humans and the Machines, as well as the necessary coordinating aspects.

II. Requisite Variety

The prerequisite for a system to be able to be in a self-organising dynamic equilibrium, is that the amount of communicatable state transitions necessary for inter-

1 Infrared energy.

2 The Orange creativity passed through the Yellow appropriateness filter.

control of the two sub-systems making up the self-organisation is great enough, and that each of the sub-systems has a necessary possibility of variation of behaviour. Consequently there is a necessity for variety. "If a system is to be able to deal successfully with the diversity of challenges that its environment produces, then it needs to have a repertoire of responses which is (at least) as nuanced as the problems thrown up by the environment" [6]. This is known as the Law of Requisite Variety by William Ross Ashby [7]. What it means is that the number of system states (that is the system's variety) necessary for control of the other system, i.e. its environment (and vice-versa), must be sufficiently great. Regarding from the level of one sub-system, it must be the same or greater than the variety of the other sub-system. Obviously, for proper self-organisation of two sub-systems, the requisite variety of both has to be the same. Additional variety is not requisite.

As a consequence of this Law of Requisite Variety, it is clear that in the future a unified Rainbow Global Service Ecosystem, as part of the Global Rainbow Ecosystem, will have to be based on providing the requisite variety for all of its sub-systems. In the Global Rainbow Ecosystem model, these sub-systems include primarily the Nature, the Humans and the Machines. To be able to cope with this, both horizontally and vertically exponentially growing variety of possible regulation needs and responses, it will not be possible to continue with present-day focus on Data and Data Processing. Such integration necessitates the use of Information and the development of a consistent Ontology, or, said in other words, the development of a semantically and syntactically correct language system.

III. The Dew-Fog-Cloud Hierarchy

The development of computing equipment started with the use of quite huge and, from our modern viewpoint, quite slow programmable processors, all of them already Turing complete. During the years huge scientific advances in electronics, followed by appropriate technological advances in production, led to the use of more and more computing (or better to say data processing) devices. Supercomputers turned into Clusters and Grids, personal computers turned into handheld mobile devices, and microcomputers turned into micro-controllers of all imaginable "things", down to menial tasks of controlling triacs for lighting and valves for irrigation. But in Clusters and Grids, the major workhorses, the individual computers are generally extremely homogeneous on the level of hardware and software architecture. The necessity of working with heterogeneity, specifically on the level of user necessities, led to the development of Clouds. The transformation of personal computers into a variety of devices, from desktops, laptops, tablets to mobile communicators (not to call them mobile phones, as they are mostly not primarily used as telephones), added a huge amount of heterogeneity on all levels, from hardware to usage scenarios, and forced the Computer Science to introduce the paradigmatic level of Fog Computing. And the complex world at and below the Edge of Internet necessitates the introduction of the lowest level architectural principles, the level of the Dew [8][9].

The Dew Computing[3] Paradigm is primarily oriented towards the physical-edge, human environment control devices i.e. lighting, traffic control, heating, cooling, energy distribution etc., where human control of the environment must take precedence over possible higher-level requests, or at least be coordinated with them, without disturbing the immediate human environment. This is the reason why in Dew-Computing there are two basic notions, which do not exist in the rest of the hierarchy, in the Fog and in the Cloud: self-sufficiency and co-operation [10] [11].

The Fog-Computing[4] Paradigm aspires to enable the use of a collaborative multitude of end-user clients or near-user edge devices (e.g. also the Mobile-Edge) to significantly enhance and expand the communication, computing and data resources available, with special attention to the proximity envelope of the end-users and of their specific objectives.

The Cloud-Computing[5] Paradigm aspires to enable a ubiquitous and convenient on-demand network access to a shared pool of configurable computing and data resources [12].

IV. The Rainbow

As a viable system model, the Rainbow includes 9 levels, named after the colours of the light spectrum. Those colours represent, as mentioned, pragmatical areas of Philosophy of Computing, and include all three aspects of present and future human life (and endeavours): Nature, Humans and Machines. Therefore not every colour is applicable to all three of those aspects. The viability of the system based on this model comes from a huge variety of elements and actors, and is internally intertwined. This allows different colour areas to act as both controlled and controlling fields. The generic meanings of the Rainbow colours are as follows [13][5]:

Ultraviolet – Visions: *The area of Wisdom, Prudence, Conscience, Responsibility, Holism.* The necessary driver for our human survival and advancement.

Violet – Interference: *Security, Limits of Expansion, Nature-Human-Machine Interference.* A very wide, very important area of further exploration and proper tackling.

Indigo – Cooperation: *Ethics, Information Use, Redundancy, Knowledge Gathering and Preservation, Behaviour.* The future of our civilisation is extremely dependent on these, on the level of individuals, societies, politics, economy and global ecology.

[3] On a global level Dew Computing comes under Green, as it is primarily connected with the human and natural environment.

[4] On the global level Fog Computing is under Blue, as its main area is communication.

[5] Globally the Cloud Computing layer is in the domain of Indigo, the area of cooperation.

Blue – Communication: *Information, Knowledge, Interaction, Languages.* The requisite base for the requisite variety. A future integration of the global system, including the Nature, Humans and Machines is impossible without huge advancements in consistent ontological development.

Green – Nature: *Environment, Health, Well-Being, Backup Systems, Global Ecosystem.* The essential area of our survival.

Yellow – Appropriateness Filter: *Consequences, Reasoning.* What consequences each of our actions (ideas, developments, technology...) have on the Global Ecosystem. Every vision, idea, creation, product or process has to pass this Filter. This is in the viable system model the highest contol element.

Orange – Creativity: *Stimulation, Ideas, Education.* Without education and stimulation of creativity we will not be able to survive in the natural environment that we forced to change so much. However, the education has to be primarily oriented towards principles and ethics.

Red – Matter: *Resources, Products.* We have to organise resources and products in such a way that we achieve sustainability of the Global Ecosystem. Presently we just throw them away.

Infrared – Energy: *Efficiency, Garbage, Quality.* Throwing away non-renewable energy, and producing ever more garbage, partly due to the lack of quality of products (of all types) costs irreparable damage to the Global Ecosystem.

V. Dew and the Rainbow Services

In the possibility of developing integrated home systems which do all the management and maintenance and keep us entertained, developing self-organising traffic-control systems and intelligent driver suggestion systems, coordinated building/car/traffic pollution control systems, real-time hospital systems with all patient and equipment status and control collaborating with the medical staff, fully consistent synesthetic artistic performances including artists and independent individuals ("active public") from wide apart, power distribution peek filtering, self-reorganisation and mutual cooperation systems based on informed behaviour of individual power consumption elements, etc., the Dew-Computing paradigm allows the development of the Distributed Information Services Environment, and finally the present civilisation's aim of establishment of a Global Service Processing Environment, Rainbow service [13].

The Rainbow service, seamlessly connecting all computing hierarchy layers, will have to be established through horizontal self-organising coordination of individual Dew-droplets (i.e. the dew devices with communication/self-organisation possibilities) and vertical Fog/Cloud integration based on the newest developments of the Fog computing paradigm. The collaboration of Dew-droplets shall primarily be established on the level of their information exchange and self-organising coordination by (intra-) networking them, using proper ontologies/protocols and security/privacy and QoS based channels [3]. By having a common ontological basis on all hierarchical levels, i.e. throughout the "rainbow" of computing, and by developing common interaction protocols and information/instruction languages, such self-coordinating and collaborating systems of individual Dew devices will then, as micro services, be seamlessly integrated higher up into the computing hierarchy through Internet gateways into the Fog and Cloud, enabling the coordination of the physical edge control through common strategies and multi-level intelligent behaviour, including adaptive-learning, predicting and other aspects of cognitive computing, with multi-modal user interfaces to facilitate appropriate Human/Computing interaction [14].

The merits of seamlessly integrating those "dew" devices into the Dew-Fog-Cloud computing hierarchy in form of Rainbow services are enormous, for individuals, the public and industrial sectors, the scientific community and the commercial sector, by bettering the physical and communication, as well as the intellectual, immediate human environment.

VI. The Lens of Dew

Infrared – Energy

Dew Computing is generally much intersecting with the Infrared area. On one hand it is the individual energy consumption of many Dew devices that is a big concern for their sustainability, as much of them may not be constantly connected to the electricity grid, and many of them are battery powered [15]. This concern necessitates specific Orange approaches in the Red hardware area, as also in the Indigo software and Blue communication areas.

On the other hand we have a growing amount of garbage generation of all kinds of Dew (and Fog and Cloud!) equipment, specifically from the area of mobile devices. This problem will necessitate a stabilisation and standardisation of equipment generations, reuse of older devices and Orange education of general public on the necessity of long-term use of all types of machines. And finally, we have the Indigo ethical problem of Quality, which has to be solved by changing the Red production principles by installing norms of Yellow appropriateness of life-cycle, and therefore push the producers to change their Indigo behaviour in the direction of taking global Ultraviolet responsibility.

Red – Matter

In the area of hardware development much has to be done with specific architectures for specific applications, with special attention given to the Infrared energy efficiency. Proper Orange education on basic working principles will be necessary for the development of novel solutions based on the Blue knowledge. The Dew devices may profit from alternative Orange developments, specifically tuned towards the Yellow appropriateness for the task.

Another aspect of Red through the lens of Dew is the possibility of certain Dew devices to directly influence the Green material world, as, for example, in the case of water supply control, entrance/exit/window... control, production control etc.

Orange – Creativity

The ever growing global environmental disruptions caused by our misuse of Infrared and Red areas, without the understanding of the cybernetic principles of Violet limits of expansion necessitate a huge effort in creativity and appropriate education, with the main aim of understanding and Yellow balancing the global Ecology and local Economy.

Introduction of Dew devices into our everyday living environment and our wider Rainbow natural habitat may lead to outstanding positive results, but may also lead to Yellow consequences we would never like to experience.

Yellow – Appropriateness Filter

Yes, the consequences. To understand the consequences of each and every action, development, introduction of technics and technology, of resources, principles, products, services etc. is almost impossible. Yet by applying the appropriateness filter with highest Ultraviolet responsibility, based on Indigo ethics, Blue knowledge, Violet limits and Green ecology, we have to aim towards a civilisation which is more appropriate for the Green well-being of our planetary habitat. Many Dew devices may easily pass this filter, but many Rainbow Services some Orange idea imagines to be based on those devices may easily not pass this Yellow filter.

Green – Nature

Due to the fact that Dew Computing is directly connected to the human and natural environment, Green is the colour where much of Dew devices will have to be developed and organised. They will deal with our environment and the global ecosystem, with our and nature's health and well-being. For all those areas it is essential in development of Red hardware and Indigo software to pass the Yellow appropriateness filter using Indigo ethics, and specifically taking into account possible Violet interference problems [security, (sub-)systems interference] and the Violet limits of growth.

Another very important aspect of Violet security are unavoidable situations where, due to any reason, like Green natural disaster, Ultraviolet political upheaval, Infrared, Red and/or Indigo technical problems etc., the computing support of our environmental, social, political and economic systems is broken (not-working). Therefore it is absolutely essential to provide backup systems which do not necessitate modern or complex technology[6]. It must be noted that this life-saving aspect of Violet Security is mostly never tackled in Computer Science. Only systems providing, specifically on the level of Dew and Fog, full non-technological (or low-technological) system backup solutions can be regarded as complete and pass the Yellow appropriateness filter for a viable system.

And finally the Green area includes the balancing and preservation of the Global Ecosystem, in which all three computing architectural levels, the Dew, the Fog and the Cloud must be seamlessly thoroughly interconnected in all aspects of the Rainbow nature-human-machine ecosystem.

Blue – Communication

The whole Rainbow world, actually the whole Universe, is based primarily on communication in its widest sense – this communication can be on all levels, from Infrared physical (like repulsing of electrons), Red material (like moving objects), Orange mind-internal (like ideas), Yellow reasonable (like thought experiments), Green temporal (like streams), recursively lower level Blue communicating (like language or sign communication, or, one Blue level deeper, data communication), over Indigo cooperating (like transfer synchronisation), dangerously Violet interfering (like the so-called "noise"), all the way up to Ultraviolet conscientious.

On the level of all three architectural computing layers the communication necessitates a common frame of reference, and this will have to be developed as an ontology encompassing the whole Rainbow nature-human-machine ecosystem. The introduction of the exponentially growing heterogeneity of present and future Dew (and partially Fog) device architectures and their enormous area of use in all aspects of human endeavours and knowledge application, necessitates the introduction of the next higher recursive level of Blue – the use of ontologically, semantically and syntactically properly defined and dynamically expandable Information Communication. This will allow a seamless integration of nature, humans and computers, by sharing a consistent linguistic approach. For that to be achieved, a wider aspect of the viable Rainbow system has to be employed, by Indigo cooperation of a wast area of different sciences.

The basic Dew Computing notion of independence and self-reliance means that communicationally the Dew devices must be able to be interconnected by means of very different Indigo protocols and Infrared/Red communication means. This than also necessitates the above-mentioned information and not data exchange and processing. These processes are also directly connected with the need for Rainbow requisite variety.

[6] Like, for example, providing alternative Red material means of payment (e.g. paper money supply without computerised records), or, as another example, providing alternative, old technology based communication and dissemination means (classical radio, in extreme cases even middle or short wave, independently wired telephone exchanges etc.)

INDIGO – COOPERATION

The second basic Dew Computing notion, in addition to the Blue independence, is cooperation. This will be enabled by solving the Blue problems, but computer science, regarding the integration of the Rainbow system, is still, on all three layers, and specifically at the Dew layer, in the phase of Ultraviolet visionary exploration and Orange creativity and new ideas. However, for a system to be actually viable, the interchange between all areas of the Rainbow viable system has to be with as short Violet latency as possible. Therefore, even during the Ultraviolet and Orange phases the Yellow appropriateness filter has to be constantly employed.

To be able to develop a viable human civilisation, we will have to develop a global viable Rainbow Indigo cooperation between computer science and all other Rainbow sciences, techniques and technologies, and the complete available human Rainbow Blue knowledge. This will necessitate an important push in Orange educational efforts towards undestanding the Raibow ecosystem, which, per se, necessitates the prime understanding of ethics and the Ultraviolet wisdom, prudence, conscience, responsibility and holism. It will also be necessary to Orange teach the basic Yellow reasoning principles and procedures to a vast population of present and future Human Earthlings.

VIOLET – INTERFERENCE

At the level of Dew all kinds of interferences can be expected, from all Colours of the Rainbow system. These are quite different from interferences in the Fog and Cloud layers, as the viable Dew sub-system has to have in each Dew Droplet the requisite variety to cope with its own environment, and with the Rainbow Services System Indigo co-operation.

The interference between the nature, humans and machines we already experience, and is exponentially growing. Here the interference is happening on all levels, from the misunderstandings between Rainbow users and Indigo programmes, to the interference of Infrared garbage and energy consumption, the Red natural resources[7] usage, etc., with the Rainbow Earth's ecosystem.

Therefore the security, not only, as already stated in Green, the security of Blue information, but also the security of having Green non- or low-technological backup systems must be consistently developed.

And, finally, with all our Ultraviolet visions of a "better" future in constant development and expansion, and all our marvelous Orange ideas, it is essential that we do not forget the limits of expansion. In cybernetics it is a well known fact that any system with positive feedback is unstable, and will finally break (stop being viable) at a certain point of individual limit of expansion. And that is presently the state of our civilisation aims and developments, primarily driven by the economy, and by wishful Ultraviolet thinking and Orange thinking out..., and not by the ecology and Ultraviolet responsibility and Orange solutions to problems. However, it has to be mentioned that the fully opposite, a system with only negative feedback greater than its positive effort will also break down, in non-operation (thus stop being viable). Therefore the Yellow balancing of positive and negative feedback is absolutely essential in the Rainbow global ecosystem.

ULTRAVIOLET – VISIONS

The visions are something that drives us to be alive. Any effort we do starts with a vision, and then it is channeled into other Colour areas. However, visions without the Orange imagination with the guidance of Yellow reason, and without the use of prudence and wisdom, being conscious as to what Yellow consequences any vision can have on the Rainbow ecosystem, and being responsible towards the Green well-being and survival (i.e. not destruction) of our Rainbow civilisation on our Rainbow planet Earth, as human Earthlings, those visions are extremely Violet dangerous.

VII. THE DEW THROUGH THE LENS OF RAINBOW

Now, looking down on Dew from the perspective of the Rainbow, we can see that there is an enormous amount of present-day civilisation problems that can be solved at that level. The use of Dew devices in direct cooperation with our human and natural environment can help in many areas of security, health, traffic, well-being, food production, resource management, industrial production and handicrafts, sub-systems stabilisation, lowering of energy consumption and balancing, through the Rainbow Services Ecosystem, the local necessities and products in the local economic ecosystem with, by balancing the local and global production and supply chains, the global ecological ecosystem. From the view of the Rainbow, it is absolutely obvious that our civilisation will have to rearrange its economic habitus, and insist on Local Economy and Global Ecology.

And finally, the ray of Ultraviolet light screems that Computer Science will have to take on itself the heavy, but honourable burden of Ecological Responsibility.

REFERENCES

[1] Stafford Beer, "The Intelligent Organization", https://www.youtube.com/watch?v=7COX-b3HK50 [14/6/2020, 17:19].

[2] Stafford Beer, "Fanfare for effective freedom – Cybernetic Praxis in Government", Third Richard Goodman Memorial Lecture, Brighton 14/2/1973, https://www.scribd.com/document/2310055/Fanfare-for-Effective-Freedom-Cybernetic-Praxis-in-Government-Beer [24/5/2020, 21:08].

[3] Karolj Skala, Davor Davidović, Enis Afgan, Ivan Sović, Zorislav Šojat: "Scalable distributed computing hierarchy: cloud, fog and

[7] Natural resources as a whole are actually Rainbow natural resurces. The usage comes under Red usage, and the mis-usage under Yellow unreasonable, Violet heavily interfering and Indigo unethical.

dew computing", Open Journal of Cloud Computing (OJCC), 2 (1). pp. 16-24. ISSN 2199-1987, 2016.

[4] Pooja Kukreja, Deepti Sharma: "A Detail Review on Cloud, Fog and Dew Computing", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 5, Issue 5, 2016.

[5] Zorislav Šojat, "From Dew Over Cloud Towards the Rainbow - Ecosystem of the Future: Nature — Human — Machine", Intelligence in Big Data Technologies - Beyond the Hype, Chapter 1, Advs in Intelligent Syst., Computing, Vol. 1167, Springer Nature, Singapore, 2020, ISBN 978-981-15-5284-7.

[6] John Naughton, "Ashby's Law of Requisite Variety", Edge, 2017, https://www.edge.org/response-detail/27150 [15/6/2020, 7:04].

[7] W. Ross Ashby, "An Introduction to Cybernetics", Chapman & Hall, London, 1956, Internet (1999), http://pcp.vub.ac.be/books/IntroCyb.pdf [15/6/2020, 7:01].

[8] Karolj Skala, Zorislav Šojat: "Cloud, Fog and Dew Computing: A Distributed Hierarchy", IRB CIR, March 2015, DOI: 10.13140/RG.2.2.26021.52963.

[9] Y. Wang, "Cloud-dew architecture," International Journal of Cloud Computing, vol. 4, no. 3, pp. 199–210, 2015.

[10] Pooja Kukreja, Deepti Sharma: "A Detail Review on Cloud, Fog and Dew Computing", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 5, Issue 5, 1412, 2016.

[11] Zorislav Šojat, Karolj Skala, "The dawn of Dew: Dew Computing for advanced living environment", MIPRO 2017, pp. 375-380, DOI: 10.23919/MIPRO.2017.7973447.

[12] "Cloud Computing Definition", ssh.org, https://www.ssh.com/cloud/computing/definition [27/8/2020, 2:05]).

[13] Karolj Skala, Zorislav Šojat, "The Rainbow Global Service Ecosystem", Proceedings of the 3rd International Workshop on Dew Computing (DewCom 2018), IEEE DewCom STC, Toronto, Canada, 2018, pp. 25-30.

[14] Zorislav Šojat, Karolj Skala, "Views on the Role and Importance of Dew Computing in the Service and Control Technology", MIPRO 2016, pp. 175-179, DOI: 10.1109/MIPRO.2016.7522131.

[15] Matías Hirsch, Cristian Mateos, Alejandro Zunino: "Practical Criteria for Scheduling CPU-Bound Jobs in Mobile Devices at the Edge", 2018 IEEE International Conference on Cloud Engineering (IC2E), Orlando, FL, 2018, pp. 340-345, DOI: 10.1109/IC2E.2018.00065.

A Disaster-Resilient Messaging Protocol Based on Dew Computing

Yingwei Wang
School of Mathematical and Computational Sciences
University of Prince Edward Island
Charlottetown, Canada
Email: ywang@upei.ca

***Abstract*—Communication is vitally important during disasters. This paper proposes a new messaging protocol that is resilient to the damaged network conditions caused by disasters based on the principles of dew computing. The goal of this messaging protocol is to make messengers adapt to both normal conditions and disastrous conditions. In a normal condition, the client-server architecture supports efficient communications; in a disastrous condition, the messengers will make full use of P2P and other possible communication links so that vital information can be transmitted whenever it is possible.**

***Index Terms*—Communication; Disaster resilient; Dew computing; Cloud-dew architecture; Client-server; Peer-to-peer; P2P.**

I. Introduction

Communication is vitally important when a disaster happens, and much research has been performed regarding to communication systems in disastrous conditions [1][2][3][4][5]. Most of these research covers satellite communications, controller area networks, communications within rescue teams, and policy making. Here, we would like to discuss this topic from the angle of social media communications.

Various forms of social media play important roles in modern society. They are the communication tools people use most often. An instant messenger is the basic form of social media communication tools, and is the focus of our discussion.

People are familiar with messengers that are running with an assumption that network connections are available. In a disastrous condition, however, network connections may be unavailable or limited.

Here we list the scenarios of network connections in a disastrous condition. We call them network connection levels.

1) The Internet works well.
2) The server is not available, but the Internet is still available.
3) The Internet is not available, but Local Area Network (LAN) connections are still available.
4) The Internet and LAN connections are not available, but battery-operated mobile devices are still able to provide Personal Area Network (PAN) or Near-Field Communication (NFC) connections.

When a disaster happens, the condition of network connections could deteriorate quickly; communications get more and more difficult. The network connection level could change from level 1 to higher levels, even reach level 3 or level 4. The major goal of the proposed messaging protocol is that the new messengers can provide some communication support in all these network connection levels by using all possible connections.

The proposed messaging protocol should be a simple and lightweight one so that it can be widely used.

With such goals in mind, we would like to discuss the proposed messaging protocol from various angles. The rest of the paper is organized as follows: Section II describes the architecture of the proposed protocol; Section III introduces all the auxiliary data links and their roles in the proposed protocol; Section IV introduces a few existing open messaging protocols and their features; Section V elaborates the considerations and decisions in the design of the proposed protocol; Section VI gives the conclusions.

II. Architecture: Integration of Client-server and P2P

First, we need to decide the architecture of this proposed protocol. There are two architectures could be considered: one is the widely-used client-server architecture, and the other is the peer-to-peer (P2P) architecture.

In the client-server architecture, tasks are divided between the providers of a resource or service, called servers, and service consumers, called clients. The client-server architecture is a very efficient architecture where servers provide resources and services so that clients can be relatively simple. The biggest problem for the client-server architecture is that the server is the single point of failure. Once the server is down or unavailable, the system cannot work at all.

A P2P network is designed around the principle of equal peer nodes simultaneously functioning as both "clients" and "servers" to the other nodes on the network. P2P does not rely on a server or any single point of failure, but each client may have to take a relatively heavier duty in terms of maintaining communications. P2P mainly concerns the application level. In the network level, MANET (Mobile Ad-hoc NETwork) and other similar frameworks provide detailed technical support. Both P2P in the application level and MANET in the network lever are faced with the same fundamental challenge, that is, to provide connectivity in a decentralized, dynamic environment.

In terms of a messaging system, these two architectures both have advantages and disadvantages.

The great features of the client-server architecture are that the server manages various information such as accounts, groups; provides services such as message relaying and mutual isolation. Overall speaking, the most efficient messaging system is the one based on the client-server architecture. The problem of using the client-server architecture is that the server may be not available in a disastrous condition.

The biggest advantage of the P2P architecture is that the system does not have a single point of failure. Thus, P2P is very suitable in a disastrous condition. The problem of the P2P architecture is that it puts heavy tasks to each client to keep multiple connections, to maintain routing information, and so on.

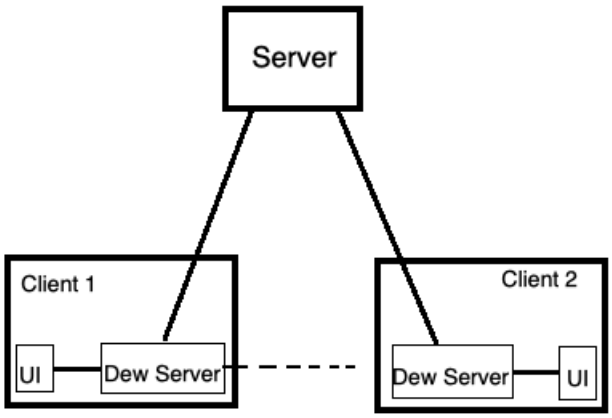
Considering the requirements of a messenger in a disastrous condition and the features of the client-server architecture and the P2P architecture, we propose to combine client-server and P2P using the cloud-dew architecture [6].

Our idea is that the messaging system works in the mode of client-server in normal conditions; when the server is not available during a disaster or similar conditions, the clients will automatically connect to each other and the system will work in the mode of P2P.

For this purpose, we would like to introduce a server in the messenger client. (This sentence might be confusing: the messenger client is an application, it is a client to the messenger server. The introduced server is just a component of this messenger client.) In dew computing, this server is called a dew server. In the messaging system, a proper name, such as a communication server or a mini router, should be given to make it more meaningful and less confusing. Comparing with the main server in the messaging system, a dew server only provides limited services and it is easier to disappear, as the name suggests.

Fig. 1 describes the architecture of the proposed new messenger protocol.

Fig. 1. Messenger Architecture



The whole messaging system works in the client-server mode if clients can connect to the server. The dashed line shows that clients connect to each other only when the server is not available. Each client consists of two components: the UI (user interface) and the dew server. This architecture was inspired by the cloud-dew architecture [6]. This architecture satisfies the key features of dew computing [7]: independence (the dew sever can provide connections to other clients even the server is not available) and collaboration (the dew server will connect to the server to realize the potential of the client-server mode when the server is available).

In Fig. 1, a client is divided into two components: the UI and the dew server. The user communicates with the UI. The UI communicates with the dew server. From the viewpoint of the user or the UI, as long as a message is sent to the dew server, it is already sent; the rest will be the work of the dew server. The dew server will try to communicate with the server or the other clients. UI only works when the user needs it, but the dew server works all the time. The dew server only deals with message transmission, it does not deal with privacy and encryption; these issues should be dealt with by the UI portion of the client.

In a normal condition, a dew server does not need to keep connections with other clients; in a disastrous condition, a dew server should keep connections with other related clients and also provides services to any other clients, even this client is not related to them. The UI only serves this user, but the dew server will serve both this user and other clients. The is vitally important because people (clients) should provide help to others in a disastrous condition.

In the client-server mode, the server will maintain accounts information, establish connections among clients. In the P2P mode, we suppose it is in an emergency condition so that the system does not need to provide complete administrative services. In this way, the clients can be kept as light-weighted.

III. Auxiliary Links

In a normal condition (the network connection level is 1), Internet connections rely on major links such as cellular data connections, cable connections, fiber connections, satellite connections. In a disastrous condition, major links could be broken due to damages to the wire or equipment; available network connections could be deteriorate quickly. If the network connection level is 2, client-server systems will stop working, but P2P systems will still work; if the network connection level is 3 or 4, P2P systems will also stop working. As described in Section I, even when the network connection level is 3 or 4, we still hope the messenger be able to transmit information to near-by messenger clients. To reach these goals, we have to make full use of all the possible auxiliary links (LAN, PAN, or NFC).

It is reasonable to assume that power is available in the disastrous areas: besides survived domestic power supplies, UPS units and battery-operated mobile devices also make it possible for some auxiliary data links to be in working conditions. In the following, we would like to discuss some auxiliary data links that could be available during a disaster.

A. LAN Links

When the Internet connections are not available but local WiFi is available (the network connection level is 3), the proposed messenger could still play a vital role in such a disastrous condition. For example, a family or a group of people were inside a building with WiFi when a disaster happened suddenly. The Internet became not available but the WiFi is still working. A person sends a message to a person or a group through the proposed messenger. Although the targeted person or the targeted group of people were not inside this building and there is no links that can reach them, the message could be transmitted to the mobile devices of some people inside this building. Later, people could be scattered and went to different directions. Some people may reach an area where Internet connections are available. Then the messages will be transmitted to the targeted people or the targeted group. The person who brought the messages out may or may not have the knowledge that he or she has brought the messages out.

B. PAN Links

As described above, WiFi links could be very useful in disastrous conditions, but it relies on power supplies. If WiFi is not available (the network connection level is 4), people have to face even harsh conditions. Another kind of possible links are Personal Area Network (PAN) links. Bluetooth is the major form of PAN connections that is widely available in mobile devices. In most cases, class 2 Bluetooth will cover 5 to 10 meters.

Other forms of PAN links are also possible, such as Infrared Data Association (IrDA) and Wireless USB, but we may not need to use them because Bluetooth is already widely available in most mobile devices.

C. NFC Links

Near-Field Communication (NFC) refers to a contact-less communication technology between two electronic devices over a distance of 4 cm or less. When we tap a credit card, we are using a NFC link. NFC technology is perfectly designed to exchange data between two devices through a simple touch gesture. The messages exchanged through NFC should conform with the NFC Data Exchange Format (NDEF).

NFC link is especially useful when the two parties trust each other; they can use a simple touch gesture to confirm such trust. Thus, NFC links should be incorporated in the proposed messaging protocol.

IV. Messaging Protocols

There exist quite a few instant messaging protocols. Some of them are open protocols. For example, Extensible Messaging and Presence Protocol (XMPP), Internet Relay Chat (IRC), Matrix, Tox, Bitmessage, and so on. We believe that the proposed protocol should not be designed completely from scratch. Instead, it should be designed on top of these existing protocols or make full of the features and mechanisms of these existing protocols.

In this section, we would like to introduce two existing messaging protocol, Matrix and Tox, and discuss their features and mechanisms.

A. Matrix Protocol

Matrix [8] is an open instant massaging protocol. It is designed to allow users with accounts at one communication service provider to communicate with users of a different service provider. We use the email system as an example to illustrate this idea: There are many email service providers: Hotmail, Gmail, Outlook, and so on; we can also set up our own email server to provide our own email service. As a user, we may obtain an account in one of these email services and communicate with any one who has an account in any email service. Such communication is possible because of the Simple Mail Transfer Protocol (SMTP). Matrix is a similar protocol in the area of instant messaging.

In Matrix, a server is called a homeserver. There are many homeservers in the Matrix system. Each homeserver can be considered as a messaging service provider. A user needs to obtain an account from one homeserver. Homeservers connect to each other so that messages can be communicated among different homeservers. To perform communication, one user needs to join a “room”. Room data is replicated across all of the homeservers whose users are participating in a given room. As such, no single homeserver has control or ownership over a given room. The process of synchronising shared conversation history between homeservers run by different parties is called “Federation”.

Matrix also supports bridging messages from different chat applications into Matrix rooms. These bridges are programs that run on the server and communicate with the non-Matrix servers. Bridges can either be acting as puppets or relays, where in the former the individual users account is visibly posting the messages, and in the latter a bot posts the messages for non puppeteered user accounts. Currently there are official bridges for Gitter, Slack, and so on. Bridges for other applications are also available, such as Apple iMessage, Discord, Facebook Messenger, Google Hangouts, Skype, Telegram, Whatsapp, and so on.

Matrix also has some other great features, such as voice chat, video chat, privacy, and end-to-end encryption. Matrix has a RESTful API which can be accessed by clients.

B. Tox Protocol

Tox [9] is an open P2P instant messaging protocol that offers end-to-end encryption with the feature of Perfect Forward Secrecy (PFS) [10].

To implement P2P messaging, each client needs to find the location (IP address) for each of its friends. In Tox, the ID of each client is its public key. Tox maintains a Distributed Hash Table (DHT) [11] which can be used to search the ID of the targeted Tox client to find its IP address. A measure of distance is defined in DHT thus the search will get closer to the target. A data structure called k-buckets was introduced to facilitate such search.

Tox does not rely on a server or any single point of failure. However, the lack of central servers has the consequence that both parties of the chat need to be online for the message to be sent and received.

PFS is a good feature in terms of encryption. With this feature, each session has its unique session key; if a single session key is compromised, it will not affect any data other than that exchanged in the specific session protected by that particular key.

To implement end-to-end encryption with the feature of PFS, it is necessary for Tox clients to change their encryption keys at the beginning of each session. The challenge is how to distribute keys securely through unsecured connections. Tox uses elliptic curve Diffie–Hellman (ECDH) [12] key agreement scheme to distribute keys securely among clients. ECDH is able to make sure both parties of the session get a new encryption key, but no one else can obtain this key even all the communication is monitored.

Although Tox provides end-to-end encryption, it does not hide the client's IP address when communicating with friends, as the whole point of a P2P network is to connect the client directly to the client's friends.

V. The Architecture and Mechanisms of the Proposed Protocol

For convenience purposes, we temporarily use Matrix-Mesh to refer the proposed protocol. In this section, we would like to introduce Matrix-Mesh's architecture and its major mechanisms.

A. Matrix-Mesh Architecture

A Matrix-Mesh client should behave in the following ways:

1) It communicates with its server when the network connection level is 1.
2) It communicates with other clients in the P2P mode when the network connection level is 2.
3) It communicates with other clients through auxiliary links when the network connection level is 3 or 4 or the user chooses to do so.

Matrix-Mesh generally follows the specification of Matrix protocol with some extra features.

Matrix-Mesh clients should be able to connect to each other in the P2P mode when their homeserver is not available.

The architecture of Matrix-Mesh is depicted in Fig. 1. A Matrix-Mesh server should be similar to a Matrix homeserver. A Matrix-Mesh client should contain two components: the UI and the dew server. Here we temporarily use the term dew server; later it could be called communication server, mini router, or some other names.

All the features and mechanisms of a Matrix client should be placed into the UI portion or the dew server portion: UI provides all the functions of the instant messenger and takes care of message encryption and privacy; the dew server takes care of the server connection, P2P connections, auxiliary link connections, DHT maintenance, and so on.

To reduce the complexity and power consumption of the messaging client, the dew server connects with the server using TCP, and connect to other clients using UDP; UDP packets will only be sent out when the server is not available. The goal of the dew server is similar to the homeserver: keeping and syncing all the rooms this client uses; but the structure of the dew server is different from the homeserver: it does not need a domain name; it works using UDP instead of TCP.

B. P2P Message Relay

Tox requires that both parties of the chat need to be online for the message to be sent and received. This feature is not acceptable in Matrix-Mesh. In a disastrous condition, major data links could be broken and auxiliary links could be limited, but we still hope messages can be transmitted to other clients nearby so that these messages could have chances to reach their destination with some delay. To achieve this goal, some clients, neither the sender nor the receiver, may need to take the responsibilities to relay these messages.

C. The Establishment of P2P Connections

The primary challenge in P2P communication is equipping each device to continuously maintain the information required to properly route traffic. In Tox, each client dynamically maintains and updates its DHT to keep connections with other related clients. This is a relatively heavy burden to the client. If we keep this mechanism in Matrix-Mesh, the client dew server has to connect with the server and also maintain the DHT, the whole client will be a heavy client and won't be very practical.

Because in normal conditions each client works in the client-server mode, the client only connects to the server instead of other clients. It is not necessary for each client to keep connections with other clients. But when the server suddenly becomes unavailable, each client needs to have some information about the clients they are related to so that it can connect to them through P2P connections.

How can each client keep up with the dynamic environment so that it can connect to other related clients when necessary? In this regard, the client-server architecture could provide some assistance. We might need to modify the Matrix protocol regarding to the communication between a client and the homeserver so that the server sends to a client some minimal information about this client's related clients. In other words, the client-server architecture provides support to the establishment of P2P connections among clients. This mechanism reflects the architectural integration between client-server and P2P.

D. The Development of Matrix-Mesh Protocol

Matrix-Mesh is a new instant messaging protocol which is under development. We plan to design Matrix-Mesh on the basis of existing open instant messaging protocols. The architecture of Matrix-Mesh will be similar to the Matrix protocol, some components of the Tox protocol and other protocols will also be integrated into the Matrix-Mesh protocol. We will

develop the protocol specification and implement the protocol by providing a reference server and a reference client. The reference server and reference client will be used as Proof of Concept (PoC) and the basis of further development. The protocol specification, the source code of the reference server and reference client will also be provided to public as an open source project.

We welcome anyone who is interested in the development of the Matrix-Mesh project to contact the author of this paper to join the project.

VI. Conclusions

Matrix-Mesh is a proposed open instant messaging protocol. Its goal is to provide an instant messenger system that can satisfy normal messaging activities and can also accomplish vital communications in a disastrous condition. The idea of Matrix-Mesh protocol was inspired by the principles of dew computing, especially inspired by the cloud-dew architecture. Matrix-Mesh is based on many existing open instant messaging protocols, such as Matrix and Tox. Matrix-Mesh itself will also be open to the public.

References

[1] Kai Tang. (2018, Jun.) Reliable communication key in times of disaster. [Online]. Available: https://psc.apcointl.org/2018/06/19/reliable-communication-key-in-times-of-disaster/

[2] L. M. Pinho and F. Vasques, "Timing analysis of reliable real-time communication in can networks," in *Proceedings 13th Euromicro Conference on Real-Time Systems*, 2001, pp. 103–114.

[3] H.-C. Jang, Y.-N. Lien, and T.-C. Tsai, "Rescue information system for earthquake disasters based on manet emergency communication platform," in *Proceedings of the 2009 International Conference on Wireless Communications and Mobile Computing: Connecting the World Wirelessly*, ser. IWCMC '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 623–627. [Online]. Available: https://doi.org/10.1145/1582379.1582514

[4] S. Yodmani and D. Hollister, "Disasters and communication technology: Perspectives from asia," in *the Second Tampere Conference on Disaster Communications*, 01 2001.

[5] Justin Veuthey. (2016, Jul.) Disasters and mobile phones in china. [Online]. Available: https://www.asiapacific.ca/blog/disasters-and-mobile-phones-china

[6] Y. Wang, "Cloud-dew architecture," *International Journal of Cloud Computing*, vol. 4, no. 3, pp. 199–210, 2015.

[7] Yingwei Wang, "Definition and categorization of dew computing," *Open Journal of Cloud Computing (OJCC)*, vol. 3, no. 1, pp. 1–7, 2016.

[8] The Matrix.org Foundation. (2020, 07) [matrix]. [Online]. Available: https://matrix.org/

[9] Tox contributors. (2020, 07) Tox: A new kind of instant messaging. [Online]. Available: https://tox.chat

[10] G. C.G., "An identity-based key-exchange protocol," in *Advances in Cryptology — EUROCRYPT '89. EUROCRYPT 1989. Lecture Notes in Computer Science*, Q. JJ. and V. J., Eds. Berlin, Heidelberg: Springer, 1990, vol. 434.

[11] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, "Chord: A scalable peer-to-peer lookup service for internet applications," *SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 4, p. 149–160, Aug. 2001. [Online]. Available: https://doi.org/10.1145/964723.383071

[12] E. Barker, L. Chen, A. Roginsky, A. Vassilev, and R. Davis. (2018, 04) Recommendation for pair-wise key-establishment schemes using discrete logarithm cryptography. [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-56a/rev-3/final

Ensuring Resource Availability with MRU/FRU Caching: A Dew-Blockcloud Model

1st Chukwuka Chukwuocha
dept. of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada.
chukwuoc@myumanitoba.ca

2nd Ruppa K. Thulasiram
dept. of Computer Science)
University of Manitoba
Winnipeg, Manitoba, Canada.
tulsi@cs.umanitoba.ca

3rd Parimala Thulasiraman
dept. of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada.
thulasir@cs.umanitoba.ca

4thYingwei Wang
dept. of Computer Science
University of Prince Edward Island
Charlottetown, PEI, Canada.
ywang@upei.ca

***Abstract*—Combining the blockchain and cloud computing systems in different ways in order to address computing issues such as decentralization and security has slowly become a norm in recent times. These Block-Cloud model(s), which are becoming prominent in the literature seem to ensure integrity in the Cloud but lacks in exhibiting availability of resources, which is essential for security.**

In this study, a dew-blockcloud model is designed to address this problem. We propose a hybrid MRU/FRU caching algorithm for the dew server that reduces the number of repeated cloud accesses while maintaining relevant updated data.

Using existing literature and postulation we show that dew computing might be the best availability system for the blockcloud model.

***Index Terms*—Blockcloud, Decentralized approach, Resource availability, Dew computing, MRU & FRU caching.**

I. Introduction and Motivation

It is generally argued that data might be one of the most valuable resources in the current era of information technology. Hence, data security has become the driving force of the development of most of the sectors of our society. Breach of data security can cripple companies and sometimes even nations. An example of this is the infamous Facebook Cambridge Analytica data scandal of 2018 when the giant social media company was accused of leaking its users data to Russia who used it for a targeted social engineering attack in the 2016 USA election [1]. Many of the USA security agencies still believe that this attack from Russia influenced the election results and have subsequently led to some of the current problems the country is facing. Regardless of the validity of these claims, one thing is clear, securing data has become one of the most important challenges of the 21st century. Confidentiality, Integrity and Availability (CIA) are regarded as the three major aspects of data security.

The widespread use of cloud computing in recent years has propagated the use of required resources by individuals and companies when needed without owning the infrastructure. While this provided access of certain resources to people who cannot afford it, it also meant that these cloud users had to deligate their data to cloud vendors. These cloud vendors required their customers to trust them with their data while promising security of this data through several sophisticated methods. However, the Facebook Cambridge Analytica data scandal [1] mentioned previously is a clear indication that cloud vendors as well as other big companies cannot be trusted completely because there is always a possibility of customer data misuse.

Eliminating the need to trust cloud vendors could be one of the most important applications of blockchain due to its decentralized nature [2]. A lot of research has gone into integrating the blockchain into the cloud computing environment and it has proved to be difficult because of the intensive nature of the blockchain. This has spawned several blockcloud architectures in which the developers find techniques to circumvent the computational intensiveness of the blockchain.

Gaetani et al. [3] developed a blockchain database for cloud computing environments that focused on ensuring data integrity in these environments. This model was intended for the European SUNFISH project which was proposing a new service FAAS (Federation as a service) in order to increase data security in various cloud systems [4]. The model contained two blockchain layers, the first being a permissioned lightweight layer that supports fast computation and data access. The second layer is more of a typical blockchain database (bitcoin blockchain) that receives the hashed data state of the first layer at intervals for permanent irrefutable storage. Gaetani et al. model [3] eliminated the need for trust while ensuring integrity in the cloud environment. Data stored in the second layer, which is similar to a bitcoin blockchain is tamper proof, hence integrity is implicit. However, in order to provide integrity for the first layer which lacked the proof of work consensus (PoW) that gives a blockchain the tamperproof

Natural Sciences and Engineering Research Council (NSERC) Canada

feature [2], Gaetani created a system where data updates and access in this layer requires the authentication key of every member of the federated cloud. The PoW was eliminated from the first layer to make it computationally fast.

While this model successfully ensured integrity, it was lacking in another important security property called availability. This is because in order to access or update data in the cloud blockchain database, the signature of all the participants (of the federation) are needed.

The focus of this research is to extend Gaetani's model to ensure availability without compromising its existing security properties. The contributions of this work includes the (1) Design of a dew-block architecture that ensures resource availability (2) Proposal of a hybrid MRU/FRU caching algorithm for the dew server and (3) Analysis of the proposed architecture, its benefits, drawbacks and possible modifications.

The rest of this paper is arranged as follows. In section 2, we present the related work and review Gaetani et al. model. In section 3, we explain our proposed architecture. In section 4, we analyze this architecture and possible implementation methods. In section 5, we conclude our study and present possible areas for future work.

II. Background and Related Work

The need for users to be able to perform computing operations by requesting for the needed resources without worrying about owning the hardware have always been a major research focus. Several attempts have been made to solve this need such as time-sharing utilities like the IBM 704 [5] and grid computing systems.

The advent of cloud computing in the early 2000's and the developmental stages it has gone through since then has facilitated a world of on-demand computing where several scalable and distributed resources such as storage and computing power are provided to a user on demand in a pay-as-you-go pricing format [6]. The bedrock of cloud computing is virtualization, which entails the logical separation of different resources [7]. It is used to simulate logical representations of physical resources through the hypervisor.

Ensuring security in a cloud computing environment has proven to be a major challenge in furthering its development. The nature of cloud computing entails that a user has to give up control of his data or programs to a cloud provider to carry out certain operations. This implies that the users should have absolute trust in the cloud provider to ensure security of their data. This could prove detrimental to the user because a cloud provider could become a security threat with ease due to its complete control over the user's data. Hence, studies on how to circumvent the concept of trust in cloud security are some of the more prevalent research areas in cloud computing.

The sections below include research work being done on the current methods of providing cloud security, Blockchain as a means of avoiding trust in cloud security and resource availability in cloud computing models.

A. Current Methods of Cloud security

Confidentiality, Integrity and Availability (CIA) are the three main attributes of cloud computing security [8]. Confidentiality refers to enabling only eligible entities to access protected data. It encompasses a wide range of security concerns such as (a) multitenancy [9], which concerns when cloud users share resources; (b) data remanence, which may lead to data divulgence through scavenging of residual data by other users; and lastly, privacy [10]. Integrity refers to preventing data alteration or expulsion by unauthorized parties.

The cloud providers usually control the provision of these security features and a prominent example is the amazon AWS cloud which uses encryption with transport layer security [11], firewalls, authorization keys and persistent storage volumes to ensure security. Most of other cloud providers mirror these techniques or have their own variations. By establishing a trusted third party Zissis and Lekkas [12] proposed a method of solving the security issues associated with cloud computing environments. Since then several security issues like multitenancy have been combated by the trusted third party. One of the tools of this trusted third party includes the public key infrastructure which issues certificates for authentication and does cryptographic separation to ensure data confidentiality on a single machine that serves many clients [12].

Techniques like sanitation or degaussing that are used to solve data remanence problem in cloud servers are also handled by this trusted third parties [13]. The service level agreement (SLA) contains information on if trusted third parties are used for external auditing and other security practices [14].

B. Blockchain as a means of avoiding trust in cloud security

The blockchain is basically a trustless system [2], which makes it a perfect method of circumventing the need for trust in cloud security. This is because the client knows every single change his data undergoes and is needed for the verification of any transaction that occurs in the system. Its architecture as a public distributed ledger without a central authority has led to its adoption as a cloud storage option in many recent research projects on cloud computing and internet of things.

In 2017 Liang et al. [15] discussed the use of blockchain to ensure data security for IOT cloud-based applications using a Cloud based drone system to demonstrate their research work. They called this the dronechain and the blockchain network used here was more of a secondary storage unit for cloud auditing and to guarantee data security by periodical data validation. The original data was still sent to the primary cloud storage while the hashed data was sent to the blockchain network. Cloud auditing refers to verifying data validity and it has been historically done by trusted third parties [16]. The use of blockchain simplifies this by eliminating the need for an external auditor and hence saving cost and increasing assurance in the system.

Sharma et al. [17] used fog nodes in a distributed cloud blockchain based architecture. Fog computing is one of the new computing paradigms that is focused on bringing computing resources to the edge of the IOT network [18]. The

proximity of the fog nodes to the IoT devices enable the provision of faster cloud services as well as low latency. Sharma et al. [17] used the blockchain architecture for the controller network in the fog node as well as the distributed cloud architecture. This guaranteed resilience, fault tolerance, availability and complete security on both the fog layer and the cloud layer.

Ming et al. [19] proposed a service-centric blockchain architecture, which gave mobility, scalability, security and fairness of service distribution in IoT devices [19]. Service-centric networking is a proposed evolution of the internet from being host centric with a never-ending connection to being focused on services [20]. Applications on this proposed paradigm communicate with service names and not with IP addresses. This architecture is meant to improve scalability while circumventing the need for an internet connection to obtain certain services [19].

Most of the research work we have seen so far has to do with IoT cloud enabled devices; however, in 2017 a group of researchers namely Gaetani et al. [3] wrote a research paper on using block chained database in regular cloud computing environments . This paper was written as a case study of the European SUNFISH project which was proposing a new service Federation as a Service (FaaS) in order to increase data security in various cloud systems [21].

As discussed earlier, Gaetani et al. model has two blockchain layers, the first being permissioned and lightweight to support fast computations or access to data. This is made possible by eliminating the proof of work concept which is responsible it's computational intensiveness [2]. The second layer is more like a typical blockchain database in which the hashed data state of the first layer is sent to it at intervals for permanent irrefutable storage. Together these layers ensured integrity and confidentiality in the cloud computing system.

C. Resource availability in cloud computing models

This is one of the most important security concepts as it involves resources accessible when required [22]. In the cloud computing environment this means that the system should be fault tolerant and there should be a connection from the client to the cloud on demand [8]. Fault tolerance in this case refers to the capability of the system to continue its operations on another server or storage unit despite the failure of the current server [23].

Techniques like redundancy, failover and proximity of servers are used to ensure availability in the cloud. Amazon EC2 cloud uses regions and availability zones to ensure redundancy in the cloud system [24]. This simply involves a failover cluster for the database, which replicates its data and hence eliminating central failure points.

It is evident that availability in cloud computing environments is usually handled by the cloud provider or a trusted third party [8]. However, the introduction of the blockchain to improve cloud security by eliminating the need for trust also presents certain issues of availability in the cloud. A major example of this is Gaetani et al. model that ensures integrity and confidentiality while compromising availability [3]. In this model, even though redundancy is present due to the distributed ledger nature of the blockchain, accessibility of data can easily be hampered by compromising a single participant. This is because the first layer is a permissioned blockchain that requires the signature of every participant of the cloud system to validate an operation.

The limited accessibility of resources involved when using a lightweight blockchain in a cloud computing environment demonstrates a tradeoff between performance and availability. Performance is usually prioritized in the cloud [25], hence in order to use blockchain layers for cloud computing at least one of the layers should be a permissioned blockchain. This is because they perform better than the typical permission-less blockchains like Ethereum [26] but have authentication key-pairs [27], which limits resource availability.

D. Current Research Problem

A cloud federation refers to the pooling together of certain cloud resources by multiple cloud vendors [28]. Companies that use the same data or resources tend to use the cloud federation. Gaetani et al developed an architecture for a cloud federation system with multiple members and is shown in Figure 1. This architecture has two layers; a permissioned blockchain and a permission-less blockchain. A permissioned blockchain as the name implies requires authentication for access. It generally lacks the proof of work consensus (PoW) in order to improve computational speed. This is because basic security features like privacy are ensured by the authentication process. Gaetani used this model as the computational layer due to its low latency. Security features like integrity on this layer were implemented using a timed rotary system called mining rotation [3] in which at intervals each member became a leader that collects the encrypted signatures of the other members including their own signature in order to update data stored in this layer. Hence, modification of data stored here can only be done by the approval of every miner (member).

This security mechanism limited availability of resources to the users and is the motivation for this research problem of ensuring availability in the cloud that uses a layered blockchain database without compromising its existing security properties.

III. Solution Strategy and Implementation

Dew computing has been defined as a type of computing where the local device (on-premises computer) can provide services without relying on the network connectivity [29], [30]. It is usually cooperative with the cloud space. It differs from other computing paradigms such as fog computing as they need extra infrastructure (like routers) to function while Dew just requires a local computing device like a computer you already own. A prominent example of this computing paradigm is in the dropbox storage application found in our local systems.

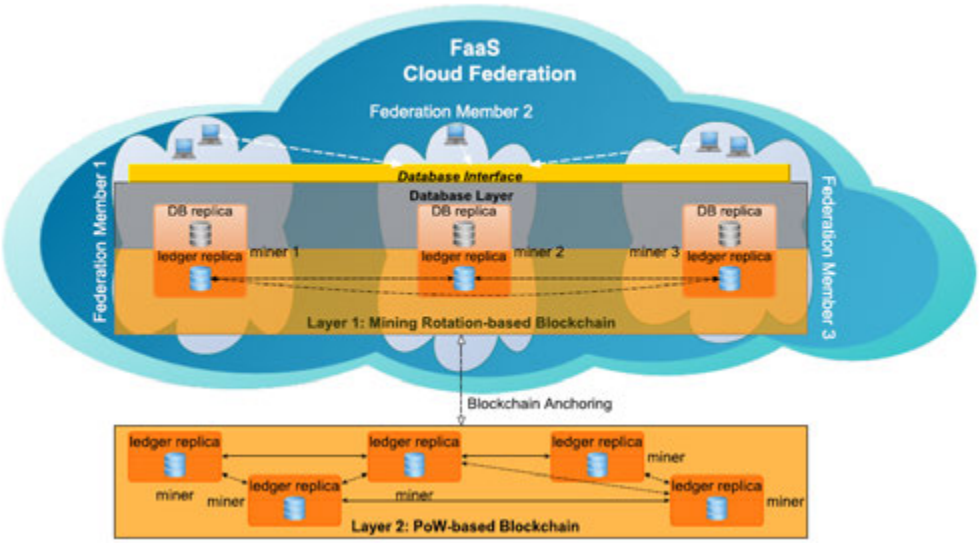


Fig. 1. Gaetani et al. Architecture (This figure is from [3])

A. Proposed Architecture

We propose and design a dew-blockcloud architecture that provisions Cloud resources with availability. This architecture is presented in Figure 2.

Our dew architecture comprises of three major components: Dew Client service, Dew Database management system and Dew server.

- **Dew Client service:** This consists of the services provided by dew server. It includes dew script analyzer that acts as a compiler for the script used to design and alter the dew site [31], synchronization services between the cloud and dew server and identity mapping, which is used for authorization and other security services. Since the cloud users in the model given in figure 1 already have keys, identity mapping here simply means mapping their private keys to the unique ID of their dew servers.
- **Dew Database Management System**: This basically handles management of data storage in the dew architecture. Local copies of cloud data are stored in the on-premises computer and are displayed in form of dew sites [31].
- **Dew server:** The dew server in this case mimics a cloud server, the only difference is that you do not require the internet to access services or data [29]. It is also lightweight and much smaller than the cloud server as expected as it provides services for fewer individuals and stores far less data. The dew server basically uses an on-premise computer as its physical component on which Dew is installed. The major features of our dew server includes the following:
  - Independence.
  - Collaboration.

Independence: This refers to the user being able to obtain services and perform computations without having any access to the cloud environment. In a general sense, it entails the use of cloud services without any internet connection [31]. However, in our architecture, we focused on the user being able to access certain data without the need for requesting data from the first layer of the blockchain database in Figure 1's model.

Collaboration: In simple terms, it outlines the transfer of information between the dew server and the cloud. This information transfer is very important as it keeps the user data updated at each given time. It is also subdivided into two processes namely (1) Upload transfer and (2) Download transfer.

Upload transfer (synchronization): This is a periodical process done at intervals to update data stored in the cloud database with data from the dew server. The dew server sends a request signal RQ to other members of the federated cloud asking for their signatures (approval). Then, it uses these signatures to access the permissioned cloud blockchain database and do the needed updates. It is an automatic periodic process and the request signal is usually sent without the knowledge of the owner of the dew server. This reduces the complexity of the system. It also helps in making local operations carried out by the client run smoothly with little to no latency given that updates in the cloud are done much later by the dew server.

Download transfer (caching): This refers to the obtaining of data from the cloud by the dew server. It is usually done concurrently with a user's access session just like a cache. The dew server uses the authorization to access the first layer obtained by the user (permissioned blockchain

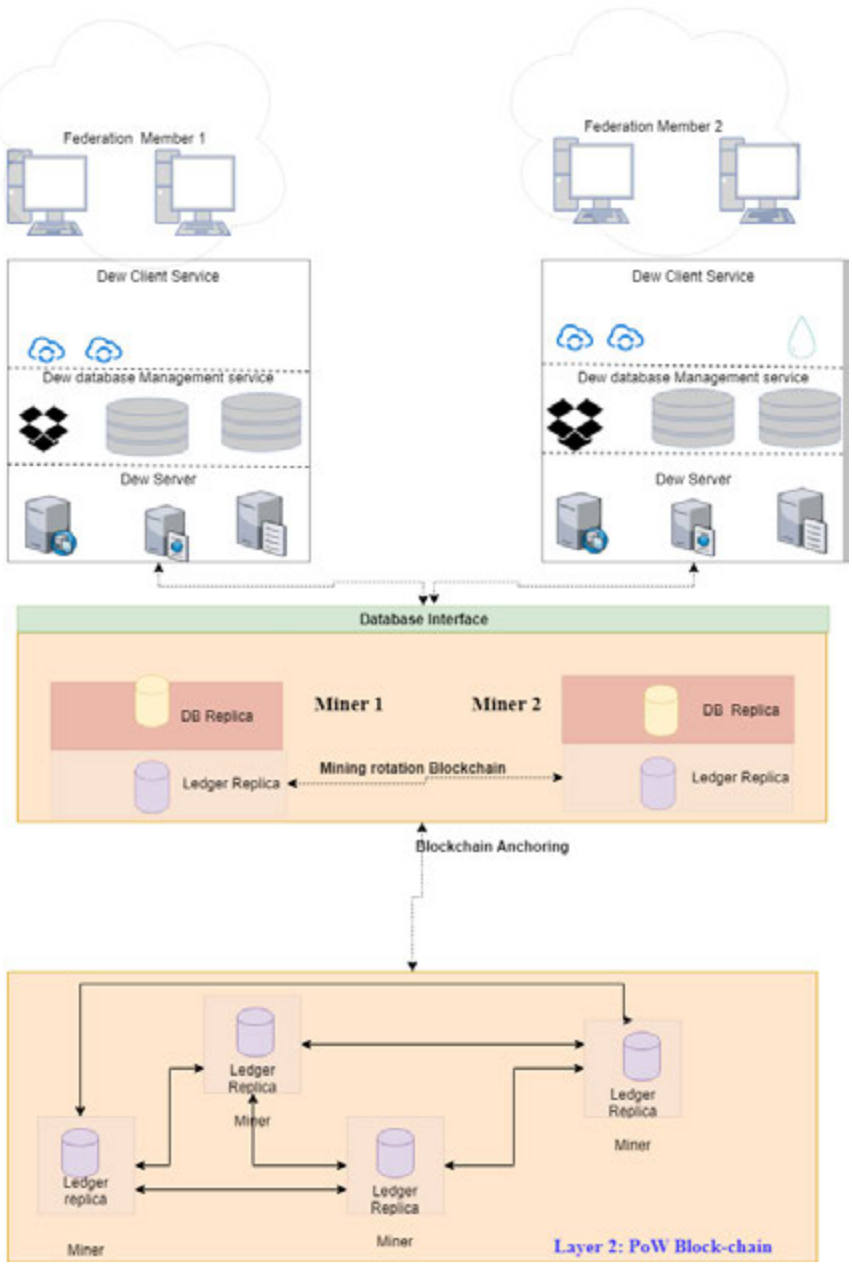


Fig. 2. The Dew-Blockcloud Architecture

database) in order to download required data onto the on-premise computer. The type of data downloaded plays a big part in the validity of subsequent computations. The download of the most recently used data (MRU) is the staple caching style used in dew systems as it always keeps local data on the dew server up-to-date especially in a federated cloud with assets shared by different users/organizations. [32]. However, it is plagued by the high number of cloud accesses as each time a federated data is modified by a different user on the federated cloud, it has to be re-downloaded by the rest of the users. The logical solution to this is to employ another widely used caching algorithm called frequently used data (FRU) caching in order to reduce the number of cloud accesses by the dew server. However, this method raises the problem of data validity in the cache if the corresponding data in the federated cloud is modified by a different user.

We proposed a caching algorithm that solves these problems outlined above by combining MRU and FRU caching styles. For example, consider two trading companies company A and company B that are in a federated cloud. If the data in the federated cloud is personalized i.e company A's employee files or resource allocation files

then most frequently used data is cached, however, if the data is a federated asset such as trading files between company A and Company B (which can also be modified by either of them at any time), then most recently used data is cached. This method retains the caching speed of MRU for the users data without compromising the validity of the cache data when a shared asset that can be modified by both parties is involved. This hybrid caching algorithm is given in Algorithm 1.

Algorithm 1 Hybrid caching algorithm

Result: Most recently used data + frequently used data
initialization: userSession(keys)
while *session is active* **do**
  if *data accessed is user's private data & data frequency > any cache data frequency* **then**
    Replace that cache data with the new more frequently used data
  else
    if *data accessed is federated asset* **then**
      Replace oldest data in cache with the new most recently used data
    end
  end
end

IV. Analysis of our Architecture and Results

The dew-blockcloud architecture we designed in this project is intended to fix the availability issue found in Gaetani et al. model [3] while providing a lot of added benefits. In order to make this evident, we analyse the resource availability of our architecture using some conceptual realistic scenarios and compare it Gaetani's model using the same scenarios.

Consider two federated cloud testbeds; Cloud 1 with Gaetani et al. model [3] as its database and Cloud 2 uses our dew-blockcloud architecture. These cloud environments provide storage for multiple users's pictures as well as editing services. it also stores shared pictures between users.

Scenario 1: A user wants to edit a picture stored on the cloud instantly and post on a social media app . In Gaetani's model given in figure 1, a user cannot perform this operation because the user needs to get approval from other members of the cloud federation before the user accesses any of their pictures. Resource availability here is pretty limited. However, in our dew-blockcloud architecture there is a high probability that the picture has already been saved by the dew server on the user's device through caching, hence the user performs the operation with little to no delay. The dew server then updates the edited picture in the cloud at a later time. Thus, availability here is very high.

Scenario 2: A malicious user wants to delete another user's picture stored on the cloud database. As seen earlier Gaetani's model is pretty tamper proof, so this integrity attack is easily thwarted. Our dew-blockcloud architecture doesn't compromise this integrity feature either, because even if the user deletes the picture in question on their local device, their dew server still needs the approval of other users in order to update the operation in the cloud blockchain database.

A. Other Benefits of the dew-blockcloud architecture

Other expected benefits of the dew-blockcloud architecture includes the following:

- The architecture provides Any Time Any How Accessibility (ATAHA) [29]. This essentially means that one can access most cloud services without the need for an internet connection at any time. This feature results in a system that has very high availability.
- Cost is minimized drastically as little to no infrastructure is needed unlike other systems that require a lot of network components. For example, Google espresso, which uses edge computing for their voice assistant application uses a lot of smart routers and server pools. [33].
- Storage cost is also minimal as size of data stored is far smaller than the stored data in the cloud.
- The architecture is failure tolerant. This is because despite the fact that the dew server is not immune to data loss due to its cache nature, the data can be easily recoverable from the cloud database.

B. Challenges and possible solutions

Outlined below are the challenges faced that could be faced by our dew-blockcloud architecture and possible techniques for solving them.

- There might be a problem of how the cloud service providers can track the service usage of a user when the user is offline (i.e exclusively using the dew server). A dedicated tracking system such as those seen in public internet centers can be installed alongside the dew server to monitor service usage. Another method could be a locking mechanism that is triggered when the user's cloud contract time expires.
- There is a trade-off between installing dew servers on each system in the federation or using a single dew server for several members of the federated cloud. The former promotes performance, speed and ease of security implementation; however, installing for every participant (e.g. every employee in a company) may be expensive and time consuming. The latter could prove advantageous especially for an organization. The method used is specified in the service level agreement document [34]
- There could be excess power usage as the dew server like a typical server is expected to work 24/7. This can be rectified by implementing a stand-by-mode for the dew server.

V. Conclusion

We proposed and designed a dew-blockcloud that is expected to not only provide resource availability without compromising the functions of the Gaetani et al. model [3] but also providing some additional benefits to the system. We also proposed a hybrid MRU/FRU caching algorithm that reduces the amount of cloud accesses by the dew server without

loosing data validity especially on a federated cloud.
We postulated that dew computing might be better than other resource availability systems due to its minimal cost of deployment characterized by the use of local devices. Our architecture also indicates that blockchain might be more compatible with newer computing paradigms than it ever was with cloud computing. Future work will likely be an actual deployment of this architecture on a real time blockchain with proper resources.

ACKNOWLEDGMENT

The second and third authors acknowledge Natural Sciences and Engineering Research Council (NSERC) Canada for partial financial support for this research through Discovery Grants. The first author acknowledges the International Graduate Student Scholarship from Faculty of Graduate Studies, University of Manitoba.

REFERENCES

[1] Solon, "Cambridge analytica closing after facebook data harvesting scandal." [Online]. Available: https://www.theguardian.com/
[2] Nakamoto, 2019. [Online]. Available: https://bitcoin.org/bitcoin.pdf
[3] E. Gaetani, L. Aniello, R. Baldoni, F. Lombardi, A. Margheri, and V. Sassone, "Blockchain-based database to ensure data integrity in cloud computing environments," 2017.
[4] SUNFISH, "Sunfish project." [Online]. Available: http://www.sunfishproject.eu/
[5] J. McCarthy, "Reminiscences on the history of time sharing," *URL: http://www-formal. stanford. edu/jmc/history/timesharing. dvi*, 1983.
[6] P. M. Mell and T. Grance, "The nist definition of cloud computing," 2011.
[7] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, "Xen and the art of virtualization," *ACM SIGOPS Operating Systems Review*, vol. 37, no. 5, p. 164, 2003.
[8] R. L. Krutz and R. D. Vines, *Cloud security: a comprehensive guide to secure cloud computing*. Wiley, 2010.
[9] A. Jasti, P. Shah, R. Nagaraj, and R. Pendse, "Security in multi-tenancy cloud," *44th Annual 2010 IEEE International Carnahan Conference on Security Technology*, 2010.
[10] H. Tianfield, "Security issues in cloud computing," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2012, pp. 1082–1089.
[11] T. Dierks and E. Rescorla, "The transport layer security (tls) protocol version 1.2," 2008.
[12] D. Zissis and D. Lekkas, "Addressing cloud computing security issues," *Future Generation Computer Systems*, vol. 28, no. 3, p. 583–592, 2012.
[13] K. Aissaoui, H. A. Idar, H. Belhadaoui, and M. Rifi, "Survey on data remanence in cloud computing environment," *2017 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, 2017.
[14] M. Alhamad, T. Dillon, and E. Chang, "Sla-based trust model for cloud computing," *2010 13th International Conference on Network-Based Information Systems*, 2010.
[15] X. Liang, J. Zhao, S. Shetty, and D. Li, "Towards data assurance and resilience in iot using blockchain," *MILCOM 2017 - 2017 IEEE Military Communications Conference (MILCOM)*, 2017.
[16] C. Wang, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for data storage security in cloud computing," *2010 Proceedings IEEE INFOCOM*, 2010.
[17] P. K. Sharma, M.-Y. Chen, and J. H. Park, "A software defined fog node based distributed blockchain cloud architecture for iot," *IEEE Access*, vol. 6, p. 115–124, 2018.
[18] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," *Proceedings of the first edition of the MCC workshop on Mobile cloud computing - MCC 12*, 2012.
[19] Z. Ming, S. Yang, Q. Li, D. Wang, M. Xu, K. Xu, and L. Cui, "Blockcloud: A blockchain-based service-centric network stack."
[20] A. Sathiaseelan, L. Wang, A. Aucinas, G. Tyson, and J. Crowcroft, "Scandex," *Proceedings of the 2015 Workshop on Do-it-yourself Networking: an Interdisciplinary Approach - DIYNetworking 15*, 2015.
[21] F. P. Schiavo, V. Sassone, L. Nicoletti, A. Reiter, and B. Suzic, "Faas: Federation-as-a-service: The sunfish cloud federation solution," in *FaaS: Federation-as-a-Service: The SUNFISH Cloud Federation Solution*, 2016.
[22] S. Samonas and D. Coss, "The cia strikes back: Redefining confidentiality, integrity and availability in security." *Journal of Information System Security*, vol. 10, no. 3, 2014.
[23] W. Zhao, P. Melliar-Smith, and L. Moser, "Fault tolerance middleware for cloud computing," *2010 IEEE 3rd International Conference on Cloud Computing*, 2010.
[24] Amazon, 2019. [Online]. Available: https://docs.aws.amazon.com
[25] C. Vecchiola, S. Pandey, and R. Buyya, "High-performance cloud computing: A view of scientific applications," *2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks*, 2009.
[26] G. Wood *et al.*, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum project yellow paper*, vol. 151, pp. 1–32, 2014.
[27] H. Li, Y. Dai, L. Tian, and H. Yang, "Identity-based authentication for cloud computing," *Lecture Notes in Computer Science Cloud Computing*, p. 157–166, 2009.
[28] A. Celesti, F. Tusa, M. Villari, and A. Puliafito, "How to enhance cloud architectures to enable cross-federation," *2010 IEEE 3rd International Conference on Cloud Computing*, 2010.
[29] Y. Wang, "Cloud-dew architecture," *International Journal of Cloud Computing*, vol. 4, 2015.
[30] S. Srivastava, S. K. Tadepalli, Y. Wang, R. K. Thulasiram, and P. Thulasiraman, "Dew text application development," in *Proc. of the 4th International Workshop on Dew Computing*, 2019, pp. 18–29.
[31] P. P. Ray, "An introduction to dew computing: Definition, concept and implications," *IEEE Access*, vol. 6, pp. 723–737, 2018.
[32] L. Fan, P. Cao, J. Almeida, and A. Broder, "Summary cache: a scalable wide-area web cache sharing protocol," *IEEE/ACM Transactions on Networking*, vol. 8, no. 3, pp. 281–293, 2000.
[33] keycdn, "what is google espresso?" 2019. [Online]. Available: https://www.keycdn.com/support/google-espresso
[34] A. Mosallanejad and R. Atan, "Ha-sla: A hierarchical autonomic sla model for sla monitoring in cloud computing," *Journal of Software Engineering and Applications*, vol. 06, no. 03, pp. 114–117, 2013.

THE RELEVANCE OF BLOCKCHAIN WITH DEW COMPUTING: A REVIEW

Rexford Nii Ayitey Sosu*, Chindanur Narendra Babu **, Samuel Akwasi Frimpong ***, John Essuman****
* Ghana Technology University College/Information Systems Departmant, Accra, Ghana
** M.S Ramaiah University of Applied Sciences/Department of Computer Science, Bangalore, India
***Ghana Technology University College/Computer Engineering Department, Accra, Ghana
***Jiansu University/School of Computer Science and Communication Engineering, Zhenjiang, China
****Ghana Technology University College/Mathematics Department, Accra, Ghana
rnasosu@gtuc.edu.gh, narendrababu.et@msruas.ac.in, sfrimpong@gtuc.edu.gh, jessuman@gtuc.edu.gh

***Abstract* - Blockchain after its discovery and application to bitcoin has become a vital platform for data validation and verification across diverse sectors such as healthcare, supply chain, governance, and many others. The era of cloud computing has facilitated the access of one's data on the go making it almost impossible to lose data as your work. However, the use of the cloud demands constant aces to the internet warranting the relevance of the dew computing paradigm with terms to give access to one's data locally without the use of the internet and then updating the document on the cloud based on the internet restoration. In this paper, we seek to conduct a review of the above areas of blockchain and dew computing drawing the need for the integration of both technologies.**

Keywords - blockchain, cloud computing, dew computing, public blockchain, private blockchain, proof of work, proof of stake, depreciated proof of stake.

I. INTRODUCTION

Blockchain is one of the advances, which showed up in the most recent decade and carried a great deal of guarantee with it. Numerous sorts of research are being directed effectively to investigate the full capacities of Blockchain. Some accept that Blockchain is key for a decentralized society. [1][2][3][4]

A. *Blockchain Technology*

Blockchain is being lauded as a mechanical advancement that permits changing how society exchanges and connects [5][6]. This notoriety is specifically owing to its properties of permitting commonly doubting elements to trade budgetary esteem and cooperate without depending on a confided in outsider. A blockchain additionally gives uprightness ensured information stockpiling and permits giving procedure straightforwardness. The ongoing increment in revealed occurrences of reconnaissance and security ruptures trading off clients' protection raise doubt about the present model, in which outsiders gather and control gigantic measures of individual information.

Bitcoin has exhibited in the budgetary space that trusted, auditable processing is conceivable utilizing a decentralized system of companions joined by an open record [7][8][9].

Blockchain is an appropriated database arrangement that keeps up a constantly developing rundown of information records that are affirmed by the hubs taking part in it. The information is recorded in an open record, remembering data for each exchange at any point finished. Blockchain is a decentralized arrangement, which doesn't require any outsider association in the center. The data about each exchange at any point finished in Blockchain is shared and accessible to all hubs. This characteristic makes the framework more straightforward than brought together exchanges including an outsider. Likewise, the hubs in Blockchain are for the most part mysterious, which makes it increasingly secure for different hubs to affirm the exchanges [10][11][12].

The architecture below [13][14] shows the structure of the Blockchain formation where we have the genesis block comprising of contents such as the Hash of current block and its transactional data. The subsequent blocks bearing the Hash of the previous block, the Hash of current block and its transactional data then follow this block.

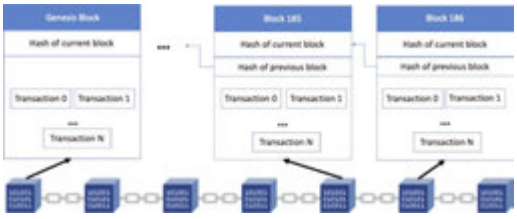


Figure. 1. Blockchain, Data Chain Architecture

B. *Dew Computing Paradigm*

The technological progress of widespread smart devices as well as universal networks, and the associated innovative network services, cloud computing paradigm is having certain challenges based on its centralized storage model as well as processing in adjusting and being applied to various types of applications as well as technological setups [15] [16][17]. Dew Computing architectural image is illustrated in figure 2 below [18].

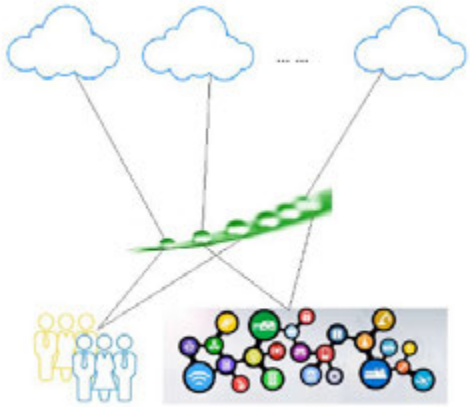


Figure 2. The Architectural Structure of Dew Computing

Advancements such as artificial intelligence, Internet of things and wireless communication devices have caused the significance of the cloud-computing paradigm to either increase slowly or not at all. For this cause, many believe that the era of post-cloud computing is therefore upon us. This era spans across some models related to cloud computing such as Dew Computing, Fog Computing, Edge Computing as well as Cloudlet. [19][20][21]

II. RELATED WORKS

A. BLOCKCHAIN ALGORITHMS

A.1 *Proof of Work (PoW) Algorithm*

In the concept of blockchain generation, every data added would be tested for its data validity, which is usually referred to as the process of mining or also known as Proof of Work. [22] [23] devise optimal framework based on PoW for adversarial strategies for selfish mining and double-spending while considering real-world constraints such as different block sizes, network propagation, block generation intervals, information and the impact of eclipse attacks. [24] Proposed a PoW based consensus distributed ledger system to prevent malicious attacks utilizing certain variations. In their research [25] created a formalized proof-of-work algorithm by the introduction of a verified primitive called (NIPoPoWs) Non-Interactive Proofs of Proof-of-Work.[26][27]

A.2 *Proof of Stake (PoS) Algorithm*

In the later Proof of State arranges, the arrangement looking is evacuated, and the square chiefs are never again chosen by computational force. The main Proof of Stakes arrange, Peercoin, was created as a PoS consensus mechanism aimed at decreasing the computational conditions of Proof of Work. Members with higher coinage, i.e., a result of system tokens and their holding time, have higher opportunities to be chosen. [28][29][30]

Rather than consuming a great deal of energy in search of the procedure using the Proof of Work method, a primary will be chosen dependent on its stakes to play out the mining procedure and add another block to the chain. [31][32]

A.3 *Delegated Proof of Stake (DPoS) Algorithm*

In their paper [33] established that in 2014, Larimer D proposed a DPoS agreement system that could understand square second-level check, which could meet a wide scope of business needs.

In the DPoS agreement instrument, each hub holding tokens is a competitor hub, and every hub chooses a few operator hubs by casting a ballot, and the specialist hubs alternate to create and confirm the squares as indicated by the setup plan. [34][35]

While the democratic model can clarify the vague set naturally, to make the operator hub determination of the accord system of DPoS like the democratic model, in reality, we proposed the improvement of DPoS agreement component dependent on Vague set.[36][37][38]

B. BLOCKCHAIN CATEGORIES AND CLASSIFICATIONS

B.1 *Consortium Blockchain*

Consortium blockchains are some of the time considered a different assignment from private blockchains. Examples of consortium blockchains include Quorum and R3 Corda. [39]

Regularly, consortium blockchain are related to big business use, where a gathering of associations working together to use blockchain innovation to improve organizations. Consortium blockchain can be comprehended when it is contrasted and open and private blockchain. [40]

A consortium blockchain isn't conceded to a solitary element as a private blockchain; rather, it is allowed to a gathering of affirmed people.

Besides, consortium blockchain is a gathering of pre-characterized hubs on the system. [41]

B.2 *Private Blockchain*

Private blockchains are more concentrated than Public blockchains and are significant for undertakings need to team up and share information however don't need their delicate business information unmistakable on an open blockchain. Examples of private blockchain include Monax and Multichain.[42][43][44]

Members need agree to join the systems Transactions are private and are just accessible to environment members that have been offered consent to join the system. [45][46]

B.3 *Public Blockchain*

Public blockchains are open source applications that intend to be completely decentralized, with nobody individual or substance controlling which exchanges are recorded in the blockchain or the request wherein they are prepared. Examples of public blockchains include Etherum and Bitcoin.

They can be profoundly restriction safe, since anybody is available to join the system, paying little heed to area, nationality, and so forth.

All exchanges that happen on public blockchains are completely straightforward, implying that anybody can inspect the exchange subtleties.[47][48][49]

TABLE I. DIFFERENCES AMONG BLOCKCHAIN CATEGORIES

Consortium	***Private***	***Public***
Multiple Algorithms of consensus	Multiple Algorithm consensus	Consensus by Proof of Work
Recognized participants from multiple organizations	Recognized participants from one organization	Many unrecognized participants
Writes require consensus of several participants	Write permissions centralized	Writes by all participants
Reads may be public or restricted	Reads may be public or restricted	Reads by all participants

C. BLOCKCHAIN APPLICATIONS AND THEIR ARCHITECTURES

C.1 *Blockchain in Internet of Vehicles (BIoV)*

Finding a solitary confided in substance to store and disseminate such messages can be testing, and vehicles may not be slanted to take an interest (e.g., age and appropriation of declaration messages) except if they can profit by such investment. While the vast majority of the vehicular messages just should be shared by close by vehicles, a few messages (e.g., declaration messages) may should be all the more extensively circulated, for instance to vehicles in a more extensive area. To empower/boost cooperation, vehicles that steadfastly communicate the declaration messages and additionally add to the square age will be compensated by some digital currency.[50][51][52]

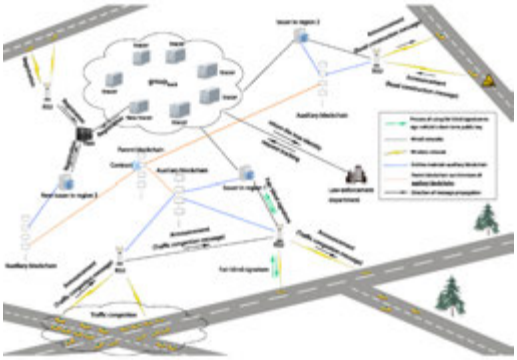


Figure. 3. A Proposed Blockchain Architecture on a Secured Data Sharing System

Figure 3. Illustrates how the proposed blockchain architecture on a secured data sharing system for Internet of Vehicles comprises of backers, vehicles, traffic the board authority, tracers and law authorization office and side of the road units.[53][54][55]

C.2 *Blockchain in the Sphere of Healthcare*

As of late, there have been expanding calls for human services suppliers to give controls to patients over their own wellbeing records. All things considered, security issues concerning how extraordinary medicinal services suppliers trade human services data have caused a failure in the sending of such frameworks. The capacity to trade information safely is significant so new borderless coordinated social insurance administrations can be given to patients. [56][57][58]

Because of its decentralized nature, blockchain innovation is an appropriate driver for the genuinely necessary move towards incorporated medicinal services, giving new bits of knowledge and tending to a portion of the fundamental difficulties of numerous human services regions. Blockchain permits medicinal services suppliers to record and oversee distributed exchanges through a system without a focal power. [59][60][61]

C.3 *Blockchain in Media*

In their paper [62] explored characterizing blockchain within digital subjects such as, media and interactive media as media blockchain. There are numerous examinations about media blockchain. These investigations guarantee the trustworthiness, authentication, or control of digital substance, and are characterized into two classes: One is content security that implies verifying substance utilizing blockchain. The other is the substance accurate data that implies tracking and working with the correct data by applying blockchain. The image in Figure 4 shows a blockchain architecture in the social media environment.[63][64]



Figure. 4. Blockchain in Social Media Architecture

With the execution of blockchain innovation, the technique for correspondence in the business would get decentralized. Under this framework, specialists and their crowds would then have the option to can communicate straightforwardly without the previously mentioned middle people. Since blockchain has the option to track all communications of a specific resource, approved audiences can evidently follow all the relative exercises, for example, utilization of substance, move of proprietorship, and keen agreements related with the substance. Content makers could deal with the rundown of approved watchers for their own needs without being influenced by the principles of an outsider. [65] [66][67]

D. BLOCKCHAIN CHALLENGES

In a blockchain, every node has to perform a comparable task for the verification of each transaction at the same time; thus, obtaining a high computation cost. In their paper [68][69] explained that this complicates the challenge of validation and verification of different methods across the nodes in the diverse networks, as well as moderation of network attacks. Therefore in order to avoid performance degradation, requiring the need to design efficient performance optimization. [70][71][72][73]

TABLE II. DISTINCTIONS BETWEEN CLOUD, DEW, EDGE AND FOG COMPUTING

Parameters	**Cloud Computing**	**Dew Computing**	**Edge Computing**	**Fog Computing**
Hardware	Scalable Capabilities	Very Limited Capabilities	Limited Capabilities	Scalable Capabilities
Service Location	Internet Based	Edge Network Based	Edge Network Based	Internet Based
Location Awareness	No	Yes	Yes	No
User Experience	Very Normal	Highly Satisfactory	Good	Normal
Target Users	Internet Users	Purely Mobile Users	Semi Mobile Users	Internet Users
Internet Dependency	Every Access Time	Not Essential	Every Access Time	Every Access Time
Geo-Distribution	Centralized	Highly Distributed	Distributed	Semi Centralized
Synchronization Feature	Not Essential	Always Essential	Not Essential	Not Essential
Deployment Scenario	Large Enterprises	Smartphone, Laptop, Pc	Gateway, Router	Small Medium Enterprises
Latency	Very High	Negligible	Low	High

E. HOW DEW COMPUTING DIFFERS FROM CLOUD, EDGE AND FOG COMPUTING

The below table illustrates the difference between Dew Computing and other computing paradigms such as Cloud, Edge and Fog Computing with regards certain parameters as listed in the table. [74][75][76]

III. THE NEED TO INTEGRATE DEW COMPUTING WITH BLOCKCHAIN

In their paper, [77] explained that blockchain data has to exist in every node and the data amount increases with time. Even though some approaches have been proposed to reduce the data size of blockchain clients, but these clients do not have the status of full nodes. Therefore Dewblock was proposed by [77] bringing a new approach that the data size of a client is reduced and the features of a full node are still kept. With Dewblock, each blockchain user needs to deploy a cloud server to a cloud service. While a client is light-weighted and is appropriately controlled on a mobile device or a personal device, the client works with a distant cloud server to act as a full node. However, a dew client operates independently to perform blockchain activities; it also cooperates with the cloud server to sustain the integrity of the whole blockchain network. Based on the wide use of cloud services and the rise of Internet of things by individual users, including blockchain users, making blockchain in dew computing feasible and affordable from an economical and technical point of view. [78][79][80][81][82]

A. A PROPOSED FRAMEWORK OF DEW COMPUTING AND BLOCKCHAIN INTEGRATION

Blockchains bear the capacity to be implemented across a large scope of, with many transactions being streamlined to accommodate blockchain integration. It is therefore critical to have a universal blockchain that supports user architecture enabling these users to deploy information across mobile platforms as well as personal devices while they possessing the entire capabilities such that full nodes do base on Dew computing architecture as well as principles. [77][83][84]

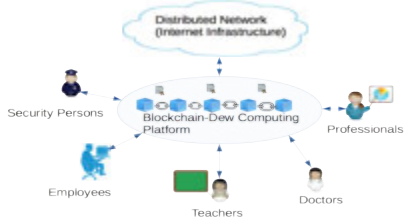


Figure. 5. A Proposed Blockchain Dew Computing Architecture

IV. CONCLUSION

In this paper, the authors reviewed other related and relevant research on the concept of blockchain and dew computing highlighting the importance of blockchain integration with dew computing with areas concentrated on topics such as introduction to blockchain technology, blockchain consensus algorithms, blockchain categories, and classifications, blockchain applications and their simple architectures, blockchain challenges, how dew computing differs from other computing paradigms such as cloud computing, edge computing, fog computing and a proposed architecture representing an integrated environment of blockchain with due computing.

V. RECOMMENDATION

The authors recommend a further independent study be conducted in the fields of blockchain and dew computing to implement proposed integration highlighted in this review study.

REFRENCES

[1] Q. E. Abbas and J. Sung-Bong, "A Survey of Blockchain and Its Applications," *1st Int. Conf. Artif. Intell. Inf. Commun. ICAIIC 2019*, pp. 1–3, 2019.

[2] J. Z. Garrod, "The Real World of the Decentralized Autonomous Society," *tripleC Commun. Capital. Crit. Open Access J. a Glob. Sustain. Inf. Soc.*, vol. 14, no. 1, Feb. 2016.

[3] M. Atzori, "Blockchain technology and decentralized governance: Is the state still necessary?," *J. Gov. Regul.*, 2017.

[4] A. Essén and A. Ekholm, "Centralization vs. Decentralization on the blockchain in a health information exchange context," in *Digital Transformation and Public Services: Societal Impacts in Sweden and Beyond*, 2019.

[5] K. Wüst and A. Gervais, "Do you need a Blockchain ?," no. i, 2018.

[6] M. J. Casey and P. Vigna, "In blockchain we trust - MIT Technology Review," *MIT Technology Review*, 2018. .

[7] G. Zyskind, O. Nathan, and A. S. Pentland, "Decentralizing privacy: Using blockchain to protect personal data," *Proc. - 2015 IEEE Secur. Priv. Work. SPW 2015*, pp. 180–184, 2015.

[8] I. Zikratov, A. Kuzmin, V. Akimenko, V. Niculichev, and L. Yalansky, "Ensuring data integrity using blockchain technology," in *Conference of Open Innovation Association, FRUCT*, 2017.

[9] M. (Google) Crosby, N. (Yahoo), P. (Yahoo) Pattanayak, S. (Samsung R. A. Verma, and V. (Fairchild S. Kalyanaraman, "Blockchain Technology Explained - Beyond Bitcoin," *Sutardja Cent. Entrep. Technol. Tech. Rep.*, 2015.

[10] J. Yli-Huumo, D. Ko, S. Choi, S. Park, and K. Smolander, "Where is current research on Blockchain technology? - A systematic review," *PLoS One*, vol. 11, no. 10, pp. 1–15, 2016.

[11] B. T. Rao, V. Lakshman Narayana, V. Pavani, and P. Anusha, "Use of Blockchain in Malicious Activity Detection for Improving Security," *Int. J. Adv. Sci. Technol.*, 2020.

[12] Z. Dong, F. Luo, and G. Liang, "Blockchain: a secure, decentralized, trusted cyber infrastructure solution for future energy systems," *J. Mod. Power Syst. Clean Energy*, 2018.

[13] J. Zhang, "Deploying Blockchain Technology in the Supply Chain," in *Blockchain and Distributed Ledger Technology (DLT) [Working Title]*, IntechOpen, 2019.

[14] S. Kushch, Y. Baryshev, and S. Ranise, "Blockchain tree as solution for distributed storage of personal id data and document access control," *Sensors*

(Switzerland), 2020.

[15] Y. Zhou, D. Zhang, and N. Xiong, "Post-cloud computing paradigms: A survey and comparison," *Tsinghua Sci. Technol.*, vol. 22, no. 6, pp. 714–732, 2017.

[16] Y. Wang, "Definition and categorization of dew computing," *Open J. Cloud Comput.*, 2016.

[17] D. E. Fisher and S. Yang, "Doing More with the Dew: A New Approach to Cloud-Dew Architecture," 2016.

[18] Y. Wang, "Definition and Categorization of Dew Computing," *Open J. Cloud Comput.*, vol. 3, no. 1, pp. 1–7, 2016.

[19] Y. Wang, "What is Post-cloud Computing ?," no. November, 2018.

[20] K. Dolui and S. K. Datta, "Comparison of edge computing implementations: Fog computing, cloudlet and mobile edge computing," in *GIoTS 2017 - Global Internet of Things Summit, Proceedings*, 2017.

[21] H. Elazhary, "Internet of Things (IoT), mobile cloud, cloudlet, mobile IoT, IoT cloud, fog, mobile edge, and edge emerging computing paradigms: Disambiguation and research directions," *Journal of Network and Computer Applications*. 2019.

[22] I. G. A. K. Gemeliarana and R. F. Sari, "Evaluation of proof of work (POW) blockchains security network on selfish mining," *2018 Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2018*, pp. 126–130, 2018.

[23] A. Gervais, G. O. Karame, K. Wüst, V. Glykantzis, H. Ritzdorf, and S. Capkun, "On the security and performance of Proof of Work blockchains," in *Proceedings of the ACM Conference on Computer and Communications Security*, 2016.

[24] S. Dey, "Securing Majority-Attack in Blockchain Using Machine Learning and Algorithmic Game Theory: A Proof of Work," in *2018 10th Computer Science and Electronic Engineering Conference, CEEC 2018 - Proceedings*, 2019.

[25] A. Kiayias, A. Miller, and D. Zindros, "Non-Interactive Proofs of Proof-of-Work," in *Financial Cryptography and Data Security*, 2020.

[26] A. Covaci, S. Madeo, P. Motylinski, and S. Vincent, "NECTAR: Non-interactive smart contract protocol using blockchain technology," in *Proceedings - International Conference on Software Engineering*, 2018.

[27] R. Zhang, R. Xue, and L. Liu, "Security and privacy on blockchain," *ACM Comput. Surv.*, 2019.

[28] C. T. Nguyen, D. T. Hoang, D. N. Nguyen, D. Niyato, H. T. Nguyen, and E. Dutkiewicz, "Proof-of-Stake Consensus Mechanisms for Future Blockchain Networks: Fundamentals, Applications and Opportunities," *IEEE Access*, vol. 7, pp. 85727–85745, 2019.

[29] J. Spasovski and P. Eklund, "Proof of Stake Blockchain," 2017.

[30] D. K. Tosh, S. Shetty, X. Liang, C. Kamhoua, and L. Njilla, "Consensus protocols for blockchain-based data provenance: Challenges and opportunities," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2017*, 2017.

[31] W. Li, S. Andreina, J. M. Bohli, and G. Karame, "Securing proof-of-stake blockchain protocols," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017.

[32] P. Gazi, A. Kiayias, and D. Zindros, "Proof-of-stake sidechains," in *Proceedings - IEEE Symposium on Security and Privacy*, 2019.

[33] G. Xu, Y. Liu, and P. W. Khan, "Improvement of the DPoS Consensus Mechanism in Blockchain Based on Vague Sets," *IEEE Trans. Ind. Informatics*, vol. PP, no. 8, pp. 1–1, 2019.

[34] F. Yang, W. Zhou, Q. Wu, R. Long, N. N. Xiong, and M. Zhou, "Delegated Proof of Stake With Downgrade: A Secure and Efficient Blockchain Consensus Algorithm With Downgrade Mechanism," *IEEE Access*, 2019.

[35] Q. Deng, "Blockchain Economical Models, Delegated Proof of Economic Value and Delegated Adaptive Byzantine Fault Tolerance and their implementation in Artificial Intelligence BlockCloud," *J. Risk Financ. Manag.*, 2019.

[36] T. Do, T. Nguyen, and H. Pham, "Delegated proof of reputation: A novel blockchain consensus," in *ACM International Conference Proceeding Series*, 2019.

[37] X. Fan and Q. Chai, "Roll-DPos: A randomized delegated proof of stake scheme for scalable blockchain-based Internet of Things systems," in *ACM International Conference Proceeding Series*, 2018.

[38] T. Zhou, X. Li, and H. Zhao, "DLattice: A Permission-Less Blockchain Based on DPoS-BA-DAG Consensus for Data Tokenization," *IEEE Access*, 2019.

[39] Manav Gupta *et al.*, "A Survey of Blockchain Security Issues and Challenges," *Int. J. Netw. Secur.*, vol. 1919, no. 55, pp. 653–659, 2017.

[40] "What Different Types of Blockchains are There? | Dragonchain - Blockchain as a Service." [Online]. Available: https://dragonchain.com/blog/differences-between-public-private-blockchains/. [Accessed: 25-Feb-2020].

[41] "Everything You Need to Know About Public, Private, and Consortium Blockchain." [Online]. Available: https://medium.com/swlh/everything-you-need-to-know-about-public-private-and-consortium-blockchain-54821c159c7a. [Accessed: 25-Feb-2020].

[42] A. Kosba, A. Miller, E. Shi, Z. Wen, and C. Papamanthou, "Hawk: The Blockchain Model of Cryptography and Privacy-Preserving Smart Contracts," in *Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016*, 2016.

[43] S. Pongnumkul, C. Siripanpornchana, and S. Thajchayapong, "Performance analysis of private blockchain platforms in varying workloads," in *2017 26th International Conference on Computer Communications and Networks, ICCCN 2017*, 2017.

[44] Y. Hao, Y. Li, X. Dong, L. Fang, and P. Chen, "Performance Analysis of Consensus Algorithm in Private Blockchain," in *IEEE Intelligent Vehicles Symposium, Proceedings*, 2018.

[45] D. E. O'Leary, "Configuring blockchain architectures for transaction information in blockchain consortiums: The case of accounting and supply chain systems," *Intell. Syst. Accounting, Financ. Manag.*, 2017.

[46] S. Rouhani and R. Deters, "Performance analysis of ethereum transactions in private blockchain," in *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, 2018.

[47] Bitfury Group and J. Garzik, "Public versus Private Blockchains. Part 2: Permissionless Blockchains," *Bitfury*, 2015.

[48] R. Lai and D. Lee Kuo Chuen, "Blockchain-From Public to Private," in *Handbook of Blockchain, Digital Finance, and Inclusion*, 2018.

[49] V. Buterin, "Ethereum State Transition Function," *Etherum*, 2018.

[50] R. Grinberg, "Bitcoin: An Innovative Alternative Digital Currency," *Hast. Sci. Technol. Law J.*, 2012.

[51] BIS, “Proceeding with caution - a survey on central bank digital currency,” *Bank Int. Settlements*, 2019.

[52] V. Sapovadia, “Financial Inclusion, Digital Currency, and Mobile Technology,” in *Handbook of Blockchain, Digital Finance, and Inclusion*, 2018.

[53] L. Zhang *et al.*, “Blockchain based secure data sharing system for Internet of vehicles: A position paper,” *Veh. Commun.*, vol. 16, pp. 85–93, 2019.

[54] J. Contreras-Castillo, S. Zeadally, and J. A. Guerrero-Ibanez, “Internet of Vehicles: Architecture, Protocols, and Security,” *IEEE Internet Things J.*, 2018.

[55] W. Xu *et al.*, “Internet of vehicles in big data era,” *IEEE/CAA J. Autom. Sin.*, 2018.

[56] I. Radanović and R. Likić, “Opportunities for Use of Blockchain Technology in Medicine,” *Appl. Health Econ. Health Policy*, 2018.

[57] P. Zhang, D. C. Schmidt, J. White, and G. Lenz, “Blockchain Technology Use Cases in Healthcare,” in *Advances in Computers*, 2018.

[58] H. D. Zubaydi, Y.-W. Chong, K. Ko, S. M. Hanshi, and S. Karuppayah, “A Review on the Role of Blockchain Technology in the Healthcare Domain,” *Electronics*, vol. 8, no. 6, p. 679, 2019.

[59] R. N. A. Sosu, K. Quist-Aphetsi, and L. Nana, “A Decentralized Cryptographic Blockchain Approach for Health Information System,” *Int. Conf. Comput. Comput. Model. Appl.*, pp. 120–1204, 2019.

[60] T. McGhin, K. K. R. Choo, C. Z. Liu, and D. He, “Blockchain in healthcare applications: Research challenges and opportunities,” *Journal of Network and Computer Applications*. 2019.

[61] B. Arunkumar and G. Kousalya, “Blockchain-Based Decentralized and Secure Lightweight E-Health System for Electronic Health Records,” in *Advances in Intelligent Systems and Computing*, 2020.

[62] S. Cho and C. Jeong, “A blockchain for media: Survey,” *ICEIC 2019 - Int. Conf. Electron. Information, Commun.*, pp. 1–2, 2019.

[63] “6 Trends Showing How Blockchain is Changing Social Media.” [Online]. Available: https://bitnewstoday.com/news/6-trends-showing-how-blockchain-is-changing-social-media/. [Accessed: 11-Feb-2020].

[64] A. Chakravorty and C. Rong, “Ushare: User controlled social media based on blockchain,” in *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication, IMCOM 2017*, 2017.

[65] “Blockchain: A New Solution for the Media and Entertainment Industry - By 8 Decimal Capital.” [Online]. Available: https://hackernoon.com/blockchain-a-new-solution-for-the-media-and-entertainment-industry-f8c83ad3dc07. [Accessed: 11-Feb-2020].

[66] V. Morabito, *Business innovation through blockchain: The B3 perspective*. 2017.

[67] D. W. E. Allen, C. Berg, B. Markey-Towler, M. Novak, and J. Potts, “Blockchain and the evolution of institutional technologies: Implications for innovation policy,” *Res. Policy*, 2020.

[68] S. Aggarwal, R. Chaudhary, G. S. Aujla, N. Kumar, K. K. R. Choo, and A. Y. Zomaya, “Blockchain for smart communities: Applications, challenges and opportunities,” *J. Netw. Comput. Appl.*, vol. 144, no. June, pp. 13–48, 2019.

[69] E. Heilman, A. Kendler, A. Zohar, and S. Goldberg, “Eclipse attacks on Bitcoin’s peer-to-peer network,” in *Proceedings of the 24th USENIX Security Symposium*, 2015.

[70] P. Thakkar, S. Nathan, and B. Viswanathan, “Performance benchmarking and optimizing hyperledger fabric blockchain platform,” in *Proceedings - 26th IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, MASCOTS 2018*, 2018.

[71] R. Wang, K. Ye, and C. Z. Xu, “Performance Benchmarking and Optimization for Blockchain Systems: A Survey,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019.

[72] Z. Zheng, S. Xie, H. N. Dai, X. Chen, and H. Wang, “Blockchain challenges and opportunities: A survey,” *Int. J. Web Grid Serv.*, 2018.

[73] A. Alketbi, Q. Nasir, and M. A. Talib, “Blockchain for government services-Use cases, security benefits and challenges,” in *2018 15th Learning and Technology Conference, L and T 2018*, 2018.

[74] P. P. Ray, “Dew Computing : A New Era of Computing Implying Minimization Over Internetwork Backhaul,” vol. 14, no. 1, pp. 53–57, 2019.

[75] K. Skala, D. Davidovic, E. Afgan, I. Sovic, and Z. Sojat, “Scalable Distributed Computing Hierarchy: Cloud, Fog and Dew Computing,” *Open J. Cloud Comput.*, 2015.

[76] P. Kukreja and D. Sharma, “A Detail Review on Cloud, Fog and Dew Computing,” *Int. J. Sci. Eng. Technol. Res.*, 2016.

[77] Y. Wang, “Dewblock : A Blockchain System Based on Dew Computing,” *Proc. 3rd Int. Work. Dew Comput.*, vol. 3, pp. 34–38, 2018.

[78] S. Ristov, K. Cvetkov, and M. Gusev, “Implementation of a horizontal scalable balancer for dew computing services,” *Scalable Comput.*, 2016.

[79] Y. Wang and D. LeBlanc, “Integrating SaaS and SaaP with dew computing,” *Proc. - 2016 IEEE Int. Conf. Big Data Cloud Comput. BDCloud 2016, Soc. Comput. Networking, Soc. 2016 Sustain. Comput. Commun. Sustain. 2016*, pp. 590–594, 2016.

[80] H. M. Patel, R. R. Chaudhari, K. R. Prajapati, and A. A. Patel, “The Interdependent Part of Cloud Computing: Dew Computing,” in *Lecture Notes in Networks and Systems*, 2018.

[81] H. Patel and K. Suthar, “A novel approach for securely processing information on dew sites (Dew computing) in collaboration with cloud computing: An approach toward latest research trends on Dew computing,” *2017 Nirma Univ. Int. Conf. Eng. NUiCONE 2017*, vol. 2018-Janua, pp. 1–6, 2018.

[82] P. P. Ray, D. Dash, and D. De, “Internet of things-based real-time model study on e-healthcare: Device, message service and dew computing,” *Comput. Networks*, 2019.

[83] Y. Wang, “Cloud-dew architecture,” *Int. J. Cloud Comput.*, 2015.

[84] Y. Wang, “A Blockchain System with Lightweight Full Node Based on Dew Computing,” *Internet of Things*, 2020.









KONČAR

www.koncar.hr

Energija koja pokreće Hrvatsku



HEP
Više od struje

sponsors

golden sponsors zlatni sponzori

HEP

KONČAR

bronze sponsor brončani sponzor

STORM GRUPA

sponsors sponzori

INFODOM

A1

m.TEH mjerne tehnologije d.o.o.

Zavod SDT
Institute for Superior Digital Technology

SELMET

nomen d.o.o.
MES and ERP Systems since 1994

MIPRO
Kružna 8/II, HR-51000 Rijeka, Croatia

phone: +385 51 423 984
mipro@mipro.hr • www.mipro.hr